SIZE OF CORPORA AND COLLOCATIONS: THE CASE OF RUSSIAN

Maria KHOKHLOVA
St Petersburg State University

Vladimir BENKO
Slovak Academy of Sciences

Khokhlova, M., Benko, V. (2020): Size of corpora and collocations: the case of Russian. Slovenščina 2.0, 8(2): 58–77

DOI: https://doi.org/10.4312/slo2.0.2020.2.58-77

With the arrival of information technologies to linguistics, compiling a large corpus of data, and of web texts in particular, has now become a mere technical matter. These new opportunities have revived the question of corpus volume that can be formulated in the following way: are larger corpora better for linguistic research or, more precisely, do lexicographers need to analyze bigger amounts of collocations? The paper deals with experiments on collocation identification in low-frequency lexis using corpora of different volumes (1 million, 10 million, 100 million and 1.2 billion words). We have selected low-frequency adjectives, nouns and verbs in the Russian Frequency Dictionary and tested the following hypotheses: 1) collocations in low-frequency lexis are better represented by larger corpora; 2) frequent collocations presented in dictionaries have low occurrences in small corpora; 3) statistical measures for collocation extraction behave differently in corpora of different volumes. The results prove the fact that corpora of under 100 M are not representative enough to study collocations, especially those with nouns and verbs. MI and Dice tend to extract less reliable collocations as the corpus volume extends, whereas t-score and Fisher's exact test demonstrate better results for larger corpora.

Keywords: collocations, Russian corpora, corpus size, corpus linguistics, statistical measures

1 INTRODUCTION

Over the past 10 years, corpora have dramatically increased in size, giving lexicographers much more data than ever before. At the same time, however, this has brought up the question whether we really need those amounts of texts or we can be satisfied with less. The issue is not that simple: corpora, on the one hand, are expected to attest such units by generating a sufficient number of examples; on the other hand, lexicographers and language users should not be overloaded with large bulks of examples.

The size of corpora is also relevant when applied to the task of describing collocability. Is there any correlation between the size of the corpus and the extracted collocations? Can we find more collocations in larger corpora?

We would like to answer the following question: What would be the benefit of using larger corpora? In our study, we analyze the behaviour of Russian collocations using corpora of different volumes. The aim of the paper is threefold. First, to conduct a case study of low-frequency lexemes and analyze their collocations. Secondly, to investigate a number of frequent collocations presented in several dictionaries. Thirdly, to apply statistical measures to collocation extraction from corpora and to interpret possible interrelation between the results and volume.

2 BACKGROUND

The issue of data volume is of importance. For a long time, the amount of data was objectively limited by technical capacities. The Brown corpus comprised 1 million words, the British National Corpus (BNC) amounted to 100 million words, the Russian National Corpus (RNC) has more than 600 million words. The volumes of newly compiled Giga-word corpora can exceed dozens of billions of words.

Linguists understand volume as a concept in different ways. Earlier, a compilation of frequency dictionaries was associated with the question of what amount of data would suffice to describe most frequent lexical units in a language. This question is also relevant in the context of sample reliability or in the context of (foreign) language learning, i.e. what is the minimal amount of lexical units – and, hence, the minimal corpus volume – that students should memorize to learn a language.

Speaking about corpora as samples from larger populations we can mention that the Russian frequency dictionary by Steinfeld (1963) required a 400-thousand-word sample, whereas dictionaries compiled by Zasorina (1977) and Lenngren (1993) are based on a 1 million-word sample; the new dictionary by Lyashevkaya and Sharoff (2009) features a sample of approximately 100 million words. It should be noted that Piotrowski et al. (1977) showed that 1600–1700 most frequent words can be reliably described using a sample of 400 thousand words.

Different works discuss the question of how large a corpus should be. This question is especially crucial in the studies of rare words and word combinations. Sinclair (2005) rightly points out that the occurrences of two or more words are far less frequent than ones of a single word. There are not too many works dealing with the ideal volume of texts required to search collocations. Brysbaert and New (2009) discuss the sufficient corpus volume depending on word frequency distinguishing between high- and low-frequency lexis. Piperski (2015) performs a case study of the same words in two corpora of different sizes, namely the main subcorpus from RNC (230 million words) and ruTenTen (14.5 billion words). The author claims that corpora cannot provide evidence for non-existence of collocations but they can be used to prove their existence. And in this case, even a single example in a corpus is enough.

Finding suitable collocation candidates is quite popular in linguistic research and statistical association measures are widely used for this task. They have their practical application to collocation selection and identification adopted in corpus tools. The dependency between the behaviour of association measures and corpus size was the main focus of a number of research studies. Daudaravičius (2008, p. 650) mentions that “the values of MI grow together with the size of a corpus, while the Dice score is not sensitive to the corpus size and score values are always between 0 and 1”. Rychly (2008) proposes logDice as the measure that is not affected by the size of the corpus and takes into account only frequency of a node and of a collocate. It can be used for collocation extraction from large corpora and is successfully implemented in Sketch Engine (Kilgarriff et al., 2014). Also relevant is the study by Evert et al. (2017) who evaluated not only association measures but also various corpora, co-occurrence contexts and frequency thresholds applied to automatic

collocation extraction and thus tuning statistical methods. The results show that sufficiently large Web corpora (exceeding 10 billion words) perform similarly or even better than the carefully sampled BNC.

Taking these findings into account, a new question is to be considered: how do corpora of different sizes represent multi-word expressions or collocations? In our paper, we analyze quantitative properties of collocations that were found in corpora of different sizes and present some findings on low-frequency collocations.

3 METHODOLOGY

Our previous experiments showed that high-frequency nouns (Khokhlova, 2017) and their ranking positions in both 1-billion-token and 14 billion-token subsets produced the same results, but this was different for low-frequency nouns. For low-frequency data, three corpora did not show much coincidence with ranking shown in the Russian frequency dictionary by Lyashevskaya and Sharoff (2009). Hence, this issue requires a more detailed investigation.

In this study, we use a collection of Russian corpus data developed within the framework of the Aranea Project (Benko, 2014). We randomly sampled the largest Araneum Russicum Maximum corpus to obtain three smaller subcorpora of total 1 million words (1 M hereafter), 10 million words (10 M hereafter), and 100 million words (100 M hereafter) respectively. The sampling procedure was document-based and worked on sets of 1,000 documents. Out of each set, the first 1,000-n documents were obtained, and the 1,000-n ones were deleted. This approach allowed to preserve all document metadata in the sampled corpus. Although the procedure is not strictly random, it proved to be sufficient for large corpora without extra sophisticated randomization required.

The aim of our experiments was to test the following hypotheses:

1. Low-frequency lexis and its collocations are better represented in large corpora (exceeding 100 million words);
2. Frequent collocations presented in dictionaries have low occurrences in small corpora;
3. Certain statistical measures perform better on small corpora, whereas others require larger corpora.

It can be somewhat problematic to find data about low-frequency lexis or at least to understand what kind of collocations belong to the low-frequency group. Authors of the Macmillan English Dictionary for Advanced Learners (2002) make a clear distinction between high-frequency core vocabulary and less common words using different fonts and the star symbol.

Russian dictionaries, on the other hand, do not provide such information. Thus, frequency dictionaries are the only ones that can provide quantitative data for individual words (but not collocations). The dictionary by Lyashevskaya and Sharoff (2009) provides data for 20,000 lemmata. In the first part of our experiment, we selected lexical items from the end of the list that can produce collocations. Those were ranked between position 19,687 to 20,004 and had the same frequency, i.e. 2.6 instances per million (ipm). Nouns and adjectives were the most representative groups, but verbs and adverbs were also analyzed.

When developing a gold standard for Russian collocability (Khokhlova, 2018a), we produced a list of collocations presented in different Russian dictionaries and introduced a notion of dictionary index, i.e. the number of dictionaries that include a given collocation. The higher the dictionary index, the more frequent and widely used the collocation is. Less frequent collocations have lower dictionary index scores. In the first experiment of our study, we evaluate corpora with those collocations that have minimal dictionary index score.

Along with studying the behavior of low-frequency lexemes and their collocations, we conducted a case study of frequent collocations from the gold standard, i.e. the ones that showed the highest dictionary index scores. For this task we selected 20 collocations which were described in four different Russian dictionaries (explanatory and specialized ones, for example, for language learners).

In the last phase of our experiment, we extracted *adjective+noun* collocations (based on the morphosyntactic annotation by TreeTagger (Schmid, 1994) from each of the above mentioned subcorpora using four association measures (t-score, MI, Dice coefficient and Fisher's exact test) (Evert, 2004; Pecina, 2009) and compared top 500 candidates. These measures were chosen as they are based on different statistical principles and have demonstrated efficiency in prior experiments (Khokhlova, 2018b). Having applied the

frequency threshold (at least 3), we extracted bigrams[1] from three subcorpora. Here are some examples: *Rossiyskaya Federatsiya*[2] 'Russian Federation', *elektronnaya pochta* 'e-mail', *vannaya komnata* 'bathroom', *rabochiy stol* 'work table', *evropeyskaya strana* 'European country' etc. Collocations that were used for evaluation are largely based on the gold standard and insufficient; therefore, we had to rely on linguistic assessment as well.

Then, we analysed the top 500 candidates. Altogether, we extracted the following number of bigrams:

- 1 M: 9,862;
- 10 M: 51,745;
- 100 M: 368,055.

There were no dictionaries of Russian collocations that would be large enough in volume and, thus, information on collocational restrictions (that can be used for data evaluation) had to be obtained from other types of dictionaries and resources.

4 RESULTS

4.1 Results for low-frequency collocations

For our case study we selected 25 adjectives, 8 nouns, 10 verbs and 8 adverbs and thus investigated the following lexical items: adjectives *bezotkaznyy* 'failproof, unfailing', *daveshniy* 'recent', *kinetisheskiy* 'kinetic', *neprerekayemyy* 'incontestable', *priglushennyy* 'muted', *slovarnyy* 'lexicographic', *neproglyadnyy* 'impenetrable', *okkupatsionnyy* 'occupational', *opryatnyy* 'neat', *pogrebal'nyy* 'funeral', *rassuditel'nyy* 'sober', *tyagovyy* 'tractive', *bezdumnyy* 'thoughtless', *vitoy* 'twisted', *neproshenyy* 'undesired', *nerazlichimyy* 'indiscernible', *bessrochnyy* 'perpetual', *mezhlichnostnyy* 'interpersonal', *orkestrovyy* 'orchestric', *zazhitochnyy* 'prosperous', *neprelozhnyy* 'inviolable', *obsharpannyy* 'shabby', *smertonosnyj* 'pestilent', *kishechnyj* 'intestinal', *tseleustremlennyy* 'purposeful'; nouns *inkvizitsiya* 'inquisition', *rassloyeniye*

1 The term "bigram" denotes combinations of two adjacent words.

2 Henceforth, the examples originally written in Cyrillic are given in Latin transliteration.

'stratification', *eroziya* 'erosion', *podlodka* 'submarine', *pischevareniye* 'digestion', *sedmitsa* 'week', *ontologiya* 'ontology", *kholuy* 'toady'; verbs *vydelyvat'* 'to curry', *zavyvat'* 'to wail', *pronzat'* 'to pierce, to impale', *teshit'* 'to amuse, to please', *vlepit'* 'to slap', *pokolebat'* 'to shake', *zayedat'* 'to eat', *poloskat'* 'to rinse, to gargle', *ostudit'* 'to cool', *privivat'* 'to implant, to instil'.

We scrutinized and evaluated the concordance output against the gold standard.

Table 1 represents the results of the analysis for collocations with low-frequency adjectives. The first column lists the lemmata, other columns give the number[3] of concordance lines in total (in the 1 M, 10 M and 100 M corpora) and with appropriate nouns (marked as collocations) for the 1 M, 10 M and 100 M corpora respectively. We considered as appropriate those lexical combinations that are recurrent in the written language. Thus, out of 20 concordance lines of output, all 20 may turn out to contain interesting word form collocates.

Table 1: *Results for low-frequency adjectives*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
bessrochnyy	2	2	24	23	249	248
bezdumnyy	0	0	18	8	51	32
bezotkaznyy	2	1	15	13	132	120
daveshniy	0	0	0	0	21	14
kineticheskiy	10	10	25	23	180	178
kishechnyy	11	11	101	95	210	208
mezhlichnostnyy	0	0	34	34	148	148
neprelozhnyy	0	0	9	9	82	78
neprerekayemyy	0	0	5	4	34	34
neproglyadnyy	0	0	5	5	33	32
neproshenyy	2	2	7	7	26	20
nerazlichimyy	0	0	1	0	41	11
obsharpannyy	0	0	1	1	35	35
okkupatsionnyy	0	0	7	7	92	88

3 Here and in the following tables we mean instances (i.e. absolute frequencies) in columns with numbers.

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
opryatnyy	0	0	12	6	130	88
orkestrovyy	0	0	4	4	69	69
pogrebal'nyy	2	2	17	17	149	149
priglushennyy	1	1	7	7	239	187
rassuditel'nyy	0	0	6	1	84	31
slovarnyy	1	1	47	47	447	441
smertonosnyj	3	3	18	18	114	104
tseleustremlennyy	4	2	48	23	221	133
tyagovyy	0	0	2	2	205	203
vitoy	3	3	14	14	156	147
zazhitochnyy	3	3	18	18	133	116

One can observe that despite the same low-frequencies found in the dictionary by Lyashevskaya and Sharoff (2009), lexical items show a significantly different behaviour, i.e. their frequencies vary as well as the number of collocates. The analysis suggests that a 1 M corpus is evidently not enough to produce a sufficient number of examples illustrating low-frequency collocations. More than 50% of adjectives were missing in the given sample. In the 1 M corpus only two lexical items (*kineticheskiy* 'kinetic' and *kishechnyy* 'intestinal') produced 10 and 11 collocations respectively (ranging from 1 up to 3 instances) that can be accounted for their narrow semantic meaning and hence restricted collocability (e.g. *kishechnaya infektsiya* 'enteric infection', *kishechnaya muskulatura* 'intestinal muscles', *kineticheskaya energiya* 'kinetic energy'). More extensive corpora would likely yield larger numbers of relevant examples.

More than a half of concordance lines in the 10 M and 100 M corpora can be seen as a source of collocations without any filtration (e.g. *priglushennyy, slovarnyy, neproglyadnyy* etc). This fact can suggest that in case of low-frequency lexis the increase of texts does not necessarily result in overflow with data and false examples.

Among irrelevant candidates one can find also other instances, i.e. errors in lemmatization (e.g. *vitoj* 'twisted' in *dolche vitoj* 'dolce vita' was lemmatized

as *vitoj* 'twisted' instead of Latin *vita* 'vita'), erroneous part-of-speech tagging (e.g. adjectives instead of adjectival nouns), mistakes and typos.

The findings of the case study for a number of adjectives are reported next.

Priglushennyy 'muted': in the 1 M corpus we found only one rare occurrence *priglushennoye urchaniye* 'muted growl'. The 10 M and 100 M corpora contained collocates representing one lexical group of colour, e.g. *tsvet* 'colour', *gamma* 'colour scheme', *ottenok* 'tint', *pigment* 'pigment', *terrakotovyy* 'terracota' and *zelenyy* 'green'. There were also examples with *golos* 'voice', *shum* 'noise', *zvon* 'toll of the bell'.

Orkestrovyy 'orchestric': only two collocations occured in the 10 M corpus, namely *orkestrovaya jama* 'orchestra pit' and *orkestrovaya partitura* 'orchestra score'. The 100 M corpus gave a wide range of collocates with the sememe 'music', e.g. *aranzhirovka* 'arrangement', *partiya* 'play', *rakovina* 'shell', *syuita* 'suite'.

The evidence suggests that the results obtained for the 1 M corpus include collocates that belong to lexical periphery – not the frequent ones. This is somewhat unexpected, hence the most frequent collocates tend to be found only in larger corpora.

Table 2 shows the results for low-frequency nouns.

Table 2: *Results for low-frequency nouns*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100 M	100 M (collocations)
eroziya	4	2	109	75	484	421
inkvizitsiya	2	1	29	14	134	64
kholuy	0	0	0	0	11	5
ontologiya	2	0	35	20	65	36
pischevareniye	6	6	126	108	1,044	725
podlodka	1	1	18	11	117	51
rassloyeniye	2	2	29	22	239	211
sedmitsa	4	4	11	8	109	100

Rassloyeniye 'stratification': there are only two occurrences in the 1 M corpus, a term *rassloyeniye vina* 'wine stratification' and *sotsial'noye*

rassloyeniye 'social differentiation'. The former has a highly specific and narrow meaning while the latter can be called a collocation. In the 10 M corpus one can find other meaningful examples, e.g. *rassloyeniye strany* 'stratification of country' or *obschestva* 'of society', *rassloyeniye nogtey* 'nail splitting' or *komponentov* 'segregation of components'.

Podlodka 'submarine': the most frequent collocate turns to be *atomnyy* 'atomic' that can be found both in the 1 M and 10 M corpora. The 10 M corpus also contains two verbal collocates, e.g. *zatonut'* 'to founder' and *topit'* 'to sink'. The 100 M corpus gives more examples, e.g. *prishvartovat'* 'to moor', *unichtozhit' 'to destroy'*, *stoyat'* 'to stay, *idti 'to go'*, *chodit' 'to go'*.

Pischevareniye **'**digestion**'**: the given noun is the only one showing wide collocability, i.e., we find collocates among adjectives, nouns and verbs. Compared to other nouns it has the highest frequency.

Sedmitsa 'week': The 1 M corpus shows only adjective collocates, e.g. *Svetlyy* 'Easter' and *Strastnoy* 'Holy'. The 10 M corpus does not add any valuable collocations with adjectives, except for one occurrence of *syrnaya sedmitsa* 'shrovetide'. The 100 M corpus includes only one example of noun collocate *sedmitsa mytarya i fariseya* 'the week of the Publican and the Pharisee'.

Kholuy 'toady': among all the nouns, it proved to have the lowest frequency; no occurrence was found in the 1 M and 10 M corpora.

It is also true for nouns (as it was the case for adjectives) that although we see the same low-frequency according to the frequency dictionary (Lyashevkaya and Sharoff, 2009), the number of examples and hence collocations is different. The noun *pischevareniye*, for example, shows more than 1,000 occurrences.

We can see that small corpora produce even fewer collocates for nouns than for adjectives. There are virtually no collocations with verbs, whereas those with nouns and adjectives prevail.

Table 3 presents the results for low-frequency verbs and their collocations.

Table 3: *Results for low-frequency verbs*

	1 M	1 M (collocations)	10 M	10 M (collocations)	100M	100M (collocations)
ostudit'	3	3	21	9	208	156
pokolebat'	2	2	10	9	68	46
poloskat'	1	1	22	21	170	123
privivat'	4	4	28	28	260	209
pronzat'	1	1	4	3	47	42
teshit'	0	0	9	6	76	63
vlepit'	0	0	2	0	16	8
vydelyvat'	0	0	3	3	41	37
zavyvat'	0	0	3	2	25	19
zayedat'	0	0	9	6	103	79

Despite the fact that the verbs selected for the experiment are polysemous and should therefore demonstrate wide collocational preferences, they tend to get the lowest number of collocations in smaller corpora, as opposed to nouns and adjectives. Both the 1 M and 10 M corpora do not yield a sufficient number of examples.

Although the frequency of the verbs is the same (2.6 ipm) in the dictionary (Lyashevkaya and Sharoff, 2009), it varies widely in corpora, e.g. from 0.16 up to 2.25 ipm.

Vydelyvat' 'to curry': only the 100 M corpus shows collocability of verbs with nouns.

Zavyvat' 'to wail': in the 10 M corpus there are two examples of a subject collocating with a verb, e.g. *v'yuga* 'snowstorm' and *veter* 'wind'.

The average percentage of the data filtering for nouns and verbs is higher than for adjectives, i.e. the output results show irrelevant occurrences, mistakes, typos, other noise or word usage without any collocates. Adjectives tend to be part of noun groups (not always, though), whereas nouns and verbs can be used more often as independent lexical units. Therefore, corpora exceeding 100 M are more efficient in representing collocability of low-frequency nouns and verbs.

Having come to a preliminary conclusion that there is a need to further expand the volume of corpora, we also studied a number of syntactic relations[4] based on 100 M and 1.2 G corpora. We looked at the neighborhood of low-frequency nouns and analyzed the output by filtering out typos, errors in lemmatization etc. in order to count lemmata examples only. Table 4 represents the number of attributive and verbal collocations.

Table 4: *Number of different collocations for nouns*

	adjective + noun (100 M)		adjective + noun (1.2 G)		verb + noun, noun + verb (100 M)		verb + noun, noun + verb (1.2 G)	
	all forms	lemmata	all forms	lemmata	all forms	lemmata	all forms	lemmata
eroziya	77	31	1,328	78	106	46	2,919	79
inkvizitsiya	26	16	564	72	26	19	1,225	87
kholuy	6	6	13	10	0	0	9	3
ontologiya	20	16	246	43	9	4	298	22
pischevareniye	53	19	1,784	73	266	41	6,945	57
podlodka	32	18	582	62	30	18	964	81
rassloyenoye	72	30	743	66	64	33	1,230	82
sedmitsa	64	12	688	22	11	8	501	55

With the expansion of corpus volume, the number of collocations increases as well as the amount of noise or irrelevant cases. Additional data filtering is therefore needed. When the corpus volume increases by 10 times, the number of concordance lines per collocation also increases by at least 10 times (strictly speaking, on average, 18 times for the nouns under consideration).

To be more specific, preliminary results of our study have shown that higher absolute frequency of a particular lexical item does not always mean a larger number of syntactic relations for the lexical item (despite the greater number of collocates typical of each relation).

4.2 Results for frequent collocations from dictionaries

The dictionary index (Khokhlova, 2018a) designates the number of dictionaries which present the given collocation. Large values of the index imply

4 The analysis was made on the Russian word sketch grammar in Sketch Engine (Khokhlova, 2010; Kilgarriff et al., 2014).

that the collocation is reproduced quite often and thus should be learnt by heart (if we speak about the learners of Russian). Theoretically, the maximum is equal to the number of dictionaries, that is 6 for the *adjective + noun* model, but in practice the maximum number of dictionaries in which the collocation was fixed was 4. The gold standard comprises more than 15,000 collocations for the given model and only 61 examples were described in 4 dictionaries (so there is no example to be recorded in all 6 dictionaries). We randomly selected 20 frequent collocations from this list and analyzed them across the corpora. Table 5 presents the results sorted by the number of occurrences in the 100 M corpus.

Table 5: *Frequency distribution of selected collocations from the gold standard*

		1 M	10 M	100 M
yarkiy primer	'vivid example'	3	65	533
vysokiy rezul'tat	'high result'	1	43	532
bol'shoy uspekh	'big success'	6	50	357
grubaya oshibka	'great error'	1	8	125
vysokaya pribyl'	'high profit'	0	15	79
glubokaya blagodarnost'	'deep gratitude'	0	3	68
polnaya tishina	'complete silence'	1	11	62
polnaya pobeda	'complete victory'	1	12	55
bogatyy urozhay	'bountiful harvest'	0	9	50
glubokiy krizis	'deep crisis'	0	5	44
glubokoye udovletvoreniye	'deep satisfaction'	0	1	31
shirokiy razmakh	'wide scope'	0	0	24
ostraya bor'ba	'fierce struggle'	0	1	21
general'noye srazheniye	'decisive battle'	0	1	15
goryachaya lyubov'	'hot love'	0	4	14
zheleznaya distsiplina	'iron discipline'	1	3	10
gomericheskiy khokhot	'homeric laughter'	0	1	8
zhguchiy vopros	'burning question'	0	0	6
shirokoye sotrudnichetsvo	'wide cooperation'	0	0	2
zheleznyy kharakter	'strong character'	0	0	2

Even in the case of frequent collocations from the gold standard the 1 M corpus yields no results and hence cannot be used as a source of linguistic

evidence. The 10 M corpus also contains a small number of collocations. The collocation frequencies are significantly higher in the 100 M corpus and this can be accounted for by high frequencies of either the node or the collocate.

4.3 Results of automatic extraction

In the course of further experiments we used statistical measures to extract bigrams setting frequency cutoff threshold of f=3 and then the bigrams were evaluated bigrams against the dictionary data, and by native-speaker inspection. The analysis also revealed a large amount of morphological mistakes and errors in lemmatization. For example, *zloy dukhi* 'evil perfume' instead of *zloy dukh* 'evil spirit'; *pal'movom masle* 'palm oil' (the lemma for the adjective stands in the prepositional case) instead of *pal'movoye maslo*.

Table 6 presents the number of collocations extracted by each of the association measures from the 1 M, 10 M and 100 M subcorpora respectively.

Table 6: *Number of collocations per subcorpus*

	1 M	10 M	100 M
MI	229	97	54
t-score	484	492	495
Dice	301	186	114
Fisher	454	490	499

The analysis suggests that MI and Dice tend to extract fewer collocations from a larger corpora, retrieving examples with typos and mistakes. This can lead us to the hypothesis that vast collections of text data will have more non-collocations (for example, free phrases) and, thus, top lists will also contain such senseless word combinations (or even hapax legomena, if there is no frequency threshold). Dice coefficient also focuses predominantly on terms, proper names and set phrases, e.g. *nashatyrny spirt* 'liquid ammonia', *gadkiy utenok* 'ugly duckling'. Compared to other measures, Fisher's exact test extracted the largest number of collocations.

Table 7 shows numbers of shared bigrams found by each measure in different corpora.

Table 7: *Numbers of shared bigrams (by subsets)*

	1 M/10 M	10 M/100 M	1 M/100 M
MI	38	31	1
t-score	275	427	262
Dice	96	63	13
Fisher	241	424	233

When we compare lists extracted by different measures, we can see that MI and Dice do not tend to extract the same collocations in the corpora of different volumes. The percentage of the intersection declines with the increase of difference between corpus volumes, resulting in a smaller amount of bigrams. T-score and Fisher's exact test demonstrate contrasting behaviour, i.e. the highest number of the identical bigrams is extracted from the 10 M and 100 M corpora while the 1 M/10 M and 1 M/100 M pairs show almost the same number.

Table 8 demonstrates the number of the same bigrams found in the 1 M, 10 M and 100 M corpora, respectively. Here the results suggest that the measures can be again divided into two groups according to the behaviour, namely, the first group contains MI and Dice, whereas in the second are t-score and Fisher's exact test.

Table 8: *Number of the shared bigrams (breakdown by measures)*

	1 M				10 M				100 M			
	MI	t-score	Dice	Fisher	MI	t-score	Dice	Fisher	MI	t-score	Dice	Fisher
MI	500	8	350	32	500	0	347	0	500	0	366	0
t-score		500	80	385		500	46	393		500	4	396
Dice			500	134			500	71			500	8
Fisher				500				500				500

Tables 9 to 11 show the number of the identical bigrams that were found in the 1 M, 10 M, and 100 M corpora, respectively, by measures. The comparison was made between corpora of different sizes. Measures from the above mentioned two groups show lower numbers of identical bigrams with the increase of corpus size.

Table 9: *Number of identical bigrams (1 M vs 10 M by measures)*

	MI (1 M)	t-score (1 M)	Dice (1 M)	Fisher (1 M)
MI (10 M)	38	0	35	4
t-score (10 M)	6	275	43	222
Dice (10 M)	62	35	96	57
Fisher (10 M)	15	248	62	241

Table 10: *Number of identical bigrams (1 M vs 100 M by measures)*

	MI (1 M)	t-score (1 M)	Dice (1 M)	Fisher (1 M)
MI (100 M)	1	0	1	0
t-score (100 M)	2	262	33	211
Dice (100 M)	11	2	13	6
Fisher (100 M)	25	241	57	233

Table 11: *Number of identical bigrams (10 M vs 100 M by measures)*

	MI (10 M)	t-score (10 M)	Dice (10 M)	Fisher (10 M)
MI (100 M)	31	0	31	0
t-score (100 M)	0	427	38	370
Dice (100 M)	54	5	63	8
Fisher (100 M)	0	375	60	424

5 CONCLUSION AND FURTHER WORK

Though final conclusions might be too early to formulate, we can say that larger corpora do not always have an advantage, especially in situations when most frequent phenomena are studied. Depending on the mode of analysis, larger amounts of data may even turn into an obstacle, especially if the research has to observe time limits. Nevertheless, the results for low-frequency lexis prove the fact that corpora of less than 100 million words are not sufficient to represent collocations. In terms of our study, this can be partly accounted for by rich flectional nature of Russian morphology and a relatively free word order.

We should mention that frequent collocations which are described in several

dictionaries cannot be found in smaller corpora. The results suggest that in order to properly represent these collocations in dictionaries, one needs corpora exceeding 100 million words.

The results are largely based and depend on the quality of data, which raises again the question of how to prepare a corpus, especially to study low-frequency phenomena. The evidence obtained for infrequent lexis can differ for other text types or domains and, thus, metatextual annotation can be taken into account in further experiments.

From the perspective of various association measures used to identify collocations, we have shown that not all of them work well for larger corpora. Our observation can be summarized as follows:

- MI and Dice extract more terms, typos, hapax legomena, errors in lemmatization with the increase of volume, and thus perform better on smaller corpora;
- t-score and Fisher's exact test extract more good collocations from larger corpora.

We believe that the relationship between the corpus size, and the number and "quality" of extracted collocations is a fascinating topic to study; a similar research should be performed on different corpora and/or languages as well.

Acknowledgments

This work was supported by the grant of the Russian Science Foundation (Project No. 19-78-00091).

REFERENCES

Dictionaries, corpuses and digital resources

Lyashevskaya, O., & Sharoff, S. (2009). *The Frequency Dictionary of Modern Russian based on the Russian National Corpus data* [Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo Korpusa Russkogo Yazyka)]. Moscow: Azbukovnik.

Macmillan English Dictionary for Advanced Learners. (2002). Macmillan Education.

Steinfeld, E. (1963). *Frequency dictionary of the Contemporary Russian language* [Chastotnyy slovar' sovremennogo russkogo literaturnogo yazyka]. Tallin.

The British National Corpus, (Version 3) (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/ (1. 5. 2020)

The Russian National Corpus [Natsional'nyy korpus russkogo yazyka]. Retrieved from http://www.ruscorpora.ru (1. 5. 2020)

The Brown Corpus. Retrieved from http://korpus.uib.no/icame/manuals/brown/index.htm, https://www.sketchengine.eu/brown-corpus/ (1. 5. 2020)

Zasorina, L. (1977). *Frequency dictionary of the Russian language* [Chastotnyy slovar' russkogo yazyka]. Moscow: Russkiy yazyk.

Other

Benko, V. (2014). Aranea Yet Another Family of (Comparable) Web Corpora. *Text, Speech and Dialogue. Proceedings of the 17th International Conference, TSD 2014, 8–12 September, 2014, Brno, Czech Republic.* LNCS 8655 (pp. 257–264). Springer International Publishing Switzerland.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.

Daudaravičius, V. (2010). The influence of collocation segmentation and top 10 items to keyword assignment performance. *Computational Linguistics and Intelligent Text Processing*. *Proceedings* of the *11th International Conference, CICLing 2010, 21–27 March, 2010, Iasi, Romania* (pp. 648–660). Berlin: Springer.

Evert, S. (2004). The Statistics of Word Cooccurrences Word Pairs and Collocations. Dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart. Available at http//purl.org/stefan.evert/PUB/Evert2004phd.pdf (20. 2. 2020)

Evert, S., Uhrig P., Bartsch S., & Proisl, T. (2017). E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference, 19–21 September, 2017, Leiden Netherlands* (pp. 531–549). Leiden: Lexical Computing.

Khokhlova, M. (2010). Building Russian Word Sketches as Models of Phrases. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress, 6–10 July, 2010, Leeuwarden* (pp. 364–371). Ljouwert: Fryske Akademy – Afûk.

Khokhlova, M. (2017). Big data and word frequency: Measuring the consistency of Russian corpora. *Quantitative Approaches to the Russian Language* (pp. 30–48). Routledge, Taylor & Francis.

Khokhlova, M. (2018a). Building a Gold Standard for a Russian Collocations Database. In J. Čibej et al. (Eds.), *Lexicography in Global Contexts. Proceedings of the XVIII EURALEX International Congress* (pp. 863–869). Ljubljana: Ljubljana University Press, Faculty of Arts.

Khokhlova, M. (2018b). Similarity between the Association Measures a Case Study of Noun Phrases. In *Proceedings of the 12th Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018* (pp. 21–27). Brno Tribun EU.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*, 7–36.

Pecina, P. (2009). *Lexical Association Measures. Collocation Extraction*. Prague Institute of Formal and Applied Linguistics.

Piotrowski, R. G., Bektaev, K. B., & Piotrowskaya, A. A. (1977). *Mathematical Linguistics* [Matematicheskaya lingvistika]. Moskva: Vysshaya shkola.

Piperski, A. (2015). To be or not to be: Corpora as Indicators of (Non-)Existence. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue"*, *1*(14), 515–522.

Rychly, P. (2008). A lexicographer-friendly association score. *Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Language Processing RASLAN 2008* (pp. 6–9). Brno: Masaryk University.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Sinclair, J. (2005). Corpus and Text — Basic Principles. In M. Wynne (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1–16). Oxford: Oxbow Books. Retrieved from http://users.ox.ac.uk/~martinw/dlc/chapter1.htm (1. 5. 2020)

VELIKOST KORPUSOV IN OBSEG KOLOKACIJ NA PRIMERU RUŠČINE

Potem ko se je na področju jezikoslovja razmahnila uporaba informacijskih tehnologij, je izdelava obsežnih korpusov, sploh tistih s spletnimi besedili, postala zelo enostavna naloga. Nove priložnosti pa so zopet oživile vprašanja o velikosti korpusa: so večji korpusi boljši za jezikoslovne raziskave, natančneje, ali morajo leksikografi posledično analizirati večje količine kolokacij? Prispevek predstavi eksperimente, v katerih smo iskali kolokacije redkejših besed s pomočjo korpusov različnih velikosti (1 milijon besed, 10 milijonov besed, 100 milijonov besed in 1,2 milijardi besed). Izbrali smo redke pridevnike, samostalnike in glagole iz Ruskega frekvenčnega slovarja in preverili sledeče hipoteze: 1) kolokacije redkejše leksike so bolje zastopane v večjih korpusih; 2) pogoste kolokacije iz slovarjev se redko pojavljajo v manjših korpusih; 3) statistične mere za luščenje kolokacije dajejo različne rezultate pri korpusih različnih velikosti. Rezultati dokazujejo, da korpusi, manjši od 100 milijonov besed, niso dovolj reprezentativni za preučevanje kolokacij, sploh tistih, ki vsebujejo samostalnike in glagole. Statistični meri MI in Dice sta pri luščenju kolokacij manj zanesljivi, sploh pri večjih korpusih, po drugi strani pa t-score in Fisherjev natančni test kažeta boljše rezultate prav pri večjih korpusih.

Ključne besede: kolokacije, ruski korpusi, velikost korpusa, korpusno jezikoslovje, statistične mere