Big and diverse is beautiful: A large corpus of Russian to study linguistic variation

Alexander Piperski[1], Vladimir Belikov[1], Nikolay Kopylov[2], Eugene Morozov[2], Vladimir Selegey[1, 2], Serge Sharoff[3]
[1]Russian State University for the Humanities, Russia
[2]ABBYY, Russia
[3]University of Leeds, UK
apiperski@gmail.com, vibelikov@gmail.com,
Nikolay_Ko@abbyy.com, Eugene_M@abbyy.com
Vladimir_S@abbyy.com, s.sharoff@leeds.ac.uk

Abstract

The General Internet Corpus of Russian (GICR) is aimed at studying linguistic variation in present-day Russian available on the Web. In addition to traditional morphosyntactic annotation, the corpus will be richly annotated with metadata aimed at sociolinguistic research of language variation, including regional, gender, age, and genre variation. The sources of metadata include explicit information available about the author in his/her profile, information coming from IP or URL, as well as machine learning from textual features.

1 Russian corpora: an overview

The linguists studying Russian have a wide range of different corpora available. By far the most popular resource is the Russian National Corpus,[1] which has become a de facto standard for the majority of corpus-based studies in Russian linguistics. However, this corpus is not well-suited for exploring the present-day language, since recently produced texts constitute a small proportion of it, and they are selected from a small number of sources. Other Russian corpora, such as I-RU, an Internet snapshot of Russian (Sharoff 2006) or ruTenTen[2] lack metadata. They are also often too small to capture frequencies of linguistic phenomena specific only to some part of the Russian-speaking community. Therefore many linguists have to rely on statistical data provided by search engines, such as Google or Yandex (the most popular search engine in Russia), but the drawbacks of this method are well-known (Kilgarriff 2007, Belikov et al. 2012).

2 General Internet Corpus of Russian: aims and objectives

The lack of a corpus representing the modern usage of Russian with diverse metadata gave rise to the General Internet Corpus of Russian (GICR) project which has been under development at the Russian State University for the Humanities since 2012 (cf. Belikov et al. 2012, Belikov et al. 2013). The aim of creating GICR is to provide the linguistic community with a reliable tool for studying the present-day Russian with specific information on language variation. In order to achieve this, it is necessary to collect a large amount of texts from the Web. The final version of the corpus is estimated to contain around 100 billion words by 2014.

The texts in GICR will be extensively annotated. Apart from morphological and syntactic annotation, GICR will contain a lot of metadata pertaining to the texts included in the corpus, such as gender, age, social status of the author, genre, topic and regional variety.

One specific objective is to draw attention to regional variation in Russian. It has always been acknowledged that there are village dialects in Russia (cf. Kasatkin 2005), but until recently the common opinion was that Russian of the cities is more or less homogeneous. However, this was questioned by Belikov (2006). His online dictionary *The Languages of Russian Cities*[3] shows that there are remarkable differences which span a wide range of uses, including locally produced legal texts (*vybit' chek* vs. *otbit' chek* 'issue a receipt'), professional terminology (*obnalichka* vs. *opanelka* 'door frame'), names of games or classes for schoolchildren, etc. Slight differences in morphosyntax are also existent, but a large

[1] http://ruscorpora.ru/en/

[2] http://trac.sketchengine.co.uk/wiki/Corpora/TenTen

[3] http://community.lingvo.ru/goroda/dictionary.asp

corpus with enough metadata is needed to investigate this issue.

3 Data collection and indexing

For data collection we use an adapted version of Nutch to crawl the Internet starting from the known hotspots of the Russian Web. The segments which are being investigated are blog platforms, forums, magazines and newspapers, etc. As of June 2013, GICR includes:

- the Russian-language blogs from LiveJournal.com, which is the most popular blog platform in Russia;
- the magazines form the *Magazine Reading Room* (*Zhurnal'nyj zal*, http://magazines.russ.ru/), a large online collection of Russian fiction magazines;
- the travel forum *Vinsky forum* (*Forum Vinskogo*, http://forum.awd.ru/)

Since we aim to study present-day Russian, only texts that are less than 5 years old are included in the corpus. Further on, we plan to adhere to this policy to keep GICR up-to-date.

At present, the size of the corpus is 1.38 billion words. However, it can be easily expanded, because making the corpus larger involves only a minimal amount of manual work.

Using blogs and forums, we expect to get most efficient results by keeping user profile statistics (date of registration, number of messages) together with user messages, so that we can get more benefits in analyzing site-specific user activity. Our algorithms rely on the idea that the more texts of a specific user we take into consideration, the more reliable the results of spam detection, age, and gender classification are.

Boilerplate removal algorithm is based on whether or not we know the web page structure (cf. Gibson et al. 2011). For known pages created using a well-known blog platform, content management system or forum platform, we can get only texts from the DOM element with well-known XPATH signatures. This also helps to separate the message body from the comments. For other pages we aim to employ a mixed strategy of taking the biggest contiguous block of text (Pomikálek 2011) or to use site-level boilerplate removal algorithms.

The page crawling strategy assumes that we collect all available web pages without using any page ranking function, but we only keep the content of those pages which have been created for humans, not for search engines. We put precision before recall, since the Russian Internet currently contains over 100 times more text than we plan for GICR.

The existing web interfaces, like Intellitext (Wilson et al. 2010) based on IMS Corpus Workbench (Christ 1994), or Manatee (Rychlý 2007) do not scale well to large corpora. Therefore we opted for development of a new system, using POS and shallow syntax annotated corpus in plain XML files indexed using what we can call *Narrowing Index*. Each sentence can be represented with some relatively small number, calculated as a product of prime numbers representing text features. Primes are assigned to word forms, lemmas, parts of regular expressions, and frequent bigrams of lemmas and RE in the descending order of their frequency. Each character number is connected with block number, by which we can reference the physical block in plain XML corpora. When we need to test a condition, the first step preceding plain corpus scan is finding block numbers in text corpora which may meet the condition of query. Random queries on an SSD storage are very fast, so that selective block retrieval becomes reasonably fast with a relatively small index.

An important type of queries concerns grouping the results, e.g. the `group` command in the CorpusWorkbench for producing collocations. Pre-caching of search results is not efficient because we cannot do set-theory arithmetic on partial results of sub-queries, but our users can be satisfied with the relative frequencies of the studied phenomena. We will collect partial results as soon as the frequencies converge to their practical limits. Since the Narrowing Index supplies us with a constant number of blocks where the query conditions are presumably met, the grouping queries can perform in constant time.

4 Text representation

The texts included in GICR are supplied with morphosyntactic annotation as well as metadata. We collected the web pages themselves (posts and comments are treated separately) as well as the following data from the user profile where available:

- username;
- user-chosen identification name (often identical with real name);
- year of birth;
- gender;

- region (this was unified to a standard form, also generalized to the respective administrative region)

Some of the authors provide only some part of this data. However, we had sufficient amount of training data even from this subset.

5 Text annotation

5.1 General issues

The size of the corpus implies that no manual annotation is possible, and for this reason it is crucial to choose fast and reliable automated annotation strategies. It is important to note that absolute accuracy cannot be achieved using such methods, but it is not a problem as long as corpus users are aware of this deficiency.

5.2 Morphosyntactic annotation

For morphosyntactic annotation we use an adapted version of the pipeline by Sharoff & Nivre (2011), which uses more Web-specific examples for training the POS tagger and the parser. The lexicon, especially for proper nouns and abbreviations, will be enriched as well.

5.3 Metadata annotation: processing pipeline

The starting point for metadata annotation is the explicit information about the author available in a standardized form in the profile of many blogging and forum platforms. Some information can be extracted from the IP address (server location for region determination) and URL (helpful for genre classification). All metadata of this kind are partial (not all bloggers provide it, IP addresses can be misleading, etc.), but this gives a source for training machine learning using textual features available on the page.

Text classification was based on standard extraction of lexical and POS features which provide sufficient reliability for this process (Sharoff et al. 2010), selection of keywords using the log-likelihood ratio (Rayson and Garside 2000) and using logistic regression and SVM for training. Because of the large amount of (sparse) training data, the Liblinear package (Fan et al. 2008) was used.

The dataset contains some amount of noise, which primarily includes spam pages (often automatically generated for search engine optimization), catalogues and other lists of objects, poems (which have very unusual text structure and linguistic properties). The majority of such instances have been cleaned by detecting the outliers using the values beyond 1.5 * *IQR* where *IQR* is the inter-quartile range for the following simple indicators:

- coverage by the most frequent words;
- average sentence length;
- text length in words.

5.4 Regional classification

There were two types of features used for machine learning. One comes from a specially compiled dictionary[4] which contains 710 words specific to different Russian-speaking regions. Other features were produced by selecting the keywords distinguishing each individual region from other regions using the standard log-likelihood keyness index (Rayson and Garside 2000). This procedure uses the top 800 words for each region (some words were specific for more than one region). For the preliminary classification, we selected 17 regions out of the complete set of Russian-speaking regions spanning over Russia and Ukraine. These regions are listed in Table 1. Moscow was excluded, since it is a melting pot for a large number of dialects.

Region	Country	Docs	%
Bashkortostan	Russia	53,420	4.29%
Chelyabinsk Oblast	Russia	49,798	4.00%
Donetsk Oblast	Ukraine	39,080	3.14%
Kiev	Ukraine	114,736	9.21%
Krasnodar Krai	Russia	50,544	4.06%
Krasnoyarsk Krai	Russia	41,032	3.29%
Moscow Oblast	Russia	119,328	9.58%
Novosibirsk Oblast	Russia	78,106	6.27%
Omsk Oblast	Russia	32,396	2.60%
Perm Krai	Russia	55,226	4.43%
St. Petersburg	Russia	300,814	24.15%
Rostov Oblast	Russia	64,340	5.17%
Samara Oblast	Russia	82,450	6.62%
Saratov Oblast	Russia	31,706	2.55%
Sverdlovsk Oblast	Russia	97,894	7.86%
Tatarstan	Russia	34,684	2.78%
	Total:	1,245,554	100%

Table 1: Regions and number of documents

For these regions, we used two sets of texts. One consisted of all texts longer than 20 words, the other included only texts longer than 300 words. The accuracy of regional classification (with 10-fold cross-validation) in the first case was about 15%, which is far from acceptable (the random baseline for the 17 regions would have been 6%). In the second case, it improved to

[4] http://community.lingvo.ru/goroda/dictionary.asp

35%, which shows that regional attribution for a very short text is very unlikely.

5.5 Gender classification

For gender classification, we used a collection of texts downloaded from the *Vinsky Forum* (http://forum.awd.ru). All posts by the same author were concatenated into a single text which was assigned the gender indicated by the author. All the words were lemmatized and supplied with grammatical features. The overall size of the collection using the format described above is 58,835 texts (28,200 texts by women and 30,635 texts by men). It is noteworthy that men tend to write more posts than women, because before concatenation we had 1,270,341 posts by men and 638,170 posts by women.

The best-suited machine learning algorithms for this purpose turned out to be logistic regression and SVM. Their results differed insignificantly, and only the results of logistic regression are provided here. All experiments included 5-fold cross-validation.

First, we tested POS-features such as the proportion of nouns, verbs, adjectives, pronouns, adverbs, prepositions, as well as the density of punctuation marks and the proportion of active voice verbs in a text. The results were low (precision = 0.574, recall = 0.575, F = 0.572). The average frequencies of different parts of speech are almost the same for men and women (the difference never exceeds 1%). This means that these features can hardly be helpful for gender classification.

Second, we chose three other features, namely the relative frequency of Adverb + Adverb bigrams (e.g., *very nicely*), of Adverb + Adjective bigrams (e.g., *very nice*) and of superlative adjectives. The accuracy of classification remained almost the same, but the number of features was significantly reduced.

Another approach was to use lexical classes described by Babych et al. (2007). The idea is to map words to general classes and to use the frequency of these classes as features. We excluded the lexical classes for which the frequency in male and female texts differed by less than 10%. The remaining classes with the correspondding male-to-female frequency ratios are represented on Graph 1. Swearwords also constituted a separate lexical class.

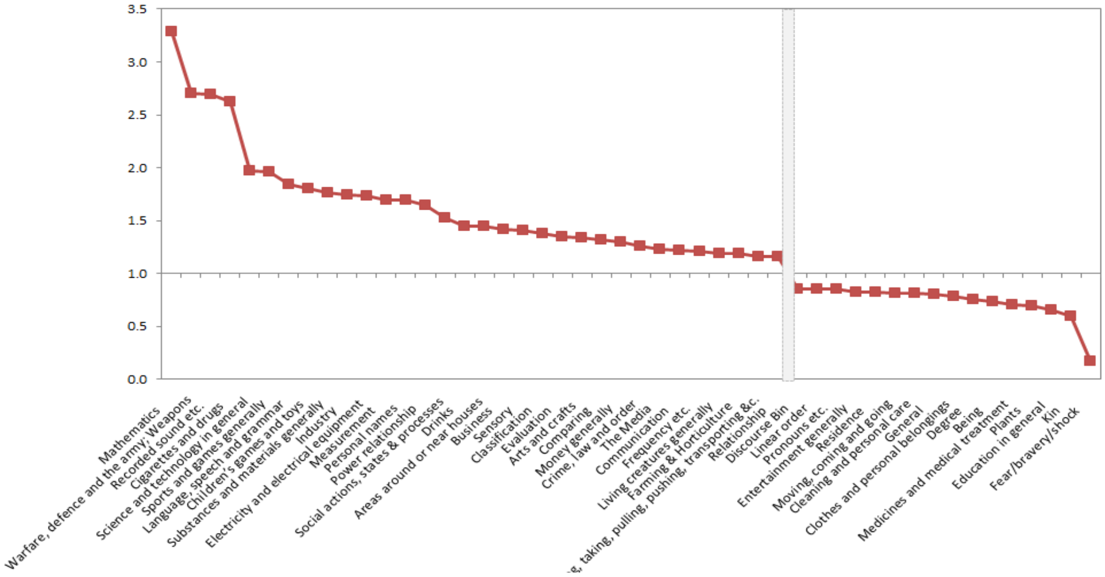
However, the problem is that some words belong to more than one lexical class, and these words had to be discarded. Unfortunately, such words are common among the most frequent ones. For example, *kuritsa* 'chicken' can belong to Food or Animals. A fast word-sense disambiguation algorithm would be useful for our purposes, but as long as such algorithms are still unavailable, we had to limit ourselves to a list of unambiguous words which contained 3000 items.

Further experiments combined lexical class features and POS features. Without imposing the lower limit on text length, we achieved the accuracy of 58%. However, such limits make it possible to improve the accuracy:

Number of words	Accuracy
≥ 1	58%
≥ 30	61%
≥ 200	67%
≥ 400	69%
≥ 1000	73%

Table 2: Accuracy of gender classification for texts of different length

There is one more grammatical feature of Russian that is useful for improving gender classification. Russian verbs are conjugated for gender in past tense (e.g., *ja*, *ty*, *on skazal* 'I (masc.), you (masc.), he said' vs. *ja*, *ty*, *ona skazala* 'I



Graph 1. Male-to-female frequency ratio of different lexical classes

(fem.), you (fem.), she said'). For this reason, the bigram *ja* 'I' + past tense is highly indicative of gender. Of course, it may sometimes be misleading, but using this bigram as a feature for machine learning we were able to reach the accuracy of 77%.

6 Conclusions

GICR is a new corpus of Russian that will contain 100 billion words by 2014, which will make it a valuable resource for studying present-day Russian. Even now, it is already larger than the Russian National Corpus contains about 500 million words. GICR aims at raising awareness of sociolinguistic variation within Russian language, and the rich metadata in the corpus will provide a basis for studying this variation.

7 Acknowledgements

The present study was supported by the Minsitry of Education and Science of the Russian Federation and by the Program of Strategic Development of the Russian State University for the Humanities.

References

Babych, B., Hartley, A., Sharoff, S. & Mudraya, O. 2007. Assisting Translators in Indirect Lexical Transfer. In: Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Prague, Czech Republic, June 23–30 2007.

Belikov, V. 2006. The examples for the dictionary of the varieties of urban Russian and the WWW. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2006*, Bekasovo, 57–60. (in Russian)

Belikov, V., Kopylov, N., Piperski, A., Selegey, V., Sharoff, S. 2013. Corpus as language: from scalability to studying variation. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2013*, Bekasovo, 84–96. (in Russian)

Belikov, V., Selegey, V., Sharoff, S. 2012. Preliminary considerations towards developing the General Internet Corpus of Russian. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2012*, Bekasovo, 37–50. (in Russian)

Christ, O., 1994. A modular and flexible architecture for an integrated corpus query system. In *Proc. COMPLEX'94*, Budapest.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Gibson, D., Punera, K., and Tomkins, A. 2005. The volume and evolution of web page templates. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, 830–839.

Kasatkin, L. 2005. Russian dialectology. Academia, Moscow. (in Russian)

Kilgarriff, A. 2007. Googleology is Bad Science. *Computational Linguistics* 33 (1): 147–151.

Pomikálek, J. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*(PhD Thesis, Masaryk University, Brno, Czech Republic).

Rayson, P. and Garside, R. 2000. Comparing corpora using frequency profiling. In *Proc. of the Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Rychlý, P. 2007. Manatee/Bonito–a modular corpus manager. *Recent Advances in Slavonic Natural Language Processing (RASLAN)*, Masaryk University, Brno, 65–70.

Sharoff, S. 2006. Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics* 11 (4), 435–462

Sharoff, S., Wu, Z., & Markert, K. 2010. The Web Library of Babel: evaluating genre collections. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, Malta.

Sharoff, S., & Nivre, J. 2011. The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. In *Proc. Computational Linguistics and Intelligent Technologies DIALOGUE2011*, Bekasovo, 591–604.

Wilson, J., Hartley, A., Sharoff, S., & Stephenson, P. (2010). Advanced corpus solutions for humanities researchers. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation*. Sendai.