Computing of frequency dynamics of Russian moral vocabulary in the 20th-21st centuries: socio-psychological factors

Anna Ivleva[1][0000-0002-2670-6795] Timofey Nestik[2][0000-0002-1410-4762], and Valery Solovyev[1][0000-0002-7780-9823]

[1] Laboratory of Linguistics and AI, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru
[2] Institute of Psychology of the Russian Academy of Sciences
nestik@gmail.com

Abstract. Morality, being a system of values by which people determine the (im)propriety of thoughts and actions, is relative and depends on the specific society and time period. A large number of studies by philosophers, sociologists, psychologists, economists and lawyers are devoted to the moral state of modern Russian society. In the present paper, we investigate this subject by the methods of computer linguistics. The article analyzes the frequency dynamics of words belonging to the semantic group "morality, duty, conscience" in the Russian language during the 20th – the beginning of the 21st century. The methods of mathematical statistics, computer linguistics are applied to the Google Books Ngram corpus that contains 80 billion words of the Russian language, as well as to the Russian National Corpus. The Google Books Ngram data are first processed by a reasoned algorithm of frequency dynamics correction. The algorithm takes into account the imbalance we detected for the main literary styles – fiction, publicistisc and other non-fiction. The words from the above semantic group, which significantly change their frequency, as well as time intervals of the most significant changes in frequency trends are singled out. Subgroups of words with similar frequency dynamics are identified. We also discuss possible socio-psychological mechanisms of the dynamics of positive and negative moral lexemes in the Russian language. Some correlations of linguistic markers of the moral state of society with crisis periods, negative individual and collective experiences, and the dynamics of group identity are revealed. The methodology can be applied to different semantic groups and languages. The results confirm the correlation between the dynamics of word frequencies in the Russian language and socio-psychological processes in the society.

Keywords: moral vocabulary, large text corpora, word frequency, socio-psychological factors.

1 Introduction

Recently, large text corpora have been applied for quantitative research of the evolution of both language and various spheres of society, for example [1–3]. Based on the large language corpora [4], the methods of cognitive linguistics allow us to make a quantitative analysis of the frequency of words that are markers of various social, cultural, scientific, technical phenomena, etc.

In the present research, we study the use of positive and negative moral vocabulary in the Russian language in the 20^{th}–21^{st} centuries. The frequency of using positive and negative moral vocabulary in texts reflects the process of moral formation, as well as the development of appropriate moral norms in various fields of public relations. The latter process inevitably provokes discussions of ethical issues and morality, as well as assessment of social phenomena, both moral and immoral, in interpersonal and public discourse [5].

One of the key points of our study is to investigate the trends of moralization and dichotomous, radical thinking and assessments of behavior of people and groups of people in the Russian society in the 20^{th}–21^{st} century, especially in times of crisis [6, 7].

2 Data and methods

To analyze the moral vocabulary in the Russian language, we compiled a list of 183 nouns related to the moral theme. Only lemmas (the basic word forms) were taken into account. The sources of the lemmas are:

- Semantic group “morality, duty, conscience” from the Semantic dictionary (by Shvedova) [8].
- The closest neighbors of the words “morality”, “duty”, “conscience” in the FastText model we trained (using Gensim library in Python 3) on the data from the Russian subcorpus of the Google Books Ngram (2-, 3-, 4-, 5-grams). We took the data for the 3 latest decades (since 1990) and trained the model for 80 epochs. The closest neighbors were chosen based on the cosine similarity.
- Manual selection from articles on the topics corresponding the theme of morality.

By means of publicly available sentiment vocabularies KFU_Bert (https://kpfu.ru/tehnologiya-sozdaniya-semanticheskih-elektronnyh.html) and KartaSlovSent [9], we automatically labelled the lemmas as negative (82) and positive ones (101). For the lemmas either not present in any vocabulary or having neutral mark, manual labels were given.

We use the word frequency data for the 1920-2022 time period from the Google Books Ngram (GBN, https://books.google.com/ngrams) corpus and the Russian National Corpus (RNC, https://ruscorpora.ru/en) provided by its authors. We do not take into account the earlier periods due to a number of errors of the old Russian spelling processing before the language reform in 1918.

The GBN is created by an overall scanning of books from the biggest world university libraries. It is a diachronic corpus presenting uni-, 2-, 3-, 4-, 5-grams from the books for more than five hundred years. The Russian subcorpus contains more than 80 billion words. This allows to make statistically confident research of the evolution of various language and societal phenomena. One can find a detailed description of the corpus in [10, 11].

The Russian National Corpus is a representative collection of texts in Russian, counting more than 2 billions of tokens having linguistic annotation and search tools. We used a part of the general corpus containing about 300 millions of tokens.

In [12], we made a detailed analysis of the genre shares of the Russian subcorpus of the GBN corpus based on the frequency dynamics of 7420 lemmas from the Frequency Vocabulary of Fiction, Frequency Vocabulary of Publicistics and Frequency Vocabulary of Other Non-fiction [13]. It showed significant imbalance of genres with highly synchronous unjustified trends and outbreaks. In [14], we proposed an algorithm of word frequency correction to compensate for the previously found imbalances and gave some examples of its application. The algorithm is based on correction of excessive noise of the frequency series for the words specific to 3 main literary styles: fiction, publicistics and non-fiction. Also, it includes detrending after 2004. In the present research, we apply this algorithm to the GBN data.

3 Results

The frequencies obtained in GBN for the period 1920–2022 were normalized using the Euclidean norm (L2 normalization), and then averaged. Figure 1 shows the resulting relative frequency for the two groups of words – positive and negative.

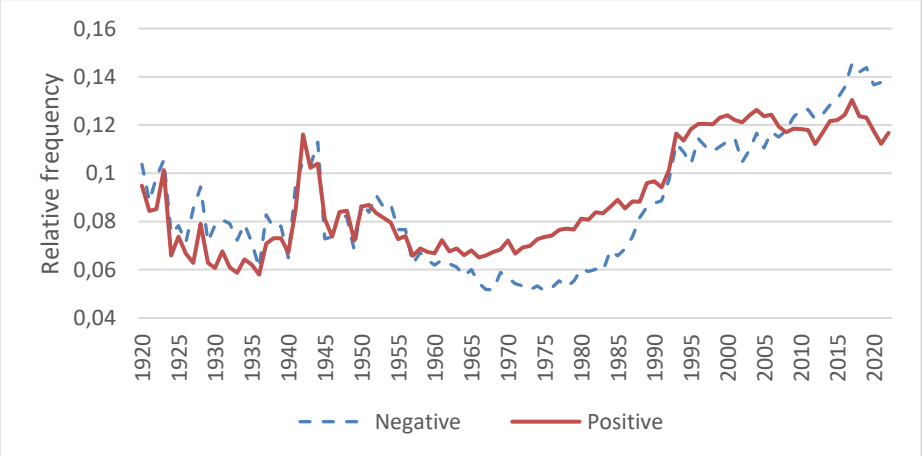

Fig. 1. Graphs of the average frequencies of moral vocabulary (GBN).

The following time periods attract attention:

1. The period of the Great Patriotic War.

This period is characterized by a rapid and sharp increase in moral vocabulary from 1940 to 1942, a plateau in 1943 and an equal sharp decline to the previous level in the period 1944–1946.

10 negative moral words with the greatest increase in frequency observed from 1940 to 1942 are: человеконенавистничество (misanthropy), сварливость (protervity), невоспитанность (discourtesy), женоненавистничество (misogyny), вероломство (treachery), злодеяние (villainy), вандализм (vandalism), лгун (liar), заносчивость (arrogance), измена (treason).

10 positive words with the greatest increase in frequency from 1940 to 1942 are: бесстрашие (fearlessness), Родина (Motherland), мужество (courage), бдительность (vigilance), храбрость (bravery), самоотверженность (dedication), патриотизм (patriotism), подвиг (feat), дружба (friendship), гуманность (humaneness).

10 negative words with the greatest frequency decline from 1944 to 1946 are: вандализм (vandalism), хвастовство (boasting), неправдивость (untruthfulness), сварливость (protervity), сквернословие (ribaldry), злословие (backbiting), невоспитанность (discourtesy), безжалостность (ruthlessness), вина (guilt), небрежность (carelessness).

10 positive words with the greatest frequency decline from 1944 to 1946 are: извинение (apology), пожертвование (donation), сердечность (cordiality), самоотверженность (dedication), бесстрашие (fearlessness), обязательство (commitment), храбрость (bravery), добросердечие (kindhearted), благожелательность (benevolence), мужество (courage).

2. 1966–1988 – the period of maximum predominance of positive moral vocabulary over negative.
3. 1991–1993 – simultaneous rapid growth of both positive and negative moral vocabulary.

10 negative words with the greatest frequency increase observed from 1991 to 1993 are: женоненавистничество (misogyny), мужененавистничество (man-hating), сквернословие (ribaldry), безбожие (godlessness), своеволие (headstrongness), безрассудность (recklessness), безжалостность (ruthlessness), праздность (idleness), грех (sin), сварливость (protervity).

10 positive words for which the greatest increase in frequency was observed from 1991 to 1993: дипломатичность (diplomacy), духовность (spirituality), этика (ethics), кротость (meekness), старание (endeavours), целомудрие (chastity), человеколюбие (benevolence), пожертвование (donation), альтруизм (altruism), заповедь (commandment).

4. Since 2017, there has been a slight decrease in the frequency of positive moral vocabulary.
5. Since 2008, there has been a unique period when the average frequency of positive moral vocabulary is less than the negative one.

This trend continues to increase up to 2022. This dynamics is slightly similar to the period of 1929–1935.

Negative words with the maximum increase since 2008 are: срам (shame), праздность (idleness), разврат (debauch), цинизм (cynicism), зло (evil), измена (treason), деспот (despot), сквернословие (profanity), несправедливость (injustice), проступок (misconduct). Positive words with the maximum decrease in frequency since 2008 are: порядочность (decency), преданность (devotion), душевность (warm-heartedness), добродушие (amiability), великодушие (generosity), обязательство (commitment), друг (friend), кодекс (code), тактичность (tact), благожелательность (benevolence).

The RNC data on the frequency of the words considered confirm that in the period of 1960–1990, the difference between the frequencies of positive and negative vocabulary is greater than in other decades, the peak being in the 70's. In the 90's, the difference decreases. The discrepancy of results for the two corpora is found after 2008. Though, the Publicistics texts from the RNC give more concordant results with the GBN (almost zero increase of the difference). Figure 2 shows the difference of average relative (L2 normalized) frequencies of positive and negative moral vocabulary.

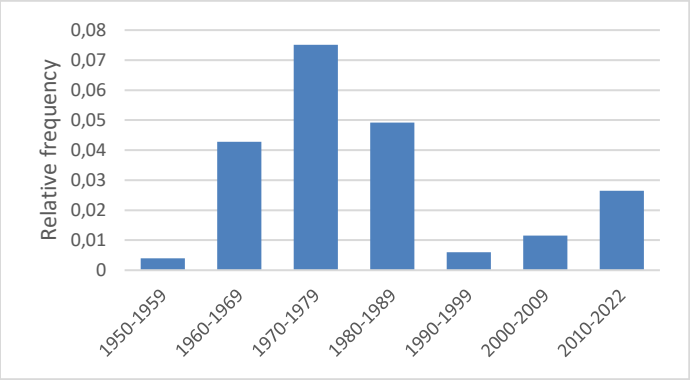

Fig. 2. The difference of average frequencies of positive and negative moral vocabulary (RNC).

We also compute the correlations of frequency dynamics of the positive and negative moral vocabulary with the dynamics of the marker words of absolutist vocabulary (все (everyone), никто (no one), никогда (never), всегда (always), должны ((they) have to)) [15]. The results are shown in Table 1. The average frequency dynamics of negative moral vocabulary correlates with the frequency dynamics of all marker words of absolutism greater than of the positive one.

Table 1. Correlations of frequency dynamics of moral vocabulary and marker words

	averaged negative	averaged positive	все (everyone)	никто (nobody)	никогда (never)	всегда (always)|	должны ((they) must)
averaged negative	1	-	-	-	-	-	-
averaged positive	0.883	1	-	-	-	-	-
все (everyone)	0.899	0.691	1	-	-	-	-
никто (nobody)	0.904	0.768	0.967	1	-	-	-
никогда (never)	0.911	0.737	0.988	0.986	1	-	-
всегда (always)|	0.889	0.694	0.984	0.953	0.981	1	-
должны ((they) must)	0.346	-0.021	0.487	0.321	0.438	0.516	1

4 Discussions

Here, we will give some explanations of the result obtained from the position of socio-psychological processes.

Some predominance of positive moral vocabulary in the Russian-language texts in the period 1960–1990 can be explained by the active development of the USSR at that time. A large number of big and small cities were being built, the military industry was actively developing, the Soviet Union began to explore space and became a leader in this field. The country also achieved significant success in sports, culture, literature, art and well-being of citizens.

The increase in the use of moral vocabulary in the 90's can be a result of value conflicts and anomie in the context of radical social changes in income levels and lifestyle of large social groups. This provoked the use of “crisis discourse” characterized by assessment and hyperbolization [16]. Secondly, the destruction of the Soviet system and transition to state capitalism exacerbated the problems of social inequality and sensitivity to injustice, which provoked a moral assessment of what was happening.

In the conditions of uncertainty, polarization, and a series of economic crises, there is usually an increase in anxiety-depressive symptoms usually accompanied by the use of absolutist vocabulary [17].

5 Conclusions

We apply quantitative research to compute the evolution of moral vocabulary frequency in the Russian language in the 20^{th} –21^{st} centuries. For this, we study word frequencies

of the nouns typical for semantic group “morality, duty, conscience”. The analysis is based on two large text corpora: the Google Books Ngram and the Russian National corpus. Before processing the Google Books Ngram data, we apply our algorithm of correction to compensate for some outliers and unjusified trends due to the genre imbalances in the Corpus. The research allows us to get some patterns and explain them from the socio-psychological point of view. We also give the lists of negative and positive moral words which most vividly demonstrate the patterns. We outline the trends of moralization, more radical and dichotomous assessment and thinking in times of crisis and some predominance of positive moral vocabulary in times of relative stability and prosperity of the society.

6 Acknowledgements

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

References

1. Pellert, M., Lechner, C.M., Wagner, C., Rammstedt, B., Strohmaier, M.: AI Psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. Perspectives on Psychological Science 19(5), 808–826 (2024).
2. Charlesworth, T.E.S., Yang, V., Mann, T.C., Kurdi, B., Banaji, M.R.: Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. Psychological Science 32(2), 218–240 (2021).
3. Bochkarev, V.V., Solovyev, V.D., Nestik, T.A. et al.: Variations in average word valence of Russian books over a century and social change. J Math Sci 285, 14–27 (2024).
4. Juola, P.: Google Books Ngrams. In: Schintler, L.A., McNeely, C.L. (eds) Encyclopedia of Big Data. Springer, Cham. 517–521 (2022).
5. Kozlovskaya, E., Kobylko, J., Medvedev, Y.: Sense-forming function of context in publicistic texts. Russian Journal of Linguistics 23, 165–184 (2019).
6. Riek, B.M., Mania, E.W., Gaertner, S.L. Intergroup threat and outgroup attitudes: A meta-analytic review. Personality and Social Psychology Review 10(4), 336–353 (2006).
7. N. Yu. Shvedova et al.: Russian semantic dictionary: An explanatory dictionary systematized by classes of words and meanings. Azbukovnik, IRYA RAS, Moscow (2014).
8. Bochkarev, V.V., Achkeev, A.A., Savinkov, A.V., Shevlyakova, A.V., Solovyev, V.D.: Large sentiment dictionary of Russian words. In: Calvo, H., Martínez-Villaseñor, L., Ponce, H. (eds) Advances in Soft Computing. MICAI 2023. LNCS, vol. 14392. Springer, Cham. (2024).
9. Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K.: The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. Erez Lieberman A.: Quantitative analysis of culture using millions of digitized books. Science 331(6014), 176–182 (2011).
10. Lin, Y., Michel, J. B., Lieberman, E. A., Orwant, J., Brockman, W., Petrov, S.: Syntactic annotations for the Google Books Ngram corpus. In: Proceedings of the ACL 2012 system demonstrations, pp. 169–174. Association for Computational Linguistics, Korea (2012).

11. Madsen, D. Ø., Slåtten, K.: The possibilities and limitations of using Google Books Ngram Viewer in research on management fashions. Societies 12(6), 171 (2022).
12. Solovyev, V., Ivleva, A. How to Detect Imbalances in the Google Books Ngram Corpus? In: Karpov, A., Delić, V. (eds) Speech and Computer. SPECOM 2024. LNCS, vol. 15300, pp 334–348. Springer, Cham. (2025).
13. Frequency Dictionary of the Modern Russian Language homepage, http://dict.ruslang.ru/freq.php, last accessed 2024/07/13.
14. Solovyev, V., Ivleva, A. (2025). An Algorithm for Genre Imbalance Correction in the Russian Subcorpus of the Last Version of the Google Books Ngram Corpus. In: Conference Proceedings “Computational Linguistics and Intellectual Technologies” (2025). (In press)
15. Pil'gun E. V. Semantika i pragmatika krizisnogo diskursa. Minsk: IVC Minfina, (2020). (In Russian)
16. Adam-Troian, J., Bonetto, E., Arciszewski, T.: Using absolutist word frequency from online searches to measure population mental health dynamics. Sci Rep 12, 2619 (2022).