Analysis of Collocations in Russian: Corpus vs Dictionary

Maria Khokhlova
Saint-Petersburg State University

The paper discusses the results of an experiment in collocation extraction in a corpus of Russian texts. The data obtained is compared to the data given for set expressions in modern Russian dictionaries in order to analyze from the standpoint of traditional lexicography what kind of phrases can be received by such an approach. The paper also explores the role of statistical measures for extracting collocations in Russian.

1. Introduction

The issue of collocability is highly important in modern linguistics. The investigation of collocability is closely connected to the study of syntagmatics as a deeper level of lexical relations. Though several criteria have been taken into consideration to describe these relations (lexical restrictions, repeatability etc), the boundary between free and set phrases often has been placed quite subjectively.

Although the term *collocation* appeared in Russian linguistics long ago (Akhmanova 1966), it is not generally recognized by Russian scholars and even is absent in the Russian Linguistic Dictionary (Yartseva 1990). There is no agreement among scholars how to call such lexical units; cf. "set verbal-noun expressions" (Deribas 1983), "analytic lexical collocations" (Teliya 1996), etc. The majority of modern authors understand under a collocation a statistically set phrase. Collocations can be put between free phrases and idioms on a scale of phrases. The monograph (Borisova 1995a) has proved to be the first work in Russian linguistics, completely devoted to the research of the concept of collocation on a material of Russian. One of the key properties of a collocation is "the impossibility of prediction of such combinations on the basis of meanings of their components" (Borisova 1995a: 13). In the "Meaning–Text" theory (Mel'čuk 1974) collocations are considered as a subclass of more extensive class of set phrases, or phrasemes.

In our meaning collocations should be defined simply as statistically set phrases.

The methods for collocation extraction proposed in most works have not been evaluated so far whether they can be applicable to Russian, and if yes, to what degree. Also there's a question what types of set phrases they allow to retrieve. The explanatory dictionaries do not always consistently reflect the information about set phrases.

According to some scientists (Mel'čuk 1960) the property of stability (for phrases) is inherent to all word combinations. A threshold of stability should be chosen to range them, above which a word combination can be called a set phrase.

We share this opinion and suggest to use different statistical measures to range word combinations, e.g. from set to free phrases.

The analysis of phrases (that is bigramms) obtained from a corpus by the statistical method can contribute to the understanding the issue what the interrelation between collocability and stability is.

2. The analysis of collocation extraction in Russian

The probabilistic nature of language is beyond any doubt. Statistical methods are widely used in corpus linguistics. Nowadays there are several ways in linguistics to calculate the degree of collocates' coherence. There are different measures based on calculation of a degree of words nearness in a text, namely, MI (mutual information), t-score, log-likelihood (henceforth LL), z-score, chi-square, Dice coefficient, odds ratio etc. (Evert 2004).

Probably, various measures extract word combinations of different types. For example, it seems to us plausible that t-score extracts not so much syntagmatically connected combinations as the common lexis of lexico-semantic fields, e. g. patient, hospital, doctor, medicine. It seems interesting to attempt to reveal functionality of different measures.

There are also other statistical methods supplementing measures of association. For example, in (Cvrček 2006) there is a description of a method to extract collocations based on bigrams rank distribution. The results of the method's application are compared with results obtained by measures of association.

So far the methods for collocation extraction proposed in most works have not been evaluated in the perspective of whether they can be applicable to Russian, and if yes, to what degree. Unfortunately there are not so many Russian corpora having this kind of tools implemented for them. We can mention here only the corpora built at the University of Leeds by S.A. Sharoff[1] and the enthusiastic project by AOT team[2].

The aim was to carry out a number of experiments in order to find a suitable measure of association for different classes of set phrases; to define opportunities of statistical methods as a whole and several measures in particular; to find ways of combination of statistical and semantic-syntactical methods in collocation extraction.

During the experiments the following ideas were tested:

- To what degree the proposed methods can be applicable to Russian;
- Whether the given methods allow revealing other classes of set phrases.

We have chosen the collocations of 19 nouns that don't have homonyms as material for our research. The nouns have been selected on the principle of their sufficient high frequency (see the Electronic Frequency Dictionary of Russian by S. Sharoff (Sharoff 2002)): *vlast'* "power", *vnimaniye* "attention", *vozmozhnost'* "opportunity", *voyna* "war", *vopros* "question", *dozhd'* "rain", *zhizn'* "life", *zakon* "law", *lyubov'* "love", *mesto* "place", *mneniye* "opinion", *mysl'* "thought", *noch* "night", *otvet* "answer", *pomosc* "help", *radost'* "joy", *slovo* "word", *sluchay* "case", *smysl* "sense".

The research has been led on the corpus of Russian newspapers (78 million words) created at the University of Leeds (Great Britain) under the guidance of S. Sharoff.

We examined bigrams as examples of collocations for each word, i.e. combinations of a given word with a word which is on its right or on its left. We compared them to the entries for these nouns in the Dictionary of Collocations (Borisova, 1995b), in the explanatory dictionaries of Russian (the Dictionary of Modern Russian (*Slovar' sovremennogo russkogo literaturnogo jazyka*, 1948-1965); the Big Academy Dictionary of Russian (*Bol'shoj akademicheskij slovar' russkogo yazyka*, 2004-2007), the Dictionary of Russian (*Slovar' russkogo jazyka*, 1957-1961)) and in the Dictionary of Synonyms and Similar Expressions (Abramov 2006).

It is necessary to mention that each element of the corpus which stands before or after a blank including punctuation marks is considered a token. Therefore there are combinations of verbs and punctuation marks, too.

2.1. *Results for Log-Likelihood*

For LL measure the following results were received.

1763 bigrams were found in total. Among them there were:

- 47 bigrams fixed in two or more dictionaries;
- 79 bigrams fixed only in [10];
- 48 bigrams fixed only in [13];
- 20 bigrams fixed only in [14];
- 11 bigrams fixed in [12];
- 6 bigrams fixed only in [11].

[1] *http://corpus1.leeds.ac.uk/ruscorpora.html*.

[2] *http://aot.ru/demo/bigrams.html*.

Also there were 15 combinations with punctuation marks.
Values of LL proved to be the largest for the collocations found in two or more dictionaries.

№	Collocation	Joint	Freq1	Freq2	LL score	Concordance
1.	обращать внимание (pay attention)	4118	12455	19714	14361.30	Examples
2.	этот вопрос (this question)	4684	476434		5130.73	Examples
3.	на вопрос (to the question)	5887	1105092		4786.25	Examples
4.	давать возможность (enable)	1904	60300		3892.76	Examples
5.	особый внимание (special attention)	1427	16112	19714	3848.17	Examples
6.	иметь место (take place)	1899	60000		3568.69	Examples
7.	весь жизнь (the whole life)	2161	130350	59718	3441.61	Examples
8.	в ответ (in response)	3543	2534398		3419.57	Examples
9.	е место (corpus failure)	1307	9896		3411.37	Examples
10.	общественный мнение (public opinion)	1066	18429		2841.35	Examples
11.	иметь возможность (have a chance)	1439	60000		2731.97	Examples
12.	привлекать внимание (attract attention)	971	9401	19714	2687.61	Examples
13.	рассматривать вопрос (consider the question)	1242	17005		2572.59	Examples
14.	первый место (the first place)	1665	111613		2499.13	Examples
15.	оказывать помощь (help)	977	17711		2491.59	Examples
16.	решать вопрос (solve a question)	1486	47147		2446.03	Examples
17.	высказывать мнение (express an opinion)	774	8475		2239.46	Examples
18.	федеральный закон (federal law)	1026	37679	49277	2152.57	Examples
19.	второй место (the second place)	1208	45762		2150.41	Examples
20.	в ночь (at night)	2363	2534398		2098.42	Examples
21.	такой мнение (such an opinion)	1227	150108		2066.85	Examples
22.	всякий случай (any case)	711	11480		2062.06	Examples
23.	свое мнение (own opinion)	853	35085		1892.07	Examples
24.	третий место (the third place)	833	19293		1686.15	Examples
25.	на место (into place)	2506	1105092		1539.12	Examples

Table 1. The first 25 collocations according to LL

2.2. *Results for MI*

1755 bigrams were found in total. Among them there were:
- 68 bigrams fixed in two or more dictionaries;
- 73 bigrams fixed only in [10];
- 27 bigrams fixed only in [13];
- 13 bigrams fixed only in [14];
- 9 bigrams fixed in [12];
- 25 bigrams fixed only in [11].

Also there were 11 combinations with punctuation marks.

№	Collocation	Joint	Freq1	Freq2	MI score	Concordance
1.	накрапывать дождь (drizzle)	7	19		14.95	*Examples*
2.	моросить дождь (drizzle)	19	67		14.57	*Examples*
3.	мифогенный любовь (mythogenic love)	4	4		14.39	*Examples*
4.	проливный дождь (downpour)	48	206		14.29	*Examples*
5.	метеорный дождь (meteor shower)	4	32		13.39	*Examples*
6.	варфоломеевский ночь (massacre of St. Bartholomew)	12	16		13.25	*Examples*
7.	метеоритный дождь (meteorite shower)	4	39		13.11	Examples
8.	вальпургиев ночь (Walpurgis-night)	5	8		12.99	*Examples*
9.	неослаблять внимание (give attention)	5	5	19714	12.57	*Examples*
10.	сакцентировать внимание (place emphasis)	5	5	19714	12.57	*Examples*
11.	утвердительный ответ (affirmative answer)	43	76		12.44	*Examples*
12.	здравый смысл (common sense)	240	1066		12.35	*Examples*
13.	замолвить слово (put in a word)	13	37		12.28	*Examples*
14.	нечаянный радость (unexpected joy)	20	219		12.18	*Examples*
15.	закрадываться мысль (creep, about a thought)	11	87		12.05	*Examples*
16.	кратковременный дождь (light rain)	32	702		11.94	*Examples*
17.	мелькнуть мысль (flit, about a thought)	17	146		11.93	*Examples*
18.	неразделять любовь (undivided love)	12	68		11.89	*Examples*
19.	крамольный мысль (rebellious thought)	12	106		11.89	*Examples*
20.	развернутый ответ (detailed answer)	11	30		11.82	*Examples*
21.	акцентировать внимание (place emphasis)	204	349	19714	11.79	*Examples*
22.	узурпировать власть (usurp power)	20	54		11.76	*Examples*

23.	шальной мысль (crazy thought)	12	118		11.74	*Examples*
24.	бытовать мнение (there is an opinion)	131	346		11.71	*Examples*
25.	платонический любовь (Platonic love)	6	39		11.69	*Examples*

Table 2. The first 25 collocations according to MI

Bigrams, extracted by MI and t-score also correlate with data of dictionaries.

Values of the MI measure are the largest ones for the collocations found only in *Slovar' russkogo jazyka*, 1957-1961, and also found in two or more dictionaries. After examination of the list of results we found out, that only two combinations were retrieved (and both were not fixed in the dictionary of collocations) within a range from 0 to 1 (according to the value of MI). It allows us making a conclusion that the combination is statistically insignificant if the MI appears in the given interval. We suppose that phrases with MI equal to 5 or higher prove to be set phrases.

2.3. *Results for t-score*

1755 bigrams were found in total. Among them there were:

71 bigrams fixed in two or more dictionaries;

73 bigrams fixed only in [10];

22 bigrams fixed only in [13];

14 bigrams fixed only in [14];

8 bigrams fixed in [12];

23 bigrams fixed only in [11].

Also there were 20 combinations with punctuation marks.

№	Collocation	Joint	Freq1	Freq2	T score	Concordance
1.	на вопрос (to the question)	5887	1105092		70.20	*Examples*
2.	этот вопрос (this question)	4684	476434		65.28	*Examples*
3.	обращать внимание (pay attention)	4118	12455	19714	64.14	*Examples*
4.	в ответ (in response)	3543	2534398		55.19	*Examples*
5.	весь жизнь (the whole life)	2161	130350	59718	45.72	*Examples*
6.	в ночь (at night)	2363	2534398		44.60	*Examples*
7.	на место (into place)	2506	1105092		43.65	*Examples*
8.	давать возможность (enable)	1904	60300		43.34	*Examples*
9.	иметь место (take place)	1899	60000		43.18	*Examples*
10.	первый место (the first place)	1665	111613		40.01	*Examples*
11.	решать вопрос (solve a question)	1486	47147		37.99	*Examples*
12.	особый внимание (special attention)	1427	16112	19714	37.71	*Examples*
13.	иметь возможность (have a chance)	1439	60000		37.60	*Examples*
14.	на помощь (in help)	1613	1105092		36.48	*Examples*
15.	е место (corpus failure)	1307	9896		36.07	*Examples*

16.	рассматривать вопрос (consider the question)	1242	17005		35.02	*Examples*
17.	такой мнение (such an opinion)	1227	150108		34.54	*Examples*
18.	второй место (the second place)	1208	45762		34.37	*Examples*
19.	общественный мнение (public opinion)	1066	18429		32.59	*Examples*
20.	свой жизнь (own life)	1164	205621	59718	32.48	*Examples*
21.	федеральный закон (federal law)	1026	37679	49277	31.84	*Examples*
22.	оказывать помощь (help)	977	17711		31.18	*Examples*
23.	привлекать внимание (attract attention)	971	9401	19714	31.11	*Examples*
24.	получать возможность (get an opportunity)	932	79406		29.98	*Examples*
25.	быть возможность (there is an opportunity)	1127	664975		29.39	*Examples*

Table 3. The first 25 collocations according to t-score

The combinations that have large values of t-score prove to be rather frequent while, unlike the previous measures, one of their parts is a preposition or a pronoun, since t-score "attracts" frequent words. And also there were more bigrams (in comparison with other measures) in which a punctuation mark is one of their parts.

We confirmed the hypothesis that t-score allows to retrieve collocations which have very frequent words, and also punctuation marks as their constituents. Thus, as well as for other languages, it is true for Russian that words with large values of t-score are frequent and can be combined with a large number of words. The right context reveals more combinations with punctuation marks than the left one.

2.4. *Conclusion*

The analysis of the data received shows that the majority of collocations (phrasemes), fixed in dictionaries, stand in the top part of the list, i.e. their parts co-occur very often.

The combinations which had not been fixed in the dictionaries before were also retrieved during the experiment. The analysis of these combinations that show both high and low values of measures of association (one or several), reveals, that bigrams which stand on the top of the list of collocations (sorted on decrease), with some degree of probability prove to be set phrases and, hence, can be included in the dictionary. The overwhelming majority of collocations that stand in the bottom part of the list prove to be free phrases.

Also it is possible to note the combinations recognized by us as collocations, but not listed in dictionaries. In case of large value of a measure for such combinations one can say to a certain degree that they belong to a class of set phrases: for example, *centr vnimaniya* "the focus of attention", *ukromnoye mesto* "secluded corner", *pokonchit' zhizn'* "to commit suicide", *drakonovskiy zakon* "draconian law", *scekotliviy vopros* "ticklish question" etc.

3. Further work

The experiment has shown the possibility to apply statistical tools in order to extract collocations from Russian texts. The results of this work (and the data about word collocability based on statistical measures), first of all, can be applied to dictionary compiling. The statistical collocations which are extracted by measures of association, and not fixed in a dictionary, can be added to the existing dictionaries after a careful analysis.

Yet it is difficult to decide on one statistical measure that could give allegedly perfect results. The work done by us has demonstrated significant discrepancies in different measures values

for the same collocations. This points to the fact that the comparative functionality analysis of different measures should be continued.

Though our experiments and results are promising, they are only the first step to demonstrate statistical methods and adopt them for the linguistic praxis. Then the received results must be manually processed within the framework of traditional linguistics, and compared to the data from dictionaries.

We are convinced that probabilistic-statistical methods should be specified and programming tools developed in the corpora study frameworks.

Acknowledgements

I am grateful to my supervisor Victor Zakharov (Saint-Petersburg State University) for his great support, and encouragement in this work. And I would like also to thank Serge Sharoff for the opportunity to work with the corpus.

References

Abramov, N. (2006). *Slovar' russkikh sinonimov i skhodnykh po smyslu vyrazhenij*. Moskva.

Akhmanova, O. S. (ed.) (1966). *Slovar' lingvisticheskikh terminov*. Moskva.

Bol'shoi akademicheskii slovar' russkogo iazyka: vol. 1-6. Sankt-Peterburg, 2004-2007. (BAS-25).

Borisova, E. G. Kollokatsii (1995a). *Chto eto takoe i kak ikh izuchat'*. Moskva.

Borisova, E. G. (1995b). *Slovo v texte. Slovar' kollokatsij (ustojchivykh slovosochetanij) russkogo jazyka s anglo-russkim slovarem kljuchevykh slov*. Moskva.

Cvrček, V. (2006). "Metoda zjišťování kolokační platnosti frekventovaných bigramů pomocí ranku". In Čermák F. (ed.). *Kolokace. Ústav Českého národního korpusu*. Praha. 36-55.

Deribas, V. M. (1983). *Ustojchivye glagol'no-imennye slovosochetanija russkogo jazyka*. Moskva.

Evert, S. (2004). *The Statistics of Word Cooccurences Word Pairs and Collocations*. Ph.D. thesis. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Mel'čuk, I. A. (1960). "O terminakh 'ustojchivost' i 'idiomatichnost'". *Voprosy jazykoznanija* 4. 73-80.

Mel'čuk, I. A. (1974). *Opyt lingvisticheskoy teorii "Smysl – Tekst"*. Moskva.

Sharoff, S. (2002). *Chastotnyj slovar' slovar' russkogo jazyka* [on-line]. *http://www.artint.ru/projects/frqlist.asp* [Access date: 23 May 2007]

Slovar' sovremennogo russkogo literaturnogo iazyka: vol, 1-17. Moskva. 1948-1965. (BAS-17).

Slovar' russkogo iazyka: vol. 1-4. Moskva. 1957-1961. (MAS).

Teliya, V. N. (1996). *Russkaja frazeologija: semanticheskij, pragmaticheskij i lingvokul'torologicheskij aspekty*. Moskva.

Yartseva, V. N. (ed.) (1990). *Lingvisticheskiy enciklopedicheskij slovar'*. Moskva.