Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

LARGE CORPORA AND FREQUENCY NOUNS

Khokhlova M. V. (m.khokhlova@spbu.ru)

St. Petersburg State University, St. Petersburg, Russia

The paper describes a new branch in corpus linguistics that deals with building and using large corpora. We introduce several new large Russian corpora that have recently become available. The paper gives a survey of the given corpora and analyzes a number of Russian nouns across the following corpora of different sizes: the Russian Web corpus by S. Sharoff (187.97 mln tokens), ruTenTen (18.28 bln tokens) and its sample (1.25 bln tokens). The research focuses on the discussion on these corpora, their comparison and the study of frequency properties for the high- and low frequency Russian nouns comparing them with data published in the Frequency Dictionary. The analysis shows the lists presented in the frequency dictionary of Russian differs from the corpus data depending on types of the nouns.

Key words: text corpora, web corpus, frequency, frequency nouns, statistics, big data

БОЛЬШИЕ КОРПУСА И ЧАСТОТНЫЕ СУЩЕСТВИТЕЛЬНЫЕ

Хохлова М. В. (m.khokhlova@spbu.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Статья посвящена использованию больших корпусов, которые стали активно развиваться в последнее время. В ней представлены результаты исследования частотных существительных русского языка на материале корпусов разных объемов. В статье дается обзор русских корпусов большого объема. Обсуждаются различия между корпусами и характеристики высоко- и низкочастотных существительных, проводится их анализ с данными частотного словаря русского языка. Анализ показывает, что данные, приведенные в словаре, и результаты, полученные на корпусной основе, отличаются для разного типа существительных.

Ключевые слова: корпус текстов, Интернет-корпус, частота, частотные существительные, статистика, большие данные

Introduction

The idea of corpora that contain big data has attracted scholars' attention for a long time. However, it was the availability of wide technical opportunities that gave impetus to the development of a new research field which enables researchers to collect data automatically from the Internet (see, for example [Kehoe, Renouf 2002; Kilgarriff, Grefenstette 2003; Belikov, Selegey, Sharoff 2012]). Researchers found it attractive to make statistical inferences and to verify the results on increasingly larger scope of data. 1 mln or 100 mln tokens are not thresholds for the corpora. At the same time access to big corpora provokes new challenges: what can we see with big data and how does it affect the results? Do linguists actually need large corpora or their appetites can be satisfied with less data? Can small corpora be viewed as little big ones [Sinclair 2004]?

1. Large Russian Corpora

Large corpora with volumes exceeding 100 mln tokens have appeared just recently. This idea is closely related to the technical resources and thus the gradually changing paradigm in corpus linguistics moving forward from "manual" approach to more automatic one. Nowadays one can speak about two types of corpora, some authors distinguish between three types [Belikov, Selegey, Sharoff 2012]. For the Russian language the most famous and popular corpus of the first type is the National Russian Corpus; altogether its subcorpora comprise 600 mln words. This corpus was built according to the "classic" style, i.e. linguists selected relevant texts, annotated them and included into the database. Corpora of the second type are collected automatically from the Web (frankly speaking, to a certain degree that holds true for the first type also). For the Russian language we can name "The General Internet-Corpus of Russian" that now exceeds 15,000 mln words and its authors aim at 50,000 mln words [Piperski et al. 2013]. Also, there is a number of corpora made within the Aranea project, which includes a Russian corpus called Araneum Russicum Maius & Minus [Benko 2014]. The Russian Web corpus was collected by S. Sharoff [Sharoff 2006] using the methodology for crawling web-based texts outlined in [Baroni, Bernardini 2004]. The developers use a list of 500 most frequent words in a language that are not function words and don't belong to a specific domain. At the next stage a program produces a list of queries (from 5,000 up to 8,000); each of them consists of 4 words from the list. As an output there is a new list of web links, 10 top results are used for downloading files. The next step requires further postprocessing, such as encoding correction, removing duplicates (filtering pages in other languages using the same alphabet, e.g. deleting texts written in Cyrillic in Slavic languages other than Russian) and technical information. The TenTen family [Jakubíček, Kilgarriff, Kovář, Rychlý, Suchomel 2013] includes corpora of various languages of the order of 10 billion words. The ruTenTen Russian corpus is one of the biggest among them along with the English, German, French and Spanish collections. Building these corpora implies that special attention is paid to the process of de-duplication in order to delete multiple copies of the same chunks of texts and thus to reduce postprocessing. Also the crawler downloads texts that have only full sentences (not navigation data).

The above-mentioned technologies can be found attractive enabling the researchers to create corpora of different languages and not demanding the time-consuming stage of collecting texts (however, this advantage can be questioned if we remember about such inherent properties of a corpus and a sample as their representativeness and balance).

To our best knowledge, there are no large corpora studies of linguistic phenomena on the Russian data, which would come up with a comparative analysis of these corpora (e.g. "big" vs. "little" corpora or "manual" vs. "automatic"). A survey of the Aranea Russian corpora was made in [Zakharov 2015]. For the English corpora [Kilgarriff 2001] suggests the χ^2-test as a suitable measure for comparing corpora that outperforms other methods. For Chinese corpora an attempt of evaluation is described in [Shu-Kai Hsieh 2014]. Comparative analysis of frequency lists can be viewed as a similar issue to a certain degree. For Russian the method was proposed in [Shaikevich 2015] describing the *Cxy* measure.

2. Experiments

The aim of our research is to compare frequency properties of a number of Russian nouns across corpora of different sizes, to identify differences, and to analyze them. We selected several corpora that had been collected and built automatically—the Russian Web corpus by S. Sharoff (187.97 mln tokens), ruTenTen (18.28 bln tokens) and its sample (1.25 bln tokens). The latter is a subcorpus of the ruTenTen corpus that was randomly generated, so it comprises the same texts but differs in size from ruTenTen. To succeed in our study we studied properties of high- and low frequency nouns that had been selected from the dictionary.

2.1. High-Frequency Nouns: Selection

The majority of Russian texts in web corpora come from news websites, blogs, commercial websites, social media groups etc. Fiction texts are less common for such corpora; therefore, we decided to focus on high-frequency vocabulary that is associated with the above-mentioned functional styles. To this end, we compiled two word lists. One word list (see Tables 1 and 2) contained nouns that the Frequency Dictionary [Lyashevskaya. Sharoff 2009] ranked top by frequency in social and political journalism and non-fiction texts (the Frequency Dictionary provides separate frequency lists for both types of texts). Hence both lists include top 10 nouns for the given style.

Table 1. High-frequency nouns in non-fiction texts

No.	Lemma	Translation	Frequency (ipm)
1	god	year	4,624.2
2	vremja	time	2,080.5
3	čelovek	man	1,945.3
4	sistema	system	1,798.0
5	rabota	job	1,766.4
6	stat’ja	article	1,363.0
7	delo	affair	1,339.5
8	slučaj	case	1,259.0
9	process	process	1,221.8
10	vopros	question	1,180.9

Table 2. High-frequency nouns in texts belonging to social and political journalism

No.	Lemma	Translation	Frequency (ipm)
1	god	year	5589.5
2	čelovek	man	2950.1
3	vremja	time	2364.6
4	žizn’	life	1548.4
5	delo	affair	1482.0
6	den’	day	1397.8
7	rabota	job	1272.4
8	strana	country	1203.9
9	vopros	question	992.0
10	slovo	word	989.7

Tables 1 and 2 show that both lists overlap; therefore, we ended up with just 14 words on the final list (indexes indicate that a word is rated top by frequency for both journalism and non-fiction texts): *god* ‘year’$^{1.2}$, *vremja* ‘time’$^{1.2}$, *čelovek* ‘man’$^{1.2}$, *sistema* ‘system’, *rabota* ‘job’$^{1.2}$, *stat’ja* ‘article’, *delo* ‘affair’$^{1.2}$, *slučaj* ‘case’, *process* ‘process’, *vopros* ‘question’$^{1.2}$, *žizn’* ‘life’, *den’* ‘day’, *strana* ‘country’ and *slovo* ‘word’.

The other list contains nouns that belong to the so-called style-specific vocabulary (i.e. typical) [Lyashevskaya. Sharoff 2009] for either social and political journalism or non-fiction texts (see Tables 3 and 4).

Table 3. High-frequency style-specific nouns on the word list for non-fiction texts (social and political journalism excluded)

No.	Lemma	Translation	Frequency (ipm)		LL-score[1]
			Corpus	Subcorpus	
1	stat'ja	article	395.0	1,363.0	10,512
2	sistema	system	617.8	1,798.0	9,943
3	federacija	federation	258.9	1,003.1	9,329
4	process	process	371.7	1,221.8	8,639
5	risunok	picture	179.2	776.2	8,451
6	virus	virus	106.5	584.1	8,388
7	issledovanie	study	200.5	799.6	7,762
8	ispol'zovanie	usage	190.3	757.9	7,342
9	sud	court	371.1	1,153.9	7,334
10	metod	method	197.0	772.3	7,312

Table 4. High-frequency style-specific nouns on the word list for social and political journalism

No.	Lemma	Translation	Frequency (ipm)		LL-score
			Corpus	Subcorpus	
1	prezident	president	311.0	634.6	2,186
2	teatr	theatre	305.3	611.9	1,944
3	god	year	3,727.5	5,589.5	1,435
4	spektakl'	play	164.7	350.0	1,429
5	pravitel'stvo	government	277.7	531.2	1,341
6	kompanija	company	392.7	699.0	1,149
7	strana	country	725.7	1,203.9	1,135
8	fil'm	film	196.8	380.1	1,009
9	reforma	reform	133.1	273.0	963
10	vybory	elections	117.7	243.4	889

2.2. High-Frequency Nouns: Results

In our research we have also analyzed 10 top-frequency nouns in the three corpora. The Russian Web Corpus list was almost identical to the list in [Lyashevskaya. Sharoff 2009]. The results for the two other corpora are more exciting. For example,

[1] The logarithmic likelihood score is a static measure used by the authors of the dictionary to identify style-specific vocabulary. In Tables 3 and 4 the results are arranged according to this parameter.

in ruTenTen *god* 'year', *rabota* 'job', *vremja* 'time', *čelovek* 'man', *kompanija* 'company', *sistema* 'system', *sajt* 'site', *den'* 'day', *mesto* 'place' and *Rossija* 'Russia' topped the frequency ranking. The lemmata *sistema* and *kompanija* were ranked 26 and 59 respectively on the high-frequency nouns list, whereas *sajt* and *Rossija* were entirely missing on this list. In the ruTenTen subset the word *sajt* was missing, whereas the lexeme *rebënok* 'baby', ranked 22 in the Frequency Dictionary, was present on the subset.

We referred to the three corpora to study frequencies of the words on the lists (see Tables 1 and 2); you can find the results on Table 5 and Fig. 1. as well as on Table 6 and Fig. 2.

Table 5. Frequencies of nouns on the non-fiction word list (journalism excluded) calculated as per three corpora

			Frequency (ipm)			
			Frequency word list for non-fiction (journalism excluded) in the Frequency Dictionary	**Russian Web Corpus**	**ruTenTen**	
No.	**Lemma**	**Translation**			**Corpus**	**Sample**
1	god	year	4,624.2	2,220.74	3,078.97	3,076.99
2	vremja	time	2,080.5	1,761.06	1,790.84	1,793.41
3	čelovek	man	1,945.3	2,343.79	1,955.40	1,950.79
4	sistema	system	1798	527.61	998.41	1,006.66
5	rabota	job	1,766.4	885.02	1,509.37	1,510.41
6	stat'ja	article	1363	257.55	293.72	292.09
7	delo	affair	1,339.5	1,037.09	813.12	809.29
8	slučaj	case	1259	632.16	750.61	752.11
9	process	process	1,221.8	294.37	473.94	478.05
10	vopros	question	1,180.9	853.94	866.03	869.27

Table 5 and Fig. 1 show the data for nouns in Table 1. We can see that both ruTenTen charts for the corpus and the subset are identical, which means that these words have identical distribution. The frequencies, indicated in the Dictionary, are the highest, except the frequency of the lemma *čelovek* which has the highest frequency in Russian Web Corpus. All the three corpora rank the words somewhat differently from the ranking in the Dictionary—two nouns in Russian Web Corpus have the same ranking as in the Dictionary, while ruTenTen (and the subset) contains four such nouns. Moreover, both corpora agree on the ranking of the four words *vremja*, *vopros*, *process* and *stat'ja*. Spearman's rank correlation coefficient between the ranked word lists in the Frequency Dictionary and in Russian Web Corpus is 0.61, whereas the coefficient between the Frequency Dictionary and the ruTenTen corpus stands at 0.78, which in the latter case reveals that the dictionary and the corpus have much more in common.

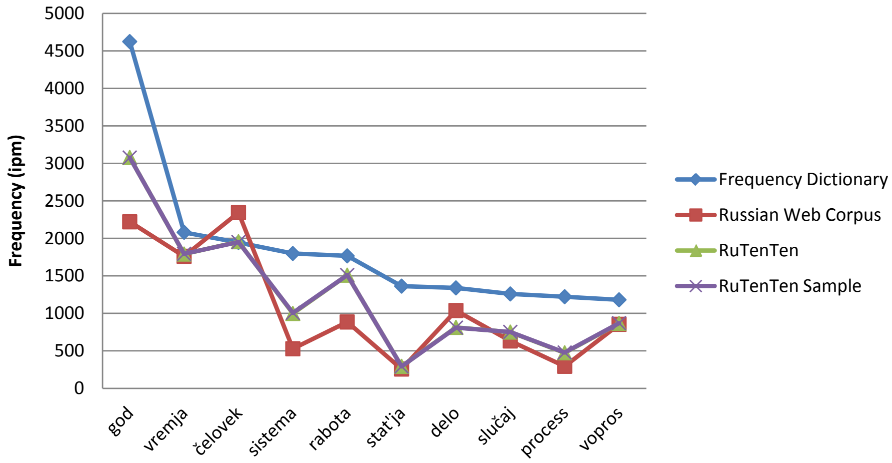

Fig. 1. Frequency distribution of nouns on the non-fiction word list (journalism excluded) as per three corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary; red—Russian Web Corpus; green—ruTenTen; purple—ruTenTen Sample)

Table 6. Frequencies of nouns on the social and political journalism word list as per the three corpora

No.	Lemma	Translation	Frequency (ipm)			
			Social & political journalism word list in the Frequency Dictionary	Russian Web Corpus	ruTenTen	
					Corpus	Sample
1	god	year	5,589.50	2,220.74	3,078.97	3,076.99
2	čelovek	man	2,950.10	2,343.79	1,955.40	1,950.79
3	vremja	time	2,364.60	1,761.06	1,790.84	1,793.41
4	žizn’	life	1,548.40	1,054.70	864.40	862.25
5	delo	affair	1,482.00	1,037.09	813.12	809.29
6	den’	day	1,397.80	1,052.35	1,089.16	1,088.18
7	rabota	job	1,272.40	885.02	1,509.37	1,510.41
8	strana	country	1,203.90	576.81	662.05	664.03
9	vopros	question	992.00	853.94	866.03	869.27
10	slovo	word	989.70	807.83	633.83	631.81

On Fig. 2 we can see the data for the nouns in Table 2; like the results on Fig. 1 it shows that both the ruTenTen corpus and subset yield identical results. The word *rabota* (see Table 6) has higher frequency in the ruTenTen corpus, than in the Dictionary; for other nouns the Dictionary shows maximum frequency values. Spearman’s rank correlation coefficient between the ranked word lists in the Frequency Dictionary and in Russian Web Corpus is remarkably high standing at 0.94 which can indicate that Russian Web Corpus has more in common with newspaper articles. Only two nouns *vremja* and *den'* have identical rankings in Russian Web Corpus and the ruTenTen corpus.

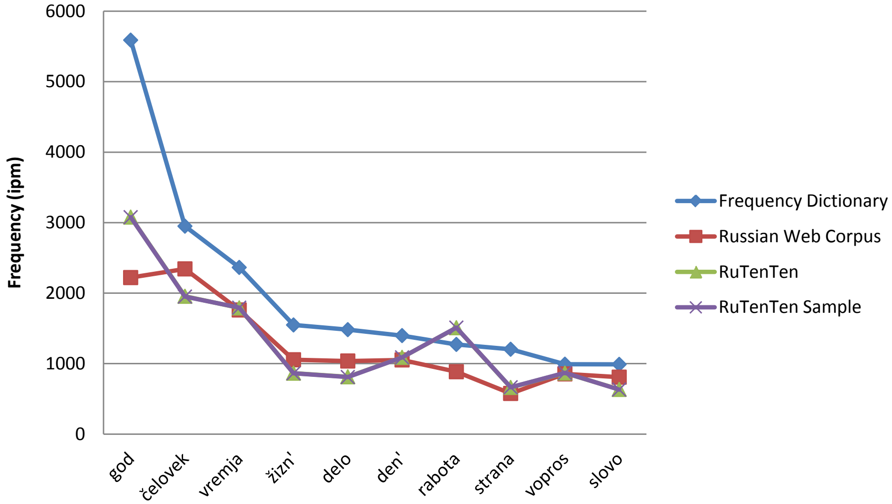

Fig. 2. Frequency Distribution of Nouns on the Social and Political Journalism Word List as per Three Corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary; red—Russian Web Corpus; green—ruTenTen; purple—ruTenTen Sample)

At the next stage of our experiment we looked at the frequencies of nouns on the style-specific word lists (Tables 3 and 4). The results for the two style-specific groups of nouns are summarized in Table 7 and Fig. 3 and in Table 8 and Fig. 4 respectively.

It can be seen from Fig. 3 and 4 that the Frequency Dictionary subcorpus shows highest frequencies (in ipm), which is hardly surprising. Although style-specific words were selected not by absolute frequencies, but rather by the LL-score, this measure is still indicative of the number of units in the subcorpus, which in turn explains the fact why some lexemes with frequencies above corpus average values are marked as style-specific. The four nouns *sistema*, *process, sud, federacija*, *risunok* and *virus* have no discrepancy in ranking across the two corpora—Russian Web Corpus and ruTenTen. This is the largest number of words with identical ranking across the corpora. Spearman's rank correlation coefficient between the ranked word lists in the Frequency Dictionary and Russian Web Corpus is 0.92 and can indicate that the data on the non-fiction word list of style-specific vocabulary and the corpus data are largely identical (though to a lesser extent, the same is true for the ruTenTen corpus data, with the equally high Spearman's correlation coefficient between the Frequency Dictionary and the corpus standing at 0.73).

Table 7. Frequencies of style-specific nouns on non-fiction word list as per three corpora

No.	Lemma	Translation	Frequency (ipm)				
			Style-specific word list for non-fiction texts (journalism excluded) in the Frequency Dictionary		Russian Web Corpus	ruTenTen	
			Corpus	Subcorpus		Corpus	Sample
1	sistema	system	617.8	1,798.0	527.61	998.41	1,006.66
2	stat'ja	article	395.0	1,363.0	257.55	293.7	292.09
3	process	process	371.7	1,221.8	294.37	473.94	478.05
4	sud	court	371.1	1,153.9	197.00	302.98	301.73
5	federacija	federation	258.9	1,003.1	88.03	198.31	197.06
6	issledovanie	study	200.5	799.6	154.44	265.11	266.86
7	metod	method	197.0	772.3	153.51	263.74	265.91
8	ispol'zovanie	usage	190.3	757.9	146.83	346.74	350.04
9	risunok	picture	179.2	776.2	45.19	77.04	76.84
10	virus	virus	106.5	584.1	21.01	36.90	36.75

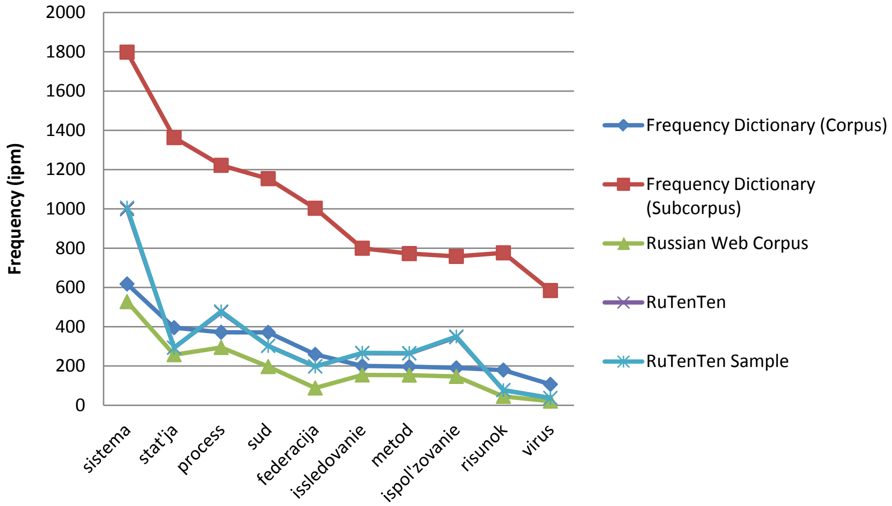

Fig. 3. Frequency distribution of style-specific nouns on the non-fiction word list as per three corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary (corpus); red—Frequency Dictionary (subcorpus); green—Russian Web Corpus; purple—ruTenTen; light blue—ruTenTen Sample)

Table 8. Frequencies of nouns on the style-specific word list for news and newspaper texts as per three corpora

No.	Lemma	Translation	Frequency (ipm)				
			Style-specific word list for news and newspaper texts from the Frequency Dictionary		Russian Web Corpus	ruTenTen	
			Corpus	Subcorpus		Corpus	Sample
1	god	year	3,727.5	5,589.5	2,220.74	3,078.97	3,076.99
2	strana	country	725.7	1,203.9	576.81	662.05	664.03
3	kompanija	company	392.7	699.0	390.72	970.15	979.11
4	prezident	president	311.0	634.6	185.6	215.07	213.81
5	teatr	theatre	305.3	611.9	91.08	102.09	99.28
6	pravitel'stvo	government	277.7	531.2	183.25	225.28	224.83
7	fil'm	film	196.8	380.1	172.15	214.16	213.10
8	spektakl'	play	164.7	350.0	37.09	44.42	42.78
9	reforma	reform	133.1	273.0	58.48	47.16	47.74
10	vybory	elections	117.7	243.4	—	62.34	63.20

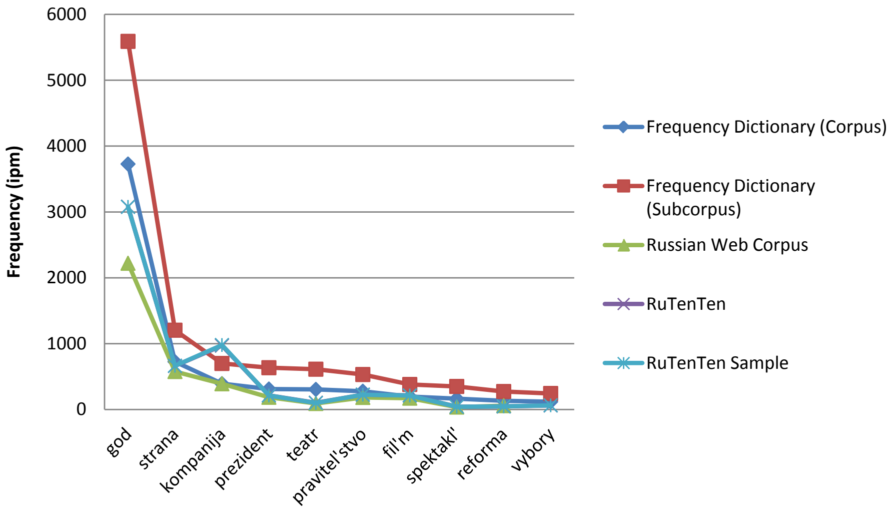

Fig. 4. Frequency distribution for nouns on the style-specific word list for news and newspaper texts as per three corpora (*y*-axis: frequency (ipm); charts: blue—Frequency Dictionary (corpus); red—Frequency Dictionary (subcorpus); green—Russian Web Corpus; purple—ruTenTen; light blue—ruTenTen Sample)

It is particularly remarkable, that the maximum initial value on the chart corresponds to the frequency of *god* and is present on each graph. The explanation is that this noun is top three by frequency in all the three corpora. The Russian Web Corpus failed to produce any results for the lexeme *vybory*, because according to its morphological token the usage of this lemma merges with the lemma *vybor*. The first four nouns *god, strana, kompanija* and *prezident* have identical ranking in the Russian Web Corpus and the Frequency Dictionary (both in the main corpus and the subcorpus). Spearman's rank correlation coefficient is equally top high for the Frequency Dictionary and Russian Web Corpus standing at 0.95.

2.3. Low Frequency Nouns: Selection

The selection of low-frequency nouns is quite tricky. For this aim we analyzed the Frequency Dictionary paying attention to the tail of the frequency list and selected 12 low-frequency nouns ranked between the following two ranges: 1) 18,940–18,965; 2) 19,955–20,000 in the Frequency Dictionary [Lyashevskaya. Sharoff 2009].

Table 9. Low-frequency nouns in the Frequency Dictionary[2]

No.	Lemma	Translation	Frequency indicated in the Dictionary (ipm)[2]	Frequency (absolute)			Frequency (ipm)		
				Russian Web Corpus	ruTenTen	ruTenTen Sample	Russian Web Corpus	ruTenTen	ruTenTen Sample
1	opala	disgrace	2.6 (4.3)	200	15,235	1,067	1.06	0.83	0.85
2	zador	fervour	2.8 (3.7)	322	26,173	1,826	1.71	1.43	1.46
3	svastika	swastika	2.6 (3.6)	496	29,270	1,902	2.64	1.60	1.52
4	šljuz	sluice	2.6 (2.9)	851	87,209	5,867	4.53	4.77	4.68
5	sčastlivec	lucky man	2.6 (2.6)	262	12,256	856	1.39	0.67	0.68
6	zlodejstvo	outrage	2.8 (2.5)	332	13,768	958	1.77	0.75	0.76
7	inkvizicija	inquisition	2.6 (2.5)	552	48,051	3,168	2.94	2.63	2.53
8	zaplata	patch	2.6 (1.9)	325	17,078	1,182	1.73	0.93	0.94
9	xoluj	groveller	2.6 (1.7)	195	8,453	540	1.04	0.46	0.43
10	tjulen'	seal	2.6 (1.3)	375	29,725	1,776	2.00	1.63	1.42
11	zagrivok	nape	2.8 (1.0)	225	9,695	655	1.20	0.53	0.52
12	sedmica	week	2.6 (0.8)	202	20,628	1,722	1.07	1.13	1.37

[2] Frequency indicated in social and political journalism list (1990–2000s) is written in parentheses.

2.4. Low Frequency Nouns: Results

The data shown in Table 9 suggests that the differences between the Frequency Dictionary and corpora are not apparently reconciled in case of low frequency words. Spearman's coefficient has the following values (as compared to the results with the high-frequency nouns): 0.33 for the Frequency Dictionary and ruTenTen sample corpus, 0.22 for the Frequency Dictionary and ruTenTen corpus and 0.21 for the Frequency Dictionary and the Russian Web corpus. However, the Russian Web Corpus shows similar high coherence in data both with the ruTenTen and ruTenTen sample corpora (Spearman's correlation coefficient is 0.80 and 0.79 respectively). Hence for the given low-frequency words the difference between the corpora and the Frequency Dictionary is more obvious than between the corpora in question. This fact can imply that corpora automatically crawled from the web share more in common with each other that with the National Russian Corpus (that served as the source for the Frequency Dictionary). Low values of the Spearman's coefficient imply that nouns differ a lot in their ranks between the latter corpus and other corpora and it is crucial in case of low-frequency words.

3. Conclusion and Future Work

The general conclusion from the obtained data suggests that texts selected for large corpora reflect the language of the web. The results published in the Frequency Dictionary are based on the Russian National Corpus, which makes them so coherent. High-frequency nouns indicated in non-fiction texts tend to be more similar to the ruTenTen corpus, whereas words fixed in the social and political journalisms subcorpus share a lot with data in the Russian Web Corpus. Thus, the Russian Web Corpus has more in common with newspaper articles. The analysis of high-frequency nouns and their ranking positions in both 1 bln subset and 14 bln corpus shows that they have produced the same results, but this is not true for the low frequency nouns. In case of low frequency data three corpora do not show much coincidence with the Frequency Dictionary lists. However we can say that in general the sample shares similar features with the total set (ruTenTen) and hence in this sense small corpora can be used for evaluating word frequencies.

The nouns (*sajt*, *sistema, kompanija,* and *Rossija*) that are ranked top by frequency in the ruTenTen corpus and its billion-size subset, but are missing among the results in the Frequency Dictionary, reveal the specific properties of web-based texts—firstly, their abundance and secondly, the focus on describing the web-page content. If we use more high-frequency nouns the Spearman's coefficient will be lower because of the diversity in ranks of the words. But the value of the coefficient will be constantly higher (than if we increase the number low-frequency words as it will be even negative).

The Russian Web Corpus appears to be more consistent with the Frequency Dictionary than the ruTenTen corpus in describing high-frequency nouns. The differences between the corpora are apparently reconciled in case of high-frequency words, but the opposite doesn't hold true for the low-frequency words.

We believe that these experiments have a future. It is crucial to study the results for low frequency words, because this group of words is the one that may produce entirely different numeric values for large corpora. To be more specific, preliminary results of our collocations study have shown that higher absolute frequency of a particular lexical item is not always conducive to a larger number of relations for the said item (despite greater number of syntagmatic partners, typical for each relation).

Acknowledgements

This work was supported by the grant of the President of Russian Federation for state support of scholarly research by young scholars (Project No. МК-5274.2016.6).

References

1. *Baroni M., Bernardini S.* (2004), BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004, Lisbon: ELDA, pp. 1313–1316
2. *Belikov V. I., Selegey V. P., Sharoff S. A.* (2012) Preliminary considerations towards developing the General Internet Corpus of Russian [Prolegomeny k proyektu General'nogo internet-korpusa russkogo yazyka (GIKRYa)]. In Computational linguistics and intellectual technologies. Vol. 11 (18). Moscow: Izd-vo RGGU, pp. 37–49.
3. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora. In: P. Sojka, A. Horák, I. Kopeček and K. Pala (Eds.): Text. Speech and Dialogue. 17th International Conference. TSD 2014. Brno. Czech Republic. September 8–12 2014. Springer International Publishing, pp. 257–264.
4. *Shu-Kai Hsieh* (2014), Why Chinese Web-as-Corpus is Wacky? Or: How Big Data is Killing Chinese Corpus Linguistics? In Proceedings of the 9th Edition of the Language Resources and Evaluation. Reykjavik, Iceland, pp. 2386–2389.
5. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family. In Proceedings of the International Conference on Corpus Linguistics, pp. 125–127.
6. *Kehoe A., Renouf A.* (2002), WebCorp: Applying the Web to Linguistics and Linguistics to the Web. In Proceedings WWW2002 Conference. Honolulu, Hawaii, available at: http://www2002.org/CDROM/poster/67.
7. *Kilgarriff, A.* (2001), Comparing corpora. In International journal of corpus linguistics. 6(1), pp. 97–133.
8. *Kilgarriff A., Grefenstette G.* (2003), Introduction to the Special Issue on Web as Corpus. In Computational Linguistics, 29 (3), pp. 333–347.
9. *Lyashevskaya O., Sharoff S.* (2009), Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data [Chastotnyj slovar' sovremennogo russkogo jazyka (na materialakh Natsional'nogo Korpusa Russkogo Jazyka)]. Moscow: Azbukovnik.
10. *Piperski A., Belikov V., Kopylov N., Morozov Eu., Selegey V., Sharoff S.* (2013), Big and diverse is beautiful: A large corpus of Russian to study linguistic variation.

In Proceedings of the 8th Web as Corpus Workshop (WAC-8) Corpus Linguistics Conference 2013, available at: https://sigwac.org.uk/raw-attachment/wiki/WAC8/wac8-proceedings.pdf

11. *Shaikevich A. Ya.* (2015), Measures of Lexical Similarity between Frequency Dictionaries [Mery leksičeskogo sxodstva častotnyx slovarej], Proc. International Conference "Corpus linguistics-2015" [Trudy Mezhdunarodnoy konferencii "Korpusnaya lingvistika–2015"], St. Petersburg: St. Petersburg State University, pp. 434–442.
12. *Sharoff S.* (2006), Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini. (eds). WaCky! Working papers on the Web as Corpus. Gedit, Bologna, available at: http://wackybook.sslmit.unibo.it/
13. *Sinclair J.* (2004), Trust the text: Language, corpus and discourse. London/New York: Routledge.
14. *Zakharov V.* (2015), Evaluation of Internet corpora of Russian [Ocenka kačestva Internet-korpusov russkogo jazyka], Proc. International Conference "Corpus linguistics-2015" [Trudy Mezhdunarodnoy konferencii "Korpusnaya lingvistika–2015"], St. Petersburg: St. Petersburg State University, pp. 219–229.