# Comparing Neural Lexical Models of a Classic National Corpus and a Web Corpus: The Case for Russian

Andrey Kutuzov[1,2] and Elizaveta Kuzmenko[1]

[1] National Research University Higher School of Economics, Moscow, Russia
[2] Mail.ru Group, Moscow, Russia
akutuzov@hse.ru, eakuzmenko_2@edu.hse.ru

**Abstract.** In this paper we compare the Russian National Corpus to a larger Russian web corpus composed in 2014; the assumption behind our work is that the National corpus, being limited by the texts it contains and their proportions, presents lexical contexts (and thus meanings) which are different from those found 'in the wild' or in a language in use.

To do such a comparison, we used both corpora as training sets to learn vector word representations and found the nearest neighbors or associates for all top-frequency nominal lexical units. Then the difference between these two neighbor sets for each word was calculated using the Jaccard similarity coefficient. The resulting value is the measure of how much the meaning of a given word is different in the language of web pages from the Russian language in the National corpus. About 15% of words were found to acquire completely new neighbors in the web corpus.

In this paper, the methodology of research is described and implications for Russian National Corpus are proposed. All experimental data are available online.

**Keywords:** corpora comparison, deep learning, semantic similarity, vector representations of lexical units, lexical co-occurrence networks, Russian National Corpus, Web as corpus, word2vec.

## 1 Introduction

Contemporary linguistics is in many aspects based on large national corpora carefully crafted by linguists. There are many examples of such 'academic' corpora: British National Corpus, Corpus of Contemporary American English, Turkish National Corpus, Russian National corpus, etc. However, the recent years saw great rise in using text corpora crawled from the Web for linguistic purposes. To some extent they compete with traditional national corpora ([1], [2]).

In this research we compare Russian National Corpus[1] (RNC) to a larger Russian web corpus. Both corpora in some sense represent the Russian language, with the first one being a product of many years of linguistic work on gathering

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


A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 47–58, 2015.
DOI: 10.1007/978-3-319-18111-0_4

texts and annotating them and the second one being a random sample of millions of web documents in Russian.

Scholars have already indicated the problem that academic corpora sometimes present researchers with counter-intuitive features, for example, improbable frequency distribution, which puts on top some peripheral scientific lexis [3], see also [4] on the comparison of genre distribution in English and Russian national and Internet corpora. The implications of incorrect representation of the Russian language in RNC were previously discussed in corpus linguistics community, with web corpora proposed as possible solution [5].

The assumption behind our work is that RNC is obviously influenced by the manual choice of constituent texts and genre proportions. Another limitation is its size (only 230 million tokens in the main corpus). That is why the corpus remains biased in various directions and for many words typical contexts (or lexical meanings, which is in fact the same thing) in RNC are different from the ones used in natural living written language. Certainly, the concept of representativeness is complicated; for the purpose of this research we define it as the ability to reflect the associations which the majority of population would have upon meeting a given lexical unit.

We hypothesize that such ability can be found in a vast and full-featured web corpus which serves as an impartial sample of a living language. It means that we should identify words with the meaning in the web corpus essentially (or totally) different from those in RNC. In this research we show that such lexical units can be discovered with the help of neural language models exploiting vectors as distributed representations of words: a paradigm, which has become quite a buzzword in computational linguistics in the last couple of years [6]. It is possible to use these representations to find topical or functional associates of lexical units. We employ them to solve the aforementioned problem. We also describe categories into which these underrepresented words fall. This knowledge is useful when considering the national corpus' architecture and further development.

In Section 2 we explain the architecture and features of the corpora we compare. Section 3 describes the methodology used in our research and provides background for the model of distributional semantics which we employ to identify differing lexical units. In section 4 we perform the comparison of the two corpora with regard to the differing nouns and observe possible causes for the discrepancies found. Section 5 offers some implications for RNC revealed during our research. Finally, in Section 6 we describe limitations of our experiment and future work.

## 2 Library of Babel vs. Selected Works: The Corpora Used

Opposition of the two corpora under analysis is to some extent similar to that of Selected Works for a writer and the Library of Babel from the famous Borges short story. Russian National Corpus consists of texts which supposedly represent the Russian language as a whole. It has been developed for more than 10

years by a large group of top-ranking linguists, who 'pick' texts and segments for inclusion into the corpus. It was extensively described in the literature[2]. Current composition of RNC is presented on its website[3]. The size of the main part of RNC (without additional sub-corpora) is 230 million word tokens. We worked with the dump containing 174 million tokens. Moreover, to exclude the influence of purely diachronic factors, we restricted ourselves only to texts created after 1950, which amounted to 115 million tokens in total.

Its 'competitor' is a large corpus of texts found on Russian web pages. It originates from a sample of the Russian Internet segment crawled in 2014. This repository contains billions of web documents and actually serves as a source of search index for one of the major search engines in the Russian market, thus is supposed to be quite representative. Consequently, the crawler was sophisticated enough to process even complex dynamic content. Spam and junk pages were also filtered out without our intervention.

To compose the corpus for analysis, we randomly selected about 9 million documents from this repository (no attention was paid to their source or any other properties). Thus, by design the corpus can contain any type of texts found in the Internet (supposedly all major types) in a fully representative proportion. In that, it functions like the Library of Babel, embracing all possible genres and styles in a uniform way.

Boilerplate and templates were filtered out to leave only main textual content of these pages. This was done with the help of *boilerpipe* library [7].

The resulting text archive contained approximately 1.8 billion word tokens. It was split into sentences, and those lacking Cyrillic letters were removed. Thus, we came up with a mainly Russian-language corpus containing about 940 million tokens and 87 million sentences.

Both corpora were lemmatized with the state-of-the-art *MyStem* tool [8]. We used version 3.0 of the software, with disambiguation turned on. It should be noted that some lemmatizer errors became visible later through the output of our language models. For example, for the word *роба* 'boilersuit' the RNC model outputs various types of clothing as topical associates, as expected. However, web corpus model outputs male proper names, obviously, because lemmatizer associated this word and many occurrences of male proper name *Rob* in Genitive, which is homonymous to *роба*. Still, such gross mistakes are rare and do not seriously influence the models in general.

At the stage of lemmatizing, stop-words were removed, as well as single-word sentences (they are useless for constructing context vectors). Then we discovered bigrams which function as integral semantic units using simple data-driven approach proposed in [9] with threshold 1000. These bigrams were joined together with the underscore sign and were treated as single tokens. For example, *углекислый газ* 'carbon dioxide' was transformed into *углекислый_газ*, etc. Such transformation allowed to detect bi-word contextual neighbors which

[2] `http://ruscorpora.ru/corpora-biblio.html`
[3] `http://ruscorpora.ru/en/corpora-stat.html`

otherwise would be split across several word elements (for example, *прямая_наводка* 'direct fire' as an associate for the word *танк* 'tank').

After this preprocessing stage, our RNC corpus was about 70 million tokens in size, and the web corpus shrank to about 620 million tokens. Thus, the web corpus is almost an order of magnitude larger than RNC. It is supported by the size of the corpora lexicons: 751 894 word types for RNC and 5 543 556 word types for the web corpus. So, it seems justified that the latter should at least in some cases provide better contexts for lexical units (and a lot more of lexical units themselves).

## 3 Learning Word Embeddings and Choosing Test Sets to Compare

Our research is performed within the framework of distributional semantics ([10], [11], [12]) and vector space modeling [13]. In particular, we used *word2vec* neural network language model algorithm [14] to learn vector word representations or neural embeddings.

First we recall the basics of neural embeddings. Lexical meaning is generally the sum of word usages, which is quite traditional for distributional semantics. Thus, the most obvious way to capture meaning is to take into account all contexts the word participates in. In other words, this means to represent each word as a vector of its 'neighborhood' to all other words in the lexicon, with various distances and weighting coefficients (Dice, etc). The matrix of $n$ rows and $n$ columns (where $n$ is the size of lexicon) with 'neighborhood degrees' in the cells is then a distributional model of language. One can compare vectors for different words (for example, calculating their cosine similarity) and find how 'far' they are from each other from the point of view of their contexts.

However, this demands operations on sparse but very large matrices. As we saw in the previous section, our RNC corpus features 750 thousand word types. It means that we would have to compute dot products of 750K-length vectors each time we need to know how similar two words are, which is computationally expensive. Vectors' dimensionality can be reduced to reasonable values using methods like singular value decomposition or principal components, but this often degrades performance or quality. This is where neural embeddings come forward. Neural models are directly trained on large corpora to produce vectors or embeddings of a comparatively small size (usually hundreds of components) which maximize similarity between contextual neighbors found in the data, while minimizing similarity for unseen contexts. The dimensions of the resulting vectors cannot be directly mapped to other words as in traditional distributional models. However, if we use them to calculate which lexical units are similar and which are not, they perform surprisingly well and clearly reveal semantic relations between words (see [15] for further details).

Once the embeddings are learned, we can find the nearest neighbors or quazy-synonyms (most topically related words) for any lexical unit. It is as trivial as to iterate through all the embeddings in the model and rank them according to

their cosine similarity with the embedding for the word analyzed. Words with top-ranking embeddings are the quazy-synonyms we looked for (throughout this paper they are also called associates). It is also possible to compare associates' lists for one and the same word produced by different models trained with different parameters or on different corpora, which is what we proposed above.

Thus, we used RNC and the web corpus to train language models with Python *word2vec* implementation[4]. The models were trained with Continuous Bag-of-Words (CBOW) architecture, vector dimensionality of 500 and context window of 5 words to the left and to the right. Also, for the web corpus we ignored the lexical units occurring only once (3 190 000 word types are known to the model) and for the RNC we ignored the lexical units occurring less than 5 times (205 610 word types are known to the model).

This set of parameters was found to be effective during our participation in Russian Semantic Similarity Evaluation track (RUSSE) [16]. The models trained with such settings performed better than the others trained on the same corpora. Thus, we hypothesize that these models are the best we can derive from corpora under analysis given our tools. Notably, the models trained on RNC outperformed web corpus models in semantic relatedness tasks but not in association tasks (see Table 1).

**Table 1.** Results of semantic similarity tasks evaluation for different models

| | Average precision for relatedness tasks | Average precision for association tasks |
|---|---|---|
| RNC model | 0.795 | 0.89 |
| Web corpus model | 0.785 | 0.91 |

Thus, the RNC-based models are better in singling out exact semantic relations (synonymy, hyponymy and hyperonymy), while the web-based ones excel at discovering associations or topical relatedness. It is also impressive that the RNC is generally on par with the web corpus, despite being an order of magnitude smaller. This is another proof that linguistic balance and well-considered composition of the corpus are of much importance and can sometimes outweigh the pompous 'big data'. See more on this in the forthcoming paper[5].

However, for the purpose of the current research it is sufficient to know that the models we have trained are indeed able to distinguish semantically similar words with state-of-the-art quality. This gives us ground to use them for the comparison of lexical units behavior in the two corpora under analysis.

For the following step, we had to select words to compare. Comparing (and interpreting the results of the comparison) for all lexical units in the models is

[4] `http://radimrehurek.com/gensim/models/word2vec.html`

[5] Kutuzov and Andreev 2015, 'Texts in, meaning out: neural language models in semantic similarity tasks for Russian'

unrealistic. What is more, the data for rare words is sparse, and thus the models are not so reliable for the associates computed. That is why we decided to restrict our experiment to 10 thousand top-frequency nominal lexical units: a quantity which one can realistically look through. Nominal units were chosen because it is usually easier to interpret their semantic relations, and neural models perform better for them, as discovered in the course of the above-mentioned experiments in RUSSE track.

Thus, we selected the top 10 thousand nominal units in RNC, which amounted to approximately $\frac{1}{45}$ of all such units in the lexicon. Then we intersected this set with the similar top $\frac{1}{45}$ set from the web corpus and left only units present in both sets. As a result, we got 9113 nouns (of which 197 are bigrams) frequent in both corpora. The nouns at the end of the lists had absolute frequency of 283 and 232 occurrences in RNC and the web corpus accordingly, which corresponds to the relative frequency of nearly 2 ipm for RNC and 0.15 ipm for the web corpus. The reason for this selection was that we did not want to compare frequent units with lots of data about their usage to rare units with very limited contexts. Also, top-ranking words are more likely to be generally important.

We computed 10 nearest neighbors, or associates, for these nouns, using both models (RNC and the web corpus). Here is an example of the output for the word *динозавр* 'dinosaur' (associates are ranked by their cosine similarity to the query word vector).

**RNC:**

1. *мамонт* 'mammoth' 0.397899210453
2. *рептилия* 'reptile' 0.360172241926
3. *млекопитающее* 'mammal' 0.328677803278
4. *ящерица* 'lizard' 0.326320767403
5. *птеродактиль* 'pterodactyl' 0.320571988821
6. *черепаха* 'turtle' 0.308944404125
7. *крыса* 'rat' 0.30866342783
8. *птица* 'bird' 0.308208823204
9. *людоед* 'cannibal' 0.303090155125
10. *вымирать* 'to become extinct' 0.295859247446

**Web Corpus**:

1. *рептилия* 'reptile' 0.496797531843
2. *мамонт* 'mammoth' 0.443771362305
3. *млекопитающее* 'mammal' 0.424831837416
4. *хищный* 'carnivore' 0.412433445454
5. *ящер* 'pangolin' 0.401978999376
6. *крокодил* 'crocodile' 0.396325200796
7. *ящерица* 'lizard' 0.393893510103
8. *черепаха* 'turtle' 0.393123477697
9. *доисторический* 'prehistoric' 0.391041249037
10. *гигантский* 'giant' 0.386854737997

It is obvious that we have two partially intersecting sets of quazy-synonyms or topical associations. The degree of difference between them can be trivially calculated using the Jaccard similarity coefficient (the size of the intersection divided by the size of the union of two sets) [17]. It takes values in the interval [0,1] and serves here as a measure of how much the meaning of a given word is different in the National corpus from the 'unsupervised' web corpus. If the Jaccard coefficient equals to 0, that means that the two sets of neighbors do not intersect and thus the meaning is supposed to be totally different. If, on the contrary, the coefficient takes the value 1, the two sets are identical: both models provided precisely the same set of 10 nearest semantic neighbors. With our data this happened only once with the word *август* 'August', for which both models output names of other months. As for the example above ('dinosaur'), its Jaccard similarity is $\frac{1}{3}$ (5 identical associates of 15 total).

Note also that this way we cannot discover lexical units which RNC lacks altogether or for which it does not provide enough frequency (such a problem can be solved without neural embeddings, by simple comparison of lexicons). Instead, we find discrepancies in the words which are present in the corpus, and even reach top positions in the frequency list.

Our trained models are available online, together with the Jaccard coefficients and example scripts[6].

## 4 What's Different: Analysis of Lexical Units' Neighbors in RNC and the Internet Corpus

We have compared associates for 9113 nouns and noun phrases, considering the Jaccard coefficient between corresponding sets for RNC and the web corpus. 1467 lexical units (about 15%) show the coefficient equal to 0, which means they do not have any neighbors in common. Almost the same number of lexical units (1463) show the Jaccard coefficient higher than 0.3, which means that at least half of their 10 semantic neighbors are the same in RNC and in the web corpus.

About 20 words are as close to maximum Jaccard value as 0.81 (only one differing neighbor). These are mostly months and female patronymics ('*степановна*', '*николаевна*', etc). Strangely, the remaining words with such a high value of Jaccard coefficient all belong to a rather threatening cluster: *бандит* 'bandit', *кинжал* 'dagger', *граната* 'grenade', *конфликт* 'conflict', *пытка* 'torture', *опасение* 'fear'. Supposedly, criminal and military topical segments in RNC are quite consistent with those found 'in the wild'.

In general, the two corpora do agree with each other in most cases. Indeed, considering unsupervised nature of our models, 3 coinciding associates out of 10 is already enough to suppose that both corpora share similar meaning. If so, more than 50% of all units under analysis should be considered as 'agreeing'.

We studied 1467 lexical units with Jaccard coefficient equal to 0, because they are the most critical cases of discrepancy: no coinciding associates at all.

[6] `http://www.cicling.org/2015/data/107`

**Table 2.** Thematic classes of most differing words

| | commerce and finance | politics and law | terminology | high register | recent concepts | Soviet era concepts |
|---|---|---|---|---|---|---|
| Absolute value | 97 | 65 | 59 | 122 | 53 | 24 |
| Percentage | 6.6% | 4.4% | 4.0% | 8.3% | 3.6% | 1.6% |

The aim of our analysis was to reveal possible patterns according to which nouns can fall in this category. The results are demonstrated in the table 2.

The first category which we found within the problematic cases was nouns describing *economics concepts* such as trade, finances, natural resources. There is little wonder that nouns from these categories differ significantly in the two corpora because economics changes rapidly, and, therefore, its concepts in a language also develop quite fast. One of the examples of 'incorrect' neighbor sets in RNC can be the word *брокер* 'broker'. Its neighbors in the web corpus include *биржа* 'exchange market', *дилинг* 'dealing', and *трейдер* 'trader', while in RNC we can find only general terms: *фирма* 'firm', *компания* 'company', *менеджер* 'manager'. Another example is *вакансия* 'vacancy, opening': its neighbors in RNC are *безработный* 'unemployed', *приработок* 'side job', *должность* 'job position', etc, while in the web corpus the word is associated with *резюме* 'resume', *соискатель* 'applicant', *трудоустройство* 'recruitment', and titles of various Russian recruiting web sites, like *job.ru*. It can be seen that in general the web corpus describes these concepts more precisely, although it can contain some unappealing language data, like titles of web sites.

The same cause of discrepancy between the corpora can also be found in the second category, where one finds nouns that refer to *political and social phenomena*. For example, the word *бюллетень* 'bulletin' in RNC produces the following associates: *газета* 'newspaper', *брошюра* 'brochure', *сводка* 'report', *некролог* 'obituary'. In the web corpus, on the contrary, we find a different list: *избирательный бюллетень* 'voting paper', *избиратель* 'voter', *открепительное удостоверение* 'absentee ballot', etc. The senses of this word extracted from both corpora demonstrate radically different shades of meaning: whereas in the web corpus the meaning of the word *бюллетень* is more politically biased, in the RNC it is more official and academic.

In the *terminology* category we find words which are associated with specific professional domains, such as chemistry, physics, or mathematics. These lexical units also differ significantly in the two corpora under analysis. In the web corpus associates are more 'terminological' and more precise than those in RNC. For example, the word *анализатор* 'analyzer' in RNC is associated with *передатчик* 'transmitter', *механизм* 'mechanism', *контроллер* 'controller', while in the web corpus the associates were *спектрометр* 'spectrometer', *глюкоза* 'glucose', *lactate* 'лактат'. Probably, this reflects more specific lexical functions. Similarly, *бензол* 'benzol' in RNC is associated with a few chemical substances like *метанол* 'methanol', but mostly with lexical units like *вата* 'cotton pellet' or

*растворитель* 'organic solvent'. In the web corpus this unit is associated with chemical substances only (e.g., *серная кислота* 'sulfuric acid' and *аммиак* 'ammonia'). It seems that in the web corpus words are employed in more particular contexts, whereas in the RNC their usage is general.

'Literary' words belonging to the *high register*, which are not normally used in speech, were found to constitute 8.3% of all differing lexical units (122 instances. We encountered some interesting examples with regard to these units. For instance, the word *кроха* 'little one, little piece' in the web corpus was associated with *малышка* 'little girl', *младенец* 'baby', *карапуз* 'little child', etc. and clearly denoted 'a kid, a baby', whereas in RNC there were such neighbors as *кусочек* 'little piece', *лихва* 'more than needed', *грош* 'farthing', *посылочка* 'delivery' and the word clearly denoted 'a crumb, a little piece of some object' and not a person. The first set of neighbors and implied meaning indeed seems to be more intuitively correct. The word *приют* 'asylum/orphanage' can serve as another discrepancy observed for words in this category. In RNC, the words like *прибежище* 'refuge', *пристанище* 'haven', *утешение* 'consolation' can be found among its associates, which implies that its meaning is closer to 'asylum'. In the web corpus, however, the associates were *бездомный* 'homeless', *сирота* 'orphan', *детдом* 'orphanage', which points at the second meaning, 'orphanage'. Other words which share similar pattern include *палата* 'chamber/ward', *химера* 'dream/chimera', etc.

However, there are other words in this category which lack the dichotomy of meaning, but are archaic or simply bookish. For these words sometimes neighbors identified in the RNC are more reasonable than those in the web corpus. One example is the word *потеха* 'merry-making', associated in RNC with quite relevant units like *гулянка* 'party' or *перекур* 'smoke-break, while the web corpus outputs less relevant neighbors (except *забава* 'jolly' and to some extent *петросятина*, derogatory derivative of the name of a famous Russian stand-up gag-man Petrosyan).

There was also a minor number of nouns that refer to Soviet era concepts, such as *комсомол* 'Komsomol', *комиссариат* 'comissariat', *партком* 'party committee', etc. With these concepts, more incorrect neighbor sets can be found in the web corpus, while in RNC the associates are better, as it features relatively higher number of texts originating from that time.

Another kind of situation can be observed with respect to nouns describing *contemporary concepts*, like IT and the Internet. RNC, supposedly, does not contain a sufficient amount of texts created in the last ten or fifteen years, when we saw a tremendous growth of Internet usage, and this is the reason why the word embeddings in the web corpus are more accurate and therefore more representative of lexical meaning. For instance, for the word *гиперссылка* 'hyperlink' there are no meaningful associates in RNC (only unclear or unrelated words), and in the web corpus we can find associates like *ссылка* 'link' and *индексировать* 'to index' (and also, interestingly, the same word misspelled: *гипперссылка*).

In spite of all the cases of discrepancy which serve in favor of the web corpus, there are also some negative examples. For instance, the word *нота* 'note/pitch' has the following associates in the web corpus: *мускус* 'musk', *амбра* 'ambergris', *пачули* 'patchouli', etc. Of course, the meaning related to music seems to be more prototypical than the one originating from perfume industry. This is probably a sign of the web corpus bias towards commercials.

Another negative example is the word *бич* 'whip', whose associates in the web corpus are *Таиланд* 'Thailand', *пляж* 'beach', *курорт* 'resort'. The possible reason is the Russian transliteration of English *beach*, frequently used in the web and homonymous to *бич*.

We note that the cases of difference which were analyzed constitute only 28.5% of all the nouns demonstrating the Jaccard coefficient equal to 0. Our analysis is limited to pointing out the most distinctive cases of discrepancy, and it seems impossible to assign an exact category to every noun, either because it falls outside all classifications or because of the fact that, due to the statistical nature of the models, some words possess a set of neighbors which are totally unrelated or are outright junk. This happens because of *MyStem* or *boilerpipe* errors, occasional spam or duplicate texts and other noise factors. However, even the categories revealed above can help to decide with which texts it would be better to augment RNC with.

## 5 Discussion

It should be emphasized that Russian National Corpus still remains the most sustainable and authoritative Russian language corpus. Its composition strategy was a success, and it indeed provides a balanced sample of the national language. This is again proved by the excellent performance of neural semantic models trained on its texts.

Having said that, we have shown that some lexical units in the corpus are surrounded by contexts that make it difficult to grasp the meaning of the word as it is used in the living written language. It means that some bias exists in the corpus data, making it less representative.

Manual discovery of such cases is extremely difficult (if possible at all). At the same time, comparing neural language models trained on the RNC to those trained on other corpora allows to perform this task in an unsupervised way, extracting necessary data automatically from lexical co-occurrences. Web corpora are good candidates for such comparisons, as in constructing them we at least partially avoid human selection bias: in our case documents are drawn randomly from the 'Babel library' of the Internet (from its representative model crawled by a search engine, to be exact). Such corpora are a useful resource to enhance 'academic' RNC corpus and make it more up-to-date through applying linguistic data from the web to augment the corpus with the texts of particular categories.

Even based on our initial research one can conclude that the RNC maintainers should pay more attention to texts related to economics, politics, law and the Internet (and possibly some other rapidly changing spheres of our life). At the

same time, detailed analysis in order to determine precise areas of expanding is certainly needed. Also, sometimes it is difficult to decide which set of associates (which meaning) is 'correct': this is the case with many 'high register' words. Should *приют* be more of a refuge for a tired soul or should it be an orphanage for children who lost their parents and pets who lost their hosts? If one of these meanings is a bit archaic, does it mean that the National corpus should reduce its presence?

Such questions are not easy to answer, but corpus linguists should at least possess the tools to measure the degree of disagreement between National corpora and other linguistic resources. One of such tools is presented here.

## 6 Limitations and Future Work

Main limitations of our experiment concern the composition of the web corpus. First of all, no duplicates were removed, thus, multiple copies of texts are undoubtedly present in the corpus. According to some estimates, about 40% of all web pages on the Internet are duplicates [18], which means that this can become a harsh problem and critically dis-balance the corpus. Thus, we are going to experiment with de-duplicating our web corpus using shingles or other established approaches, and see whether this would change the results.

Another issue is the language of web pages in the corpus. We did not apply proper language detection and considered all sentences with Cyrillic characters to be 'Russian', which led to some noise. For example, the word *свая* 'pile, pole' is characterized in the web corpus by a set of 'neighbors', all of which seem to be Belorussian ('*цяпер будза яго ён калі пра толькі гэт якать ў*'). It means that there is a sufficient amount of Belorussian texts to pop a Belorussian reflexive pronoun up above the homonymous Russian noun.

Therefore, we plan to apply a simple n-gram based language detector to the corpus and to select only Russian texts in order to avoid multilingual noise.

Another kind of noise is created by a widespread use of English lexicon in the Web. This problem is more difficult to solve as English (and non-Cyrillic in general) words are not always inappropriate in the corpus.

Finally, it would be interesting to research into the behavior of other lexical classes (especially verbs) and multi-word entities more than 2 words length. This should induce more insights about the essence of lexical differences between the RNC and the web corpus.

**Acknowledgments.** The authors cordially thank Igor Andreev of Mail.ru Search applied linguistics team for his inspiring idea. Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

## References


1. Kilgarriff, A., Grefenstette, G.: Introduction to the special issue on the web as corpus. Computational Linguistics 29(3), 333–347 (2003)
2. Baroni, M., Ueyama, M.: Building general-and special-purpose corpora by web crawling. In: Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application, pp. 31–40 (2006)
3. Belikov, V.: What are sociolinguists and lexicographers lacking in a digitized world? (in Russian). In: Proceedings of the Dialog Conference (2011)
4. Sharoff, S.: In the garden and in the jungle: Comparing genres in the bnc and the internet. In: Genres on the Web, pp. 149–166. Springer (2011)
5. Belikov, V., Kopylov, N., Piperski, A., Selegey, V., Sharoff, S.: Corpus as language: from scalability to register variation (in Russian). In: Proceeding of the Dialog Conference (2013)
6. Baroni, M., Dinu, G., Kruszewski, G.: Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 1 (2014)
7. Kohlschütter, C., Fankhauser, P., Nejdl, W.: Boilerplate detection using shallow text features. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 441–450. ACM (2010)
8. Segalovich, I.: A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In: MLMTA, Citeseer, pp. 273–280 (2003)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
10. Curran, J.R.: From distributional to semantic similarity. PhD thesis, University of Edinburgh (2004)
11. Lenci, A.: Distributional semantics in linguistic and cognitive research. Italian Journal of Linguistics 20(1), 1–31 (2008)
12. Bruni, E., Tran, G.B., Baroni, M.: Distributional semantics from text and images. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, pp. 22–32 (2011)
13. Turney, P.D., Pantel, P., et al.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37(1), 141–188 (2010)
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
15. Levy, O., Goldberg, Y.: Dependency-based word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, vol. 2 (2014)
16. Panchenko, A., Loukachevitch, N.V., Ustalov, D., Paperno, D., Meyer, C.M., Konstantinova, N.: Russe: The first workshop on russian semantic similarity. In: Proceeding of the Dialogue 2015 Conference (2015)
17. Jaccard, P.: Distribution de la Flore Alpine: dans le Bassin des dranses et dans quelques régions voisines. Rouge (1901)
18. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval, vol. 1. Cambridge university press, Cambridge (2008)