A Case Study of High-Frequency Dictionary Collocations in a Spoken Corpus

Maria Khokhlova

St Petersburg State University,
Universitetskaya emb. 7-9-11, 199034 St Petersburg, Russia
m.khokhlova@spbu.ru

Abstract. Linguists rarely focus their attention on spoken corpora to study collocations, but these resources can suggest valuable examples. This article discusses the adj-noun frequency collocations from the Russian collocation database that constitute a gold standard. The aim is to compare the usage of collocations on the material of the oral and written corpora. The results show that low frequencies characterize dictionary collocations, and in most cases, the occurrences are adjacent combinations that do not include other words.

Keywords: Collocations · Spoken corpora · Evaluation · Dictionaries · Russian language

1 Introduction

In numerous studies, MWEs, collocations, and other set phrases were considered on the material of exclusively written texts and mainly from the point of view of their frequency. Oral data remained outside the scope of these works, which can be objectively explained by small volumes of oral texts available to researchers until recently, as well as the laboriousness of their processing.

Our paper focuses on the following questions: 1) do high-frequency collocations collected from dictionaries occur in spoken texts? 2) do their frequencies differ from the ones in written corpora?

The paper is structured as follows. The Introduction presents the basic idea of the research. The next section provides a brief overview of the spoken corpora. Section 3 discusses the methods and relevant notions essential for the analysis. The next section examines the experiment results, while the conclusion ends the paper and offers future perspectives.

2 Spoken Russian Corpora

Spoken corpora are not as common as their written counterparts since their building is a difficult task. However, we cannot overestimate their importance while they provide valuable data. There are not so many projects for Russian that

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2021, pp. 161–166, 2021. © Tribun EU 2021

focus on collecting oral data. Most of the existing ones are of a small volume and were compiled for a particular task (for example, to study learners' speech).

The Spoken Corpus of Russian (SCR) is a part of the Russian National Corpus [14] and has various types of annotation (morphological and lexical features and textual information). It includes transcripts of recordings of public and private oral speech, as well as film transcripts, and comprises about 13.4 mln tokens.

The project "Night Dream Stories and Other Corpora of Oral Speech" gave birth to several spoken corpora [12]. The first one comprises stories about night dreams that were retold by children and teenagers; its volume is 14,000 tokens. The second corpus consists of 17 stories described by adult residents of Novosibirsk (from 19 to 70 years old) about exciting events in their lives (5,000 tokens). The last collection includes 40 stories presented by adults (from 18 to 60 years old) about funny incidents in their lives (10,000 tokens).

The Corpus of Russian Oral Speech was compiled to study the processes of speech perception by native speakers; its texts have spelling annotation, as well as acoustic and phonetic transcription [2]. Currently, its total volume goes beyond 22,000 tokens, representing different styles of speech: professional voice-over reading, reading by native speakers, spontaneous monologue speech, and children's speech.

The ORD Speech Corpus ("One Day of Speech") was built using the method of long hours monitoring [10]. It includes data from 128 speakers and more than 1,000 interlocutors representing different social groups in St Petersburg. The whole length of the recordings is 1,450 hours; their transcribed version reaches over 1 mln tokens.

3 Methods

The statistical patterns of collocability cannot be considered without linguistic parameters, which show the real usage of word combinations in texts. As reference data, we will focus on collocations obtained by us earlier (see, for example, [6]) and constituting the so-called "gold standard". From the Russian collocations database described in [5], we selected 50 items with different dictionary indices, i.e. which are present in explanatory and specialized dictionaries ([1], [3], [4], [7], [8], [9], [11], [13]). The first group has the dictionary index equal to 5, which means that five dictionaries describe these collocations. In contrast, the examples from the second group were found only in two dictionaries. We proceed from the fact that collocations from the first group show high frequency in lexicographic resources and hence are highly reproducible in speech. Both groups represent the adj-noun structural model. Further, we considered occurrences of these collocations in the SCR and the written disambiguated subcorpus of RNC (6 mln tokens).

In order to establish how native speakers recognize collocations, it is necessary to collect additional information about their usage in texts. These parameters include not only information already available about frequencies or parts of speech (that is, standard statistical values applied at the text or entire corpus level) but also previously unexplored parameters of the behavior of units, for example, at the clause level. We speculate that any semantic shift within a collocation (e.g., semantic non-compositionality) deals with features that may be inferred from corpus data. One of them is permeability, i.e., the ability of a collocation to be split by a foreign token in-between. Hence we will study the representation of this characteristic that can be found in corpus examples. We will consider not only adjacent bigrams but also their distance equivalents (for example, *polnaya svoboda* "complete freedom" and *polnaya i bezgranichnaya svoboda* "complete and unlimited freedom").

4 Results

4.1 Dictionary indices 5 and 2

The majority of collocations from the first group were found in specialized dictionaries. One item was described in explanatory dictionaries, namely, *zhguchiy bryunet* "burning brunette" and has idiomatic features. Among the considered examples, two nouns have more than one collocation, namely, *glubokaya tishina* "deep silence", *polnaya tishina* "complete silence", *bogatyy urozhay* "rich harvest", *vysokiy urozhay* "high harvest". The most frequent collocate is *glubokiy* "deep" (8 examples), while such adjectives as *zheleznyy* "iron", *ostryy* "sharp" and *polnyy* "complete, full" show 2 examples.

The results for the group with the dictionary index 5 are shown in Table 1 (absolute frequencies). We can note a low correlation between two distributions (0.36 according to the Spearman coefficient, $p > 0.05$). However, the frequencies are small and do not differ in the corpora ($V=80$ according to the Wilcoxon test, $p > 0.05$).

For distance n-grams, we searched up to five words between a node and a collocate (the last column in Table 1). The selected collocations show low permeability. The average frequency is 0.68 and 0.80 for spoken and written texts, respectively. The following cases exemplify the longest n-grams: *tverdaya, khotya i mgnovenno sozrevshaya uverennost'* "firm, albeit instantly ripe, confidence"; *polnoy i ravnoy dlya vsekh svobody* "full and equal freedom for all".

Table 2 presents absolute frequencies for the collocations registered in two dictionaries. More than half of collocations from this group had no examples in corpora. They tend to occur rarer than the collocations mentioned above. Long n-grams were not found with only four exceptions, that are trigrams, e.g. *dlinnaya avtomatnaya ochered'* "a long gun burst", *chrezmernoye issledovatel'skoye svimaniye* "excessive research attitude", *bol'shoy vas poklonnik* "a big fan of you" and *svezhaya nemetskaya gazeta* "a fresh German newspaper".

The results suggest that both corpora are not sufficient in their volume to study collocations. The collocations from the second group tend to occur only in their adjacent forms.

Table 1: Results for the dictionary index 5.

Collocation		Freq (SCR)	Freq (RNC)	Dist(SCR)
bogatyy urozhay	"rich harvest"	3	3	1
bol'shoy avtoritet	"great authority"	12	0	1
vysokiy urozhay	"high yield"	5	0	0
glubokaya blagodarnost'	"deep gratitude"	4	2	1
glubokoye vliyaniye	"deep influence"	0	2	0
glubokoye znaniye	"deep knowledge"	6	7	1
glubokiy interes	"deep interest"	1	3	0
glubokiy krizis	"deep crisis"	3	3	4
glubokaya tishina	"deep silence"	3	3	0
glubokoye ubezhdeniye	"deep refuge"	25	9	1
glubokoye chuvstvo	"deep feeling"	1	5	1
goryachaya lyubov'	"hot love"	6	1	1
grubaya oshibka	"big mistake, blunder"	12	5	0
zhguchiy bryunet	"hot brunette"	2	3	0
zheleznaya distsiplina	"iron discipline"	2	8	0
zheleznyy kharakter	"iron character"	2	0	0
krepkaya druzhba	"strong friendship"	2	1	1
nesterpimaya bol'	"unbearable pain"	1	4	0
ozhestochennyy boy	"fierce battle"	11	2	0
ostraya kritika	"sharp criticism"	1	2	1
ostraya nuzhda	"urgent need"	0	0	0
polnaya svoboda	"complete freedom"	22	13	2
polnaya tishina	"complete silence"	9	21	0
tverdaya uverennost'	"firm confidence"	6	4	0
tyazhelaya bolezn'	"serious illness"	11	9	2

4.2 Textual and syntactic characteristics

Based on the main corpus of the RNC and its textual annotation, it was found that the selected collocations are more characteristic of journalistic texts (compared to fiction). The use of the collocations prevails in the position of the end of the clause. Obviously, it is impossible to use the considered units in plural since abstract nouns cannot be counted, so most examples were found in the singular form. It can also be noted that examples of collocations are more typical for texts written by men.

5 Conclusion

The analyzed collocations are characterized by low occurrences in the corpus. It can be assumed that, on the one hand, dictionary collocations are rare linguistic phenomena, and on the other hand, dictionaries themselves are not an ideal source of data compared with corpora.

The results of this and future work in this direction are essential for developing applications related to speech processing. Creating a full-fledged descrip-

Table 2: Results for the dictionary index 2.

Collocation		Freq (SCR)	Freq (RNC)	Dist(SCR)
bezmernaya glubina	“immeasurable depth”	0	0	0
bezumnaya otvetstvennost′	“terrible responsibility”	0	0	0
bol′shoy poklonnik	“big fan”	7	8	1
vysokiy spros	“high demand”	1	2	0
gromadnaya bystrota	“tremendous speed”	0	0	0
dlinnaya ochered′	“long queue”	6	11	1
doskonal′nyy analiz	“thorough analysis”	0	0	0
isklyuchitel′naya vezhlivost′	“exceptional politeness”	0	1	0
kolossal′naya stoimost′	“colossal cost”	0	0	0
nastoychivaya pros′ba	“insistent request”	1	1	0
nezyblemyy avtoritet	“unshakable authority”	0	0	0
neissyakayemaya vera	“inexhaustible faith”	0	0	0
neistovyy azart	“frantic excitement”	0	0	0
ogromnoye zhelaniye	“great desire”	6	1	0
ogromnyy rost	“huge growth”	5	5	0
ostraya zhalost′	“keen pity”	0	1	0
plamennaya strast′	“fiery passion”	0	0	0
polnoye bezvetriye	“complete calm”	0	1	0
porazitel′naya tishina	“astonishing silence”	1	0	0
reshitel′nyy kharakter	“decisive character”	0	1	0
svezhaya gazeta	“fresh newspaper”	5	1	1
tverdoye obyazatel′stvo	“firm commitment”	0	0	0
tyzhelyy krizis	“severe crisis”	2	0	0
chistoye bezumiye	“pure madness”	2	2	0
chrezmernoye vnimaniye	“excessive attitude”	0	0	1

tive base of Russian oral speech requires a description devoted to stable word combinations. This part is a necessary condition for developing those areas of linguistics and information technologies that take into account a speaker and his (or her) speech behavior.

Acknowledgements. The database was compiled with the support from the Russian Science Foundation (Project No. 19-78-00091). The work on collocation evaluation in spoken corpora was supported by St Petersburg State University, project No. 75254082 “Modeling of Russian megalopolis citizens′ communicative behavior in social, speech and pragmatic aspects using artificial intelligence methods”.

References

1. 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).

2. Corpus of Russian Oral Speech, http://russpeech.spbu.ru/. Last accessed 14 Nov 2021.
3. Deribas, V.: Verb–Noun Collocations in Russian [Ustoychivye glagol'no-imennye slovosochetaniya russkogo yazyka]. Russkiy yazyk, Moscow. (1983)
4. 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).
5. Khokhlova, M.: Building a Gold Standard for a Russian Collocations Database. In: Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, pp. 863—-869. Ljubljana (2018)
6. Khokhlova, M.: Collocations in Russian Lexicography and Russian Collocations Database. In: Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France, pp. 3191–3199. European Language Resources Association (2020)
7. Kustova, G.: Dictionary of Russian Idiomatic Expressions [Slovar' russkoyj idiomatiki. Sochetaniya slov so znacheniyem vysokoy stepeni] (2008), http://dict.ruslang.ru. Last accessed 14 Nov 2021.
8. The Large Explanatory Dictionary of the Russian Language [Bol'shoy tolkovyy slovar' russkogo yazyka]. S.A. Kuznetsov (ed.). Norint: St. Petersburg (1998).
9. Mel'čuk, I., Zholkovsky, A.: Explanatory Combinatorial Dictionary of Modern Russian [Tolkovo-kombinatornyy slovar russkogo yazyka]. Vienna. (1984)
10. ORD Speech Corpus ("One Day of Speech"), https://ord.spbu.ru/. Last accessed 14 Nov 2021.
11. Oubine, I.: Dictionary of Russian and English Lexical Intensifiers [Slovar' usilitel'nykh slovoso-chetaniy russkogo I angliyskogo yazykov]. Russian Language: Moscow (1987).
12. Project "Night Dream Stories and Other Corpora of Oral Speech", http://spokencorpora.ru. Last accessed 14 Nov 2021.
13. 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).
14. Russian National Corpus, http://ruscorpora.ru. Last accessed 14 Nov 2021.