Linguistic Big Data: Problem of Purity and Representativeness

Valery Solovyev [1[0000-0003-4692-2564]] and
Svetlana Akhtyamova [2[0000-0001-5863-9841]]

[1] Kazan (Volga Region) Federal University, Kazan, Russia
[2] Kazan National Research Technological University, Kazan, Russia
Maki.solovyev@mail.ru

Abstract. This paper deals with the quality problem of linguistic big data exemplified by the corpus of Google Books Ngram. The criticism of this corpus has been summarized and discussed. Special attention is paid to the matters of the corpus balance, spelling errors, and errors in metadata. It is also compared to the Russian National Corpus and to the General Internet-Corpus of Russian. A new concept, "diachronically balanced corpus", has been introduced. The methods are discussed for enhancing the quality of Google Books Ngram.

Keywords: Text Corpora, Data Representativeness, Time-series, Data Noisiness.

1 Introduction

Modern linguistic studies can hardly go without using large text corpora or linguistic databases of various types or without applying to them computer-based or mathematical methods to obtain valid statistics. For the Russian language, the Russian National Corpus [1] (abbreviated as RNC) is well known, which is specified in [2, 3]. RNC contains over 350 million words [1]. It is carefully checked, and its part has been disambiguated manually. All this make it exceptionally useful for research in the Russian language.

A new, interesting resource appeared recently, the General Internet-Corpus of Russian (http://www.webcorpora.ru/, abbreviated as GICR). It already contains over 20 billion words and is thought to be further expanded. GICR includes the contents from the largest resources of the Runet, such as Zhurnalny Zal (Room of Thick Literary Journals), Novosti (News), VKontakte, LiveJournal, and Blogs at Mail.ru.

Since 2009, there has been an even larger corpus, Google Books Ngram (https://books.google.com/ngrams, abbreviated as GBN). It contains data on 9 languages, including Russian. The volume of the Russian GBN sub-corpus is over 67 billion words, while it exceeds 500 billion for the English one. GBN was created by fully scanning, followed by text recognition, all books from over 40 largest libraries in the world, including those of Harvard University and Oxford University. As a result, 30 million books were digitalized, of which the 8 million best digitalized ones were selected to form the corpus, which amounted to 6 % of all the books published worldwide [4]. A detailed description of the GBN can be found in [4–6]. For so large corpora, there seems no escaping the matters of their quality and of the possibility of errors in creating them. In this paper, we are focusing on GBN as the largest corpus of all the corpora existing in the world. Some publications noted errors in the corpus [7–9]. There are three core GBN problems discussed in literature: OCR errors, balance of the corpus, and errors in metadata. Herein, we are considering the above problems and the possible ways of improving the corpus. The paper presents a review of key publications related to the GBN issues, as well as the unique results obtained by the authors in this area.

2 OCR Errors

There are recognition errors in the GBN corpus, which are primarily related to ancient books characterized by poor print quality. In the first GBN version released in 2009, things were really in a bad way. Thus, in ancient English books, letter *s* was frequently recognized as *f*. For example, the word *best* was mistaken for *beft* in up to 50 % of the 17th-century books recognized. The creator of this resource, Google, has considered the criticism and considerably improved the recognition quality. Scanning devices were upgraded every six months [6]. As a result, in the second version in 2012, *best* was incorrectly recognized as *beft* in just 10% of cases in the 17th-century books, while in contemporary books of 2000, the amount of errors only made 0.02%, that is, it was rather low and could not affect any statistics regarding the frequency of using the word *best*.

Similarly, that is the case for the Russian language. We have considered several dozens of randomly chosen words containing recognition errors. Typically, the error rate does not exceed 0.1 %. For example, letter н is sometimes recognized as и. In Fig, 1, the exemplary frequency diagrams are shown for the word “иней” (hoarfrost). The frequency of its incorrect recognition as “иией” is lower than 0.1% of the correct one.

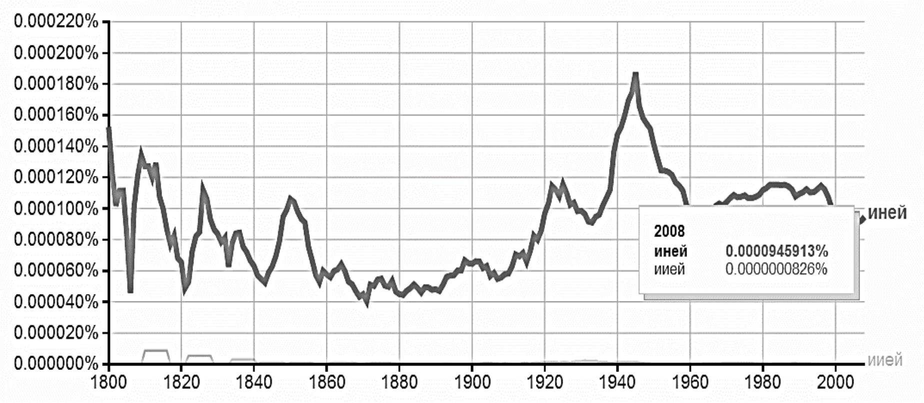

Fig. 1. Frequency of *иней* and *иией*

For Russian, certain difficulties occur with the pre-reform (before 1918) orthography. The Russian language previously used letters, such as Ѣ (yat) and Ѳ (fita), that are incorrectly recognized in GBN. For the data beyond 1918, the problem is eliminated. For many words from texts issued before 1918, the data will also be correct, since letters Ѣ and Ѳ, as well as other elements of the old orthography, are rather uncommon.

Thus, there are no apparent reasons for considering that recognition errors may essentially affect the results of counting the frequency of word usage, except for some probable rare cases with ancient books, where certain care must be exercised.

3 Balance

As a matter of principle, the problem of corpus balance is considerably more complicated. Balanced should be considered a corpus, in which all types of texts, i.e., literary, journalistic, pedagogic, scientific, business, and other texts, occur in the corpus proportionally to their shares within the texts of the chosen period [1]. It is commonly supposed that the RNC is well-balanced, which is ensured by the efforts of its developers who have "hand-picked" the texts for the corpus. GBN was created by a very different technique, its composition was not specially designed, so GBN is often faulted for being unbalanced [7, 8].

In [8], a radical opinion is expressed: "Therefore, instead of speaking about general linguistic or cultural change, it seems to be preferable to explicitly restrict the results to linguistic or cultural change 'as it is represented in the Google Ngram data'". In fact, the author of [7] is on the same side of the fence, suggesting that a well-balanced corpus is a utopia and that any data obtained based on a corpus reflects the content of that corpus rather than the language state.

If this were really the case, the creation of text corpora, on which many efforts have been focused, would become a little promising activity. Then corpora would be just a set of examples linguists can extract to quote in their articles, and they would not be suitable as a tool for fundamental research in the essence of a language. Fortunately, this is not the case. The best proof of the entire language representation adequacy in large corpora is the reproducibility of results demonstrated on different corpora. Let us give a simple example.

In [10], the changes were considered regarding the frequency of the members of a synonymic row that includes the words *стараться* (try) and *пытаться* (attempt). In Figures 2 and 3, the graphs are shown for the most frequently used words from the inflectional paradigm, i.e., *старался* and *пытался* (both are past-tense third-person singular masculine verbs), for GBN and RNC. The trend of the last two centuries is clearly in evidence for both corpora – *пытался* becomes more frequent than *старался*. The appearance the graphs in GBN is smoother since this corpus is larger. Even the period where the word *пытался* becomes more frequent is the same – around the year 1960. For highly-frequent words, the graphs of GBN and RNC are usually similar, as well.

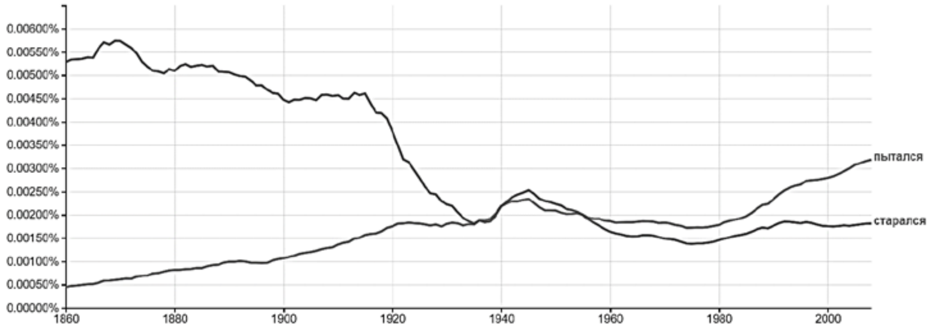

Fig. 2. Frequencies of the words *старался* and *пытался* for GBN

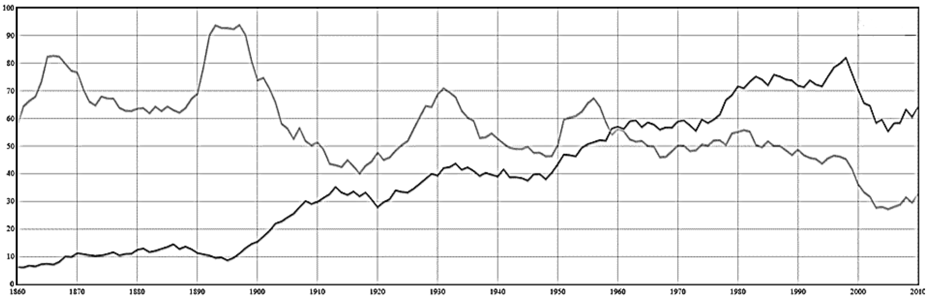

Fig. 3. Frequencies of the words *старался* and *пытался* (black line) for RNC

Such agreement of results obtained on different corpora both validates the results as such and indicates the high quality of the corpora and their consistency. Unfortunately, we cannot directly check everything on GICR within the above timeframe, since almost all GICR texts are dated the 21st century and it has just started growing deeper recently.

However, there is a curious possibility to perform an indirect comparison. Time series generated based on diachronic corpora provide a great opportunity to predict the development of the language. So far, no quantitative predictions regarding language changes have been made based on corpora. In this paper, we are probably making one of the first attempts of this kind. The scheme proposed for extrapolating time series can be useful to the research in various language-specific phenomena.

In Table 1, the GBN-based frequencies of the words *пытался* and *старался* are shown in 1978 and in 2008, as well as the ratio of the former of those values to the latter one. Using the linear regression method, we compute the expected values of the frequencies for the year 2014. That year was chosen to compare our predictions with the data of [6], in which the time interval is limited to the years 2014–2015 being available to the authors at the time of writing their work. The increased number of uses of *пытался* as compared with *старался* over a 30-year period in 1978–2008 allows expecting its further growth by 2014.

Table 1. Known and predicted frequencies of the words *пытался* and *старался* in GBN

Word	1978	2008	Prediction for 2014
пытался	0.00173	0.00318	0.00347
старался	0.00140	0.00181	0.00189
пытался/старался	1.24	1.75	1.84

Let us consider the GICR-based data from [7]. We divide all the GICR sub-corpora into three groups that differ in their genres and styles: 1) Zhurnalny Zal that contains texts from literary journals and is the closest one to GBN; 2) Novosti that contains the texts of another genre, and 3) LiveJournal and VKontakte, both containing texts that fundamentally differ from book texts. Hence, we can expect that data for Zhurnalny Zal will be similar to that of GBN, while the data for Novosti and for the sub-corpora of the third group will be different.

Indeed, the ratio of the *пытался* usage number to the *старался* usage number makes 1.94 in Zhurnalny Zal, 10.41 in Novosti, and 3.30 in the sub-corpora of the third group [7]. Thus, we can see that the text genres and styles are, of course, of great importance. At the same time, for the texts of a similar nature, such as books from GBN and articles from literary journals, the values predicted based on GBN and the real values have turned out to be very close to each other, differing by less than 6%. This also indicates the high quality of the corpora being compared and the possibility to obtain rather correct predictions based on GBN.

Unfortunately, not all the studies performed on GBN can be repeated on RNC or GICR. This is because, unlike GBN, the RNC and GICR corpora are not available to users for downloading. This limits the possibilities of processing the RNC and GICR data with simple queries and does not allow applying complex computer-aided and mathematical data-processing methods that are widely used in contemporary research. The latter ones include measuring the distances between languages at some point in time or between the states of a single language at different time instants, using Kullback-Leibler's metrics [14].

In our opinion, GBN is exactly an example of the best-balanced corpus, as well-balanced as possible. Since all the books from the largest libraries were scanned when creating it, this results in all types of texts being represented in GBN proportionally to the representativeness thereof in the libraries. So GBN is as balanced as the entire human-created library system is balanced.This result cannot be achieved by manually selecting texts.

Let us discuss the term of balance. Here, the balance shall mean the generally balanced corpus as a cohesive whole. In the paper, we introduce a new term: Diachronic balance corpus.

We will apply the concept of "diachronically balanced" to a diachronic corpus that is balanced for any given moment of time, ideally for every year or decade. That is, a corpus sample within any small timespan shall already be balanced as such.

Until now, the problem of creating diachronically balanced corpora has not even been stated. However, the giant volume of GBN, as well as the adopted ideology of total scanning, make this corpus exactly like that. For Russian, over the past decades, the volume of the corpus has made about 1 billion words per year, which is triple as much as the volume of the entire RNC. For English, the corpus volume is 10 times more. Naturally, we cannot prove the diachronic balance of GBN, since there is no operational definition of balance, which would allow us to consider corpora as balanced or unbalanced ones.

Let us consider a specific example demonstrating the degree of the GBN balance as compared with RNC. In the USSR of late 1980s, the word *ускорение* (acceleration) borrowed from physics was embedded in the political vocabulary, which word meant the accelerated development of the country's economy. This term started to be widely used in political essays after April 23, 1985 on which day M.S. Gorbachev declared at the Plenum of Central Committee of the Communist Party of the USSR (CC CPSU) a large-scaled program of reforms under the slogan of accelerating the social and economic development of the country. However, just 2 years later, in January 1987, at the Plenum of the CC CPSU the task was stated aimed at cardinally reconstructing the economy management. The new slogan of *перестройка* (reconstruction) appeared, and *ускорение* started becoming irrelevant. Let us have a look at how frequently the word *ускорение* was used in GBN and RNC.

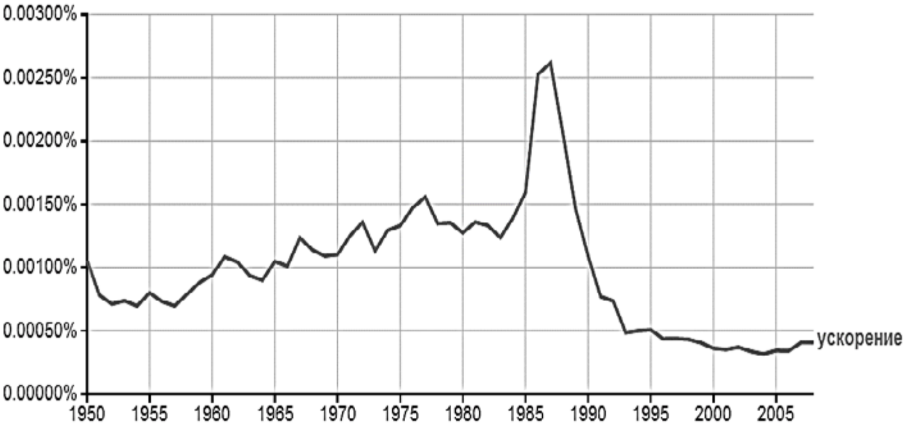

Fig. 4. Frequency of the *ускорение* uses in GBN

In Fig. 4 above, we can see that the sharp rise in the frequency of using the word *ускорение* falls exactly within the year 1985, and its frequency sharply decreases, starting from 1987. Thus, GBN reflects adequately the volume of socially- and politically-focused literature at that time and exactly reflects the processes running in the society. It is also noted in [34] that changes in languages registered in GBN correlate with social events. What about RNC?

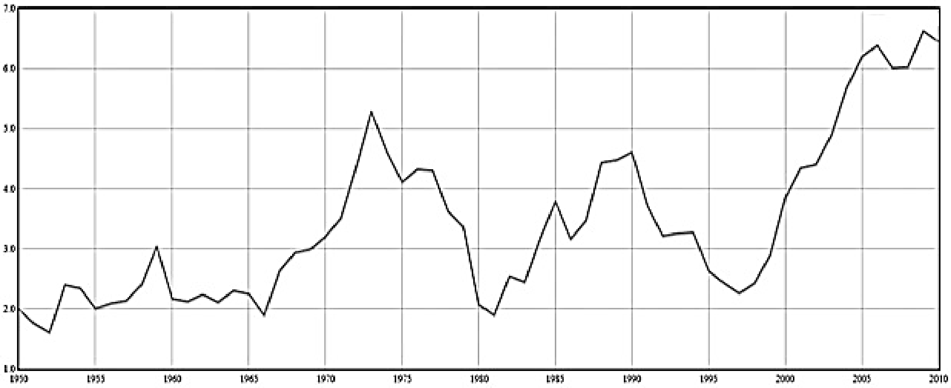

Fig. 5. Frequency of the *ускорение* uses in RNC

In Fig. 5 above, no growth in the frequency of the word *ускорение* can be seen in RNC for that period. Moreover, the frequency of *ускорение* starts falling in 1985 and growing in 1987. At the same time, no political texts are found among the specific ones containing the word ускорение in RNC in those years. This is, of course, just one example. However, it makes it clear how difficult it is to ensure the diachronic balance in manually assembling the corpus, and how naturally it occurs by itself in total digitalizing.

4 Errors in Metadata

In [9], a metadata error was found in the English sub-corpus Fiction. In the first version, many scientific books got into it. This was found based on considering the use in the Fiction corpus the word typical of scientific texts, i.e., ‘Figure’, compared to the word ‘figure’ (lowercased) that may occur in literary works, as well. In Fig. 6, you can see well the unnatural growth of the uses of ‘Figure’, which corresponds in time with the exponential growth of scientific publications.

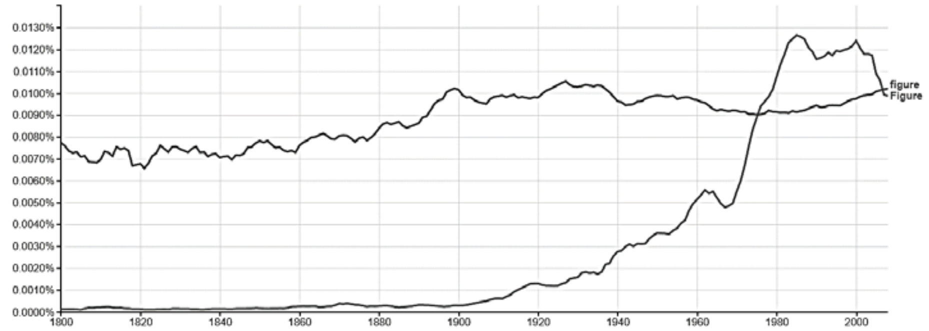

Fig. 6. Frequencies of using ‘Figure’ and ‘figure’ in the Fiction corpus, the version of 2009

This was considered in the second version of the corpus, and the books were classified correctly. Therefore, the frequency of using the word ‘Figure’ in the Fiction subcorpus fell 20 times (Fig. 7).

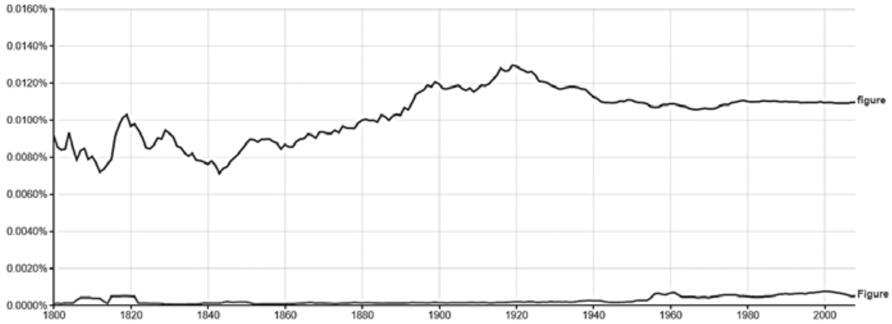

Fig. 7. Frequencies of using ‘Figure’ and ‘figure’ in the Fiction corpus, the version of 2012

Thus, in that case, again, Google rapidly responded to criticism, and the error was corrected.

Further, the authors of [8, 9] returned to using GBN in their studies of the language evolution [11, 34].

5 Using GBN

Despite the problems of the corpus mentioned above, it is widely used in various linguistic and culturological studies. There are over 6,500 articles mentioning GBN in the Google Scholar system. 187 works have already been published within the first 3.5 months of 2019. Any review of those works is far beyond the scope of this paper. However, we would like to note some of the most interesting and typical, in our opinion, trends in research, demonstrating the considerable room for using GBN in Digital Humanities.

In linguistics, the matters have been studied, such as the number and the changes in the number of words within the language vocabulary [12], the dynamics in the “births” and “deaths” of words [13], the speed of evolving the languages and their vocabularies [9, 14], the mechanisms of competing the regular and irregular forms of verbs in English [5], and comparing British English and American English [14].

In psychology, emotions [15–18] and cognitive processes [19–21] have been studied. One of the most popular matters turned out to be the changes in the psychology of collectivism/individualism. Of many works in this area, we would like to note articles [22–25], in which the growth of individualism was tracked in different countries, as exemplified by English, German, Russian, and Chinese.

In social studies, the research has been performed in gender differences and diversity [26, 27] and in global cultural trends [28].

6 Directions in Enhancing the Results

We can propose two ways of enhancing the reliability of the results obtained using GBN. The first one consists in using and comparing all types of data that can be extracted from GBN. This particularly includes considering, along with the word itself, its various inflectional forms [29] and synonyms [24]. In [29], this is exemplified with the German word *eigen* (own, peculiar), which is relatively rare to occur in this form, while it 35 times more frequently occurs in the form of *eigenen*. In [24], it is recommended to use each word with three synonyms selected in the relevant dictionaries of synonyms. Should your research be of intercultural nature, then it is natural to use corpora for several languages represented in GBN [29] in order to compare the dynamics of the frequencies of the same or close terms. For research in English, GBN provides several corpora, such as general English, American English, British English, and the Fiction corpus. They can also be used to compare and verify the results obtained. For example, in [30], the dynamics of the first-person pronoun frequencies can be tracked using both the English corpus and the fiction corpus.

The second way consists in preprocessing raw data provided by GBN. Although this way is rather labour-consuming, it can still be recommended. In [14], the corpus preprocessing is described that consists in removing all tokens (character strings) that are not words. All tokens are deleted that contain numbers or other non-alphabetic characters, except for apostrophes. (The '–' symbol is processed by the GBN system itself.) This is probably especially topical for the languages that have undergone spelling reforms, such as the Russian language. The 1918 reform removed ъ (hard sign) at the ends of masculine words ending with a consonant. To process those words correctly, it would be reasonable to delete ъ at the ends of all such words. This is just a realistic way for a researcher, which allows correctly processing an enormous number of Russian words – practically all masculine nouns.

We can find other systemic changes in spelling the words and correct the corpus accordingly in compliance with the current spelling rules. Replacing ancient orthography with the modern spelling is adopted in RNC. This is reasonable, of course, for the studies only that do not focus on researching in ancient orthography.

It is unreal, of course, to eliminate all the errors in a multibillion-word corpus. Therefore, it would be reasonable to try and apply the recently-developed methods of working with noisy language data [31, 32].

7 Conclusion

Creating very large specialized and multi-use text corpora is important for both theoretical and applied research in linguistics and allied areas of knowledge. Very large text corpora, especially diachronic ones, create fundamentally new opportunities for studies that just could not have been performed without them. GBN corpus presents very accurately both the language changes and the processes occurring in the society and reflected in the language. This allows using this corpus in various humanities research. Diachronic corpora provide a researcher with the opportunities for both describing the language properties observed and reasonably predicting about their further developments.

Creating such corpora is an extremely complicated and labour-consuming activity, and the matters regarding the quality of the corpora created emerge inevitably. If texts recently published using computer-aided techniques are quite "pure," then the ancient books and periodicals must be scanned followed by recognizing the characters, which inherently leads to errors in the corpus. In this paper, we have considered the case of the currently largest diachronic corpus, GBN. It is shown that the main errors of the earlier version have already been eliminated in the next version of the corpus. The remaining specific minor errors are invalidated in statistical computations on a big data array. However, for Russian, some problems persist, which are related to the ancient spelling and which it would be reasonable to solve.

As to the most important issue regarding the balance/representativeness of GBN, the conceivable case for the fact has been made out that the corpus is highly balanced. GBN is specifically compared with RNC and GICR, which comparison has demonstrated their high consistency. The latest versions of spelling correction systems may be used, as well [33].

Acknowledgment

This research was financially supported by the Russian Foundation for Basic Research (Grant No. 17-29-09163), the Government Program of Competitive Development of Kazan Federal University, and through the State Assignment in the Area of Scientific Activities for Kazan Federal University, agreement No 2.8303.2017/8.9.

References

1. Russian National Corpus. http://www.ruscorpora.ru/. (2019)
2. Russian National Corpus: 2003–2005. Indrik, Moscow. (2005). (in Russian).
3. Russian National Corpus: 2006-2008. V.A. Plungyan, ed. Nestor-Istoriya. St. Petersburg. (2009) (in Russian).
4. Lin, Y., Michel, J.-B., Aiden, E., Orwant, J., Brockman, W., and Petrov, S.: Syntactic Annotations for the Google Books Ngram Corpus. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics vol, 2: Demo Papers (ACL '12) (2012).
5. Michel, J. et al.: Quantitative Analysis of Culture Using Millions of Digitized Books. Science **331 (**6014), 176–182 (2011).
6. Aiden, E. and Michel, J.-B.: Uncharted Big Data as a Lens on Human Culture. Russian edition: Moscow. AST. 352 p. (2016) (In Russian).
7. Belikov, V.I.: What and how can a linguist get from digitized texts? In: Sibirsky philologichesky zhurnal [Siberian Journal of Philology] (3), 17–34 (2016) (In Russian).
8. Koplenig, A.: The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German corpus in times of WWII, *Digital Scholarship in the Humanities* **32**, 169–188 (2017). https://doi.org/10.1093/llc/fqv037
9. Pechenick, E.A., Danforth, Ch.M., Dodds, P.Sh., and Barrat, A.: Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. PLOS ONE. 10 (10): e0137041 (2015).
10. Solovyev, V.D.: Possible mechanisms of change in the cognitive structure of synonym sets. In: Language and Thought: Contemporary Cognitive Linguistics. A collection of articles. Languages of Slavic Culture. Moscow, 478–487 (2015).

11. Pechenick, E.A., Danforth, Ch.M., and Dodds, P.Sh.: Is language evolution grinding to a halt? The scaling of lexical turbulence in English fiction suggests it is not. J. Comput. Science **21**, 24–37 (2017).
12. Petersen, A.M., Tenenbaum, J., Havlin, S., Stanley, H. E., and Perc, M.: Languages Cool as They Expand: Allometric Scaling and the Decreasing Need for New Words, Sci. Rep. 2, 943 p. (2012).
13. Petersen, A.M., Tenenbaum, J., Havlin, S., and Stanley, H.E.: Statistical laws governing fluctuations in word use from word birth to word death. Scientific Reports 2, 313 p. (2012).
14. Bochkarev, V., Solovyev, V., and Wichmann, S.: Universals versus historical contingencies in lexical evolution. J. R. Soc. Interface **11** (101). DOI: 10.1098/rsif.2014.0841. (2014)
15. Acerbi, A., Lampos, V., Garnett, P., and Bentley, R.A.: The expression of emotions in 20th century books. PloS One **8** (3), (2013), e59030. https://doi.org/10.1371/journal.pone.0059030 PMID: 23527080
16. Mohammad, S.M.: From once upon a time to happily ever after: Tracking emotions in mail and books. Decision Support Systems **53** (4), 730–741 (2012).
17. Morin, O. and Acerbi, A.: Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. Cognition and Emotion **31** (8), 1663–1675 (2017). https://doi.org/10.1080/02699931. (2016). 1260528 PMID: 27910735
18. Scheff, T.: Toward defining basic emotions. Qualitative Inquiry **21** (2), 111–121 (2015).
19. Ellis, D.A., Wiseman, R., and Jenkins, R.: Mental representations of weekdays. PloS One **10** (8), e0134555 (2015). https://doi.org/10.1371/journal.pone.0134555 PMID: 26288194.
20. Hills, T.T. and Adelman, J. S..: Recent evolution of learnability in American English from 1800 to 2000. Cognition **143**, 87–92 (2015). https://doi.org/10.1016/j.cognition.2015.06.009 PMID: 26117487.
21. Virues-Ortega J. and Pear J.J.: A history of “behavior” and “mind”: Use of behavioral and cognitive terms in the 20th century. The Psychological Record **65** (1), 23–30 (2015).
22. Greenfield, P.M.: The changing psychology of culture from 1800 through 2000. Psychological Science **24** (9), 1722–1731 (2013). https://doi.org/10.1177/0956797613479387 PMID: 23925305
23. Zeng, R. and Greenfield, P.M.: Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values. International Journal of Psychology **50** (1), 47–55 (2015). https://doi.org/10.1002/ijop.12125 PMID: 25611928
24. Younes, N. and Reips, U.-D.: The changing psychology of culture in German-speaking countries: A Google Ngram study. International Journal of Psychology **53**, 53–62 (2018). https://doi.org/10.1002/ijop. 12428 PMID: 28474338
25. Velichkovsky, B.B., Solovyev, V.D., Bochkarev, V.V., and Ishkineeva, F.F.: Transition to market economy promotes individualistic values: Analysing changes in frequencies of Russian words from 1980 to 2008. International Journal of Psychology (2017).
26. Del Giudice, M.: The twentieth century reversal of pink-blue gender coding: A scientific urban legend? Archives of Sexual Behavior **41** (6), 1321–1323 (2012). https://doi.org/10.1007/s10508-012-0002-z PMID: 22821170
27. Ye, S., Cai, S., Chen, C., Wan, Q. and Qian, X.: How have males and females been described over the past two centuries? An analysis of Big-Five personality-related adjectives in the Google English Books. Journal of Research in Personality **76**, 6–16 (2018).
28. Bochkarev, V.V., Shevlyakova, A.V., and Solovyev, V.D.: The average word length dynamics as an indicator of cultural changes in society. Social evolution & History **14** (2), 153–175 (2015).
29. Younes, N. and Reips, U.-D.: Guideline for improving the reliability of Google Ngram studies: Evidence from religious terms. PLoS ONE **14** (3), e0213554 (2019). https://doi.org/10.1371/journal.pone.0213554.

30. Twenge, J.M., Campbell, W.K., and Gentile, B.: Changes in pronoun use in American books and the rise of individualism, 1960–2008. Journal of Cross-Cultural Psychology **44** (3), 406–415 (2013).
31. Malykh, V. and Lyalin, V.: Named Entity Recognition in Noisy Domains. In: The Proceedings of the 2018 International Conference on Artificial Intelligence: Applications and Innovations. ISBN: 978-1-7281-0412-6 (2018).
32. Malykh, V. and Khakhulin, T.: Noise Robustness in Aspect Extraction Task. In: The Proceedings of the 2018 Ivannikov ISPRAS Open Conference. (2018). ISBN: 978-1-7281-1275-6.
33. Anisimov, I., Polyakov, V., Makarova, E., and Solovyev, V.: Spelling correction in English: Joint use of bi-grams and chunking. 2017 Intelligent Systems Conference, IEEE Xplore Digital library (2018). https://ieeexplore.ieee.org/document/8324234
34. Koplenig, A.: A fully data-driven method to identify (correlated) changes in diachronic corpora. https://arxiv.org/ftp/arxiv/papers/1508/1508.06374.pdf (2015).