Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2025”

April 23–25, 2025

An Algorithm for Genre Imbalance Correction in the Russian Subcorpus of the Google Books Ngram Corpus

Anna Ivleva
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
ivleva.anna.igorevna@gmail.com

Valery Solovyev
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru

Abstract

Recently, the latest version of the Google Books Ngram was shown to be imbalanced. This corpus being an important and widely used database, the imbalance can affect the results of research in different areas. In the paper, we present an algorithm for correcting the dynamics of word frequency in the Google Books Ngram corpus. The algorithm takes into account the discovered imbalance of the main literary styles—fiction, publicistics, and non-fiction. The rationale of the algorithm is given, as well as examples of correction.

Keywords: the Google Books Ngram corpus; genre imbalance; word frequency; time series correction
DOI: 10.28995/2075-7182-2025-23-XX-XX

Алгоритм коррекции жанрового дисбаланса в подкорпусе русского языка Google Books Ngram

Аннотация

Недавно было показано, что последняя версия Google Books Ngram — важного исследовательского цифрового ресурса — не сбалансирована, что может исказить результаты использующих его работ. В статье представлен алгоритм коррекции динамики частоты слов в корпусе Google Books Ngram с учетом обнаруженного дисбаланса основных стилей литературы — художественного, публицистического и нехудожественной литературы. Приводится обоснование алгоритма, а также примеры коррекции частот слов.

Ключевые слова: корпус Google Books Ngram, дисбаланс стилей, частота слова, коррекция временных рядов

1 Introduction

The Google Books Ngram corpus (https://books.google.com/ngrams, hereinafter abbreviated GBN) continues to be a popular research tool in linguistics, NLP, as well as in many social and cultural studies [1, 2, 3]. Up-to-date review of the corpus application is given in [4]. It gives access to large diachronic data on word frequencies significant for quantitative studies of the evolution of both language and society. The GBN corpus was made by overall scanning of books from the university libraries [1, 5]. Now there are three GBN versions – 2009, 2012, and the last of 2020 recently replenished with the data up to 2022.

The diachronic genre representativeness and stability is one of the most significant parameters. This is due to the fact that word frequency trends can be treated as indicators of societal and/or language changes. But changes in word frequency can be also caused by the imbalanced genre composition of the corpus. The issues of the GBN balance and representativeness were debated in [6, 7]. Some problems were pointed out, which Google promptly took into account and made changes to the corpus structure. As a result, [6] concluded that the second version of GBN became up-to-date, representative and balanced.

Previously, we came across some peculiarities of the word frequency in the Russian subcorpus of the last version of GBN. First of all, we found the frequency dynamics of the words specific to fiction to be highly synchronous after 1999. In [8], the authors gave specific examples of this sort and provided a more detailed and general analysis of the corpus based on the frequency dynamics of 7420 words from the Frequency Vocabulary of Fiction, Frequency Vocabulary of Publicistics and Frequency Vocabulary of Non-fiction [9]. We analyzed the average variation coefficient, the average autocorrelations, as well as some other statistical parameters of the frequency dynamics. Specific years and intervals with significant imbalance of genres were obtained. We concluded unjustified overall increasing trend in the share of fiction from 2000 to 2019, with essential outbreaks in 2000, 2011–2012 and 2018–2019.

If we take the frequency data directly from the third version of the GBN corpus, it can result in incorrect conclusions. For example, we can see the frequency of the pronoun *я (I)* increasing. As noted in [8], according to the general approach developed in [9], this can be interpreted as increase in individualism in society. But it can also be caused by the growth of the share of fiction, where the pronoun *я (I)* is much more frequent than in other genres. The issue gets even more complex and complicated due to the fact that Google does not publish the shares of texts genres and styles.

The present research is devoted to the algorithm of word frequency correction to compensate for the previously found imbalances.

The second section of the article describes the algorithm for genre imbalance correction and some examples of its application. In the 3rd section, we discuss the algorithm operation and results. In conclusion, we summarize the results of our research.

2 Results

We correct the word frequency dynamics by applying the time series analysis.

At the first stage, standalone outbreaks of word frequencies are corrected. We consider that in case of balance, the average noise – the fluctuations that do not relate to any trends or other systematic changes – is to be close to zero for large lists of words as individual frequency dynamics compensate for one another. That is why outbreaks in the averaged noise are treated as indicators of imbalance.

The second stage is the correction of word frequency trends. We analyze the trends for the words specific to different genres and correct the unjustified trends of the XXI century.

2.1 Outliers correction

First, we made preliminary analysis of highly specific words. The lists of the words (930 fiction, 434 publicistics and 981 non-fiction words) were taken from [9], auxiliary parts of speech were excluded. Also, the list of out-of-genre words was prepared on the basis of 20000 most frequent words from [9]. Words from the above lists of highly specific genre words, as well as auxiliary parts of speech were excluded from it. As a result, the list of out-of-genre words contained 17468 words. We applied the method of singular spectral analysis (SSA) [10] to extract the noise from the data. In Figure 1, we can see the average noise for out-of-genre words. Several outbreaks down in 1964, 2013 and up in 2000 show some general disbalance of the corpus composition or some mistakes not connected with the imbalance of the three genres we consider. Still, the variability of the noise in Figure 1 as well as the magnitude of the outbreaks are not as great as in Figure 2 that shows the average noise in frequency dynamics for words specific to different genres.

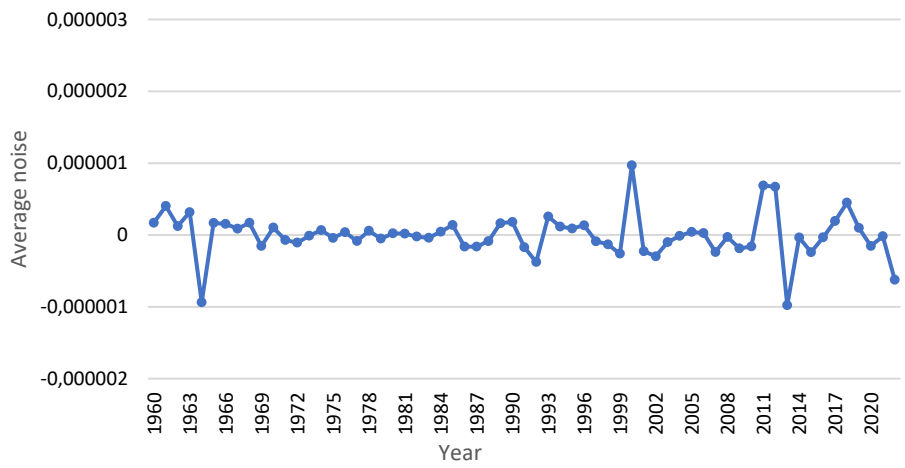

Figure 1: Average noise in frequency dynamics for out-of-genre words

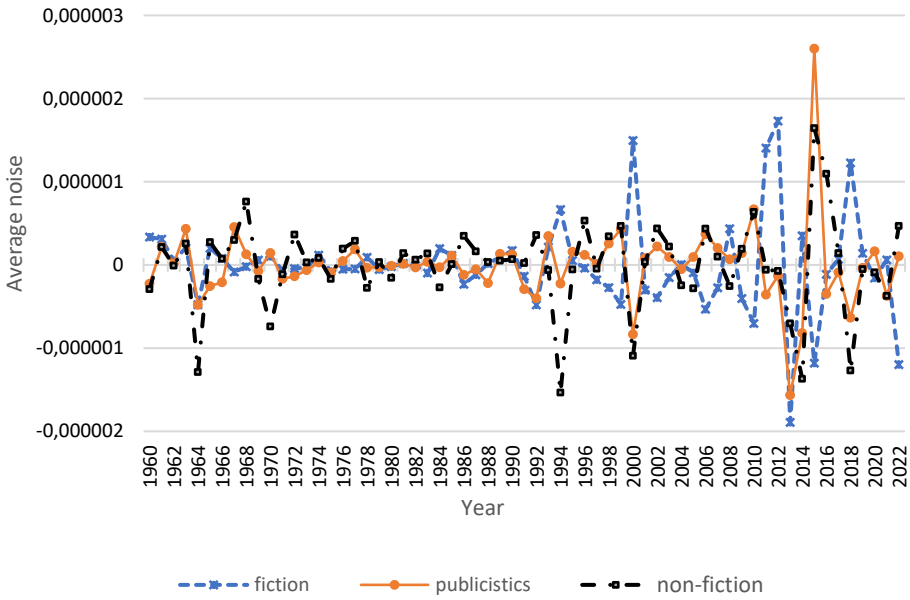

Figure 2: Average noise in frequency dynamics for words specific to different genres

The noise outliers were identified by means of the Hampel filter (the method that treats outliers as values outside the interval formed by the median value plus and minus 3 median absolute deviations). It gives the same results as the inter quartile range approach in case of 1*IQR value (25, 75 percentiles range). Table 1 gives the lists of years we identify as outliers.

Fiction		Publicistics		Non-fiction	
Outbreaks up	Outbreaks down	Outbreaks up	Outbreaks down	Outbreaks up	Outbreaks down
1994, 2000, 2008, 2011, 2012, 2018	1964, 1992, 1999, 2006, 2010, 2013, 2015	1963, 1967, 1999, 2010, 2015	1964, 2000, 2013, 2014, 2018	1968, 2015, 2016	1964, 1970, 1994, 2000, 2013, 2014, 2018

Table 1: The lists of average noise outbreaks

The first stage of our algorithm corrects the word frequency data for the period 1960–2022 according to the following formula:

$$data_{corr}^{1,year} = a \cdot data_{GBN}^{year} + b \cdot data_{SSA}^{year},$$

where

$data_{GBN}^{year}$ is the raw data taken from the GBN corpus;
$data_{SSA}^{year}$ is the value obtained by the SSA method;
a is the sum of shares of RNC word frequencies for the genres that do not need correction according to Table 1;
b is the sum of shares of RNC word frequencies for the genres that need correction according to Table 1.

For example, the correction formula for 2008 is:

$$data_{corr}^{1,2008} = (a_1 + a_2) \cdot data_{GBN}^{2008} + b_1 \cdot data_{SSA}^{2008},$$

where a_1 is the word frequency share in publicistics, a_2 is the word frequency share in non-fiction, b_1 is the word frequency share in fiction. The coefficients are like this because in 2008 the outbreak is found for fiction only.

The shares of word frequencies are calculated based on the genre subcorpora of RNC (https://ruscorpora.ru/) provided by the authors of RNC. Subcorpus frequencies for about 47 thousand of most frequent words are calculated in advance, so that the algorithm works faster. In case the word is not present in the database, the direct search in the RNC database takes place.

2.2 Trend correction

The second step of the correction algorithm is the trend correction applied to the data obtained in the first stage.

The average trends of frequencies for words specific to different genres are shown in Figure 3.

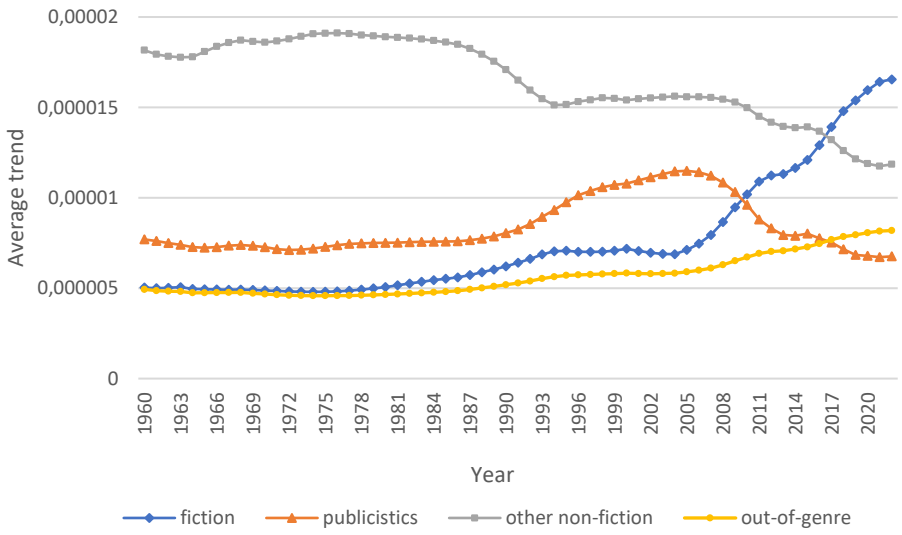

Figure 3: Average trends of frequencies for words specific to different genres

The trends are not constant. Also, we can see they compensate for one another to some extent. In the 80's of the XX century, both fiction and publicistics trends increase, whereas non-fiction trend starts to decrease. In the last decade of the XX century, the trends of publicistics and non-fiction decrease, while the fiction trend continues to vividly grow. In general, the sharpest and greatest in magnitude is the

fiction trend increase starting at the beginning of the XXI century. The difference in trend between 2000 and 2022 is almost 3 times greater than the maximum variation of other trends.

As stated in [4], it is not due to increase in the frequency of words typical for fiction in the Russian. The word frequency dynamics in the RNC shows the opposite tendencies for fiction and non-fiction as well as increase in the use of slang, vernacular, and technical terms in the fiction described in [7]. Also, over the beginning of the XXI century the share of the Russian-language books published has not increased [8].

Due to the above arguments in favor of the sharpest and unjustified change of trends in the beginning of the XXI century, we include the step of detrending in our algorithm.

To get the point where the strongest trend change starts, we divide the time period into two parts and obtain slopes of linear trends for both parts. Then we average the slopes through all the 930 fiction, 434 publicistics and 981 non-fiction words. Figure 4 presents the difference between the two slopes for each point of division. The maximum for fiction and minima for publicistics and non-fiction are at the point of 2004-2005. That is why we detrend the data in the period 2004–2022.

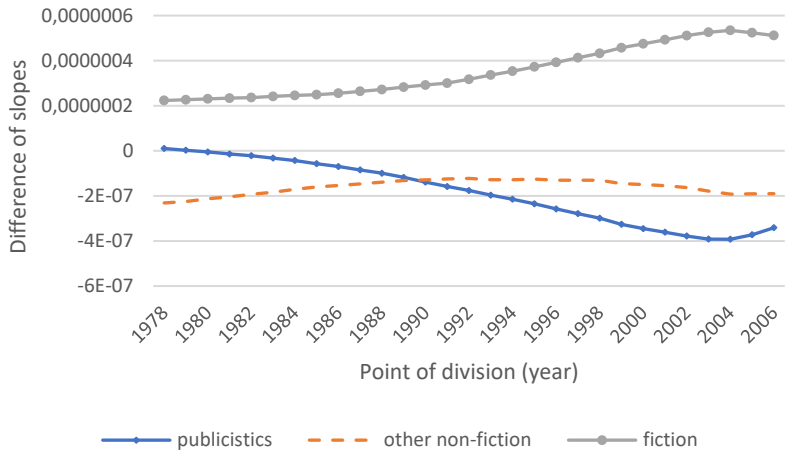

Figure 4: Average difference of slopes in the point of division

First, we partially detrend the data in the period of 2004–2022 by the SSA method leaving the baseline (the average value of the data in the period until 2004) and the trend until 2004.

$$data_{corr}^{2,year} = data_{corr}^{1,year} - data_{SSA}^{year} + baseline + data\_trend_{SSA}^{1960-2004}.$$

Here,

$data_{corr}^{1,year}$ is the result of correction at the first stage;
$data_{SSA}^{year}$ is the trend we remove;
$baseline$ is the average value of the word frequency within the period 1999–2004 (to conjoin the corrected data);
$data\_trend_{SSA}^{1960-2004}$ is the linear trend obtained by the Mann-Kendall test for the period before increase.

2.3 Specific examples of correction

The correction algorithm is implemented in Python 3. In this section, we give the illustration of its application. Frequency plots for several lexemes specific to different genres are given before and after correction. The words *комната (room)* and *подумать (to speculate)* are specific to fiction (fiction shares are 0,69 and 0,80 accordingly). The shares are calculated using the RNC. In Figure 5, we can see the smoothing of peaks after the first stage of algorithm application and partial detrending of the word frequency in the XXI century after the second stage. As a result of correction, the word frequencies increase in the XXI century, but not that much (Figure 5).

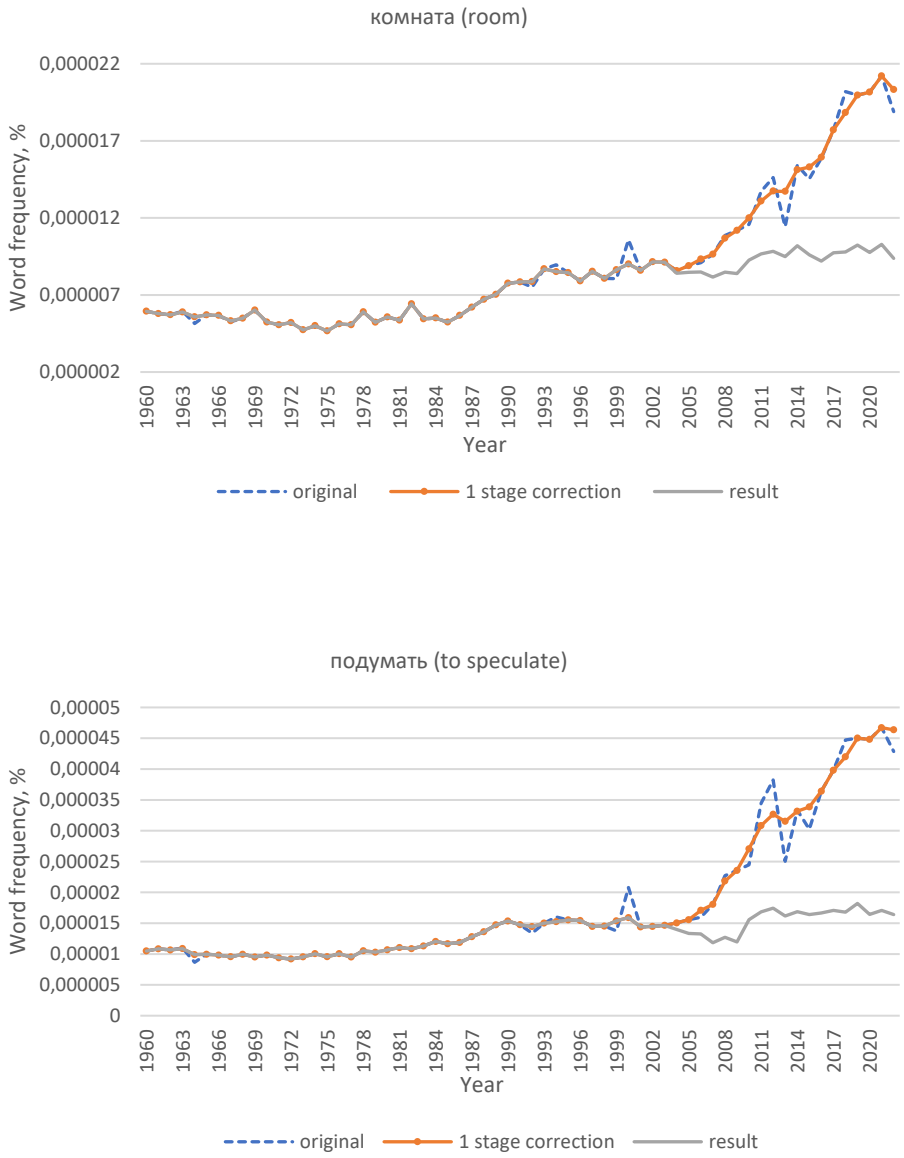

Figure 5: Correction of the words *комната (room)* and *подумать (to speculate)* specific to fiction

The word *президентский (presidential)* is specific to publicistics with the share equal to 0,72. The same publicistics coefficient is for the word *акция (shareholding)*. The sharp unnatural frequency jumps of these words as well as the general decrease are caused by the drop of the share of publicistics in the corpus. These effects are corrected (Figure 6).

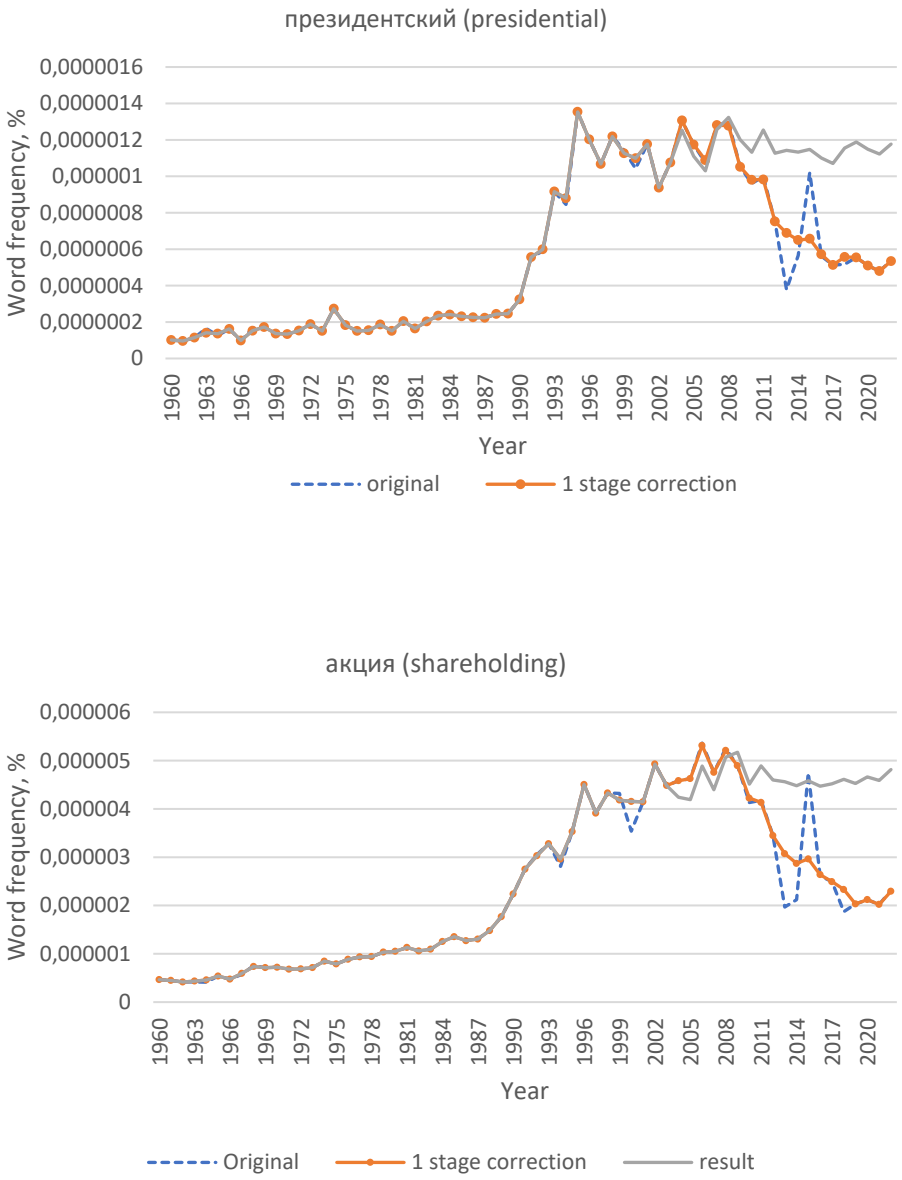

Figure 6: Correction of the words *президентский (presidential)* and *акция (shareholding)* specific to publicistics

The words *теоретический (theoretical)* and *таблица (table)* are specific to non-fiction (including scientific and academic literature) – non-fiction shares are 0,67 and 0,78 accordingly. Our algorithm removes unnatural frequency jumps and corrects the overall trend after 2004 (Figure 7).

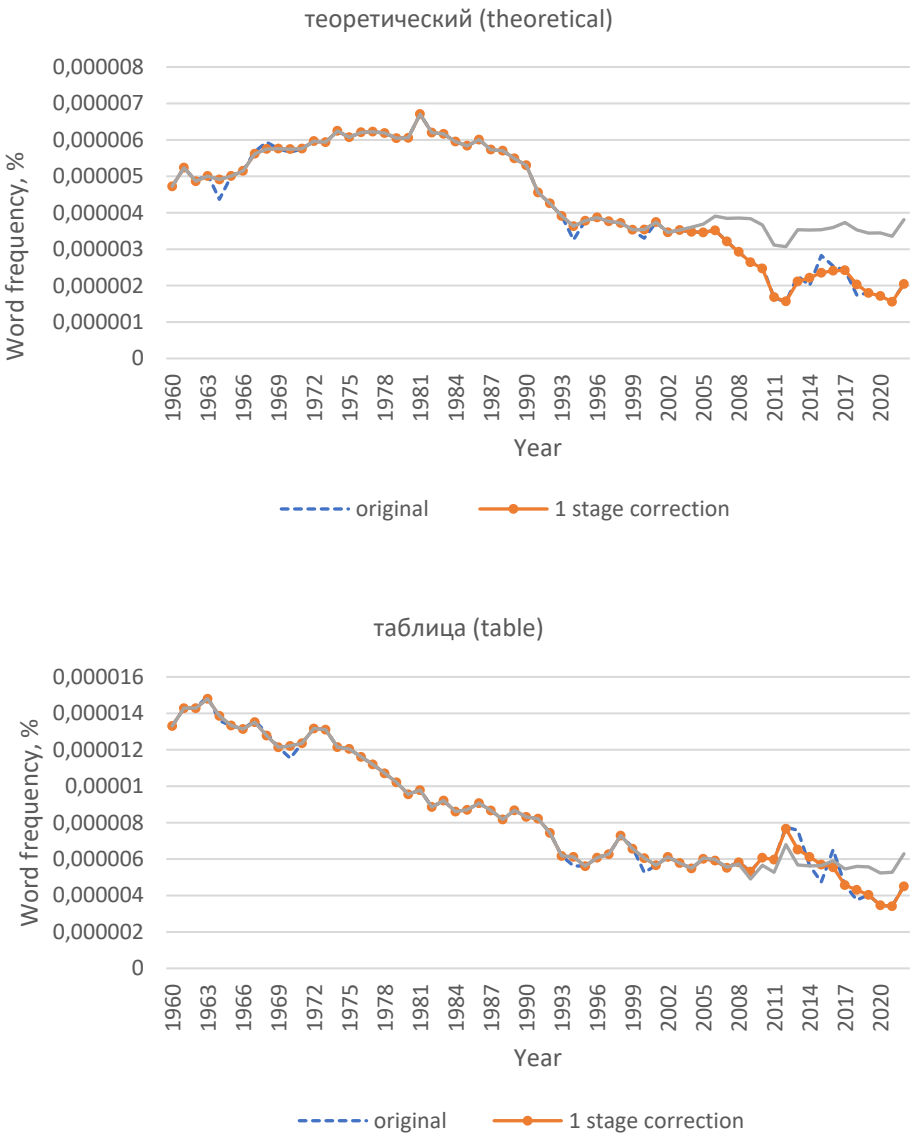

Figure 7: Correction of the words *теоретический (theoretical)* and *таблица (table)* specific to non-fiction

The algorithm can be applied to the words that are different parts of speech. No qualitative or quantitative difference in the result obtained was found. Here is an example of personal pronouns correction. The Russian pronoun *я (I)* is more typical for fiction. The fiction share for *мы (we)* is 0.30, for *я (I)* it is 0.68. According to the frequency dictionary, the pronoun *я (I)* is more frequent than *мы (we)*. Before the use of correction algorithm, the frequency of these pronouns, especially *я (I)*, was growing significantly in the XXI century: from 1999 to 2020 the frequency of the pronoun *я (I)* increased by about 3 times and the frequency of *мы (we)* increased by 1.5 times. This growth can naturally be explained by the growing share of fiction in the GBN, but it can also be due to the growth of individualism in the Russian society [3]. After our correction, as seen in Figure 8, the frequency gap between *я (I)* and *мы (we)* is significantly reduced. After eliminating the effect of the genre imbalance, the remaining increase in the frequency of the pronoun *я (I)* can be interpreted as the increase in the degree of individualism.

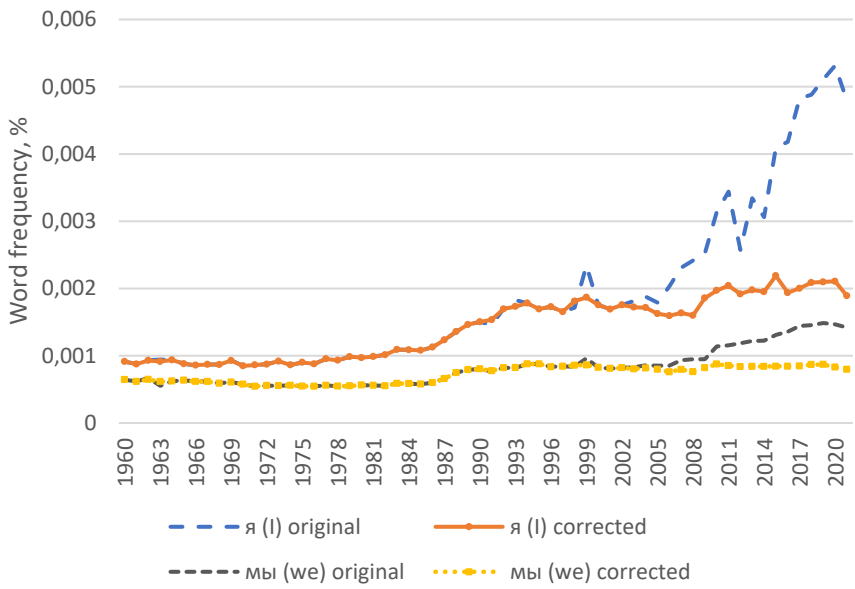

Figure 8: Correction of the pronouns *я (I)* and *мы (we)*

3 Limitations

We should note two points.

1. The main difficulty in frequency correction is unknown proportion of different genres in the GBN corpus. The above correction formulas use the data from the RNC corpus. This definite simplification – an enforcement action – is the best approximation currently available. We can only note that the resulting corrected data looks much more realistic than the original ones.

2. The discovered serious imbalance of genres is primarily in the XXI century. Previously, only individual noise peaks were identified. Therefore, the purpose of the present research is to adjust the frequencies in this particular time interval using the data from the previous period of 1960–1999 as the background. Errors in the frequencies of words before 1917 can have reasons completely different from the genre imbalance, namely, incorrect recognition of letters in the old spelling of the Russian language. Also, the problem of correction of the frequency data in the XIX century (and earlier) is complicated by a small volume of the corpus those years. Since 1960, the volume of the corpora has reached a fairly significant amount, about 1 billion words per year and it is stabilizing.

4 Conclusion

The essential genre imbalance that has appeared in the latest version of GBN in the XXI century is a significant obstacle to apply this corpus for diachronic sociological research. The reasons for this imbalance are not known. It is present not only in the Russian subcorpus, but also in the subcorpus of other languages, most vividly in British. The correction algorithm presented enables to more efficiently use the word frequency data from the Google Books Ngram corpus for research. The algorithm cannot provide results equivalent to a completely balanced corpus, but it gives a good approximation to the real word frequencies. Although in the present paper we deal only with the Russian language subcorpus, the approach proposed is language-independent and can be applied to other language corpora as well. Our correction algorithm is one of the options. Other approaches and improvements can be considered in future. In any case, the general recommendation for researchers is to take into account the discovered genre imbalance of GBN in the XXI century.

Acknowledgements

The research was funded by the Russian Science Foundation (project No. 24-18-00570, https://rscf.ru/project/24-18-00570/).

References

[1] Michel J.-B., Shen Y. K. et al. Quantitative analysis of culture using millions of digitized books // Science. — 2011. — Vol. 331(6014). — P. 176–182.

[2] Zeng R., 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. — 2015. — Vol. 50(1). — P. 47–55.

[3] Velichkovsky B.B., Solovyev V.D. et al. Transition to market economy promotes individualistic values: analysing changes in frequencies of Russian words from 1980 to 2008 // Int. J. Psychol. — 2019. — Vol. 54(1). — P. 23–32.

[4] Solovyev V. Using the Google Books Ngram Corpus to Study Social Evolution // Social Evolution & History. — 2024. — Vol. 23(2). — P. 144–164.

[5] Madsen D. Ø., Slåtten K. The possibilities and limitations of using Google Books Ngram Viewer in research on management fashions // Societies. — 2022. — Vol.12(6). — P. 171.

[6] Solovyev V.D., Bochkarev V.V. et al. Google Books Ngram: problems of representativeness and data reliability. // Proceedings of Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science. — Springer, Cham, 2020. — Vol. 1223. — P. 147–162.

[7] 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 // Digit. Scholarsh. Humanit. — 2017. — Vol. 32. — P. 169–188.

[8] Solovyev V., Ivleva A. How to Detect Imbalances in the Google Books Ngram Corpus? Proceedings of Speech and Computer. SPECOM 2024. Lecture Notes in Computer Science. — Springer, Cham, 2025. — Vol. 15300. — P.334–348.

[9] Frequency Dictionary of the Modern Russian Language homepage. http://dict.ruslang.ru/freq.php, last accessed 2024/07/13.

[10] Yu F., Peng T. et al. Cultural value shifting in pronoun use // Journal of Cross-Cultural Psychology. — 2016. — Vol. 47(2). — P. 310–316.

[11] Elsner J.B., Tsonis A.A. Singular spectrum analysis: a new tool in time series analysis. — Springer Science & Business Media, 2013.

[12] Solganik G.Ya. (2010), The modern linguistic situation and trends in the development of the Russian literary language [Sovremennaya yazykovaya situaciya i tendencii razvitiya russkogo literaturnogo yazyka], Bulletin of the Moscow University, Journalism Series 10 [Vestnik Moskovskogo universiteta. Seriya 10. Zhurnalistika], Vol. 10(5), pp. 122–134

[13] The Russian book market. The state, trends and prospects of development. https://bookunion.ru/upload/files/Bookmarket-2022.pdf?ysclid=lyk1taafsq277227991, last accessed 2024/07/13.