Quantitative Properties of Russian Adjective-Noun Collocations across Dictionaries and Corpora*

Maria Khokhlova[0000-0001-9085-0284]

St. Petersburg State University, 7/9 Universitetskaya nab., 199034 St. Petersburg, Russia
m.khokhlova@spbu.ru

Abstract. The paper discusses the differences between collocations extracted from a number of Russian dictionaries paying attention to their frequency characteristics based on corpora. The aim of the study was, first, to analyze how collocations and set expressions are described in Russian explanatory and specialized dictionaries and to what extent their data coincide with each other, and, secondly, to investigate how collocations presented in dictionaries are reflected in text corpora. This will make it possible to examine the interrelation between the "manually" collected data and modern corpora (the Russian National Corpus and ruTenTen). We tested the following hypothesis, i.e. high collocation frequencies correspond to the fact that the item is represented in several dictionaries. In our paper we considered 180 collocations built according to the "adjective / participle + noun" model. The results show the heterogeneity of the dictionary data while the choice of lexical items does not coincide with its frequency characteristics: the examples are low-frequency and about 34% are absent in the disambiguated subcorpus. Explanatory dictionaries and collocation dictionaries show the smallest overlap.

Keywords: Collocations, Russian Language, Dictionaries, Corpora, Statistics.

1 Introduction

Our project deals with the process of building a database that will represent Russian collocations extracted from dictionaries and corpora [8]. The results of this research can be used in various NLP tasks and also in different fields of theoretical and applied linguistics, i.e. Russian lexicology, morphology and syntax or teaching the Russian language. Data about Russian collocability can be valuable for machine translation, clustering of words and word combinations, sentiment analysis, text summarization, disambiguation etc. It is expected that the collocations extracted from dictionaries will be used for the evaluation of machine learning algorithms dealing with automatic text processing, since today there is no single standard that would include verified information and at the same time in sufficient quantity.

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

Since dictionaries are an important source of data about collocations, it is important to analyze them in order to understand the possible distinctions. At the same time, it can be tricky to compare different dictionaries.

In the paper we consider collocations which were extracted from six Russian dictionaries, analyzing how they are reflected in corpora of the Russian language. Within the framework of our research, we dwell on two tasks. First, to analyze how recurrent word combinations are presented in different dictionaries and how much they coincide with each other. Secondly, to investigate the extent to which collocations that are reflected in dictionaries can be found in corpora and, therefore, trace the intersection between "manually" collected data and modern corpora.

2 Related Work

The tradition of Russian lexicography has a rich history, however, there are not many projects dealing with collocability in Russian and moreover based on corpus data or created by automatic methods. At the same time corpora are seen as the main source of language data in Western lexicography and up-to-date projects implement them (Macmillan Dictionary).

The issue of selecting collocations is crucial in lexicography, and not only for monolingual dictionaries, representing "the most controversial and vulnerable part of almost every bilingual dictionary" [2, p. 61]. Atkins and Rundell [1] point out the difficulty of selecting examples from the corpus and suggest using collocation lists for this task. As some authors note [4, 12] the issue of differentiating phrases of various types is still controversial, which leads to the fact that "specific cases of idiomatic combinations often do not receive an unambiguous qualification, which is reflected, in particular, in dictionaries" [12, p. 2].

What kind of phrases to include in a computer dictionary is discussed, for example, in [14]. Multi-word expressions can be seen as an umbrella term and it is true for lexicographic resources. Authors list different word combinations in dictionary entries calling them idioms, phrasemes, collocations etc. A detailed overview of the available Russian dictionaries was given in the paper [9]. The paper [11] describes the machine learning procedure of selecting coefficients for searching collocations based on the examples selected from the dictionaries.

The idea of comparison between dictionaries and corpora attracted attention from scholars a few decades ago. The interest was focused on automatic extraction (either rule-based or statistical one) from the sources and their further evaluation. The results of the analysis complement each other and can be applied for constructing NLP lexicons [6]. The research presented later in [20] puts emphasis on collocations in machine translation. Their implementation improved the quality of the analyzed systems.

There is also a number of works involving comparison between text corpora (for example, [17, 10]). Most of them deal with building frequency lists and calculating statistical metrics based on several corpora trying to find a suitable test "supporting the comparison of small and large corpora" [10, p. 258].

3 Experiment: Description

The merging of dictionaries from different sources implies not only a single lexicographic format, but also raises the question of data relevance. When describing collocations, a lexicographer needs to select examples taking into account their representativeness in corpus, coverage in dictionaries, and also suitability for language users and their purposes.

We identified items from several Russian dictionaries of different types:

1. Explanatory dictionaries, i.e. the Dictionary of the Russian Language (DRL [5]); the Large Explanatory Dictionary of the Russian Language (LEDR [13]);
2. Collocation dictionaries [18, 16, 3];
3. Online dictionary [12].

In our research, we tested the following hypothesis: high collocation frequencies in the corpus correspond to high values of the dictionary index introduced in [8]. This index is understood as the number of dictionaries in which the item is recorded. That is, we expect to see a directly proportional relationship between lexicographic and corpus data and, therefore, a positive correlation between dictionaries and corpora. None of the collocations was recorded in all six dictionaries we examined, so the maximum value of the dictionary index turned out to be 4.

To assess the extracted data across text corpora, we randomly selected 20 collocations from groups with dictionary indices 2, 3 and 4 (see Table 1 for the examples) resulting in 60 collocations. During the statistical analysis we implemented nonparametric Friedman and Kruskal-Wallis tests for the comparison between corpora.

We also analyzed 20 collocations that are present only in one dictionary (i.e., their dictionary index was 1) resulting in 120 collocations.

As a material for our study we used the Russian National Corpus (RNC), i.e. a disambiguated subcorpus of 6 mln tokens and the main corpus of 321 mln tokens. The given subcorpora represent a "classical" (traditional) approach to corpus building compared to an automatic one but are not so large, as opposed to other Russian corpora. Therefore we also consulted ruTenTen corpus of more than 18 bln tokens that was crawled automatically [7].

Table 1. Examples of collocations

	Collocation	Borisova 1995	Kustova 2008	Oubine 1987	DRL	Reginina, Tyurina, Shirokova 1980	LEDR	Dictionary index
1.	*adskaya bol'* 'hellish pain'	0[1]	1	1	0	0	0	2
2.	*glubokaya drevnost'*	0	1	1	0	1	0	3

[1] 1 and 0 stand for presence or absence of the collocation in the dictionary.

	'great antiquity'						
3.	*goryachaya lyubov'* 'burning love'	1	1	1	0	1	0	4
4.	*ostraya diskussiya* 'heated discussion'	1	1	1	0	1	0	4
5.	*vysokoye masterstvo* 'superior skill'	0	1	1	0	1	0	3
6.	*zverinaya zhestokost'* 'monstrous cruelty'	0	1	1	0	0	0	2

In our study we will also address to the following questions: 1) can we use corpora of a smaller volume for collocation analysis or in tasks dealing with their automatic processing? 2) do "traditional" and large web corpora produce the same results? 3) do dictionaries present homogeneous data, i.e. collocations of the same language nature demonstrate similar quantitative features?

4 Experiment: Results

4.1 Dictionary index 4

Even for the high value of dictionary index, the results are not uniform (see examples in Table 2). 4 collocations are absent in the disambiguated subcorpus, although they are present in several dictionaries. Spearman's rank correlation coefficient is about 0.81 for all three pairs of corpora, which indicates a strong positive relationship between them and their similar ranking of collocations.

Table 2. Collocations with dictionary index 4

№	collocation		RNC, subset	RNC	ruTen-Ten
1.	bol'shaya raznitsa	'big difference'	1.33	2.06	1.85
2.	bol'shoy uspekh	'great success'	5.16	7.61	4.92
3.	bol'shoye znacheniye	'great meaning'	7.33	9.27	12.65
4.	glubokiy smysl	'deep meaning'	2.17	1.52	1.23
5.	glubokoye udovletvoreniye	'deep satisfaction'	0.83	0.55	0.34
6.	glubokoye uvazeniye	'deep respect'	0.50	1.88	1.10
7.	goryachaya lyubov'	'burning love'	0.17	1.20	0.24

8.	krepkaya druzhba	‘strong friendship’	0.17	0.22	0.29
9.	ostraya diskussiya	‘heated discussion’	0.00	0.28	0.33
10.	ostraya kritika	‘sharp criticism’	0.33	0.25	0.21
11.	ostraya nuzhda	‘desperate need’	0.00	0.42	0.16
12.	polnaya svoboda	‘complete freedom’	2.17	4.76	2.24
13.	shirokaya diskussiya	‘broad discussion’	0.17	0.15	0.15
14.	shirokaya izvestnost’	‘great fame’	1.00	1.00	1.26
15.	shirokaya podderzhka	‘broad support’	0.00	0.25	0.39
16.	shirokiy razmakh	‘wide scope’	0.50	0.72	0.31
17.	slaboye mesto	‘weak point’	1.50	2.20	3.14
18.	vysokiy rezul’tat	‘high score’	1.00	0.53	3.83
19.	vysokiy urozhay	‘high yield’	0.00	0.92	0.75
20.	yarkiy primer	‘vivid example’	2.50	2.83	4.84

The average frequency in the subcorpus of the RNC was 1.34, the main corpus of the RNC and ruTenTen showed 1.93 and 2.01 respectively, but the differences between the data are statistically insignificant ($p > 0.05$ according to the Friedman test). Hence, the frequencies are homogeneous but two collocations with the lexeme bol’shoy ‘large’ (bol’shoye znacheniye ‘great meaning’ and bol’shoy uspekh ‘great success’) show outliers.

4.2 Dictionary index 3

The frequencies of this group of collocations (see Table 3) show lower correlation (Spearman’s rank correlation coefficient varies between 0.62 and 0.79), while the differences between them in corpora are significant ($p < 0.05$ according to the Friedman test).

Table 3. Collocations with dictionary index 3

№	collocation		RNC, subset	RNC	ruTen-Ten
1.	bol’shaya beda	‘big trouble’	0.83	1.76	0.60
2.	bol’shaya pol’za	‘great benefit’	0.50	2.42	1.13
3.	bol’shaya pomosch	‘great help’	0.66	0.98	1.23
4.	bol’shaya vazhnost’	‘great importance’	0.33	0.93	0.25
5.	gigantskiy shag	‘giant step’	1.00	0.62	0.14
6.	glubokaya drevnost’	‘great antiquity’	1.83	1.80	1.62
7.	glubokoye vliyaniye	‘deep influence’	0.33	0.19	0.17
8.	korennoy interes	‘core interest’	0.50	0.20	0.14
9.	lyutaya nenavist’	‘fierce hatred’	0.83	0.45	0.21
10.	lyutyy moroz	‘bitter frost’	1.50	0.73	0.42
11.	nabityy durak	‘perfect fool’	0.00	0.13	0.01
12.	posledneye izvestiye	‘last news’	2.00	2.23	0.24
13.	ravnoye pravo	‘equal right’	1.00	1.45	1.63

14.	tesnaya druzhba	‘close friendship’	0.50	0.81	0.16
15.	tyazhelaya zadacha	‘difficult task’	0.00	0.21	0.13
16.	velikoye pereseleniye	‘great relocation’	0.33	0.45	0.34
17.	vysokaya tre-bovatel’nost’	‘high exactingness’	0.17	0.14	0.13
18.	vysokoye masterstvo	‘high skill’	0.50	0.38	0.68
19.	zhguchiy styd	‘burning shame’	0.50	0.23	0.04
20.	zhiznennyy put’	‘life path’	2.83	4.03	3.87

Thus, in the case of the above given examples that are present in three dictionaries, the frequencies tend to reveal more diversity.

4.3 Dictionary index 2

For collocations selected from two dictionaries, we see that 12 units out of 20 (60%) were not found in the smallest corpus, and 3 of them were not recorded in the main RNC corpus either (see Table 4 for the results).

Table 4. Collocations with dictionary index 2

№	collocation		RNC, subset	RNC	ruTen-Ten
1.	adskaya bol’	‘hellish pain’	0.50	0.17	0.15
2.	bezgranichnaya toska	‘boundless longing’	0.00	0.01	0.01
3.	bezmernaya glubina	‘immense depth’	0.17	0.03	0.01
4.	isklyuchitel’noye mnogoobraziye	‘exceptional diversity’	0.00	0.01	0.01
5.	l’vinaya chast’	‘lion's share’	0.00	0.12	0.13
6.	mestnyy padezh	‘local case’	0.00	0.02	0.01
7.	nervnaya sistema	‘nervous system’	9.66	9.46	17.72
8.	neukrotimaya zloba	‘indomitable malice’	0.17	0.05	0.01
9.	ogromnyy diapazon	‘huge range’	0.00	0.07	0.09
10.	otchayannaya khrabrost’	‘desperate courage’	0.00	0.23	0.04
11.	polnyy vostorg	‘complete delight’	0.33	0.98	0.88
12.	porazitel’naya predos-torozhnost’	‘astounding precaution’	0.00	0.00	0.01
13.	putevodnaya nit’	‘guiding thread’	0.17	0.40	0.17
14.	total’naya slezhka	‘total surveillance’	0.00	0.04	0.08
15.	tsepnaya reaktsiya	‘chain reaction’	1.50	1.97	1.32
16.	uzhasnaya groza	‘terrible thunderstorm’	0.17	0.11	0.02
17.	zhestokoye nakazaniye	‘cruel punishment’	0.00	0.71	0.21
18.	zhguchaya zlost’	‘burning anger’	0.00	0.00	0.01
19.	zverinaya skuka	‘animal boredom’	0.00	0.00	0.01
20.	zverinaya zhestokost’	‘bestial cruelty’	0.00	0.07	0.04

Spearman's correlation coefficient increased up to 0.92 for ruTenTen and the main RNC corpus while it decreased to 0.53 for ruTenTen and the disambiguated RNC subcorpus. Hence we can register differences in ranking in the latter case. But the fluctuations in the frequencies between all three corpora are not significant ($p > 0.05$ according to the Friedman test). This result enables us to suggest that collocations found only in two dictionaries are rare.

4.4 Dictionary index 1

Despite the fact that the online dictionary of idiomatic expressions [12] was compiled on the basis of the RNC, only half of the collocations were recorded in the disambiguated subcorpus. Collocations extracted from the dictionary are characterized by extremely low frequencies in all three corpora and show minimal values compared to other lexicographic resources. This is the poorest result among collocations obtained for all dictionaries. The differences in frequency values between corpora are insignificant ($p > 0.05$ according to the Friedman test), and the standard deviation values are also low, i.e. one can assume some homogeneity of noun collocations in this dictionary.

The collocations extracted from the dictionary of lexical intensifiers [16] are also characterized by low frequencies in corpora and insignificant differences.

The dictionary of set expressions [18] shows the highest results for the collocation frequencies (the differences between corpora are also insignificant, $p > 0.05$ according to the Friedman test), i.e. it can be concluded that this lexicographic resource reflects more frequent collocations. For example, *vyssheye obrazovaniye* 'higher education' and *dukhovnaya zhizn'* 'spiritual life'.

Analysis of the data in the collocations dictionary [3] suggests that the selected items occupy an intermediate position according to their frequency characteristics, i.e. 8 items are not recorded in the disambiguated subcorpus, and 5 collocations have only 1 occurrence in the main RNC corpus. But nevertheless the collocations extracted from the given dictionary prove to be the only ones showing significant differences between corpora.

The results for collocations from both explanatory dictionaries suggest that the sources differ to a certain degree in how they represent unique phrases. For units from the LEDR, the distribution of frequencies in three corpora is characterized by outliers and a large range of values (for example, *aktsionernoye obscestvo* 'joint stock company', *organicheskoye veschestvo* 'organic matter', *pochtovyy yaschik* 'letterbox'). Collocations from the DRL have smaller deviations from the mean values. Both explanatory dictionaries show very little overlap with other lexicographic sources. This can be explained by the fact that dictionaries are aimed at describing different vocabulary: for example, the dictionary [12] represents only phrases with the meaning of high intensity, while DRL and LEDR are aimed at a more complete presentation of vocabulary and dictionary entries list phraseological units.

5 Discussion

The analysis shows that in total there are no significant differences between corpora in frequencies of collocations with the same dictionary indexes (or from the same dictionary). Thus the analyzed items prove to be rare units. About 34% of the considered collocations are absent in the RNC disambiguated subcorpus, i.e. it can be assumed that the volume of 6 mln tokens is not enough to study collocability. About 12% of the analyzed collocations yield less than 0.01 occurrences per million even in the largest ruTenTen corpus.

It should be mentioned that collocation frequencies in corpora are steadily decreasing with the decrease of dictionary index (the differences are statistically significant, $p < 0.05$ according to the Kruskal-Wallis test), and the value of the Spearman's rank correlation coefficient also decreases. Collocations represented in four dictionaries tend to be more widespread in corpora but also have low frequencies.

It is worth noting that with the rise of corpus volume the unique collocations (with dictionary index equal to 1) tend to show more diversity in their frequencies. Only the collocation dictionaries [12] and [18] demonstrated significant differences on the smallest RNC corpus while the main RNC corpus could exemplify one more pair, e.g. the dictionary of set expressions [18] and the dictionary of lexical intensifiers [16].

The ruTenTen corpus proved to have the largest number of pairs with significant differences in frequencies (here we can name additionally, firstly, the dictionary of idiomatic expressions [12] and the collocations dictionary [3], and secondly, the former [12] and LEDR [13]). This can suggest that the dictionary of set expressions [18] includes more frequent phrases compared to other sources, while the dictionary of Russian idiomatics [12] contains the least recurrent units.

With the exception of a few collocations (*bol'shoye znacheniye* 'great importance', *bol'shoy uspekh* 'great success', *vyssheye obrazovaniye* 'higher education' and *nervnaya sistema* 'nervous system' the examples turn out to be low-frequency in all three corpora. The hypothesis is confirmed that with the decrease of the dictionary index, the relative frequencies of collocations in the corpus decrease (with the exception of unique collocations in the dictionary [18], whose frequencies, on the contrary, exceed the others). The presence of collocations in several dictionaries indicates their higher frequencies and hence possible prediction by automatic methods.

6 Conclusion

In our study we examined the Russian collocations which were extracted from six dictionaries. Their quantitative characteristics obtained on corpora of different volumes show that the analyzed examples turn out to be low-frequency and demonstrate their ambiguous nature. The overwhelming majority of dictionary collocations are unique, i.e. presented in only one dictionary; hence such items are difficult to be identified in corpora by using automatic methods.

The issue of data volume deserves much more attention, and the very phenomenon of collocability must be investigated in larger corpora as small volume does not show

any occurrences for a number of collocations. Automatically crawled large corpora reveal more fascinating findings as well as peculiarities obscured in smaller text collections. Hence we can assume that machine learning algorithms that process word combinations should be based on large datasets counting several bln tokens.

The number of dictionary collocations depends on the dictionaries used. Unfortunately, despite the processing of several sources, the volume of the extracted data is still insufficient, therefore it is important to analyze other dictionaries and lexicographic sources and extract examples from them. Explanatory dictionaries may contain set expressions in other parts of dictionary entries (in the texts of quotations or illustrative examples), therefore, their further analysis is necessary, which will be performed at the next stage of our work. In future we plan to consider other resources and to study collocations based on other syntactic models.

References

1. Atkins, S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford U.P., Oxford (2008).
2. Berkov, V.: Bilingual lexicography [Dvuyazychnaya leksikografiya]. 2nd edition. AST, Moscow (2004).
3. Borisova, E.: A Word in a Text. A Dictionary of Russian Collocations with English-Russian Dictionary of Keywords [Slovo v tekste. Slovar' kollokatsiy (ustoychivykh sochetaniy) russkogo yazyka s anglo-russkim slovarem klyuchevykh slov]. Filologiya, Moscow (1995).
4. Calzolari, N., Fillmore, Ch., Grishman, R., Ide, N., Lenci, A., MacLeod, C., Zampolli, A.: Towards Best Practice for Multiword Expressions in Computational Lexicons. In: Proceedings of LREC – 2002, 1934–1940 (2002).
5. The Dictionary of the Russian Language [Slovar' russkogo jazyka v 4 tomakh]. Yevgen'yeva, A. P. (ed.-in-chief). Vol. 1–4, 2nd edition, revised and supplemented. Russkij jazyk, Moscow (1981–1984).
6. Fontenelle, T.: Collocation acquisition from a corpus or from a dictionary: a comparison. In: Proceedings I-II Papers submitted to the 5th EURALEX International Congress on Lexicography in Tampere, 221–228 (1992).
7. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In: Proceedings of the 7th International Corpus Linguistics Conference CL 2013, the United Kingdom, July 2013, 125–127 (2013).
8. Khokhlova, M.: Building a Gold Standard for a Russian Collocations Database. In: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts. Ljubljana, 863–869 (2018).
9. Khokhlova M.: Collocations in Russian Lexicography and Russian Collocations Database. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association, 3191–3199 (2020).
10. Kilgarriff, A.: Comparing Corpora. International Journal of Corpus Linguistics, 6 (1), 97–133 (2001).
11. Klyshinsky, E., Khokhlova, M.: In Search of Lost Collocations: Combining Measures to Reach the Top Range. In: Internet and Modern Society: Proceedings of the International Conference IMS-2017 (St. Petersburg; Russian Federation, 21–24 June 2017). Radomir V. Bolgov, Nikolai V. Borisov, Leonid V. Smorgunov, Irina I. Tolstikova, Vic-

tor P. Zakharov (eds.). ACM International Conference Proceeding Series, 160–163. ACM Press, N.Y. (2017).
12. Kustova, G.: Dictionary of Russian Idiomatic Expressions [Slovar' russkoyj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni] (2008), http://dict.ruslang.ru, last accessed 2020/10/14.
13. The Large Explanatory Dictionary of the Russian Language [Bol'shoy tolkovyy slovar' russkogo yazyka]. Kuznetsov, S. (ed.). Norint, St. Petersburg (1998).
14. Lukashevich, N., Dobrov, B., Chuyko, D.: Selecting word phrases for an automatic text processing system dictionary [Otbor slovosochetaniy dlya slovarya sistemy avtomaticheskoy obrabotki tekstov]. In: Computational linguistics and intellectual technologies: Proceedings of Int. Conf. "Dialog–2008", 339–344. RSUH, Moscow (2008).
15. Macmillan Dictionary, https://www.macmillandictionary.com, last accessed 2020/10/14.
16. Oubine, I.: Dictionary of Russian and English Lexical Intensifiers [Slovar' usilitel'nykh slovoso-chetaniy russkogo I angliyskogo yazykov]. Russian Language, Moscow (1987).
17. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora. Association for Computational Linguistics, 1–6 (2000).
18. Reginina, K., Tjurina, G., Shirokova, L.: Set Expressions of the Russian Language. A Reference Book for Foreign Students [Ustoychivye slovosochetaniya russkogo yazyka: Uchebnoye posobiye dlya studentov-inostrantsev]. Shirokova, L. I. (ed.). Moscow (1980).
19. Russian National Corpus, http://ruscorpora.ru, last accessed 2020/10/14.
20. Wehrli, E., Seretan, V., Nerima, L., Russo, L.: Collocations in a rule-based MT system: A case study evaluation of their translation adequacy. In: Proceedings of the 13th Annual Conference of the EAMT, Barcelona, 128–135 (2009).