Crosslingua'2015 <08.10 — 09.10. 2015 Crimea, Simferopol

МЕТОДЫ КОРПУСНОЙ ЛИНГВИСТИКИ В КУЛЬТУРОМЕТРИИ

Виктор Павлович Захаров

кандидат филологических наук
доцент кафедры математической лингвистики
Санкт-Петербургский государственный университет
Санкт-Петербург, Россия
E-mail: vz1311@yandex.ru

Андрей Цезаревич Масевич

старший преподаватель кафедры менеджмента
библиотечно-информационного факультета
Санкт-Петербургский государственный институт культуры
Санкт-Петербург, Россия
E-mail: andmasev@mail.ru

Приводятся результаты диахронических исследований военной и политической лексики с помощью системы Google Books Ngram Viewer. Изучались изменения частоты употребления лексических единиц в период1920-2000 гг. .Результаты исследования указывают на связь изменения частотности лексических единиц в текстах печатных документов с историческими событиями. Рассмотрены также некоторые ограничения системы. Делается вывод, что предлагаемый метод является перспективным как для историко-культурных, так и для лингвистических исследований.

Введение

Корпусная лингвистика – раздел компьютерной лингвистики, занимающийся разработкой общих принципов построения и использования лингвистических корпусов (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 снабженным поисковым инструментарием. Тип разметки и поисковый аппарат определяются лингвистическими исследовательскими задачами, которые предположительно будут решаться посредством этого корпуса [1].

Язык, как известно, динамичная система; в исторические периоды разной продолжительности на всех его уровнях (в фонетике и письме, морфологии и лексике, синтаксисе и семантике) происходят изменения: частота встречаемости одних элементов, **форм или** явлений уменьшается, а бывает, что явления, **формы или** элементы полностью выходят из употребления, другие же возникают или становятся более частотными, чем прежде. Изменения обусловлены действием факторов различной природы – прежде всего психологических, социальных и культурных.

Диахронические исследования языка позволяют выявить 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 её развития во времени посредством количественного анализа слов и словосочетаний в очень больших корпусах оцифрованных текстов» [1] [3]

Компьютерные технологии и корпусная лингвистика дают принципиально новые инструменты диахронического исследования языка. В частности, можно проследить поведение лексической единицы во времени, а точнее, изменения частоты её употребления в письменном языке. К таким инструментам относится система Google Books Ngram Viewer, являющаяся основным инструментом нашего исследования.[4, 6, 12, 13]

В настоящей публикации описывается возможная методическая модель сравнительно-диахронического исследования на примере военной и политической лексики на примере периода Второй мировой войны

Сервис Google Books Ngram Viewer доступен в Интернете с 2010 г. Он включает в себя корпуса девяти языков. Суммарный объём 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 591310 текстов (книг), образующих корпусный массив более 67 137 666 353 словоупотреблений. Временной охват русского корпуса – с 20-х гг. XVIII века 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, т.е. умноженное на 100 отношение абсолютного числа употреблений данной N-граммы за определенный год к общему числу словоупотреблений в корпусе в этом же году. Например, число словоупотреблений слова «slavery» (рабство) в 1861 г. в английском корпусе

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.

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].

Система снабжена развитым поисковым аппаратом и возможностями представления данных. Система осуществляет поиск словоформ, грамматическая нормализация поисковых терминов не осуществляется.

В данном исследовании использовались следующие возможности Ngram Viewer :

Суммирование (сложение) графиков. Операция позволяет суммировать значения каждой точки двух или более графиков. Для осуществления операции поисковые слова вводятся в окно через знак +, например: *танк* + *танки*. (рис.3,11)

Тэг «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы или до неё (рис. 5) позволяет строить график встречаемости десяти наиболее частотных сочетаний N-граммы и слова следующего за нею или ей предшествующего.

Результаты исследования

Поскольку исследование заявлено как сравнительно-диахроническое, зададим временной промежуток для построения графиков с 1920 по 2000 г. Рассмотрим лексику, обозначающую рода войск: «танки», «артиллерия», «авиация», «пехота», «кавалерия» (рис.1).

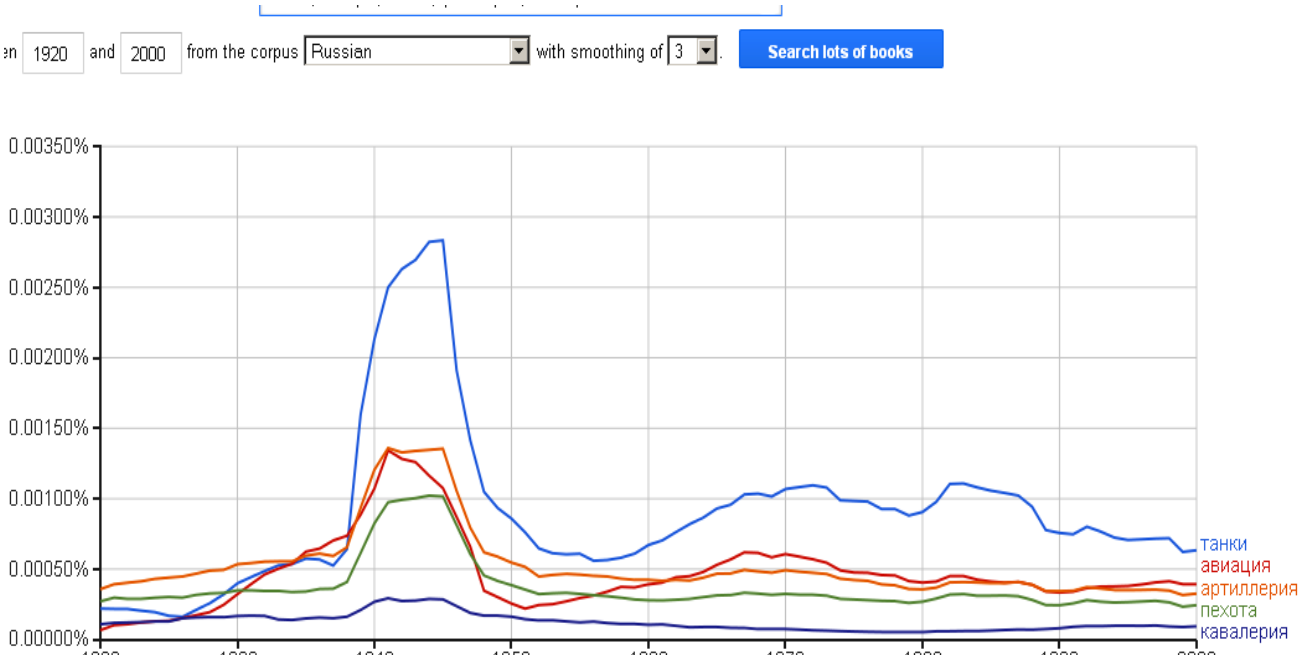

Рис.1. График частоты встречаемости названий родов войск в текстах книг на русском языке

На рис.1 отчетливо видно, что частота употребления названий родов войск в текстах книг в период войны (1941-1945) существенно возрастает по сравнению с периодами мирного времени. Особенно выражен рост употребляемости слова «танки». Значительно

меньше употребляемость слова «кавалерия». Как известно, этот род войск во время Второй мировой войны использовался сравнительно редко.

Необходимо отметить, что слово «танки» может иметь два значения: род войск (напр. «Если пехота или *танки* наступающего прорвут фронт обороны, то *танки* обороняющегося бросаются внезапной контратакой во фланги и в тыл прорвавшемуся противнику»), а также форма множественного числа от слова «танк», боевая машина (*ср. «Танки* стояли хорошо замаскированные, под прикрытием буковой рощи, в русле пересохшей горной реки»). Этим, в частности, может быть объяснена более высокая употребляемость словоформы танки по сравнению с другими терминами, обозначающими роды войск.

Если построить график с двумя кривыми взяв существительное «танк» в единственном числе (в такой форме полисемия снимается), и слово «танки», то, как это отмечает также В.Д. Соловьев [7,8], кривые встречаемости разных грамматических форм одного слова меняются синхронно (рис.2).

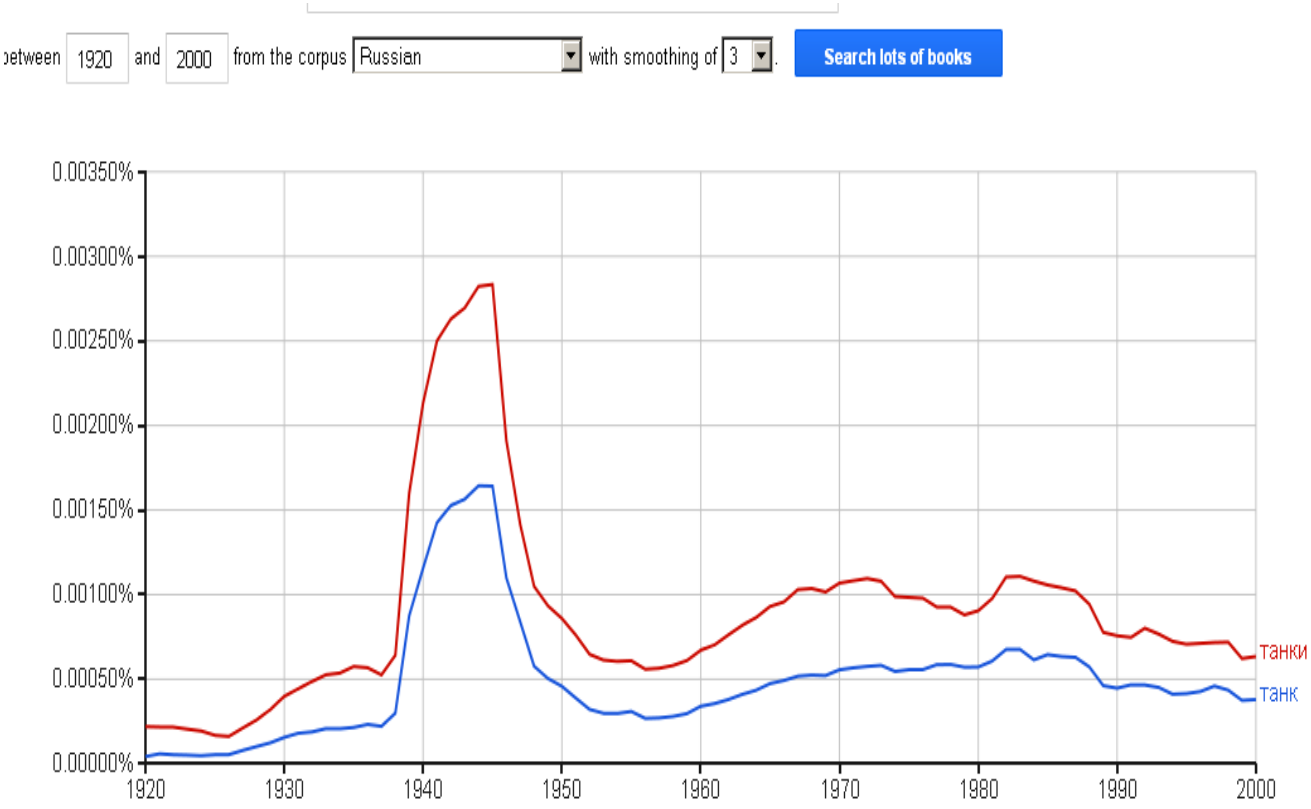

Рис. 2. Кривые встречаемости слов «танк» и «танки»

В контексте нашего исследования оба значения слова «танки» представляется нам условно эквивалентными.

Также допустимо принять как условно эквивалентные слова, находящиеся в отношении гиперонимии - гипонимии («род-вид», «общее-частное») и суммировать графики из таких слов. Тогда частота встречаемости названий родов войск будет такая, как показано на рис. 3, т.е. чаще всего в книгах военных лет говорится об артиллерии и танках.

Рис. 3. Суммарные графики встречаемости названий родов войск авиация, артиллерия и танки.

На рис. 4 построен график из шести кривых, отражающих динамику частотности употребления названий войсковых подразделений.

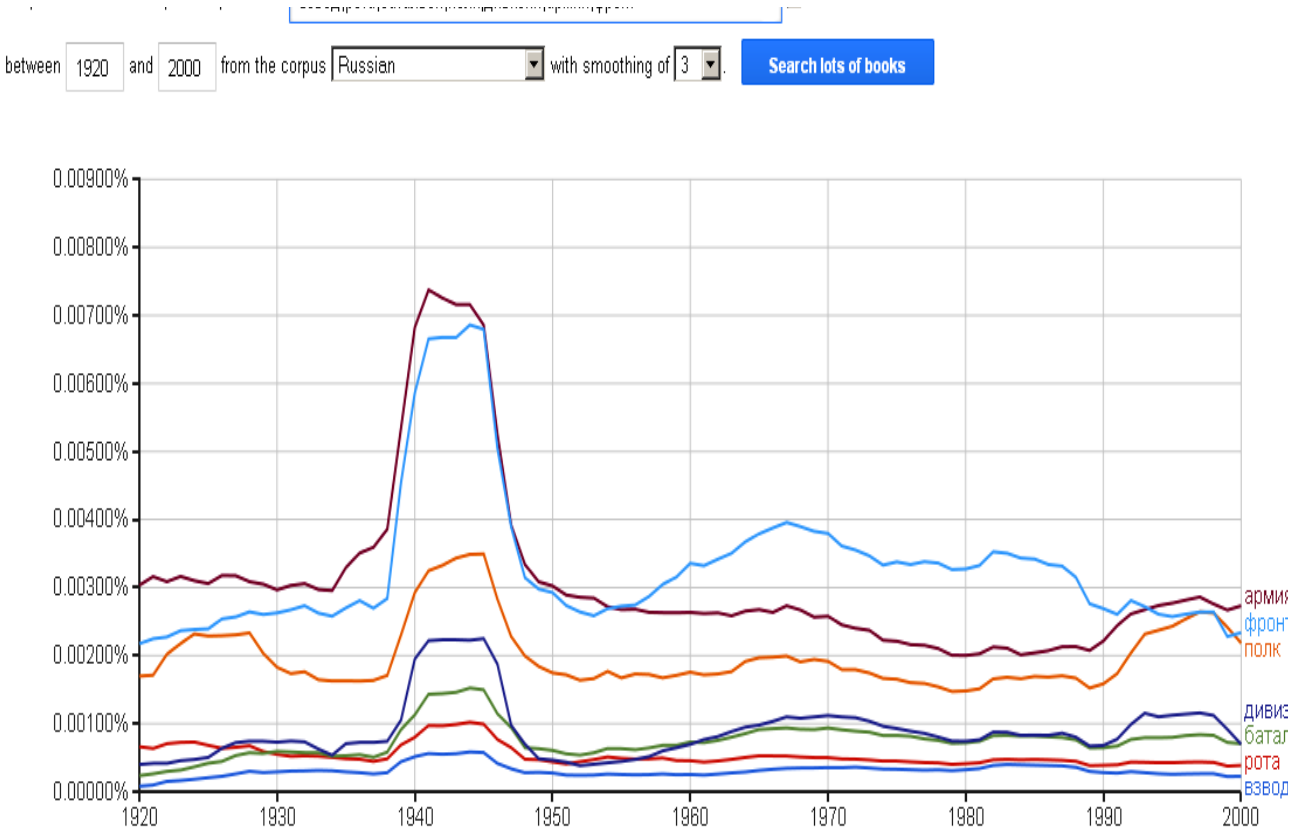

Рис. 4. График частоты встречаемости названий войсковых подразделений

На рисунке можно увидеть две закономерности. Во-первых, как и на предыдущих рисунках, все кривые дают выраженный подъем в военные годы. Во-вторых, на первый взгляд представляется, что в целом в текстах книг, изданных в 1941-45 гг., более частотными являются названия более крупных подразделений (армия, фронт). Заманчиво связать это с тем, что книги чаще описывали стратегические боевые операции. Однако и здесь, вероятнее

всего, причина лингвистического, а не исторического характера. В этих словах присутствует полисемия. Так, слово «армия» помимо названия конкретного войскового формирования (напр. 10-я гвардейская армия) может иметь значение «вооруженные силы страны» (призвать в армию, служить в армии и т.д.), а слово «фронт» помимо названия конкретного войскового формирования, объединения армий (напр., 2-ой Белорусский фронт) может иметь значение места, территории проведения боевых действий (напр., попасть на фронт, уйти на фронт).

Посмотрим наиболее частотные словосочетания с этими словами. Система Google Books Ngram Viewer предоставляет возможность построить графики частотности для десяти самых частотных словосочетаний (биграмм), включающих данное слово в начальной или конечной позиции. На рис.5 представлен график десяти наиболее частотных биграмм, содержащих слово «фронт» в постпозиции.

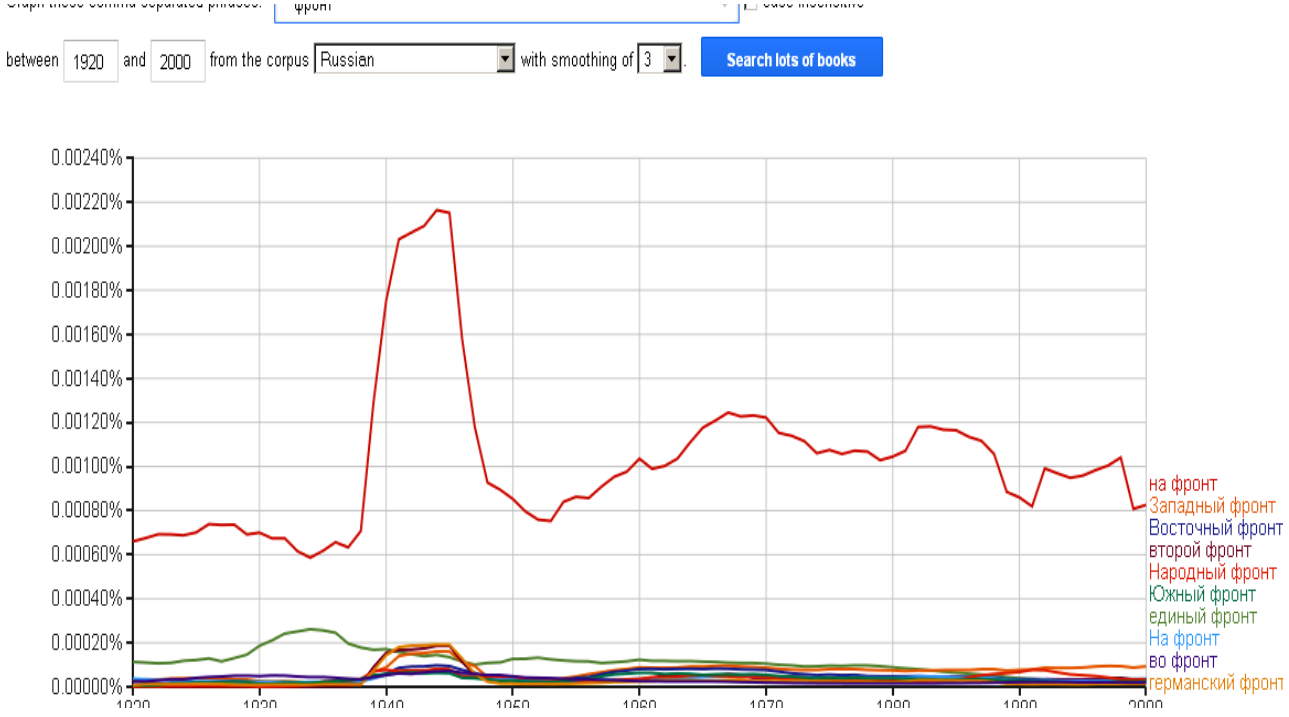

Рис. 5. График частоты встречаемости в текстах книг N-грамм, включающих слово «фронт» в постпозиции

И действительно, более частотным из десяти биграмм является выражение «на фронт».

Из названий военных формирований в число самых частотных биграмм входят Западный фронт, Восточный фронт и Южный фронт. Однако и здесь следует критически подойти к полученным данным. Дело в том, что названия указанных фронтов относятся не ко времени Отечественной войны, а ко времени первой мировой и гражданской войн. Это можно проверить, обратившись собственно к базе данных Google Books (рис. 6, 7).

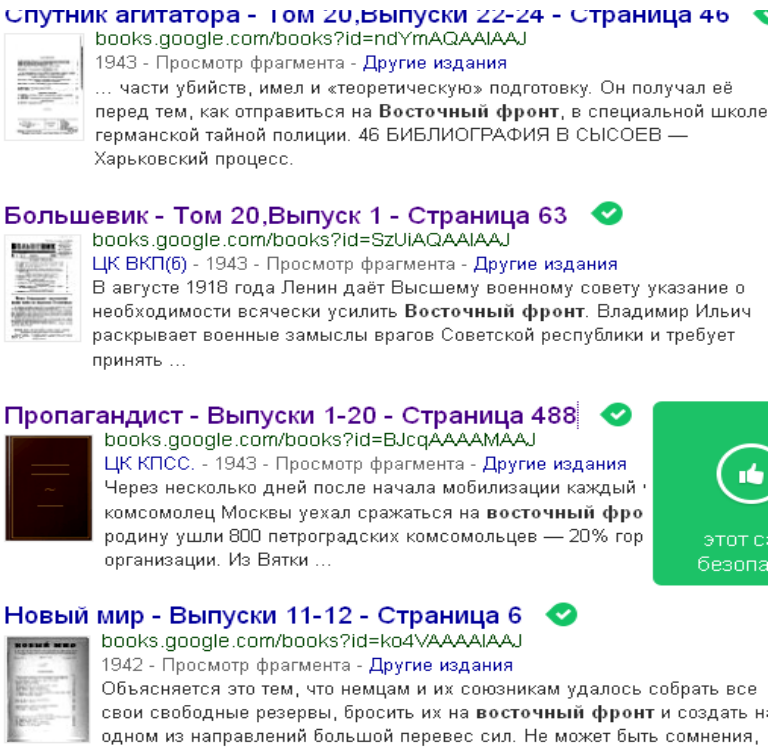

Рис. 6. Страница результатов поиска в Google books по запросу «Восточный фронт»

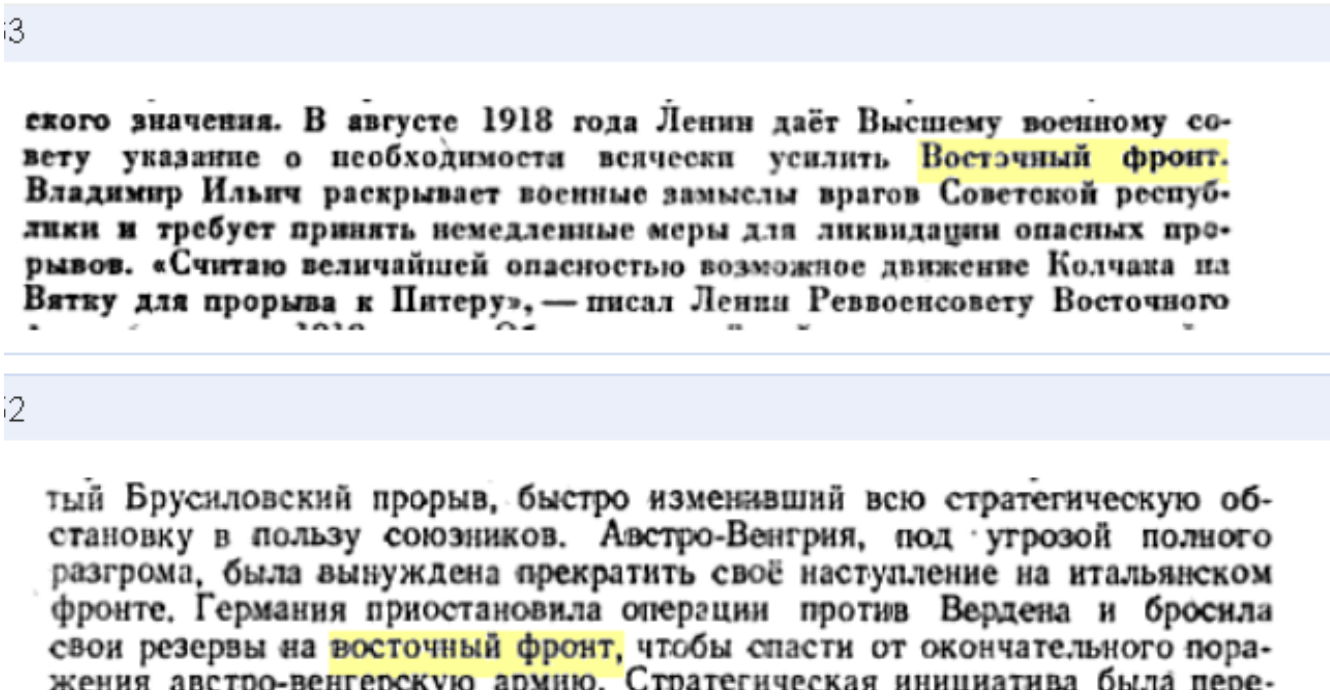

Рис. 7. Фрагменты из сборника «Большевик» 1943 г

Дело, по-видимому, в том, что в годы войны, особенно в ее начальный период, в значительном количестве издавались книги по военной истории. То же подтверждает и анализ употребления имен собственных. В книгах, изданных в годы Великой отечественной войны, очень часто упоминаются Суворов и Кутузов, как герои войн прошедших (рис. 8). И это понятно. Непонятно другое: еще чаще встречается имя Наполеон. При этом частотность

имени Наполеон выше частотности имен русских полководцев. Воздержимся пока от комментариев по этому поводу, хотя можно вспомнить выражение «культ Наполеона в России» или периодически встречающиеся в литературе сопоставления фигур Гитлера и Наполеона.

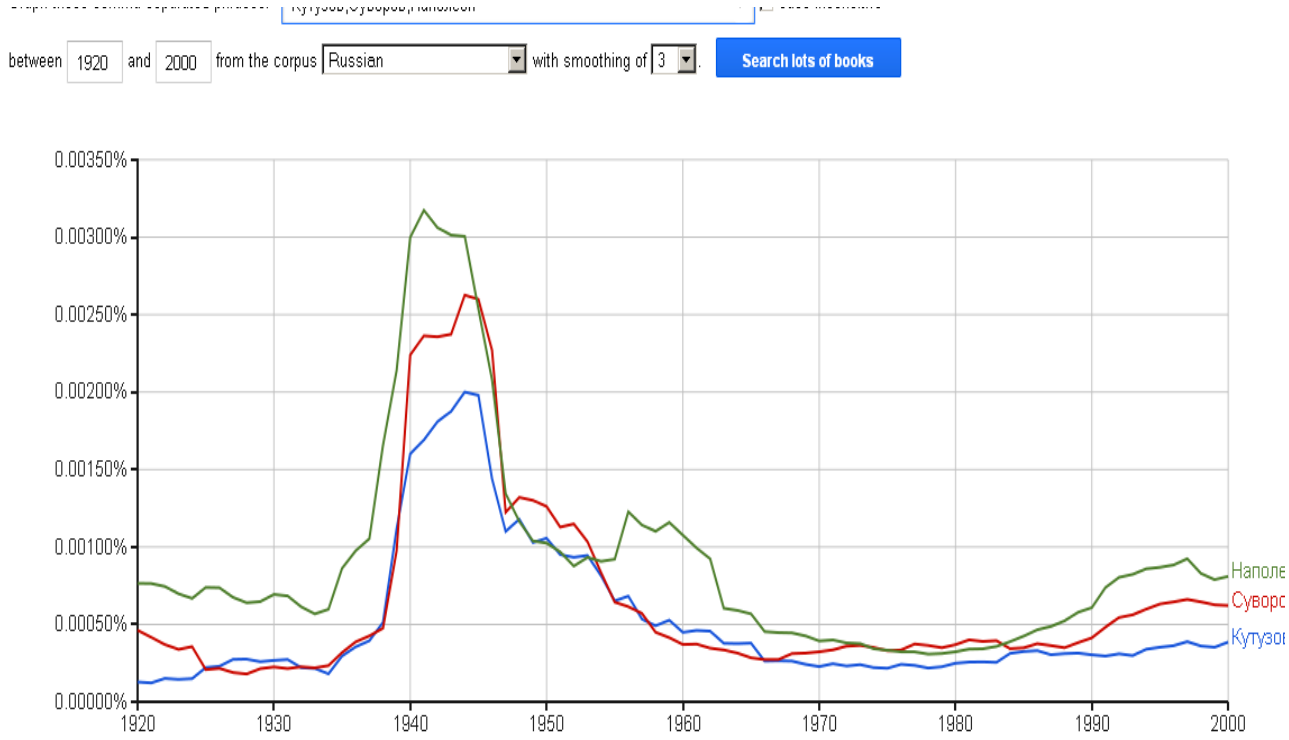

Рис. 8. Динамика встречаемости имен Наполеон, Суворов, Кутузов

Хорошо известно, что любой текст имеет характеристику, называемую предметом или темой. В англоязычной литературе по информатике для нее употребляется другой, более удачный, на наш взгляд, термин «aboutness» (о-чём-ность). Эта характеристика представляет собой одно из ключевых понятий не только в информатике, но и в философии. В подтверждение можно привести следующие цитаты: «В самом деле, в каждом познании объекта имеется единство понятия, которое можно назвать качественным выводом, поскольку под ним подразумевается лишь единство сочетания многообразного в знаниях, каково, например, единство **темы** в драматическом произведении, разговоре, сказке» (И.Кант. «Критика чистого разума»). «То, что мы называем содержанием, значением есть нечто простое внутри себя, сам предмет, сведенный к своим простейшим, хотя и всеохватывающим определениям, в отличие от выполнения. Так, например, можно указать содержание книги в нескольких словах или предложениях, и в книге не должно встречаться ничего другого, кроме того, что мы в общих чертах уже указали в изложении ее содержания. Это **тема,** образующая основу выполнения, есть нечто абстрактное, и лишь выполнение представляет собой нечто конкретное» (Г.-Ф. Гегель «Эстетика»).

Кроме «предмета» - характеристики постоянной и более или менее объективной, текст обладает другой характеристикой - *значением* (meaning) - динамичной, меняющейся в зависимости от модальности текста, от историко-культурного контекста и в большой степени субъективной. Представляется значимым, что корпусы текстов при диахронических исследованиях позволяют выявить влияние на поведение во времени не только лексики, описывающей предмет (aboutness), но и значение (meaning) текстов. Конкретный лексический материал, который мы исследуем, в большей степени, на наш взгляд, отражает именно «meaning». Но это тема отдельного исследования.

Если обратиться к названиям фронтов Великой Отечественной войны, то статистическая картина получается не менее интересная, а именно: эти названия чаще упоминаются после войны, чем во время войны (рис 9).

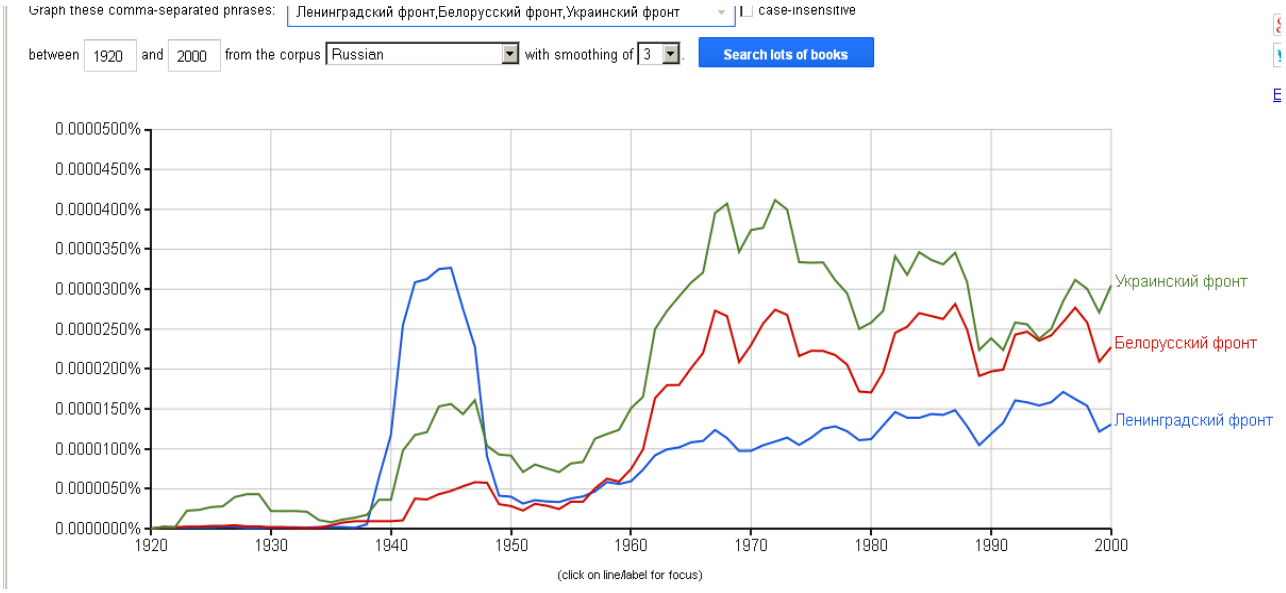

Рис.9. Динамика изменения во времени частотности названий фронтов

Биграмма «Белорусский фронт» охватывает 1-й Белорусский, 2-й Белорусский и 3-й Белорусский фронты, которые были сформированы в 1943-44 гг. и прекратили существование в 1944 – 45 гг., а биграмма «Украинский фронт» - 1-й, 2-й, 3-й и 4-й украинские фронты, которые существовали примерно в те же годы. Ленинградский фронт был сформирован 27.8.1941 и прекратил существование 24.7.1945 [14].

Почему в 1960–1970-е годы отмечается рост частотности упоминания всех фронтов? Понятно, что боевые действия текущего момента войны чаще отражались в газетах, а не в книгах. В то же время в 1960–70-е гг. наблюдается рост публикаций научной и мемуарной литературы, посвященной Великой Отечественной войне, и отсюда рост частоты упоминания фронтовых соединений. Интересно отметить, что «пики» встречаемости этих биграмм, приходятся на 1965, 1975 и 1985 гг., т.е. на годы юбилеев победы.

Следующий график (рис.10) отражает изменение частотности слов, обозначающих воинские звания.

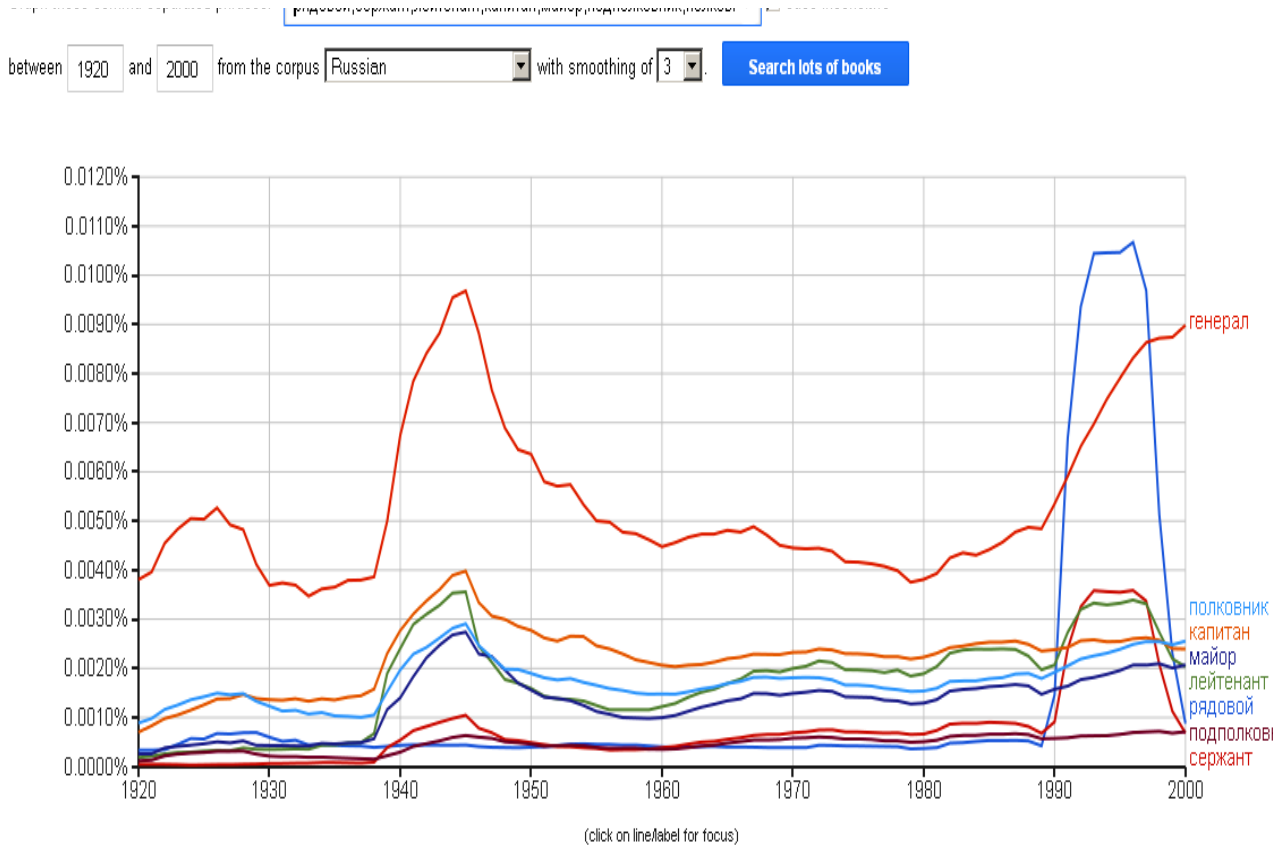

Рис. 10. График частоты встречаемости в текстах книг слов, обозначающих воинские звания

Видно, что в промежутке 1941-45 гг. частота употребления слов распределяется следующим образом: наиболее высокая частота употребления у слова «генерал», затем с очень небольшой разницей идут кривые для офицерских званий (за исключением звания подполковник) и наименьшую частотность имеют звания рядового состава.

График станет наглядней, если использовать имеющуюся в Google Books Ngram Viewer функцию суммирования кривых отдельно для наименований генеральских званий, отдельно для офицерских званий и отдельно для званий рядового состава (рис. 11).

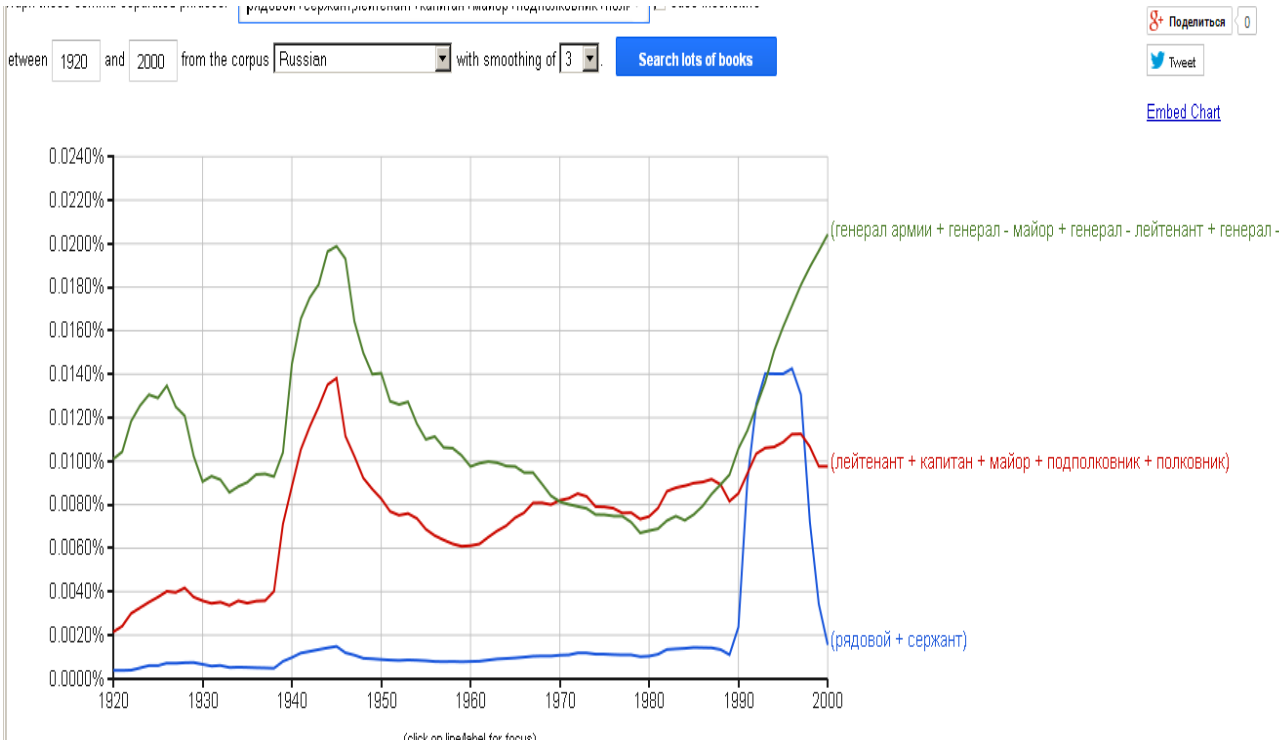

Рис. 11. Суммирование кривых наименований званий по группам званий (генералы, офицеры, рядовой состав)

По частоте употребления «лидируют» генералы, потом идут офицеры, намного опережая рядовой и сержантский состав. Но в 1990-е годы мы вдруг видим пик в использовании слов «рядовой» и «сержант» (рис. 11). Обратившись к текстам корпуса (рис.12, 13, 14), мы находим разгадку.

1920 - 1941	1942 - 1943	1944 - 1968	1969 - 1994	1995 - 2000	
1920 - 1934	1935 - 1942	1943 - 1947	1948 - 1993	1994 - 2000	
1920 - 1941	1942 - 1944	1945 - 1966	1967 - 1994	1995 - 2000	
1920 - 1930	1931 - 1941	1942 - 1945	1946 - 1993	1994 - 2000	
1920 - 1928	1929 - 1940	1941 - 1944	1945 - 1994	1995 - 2000	

Рис.12. Ссылки к результатам поиска в базе данных Google Books

При актуализации ссылки на книги 1995-го года в строке «рядовой» получаем следующую картину:

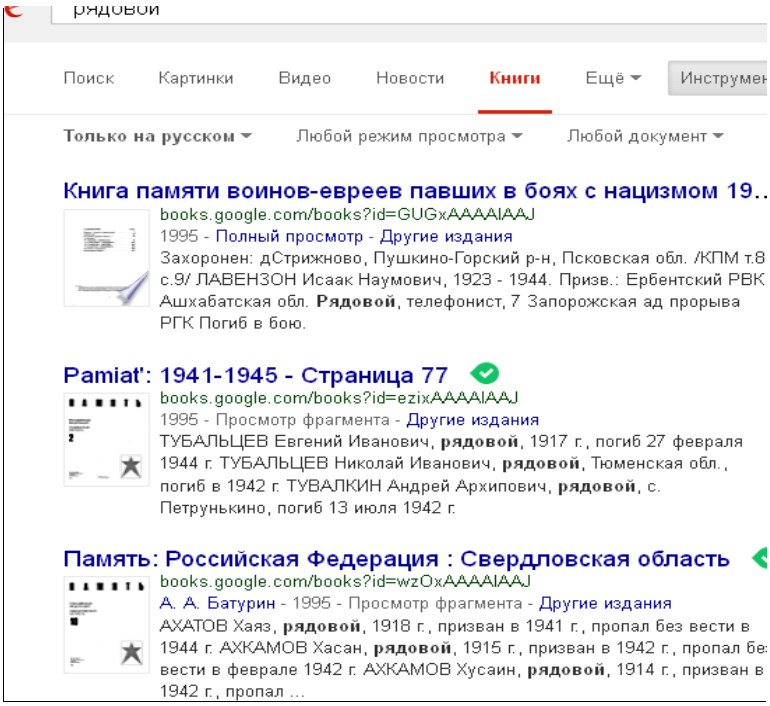

Рис.13. Результаты поиска в базе данных Google Books

Большинство ссылок в результатах поиска по запросу «рядовой» в книгах 1995-го года ведут к книгам памяти, т.е. к спискам погибших и пропавших без вести во время Великой Отечественной Войны, которые издавались в большом количестве в середине 1990-х в связи с 50-летием Победы. И естественно, среди погибших больше всех рядовых и сержантов.

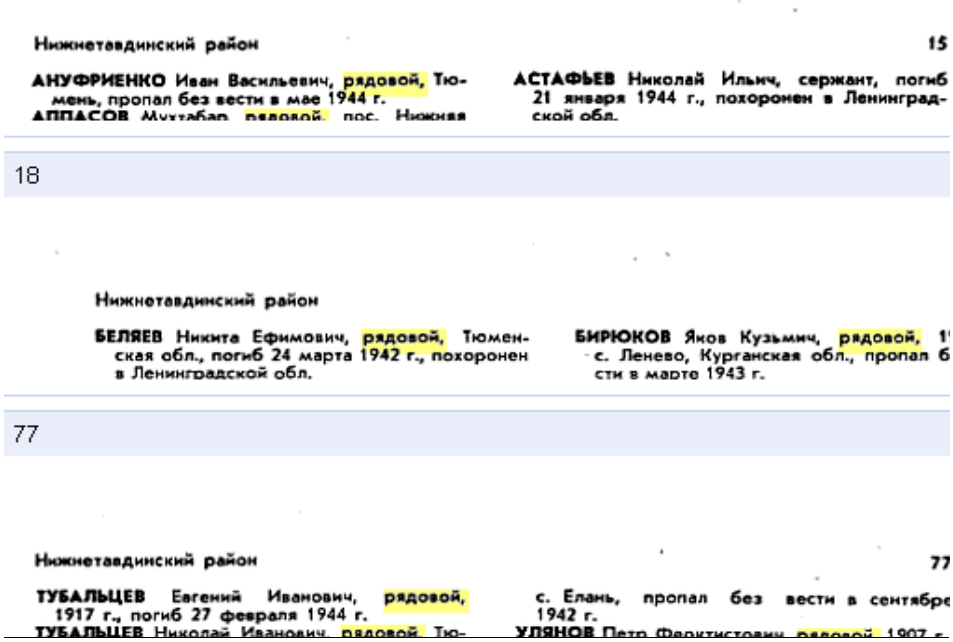

Рис.14. Скриншот страниц текста книг памяти

Из анализируемых текстов через корпус можно получить и другую информацию. Каждая статья книги памяти содержит единообразно организованные сведения о погибших, в т.ч. дату гибели. Если построить графики по цифрам военных лет получаем следующую картину (рис. 15).

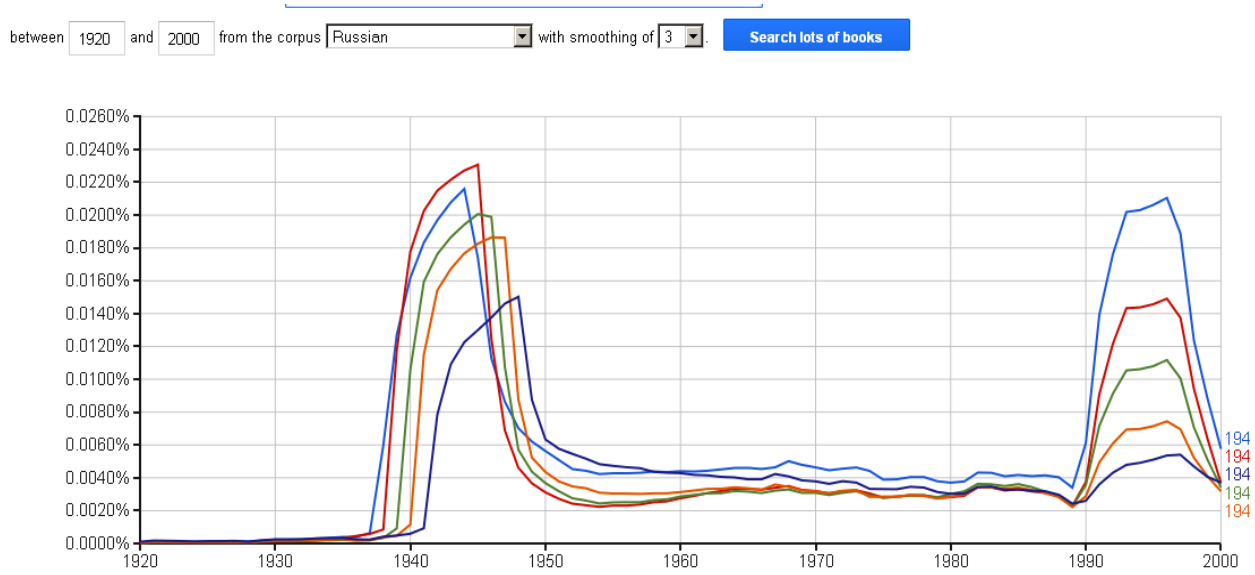

Рис.15. График встречаемости цифр военных лет в текстах книг

Кривые на рис.15 образуют два пика, которые объясняются следующим образом. Пики числа упоминаний года приходится на дату двумя – тремя годами позже самого года. Так, пик упоминаний 1941 года приходится на 1944, 1942 и 1943 на 1945 и т.д. Видимо, это общая закономерность встречаемости цифр, означающих годы, которая не зависит от исторических событий.

Второй пик приходится на середину 1990-х годов, что имеет то же объяснение: в эти годы были изданы книги памяти. На графике отчетливо видно, как с каждым годом число погибших в течение года уменьшалось. Уже в 1943 году число погибших примерно в два раза меньше, чем в 1941.

Заключение

Система Google Books Ngram Viewer, а по-видимому, и 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 и культурологии. Разумеется, предложенная методическая модель подобных исследований нуждается в детальной проработке.

На наш взгляд, проведенное нами исследование отчетливо демонстрирует, что изменение частотности N-грамм в печатных документах связаны с определенным

историческим событием, а также с политическим режимом государства, на территории которого издаются документы, тексты которых образуют корпус.

Нам удалось, как представляется, на основе корпусных данных показать отдельные примеры лингвистических историко-культурных закономерностей. Наиболее остро при этом стоит проблема омонимии, полисемии и синонимии, но и она в первом приближении с определенной степенью погрешности решаема. Наше исследование показывает, что слова со значениями, определяющими семантическое поле интересующего нас понятия, во-первых, имеют сходное поведение, т.е. изменение их частотности происходит по сходной модели, а во вторых, используя операцию сложения Google Books Ngram Viewer, как правило, можно подобрать достаточное число слов из этого поля, так чтобы они достаточно полно отразили лексический образ некоторого понятия. Таким образом, модель изменения частотности лексических единиц во времени на графиках характеризует именно понятия, что позволяет использовать описанную методологию для историко-культурных исследований.

Наконец, иногда корпус даёт неожиданные находки, которые позволяют выявить некоторый аспект исторического события, например, в случае с книгами памяти.

В целом представляется, что настоящая работа позволяет говорить о новом направлении в историко-культурных и лингвистических исследованиях.

Ключевые слова: корпусная лингвистика, Google Books Ngram Viewer, диахронические исследования, историко-культурные исследования

ЛИТЕРАТУРА

1. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. 2-е изд., перераб. и дополн. – СПб.: СПбГУ. РИО. Филологический факультет, 2013 - 148 с.
2. Лотман Ю.М. Символ в системе культуры // Статьи по семиотике и топологии культуры, - Таллин: «Александра», 1992. – Т. 1. – С. 191-199.
3. Culturomics/ Dictionary.com[Электронный ресурс] http://dictionary.reference.com/browse/culturomics (дата обращения 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. Выпуск 10. - СПб.: СПбГУ, 2014. -С. 303-327.
5. Захаров В.П., Масевич А.Ц. Диахронические исследования терминологической лексики // Прикладная лингвистика в науке и образовании: Сб. трудов VII Международной научной конференции 10-12 апреля 2014 г. г. Санкт-Петербург . – СПб: ООО «Книжный знак», 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 всероссийской научно-практической конференции «Электронные ресурсы библиотек, музеев, архивов» (30–31 окт. 2014 г., Санкт-Петербург) / ЦГПБ им. В. В. Маяковского. – СПб., 2014. – С. 43–58 : ил.

7. Соловьев В.Д. Частотно-основанный подход к языковой динамике // Труды междунар. конференции «Корпусная лингвистика-2013». -СПб., 2013.- С. 424-431.
8. Соловьев В.Д. Частотность как объект корпусных исследований // Труды междунар. конференции «Корпусная лингвистика-2011». - СПб., 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. – Vol. 19, №:3. – P. 401-416.
11. Mann J. et al. Enhanced Search with Wildcards and Morphological Inflections in the Google Books Ngram Viewer // Proceedings of ACL Demonstrations Track Association for Computational Linguistics 2014. [Электронный ресурс] http://www.dipanjandas.com/files/acl2014ngrams.pdf (дата обращения 12.07.2015)
12. Michel J.-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science. // Science. – 2011. - 14 January 2011. – P. 176-182 [Электронный ресурс] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3279742/ (дата обращения 12.07.2015)
13. Google books Ngram Viewer [Электронный ресурс] https://books.google.com/ngrams (дата обращения 14.07.2015).
14. Список фронтов вооружённых сил РККА (1941—1945) [Электронный ресурс] https://ru.wikipedia.org/wiki/%D0%A1%D0%BF%D0%B8%D1%81%D0%BE%D0%BA_%D1%84%D1%80%D0%BE%D0%BD%D1%82%D0%BE%D0%B2_%D0%B2%D0%BE%D0%BE%D1%80%D1%83%D0%B6%D1%91%D0%BD%D0%BD%D1%8B%D1%85_%D1%81%D0%B8%D0%BB_%D0%A0%D0%9A%D0%9A%D0%90_%281941%E2%80%941945%29 (дата обращения 12.07.2015)