

Вариативность представления имен политических деятелей при диахронических исследованиях на основе корпусов текстов

А.Ц. Масевич¹, В.П. Захаров²

¹ Санкт-Петербургский государственный институт культуры

² Санкт-Петербургский государственный университет

andmasev@mail.ru, v.zakharov@spbu.ru

Настоящая публикация продолжает ряд публикаций, посвященных диахроническим исследованиям частотного поведения политической лексики в текстах книг. В определенной степени она носит методический характер. В ней рассматривается, каким образом форма представления собственного имени и статуса политического деятеля отражаются на частотном поведении этой лексики и как она должна быть учтена при осуществлении исследований.

Ключевые слова: корпусная лингвистика, имена политических деятелей, частотность, вариативность, система Google books Ngram Viewer

1. Введение

Настоящая публикация продолжает ряд публикаций, посвященных диахроническим исследованиям частотного поведения политической лексики в текстах книг. В предыдущих публикациях [1–3] мы показали на нескольких примерах, как изменение частоты встречаемости лексических единиц в текстах книг отражает реальные историко-культурные процессы, и описали несколько моделей частотного поведения единиц политической лексики [4, 5]. Настоящая публикация носит в определенной степени методический характер. В ней рассматривается, каким образом форма представления собственного имени и статуса политического деятеля отражаются на частотном поведении этой лексики и как она должна быть учтена при осуществлении исследований.

В любой поисковой системе стоит задача точной формулировки информационной потребности, т. е. перевода содержательного пользовательского запроса, сформулированного на естественном языке, в поисковое предписание на запросном языке данной системы. В случае тематических запросов это далеко не тривиальная задача. В современных поисковых системах бестезаурусного типа содержание запросов, так же, как и содержание документов, выражается с помощью слов естественного языка. Однако понятия, лежащие в основе тематического запроса, в языке могут выражаться с помощью различных слов и словосочетаний. Поскольку тема запроса и ее аспекты суть имена понятий, и мы не знаем, каким способом это понятие будет выражено в искомых документах, то необходимо в запросе «развернуть» все гнездо близких по смыслу слов и словосочетаний, описывающих это понятие (синонимичные выражения, видовые термины и т. п.). Когда объектом поиска являются имена собственные, то задача формулирования запроса, казалось бы, не представляет особой трудности. На самом деле это не совсем так. Проблемой представления имен давно занимаются специалисты по информационным, в особенности библиотечным системам. В электронных каталогах крупных библиотек существуют системы авторитетного контроля, учитывающие вариативность элементов библиографического описания, в том числе личных имен [6–8].

При поиске по именам собственным следует учитывать омонимию имен и фамилий, а также разные способы именованья персон (Сталин, товарищ Сталин, И.В. Сталин, Иосиф Виссарионович Сталин и т. п.), причем эти способы могут зависеть от страны, от соответствующего исторического периода или вообще быть сугубо индивидуальны. Эти вариации и исследуются в данной статье.

Наше исследование проводилось на основе трех корпусов системы Google Books Ngram Viewer — русского, британского английского и американского английского.

Система Google Books Ngram Viewer [9] детально описана нами в нескольких предыдущих публикациях [5, 10]. Она позволяет строить графики встречаемости слов и коллокаций за выбранный временной период. На горизонтальной оси графика показываются годы, входящие в заданный временной период. По вертикальной оси откладывается выраженная в процентах относительная частота встречаемости в корпусе заданной N-граммы (от одного слова до пяти) в соответствующем году. 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, результат умножается на 100 [10, с. 307].

При построении графика имеется возможность задать ряд условий, позволяющих делать графики более наглядными, сопоставлять кривые поведения нескольких N-грамм, выявлять наиболее частотные словосочетания с данной словоформой, их поведение во времени и др.

2. Методологические замечания

Значение относительной частоты встречаемости некоторой лексической единицы за данный год зависит от числа текстов, изданных в этом году, которые введены в базу данных Google Books. Здесь перед нами встают методологические трудности. Во-первых, мы не знаем, какая часть опубликованных в данном году текстов введена в базу данных, нам известно только общее число текстов — примерно 590 тыс. текстов для русского корпуса. Во-вторых, мы не знаем, какая часть опубликованных текстов должна быть введена в базу данных, чтобы исследование было статистически достоверным.

Тем не менее, есть уверенность, что при аккуратно выстроенной методике исследования результатам, полученным на данных системы Google Books Ngram Viewer, можно доверять. В подтверждение сказанного приведем несколько соображений.

1. Система предполагает определенный «уровень достоверности» данных за счет того, что графики могут быть построены лишь для тех N-грамм, которые встречаются в корпусе не менее 40 раз [10, с. 307].

2. В системе имеется механизм сглаживания [10, с. 307–308], позволяющий нивелировать годовые колебания в наполнении корпуса.

3. Результаты, полученные на данном корпусе другими исследователями, такими, в частности, как Ю.С. Масленникова, В.В. Бочкарев, В.Д. Соловьев, Т.И. Галеев [11–13] хорошо коррелируют с данными лингвистической науки.

4. Данное исследование, является частью большой работы, начатой авторами три года назад. В ходе этой работы построено около 500 графиков динамики частотного поведения различных слов и коллокаций, относящихся к политической лексике. Этот материал доступен в Интернет [14]. И полученные результаты показывают очевидные корреляции с реальными историческими событиями.

5. Мы провели сравнение частотного поведения лексических единиц в Google Books Ngram Viewer и в Национальном корпусе русского языка (НКРЯ), сервис «Графики» (<http://ruscorpora.ru/ngram.html>). Сравнение графиков, построенных разными системами для десяти частотных существительных русского языка (*год, человек, время, дело, жизнь, день, рука, раз, работа, слово*) [15] показало, что соотношение частотности разных единиц и модели их частотного поведения в большинстве случаев сходны или отличаются незначительно. Затем такое же сравнение было проведено и на некоторых терминах

политической лексики. Имеются случаи и существенных различий, но они, как правило, легко объяснимы: это случаи редких лексических единиц, когда объем НКРЯ (76 тыс. текстов) оказывается явно мал.

Таким образом, мы утверждаем, что частотное поведение лексики по данным системе Google Books Ngram Viewer является достоверным индикатором культурно-исторических явлений и процессов, хотя и нуждается в дополнительных методологических исследованиях.

3. Вариативность имен политических деятелей в разрезе хронологии и географии

Рассмотрим варианты именованя монархов Российской империи в пред- и послереволюционные годы (1820–1920). При этом, разумеется, была учтена дореволюционная орфография [10].

На рисунке 1 видно, что более частотными способами обозначения особы монарха были выражения «Государь», «Императорь», «Его Величество».



Рис. 1. Сопоставление изменения частоты слов «Государь» и «Императорь» и имен пяти императоров за период с 1820 г. по 1920 г.

На рис. 1 видно, что при сопоставлении кривых для слов «Государь» и «Императорь» и кривых имен императоров последние практически сливаются с горизонтальной осью, т.е. чаще указание на императора шло без имени.

Система Google Books Ngram Viewer позволяет выявлять для каждого выражения 10 наиболее частотных словосочетаний и строить графики их встречаемости за определенный период. При этом существует возможность задать часть речи в правой или левой позиции (рис. 2).

Таким образом, были выявлены наборы глаголов наиболее употребительных с данными выражениями. Во всех трех случаях это, в основном, глаголы в прошедшем времени, такие как «изволил», «отправился», «возвратился», «благоволил», «занимался», «изъявил», «пожаловал», «послал», «согласился», «утвердил», «приказал», реже употребляются формы настоящего времени «изволит», используется также инфинитив «повелеть [соизволил]).

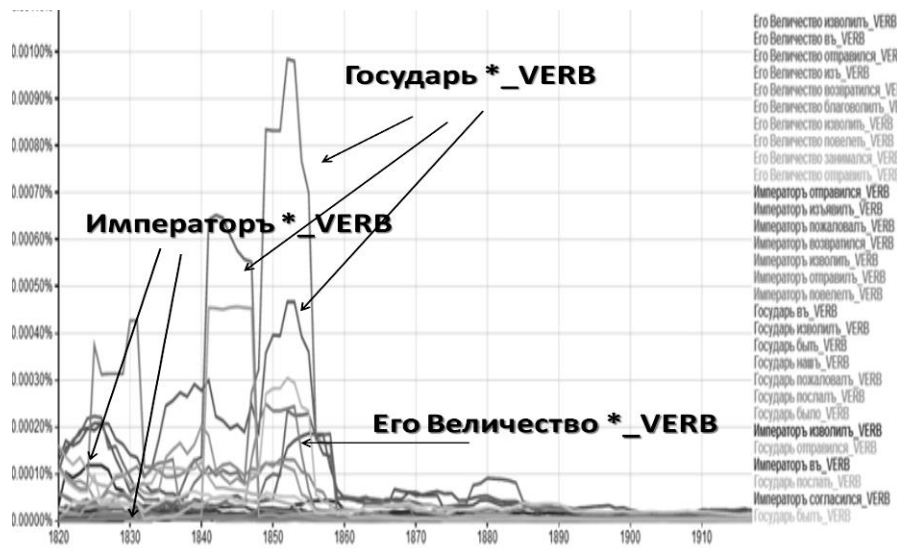


Рис. 2. Динамика частотного поведения выражений «Государь», «Его Величество» и «Императоръ» с глаголами справа (период 1820 по 1917 г.г.)

Может возникнуть опасение, что частота употребления слова «государь» завышена за счет часто используемого в XIX веке светского обращения «милостивый государь».



Рис. 3. Частотное поведение N-грамм «Государь» и «Милостивый Государь» с 1820 г.

Однако, как видно на рис. 3, частотность выражения «милостивый государь» настолько ниже частотности слова «Государь», что никак не искажает график последнего.

Рассмотрим частотное поведение имен руководителей СССР в книгах, зафиксированных в Google books Ngram Viewer (рис. 4).

Пики числа упоминаний трех руководителей СССР (Сталин, Хрущев, Брежнев) всегда приходится на время их правления. Иначе выглядит кривая имени «Ленин». Этот политик посмертно стал сакральным символом коммунистической идеи и в этом качестве упоминается в текстах книг. Для имен трех советских политиков, долгое время занимавших пост руководителя страны, форма кривой частотного поведения имеет определенные черты сходства. Подъем, достижение пика, затем спад и период забвения, продолжительность которого зависит от того, сколько прошло времени с момента прекращения политической

деятельности до середины восьмидесятых годов двадцатого века. В случае Сталина можно говорить не столько о пике, сколько о плато продолжительностью почти десять лет. Эта модель ранее описана нами в [5].



Рис. 4. Частотное поведение имен руководителей СССР (1920–2000)

Благоприятным для нашего исследования является то обстоятельство, что фамилии политических лидеров страны хоть и не являются редкими, но не слишком распространены. Подобное исследование с именем, например, Медведев, было бы невозможным.



Рис. 5. Частотное поведение имен руководителей СССР и РФ (1950–2005)

На рис. 5 показаны кривые частотного поведения имён последних пяти руководителей СССР и первых двух руководителей РФ. Обращает внимание, что кривые имен «Хрущёв», «Брежнев», «Андропов», «Черненко» следуют описанной нами выше модели, которую условно назовём «Сталинской». Модель для имен «Горбачёв» и «Ельцин» не имеет резкого спада после завершения политической деятельности лица. Рассмотрим далее формы

представления трех политиков СССР и двух политиков России. Для этого был использован тег формирования десяти самых частотных биграмм с фамилией политика и существительным в левой позиции. Из полученных наборов биграмм мы отобрали те, кривые которых различимы на рисунках.

Самой частотной формой представления И.В. Сталина в текстах печатных документов являлось выражение «товарищ Сталин» (рис. 6). Заглавная «Т», вероятно, не носит специального характера, а связана с позицией биграмм в предложении. Значительно реже использовалась форма с полным именем и отчеством.



Рис. 6. Три наиболее частотных биграммы с именем «Сталин»

Н.С. Хрущева, как видно на рис. 7, чаще всего обозначали полным именем. Варианты выражения «товарищ Хрущев» встречаются реже. Отметим также, что в конце девяностых появляется форма «Никита Хрущев»

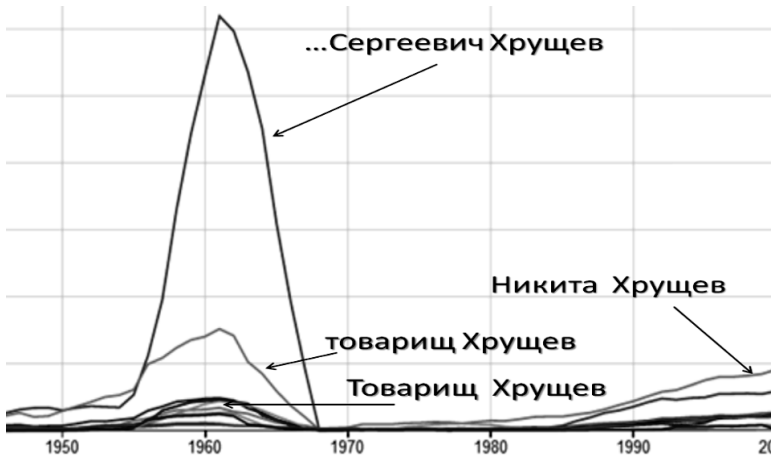


Рис. 7. Четыре наиболее биграммы с именем «Хрущев»

Биграмма «секретарь Брежнев», очевидно, является частью выражения «Генеральный секретарь Брежнев» (рис. 8). В отношении же Сталина и Хрущёва среди самых частотных биграмм указание на занимаемую должность отсутствует.



Рис. 8. Четыре наиболее частотных биграмм с именем «Брежнев»

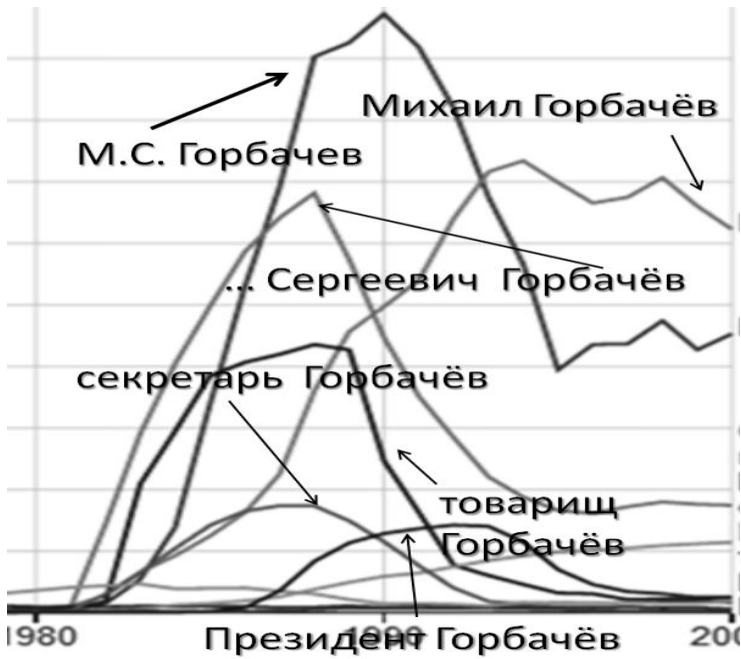


Рис. 9. Наиболее частотные существительные с именем «Горбачев»

Что касается первого и последнего президента СССР, то у него наиболее частотна форма с инициалами. Ее употребляемость выше, чем формы с полным именем. Появляется форма «президент Горбачев». При этом достаточно частотны биграммы «секретарь Горбачев» и «товарищ Горбачев» (рис. 9).

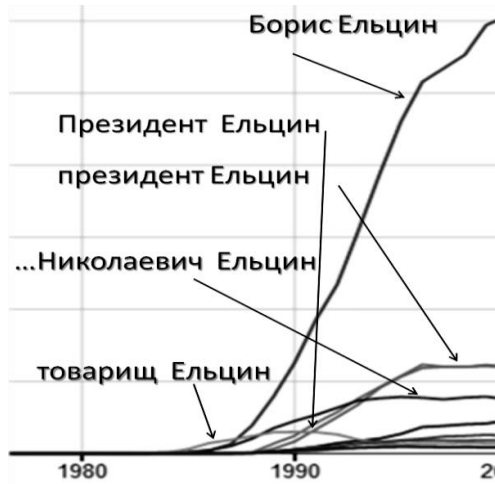


Рис. 10. Наиболее частотные биграммы с именем «Ельцин»

Интересно, что Б.Н. Ельцин чаще всего обозначается как «Борис Ельцин» (рис. 10). Кривые биграмм «Президент Ельцин» и «президент Ельцин» практически сливаются. Вполне различима кривая биграммы «товарищ Ельцин».

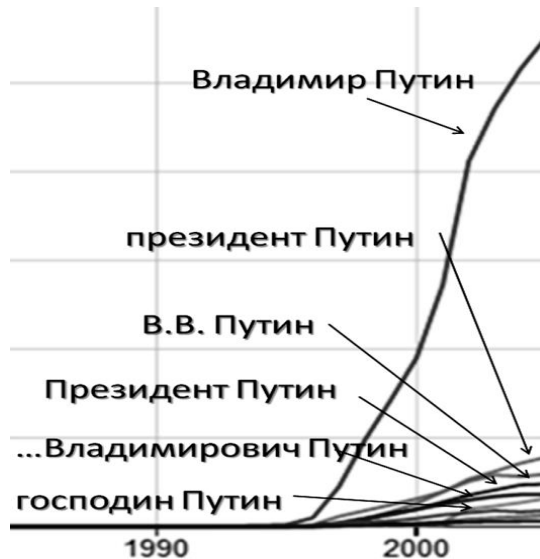


Рис. 11. Наиболее частотные биграммы с именем «Путин»

Полноценно проследить частотное поведение биграмм с именем «Путин» невозможно, поскольку корпус охватывает временной период только до 2008 года. Из рис. 11 видно, что наиболее частотной является форма «Владимир Путин», значения частотности прочих форм очень близки между собой. И, наверно, не будет ошибкой сказать, что именование публичных деятелей по имени и фамилии является особенностью русского языка нового времени.

Покажем одну особенность функционирования имен политических деятелей в функции субъектов в структуре предложения, а именно, сочетаемость их с глаголами (рис. 12).



Рис. 12. Сопоставление частотного поведения N-грамм имен глав СССР с глаголами в правой позиции словоформы «Государь» с глаголами в правой позиции

Для графика на рис. 12 были применены теги для формирования наборов биграмм с глаголами в правой позиции. В каждом из пяти образовавшихся наборов выбрана одна наиболее частотная N-грамма. Самый частотный глагол для «Государя» — «извоиль», который предполагает после себя другой глагол в инфинитиве, напр., «извоиль повелеть». Это была, очевидно, официальная словесная формула в Российской империи. Самый частотная биграмма с именем «Ленин» — «Ленин писал». Это легко объясняется — Ленин упоминается не как политик, а как теоретик, непререкаемый авторитет и, кроме того, как сакральный символ, и ссылка делается на его письменное наследие. Что касается трех других советских лидеров, то высокая частотность их имен приходится на периоды правления, и в текстах приводятся ссылки на их выступления — отсюда глагол «говорил».

Рассмотрим далее представление и частотное поведение имен зарубежных лидеров в корпусах американского английского и британского английского языков.

На рис. 13–15 показаны графики частотного поведения имен президентов США в разном представлении.

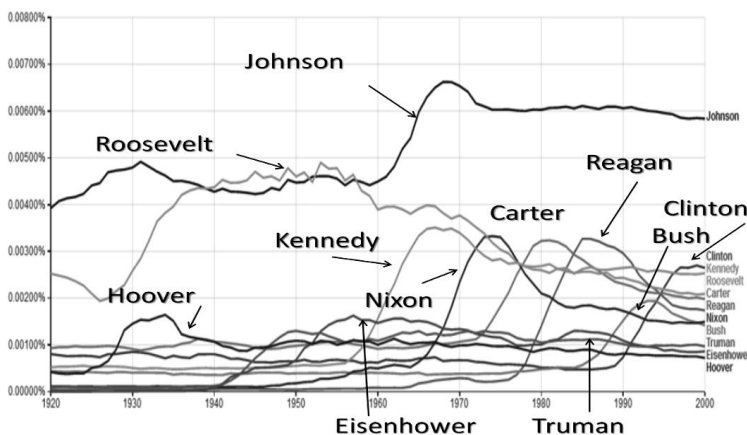


Рис. 13. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление — фамилия)

При таком представлении, безусловно, повышается полнота поиска, однако согласно известной закономерности информационного поиска снижается его точность. Так, фамилия президента, заменившего в 1963 году убитого Джона Кеннеди, Джонсон — одна из самых

распространенных фамилий в англоязычных странах. На рис. 14 видно, что фамилия «Johnson» имеет самую высокую частотность, и соответствующая кривая имеет сравнительно мало подъёмов и снижений. Тем не менее, во время президентства Линдона Джонсона (1963–1969) кривая его фамилии дает выраженный подъём.

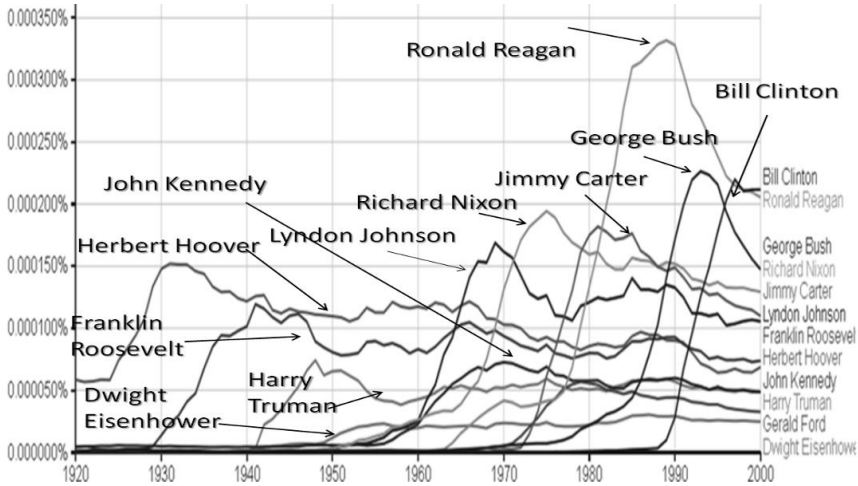


Рис. 14. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление — имя, фамилия)

Однако для сравнения частотности упоминаний каждого президента целесообразно выявить данные, наиболее точно им соответствующие. Кривые на рис. 14 в основном изоморфны кривым на рис. 13, но при этом меняется соотношение между ними. Так на рис. 13 пик кривой фамилии Reagan находится примерно на одном уровне с пиками кривых фамилий Kennedy, Nixon, Carter, а пик фамилии Bush значительно ниже этого уровня. На рис. 14 пик кривой имени Ronald Reagan значительно выше пиков кривых Richard Nixon и Jimmy Carter. Пик кривой George Bush также выше этих двух пиков.

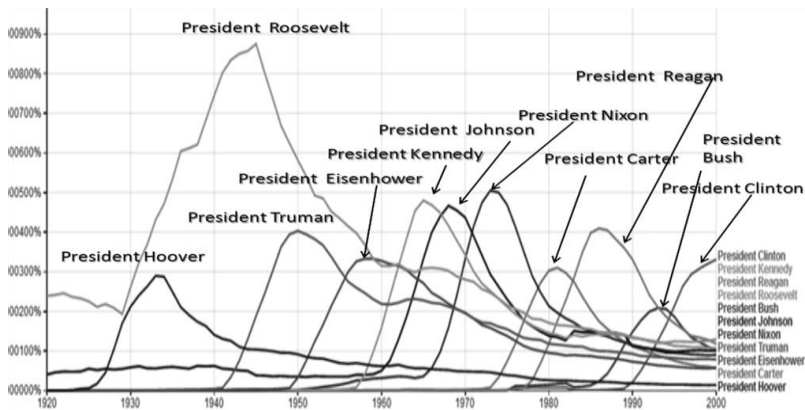


Рис. 15. Частотное поведение имен президентов США в корпусе американского английского языка Google Books (представление – статус, фамилия)

У большей части кривых после окончания срока президентства соответствующих персон отмечается некоторое снижение, которое в некоторых случаях выражено сильнее (Ronald

Reagan, George Bush), а в других слабее (Herbert Hoover, Franklin Roosevelt, Harry Truman, Dwight Eisenhower). Заметим, что более резкое снижение кривых характерно для исторически более поздних президентов.

Сочетание слова «President» (NB: с заглавной буквы) и фамилии лица дает серию изоморфных кривых, подъёмы, пики и снижения которых соответствуют срокам президентства (рис. 15). В отличие от рис. 14 снижения более выражены. По-видимому, в американской традиции слово President с заглавной буквы используется только в отношении действующего президента. На данном графике отчетливо видно, что пик кривой President Roosevelt намного выше, чем остальных кривых.

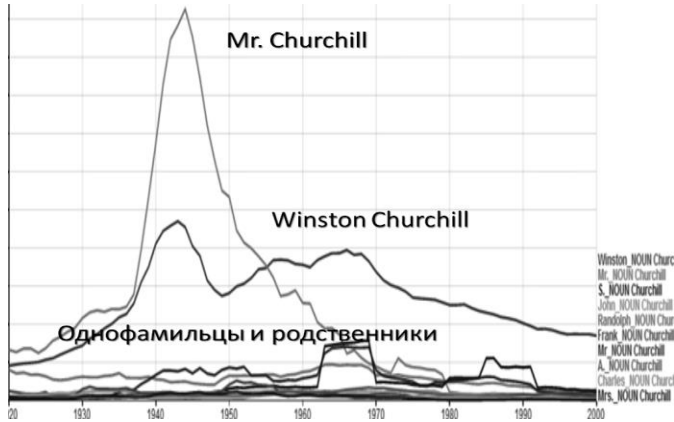


Рис. 16. Десять наиболее частотных биграмм с именем «Churchill» и существительными в левой позиции в корпусе британского английского

На рис. 16 показаны кривые частотного поведения биграмм с фамилией Churchill. Обращает внимание пик в середине сороковых годов биграммы Mr. Churchill. Высокий уровень такой формы представления вполне объясним. Черчилль был в это время премьер-министром. И это был разгар войны. Известно, что королева Елизавета II посвятила в рыцари премьер министра Черчилля во время второго срока его премьерства в 1953. Это обстоятельство отразилось в поведении кривой «Sir Winston Churchill» (рис. 17).

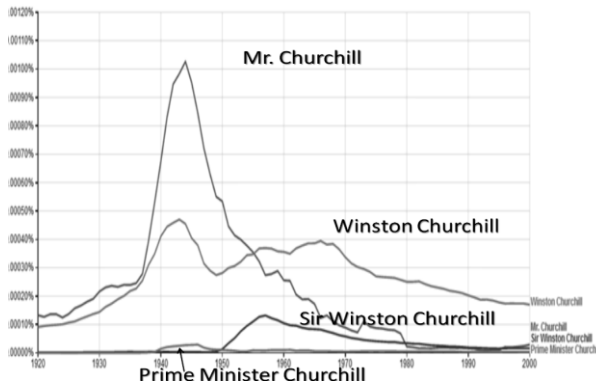


Рис. 17. Четыре варианта обозначения Уинстона Черчилля в корпусе британского английского языка

В число отобранных системой биграмм на рис. 16 вошли имена также предка премьер-министра первого герцога Мальборо Джона Черчилля (John Churchill), сына премьер-

министра Рандольфа Черчилля (Randolph Churchill), американского композитора Фрэнка Черчилля (Franc Churchill), а возможно и другого Фрэнка Черчилля — персонажа романа Джейн Остин «Эмма», британского поэта сатирика XVIII века Чарльза Черчилля (Charles Churchill) и, наконец, супруги премьер-министра, кривая которой соответствует сочетанию «Mrs Churchill». Все эти кривые, однако, плохо различимы, т.к. расположены в нижней части графика, ближе к горизонтальной оси.

Пики кривых «Prime Minister Churchill» и «Sir Winston Churchill» значительно ниже пиков остальных двух. Кривая «Winston Churchill», на наш взгляд, наиболее точно отражает биографию этого политического деятеля. Снижение упоминаний в середине 1940-х, по-видимому, соответствует поражению на выборах в 1946 году, а последующий подъем - второму сроку пребывания на посту.

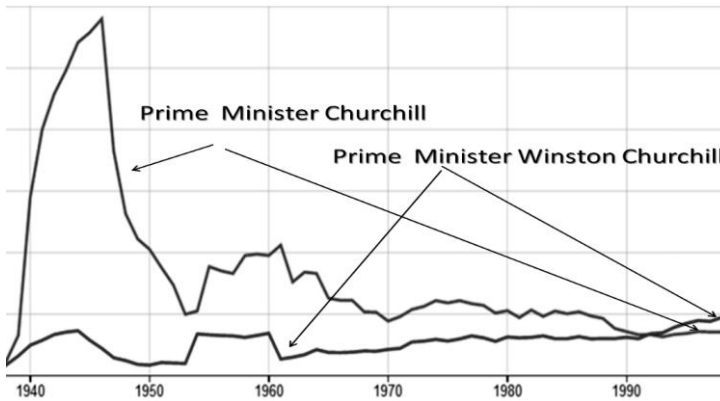


Рис. 18. Сопоставление частотного поведения N-грамм «Prime Minister Churchill» и «Prime Minister Winston Churchill»

Кривые N-грамм, содержащих название должности и имя лица в двух вариантах сходны по конфигурации с кривой «Winston Churchill» (рис. 17). В кривой «Prime Minister Churchill» подъемы и снижения частотности выражения выражены сильнее, чем в кривой «Winston Churchill», что, по-видимому, связано с включением в N-грамму названия должности, ведь именно с ней и связаны колебания кривой. Еще одно наблюдение состоит в том, что практически на всем протяжении выбранного периода кривая «Prime Minister Churchill» выше, чем кривая «Prime Minister Winston Churchill» и только в середине 1990-х частотность N-граммы с именем и фамилией становится чуть выше, чем кривая только с фамилией.

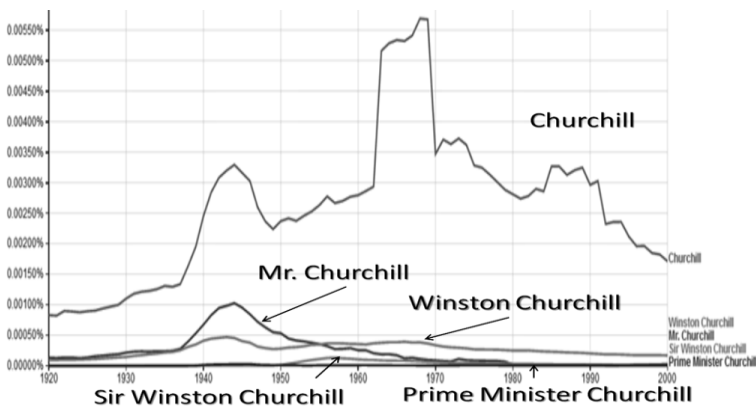


Рис. 19. Сопоставление кривых разных вариантов представления лица с кривой фамилии Churchill.

Понятно, что частотность фамилии «Churchill» значительно выше частотности всех форм (рис. 19). Однако заметим, что поведение кривой «Churchill» в целом, особенно до 1960 г., сходно с поведением кривых для других обозначений этого лица.

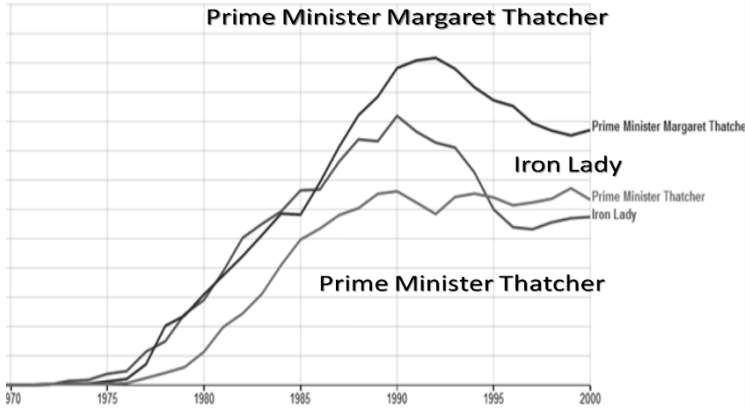


Рис. 20. Частотное поведение N-грамм с указанием должности с полным именем должности с фамилией, а также прозвища

Рассмотрим теперь несколько графиков частотного поведения имени другого британского премьер-министра — Маргарет Тэтчер (рис. 20–22).

Частотность форм обозначения лица в данном случае отличается от случая с Черчиллем, а именно, статус и полное имя более частотно, чем статус и фамилия, а частотность прозвища также довольно высока, причем частотность его снижается после прекращения полномочий и это более выражено, чем двух других N-грамм (рис. 20).

Кривые на рис. 21 показывают, что частотность формы «Mrs Thatcher» значительно выше других биграмм. Это, видимо, особенность британского английского языка или британского политического лексикона. Об этом же говорят и графики на рис. 22.

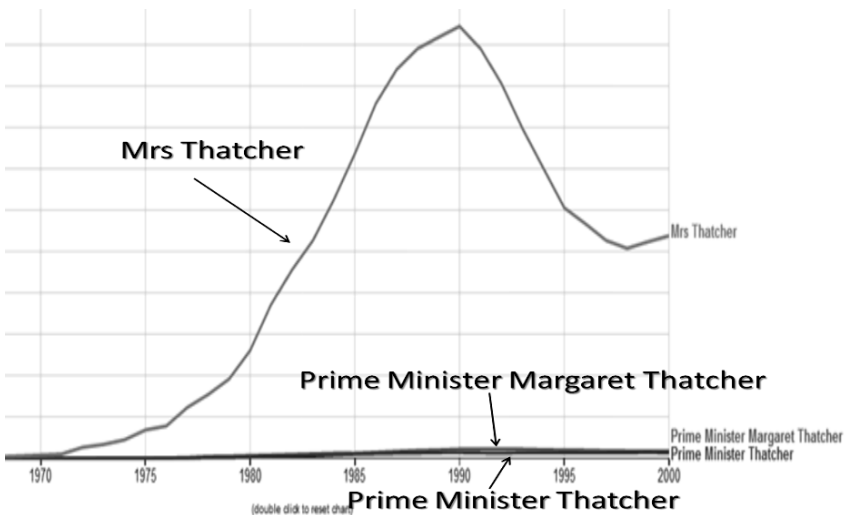


Рис. 21. Частотное поведение N-грамм с указанием должности с полным именем должности с фамилией, а также формы «Mrs Thatcher»

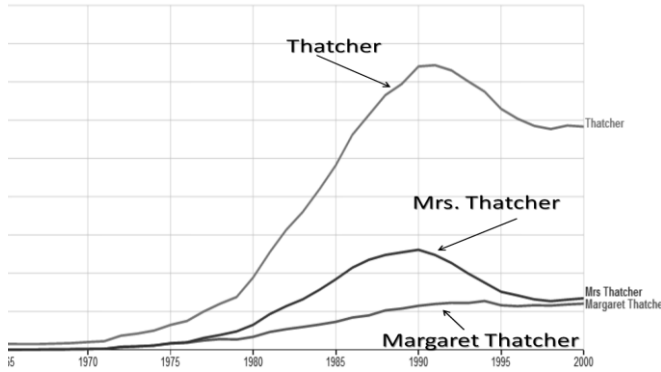


Рис. 22. Частотное поведение N-грамм «Margaret Thatcher», «Mrs Thatcher» и «Thatcher»

Ниже представлены графики частотного поведения фамилий нескольких британских премьер-министров (рис. 23–26).

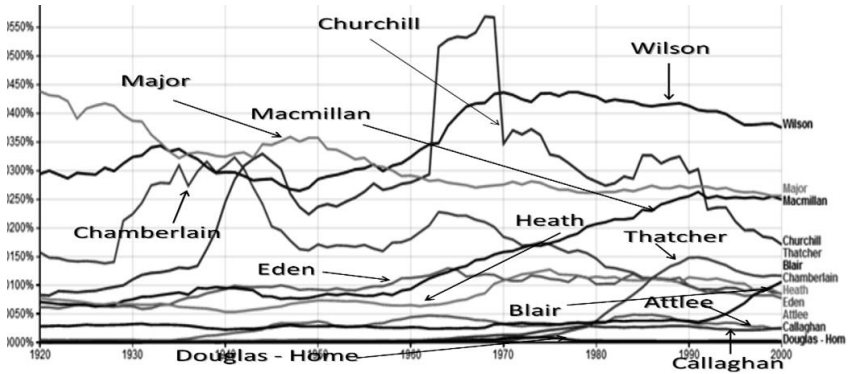


Рис. 23. Частотное поведение имен премьер министров Великобритании в корпусе британского английского языка Google Books (представление — фамилия)

Как видим, такие имена как «Wilson», «Major», являясь распространенными английскими фамилиями и, соответственно, не отражают историческую реальность, связанную с высокопоставленными носителями этих фамилий (рис. 23).

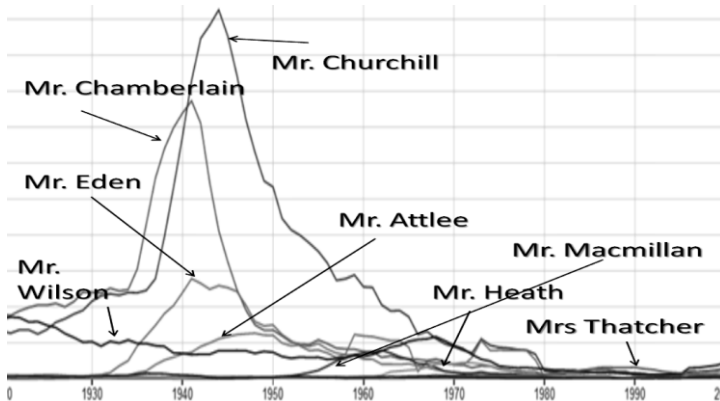


Рис. 24. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление Mr./ Mrs. фамилия)

Можно предположить, что высокий пик биграммы «Mr. Churchill» в 1944 г. связан с незаурядной личностью этого политика, его ролью во второй мировой войне (рис. 24). Вероятно, роль (положительная или негативная) Невилла Чемберлена в предвоенной политике была значительной, и потому пик его кривой – второй по высоте. Обращает внимание, насколько меняются соотношения между кривыми. Кривые имен политиков второй половины XX века значительно ниже, чем кривые первой половины. Можно было бы предположить, что это связано с неспокойной историко-политической ситуацией — войнами, возникновением национал-социализма, усилением коммунистического режима и пр., однако, нижеследующие графики не показывают такой корреляции.

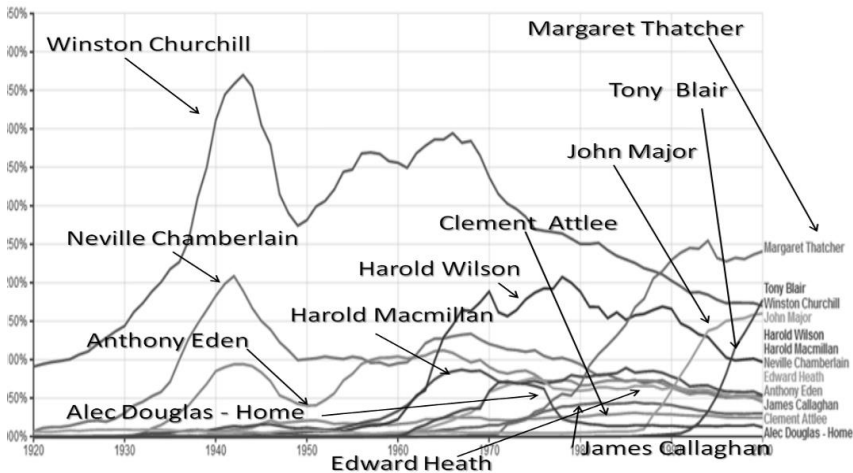


Рис. 25. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление – имя, фамилия)

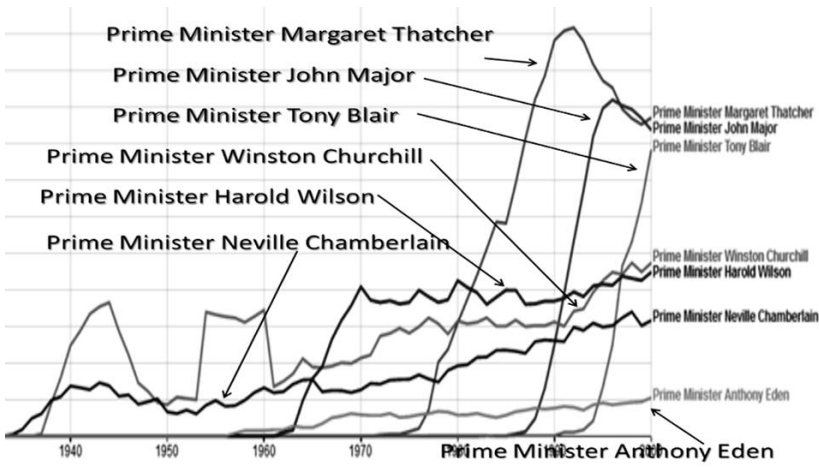


Рис. 26. Частотное поведение имен премьер министров Великобритании в корпусе Британского английского языка Google Books (представление – должность, имя, фамилия)

При указании имени и фамилии соотношения между кривыми снова меняются (рис. 25). Высота кривой биграммы «Winston Churchill» до середины 1980-х является наибольшей среди других кривых. Частотность биграммы «Margaret Thatcher» с конца 1970-х до 2000 г.

становится наиболее высокой, превышая в это время даже «Winston Churchill», что, впрочем, естественно для действующего премьер-министра. На рис. 26 отмечается значительный подъем кривых «John Major» и «Tony Blair» (форма представления – должность, имя, фамилия), которые на рис. 25 практически сливались с горизонтальной осью. Для них, как мы видим, характерно сочетание с личным именем и с должностью.

При таком представлении лидеров Великобритании мы видим картину, противоположную той, которая представлена на рис. 25. Пики кривых британских премьер-министров конца XX века значительно выше кривых премьеров начала века. Следует, однако, помнить, что эта форма представления (см. рис. 20, 22) наименее частотна среди всех рассмотренных форм.

Заключение

Частотное поведение имен политических деятелей отражает политические процессы, политические режимы, а в отдельных случаях и факты биографии политических деятелей.

При этом, как показано в настоящей статье, частотное поведение лексической единицы, означающей политического деятеля, зависит от способа представления: какая форма имени — полная или краткая — используется, указывается ли пост или должность политического деятеля.

В Российской империи имя императора указывалось намного реже, чем его пост, который обозначался лексическими единицами «Императорь», «Государь», «Его Величество». При этом по нашим данным во второй половине XIX века встречаемость этих слов в текстах русских книг значительно падает (рис. 2). Существует определенный набор глаголов, по-видимому, строго регламентированный, который употреблялся для описания действий императора (рис. 12).

В СССР кривые числа упоминаний имени лидера в печатных документах имеют подъем, некоторое плато или пик, затем после смерти или прекращения полномочий лидера некоторый период малой частотности, затем во второй половине 1980-х гг. рост частотности (Сталин, Хрущев, Брежнев, Андропов, Черненко, см. рис. 4, 5). Эта модель уже была описана нами ранее [5]. Несколько иначе выглядит кривая имени Горбачев. Частотность этой фамилии снижается лишь незначительно. Сходной модели следуют кривые имен «Ельцин» и «Путин», хотя в отношении их возможности исследования ограничены — графики могут быть построены только до 2008 года.

Частотность вариантов представления российских и советских лидеров имеет следующие особенности: у советских лидеров редко упоминается должность. В силу, вероятно, того что названия их должностей довольно длинны. Лидеры СССР обозначаются словом «товарищ» и далее фамилией. Такое обозначение наиболее частотно для И.В. Сталина. Н.С. Хрущёв и Л.И. Брежнев чаще обозначались именем, отчеством и фамилией, чем словом «товарищ» (рис. 7, 8, 9). У Брежнева, тем не менее, встречается и указание на должность.

При представлении М.С. Горбачева самыми частотными формами являются имя и фамилия, а также фамилия с инициалами. Используется также форма «президент Горбачев» но также сравнительно часто, особенно в начале правления, встречается и «товарищ Горбачев».

Самые распространённые формы представления Б.Н. Ельцина и В.В. Путина — имя и фамилия, у Ельцина встречается форма со словом «товарищ», а у Путина «господин».

В англоязычных странах имеется проблемы широко распространённых имен. Такие имена как «Johnson», «Wilson», «Major» и т. п. не могут отражать, какие-либо исторические и биографические события или процессы в силу распространенности этих фамилий. В русском корпусе такую же проблему может представить фамилия «Медведев».

Рассмотрены три варианта графиков частотного поведения имен президентов США (рис. 14–16). В первом варианте (рис. 14) использованы только фамилии. Заметно, что

распространённая фамилия Johnson имеет более-менее пологую кривую с некоторым подъёмом в годы президентства Линдона Джонсона. Кривые имен президентов второй половины XX века практически изоморфны.

График, где кривые построены по имени и фамилии президента (рис. 15), имеет следующие особенности. Кривые имен исторически более ранних президентов, достигнув определенной точки, не снижаются или снижаются незначительно. Кривые более поздних президентов достигают более высоких пиков частотности, но за пиками, как правило, годами окончания президентства, следует выраженное снижение.

На графике, построенном на биграмах из слова «President» и фамилий (рис. 15), виден ряд изоморфных кривых, каждая из которых точно соответствует периоду правления соответствующего президента. Можно утверждать, что эта форма обеспечивает высокую точность поиска.

На двух примерах рассмотрены формы обозначения премьер-министров Великобритании. Существует форма, состоящая из фамилии. Эта форма обеспечивает полноту поиска, но в то же время может дать некоторый информационный шум, иначе говоря, отразить в результатах поиска однофамильцев и родственников (рис. 17).

В британской традиции премьер-министр страны обозначается несколькими способами: Mr / Mrs и фамилия, имя и фамилия, Prime Minister имя и фамилия, Prime Minister фамилия (рис. 17–23). Наши данные показывают, что форма Mr / Mrs и фамилия более частотна в период пребывания политика на посту премьер министра (рис. 17–23).

При построении графиков на основе разных вариантов именовании мы видим изменение соотношений кривых в зависимости от выбранного варианта (рис. 24–27).

Таким образом, при проведении диахронических исследований частотности имен политических деятелей следует учитывать варианты их представления в печатных документах.

После пилотного исследования, позволяющего выявить различные варианты именовании политических деятелей, должно последовать методическое решение относительно конкретных запросов и конкретных имен. Возможно, например, включать характерные варианты в запрос как члены дизъюнкции. Другой вариант — при выраженном изоморфизме кривых использовать наиболее частотную N-грамму.

Система Google Books Ngram Viewer является мощным инструментом для диахронических исследований как отдельных языков, так для сравнительных межязыковых исследований. При этом она имеет и существенные недостатки. Необходимы дополнительные методологические разработки для уточнения связи между объемами корпусов и точностью результатов исследований. Мы, тем не менее, считаем, что Google Books Ngram Viewer — на сегодняшний день является единственным эффективным инструментом для исследований подобного рода.

Литература

- [1] Захаров В.П., Масевич А.Ц. Опыт корпусно-ориентированного историко-культурного исследования исторической и политической лексики // Библиосфера. 2016. №2. С.47–56.
- [2] Масевич А.Ц., Захаров В.П. Методы корпусной лингвистики в исторических и культурологических исследованиях // Компьютерная лингвистика и вычислительные онтологии: сб. научн. статей. Труды XIX Международной объединенной научной конференции «Интернет и современное общество» (IMS-2016), Санкт-Петербург. СПб: Университет ИТМО, 2016. С. 24–43.
- [3] Масевич А.Ц., Захаров В.П. Диахроническое исследование лексико-семантического поля «враги»// Труды международной конференции «Корпусная лингвистика – 2017». СПб.: Изд-во С.-Петерб. ун-та, 2017. С. 248–254.

- [4] Масевич А.Ц., Захаров В.П. Семантические трансформации политической лексики // Сборник научных статей XIX Объединенной конференции «Интернет и современное общество» IMS-2017. СПб: Университет ИТМО, 2017. С. 107–120.
- [5] Масевич А.Ц., Захаров В.П. Модели частотного поведения русской политической лексики XX века // Вестник Новосибирского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2017. Т. 15, № 2. С. 30–46.
- [6] Masevich A., Ostrovskaya A. V. The Authority for Names of Persons of Eighteen and Nineteen Century Russia in the Institute for Studies in Russian Literature: a Utopian Project // The Scholar and Database: paper presented on 4 November 1999 at the CERL conference hosted by the Royal library, Brussels / Ed. by Lotte Hellinga. London, 2001. P. 79–90.
- [7] Zakharov V.P., Masevich A.C., Pimenov E.N. Authority Control as a Linguistic Support Element of an Automated Library System // International Cataloguing and Bibliographic Control. 1996. Vol. 25, №. 4. P. 84–86.
- [8] Вершинина Л.П., Вершинин М.И., Масевич А.Ц. Построение модели поиска в электронном каталоге библиотеки на основе нечеткого отношения сходства // Библиосфера. 2013. №2. С.44–81.
- [9] Michel J-B. et al. Quantitative Analysis of Culture Using Millions of Digitized Books science // Science. 2011. Vol. 331. P. 176. DOI: 1126/Science.1199644.
- [10] Захаров В.П., Масевич А.Ц. Диахронические исследования на основе корпуса русских текстов Google Books Ngram Viewer // Структурная и прикладная лингвистика. 2014. Вып. 10. С. 303–327.
- [11] Galeev T.I., Solovyev V.D. Methods of Application of Modern Text Corpora in the Study of the Morphological System of Russian Verbs Unification of I Productive (irregular) Class of Verbs. Quantitative Model Based on Google Books // Modern Journal of Language Teaching Methods (MJLTM), (Dec. 2016). P. 177–180.
- [12] Bochkarev V., Solovyev V., Wichmann S. Universals Versus Historical Contingencies in Lexical Evolution // Journal of the Royal Society Interface. 2014. Vol. 11: 20140841. DOI: 10.1098/rsif.2014.0841.
- [13] Масленникова Ю.С., Бочкарев В.В., Соловьев В.Д. Вероятностная модель для оценки объема лексикона по данным корпуса Google Books Ngram. 2017. С. 255-260.
- [14] Захаров В. П., Масевич А. Ц. Лингвистическая картина российской истории XX века: корпусное исследование. URL: https://www.academia.edu/29209763/Лингвистическая_картина_российской_истории_XX_века:_корпусное_исследование_Linguistic_portrait_of_20th_century_Russian_history_a_corpora-based_study (дата обращения: 22.04.2018).
- [15] Ляшевская О. Н., Шаров С. А., Частотный словарь современного русского языка (на материалах Национального корпуса русского языка). М.: Азбуковник, 2009.

The Variability of the Representation of Politicians' Names in Diachronic Studies on the Base of Text Corpora

A.Ts. Masevich¹, V.P. Zakharov²

¹ Saint-Petersburg Institute of Culture, ² Saint-Petersburg State University

The paper continues a number of publications dealing with diachronic studies of frequency behavior of political terms using Google Books Ngram Viewer. In the preceding papers we have demonstrated how changing of frequency of certain lexical units reflects actual historical and cultural processes and described some models of frequency behavior of political terms in the massive of published texts. The present publication is written from the methodological point of view. We consider which way the form of personal name and position of a person can influence the frequency behavior of the lexical unit and how this influence should be considered in diachronic studies. The study covers a large period — from the middle of 19th century until the end of 20th century. We consider frequency behavior of more than 30 proper names of state leaders of Russian Empire, the Soviet Union, the Russian Federation, USA and UK in the texts in Russian, British and American versions of English. In the obtained material, we made certain observations, which can be useful in forthcoming diachronic studies.

Keywords: corpora, diachronic studies, political names, political terms, methodology, Google Books Ngram Viewer