Comparison of High-Frequency Nouns from the Perspective of Large Corpora

Maria Khokhlova

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

Abstract. Since the last decade a number of corpora have become available, a large part of them have been compiled automatically on web data. From traditional text collections such corpora vary both in their volume and content. The paper focuses on the discussion on these corpora and deals with two of them: ruTenTen (18.3 bln tokens) and Araneum Russicum Maximum (13.7 bln tokens). The authors discuss linguistic phenomena across the corpora examining quantitative properties of 20 high-frequency Russian nouns. The lexemes are compared between these corpora and also with data published in the Frequency Dictionary on their rank distributions. This dictionary was compiled on the subset of Russian National Corpus that represents modern Russian of the 20 th century (1950–2007) and can be viewed as an excellent example of a traditional corpus. The analysis shows promising results; there is a close correlation between traditional and web-corpora and this topic should be studied in more detail paying attention to other parts of speech.

Key words: Web corpora, big data, frequency, correlation analysis, corpus evaluation

1 Introduction

Corpus linguistics, a branch of linguistics that deals with building corpora and investigation of their data, has already celebrated its 55th anniversary counting from the appearance of the Brown corpus. The idea of corpora that contain big data has attracted scholars' attention for a long time. During the last decade more and more corpora are being compiled automatically. From traditional text collections they vary both in their volume and content. This is closely related to the growing availability of technical resources and thus the gradually changing paradigm in corpus linguistics moving forward from "manual" approach to more automatic one. By a classical or traditional approach one can understand a compilation of corpora based on a previously described methodology: selection of texts involving their representativeness and balance, their correction, annotation and upload. New corpora contain in general texts that were automatically crawled from the Web. Researchers find it attractive to make statistical inferences on increasingly larger scope of data. At

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 9–17, 2016. © Tribun EU 2016

the same time access to large corpora provokes new challenges: what can we see with big data and how does it affect the results? Are there differences between traditional and automatically compiled corpora? The paper is organized as follows. In the next section we give an overview of the related work. In Section 3 we describe our data, while Section 4 presents the experiments and results of the analysis. Finally, Section 5 closes the paper with conclusions and suggestions for future work.

2 Related work

Large corpora with volumes exceeding 100 mln tokens have appeared just recently. The idea of creating such large text collections Nowadays one can speak about two types of corpora, some authors distinguish between three types [1]. 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 can be named a traditional one and 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 (obviously, to a certain degree that holds true for the first type also). For the Russian language we can name the Aranea project, which includes a few Russian corpora that differ in their size and texts among them Araneum Russicum Maximum [2]. The TenTen family [3] 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. Here we leave aside rather large collections of texts that can't be viewed as electronic corpora from the traditional viewpoint (for example, the service Google Books). 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"). In [4] the authors present their results on studying rare Russian idioms in large corpora.

3 Data and methods

The aim of our research is to compare linguistic phenomena across different large corpora and dictionary, to identify differences, and to analyze them. We selected above mentioned two corpora that had been collected and built automatically – ruTenTen (18.3 bln tokens) and Araneum Russicum Maximum (13.7 bln tokens). In our study we used the Frequency Dictionary [5]. This dictionary was compiled on the subset of 92 mln tokens from Russian National Corpus that represents modern Russian of the 20 th century (1950–2007). It includes texts of various genres: fiction, social and political journalism, non-fiction (textbooks, social media, advertisements, technical literature) etc. 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 a word list of lemmas based on the Frequency Dictionary [5]. To succeed in our study we studied frequency properties of high-frequency nouns that had been selected from the dictionary among these corpora. As nonparametric measure of statistical dependence between our data we chose Spearman's rank correlation coefficient.

4 Experiments

We compiled two lists of nouns extracted from [5] that are typical for non-fiction (see Table 1) and social and political journalism texts (see Table 2)[1]. Each list contains 20 nouns that were ranked top by frequency the Frequency Dictionary.

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

	Lemma	Translation	Frequency (ipm)
1	god	year	4624.2
2	vremja	time	2080.5
3	čelovek	man, person	1945.3
4	sistema	system	1798.0
5	rabota	job, work	1766.4
6	stat'ja	article, clause	1363.0
7	delo	affair, business	1339.5
8	slučaj	case	1259.0
9	process	process	1221.8
10	vopros	question	1180.9
11	lico	face, person	1175.9
12	sud	court	1153.9
13	čast'	part	1153.8
14	vid	kind, aspect	1147.9
15	reshenie	decision	1122.3
16	pravo	right	1117.6
17	rebënok	baby, child	1078.4
18	otnošenie	relation	1077.5
19	razvitie	development	1059.6
20	federacija	federation	1003.1

Tables 1 and 2 show that some words are shared by both lists; they belong to the high-frequency lexemes that do not depend on the genre: *čelovek* 'man, person', *delo* 'affair, business', *god* 'year', *rabota* 'job, work', *slučaj* 'case', *vopros*

[1] The Frequency Dictionary provides separate frequency lists for both types of texts.

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

	Lemma	Translation	Frequency (ipm)
1	god	year	5589.5
2	čelovek	man, person	2950.1
3	vremja	time	2364.6
4	žizn′	life	1548.4
5	delo	affair, business	1482.0
6	den′	day	1397.8
7	rabota	job, work	1272.4
8	strana	country	1203.9
9	vopros	question	992.0
10	slovo	word	989.7
11	mesto	place	976.1
12	mir	world, peace	887.8
13	dom	house, home	879.7
14	drug	friend	850.9
15	slučaj	case	744.3
16	gorod	city, town	738.5
17	ruka	arm, hand	713.0
18	vlast′	power	711.8
19	konec	end	710.8
20	sila	strength	709.8

'question', *vremja* 'time'. It is worth mentioning that the average frequency of the nouns presented in Table 1 is higher than in Table 2. It can be supposed that the given high-frequency lexemes are more often used in non-fiction texts than in newspapers (however the volume of non-fiction subcorpus is less). In our research we have also analyzed 20 top-frequency nouns in the two corpora. The following list was compiled for ruTenTen corpus: *god* 'year', *rabota* 'job, work', *vremja* 'time', *čelovek* 'man, person', *kompanija* 'company', *sistema* 'system', *sajt* 'site', *den′* 'day', *mesto* 'place', *Rossija* 'Russia', *vid* 'kind, aspect', *vopros* 'question', *slučaj* 'case', *rebënok* 'baby, child', *žizn′* 'life', *vozmožnost′* 'opportunity, possibility', *kačestvo* 'quality', *programma* 'programme', *delo* 'affair, business', *usluga* 'service, favour'. For Araneum Russicum Maximum corpus the following nouns were most frequent: *god* 'year', *čelovek* 'man, person', *vremja* 'time', *rabota* 'job', *den′* 'day', *kompanija* 'company', *oblast′* 'region, field', *sistema* 'system', *sajt* 'site', *mesto* 'place', *vopros* 'question', *žizn′* 'life', *slučaj* 'case', *Rossija* 'Russia', *vid* 'kind, aspect', *dom* 'house, home', *delo* 'affair, business', *strana* 'country', *raz* 'time, one', *vozmožnost′* 'opportunity, possibility'. We can see that a half of the first list overlaps with Table 2 whereas twelve nouns from the second list coincide with the data in the same table. Based on this preliminary analysis of the lists we can see that two corpora share more in common with newspapers than with non-fiction texts. Comparing two lists it can be said that the majority of the nouns presents in both of them that indicates the similarity of the corpora. Lexemes *sajt* 'site' and *kompanija* 'company' were not among 20 most frequent nouns selected from the Frequency Dictionary but were ranked top by frequency in the lists for both corpora. This fact can be explained on one hand that a vast number of data for the corpora is crawled from news web-sites and on the other hand a lot of texts have description of web pages and their content. This holds particularly true for ruTenTen corpus due to other frequent lexemes: *vozmožnost′* 'opportunity, possibility', *kačestvo* 'quality', *programma* 'programme', *usluga* 'service, favour'. We referred to the two corpora to study frequencies of the words on the lists (see Tables 1 and 2); you can find the results on Table 3 and Fig. 1. as well as on Table 4 and Fig. 2.

Table 3: Frequencies of nouns on the non-fiction word list (journalism excluded) calculated as per two corpora

	Lemma	Translation	Frequency word list for non-fiction in the Frequency Dictionary	ruTenTen	Araneum Russicum Maximum
1	god	year	4624.2	3080.0	3263.0
2	vremja	time	2080.5	1791.0	1857.0
3	čelovek	man, person	1945.3	1956.0	2012.0
4	sistema	system	1798.0	999.0	1011.0
5	rabota	job, work	1766.4	1510.0	1632.0
6	stat′ja	article, clause	1363.0	294.1	446.8
7	delo	affair, business	1339.5	814.0	741.0
8	slučaj	case	1259.0	752.0	758.0
9	process	process	1221.8	474.0	491.9
10	vopros	question	1180.9	866.0	855.0
11	lico	face, person	1175.9	483.7	458.1
12	sud	court	1153.9	303.2	255.7
13	čast′	part	1153.8	677.0	650.3
14	vid	kind, aspect	1147.9	723.0	806.0
15	reshenie	decision	1122.3	558.0	556.3
16	pravo	right	1117.6	507.2	405.1
17	rebënok	baby, child	1078.4	850.0	443.1
18	otnošenie	relation	1077.5	481.2	438.4
19	razvitie	development	1059.6	587.0	570.6
20	federacija	federation	1003.1	198.4	168.0

Table 3 and Fig. 1 show the data for nouns in Table 1. We can see that both corpora show similar curves on the graph, which means that these words have similar distribution. Both corpora agree on the ranking of the seven words five of them being on the top of the lists. Spearman's rank correlation coefficient between the ranked word lists of ruTenTen and Araneum Russicum Maximum corpora is 0.89 that indicates a very high correlation. The rank coefficient between the lists in the Frequency Dictionary and in ruTenTen is 0.63, whereas the coefficient between the Frequency Dictionary and the Araneum Russicum Maximum corpus stands at 0.76, which in the latter case reveals that the dictionary and the corpus have much more in common. The frequencies, indicated in the Dictionary, are the highest, except the frequency of the lemma *čelovek* 'man, person' which has the highest frequency in ruTenTen. Two corpora rank the words differently from the ranking in the Dictionary – two nouns in ruTenTen have the same ranking as in the Dictionary, and Araneum Russicum Maximum contains three such nouns.

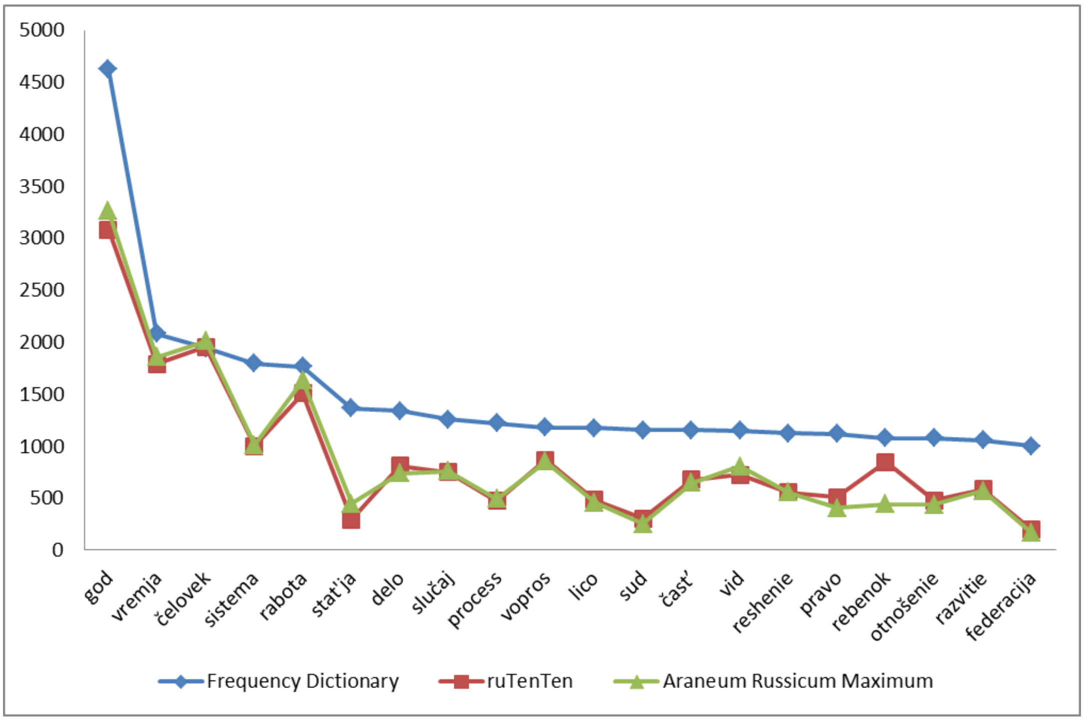

Fig. 1: Frequency distribution of nouns on the non-fiction word list (journalism excluded) as per two corpora (x-axis: nouns; y-axis: frequency in ipm)

On Fig. 2 we can see the data for the nouns in Table 2; like the results on Fig. 1. Fig 2 shows that both the ruTenTen and Araneum Russicum Maximum corpora yield to a certain degree identical results. The word *rabota* 'job, work' (as well as *slučaj* 'case' and *gorod* 'city, town', see Table 4) has higher frequency in Araneum Russicum Maximum, than in the Dictionary; for other nouns the Dictionary shows maximum frequency values. Four nouns have identical rankings in the Frequency Dictionary and both in ruTenTen and Araneum Russicum Maximum corpora. In case of two corpora the number of such nouns (that have the same ranks) is nine. Spearman's rank correlation coefficient between the ranked word lists in the Frequency Dictionary and in ruTenTen is high standing at 0.84 and it is 0.82 for the word lists in the Frequency Dictionary and in Araneum Russicum Maximum. This can indicate that both corpora more in common with newspaper articles and similar texts and moreover with Russian National Corpus (as it was the source for the Frequency Dictionary). Spearman's rank correlation coefficient between the lists in two corpora is remarkably large and equals 1 (this points to the highest correlation).

Table 4: Frequencies of nouns on the social and political journalism word list as per the two corpora

	Lemma	Translation	Social & political journalism word list in the Frequency Dictionary	ruTenTen	Araneum Russicum Maximum
1	god	year	5589.50	3080.0	3263.0
2	čelovek	man, person	2950.10	1956.0	2012.0
3	vremja	time	2364.60	1791.0	1857.0
4	žizn′	life	1548.40	865.0	899.0
5	delo	affair, business	1482.00	814.0	741.0
6	den′	day	1397.80	1089.0	1253.0
7	rabota	job, work	1272.40	1510.0	1632.0
8	strana	country	1203.90	662.0	657.6
9	vopros	question	992.00	866.0	855.0
10	slovo	word	989.70	645.0	563.3
11	mesto	place	976.10	950.0	970.0
12	mir	world, peace	887.80	626.0	655.5
13	dom	house, home	879.70	689.0	751.0
14	drug	friend	850.90	452.3	500.7
15	slučaj	case	744.30	752.0	758.0
16	gorod	city, town	738.50	757.0	792.0
17	ruka	arm, hand	713.00	466.7	430.5
18	vlast′	power	711.80	330.0	273.9
19	konec	end	710.80	417.8	344.4
20	sila	strength	709.80	467.5	438.2

5 Conclusion and Further Work

We come to the general conclusion that texts selected for large corpora feature the language of the web and their structure corresponds to newspaper texts and thus to journalistic genre. The Araneum Russicum Maximum appears to be slightly more consistent with the Frequency Dictionary than the ruTenTen corpus in describing high-frequency nouns. For the given high-frequency nouns there is a very strong association between the data obtained on two corpora. Hence it can be supposed that there is no difference between the

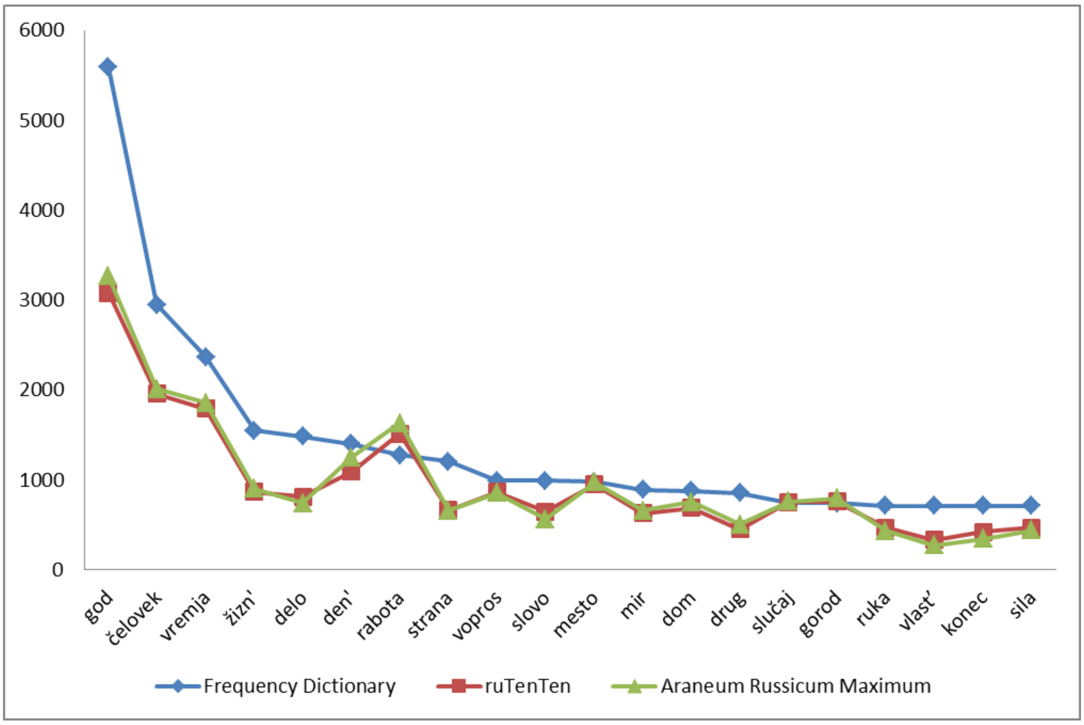

Fig. 2: Frequency distribution of nouns on the social and political journalism word list as per two corpora (x-axis: nouns; y-axis: frequency in ipm)

automatically crawled corpora in case of high-frequency lexemes. Both corpora show quite a high correspondence with the Frequency dictionary. The data selected from the Frequency dictionary were based on the Russian National Corpus and therefore the obtained results reveal a close correlation between traditional and web-corpora. Our next work will be targeted at other parts of speech as nouns can be thematically biased, and their frequencies can depend on types of texts and thus differ dramatically even among corpora compiled within the same methodology.

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

References

1. Belikov, V., Selegey, V., Sharoff, S.: Prolegomeny k proyektu General'nogo internet-korpusa russkogo yazyka (GIKRYa) [Preliminary considerations towards developing the General Internet Corpus of Russian]. In Computational linguistics and intellectual technologies. Vol. 11 (18), pp. 37-49. RSUH, Moscow (2012).
2. Benko, V.: Aranea: Yet Another Family of (Comparable) Web Corpora. In Text. Speech and Dialogue. 17th International Conference. TSD 2014, pp. 257-264. Springer, Switzerland (2014).

3. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The TenTen Corpus Family. In Proceedings of the International Conference on Corpus Linguistics, pp. 125-127. Lancaster, UK (2013).
4. Benko, V., Zakharov, V.: Very Large Russian Corpora: New Opportunities and New Challenges. In Computational linguistics and intellectual technologies. Vol. 15 (22), pp. 79-93, RSUH, Moscow (2016).
5. Lyashevskaya, O., Sharoff, S.: Častotnyj slovar′ sovremennogo russkogo jazyka (na materialax Nacional′nogo Korpusa Russkogo Jazyka) [Frequency Dictionary of Contemporary Russian based on the Russian National Corpus data]. Azbukovnik, Moscow (2009).