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

Moscow, June 1–4, 2016

VERY LARGE RUSSIAN CORPORA: NEW OPPORTUNITIES AND NEW CHALLENGES

Benko V. (vladob@juls.savba.sk)

Slovak Academy of Sciences, Ľ. Štúr Institute of Linguistics, Bratislava, Slovakia

Zakharov V. P. (v.zakharov@spbu.ru)

St. Petersburg State University;
Institute for Linguistic Studies, RAS, St. Petersburg, Russia

Our paper deals with the rapidly developing area of corpus linguistics referred to as *Web as Corpus* (*WaC*), i.e., creation of very large corpora composed of texts downloaded from the web. Some problems of compilation and usage of such corpora are addressed, most notably the "language quality" of web texts and the inadequate balance of web corpora, with the latter being an obstacle both for corpus creators, and its users. We introduce the *Aranea* family of web corpora, describe the various processing procedures used during its compilation, and present an attempt to increase the size of its Russian component by the order of magnitude. We also compare its contents from the user's perspective among the various sizes of the Russian *Aranea*, as well as with the other large Russian corpora (*RNC*, *ruTenTen* and *GICR*). We also intent to demonstrate the advantage of a very large corpus in linguistic analysis of low-frequency language phenomena in linguistics, such as usage of idioms and other types of fixed expressions.

Keywords: web corpora, *WaC* technology, representativeness, balance, evaluation

СВЕРХБОЛЬШИЕ КОРПУСЫ РУССКОГО ЯЗЫКА: НОВЫЕ ВОЗМОЖНОСТИ И НОВЫЕ ПРОБЛЕМЫ

Бенко В. (vladob@juls.savba.sk)
Словацкая академия наук, Институт языкознания им. Людовита Штура, Братислава, Словакия

Захаров В. П. (v.zakharov@spbu.ru
Санкт-Петербургский государственный университет; Институт лингвистических исследований РАН, Санкт-Петербург, Россия

В статье обсуждается одно из активно развиваемых направлений в корпусной лингвистике — создание корпусов большого объема на основе текстов из веба. Показаны их возможности в исследовании и описании устойчивых сочетаний. Описываются технология и проблемы их создания. Обсуждаются проблемы таких корпусов, которые ставят вопросы как перед разработчиками корпусов, так и перед пользователями, а именно, проблемы морфологической разметки и сбалансированности корпусов.

Ключевые слова: веб-корпусы, *WaC* технология, репрезентативность, сбалансированность, оценка

0. Introduction

Quantitative assessment of language data has always been an area of great interest for linguists. And not only for them: as early as in 1913, the Russian mathematician A. A. Markov counted the frequencies of letters and their combinations in the Pushkin's *Eugene Onegin* novel, and calculated the lexical probabilities in the Russian language [Markov, 1913]. With the advent of first computers, the usage of quantitative methods in linguistic research has accelerated dramatically [Piotrovskiy 1968; Golovin 1970; Alekseev 1980; Arapov 1988], aiding in creation of frequency dictionaries[1] and in other research activities of theoretical and applied nature [Frumkina 1964, 1973].

The next step in using quantitative methods in language research has been done within an area of corpus linguistics. The results of corpus queries are usually accompanied by the respective statistical information. Advanced corpus management systems provide for obtaining all sorts of statistical data, including those of linguistic

[1] It should be noted, however, that first frequency dictionaries have been compiled well in the pre-computer era, in the end of the19th century [Kaeding 1897].

categories and metadata. Combination of quantitative methods, distributional analysis and contrastive studies is becoming the basis of new corpus systems that could be referred to as "intellectual". Their functionalities include automatic extraction of collocations, terms, named entities, lexico-semantic groups, etc. In fact, corpus linguistics based on formal language models and quantitative methods is "learning" to solve intellectual semantic tasks.

Assuming that one of the main features of a representative corpus is its size, then a 100-million token corpus, considered a standard at the beginning of this century, now appears in many cases to be insufficient to receive relevant statistical data. To study and adequately describe multi-word expressions consisting of medium or low-frequency words, it is necessary to apply large and even very large corpora. In the context of this paper, we call a corpus "very large" if its size exceeds 10 billion tokens[2].

1. Web as Corpus

Nowadays, the "big data" paradigm became very popular. According to Wikipedia, "*Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate*"[3]. This "big data" now seem to have approached the corpus linguistics.

Compilation of traditional corpora is usually a laborious and rather slow process. As soon as the need for larger corpora has been recognized, it became clear that the requirements of the linguistic community cannot be easily satisfied by the traditional resources of corpus linguistics. This is why many linguists in the process of their research turned to Internet search services. But using search engines as corpus query systems is associated with many problems (cf. [Kilgarriff 2007; Belikov et al. 2012])— this is where the idea of *Web as Corpus* (*WaC*), i.e., creation of language corpora based on the web-derived data has been born. It was apparently for the first time explicitly articulated by Adam Kilgarriff [Kilgarriff 2001; Kilgarriff, Grefenstette 2003].

In early 2000s, a community called *WaCky!*[4] was established by a group of linguists and IT specialists who were developing tools for creation of large-scale web corpora. During the period of 2006–2009, several *WaC* corpora were created and published, including the full documentation of the respective technology, with each containing 1–2 billion tokens (*deWaC*, *frWaC*, *itWaC*, *ukWaC*) [Baroni et. al 2009].

In 2011, the *COW*[5] (*COrpora from the Web*) project started at the Freie Universität in Berlin. Within its framework, English, German, French, Dutch, Spanish and Swedish corpora have been created. In the 2014 edition (*COW14*) of the family, sizes of some corpora reached almost 10 billion tokens, while the German corpus has 20 billion tokens [Schäfer, Bildhauer 2012; Schäfer 2015]. These corpora are accessible (for

---

[2] In Russian, we suggest the term "сверхбольшой корпус".

[3] https://en.wikipedia.org/wiki/Big_data

[4] http://wacky.sslmit.unibo.it/

[5] http://hpsg.fu-berlin.de/cow/

research purposes) via the project web portal[6]. The site also provides English, German, Spanish and Swedish corpus-based frequency lists.

Large number of *WaC* corpora has been created and/or made available within the framework of the *CLARIN* Project in Slovenia (Jožef Stefan Institute). Besides the respective South Slavic languages (*bsWaC*, *hrWaC*, *slWaC*, *srWaC*) [Ljubešić, Erjavec 2011; Ljubešić, Klubička 2014], corpora for many other languages, including Japanese, are available there. Their sizes vary between 400 million and 2 billion tokens. Most of the corpora are accessible[7] under *NoSketch Engine*[8] without any restrictions.

None of the projects mentioned, however, includes the Russian language.

The largest number of *WaC* corpora was created by Lexical Computing Ltd. (Brighton, UK & Brno, Czech Republic) company that made them available within *Sketch Engine*[9] environment. At the time of writing this paper (April 2016), these corpora covered almost 40 languages, including Russian, and their sizes varied between 2 million and 20 billion tokens. The size of the largest Russian *ruTenTen* corpus was 18.3 billion tokens [Jakubíček et al. 2013].

From today's perspective, we can see that the *WaC* technology has succeeded. Related set of application programs that represent effective implementation of this technology has been published, including tools for web crawling, data cleaning and deduplication, with many of them under free or open-source licenses (*FLOSS*) that made the technology available also for underfunded research and educational institutions in Central and Eastern Europe.

There are, however, also other approaches to creation of very large corpora. One of them—based on massive digitization of books from public libraries—has been attempted by Google (available via *Google Books Ngram Viewer*[10]) [Zakharov, Masevich 2014]. Another possibility is creating corpora based on the integral web collections, such as the *General Internet Corpus of Russian*[11] (*GICR, 19.7 billion tokens*) [Belikov et al. 2013], that is composed of blogs, social media, and news.

2. WaC "How To"

To create a web corpus, we usually have to perform (in a certain sequence) operations as follows:

- Downloading large amounts of data from the Internet, extracting the textual information, normalizing encoding

---

6 https://webcorpora.org/

7 http://nl.ijs.si/noske/index-en.html

8 https://nlp.fi.muni.cz/trac/noske

9 http://www.sketchengine.co.uk

10 https://books.google.com/ngrams

11 http://www.webcorpora.ru/en

- Identification the language of the downloaded texts, removing the "incorrect" documents
- Segmenting the text into paragraphs and sentences
- Removing duplicate content (identical or partially identical text segments)
- Tokenization—segmenting the text into words
- Linguistic (morphological, and possibly also syntactic) annotation—lemmatization and tagging
- Uploading the resulting corpus into the corpus manager (i.e., generating the respective index structures) that will make the corpus accessible for the users.

With the exception of first two, all other operations have been already included (to a certain extent) in the process of building traditional corpora. It is therefore often possible to use existing tools and methodology of corpus linguistics, most notably for morphological and syntactic annotation.

Downloading data from the web is usually performed by one of two standard methodologies that differ in the way how the URL addresses of the web pages to be downloaded are retrieved.

(1) Within the method described in [Sharoff 2006], a list of medium-frequency words is used to generate random n-tuples that are subsequently iteratively submitted to a search engine. Top URL addresses delivered within each search are then used to download the data for the corpus. The process can be partially automated by the *BootCaT*[12] program [Baroni, Bernardini 2004].

(2) The second method is based on scanning ("crawling") the web space by means of a special program—crawler—that uses an initial list of web addresses provided by the user and iteratively looks for new URLs by analysing the hyperlinks at the already downloaded web pages. The program usually works autonomously and may also perform encoding/language identification and/or deduplication on the fly, which makes the whole process very efficient and allows in a relatively short time (several hours or days) download textual data containing several hundreds of millions tokens. Two most popular programs used for crawling the web corpora are the general-purpose *Heritrix*[13] and a specialized "linguistic" crawler *SpiderLing*[14] [Suchomel, Pomikálek 2012].

Each of the methods mentioned above has its pros and cons, with the former being more suitable for creation of smaller corpora (especially if the corpus is geared towards a specific domain), while the latter is usually used to create very large corpora of several billions of tokens in size.

---

[12] http://bootcat.sslmit.unibo.it/

[13] https://webarchive.jira.com/wiki/display/Heritrix

[14] http://corpus.tools/wiki/SpiderLing

3. The *Aranea* Web Corpora Project: Basic Characteristics and Current State

The *Aranea*[15] family presently consists of (comparable) web corpora created by the *WaC* technology for 14 languages in two basic sizes. The *Maius* ("larger") series corpora contain 1.2 billion tokens, i.e. approximately 1 billion words (tokens starting with alphabetic characters). Each *Minus* ("smaller") corpus represents a 10% random sample of the respective *Maius* corpus. For some languages, region-specific variants also exist that, e.g., increase the total number of Russian corpora to six. *Araneum Russicum Maius & Minus* include Russian texts downloaded from any internet domains, *Araneum Russicum Russicum Maius & Minus* contain only texts extracted from the *.ru* and *.рф* domains, and *Araneum Russicum Externum Maius & Minus* are based on texts from "non-Russian" domains, such as *.ua*, *.by*, *.kz*, etc. For more details about the *Aranea* Project see [Benko 2014].

According to our experience, a Gigaword corpus can be created by means of *FLOSS* tools in a relatively short time, even on a not very powerful computer. After the processing pipeline had been standardized, we were able to create, annotate and publish a corpus for a new language in some 2 weeks (provided that the respective tagger was available).

The situation, however, has changed when we wanted to increase the corpus size radically. We decided to create a corpus of a *Maximum* class, i.e., "as much as can get". Our attempt to create the Slovak and Czech *Maximum* corpora revealed that the limiting factor was the availability of the sufficient amounts of texts for the respective languages in Internet. With standard settings for *SpiderLing* and after several months of crawling, we were able to gather only some 3 Gigawords for Slovak and approximately 5 Gigawords for Czech.

To verify the feasibility of building very large corpora within our computing environment, we decided to create *Araneum Maximum* for a language, where sufficient amount of textual data in Internet is expected. The Russian language has been chosen for this experiment, and the lower size limit was set to 12 billion tokens, i.e., ten times the size of the respective *Maius* corpus.

It has to be noted that the work was not to be started from scratch, as the data of existing Russian *Aranea* had been utilized. After joining all available Russian texts and deduplicating them at the document level, we received approximately 6 billion tokens, i.e., seemingly half of the target corpus size. It was, however, less than that, as the data had not been dedulicated at the paragraph level yet.

The new data was crawled by the (at that time) newest version 0.81 of *SpiderLing*, and the seed URLs were harvested by *BootCaT* as follows:

(1) A list of 1,000 most frequent adverbs extracted from the existing Russian corpus was sorted in random order (adverbs have been chosen as they do not have many inflected forms and usually have rather general meaning).

[15] http://ella.juls.savba.sk/aranea_about

(2) For each *BootCat* session, 20 adverbs were selected to generate 200 *Bing* queries (three adverbs in each), and requesting to get the maximal amount of 50 URLs from each query. This procedure has been repeated five times, totalling in 1,000 *Bing* queries.

The number of URLs harvested by a single *BootCaT* session in this way was usually close to the theoretical maximum of 50,000, but it decreased to some 40,000–45,000 after filtration and deduplicaction. The resulting list was sorted in random order and iteratively used as seed for *SpiderLing*.

To create a *Maius* series corpus, we always tried to gather approximately 2 billion tokens of data, so that the target 1.2 billion can be safely achieved after filtration and deduplication. For "large" languages, this could be reached during first two or three days of crawling. As it turned out later, we were quite lucky not to reach the configuration limits of our server, most notably the size of RAM (16 GB). As all data structures of *SpiderLing* are kept in main memory, when trying to prolong the crawling time for the Russian the memory limit has been reached only after approximately 80–90 hours of crawling. Though some memory savings tricks are described in the *SpiderLing* documentation, we, nonetheless, had to opt for a "brute force" method by restarting the crawling several times from scratch, knowing that lots of duplicate data would be obtained.

In total, 12 such crawling iterations (with some of them consisting of multiple sessions) have been performed, during which we experimented with the number of seed URLs ranging from 1,000 to 40,000.

To speed up the overall process, another available computer was used for cleaning, tokenization, partial deduplication and tagging of the already downloaded lots of data. Moreover, the most computationally-intensive operations (tokenization and tagging) have been performed in parallel, taking the advantage of the multiple-core processor of our computer. The final deduplication has been performed only after all data has been joined into one corpus.

Our standard processing pipeline contains the steps described in Tables 1 and 2.

Table 1. Processing of a typical new lot (one of 12)

Operation	Output	Processing time (hh:mm)
Data crawling by *SpiderLing* (2 parallel processes) with integrated boilerplate removal by *jusText*[16] [Pomikálek 2011] and identification of exact duplicates	2,958,522 docs 39.68 GB	cca 86 hours
Deleting duplicate documents identified by *SpiderLing*	2,058,810 docs 18.15 GB	0:27

[16] http://corpus.tools/wiki/Justext

Operation	Output	Processing time (hh:mm)
Removing the survived HTML markup and normalization of encoding (Unicode spaces, composite accents, soft hyphens, etc.)		0:30
Removing documents with misinterpreted utf-8 encoding	2,054,827 docs	0:41
Tokenization by *Unitok*[17] [Michelfeit et al. 2014] (4 parallel processes, custom Russian parameter file)	1,611,313,889 tokens 19.88 GB	4:04
Segmenting to sentences (rudimentary rule-based algorithm)		0:29
Deduplication of partially identical documents by *Onion*[18] [Pomikálek 2011] (5-grams, similarity threshold 0.9)	1,554,837 docs 1,288,238,029 tokens (20.05% removed) 17.23 GB	1:23
Conversion all utf-8 punctuation characters to ASCII and changing all occurrences of "*ë*" to "*e*" (to make the input more compatible with the language model used by the tagger).		0:53
Tagging by *Tree Tagger*[19] [Schmid 1994] with language model trained by S. Sharoff[20] (4 parallel processes)	39.06 GB	8:26
Recovering the original utf-8 punctuation and "ë" characters		0:53
Marking the out-of-vocabulary (*OOV*) tokens (*ztag*)	82,786,567 tokens marked *OOV* (6.43%)	1:09
Mapping the "native" *MTE*[21] tagset to "PoS-only" *AUT*[22] tagset	46.39 GB	1:09

[17] http://corpus.tools/wiki/Unitok

[18] http://corpus.tools/wiki/Onion

[19] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[20] http://corpus.leeds.ac.uk/mocky/

[21] http://nl.ijs.si/ME/V4/msd/html/msd-ru.html

[22] http://ella.juls.savba.sk/aranea_about/aut.html

Table 2. Final processing

	Output	Processing time (hh:mm)
Joining all parts of data (old data + 12 new lots, some of them accessed via Ethernet at a different machine)	37,956,781 docs 26,720,417,271 tokens 932.80 GB	10:42
Deduplication of partially identical documents by *Onion* (5-grams, similarity threshold 0.9)	24,509,170 docs 17,322,616,899 tokens (35.17% removed) 602.33 GB	19:12
Deduplication of partially identical paragraphs by *Onion* (5-grams, similarity threshold 0.9)	13,704,863,990 tokens (20.88% removed) 482.04 GB	27:07
Compilation by *NoSketch Engine*	249.78 GB of index structures	79:54

4. Experimenting with the New Corpus

At the end of all the processing mentioned, we indeed succeeded to create a very large Russian corpus of the expected size—its characteristics (as displayed by *NoSketch Engine*) are shown in Fig. 1.

Corpus *Araneum Russicum Maximum (Russian, 16.04) 13,7 G* – statistics and info

Russian Web (crawled 2013 to 2016, version 1.3.70) 13,7 G (build #a068)

Counts	
Tokens	13,704,863,994
Words	10,945,698,722
Sentences	798,912,811
Paragraphs	299,974,845
Documents	24,509,166

General info	
Corpus description	Document
Language	Russian
Encoding	UTF-8
Compiled	04/23/2016 15:56:17
Tagset	Description

Lexicon sizes	
word	43,132,672
lemma	40,074,411
atag	12
tag	1,140
ztag	2
lc	37,345,635
lemma_lc	34,686,361

Fig. 1. New Corpus Info

Within the context of *NoSketch Engine*, a token is considered "word" if it begins with an alphabetic character (in any script recognized by Unicode). It must be also noted that the lemma lexicon contains large proportion of out-of-vocabulary items that could not have been lemmatized.

In the following text, we will demonstrate the usefulness of a very large corpus for studying rare language phenomena, such as phraseology.

4.1. Chasing Fixed Expressions

In small corpora, many idioms often appear—if ever—in singular ("hapax") occurrences that make it difficult to draw any relevant linguistic conclusions. Moreover, idioms and other fixed expressions are often subject to lexical and/or syntactic variation, where the individual members of the expressions change within a fixed syntactic formula, or the same set of lexical units create different syntactic structures [Moon 1998]. It is most likely without exaggeration to claim that idioms having lexical and syntactic variants represent the majority of cases. Lots of (Russian) examples can be shown: *беречь/хранить как зеницу ока; беречь пуще глаза; мерить одной мерой/меркой, мерить на одну меру/мерку; ест за троих, есть в три горла; драть/сдирать/содрать шкуру (три/две шкуры), драть/сдирать/содрать по три (две) шкуры; хоть в землю заройся, хоть из-под земли достань; брать/взять (забирать/забрать) в [свои] руки, прибирать/прибрать к рукам; сталкивать/столкнуться лицом к лицу, носом к носу, нос в нос, лоб в лоб.*

The description of variant multi-word expressions in dictionaries is naturally much less complete in comparison with fixed phrasemes. And, only large and very large corpora can help us to analyse and describe this sort of variability in full.

Now we shall try to demonstrate the possibilities given by *Araneum Russicum Maximum* on three examples. Let us take fixed expressions described in dictionaries and show how they behave in various corpora.

4.2. *"Щёки как у хомяка"*[23]

The *Russian National Corpus* (*RNC*[24], 265 M tokens[25]) gives 5 occurrences of "*щёки как*": *как у матери, как у бульдога, как у пророка, как у тяжко больного, как у меня*. As it can be seen, all of them are singular occurrences (hapax legomena), and no occurrence of *как у хомяка* has been found.

Let us have a look what can be found in other corpora. While the smaller *Aranea* provide even less information, *Araneum Russicum Maximum* confirms the dictionary data, and *ruTenTen* and *GICR* corpora make it even more convincing. Besides *как у хомяка*, they also add *как у бульдога*, *как у бурундука* and *как у матрешки*, as well as several other (less frequent) comparisons.

---

[23] "*cheeks like a hamster*"

[24] http://www.ruscorpora.ru/en/search-main.html

[25] This number is not directly comparable with other corpora, as punctuation characters are not considered tokens in *RNC*.

Table 3. "*Щёки как у…*"

	щёки/ щеки как у...	**хомяка/ хомячка**	**буль-дога**	**бурун-дука**	**мат-решки**
Araneum Russicum Minus	1	–	1	–	–
Araneum Russicum Maius	1	–	1	–	–
Araneum Russicum Maximum	33	6	1	4	2
ruTenTen	45	24	4	–	1
GICR	126	84	3	5	1

4.3. "*Щёки из-за спины видны*"[26]

RNC gives just one example of *щеки из-за...*: *щеки из-за ушей видны*. The other corpora give the following:

Table 4. "*Щёки из-за…*"

	щёки/ щеки из за...	**спины видны/ видать/торчат**	**ушей видны/ видать/торчат**
Araneum Russicum Minus	–	–	–
Araneum Russicum Maius	6	3	–
Araneum Russicum Maximum	27	7	5
ruTenTen	30	20	6
GICR	65	40	23

The very large corpora not only provide much more evidence, but also add several interesting variants of "*щеки из-за…*": *увидеть можно, просматриваются, вылезают, сияют румянцем; щек из-за спины видно не было*, etc.

4.4. "*Чистой воды…*"[27]

The idiomatic expression *чистой* or *чистейшей воды* is described in the dictionary as "*о ком или чем-либо, полностью соответствующем свойствам, качествам, обозначенным следующим за выражением существительным*" [BED 1998]. But if we want to extract the relevant information on the most frequent noun collocates of this expression from *RNC*, we mostly get 2–3 examples for each noun: *авантюрист, блеф, гипотеза, демагогия, монополизм, мошенничество, популизм, провокация, садизм, спекуляции, фантастика, хлестаковщина,* etc.

---

[26] "*cheeks visible from behind*"

[27] "*of the clear water*"

What can be observed in larger corpora? When comparing frequency ranks of expressions with different nouns derived from large corpora, we can see that they are more or less similar, while the data received from small corpora can differ significantly. Nouns appearing at the top positions of the ranked frequency lists derived from the large corpora (*выдумка, вымысел, лохотрон, обман, пиар, профанация, развод, спекуляция*) are usually missing in the output from smaller corpora. On the other hand, top words obtained from *Araneum Russicum Minus* (*чудодействие, грабеж, подстава*) are ranked 50, or even 500 in large corpora. We can also see that the total weight of expressions with significant frequencies (4 or more within the framework of our experiment) is greater in large corpora (Table 5).

Table 5. Frequencies of "*чистой/читейшей воды + noun*" expressions in various corpora

corpus size in tokens	**Araneum Russicum Minus 120 M**	**Araneum Russicum Maius 1.2 G**	**Araneum Russicum Maximum 13.7 G**	**ruTenTen 18.3 G**
total expressions	**146**	1,256	10,441	15,548
unique expressions	26	692	3,264	≥ 5,000[28]
total expressions with f >3	12 (8.2%)	450 (35.8%)	6,841 (65.5%)	9,370 (60.3%)
unique expressions with f >3	2 (7.7%)	54 (7.8%)	449 (13.8%)	668 (13.4%)

The corpus evidence, however, shows that the *чистой воды* expression is also used in its direct meaning. In fact, there are two direct meanings of "*чистой воды*" present there: "*вода чистая, без примесей*", and "*чистая, свободная ото льда или водной растительности*". The interesting fact is, that practically in all cases where *чистой воды* precedes the respective noun, its meaning is idiomatic (Fig. 2).

In *Araneum Russicum Maximum*, out of 449 different analysed expressions with total count of 6,841, less than 10 contained non-idiomatic use of "*чистой воды*" (associated with *объем/температура* or *озеро/море/океан*). And, the majority of the respective nouns have a negative connotation: *абсурд, авантюра, агрессия, алчность, бандит, блеф, богохульство, болтология, бред, брехня, бытовуха, вампиризм, вкусовщина, вранье, глупость, госдеповец, графоманство, демагог, диктатура, жульничество, заказняк, зомбирование, идеализм, извращение, издевательство, инквизиция, кальвинизм, капитализм, кидалово, копипаст, коррупция, лапша, липа, литература, популизм, порнография, пропаганда, развод, расизм, рвач, русофобия, садизм, фарисейство, фарс, фашизм*, etc. Some of them are receiving this negative connotation especially within this expression (*кальвинизм, капитализм, копипаст, лапша, липа, литература, пропаганда* etc.)

[28] Only first 5,000 items of frequency distributions are shown in Sketch Engine.

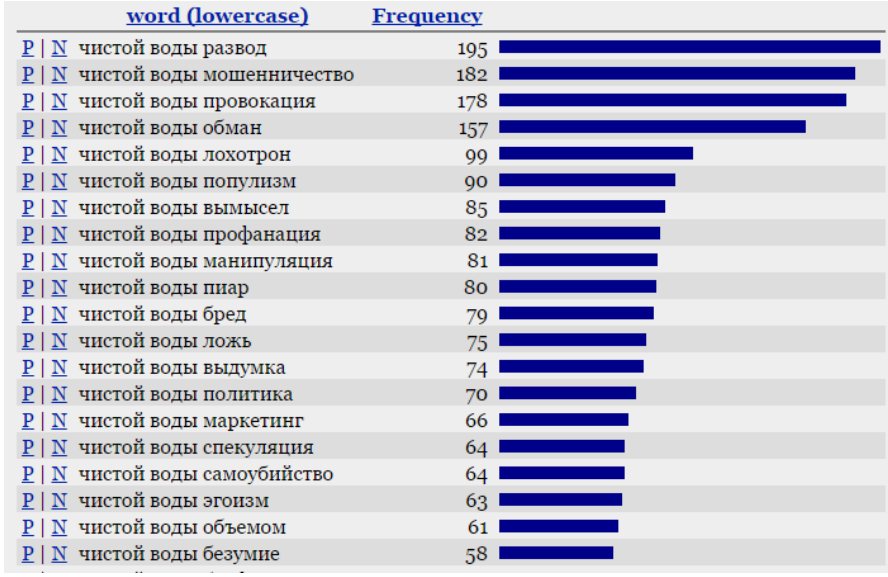

Fig. 2. Frequency distribution of right-hand noun collocates of *чистой воды* in *Araneum Russicum Maximum*

On the other hand, if *чистой воды* is located after the corresponding noun, the share of its direct meaning is as much as 80% (*литр чистой воды, стакан чистой воды, количество, подача, перекачивание, источник, резервуар, глоток, кран чистой воды*, etc.)

5. Conclusions and Further Work

As it can be seen, very large corpora enable much deeper analysis that is not possible with corpora of smaller size. We can also say that, starting from a certain size of corpora, the results of these studies can be seen as representative. On the other hand, we do not want to state that web corpora could fully replace the traditional ones. They can, however, be really very large and reflect the most "fresh" changes of the language.

Our experiment has also shown that not everything is that simple. The problems encountered can be divided into three parts: problems of linguistic annotation (lemmatization and tagging), problems of metadata (tentatively referred to as "meta-annotation"), and technical problems related to deduplication and cleaning. It is clear that the traditional TEI-compliant meta-annotation cannot be performed in web corpora, as they lack the explicit necessary bibliographic data. In practice, we can get data only with minimal bibliographic annotation in terms of web (domain name, web page publication or crawl date, document size, etc.), and traditional concepts of representativeness and/or balance are hardly applicable. What we can get is the volume,

but the question of "quality" remains without an answer. Both the nature of textual data and the imbalance of web corpora make the question of assessing the results of analyses based on such corpora open.

A new methodology based on the research has to be developed yet. We believe that such methods should include both quantitative and qualitative assessments from the perspective of applicability of very large corpora in various types of linguistic research. It might also be useful to compare contents of web corpora with the existing traditional corpora, as well as with frequency dictionaries. It is also necessary to take into account the technical aspects, such as "price vs. quality" relation.

Our experiment aimed to create the Russian *Araneum Maximum* has shown that though some technical problems related to the computing power of our equipment (two quad-core Linux machines witch 16 GB RAM and 2 TB of free disk space each, joined by a Gigabit Ethernet line, and having a 100 Mbit Internet connection), do exist, they could be eventually solved. The bottleneck of the process was the final deduplication by *Onion* that needed 56 GB of RAM, and had to be performed on a borrowed machine. After minor modifications of our processing pipeline, we were able to perform all other operations, including the final corpus compilation by the *NoSketch Engine* corpus manager using our own hardware.

The first results based on our new corpus show that in comparison the *RNC*, *Araneum Russicum Maximum* can provide much more data on rare lexical units and fixed expressions of different kinds and allows for linguistic conclusions. On the other hand, our experience shows that lexis typical for fiction and poetry seems to be underrepresented in our corpus.

Our next work will be targeted both at the increase of the size of our corpus, and also at improving its "quality"—by better filtration, normalization and linguistic annotation. Here we hope to apply methods of crowd-sourcing (e.g., verifying the morphological lexicons by students). The other serious task will be the classification of the texts according to web genres, so that the balance of the corpus could be—at least partially—controlled.

Acknowledgements

This work has been, in part, supported by the Slovak Grant Agency for Science (VEGA Project No. 2/0015/14), and by the Russian Foundation for the Humanities (Project No. 16-04-12019).

References

1. *Alexeev P. M.* (1980), Statistical lexicography [Statisticheskaya lexikografiya], Moscow.
2. *Arapov M. V.* (1988), Quantitative linguistics [Kvantitativnaya lingvistika], Moscow.
3. *Baroni M., Bernardini S.* (2004), BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

4. *Baroni M., Bernardini, S., Ferraresi A., Zanchetta E.* (2009), The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation 43 (3), pp. 209–226.
5. *BED* (1998), Kuznetsov S. A. (Ed.) Big Explanatory Dictionary of the Russian Language [Bol'shoj tolkovyj slovar' russkogo yazyka], St. Petersburg: Norint.
6. *Belikov V., Selegey V., Sharoff S.* (2012). Preliminary considerations towards developing the General Internet Corpus of Russian // Komp'juternaja lingvistika i intellectual'nye tehnologii: Trudy mezhdunarodnoj konferentsii «Dialog–2012» [Computational Linguistics and Intellectual Technologies. Proceedings of International Conference «Dialog–2012»]. Moscow, RGGU, pp. 37–49.
7. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation, [Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2013"], vol. 12 (19), Moscow, RGGU, pp. 84–95.
8. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978-3-319-10815-5.
9. *Frumkina R. M.* (1964), Statistical methods of lexica research [Statisticheskiye metody izucheniya leksiki], Moscow.
10. *Frumkina R. M.* (1973), The role of statistical methods in modern linguistic researches [Rol' statisticheskikh metodov v sovremennykh lingvisticheskikh issledovaniyakh], Moscow.
11. *Golovin B. N.* (1970), Language and statistics [Yazyk i statistika], Moscow.
12. *Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V.* (2013), The TenTen Corpus Family, 7th International Corpus Linguistics Conference, Lancaster, July 2013.
13. *Kaeding F. W.* (1897), Häufigkeitswörterbuch der deutschen Sprache. Steglitz b. Berlin.
14. *Kilgarriff A.* (2001), Web as corpus, in P. Rayson, A. Wilson, T. McEncry, A. Hardic and S. Klioja (eds.) Proceedings of the Corpus Linguistics 2001 Conference, Lancaster (29 March—2 April 2001). Lancaster: UCREL, pp. 342–344.
15. *Kilgarriff A., Grefenstette G.* (2003), Introduction to the Special Issue on Web as Corpus. Computational Linguistics 29 (3), 2003. Reprinted in Practical Lexicography: a Reader. Fontenelle, T. (Ed.) Oxford University Press. 2008.
16. *Kilgarriff A.* (2007), Googleology is Bad Science. Computational Linguistics 33 (1): pp. 147–151.
17. *Ljubešić N., Erjavec T.* (2011), hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. Text, Speech and Dialogue 2011. Lecture Notes in Computer Science, Springer.
18. *Ljubešić N., Klubička F.* (2014): {bs,hr,sr} WaC—Web corpora of Bosnian, Croatian and Serbian. Proceedings of the 9th Web as Corpus Workshop (WaC-9). Gothenburg, Sweden.

19. *Markov A. A.* (1913), An Example of statistical research on the text of Eugene Onegin illustrated trial relations in a chain [Primer statisticheskogo issledovaniya nad tekstom “Yevgeniya Onegina”, illustrirujuscikh svyaz’ ispytaniy v tsepi], Imperial St. Petersburg Academy of Sciences Transactions [Izvestiya Inperatorskoy Akademii Nauk S.-Peterburga], series VI, vol. VII, pp. 153–162.
20. *Michelfeit J., Pomikálek J., Suchomel V.* (2014), Text Tokenisation Using unitok. In Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 71–75, 2014. Brno: NLP Consulting 2014.
21. *Moon, R.* (1998), Fixd Expressions and Idioms in English. A Corpus-Based Approach. Oxford: Clarendon Press.
22. *Piotrovskiy R. G.* (1968), Information measuring in language [Informatsionnye izmereniya yazyka], Leningrad.
23. *Pomikálek J.* (2011), Removing Boilerplate and Duplicate Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno.
24. *Schäfer R., Bildhauer F.* (2012), Building Large Corpora from the Web Using a New Efficient Tool Chain. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12).
25. *Schäfer R.* (2015), Processing and querying large web corpora with the COW14 architecture. In: Proceedings of Challenges in the Management of Large Corpora (CMLC-3). Talk at Challenges in the Management of Large Corpora (CMLC-3) on July 20, 2015 in Lancaster.
26. *Schmid, H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester.
27. *Sharoff S.* (2006), Creating General-Purpose Corpora Using Automated Search Engine Queries. In: WaCky! Working Papers on the Web as Corpus. ISBN 88-6027-004-9, Bologna: Gedit Edizioni, pp. 63–98.
28. *Suchomel V., Pomikálek J.* (2012), Efficient Web Crawling for Large Text Corpora. In: Adam Kilgarriff, Serge Sharoff. Proceedings of the seventh Web as Corpus Workshop (WAC7). Lyon, 2012. pp. 39–43.
29. *Zakharov V. P., Masevich A. Ts.* (2014), Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus [Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer], Srtuctural and Applied Linguistics [Strukturnaya i prikladnaya lingvistika], vol. 10, Saint-Petersburg, pp. 303–327.