Информатика

УДК 81’373:004
ББК 81.053+81.11
DOI 10.20913/1815-3186-2016-2-47-55

ОПЫТ КОРПУСНО-ОРИЕНТИРОВАННОГО ИСТОРИКО-КУЛЬТУРНОГО ИССЛЕДОВАНИЯ ИСТОРИЧЕСКОЙ И ПОЛИТИЧЕСКОЙ ЛЕКСИКИ

© В. П. Захаров[*], А. Ц. Масевич[**], 2016

[*] *Санкт-Петербургский государственный университет, Институт лингвистических исследований РАН, г. Санкт-Петербург, Россия; e-mail: v.zakharov@spbu.ru*

[**] *Санкт-Петербургский государственный институт культуры, г. Санкт-Петербург, Россия*

Приводятся результаты диахронических исследований исторической и политической лексики с помощью системы Google Books Ngram Viewer. Изучались изменения частоты употребления имен политических деятелей в период 1920–2000 гг. Результаты исследования демонстрируют связь изменения частотности употребления имен в текстах печатных документов с историческими событиями и политическими традициями разных стран. Рассмотрены также некоторые ограничения системы.

Ключевые слова: корпусная лингвистика, Google Books Ngram Viewer, диахронические исследования, историко-культурные исследования, имена политических деятелей.

Для цитирования: *Захаров В. П., Масевич А. Ц.* Опыт корпусно-ориентированного историко-культурного исследования исторической и политической лексики // Библиосфера. 2016. № 2. С. 47–55. DOI: 10.20913/1815-3186-2016-2-47-55.

The experience of corpus-subjected historical-cultural studies of historical and political vocabulary

V. P. Zakharov[*], A. Ts. Masevich[**]

[*] *Saint-Petersburg State University, Institute for Linguistic Studies of the Russian Academy of Sciences, Saint Petersburg, Russia; e-mail: v.zakharov@spbu.ru*

[**] *Saint-Petersburg State Institute of Culture, Saint Petersburg, Russia*

The article represents results of diachronic studying the historical and political vocabulary with Google System Books Ngram Viewer. The authors investigate changes in use frequency of political figures names during 1920–2000. The study results demonstrate the relationship between changes in the frequency use of names in printed documents texts with historical events and political traditions of different countries. Some limitations of the system were also considered.

Keywords: corpus linguistics, Google Books Ngram Viewer, diachronic research, historical-cultural studies, political figures names.

Citation: *Zakharov V. P., Masevich A. Ts.* The experience of corpus-subjected historical-cultural studies of historical and political vocabulary // *Bibliosfera*. 2016. № 2. P. 47–55. DOI: 10.20913/1815-3186-2016-2-47-55.

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования корпусов текстов с применением компьютерных технологий. Существует множество определений термина «корпус». Все они так или иначе фиксируют основные компоненты этого понятия. Корпус должен быть электронным, репрезентативным, размеченным и включать тексты и фрагменты текстов, отобранные по определенному принципу в соответствии с четкими языковыми критериями, определяемыми решаемой задачей (см., например, [1, с. 5]).

Наличие в корпусах метаданных, таких как вид и жанр текста, сфера функционирования, тематика, дата создания текста, сведения об авторе и т. п., позволяет использовать их не только для изучения языка, но и для социологических, литературоведческих, культурологических и других исследований. Диахронические исследования на основе корпусных данных позволяют выявлять факты и закономерности не только лингвистического, но и историко-культурного значения.

Тексты, написанные на естественном языке, это не просто акты коммуникации, но и знаки (символы) вторичных моделирующих систем (Ю. М. Лотман). В последние годы в рамках культурологии появилось направление научных исследований, называемое «культурометрия» (синоним «квантитативная культурология»). В отечественной литературе

этот термин можно трактовать как развитие идей, высказанных Ю. М. Лотманом: «Являясь важным механизмом памяти культуры, символы переносят тексты, сюжетные схемы и другие семиотические образования из одного пласта культуры в другой. Пронизывающие диахронию культуры константные наборы символов в значительной мере берут на себя функцию механизмов единства: осуществляя память культуры о себе, они не дают ей распасться на изолированные хронологические пласты. Единство основного набора доминирующих символов и длительность их культурной жизни в значительной мере определяют национальные и ареальные границы культур» [2, с. 192].

В зарубежных исследованиях для понятия «культурометрия» используется термин culturomics. В словаре dictionary.com он определяется как «Исследование культуры человечества, направлений ее развития во времени посредством количественного анализа слов и словосочетаний в очень больших корпусах оцифрованных текстов»[1] [3].

Уже существует несколько российских публикаций [4–8], в которых описываются попытки диахронических исследований лексики русского языка методами лингвистической статистики. Имеется также некоторое количество зарубежных публикаций по методике и инструментам таких исследований [9–13]. Компьютерные технологии и корпусная лингвистика дают принципиально новые инструменты диахронического исследования языка и скрывающихся за ним реалий. В частности, можно проследить поведение лексической единицы во времени, а точнее, изменения частоты ее употребления в письменном языке. К таким инструментам относится система Google Books Ngram Viewer [14], предоставившая основной массив данных для нашего исследования.

Цели и задачи исследования

Основная задача – разработка методической модели сравнительно-диахронического исследования исторической и политической лексики на примере употребления имен политических деятелей в разное время в разных странах.

Для решения этой задачи необходимо:

1. Показать, каким образом в течение определенного периода меняется частота употребления отобранных лексических единиц. Корпусная лингвистика и система Google Books Ngram Viewer в частности предоставляют беспрецедентные возможности для выявления таких изменений и, тем самым, для историко-культурных и лингвистических исследований. Огромные массивы размеченных текстов, возможность изучения поведения лексических эквивалентов в нескольких языках одновременно, быстрота получения результатов, многие другие достоинства – все это делает такую систему исключительно ценным исследовательским инструментом, который, на наш взгляд, еще не вполне оценен;

2. Выявить ограничения корпусно-ориентированного подхода. Для того чтобы можно было делать достоверные выводы на основе корпусных данных, следует хорошо представлять себе недостатки и ограничения тех инструментов, которыми мы пользуемся;

3. Наконец, возникает вопрос интерпретации данных. Как оценивать полученные результаты с точки зрения исторической науки? Как установить связь употребления тех или иных слов с историко-политической ситуацией? Некоторые такие интерпретации с известной осторожностью попытаемся сделать в настоящей работе.

Материал и инструмент исследования (Google Books Ngram Viewer)

Сервис Google Books Ngram Viewer доступен в Интернете с 2010 г. Он включает корпуса девяти языков. Общий объем корпусов – 8 116 746 текстов и 861 877 262 497 словоупотреблений. По утверждению разработчиков, число текстов в корпусе составляет 6% всех когда-либо изданных печатных документов. Русский корпус содержит 591 310 текстов (книг), образующих корпусный массив объемом более 67 млрд словоупотреблений. Временной охват русского корпуса – с 1720-х гг. по 2008 г.

При вводе печатного документа в базу данных системы Google Books каждый текст подвергался сканированию с последующим оптическим распознаванием. Файл каждой книги снабжается метаданными, во введенных текстах осуществляется метатекстовая и частично грамматическая разметка.

Система осуществляет поиск заданной N-граммы в массиве корпуса и строит график частоты ее встречаемости по годам в период времени, определяемый пользователем. Под термином N-грамма в данном случае понимается последовательность от одного до пяти слов. На горизонтальной оси графика показаны годы, входящие в заданный временной период. По вертикальной оси откладывается относительная частота встречаемости в корпусе заданной N-граммы в соответствующем году, умноженная на 100 (то есть выраженная в процентах) (см. рис. 1). Относительная частота встречаемости N-граммы за определенный год подсчитывается следующим образом: число употреблений N-граммы в данном году делится на общее число словоупотреблений в корпусе в этом же году. Так,

[1] 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: Culturomics can pinpoint periods of accelerated language change.

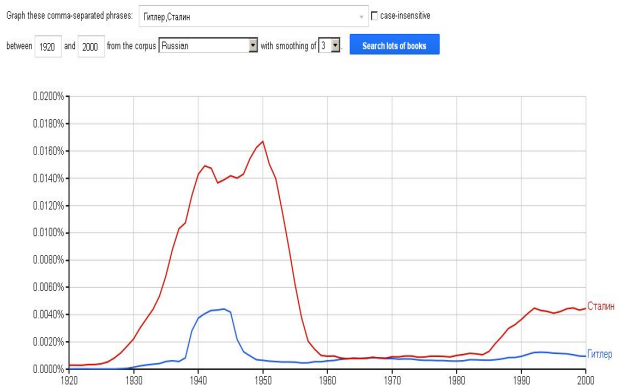

Рис. 1. Динамика частоты упоминания имен «Сталин» и «Гитлер» в книгах на русском языке

Fig. 1. Frequency dynamics of Stalin and Hitler names references in Russian-language books

например, число словоупотреблений слова «slavery» (рабство) в 1861 г. в английском корпусе 2009 г. составило 21 460 на 11 687 страницах 1208 книг. Всего в корпусе за 1861 г. содержится 386 434 758 словоупотреблений. Таким образом, значение относительной частоты использования слова «slavery» (рабство) в 1861 г. составляет 0,0055533307% [12].

При построении графиков изменения частоты употребления лексических единиц используется так называемое «сглаживание» (smoothing). При нулевом сглаживании в графике учитывается относительная частота встречаемости N-граммы за каждый год. Однако по-настоящему тенденция в динамике встречаемости слов прослеживается более отчетливо при скользящем усреднении данных, когда относительная частота усредняется для заданного интервала. Если значение коэффициента сглаживания равно 3, то это означает, что для некоторого года к числу словоупотреблений искомого слова за этот год прибавляется число словоупотреблений за три предыдущих года и три последующих, и полученная сумма делится на семь. Относительное значение этой средней величины будет отражено на вертикальной оси.

В системе нет морфологической нормализации лексических единиц, иначе говоря, поиск лексических единиц (слов или словосочетаний), для которых строится график – это поиск по словоформам. Система предусматривает использование пользовательских тегов для модификации условий построения графиков. И в их числе есть тег _INF (Inflections), который строит кривые для всех форм словоизменительной парадигмы данного слова. Однако данная функция для русского языка работает не всегда корректно.

Имеется тег «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы или до нее позволяет строить график встречаемости десяти наиболее частотных сочетаний данной N-граммы с другими словами – справа или слева от нее.

Над кривыми графиков возможны операции: суммирование, вычитание, умножение, деление. Например, суммирование (сложение) кривых, при котором поисковые слова вводятся в окно запроса через знак +, позволяет суммировать значения каждой точки по оси ординат двух или более кривых. Это может быть использовано как аналог поиска по лемме, например: *лошадь* + *лошади* + *лошадью* + ...+ *лошадях*.

Полезная операция – умножение графиков, позволяющая умножать на n значения всех точек графика (например, лемматизация*100). Данная операция позволяет сделать сопоставимым вид кривых, значения которых отличаются на несколько порядков.

Кроме построения графиков, система предоставляет ссылки к текстам, найденным по запросам, в которых встретились заданные слова. Как правило, это библиографические описания книг и фрагменты текстов с выделением в них заданных N-грамм. В некоторых случаях доступен полный текст книги в графическом формате. Более подробно о сервисе Google Books Ngram Viewer см. [4].

В данном исследовании использовалось суммирование (сложение) графиков и построение кривых по корпусам разных языков на одном общем графике.

Результаты исследования

Нижеследующее исследование связано с частотой употребления имен политических деятелей в текстах книг на разных языках. Поскольку исследование заявлено как сравнительно-диахроническое, зададим временной промежуток для построения графиков с 1920 по 2000 г. Но сами деятели были наиболее активны, а значит, и наиболее упоминаемы в печатных изданиях в середине XX века.

Может показаться, что изучать язык или реалии по советским книгам этого периода неверно, поскольку массив книг периода Великой Отечественной войны мал и не репрезентативен. Однако это не так. Приведем некоторые данные о том, что представляло собой книгопечатание в СССР в годы войны. «В целом в годы Великой Отечественной войны наблюдалось *сокращение объема издательской продукции* по сравнению с предвоенным периодом: если в 1938–1940 гг. в СССР было выпущено 130,6 тыс. названий книг тиражом 1875,4 млн экз., то в 1941–1945 гг. – соответственно 109,1 тыс. и 1691,7 млн экз. … Сыграло свою роль и резкое сокращение объемов книг. Средний объем в 1942 г. был почти в три раза меньше, чем в 1940 г., в то время как средний тираж увеличился более чем вдвое» [15].

Неизвестно, какая часть выпущенных во время войны книг вошла в 590-тысячный корпус русских

книг корпуса Google Books. Но графики, построенные для единиц общеупотребительной лексики, не показывают каких-либо отличий в «поведении» этих слов по сравнению с годами мирного времени, следовательно, можно предположить, что книг военных лет в корпусе достаточно много.

***

На рис. 1 приводится график с именами двух протагонистов Второй мировой войны – Сталина и Гитлера в книгах на русском языке.

Кривая, отражающая частотность слова «Сталин», легко поддается интерпретации и вполне соответствует современным историческим представлениям. К началу войны кривая достигает некоторого пика. В первые годы войны отмечается незначительное снижение числа упоминаний, затем снова рост и абсолютный пик в 1950 г. После чего кривая резко опускается, оставаясь до середины 1980-х гг. на одном уровне, а затем снова начинается рост числа упоминаний, но уже, по-видимому, в негативной коннотации. Имя «Гитлер» появляется в текстах русских книг с начала 1930-х гг. Во время войны число упоминаний этой фигуры резко возрастает, а после войны падает и до конца заданного нами периода находится на одном уровне. Что интересно, от начала 1960-х гг. до середины 1980-х гг. обе кривые практически совпадают.

Рассмотрим далее число упоминаний этих двух политических фигур в немецкоязычных книгах, не забывая при этом, что корпус содержит тексты книг, изданных в разных немецкоязычных государствах (рис. 2). Кроме того, существовала немецкая литература в изгнании.

На первый взгляд модель поведения слов на рис. 2 сходна с той, что показана на рис. 1. Руководитель своей страны в книгах, изданных во время войны, упоминается чаще, чем руководитель вражеского государства. Однако обратим внимание, что пик частотности собственного имени «Hitler» приходится на 1948 г., а пик частотности имени «Stalin» – на 1953 г., после чего следует снижение.

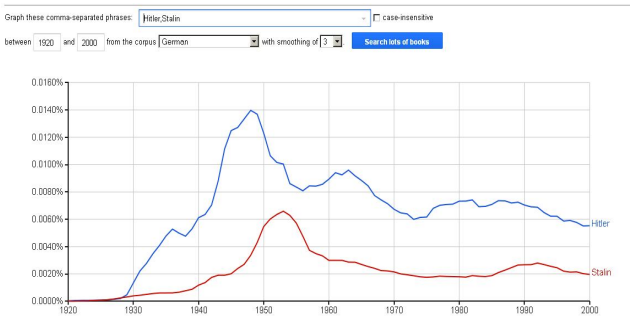

Рис. 2. Динамика частоты упоминания имен «Stalin» и «Hitler» в книгах на немецком языке

Fig. 2. Frequency dynamics of Stalin and Hitler names references in German-language books

Во время войны в книгах, изданных на территории нацистской Германии, имя «Hitler» упоминается лишь в положительной коннотации. Но, как сказано выше, существовала и обширная эмигрантская литература, которая также вошла в базу данных. Потому в немецкоязычном корпусе можно встретить и издания такого типа (рис. 3).

Рис. 3. Пример ссылки в базе данных Google Books на издание латиноамериканского комитета свободных немцев (1944 г.)

Fig. 3. A reference on a publication of Latin American Free Germans Committee in Google Books database

В конце 1940-х – начале 1950-х гг. в немецкоязычной литературе появилось много книг, обличающих Гитлера (рис. 4).

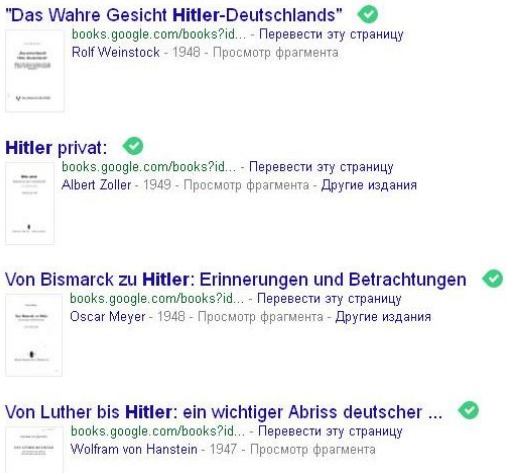

Рис. 4. Примеры немецких книг 1947–1950 гг., в которых употребляется собственное имя «Hitler»

Fig. 4. Cases of Germain books for 1947–1950 with Hitler name references

Разумеется, и фигуре руководителя враждебного государства в немецкой литературе также уделялось внимание (рис. 5).

Рис. 5. Примеры ссылок на немецкие книги, с употреблением собственного имени «Stalin» (1942 и 1944 гг.) (поиск в базе данных Google Books)

Fig. 5. References of German books with Stalin name (1942 and 1944) (search in Google Books database)

В конце 1940-х – начале 1950-х гг. отмечается рост употребления имени «Stalin» в немецкоязычных книгах. На рис. 6 видно, что уже с 1945 г. издается большое число трудов и речей Сталина, по-видимому, на территории советской оккупационной зоны, а позднее в ГДР, что обусловило рост его упоминаний в текстах книг. Пик кривой употребления этого имени приходится на 1952 г. После 1953–1956 гг. кривая идет вниз и остается на постоянном уровне с незначительными изменениями до конца выбранного периода (см. рис. 6).

Рис. 6. Ссылки на немецкие книги после 1945 г., в текстах которых упоминается имя «Stalin»

Fig. 6. References on German books after 1946 where Stalin name is mentioned in texts

Для сравнения были построены графики встречаемости имен «Hitler» и «Stalin» в другом корпусе – DWDS (Das Digitale Wörterbuch der deutschen Sprache) [16] (рис. 7, 8). Корпус имеет как некоторые ограничения по сравнению с Google Books Ngram Viewer – невозможность построить два графика на одном рисунке или задать временной период, так и преимущества – он содержит разные типы документов – беллетристика (Belletristik), «серая» литература (Gebrauchliteratur), научные (Wissenshaft) и газетные статьи (Zeitung).

Тенденции, обнаруженные в графиках, построенных в корпусе DWDS, в целом сходны с тенденциями графиков Ngram Viewer, что является косвенным подтверждением достоверности данных, полученных в предыдущих экспериментах. Специального анализа требуют кривые, отражающие встречаемость имен в разных категориях документов.

Сопоставим поведение этих собственных имен в немецком и русском корпусе. Google Ngram Viewer дает возможность построения на одном рисунке графиков встречаемости слов в разных корпусах (рис. 9).

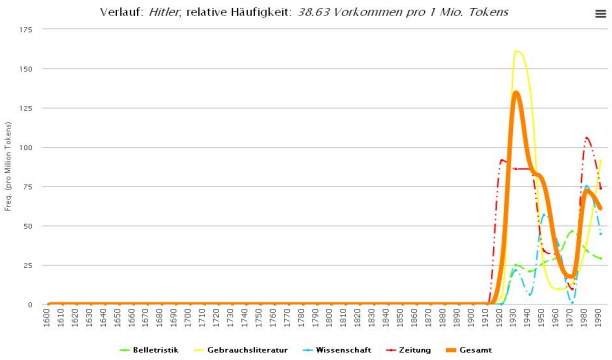

Рис. 7. Динамика упоминания имени «Hitler» в немецкоязычных документах корпуса DWDS, по категориям беллетристика, «серая» литература, научная литература, газеты, все категории (Gesamt)

Fig. 7. Dynamics of Hitler name mentioning in German-language documents of DWDS corpus in categories of fiction, «grey» literature, scientific literature, newspapers, all categories (Gesamt)

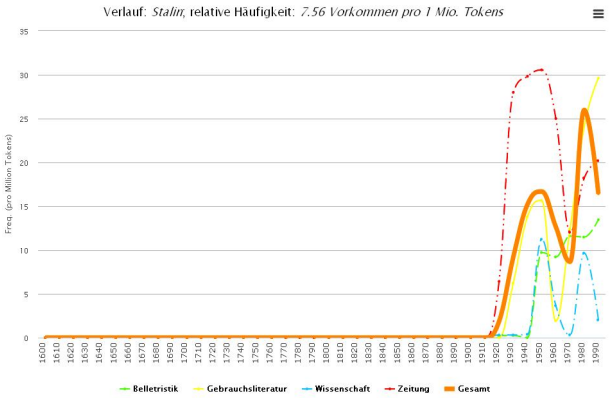

Рис. 8. Динамика упоминания имени «Stalin» в немецкоязычных документах корпуса DWDS по категориям беллетристика, «серая» литература, научная литература, газеты, все категории (Gesamt)

Fig. 8. Dynamics of Stalin name mentioning in German-language documents of DWDS corpus in categories of fiction, «grey» literature, scientific literature, newspapers, all categories (Gesamt)

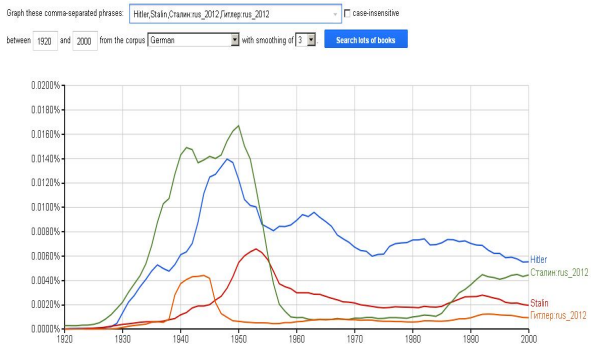

Рис. 9. Динамика частоты встречаемости собственных имен «Сталин» и «Гитлер» в русском и немецких корпусах

Fig. 9. Frequency dynamics of proper names «Stalin» and «Hitler» mentioning in Russian and German corps

Кривая динамики встречаемости имени «Сталин» в русском корпусе имеет наиболее высокий пик в 1950 г. Рост числа упоминаний начинается в 1930-е гг., а в десятилетие 1940–1950 гг. это число удерживается на стабильно высоком уровне. Причем этот уровень выше уровня упоминаний имени «Гитлер» в немецком корпусе. Снижение числа упоминаний имени «Гитлер» в немецких книгах происходит не так резко, как имени «Сталин» в русских (ср. 1950–1965 гг.). Более того, в середине 1960-х гг. число упоминаний о Гитлере растет и в дальнейшем до конца заданного периода это число остается на уровне более высоком, чем в остальных трех кривых. Выдвинем гипотезу, которая, наш взгляд, заслуживает внимания, хотя и требует проверки. В Советском Союзе негативный исторический опыт замалчивался, о чем, по-видимому, говорит низкая частотность упоминаний о Сталине в период с середины 1950-х до середины 1980-х гг. В Германии (а возможно, и в обоих германских государствах) – напротив, негативный опыт прошлого изучался и осмысливался.

Следующие графики построены на базе корпусов британского английского языка и американского английского. Сопоставлялись частоты встречаемости имен «Churchill», «Roosevelt», «Stalin» и «Hitler» в книгах на английском языке, изданных в Великобритании и в США (рис. 10, 11).

Пиковое значение встречаемости имени «Hitler» в английских книгах, которое приходится примерно на 1943 г., существенно превышает встречаемость имен «Churchill» и «Stalin» (рис. 10). Таким образом, поведение кривых встречаемости имен национальных лидеров прямо противоположно поведению их в русском и немецком корпусах. В английских книгах намного чаще упоминается руководитель враждебной страны, чем руководитель собственной страны или союзного государства. Кривая встречаемости имени «Churchill» образует совсем небольшой подъем в годы войны, затем снижается, а следующий подъем отмечается в середине 1960-х гг. Можно предположить, что это связано со смертью У. Черчилля (1965) и увековечиванием его памяти. Примерно пять лет кривая остается на высоком уровне, затем снижается.

Кривая имени «Roosevelt» – единственная из рассмотренных кривых встречаемости имени национального лидера, не имеющая выраженного пика в годы Второй мировой войны (рис. 11). Рост и снижение числа упоминаний о Рузвельте в американских книгах происходят плавно. Объяснить это обстоятельство мы пока не можем. Возможно, интерпретировать график следует, отталкиваясь от психологических, социологических особенностей общества США. Кривая частоты упоминаний имени «Hitler», в отличие от имени «Stalin», в начале 1940-х гг. образует пик, с последующим резким снижением.

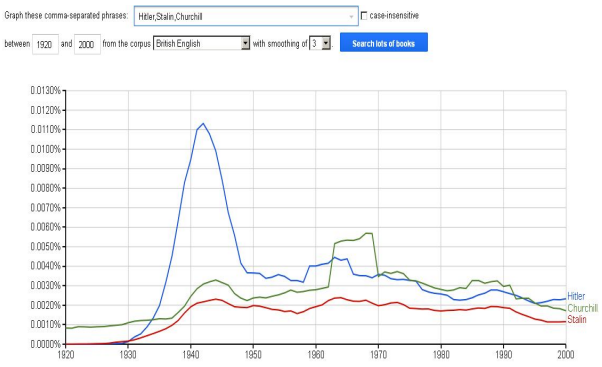

Рис. 10. Динамика частоты встречаемости имен «Churchill», «Hitler» и «Stalin» в книгах на английском языке, изданных в Соединенном Королевстве

Fig. 10. Frequency dynamics of proper names Churchill, Hitler and Stalin mentioning in English-language books published in the United Kingdom

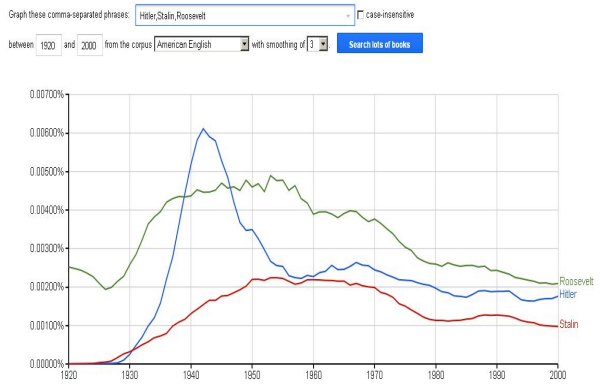

Рис. 11. Динамика частоты встречаемости имен «Roosevelt», «Hitler» и «Stalin» в книгах на английском языке, изданных в США

Fig. 11. Frequency dynamics of proper names Roosevelt, Hitler and Stalin mentioning in English-language books published in the USA

Наконец, построим совмещенные графики по корпусам разных языков. Графики строятся по именам глав государств с соответствующими языками: «Сталин», «Churchill», «de Gaulle», «Hitler», «Mussolini», «Roosevelt» (рис. 12) и «Сталин», «Hitler», «毛泽东 (Мао Цзедун)» (рис. 13).

Кривые встречаемости для имен «Сталин», «Hitler», «Mussolini» соответственно в русских, немецких и итальянских текстах дают наиболее выраженный подъем в военные годы, в эти же годы кривые для имен «Churchill» и «de Gaulle» имеют значительно меньший подъем, а кривая имени «Roosevelt» совсем не дает подъема в эти годы.

Абсолютный пик кривой имени «Мао Цзедун» в корпусе китайского языка превышает по высоте пики кривых для имен «Сталин» в русском языке и «Гитлер» в немецком корпусах. Исторически этот пик приходится на конец 1960-х гг., то есть на время «культурной революции», когда культ Мао в Китае достиг своей высшей точки.

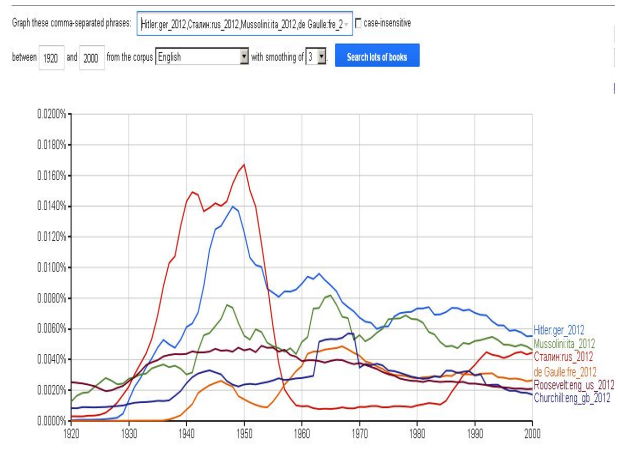

Рис. 12. Динамика упоминаний имен руководителей государств, участвовавших во Второй мировой войне, в корпусах английского британского, английского США, итальянского, немецкого, русского и французского языков Google

Fig. 12. Mentioning dynamics of names of the state heads participating in World War II in corps of the British English, US English, Italian, German, Russian and French languages Google

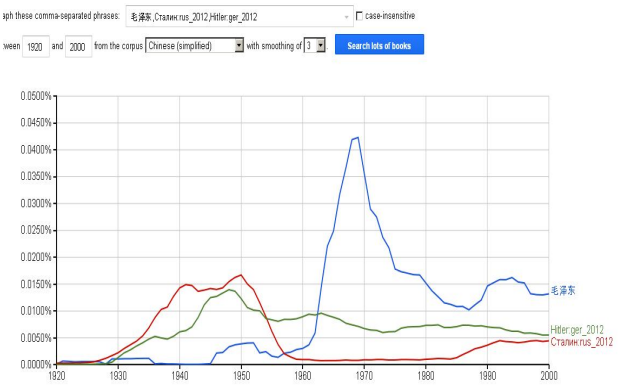

Рис. 13. Динамика упоминаний имен «**毛泽东** (Мао Цзедун)», «Hitler» и «Сталин» за период в корпусах китайского, немецкого и русского языков

Fig. 13. Mentioning dynamics of names 毛泽东 (Mao Zedong), Hitler and Stalin during the period in Chinese, German and Russian language corps

Но даже в 1990-х гг. число упоминаний Мао Цзедуна в китайских книгах выше упоминаний Сталина и Гитлера.

Заключение

Система Google Books Ngram Viewer, и, по-видимому, другие языковые корпуса, позволяют предложить принципиально новый подход к диахроническим исследованиям, результаты которых могут представлять интерес не только для лингвистики, но и для исторической науки, социологии, культурологии и др. Разумеется, методика подобных исследований нуждается в детальной проработке.

Система Google Books Ngram Viewer, безусловно, не лишена недостатков. Их анализу можно посвятить специальную публикацию. Для настоящей работы могут иметь значение следующие моменты: поиск заданных лексических единиц ведется по словоформам, а не по леммам. Впрочем, имеющиеся данные [7, 8], а также наш опыт показывают, что поведение во времени отдельных словоформ какой-либо лексемы, как правило, сходно с «суммарным» поведением всех членов ее словоизменительной парадигмы.

Корпусы построены исключительно на текстах печатных книг. На сегодняшний день мы не встретили публикаций о сопоставлении корпусов книг и периодических изданий. Разумеется, реалии временных периодов в книгах и газетах отражаются по-разному как в отношении частотности лексики, так и в отношении оперативности, это хорошо видно на рис. 7, 8. Однако достоверной количественной оценки таким различиям сделано пока не было.

Существуют также проблемы, общие для всех электронных систем, работающих с естественным языком. Это, прежде всего, явления омонимии и синонимии. В отношении личных имен они требуют изучения. Мы тем не менее считаем, что в нашем исследовании ими можно пренебречь. Синонимичные обозначения изучаемых исторических лиц можно не учитывать, так как в пределах текста (одного или нескольких) число использований основного обозначения лица (например, Сталин) достаточно для достоверной количественной оценки и построения графика, несмотря на наличие других обозначений («вождь», «отец народов» и т. д.). Омонимия в нашем случае также не играет большой роли, так как фамилии и псевдонимы исследуемых нами персон мало распространены. При выборочном просмотре текстов мы не выявили ни одного случая упоминания однофамильцев исследуемых исторических лиц. В других исследованиях такая проблема может возникнуть, так как существуют политические деятели с распространенными фамилиями, например, Жуков, Медведев и др.

На наш взгляд, проведенное исследование отчетливо демонстрирует, что изменение частотности N-грамм, в частности, имен собственных, в печатных документах связано с определенными историческими событиями, а также с политическим режимом государства, на территории которого издаются документы, тексты которых образуют корпус.

Так, по данным, полученным в ходе нашего исследования, можно предположить, что в странах с тоталитарным политическим режимом (и вследствие этого с культом вождя) в годы войны число упоминаний руководителя государства в текстах книг растет, а в либерально-демократических странах в текстах книг больше внимания уделяется

лидеру враждебного государства. Причем, как представляется, частотность упоминания главы государства в текстах книг тем выше, чем более жесток политический режим и чем более выражен культ вождя. Воздержимся, однако, от дальнейших историко-политических комментариев, так как наши данные требуют дополнительных исследований, а для комментариев было бы целесообразно привлечь профессиональных историков.

Следует отметить, что исследование позволяет, с одной стороны, подтвердить очевидные с точки зрения общих представлений явления, например, изменение частотности имени Сталин в русских текстах (см. рис. 1) кажется вполне предсказуемым. С другой стороны, в ходе исследования выявлены не столь очевидные явления, которые могли быть замечены только с помощью использованной системы (см. рис. 9 и комментарии, а также рис. 12, 13).

Нам удалось на основе корпусных данных показать отдельные модели изменения частоты употреблений лексических единиц во времени и связать эти модели с историко-политической ситуацией. В целом, на наш взгляд, настоящая работа позволяет говорить о новом направлении в историко-культурных исследованиях, которое требует дополнительных исследований и разработки методического аппарата.

Литература

1. *Захаров В. П., Богданова С. Ю.* Корпусная лингвистика. Санкт-Петербург : СПбГУ, 2013. 148 с.
2. *Лотман Ю. М.* Символ в системе культуры // Статьи по семиотике и топологии культуры. Таллинн : Александра, 1992. Т. 1. С. 191–199.
3. Culturomics // Dictionary.com. URL: http://dictionary.reference.com/browse/culturomics (accessed 12.07.2015).
4. *Захаров В. П., Масевич А. Ц.* Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer // Структурная и прикладная лингвистика. Санкт-Петербург, 2014. Вып. 10. С. 303–327.
5. *Захаров В. П., Масевич А. Ц.* Диахронические исследования терминологической лексики // Прикладная лингвистика в науке и образовании : сб. тр. VII Междунар. науч. конф. Санкт-Петербург, 10–12 апр. 2014 г. Санкт-Петербург, 2014. С. 95–100.
6. *Масевич А. Ц.* Google Books Ngram Viewer – инструмент для историко-культурных исследований // Информационные ресурсы – футурологический аспект: планы, прогнозы, перспективы : материалы X всерос. науч.-практ. конф. «Электронные ресурсы библиотек, музеев, архивов» (30–31 окт. 2014 г., Санкт-Петербург). Санкт-Петербург, 2014. С. 43–58.
7. *Соловьев В. Д.* Частотно-основанный подход к языковой динамике // Труды международной конференции «Корпусная лингвистика-2013». Санкт-Петербург, 2013. С. 424–431.
8. *Соловьев В. Д.* Частотность как объект корпусных исследований // Труды международной конференции «Корпусная лингвистика-2011». Санкт-Петербург, 2011. С. 328–332.
9. *Baroni M., Lenci A.* Distributional memory : a general framework for corpus-based semantics // Computational Linguistics. 2010. Vol. 36, № 4. P. 673–721.
10. *Davies M.* Making Google Books n-grams useful for a wide range of research on language change // International Journal of Corpus Linguistics. 2014. Vol. 19, № 3. P. 401–416.
11. *Mann J., Zhang D., Lu Yang, Dipanjan Das, Petrov S.* Enhanced search with wildcards and morphological inflections in the Google Books Ngram viewer // Proceedings of ACL Demonstrations Track Association for computational linguistics 2014. URL: http://www.dipanjandas.com/files/acl2014ngrams.pdf (accessed 12.07.2015).
12. *Michel J.-B., Shen Y. K., Aidenet A. P.* [et al.]. Quantitative analysis of culture using millions of digitized books science // Science. 2011. Vol. 331, № 6014. P. 176–182. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/ (accessed 12.07.2015).
13. Sociolinguistics and language history: studies based on the corpus of early English correspondence / eds: T. Nevalainen, H. Raumolin-Brunberg. Amsterdam : Cambridge Univ. Press, 1996. 213 p.
14. Google books Ngram viewer. URL: https://books.google.com/ngrams (accessed 14.07.2015).
15. История книги : учебник для вузов / под ред. А. А. Говорова, Т. Г. Куприяновой. Москва : Мир книги, 1998. 346 с.
16. Das Digitale Wörterbuch der deutscher Sprache. URL: http://www.dwds.de/ (accessed 12.07.2015).

References

1. *Zakharov V. P., Bogdanova S. Yu. Korpusnaya lingvistika* [Corpus liquistics]. Saint Petersburg, SPbGU, 2013. 148 p. (In Russ.).
2. *Lotman Yu. M.* A simbol in the culture system. *Stat'i po semiotike i topologii kul'tury*. Tallinn, Aleksandra, 1992, 1, 191–199. (In Russ.).
3. Culturomics. *Dictionary.com*. URL: http://dictionary.reference.com/browse/culturomics (accessed 12.07.2015).
4. *Zakharov V. P., Masevich A. Ts.* Diachronic studies on a base of the Russian texts corp in Google Books Ngram Viewer. *Strukturnaya i prikladnaya lingvistika*. Saint Petersburg, 2014, 10, 303–327. (In Russ.).
5. *Zakharov V. P., Masevich A. Ts.* Diachronic studies of terminological lvocabulary. *Prikladnaya lingvistika v nauke i obrazovanii : sb. tr. VII Mezhdunar. nauch. konf. Sankt-Peterburg, 10–12 apr. 2014 g.* Saint Petersburg, 2014, 95–100. (In Russ.).
6. *Masevich A. Ts.* Google Books Ngram Viewer – a tool for historic-cultural studies. *Informatsionnye resursy – futurologicheskii aspekt: plany, prognozy, perspektivy : materialy X vseros. nauch.-prakt. konf. «Elektronnye resursy bibliotek, muzeev, arkhivov» (30–31 okt. 2014 g., Sankt-Peterburg).* Saint Petersburg, 2014, 43–58. (In Russ.).
7. *Solov'ev V. D.* A frequency-based approach to language dynamics. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika-2013».* Saint Petersburg, 2013, 424–431. (In Russ.).
8. *Solov'ev V. D.* The frequency of an object of corp studies. *Trudy mezhdunarodnoi konferentsii «Korpusnaya lingvistika-2011».* Saint Petersburg, 2011, 328–332. (In Russ.).

9. *Baroni M., Lenci A.* Distributional memory: a general framework for corpus-based semantics. *Computational Linguistics,* 2010, 36 (4), 673–721.
10. *Davies M.* Making Google Books n-grams useful for a wide range of research on language change. *International Journal of Corpus Linguistics,* 2014, 19 (3), 401–416.
11. *Mann J., Zhang D., Lu Yang, Dipanjan Das, Petrov S.* Enhanced search with wildcards and morphological inflections in the Google Books Ngram viewer. *Proceedings of ACL Demonstrations Track Association for computational linguistics 2014*. URL: http://www.dipanjandas.com/files/acl2014ngrams.pdf (accessed 12.07.2015).
12. *Michel J.-B., Shen Y. K., Aidenet A. P. [et al.]*. Quantitative analysis of culture using millions of digitized books science. *Science,* 2011, 331 (6014), 176–182. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/ (accessed 12.07.2015).
13. *Nevalainen T., Raumolin-Brunberg H. (eds)*. Sociolinguistics and language history: studies based on the corpus of early English correspondence. Amsterdam, Cambridge Univ. Press, 1996. 213 p.
14. *Google books Ngram viewer*. URL: https://books.google.com/ngrams (accessed 14.07.2015).
15. *Govorov A. A., Kupriyanova T. G. (eds). Istoriya knigi : uchebnik dlya vuzov* [Book history : a textbook]. Moscow, Mir knigi, 1998. 346 p. (In Russ.).
16. *Das Digitale Wörterbuch der deutscher Sprache*. URL: http://www.dwds.de/ (accessed 12.07.2015).

Материал поступил в редакцию 13.11.2015 г.

Сведения об авторах: *Захаров Виктор Павлович – кандидат филологических наук, доцент СПбГУ, ведущий научный сотрудник ИЛИ РАН,*
Масевич Андрей Цезаревич – старший преподаватель кафедры информационного менеджмента