Natural Language Processing Centre
Faculty of Informatics
Masaryk University
Brno

Slavonic Natural Language Processing in the 21st Century

Aleš Horák, Klára Osolsobě, Adam Rambousek, Pavel Rychlý

Published by Tribun EU
Brno 2019

Table of Contents

Diachronic Corpora as Research Tool in Humanities

Victor Zakharov[1], Andrei Masevich[2]

[1] Saint-Petersburg State University
Universitetskaya emb. 7-9, 199034 Saint-Petersburg, Russia
v.zakharov@spbu.ru
[2] Saint-Petersburg State Institute of Culture
Dvortsovaya emb., 2, 191186 St.-Petersburg, Russia
andmasev@mail.ru

Abstract The paper presents results of diachronic research of political terms on the base of Google Books Ngram Viewer. We studied fluctuations of frequency of lexical units in texts of Russian books in the period from 1920 to 2000. The results show connection of frequency fluctuation of words with historical events, which allows various interpretations. This can be of interest both for linguistic and for culturological and historical studies.

1 Introduction

The Information technologies and corpus linguistics give brand new instruments of the diachronic research of languages. In particular, the time behaviour of a lexical item can be observed, i.e. the frequency of its use in written language. Among many types of corpora, diachronic (or historical) corpora can be mentioned. They are corpora of texts created during rather long historical periods of language development. They help to do diachronic researches, i.e. to trace behavior of language units, constructions and phenomena within historical periods. Among these corpora, a special place belongs to Ngram Viewer diachronic corpus based on the Google Books library[3] [1]. Now it seems to be the most powerful tool for diachronic research. This system contains large corpora of annotated book texts in 8 languages which covers a few centuries.

The corpus creates charts of time dynamics of a word frequency, which allows reveal possible connections between frequency behavior of words and cultural trends of the society.

Grammar can also be studied with the help of Google Books Ngram. For example, V. Solovyov [2] used it for studying the aspectual system of the Russian language.

The language is known to be a dynamic system which is changing throughout all its history, during time lapses of different length on all levels, and these

[3] https://books.google.com/ngrams

A. Horák, K. Osolsobě, A. Rambousek, P. Rychlý (eds.): Slavonic Natural Language Processing in the 21st Century, pp. 226–238. Tribun EU, Brno, 2019.

changes are based on factors of different origin, mostly psychological, social and cultural ones. Detection of these factors through texts and products of communicative correspondence is the standard research method in social studies, known as content analysis.

Diachronic research of the language lets detect the factors and laws of both linguistic and historical and cultural significance. In recent years, a new branch of scientific research called culturomics (or quantitative culturology) has appeared. The vocabulary at dictionary.com website defines it as "the study of human culture and cultural trends over time by means of quantitative analysis of words and phrases in a very large corpus of digitized texts"[4].

Nowadays, there are already some Russian and foreign publications on the methods and tools of diachronic research on the base of Google Books Ngram Viewer [3-9].

2 Materials and tools for the research

Google Books Ngram Viewer service has been available on Internet since 2010. It includes corpora of 8 languages: English, German, French, Spanish, Italian, Russian, Hebrew, Chinese, though there are much more existing corpora in it. For example, there are nine corpora only for the English language (English, American English, British English, English Fiction, English One Million, etc.). The volume of corpora is unprecedented for corpus linguistics. For example. the volume of the Russian corpus is 591,300 texts (books) forming the corpus of more than 67 billion tokens. The lapse of time is from the 18th century till 2008.

Google Books Ngram Viewer is a corpus of books. Each book file is supplied by metadata together with partial grammar tagging. The system provides the search for required ngrams (1 to 5 words) and builds the chart of their frequency down the years with the lapse of time defined by user. The horizontal axis of the chart shows the years within the required time lapse, while the vertical axis contains the percentage of the appropriate ngram in the observed year. The relative frequency of a ngram during the year is counted as the quotient of the quantity of its usage during the year and the total quantity of the word usages in the corpus during the same year. Curves for different language corpora can be shown on one total chart. The system has highly developed search engine and the opportunities for data presentation[5].

When building the charts of ngram usage frequency, the so-called smoothing is used. In the case of zero smoothing, the relative frequency of any ngram for every year is being taken into account. But the real trend in the dynamics of word frequency can be seen more clearly in the case of the moving data smoothing (year range).

[4] https://www.dictionary.com/browse/culturomics

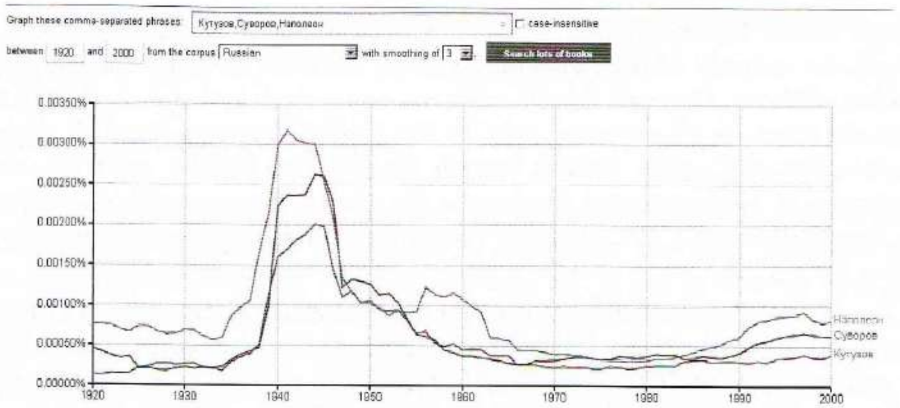
[5] see https://books.google.com/ngrams/info

3 Experimental Research

Our task is the comparative diachronic research of the political vocabulary of the mid-20th century to show the reflection of social and political events in texts.

3.1 Analysis of the use of proper nouns in Russian books published during the Great Patriotic War

During the Great Patriotic War, especially in its beginning, a lot of books on military history have been published. This can be proved by the analysis of the use of proper nouns. The famous Russian military leaders Alexander Suvorov and Michael Kutuzov have been often mentioned as heroes of past-time wars (Fig. 1), which is obvious. But the further analysis shows the facts which are not so obvious without the corpus data. The frequency of the name of Napoleon usage being mentioned is even higher than the frequency of the names of Russian warlords. No comments are needed, except the famous idiom "the cult of Napoleon in Russia".



Graph these comma-separated phrases: Кутузов,Суворов,Наполеон □ case-insensitive

between 1920 and 2000 from the corpus Russian with smoothing of 3.

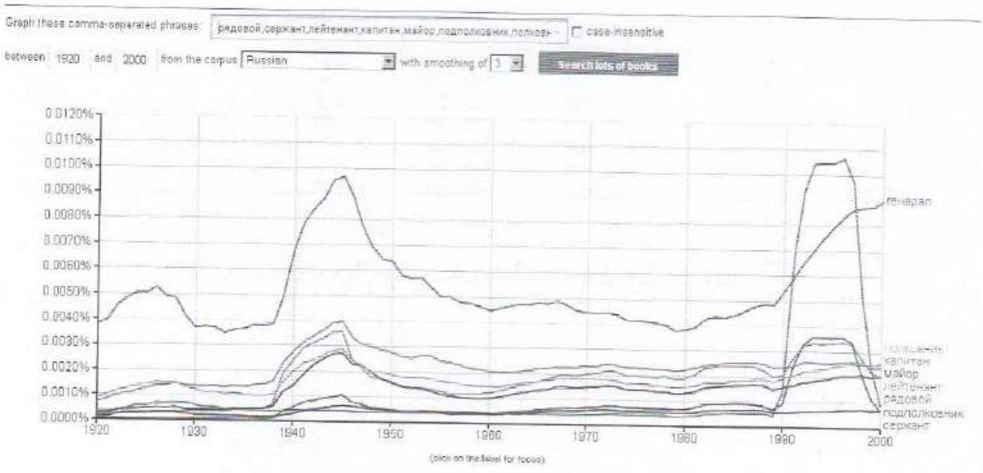
Figure 1: Frequency chart for for the names Napoleon, Suvorov, Kutuzov

3.2 Analysis of words denoting military ranks

The next chart (Fig. 2) reflects the change of frequency of words denoting military rank.

It can be seen that the word "general" (генерал' in Russian) is the most frequent of all, with curves for officer ranks following and the ranks for private corps being the less frequent of all.

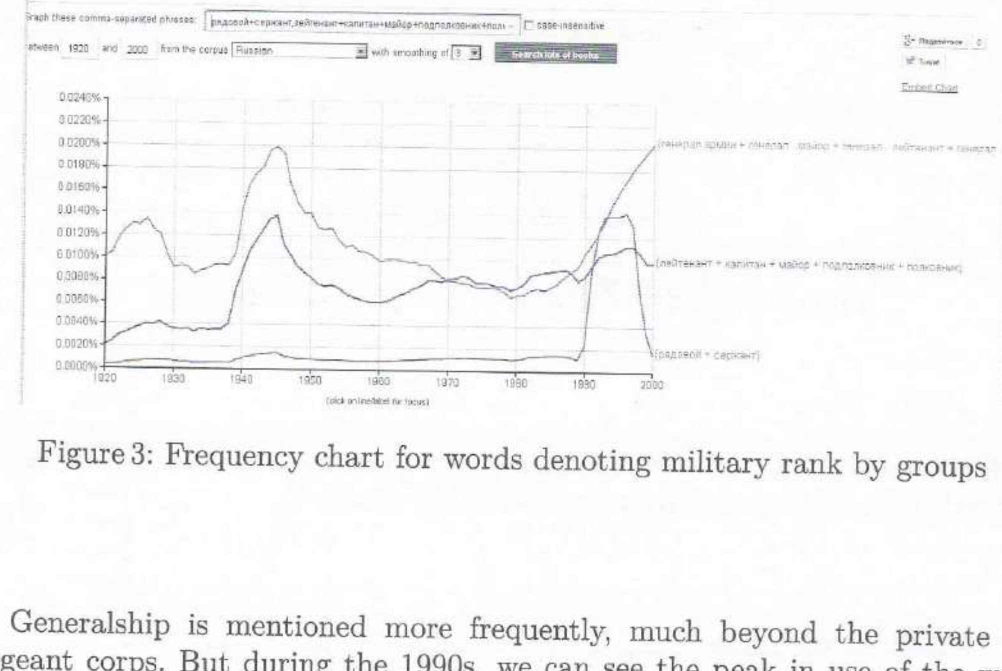
The graphic would become more obvious if we use the function of curve accumulation separately for generalship, officer ranks and ranks for private corps (Fig. 3).



Graph these comma-separated phrases: рядовой,сержант,лейтенант,капитан,майор,подполковник,полковн □ case-insensitive

between 1920 and 2000 from the corpus Russian with smoothing of 3 Search lots of books

Figure 2: Frequency chart for words denoting military rank



Graph these comma-separated phrases: рядовой+сержант,лейтенант+капитан+майор+подполковник+полк □ case-insensitive

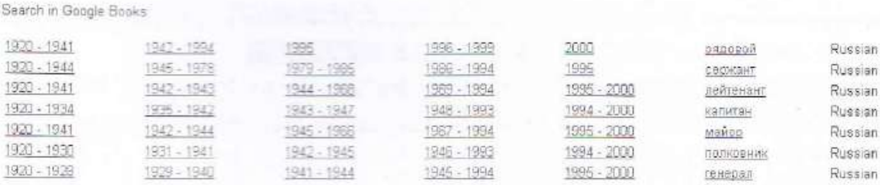
between 1920 and 2000 from the corpus Russian with smoothing of 3 Search lots of books

Figure 3: Frequency chart for words denoting military rank by groups

Generalship is mentioned more frequently, much beyond the private and sergeant corps. But during the 1990s, we can see the peak in use of the words "private" (**рядовой**) and "sergeant" (**сержант**), which is shown at Fig. 3. At first it seems incomprehensible. If we search in the Google Books database according links (Fig. 4) and observe texts from the corpus (Fig. 5), we can see the reason for that.

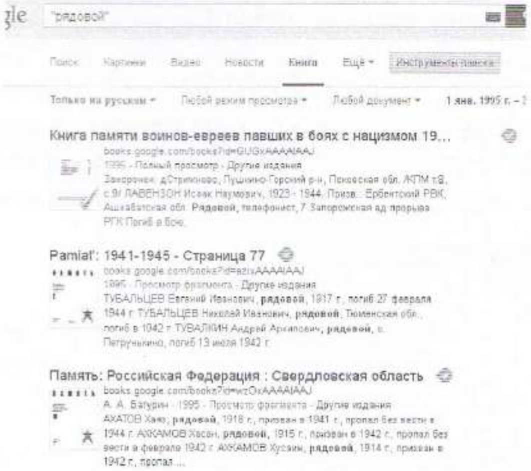
If we actualize the links to the 1995 books in the row "**рядовой**" (private), we get the following picture.

In the results of search for the query "**рядовой**" (private) for 1995 we see so called "**Книги памяти**" (memorial books), i.e. lists of those who died or went missing during the Great Patriotic War, which had been published a lot in mid-1990s according to the 50th anniversary of the victory over the Nazi Germany. Of course, most of those who had died are privates and sergeants.


Search in Google Books

1920 - 1941	1942 - 1994	1995	1996 - 1999	2000	рядовой	Russian
1920 - 1944	1945 - 1978	1979 - 1985	1986 - 1994	1995	сержант	Russian
1920 - 1941	1942 - 1943	1944 - 1968	1969 - 1994	1995 - 2000	лейтенант	Russian
1920 - 1934	1935 - 1942	1943 - 1947	1948 - 1993	1994 - 2000	капитан	Russian
1920 - 1941	1942 - 1944	1945 - 1966	1967 - 1994	1995 - 2000	майор	Russian
1920 - 1930	1931 - 1941	1942 - 1945	1946 - 1993	1994 - 2000	полковник	Russian
1920 - 1928	1929 - 1940	1941 - 1944	1945 - 1994	1995 - 2000	генерал	Russian

Figure 4: Links to the Google Books database



"рядовой"

Книга памяти воинов-евреев павших в боях с нацизмом 19...
books.google.com/books?id=GUGxAAAAIAAJ
1995 - Полный просмотр - Другие издания
Захоронен: д.Стриково, Пушкино-Горский р-н, Псковская обл. АПМ т.8, с.9/ ЛАВЕНЗОН Исаак Наумович, 1923 - 1944. Призв.: Ербентский РВК, Ашхабатская обл. **Рядовой**, телефонист, 7 Запорожская ад прорыва РГК Погиб в бою.

Pamiat': 1941-1945 - Страница 77
books.google.com/books?id=aztxAAAAIAAJ
1995 - Просмотр фрагмента - Другие издания
ТУБАЛЬЦЕВ Евгений Иванович, **рядовой**, 1917 г., погиб 27 февраля 1944 г. ТУБАЛЬЦЕВ Николай Иванович, **рядовой**, Тюменская обл., погиб в 1942 г. ТУВАЛКИН Андрей Архипович, **рядовой**, с. Петрушкино, погиб 13 июля 1942 г.

Память: Российская Федерация : Свердловская область
books.google.com/books?id=wzOxAAAAIAAJ
А. А. Батурин - 1995 - Просмотр фрагмента - Другие издания
АХАТОВ Хаяз, **рядовой**, 1918 г., призван в 1941 г., пропал без вести в 1944 г. АХКАМОВ Хасан, **рядовой**, 1915 г., призван в 1942 г., пропал без вести в феврале 1942 г. АХКАМОВ Хусаин, **рядовой**, 1914 г., призван в 1942 г., пропал ...

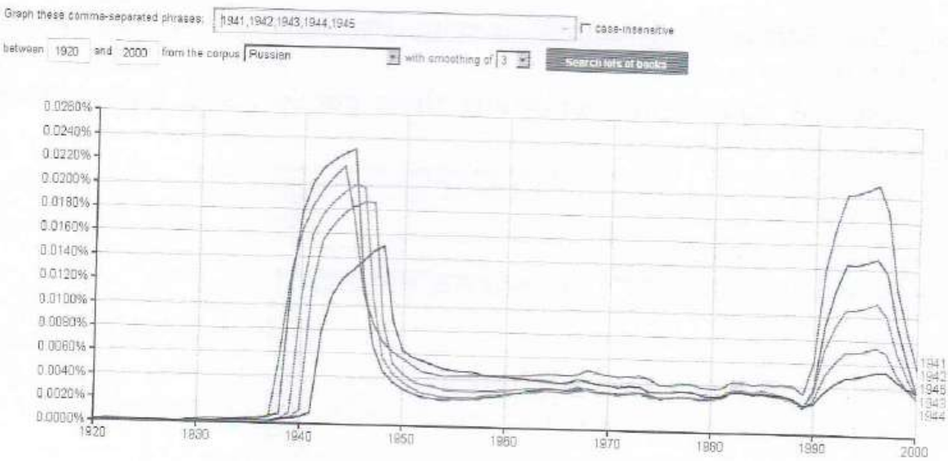
Figure 5: Results of search in Google Books database

At the same time, the sharp increase of the frequency of the word "генерал" (general) (Fig. 3) in post-Soviet times is still a puzzle.

3.3 Analysis of dates

Other information can also be received from the corpus. Each article from memorial books contains standard data on war victims, including the date of death. If we build charts according to war dates, we will get the following picture (Fig. 6).

Curves at Fig. 6 form two peaks. The first one falls at war years which can be easily explained. The frequency peaks for any certain year belong to the dates coming two or three years later. It is worth mentioning that we talk of books published later than the events described in them. The second peak falls to the mid-1990s, which can be explained by memorial books again. The chart shows obviously the annual reduce of the number of war victims. In 1943, the number of the dead was half the 1941 number.

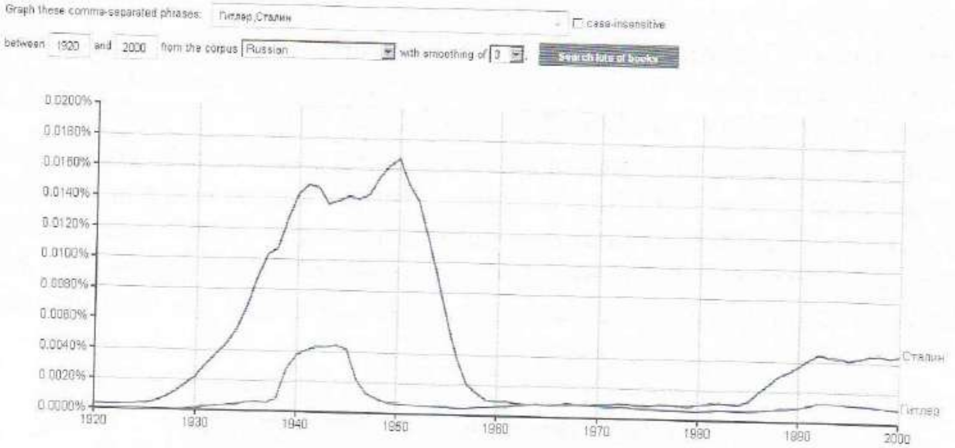

Graph these comma-separated phrases: 1941,1942,1943,1944,1945 □ case-insensitive

between 1920 and 2000 from the corpus Russian with smoothing of 3

Figure 6: Chart of war dates frequency in the corpus

3.4 Frequency of use of politicians' names

The chart with names of the two war protagonists, Joseph Stalin and Adolf Hitler, in Russian books is shown below (Fig. 7).


Graph these comma-separated phrases: Гитлер,Сталин □ case-insensitive

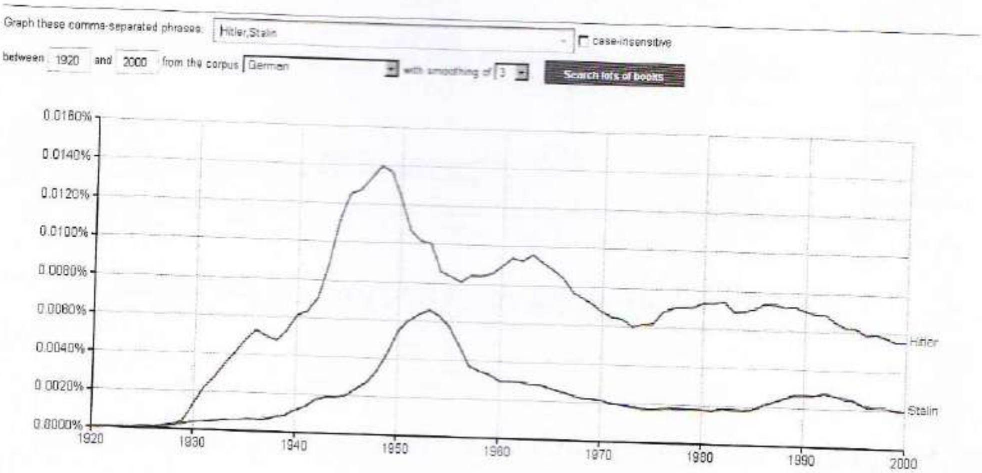
between 1920 and 2000 from the corpus Russian with smoothing of 3

Figure 7: Frequency dynamics of mentioning Stalin and Hitler in Russian books

The curve reflecting the frequency of the word "Stalin" can be easily explained. This curve reached its peak at the beginning of the war and grew until 1950, then dropped sharply, keeping the level until mid-1980s, then the frequency grew, probably with negative connotations.

Hitler's name has been appearing in Russian texts since early 1930s. During the war, the frequency of mentioning for this personality grew sharply, then fell after the war and was stable until the end of the observed period. It is worth noticing that both curves have been keeping the similar level from early 1960s till mid-1980s.

Let's see the quantity of mentioning these politicians in books written in German (Fig. 8).

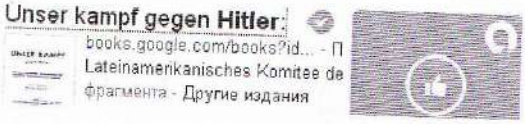

Graph these comma-separated phrases: Hitler,Stalin □ case-insensitive

between 1920 and 2000 from the corpus German with smoothing of 3

Figure 8: Frequency dynamics of mentioning Stalin and Hitler in German books

At first sight, the model of word behaviour is similar to that at Fig. 7. The country's leader is mentioned much more frequently in wartime books than the leader of the enemy state. But it is worth noticing that the peak of frequency for the proper noun "Hitler" in German books falls to 1948.

During the wartime the name of Hitler has been mentioned in positive connotations in books published in Nazi Germany. But the corpus also contains texts of books published in different German-speaking countries. There were also books of German refugees which has also been included into the database. So we can see the links to publications similar to one at Fig. 9 within the German corpus.


Unser kampf gegen Hitler:

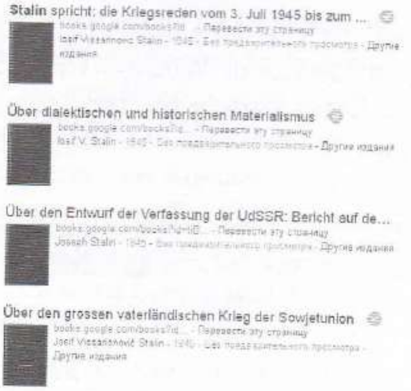
books.google.com/books?id... - П

Lateinamerikanisches Komitee de

фрагмента - Другие издания

Figure 9: The link to the publication of the Latin American Committee of Free Germans (1944)

One can see on Fig. 8 that in late 1940s and early 1950s, the frequency of mentioning Stalin in German books grew and exceeds even the level during the years of the war. How can it be explained? Fig. 10 shows that it occurred due to publications from the Soviet occupation zone and then from East Germany during the period when a lot of Stalin's works and speeches had been translated to German and published.



Stalin spricht: die Kriegsreden vom 3. Juli 1945 bis zum ...
books.google.com/books?id ... - Перевести эту страницу
Iosif Vissarionovič Stalin - 1945 - Без предварительного просмотра - Другие издания

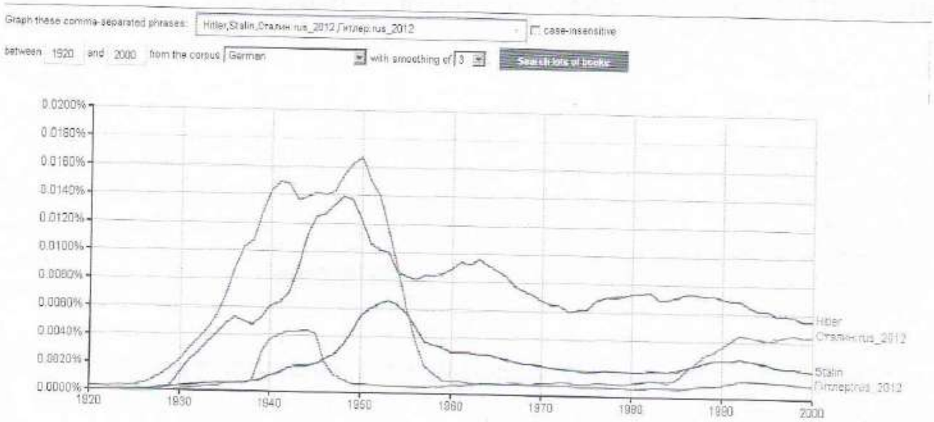
Über dialektischen und historischen Materialismus
books.google.com/books?id ... - Перевести эту страницу
Iosif V. Stalin - 1945 - Без предварительного просмотра - Другие издания

Über den Entwurf der Verfassung der UdSSR: Bericht auf de...
books.google.com/books?id=HB... - Перевести эту страницу
Joseph Stalin - 1945 - Без предварительного просмотра - Другие издания

Über den grossen vaterländischen Krieg der Sowjetunion
books.google.com/books?id ... - Перевести эту страницу
Josif Vissarionović Stalin - 1946 - Без предварительного просмотра - Другие издания

Figure 10: Links to German books after 1945 mentioning Stalin

As it was said we can show line plots from different corpora on the same chart. Let's compare the behaviour of names of both Hitler and Stalin in German and Russian corpora (four curves in Fig. 11).



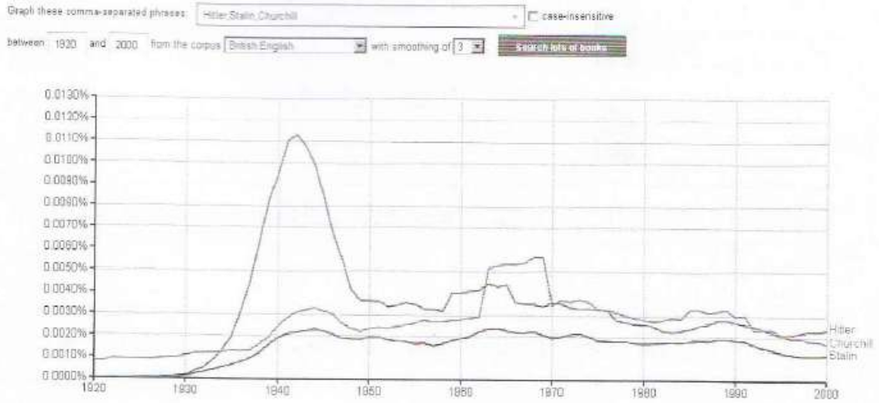
Graph these comma-separated phrases: Hitler,Stalin,Сталин:rus_2012,Гитлер:rus_2012 □ case-insensitive

between 1920 and 2000 from the corpus German with smoothing of 3

Figure 11: Frequency dynamics of mentioning Stalin and Hitler in Russian and German corpora

It can be seen that the level of mentioning Stalin in the Russian corpus is a bit higher than the level of mentioning Hitler in the German one. At the same time, the drop of frequency of mentioning Hitler in German books is not so sharp as it is for Stalin in Russian ones. Let's offer our interpretation of this fact. The chart reflects the difference in understanding the historical experience in Germany and the USSR. The Soviet Union had the trend to understatement of Stalin's personality, while in Germany, the negative historical experience had been studied precisely and the results of the studies had been published for long years after Hitler's death.

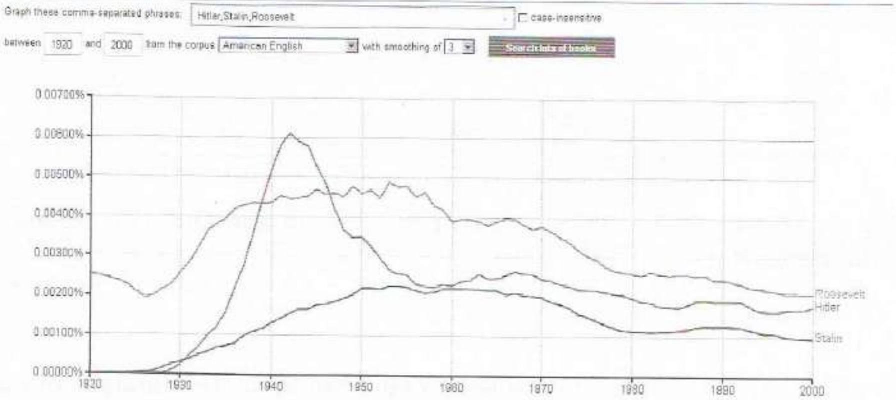
The next chart is based on the data from the British English corpus. The frequency of mentioning the names of Winston Churchill, Hitler and Stalin in English books published in Great Britain has been compared (Fig. 12).



Graph these comma-separated phrases: Hitler,Stalin,Churchill □ case-insensitive

between 1920 and 2000 from the corpus British English with smoothing of 3

Figure 12: Frequency dynamics for mentioning the names of Churchill, Hitler and Stalin in English books published in the United Kingdom



Graph these comma-separated phrases: Hitler,Stalin,Roosevelt □ case-insensitive

between 1920 and 2000 from the corpus American English with smoothing of 3

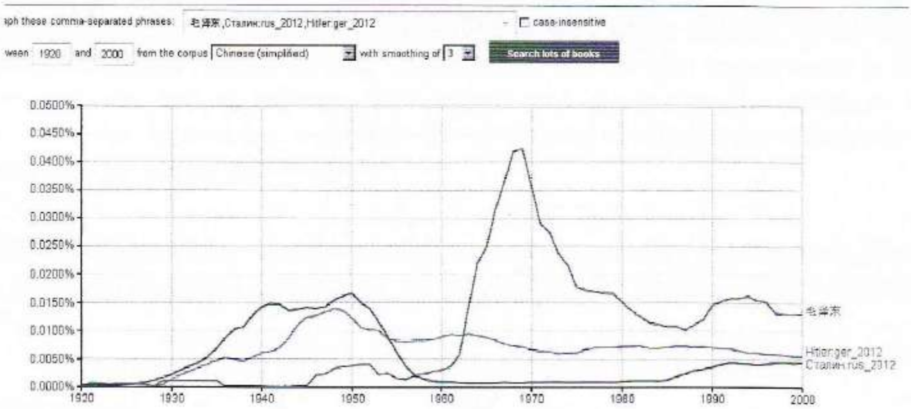
Figure 13: Frequency dynamics for mentioning the names of Roosevelt, Hitler and Stalin in English books published in the USA

The peak frequency amount for Hitler's name falls to 1943, exceeding significantly the frequency for Churchill's and Stalin's names. We can see that the leader of the enemy state is mentioned much more frequently in British books than the leader of their own country, which is very different from Russian and German corpora. The rise of the curve for Churchill can also be seen in mid-1960s. We can suppose that it has been so due to Winston Churchill's death in 1965 and the perpetuation of his memory.

The similar things when Hitler's name occurrences exceeds others can be seen within the data from the American English corpus (Fig. 13).

We see the frequency for Hitler's name exceeds the frequency for Roosevelt and Stalin's names. At the same time, the curve for Franklin Roosevelt's name is the only one from the observed curves that had no peak in home literature for the time of World War II.

Let's return for a moment to Fig. 11. We see very high level of mentioning Stalin in the Russian corpus and Hitler in the German one during 1940s. It is due to the leader's personality cult in both countries and due to an extraordinary period in the life of countries, namely, the war. However, let's see one more chart (Fig. 14).



aph these comma-separated phrases: 毛泽东,Сталин:rus_2012,Hitler:ger_2012 □ case-insensitive

ween 1920 and 2000 from the corpus Chinese (simplified) with smoothing of 3

Figure 14: Frequency dynamics for mentioning the names of 毛泽东（Mao Zedong), Hitler and Stalin from 1920 to 2000 in Chinese, German and Russian corpora

The chart shows obviously that the peak for the name of 毛泽东（Mao Zedong) in the Chinese corpus during the Cultural Revolution in China exceeds peaks for Stalin in the Russian corpus and Hitler in the German one. Against the background of Mao Zedong Stalin's and Hitler's peaks seem to be small "hillocks". This shows the level of the personality cult in China in the period 1966-1976. It wouldn' t have been so obvious without the data from corpora.

3.5 Historical and political realities and language vocabulary

In [9] we showed how the socio-political situation in the country is reflected in the language. During the years of Soviet power, many words and phrases have penetrated into the Russian language, reflecting the realities of those years. We show also how collocations with the word "enemy" were "born" and "died" (**классовый враг** 'class enemy', **злейший враг** 'worst enemy', **враги народа** enemies of the people, etc.), how the semantic field of the concept "enemy" was formed and how it changed over time.

In the same paper [9], we compared the frequency behavior of the word "enemy" in different languages (Russian, English, German, French, Chinese). In three European languages, the usage curves of this lexeme have two features. First, they have two peaks, the first of which is in 1918, i.e. at the end of the First World War, the second - at the time of the World War II, with the second peak being lower than the first. Secondly, in all three languages, the frequency of the word "enemy" tends to decrease.

The frequency of the lexeme "enemy" in the Russian language during the World War II is much higher than its equivalents in three European languages. It is very characteristic that it was not possible to reveal reflections of the events of the First World War in the Russian corpus.

The Chinese language demonstrates another model that reflects the socio-political realities that took place in the People's Republic of China (Fig. 15).

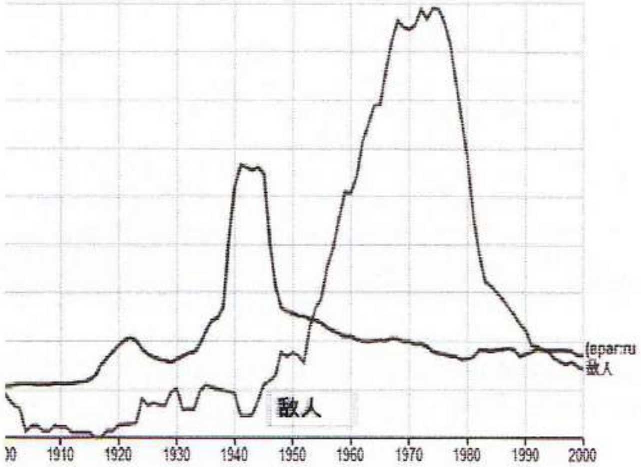


Figure 15: Frequency chart for the lexeme "enemy" in the Russian and Chinese corpora

Moreover, on the chart, we see that the policy of searching and destroying enemies and the intensity of propaganda in China during the cultural revolution in 1966-1976 was even higher than the anti-fascist propaganda in the USSR during the World War II. This is another confirmation of the connection between the frequency behavior of words and historical and political reality.

4 Conclusion

Corpus linguistics as a whole and the Google Books Ngram Viewer system in particular give unprecedented opportunities for scientists. Large amounts of annotated texts, the possibility of studying the data in several languages at the same time together with other qualities make the Google Books Ngram Viewer the remarkable research tool. Corpus linguistics gives a new approach to diachronic research, having the results that can be interesting not only for linguistics, but also for other humanitarian studies.

Our research shows that frequency changes for ngrams in printed documents are often related to some historical events and the political systems of the certain countries where documents for the corpus had been published. According to data from our research, it can be suggested that in the countries with totalitarian political regimes having personality cults, the frequency of mentioning the leader in book texts is much higher compared to liberal and democratic countries. The more severe the state regime is during the observed period of time, the more frequently the leader of the state is mentioned.

At the same time, the disadvantages and the limitations of the corpus data and the used tools should also be noticed. If we talk of Google Books Ngram Viewer, we should remember that the search of the specific lexical units is based on word forms, not on lemmas. The second problem is that the corpus is built only on books, so it is not balanced. The problems of homonymy and synonymy should also be taken into account.

Acknowledgements. This paper has been supported by the Russian Foundation for Basic Research, project No. 18-012-00474 , and partly project 17-04-00552-ОГН-А.

References

1. Michel, J.-B. et al. Quantitative analysis of culture using millions of digitized books. In: Science, 14 January 2011, pp. 176-182 (2011)
2. Solovyev, V. et al. The aspectual system in the Russian language: the semantic contribution of prefixes and suffixes [Aspektual'naya sistema v russkom yazyke: semanticheskiy vklad pristavok i suffiksov]. In: 15th Slavic Cognitive Linguistics Conference, 12–14 October 2017, Institute for Linguistic Studies, St. Petersburg, Russia. Book of abstracts, pp. 75-76 (2017)
3. Bochkarev, V. et al. Average word length dynamics as indicator of cultural changes in society. CoRR abs/1208.6109 (2012)
4. Zakharov, V., Masevich, A. Diachronic studies based on the corpus of Russian texts of Google Books Ngram Viewer [Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov]. In: Structural and Applied Linguistics. Vol. 10. Saint-Petersburg, pp. 303-327 (2014)

5. Zakharov, V., Masevich, A. Methods of corpus linguistics in culturomics [Metody korpusnoy lingvistiki v kul'turometrii]. In: CROSSLINGUA 2015: International Cross-Disciplinary Forum on Applied Cognitive Science. Cognition. Communication. Culture, Republic of Crimea, September 8-9, 2015, Feodosia, http://crosslingua.cfuv.ru/publications/2016_3_zakharov_masevich.pdf (2015)
6. Solovyev, V. Frequency-based approach to language dynamics [Chastotno-osnovannyy podkhod k yazykovoy dinamike]. In: Proceedings of the International Conference "Corpus linguistics 2013". Saint-Petersburg, pp. 424431 (2013)
7. Davies, M. Making Google Books n-grams useful for a wide range of research on language change. In: International Journal of Corpus Linguistics, 19 (3), pp. 401416 (2014)
8. Mann, J. et al. Enhanced Search with Wildcards and Morphological Inflections in the Google Books Ngram Viewer. In: Proceedings of ACL Demonstrations Track Association for Computational Linguistics 2014, http://www.dipanjandas.com/files/acl2014ngrams.pdf (2014)
9. Masevich, A., Zakharov, V. Models of frequency behavior of the Russian political vocabulary of the XX century [Modeli chastotnogo povedeniya russkoy politicheskoy leksiki XX veka]. In: Bulletin of the Novosibirsk State University. Series: Linguistics and Intercultural Communication, 15 (2), pp. 30-46 (2017)