DOI: 10.24833/2410-2423-2024-2-39-125-140 Review article

GETTING A HANDLE ON HANSARD WITH PYTHON AND NLTK, OR HOW TO TAME THE LINGUISTIC PICTURE OF BRITISH POLITICS WITH NLP

Sergey N. Gagarin

MGIMO UNIVERSITY
76, Prospect Vernadskogo, Moscow, 119454, Russia

Abstract. The article proposes an optimised starter's set of basic Python and NLTK (Natural Language Toolkit) methods that are essential in the analysis of massive textual corpora conducted as part of research investigating linguistic images of the world. The need to specify and detail these applied techniques stems from the nature and scope of the inexorable challenges confronted by contemporary cognitive linguistics and lexicology in the realm of unstructured big data analysis. Their viability and practical value are demonstrated in a series of illustrative examples where they are applied to the processing of continuous parallel diachronic corpora of Hansard that capture the discourse of both chambers of the British parliament produced in the years 2006-2023 and jointly amounting to over a third of a billion tokens.

The article suggests that the methods it outlines and classifies can be seen as forming an indispensable minimum of IT competences that is capable of delivering a substantial boost to the level of research both as regards its overall quality and its competitive edge. The proposed toolkit includes an essential set of instruments for target vocabulary processing as well as for the assessment and visualisation of word and phrase frequency and collocation.

The author presumes that, urged by the need to keep abreast of prevailing trends, the contemporary Russian researcher of linguistic images of the world is highly likely to find themselves compelled at some point to embrace the quantitative analysis methods made possible by combining Python and NLTK. As part of its substantial and varied range of benefits, the latter would arguably help them design and customise research protocols, adapting them with ease and versatility. Lastly and most importantly, the author suggests that Python and NLTK skills may serve as a comfortable gateway towards eventually upgrading one's linguistic research to cutting-edge global standards of technological sophistication and marketability.

Keywords: corpus linguistics, natural language processing, big data, cognitive linguistics, parliamentary discourse

For citation: Sergey N. Gagarin (2024). Getting a handle on a Hansard with Python and NLTK, or how to tame the linguistic picture of British politics with NLP. *Linguistics & Polyglot Studies, 10(2),* pp. 125–140. https://doi.org/10.24833/2410-2423-2024-2-39-125-140

Обзорная статья

БАЗОВЫЕ МЕТОДИКИ АНАЛИЗА ЯЗЫКОВЫХ КАРТИН ПОЛИТИКИ С ПОМОЩЬЮ ЯЗЫКА ПРОГРАММИРОВАНИЯ PYTHON И БИБЛИОТЕКИ NLTK (НА МАТЕРИАЛАХ КОРПУСОВ БРИТАНСКОГО ПАРЛАМЕНТСКОГО ДИСКУРСА)

С.Н. Гагарин

Московский государственный институт международных отношений (университет) МИД России, 119454, Россия, Москва, пр. Вернадского, 76

Аннотация. В рамках данной статьи предлагается один из возможных вариантов оптимального набора базовых методик, необходимых для изучения языковых картин мира на материалах крупных текстовых корпусов с использованием сочетания языка высокоуровневого языка программирования Python и библиотеки NLTK (Natural Language Toolkit). Необходимость выделения и конкретизации означенного методологического инструментария проистекает из характера тех вызовов, которые стоят перед современной когнитивной лингвистикой и лексикологией в сфере анализа больших неструктурированных данных. Работоспособность и практическая ценность предлагаемых методик демонстрируется на примере составленных автором сплошных параллельных диахронических корпусов дискурса обеих палат британского парламента за период с 2006 по 2023 гг., совокупный объём которых превышает треть миллиарда токенов. Набор предлагаемых методик включает в себя инструменты анализа базовых параметров вокабуляра, инструменты извлечения целевого вокабуляра, а также обработки и визуализации его частотных параметров и сочетаемости.

Целесообразность овладения предлагаемыми и систематизируемыми в рамках данной статьи методиками автоматического анализа текста обосновывается с позиции их необходимости как компетентностного минимума в области компьютерных технологий, который способен значительно повысить уровень лингвистических исследований и их научную конкурентоспособность.

Автор приходит к выводу о том, что в силу объективно сложившихся условий современному отечественному исследователю языковых картин мира с высокой долей вероятности придётся включить в свой прикладной инструментарий сочетание языка программирования Python и библиотеки NLTK. Предлагаемые в данной статье методики делают возможным гибкое формирование исследовательских протоколов с учётом широкого разнообразия возможных приоритетов. В качестве главного преимущества предлагаемого автором набора методов машинной обработки и количественного анализа текста видится возможность использования практических навыков, полученных в результате её освоения как комфортной компетентностной основы для последующей интеграции овладевшего ими лингвиста в сообщество исследователей наиболее высокотехнологичных и наиболее актуальных на сегодняшний день направлений науки о языке.

Ключевые слова: корпусная лингвистика, обработка естественного языка, большие данные, когнитивная лингвистика, парламентский дискурс

Для цитирования: Гагарин С.Н. (2024). Базовые методики анализа языковых картин политики с помощью языка программирования Python и библиотеки NLTK (на материалах корпусов британского парламентского дискурса). *Филологические науки в МГИМО. 10(2),* С. 125–140. https://doi.org/10.24833/2410-2423-2024-2-39-125-140

1. Введение

В каждой стране у лингвистики есть своё уникальное своеобразие, и Россия в этом плане не исключение. В отечественной лингвистике на протяжении многих десятилетий в числе наиболее приоритетных и определяющих направлений остаётся когнитивное. Оно традиционно отводит одну из ключевых ролей исследованию вербализации концептов. Поскольку эти ментальные конструкты, позволяющие человеческому сознанию упорядочивать восприятие реальности, не могут быть наблюдаемы непосредственно в силу естественных причин, их изучение в значительной мере сводится к анализу и категоризации той лексики, которая отражает их в языках, и тех языковых картин, которые из этой лексики складываются.

С одной стороны, можно с уверенностью утверждать, что научный интерес российских исследователей к системному осмыслению особенностей вербализации концептов не ослабевает. С другой стороны, очевидно и то, что результативность и глубина этих исследований страдают в силу необъятности эмпирического материала. Здесь можно с долей иронии говорить о существовании своего рода "проклятия лексиколога". Оно состоит в том, что обеспечение высокой репрезентативности исследования нередко сопряжено с такими затратами времени и усилий, которые плохо совместимы со здравым смыслом. Доступный к изучению материал настолько огромен, что на современном этапе практически любые работы по вербализации концептов, где выдвигаются гипотезы о закономерностях, характерных для того или иного языка в целом, оказываются в весьма уязвимом положении в связи с нереалистичностью полноценного выполнения исследовательской задачи такого уровня. Иными словами, исследователи языковых картин мира нередко оказываются в положении человека, пытающегося выпить море.

Очевидным способом решения проблемы обеспечения приемлемого уровня репрезентативности является разумное сужение эмпирической базы исследования. В связи с этим встаёт вопрос об оптимальных пределах подобного сужения. Проблема состоит в том, что, при всей своей привлекательности, компактный нишевой материал может не быть оптимальным выбором. В силу своего удобства он может привлекать внимание слишком большого количества исследователей и, как следствие, быть уже достаточно полно изученным на текущий момент. Это может делать его неподходящим для тех лингвистов, кому новизна их научных работ важнее простого факта их публикации. Таким образом, можно предположить, что оптимально суженная эмпирическая база для исследования вербальной репрезентации концептов должна быть достаточно обширной, чтобы её тщательный анализ отталкивал большинство исследователей, и в то же время не слишком крупной, чтобы кто-либо рискнул заняться её изучением на современном этапе развития науки и техники. В случае успешной реализации, эта концепция могла бы дать важное преимущество: зону собственного исследовательского комфорта, очерченную за рамками зоны комфорта большинства научных конкурентов, и лежащую в области больших данных, таких как крупные корпусы текста, содержащие сотни миллионов слов. Анализ подобных объёмов неструктурированных больших данных требует их машинной обработки, а она принципиально невозможна без знания языков программирования и специализированных библиотек. Для задач, предполагающих обработку массивных корпусов, их оптимальным сочетанием можно считать язык программирования Python и библиотеку NLTK (Natural Language Toolkit).

Умение или неумение использовать эти два инструмента можно с уверенностью назвать одной из причин, способных в значительной степени определить характер исследования одних и тех же корпусных данных в разных странах. В качестве доказательства данного тезиса можно привести пример того, каким образом на современном этапе развития науки проводятся корпусные исследования официальных отчётов о заседаниях британского парламента, также известных как "Хэнсард" (Hansard).

За рубежом в настоящее время существуют три основных направления исследования "Хэнсарда". Это анализ языковой репрезентации концептов и конструктов [23], [29], [30], [33], [34], [36], [37], [42], [44], [45], тональности и эмоциональной окраски текста [16], [17], [18], [19], [25], [41], а также типов риторики и коммуникативных стратегий [21], [31], [32], [39]. В силу масштабности своих целей, задач и эмпирических баз, эти исследования в своей подавляющей массе не могут обойтись без применения методов количественного анализа с опорой на возможности Python и NLTK. При этом ряд основанных на их использовании и прочно укоренившихся в зарубежной практике высокотехнологичных направлений исследования "Хэнсарда" пока в принципе не встречаются в отечественной науке. Это прежде всего анализ точности транскрибирования парламентских заседаний [24], [40] и анализ технических аспектов сбора и обработки больших данных, представленных текстовым материалом парламентских дебатов [35], [38], [43].

В нашей стране основным и наиболее массовым направлением изучения "Хэнсарда" является то, которое не требует опоры на языки программирования и специализированные библиотеки для них. Это историческое направление, и в его рамках материалы парламентских транскриптов выступают в роли исторических источников [1], [2], [3], [9], [10], [12], [13], [14].

Необходимо особо отметить, что отечественную практику изучения "Хэнсарда" объединяют с зарубежной три основных направления. Это анализ типов риторики и коммуникативных стратегий [4], [6], исследования языковой репрезентации концептов [7], [15], а также лексикологические исследования [5], [8], [11], [20], [22], [26]. Объединяет отечественную и зарубежную науку само наличие этих направлений, но в то же самое время их разделяет тот факт, что в отечественной практике использование Python и NLTK в их рамках с высокой долей вероятности никогда не имело места, о чём можно сделать вывод по отсутствию информации о подобных исследованиях в научных публикациях.

Проиллюстрированный выше пример различий между зарубежным и отечественным подходом к корпусным исследованиям свидетельствует о том, что потенциал отечественных научных школ в данной области на настоящий момент не реализован в полной мере. В этой связи можно говорить об определённом отставании отечественной науки. Тем не менее, представляется вполне возможным нивелировать его с помощью внедрения культуры использования передовых технологий машинной обработки текста. В этой связи возникает закономерный вопрос о том, как именно это сделать и с чего следует начать в первую очередь.

2. Цель работы

Целью данной работы является определение спектра базовых методик анализа с помощью Python и NLTK крупных корпусов англоязычного дискурса, которые были бы оптимальны для выявления особенностей и закономерностей вербализации концептов. В качестве подвергаемых машинной обработке материалов используются составленные автором данной работы сквозные диахронические параллельные корпусы отчётов о заседаниях Палаты общин и Палаты лордов британского парламента [27], [28]. По своему качеству эти отчёты максимально приближены к стенограммам, фиксирующим практически дословно всё пронесённое в обеих палатах парламента в ходе каждого дня заседания. На сегодняшний день оба этих корпуса охватывают период с июля 2006 г. по июль 2023 г. включительно. Это объясняется тем, что архив "Хэнсарда", доступный на официальном сайте парламента, не является полным. Часть материалов за период, предшествующий июлю 2006 года, отсутствует или же недоступна по причине сбоев интернет-ресурса.

Таким образом, означенный временной промежуток пока остаётся единственным и наиболее современным периодом, который полностью и без пробелов отражён в доступном, целостном и параллельном дискурсе обеих палат британского парламента.

Объём корпуса Палаты общин составляет 194731646 токенов, а объём корпуса Палаты лордов – 150866492 токенов. Их совокупный объём превышает треть миллиарда токенов. Без машинной обработки анализ подобного материала лишён практической целесообразности, поскольку на него не хватило бы одной человеческой жизни. Вместе с тем работа с данными корпусами помогла на практике выявить ту базовую совокупность методов, основанную на сочетании Python и NLTK, которая полностью соответствует задачам исследования больших неструктурированных языковых данных, стоящих перед современной российской лексикологией и когнитивной лингвистикой.

3. Базовый инструментарий Python и NLTK

3.1 Инструменты для анализа базовых параметров вокабуляра

На начальном этапе исследования корпуса зачастую возникает необходимость в оценке его количественных параметров, таких как размер вокабуляра и уровень лексического разнообразия. Данные задачи решаются с помощью таких базовых функций языка Python как ***len*** и ***set*** (Таблица 1).

jupyter House_of_Lords_lexical_diversity (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

```
In [8]: lords_count = my_lords_corpus.words()

In [9]: len(lords_count) ### общее число токенов в корпусе Палаты лордов
Out[9]: 150866492

In [10]: len(set(lords_count)) ### число уникальных токенов в корпусе Палаты лордов
Out[10]: 150991

In [13]: ### функция для определения уровня лексического разнообразия корпуса Палаты лордов

         def lords_lexical_diversity(lords_count):
             word_count = len(lords_count)
             vocab_size = len(set(lords_count))
             diversity_score = word_count / vocab_size
             return diversity_score

In [14]: lords_lexical_diversity(lords_count)  ### показатель лексического разнообразия корпуса Палаты лордов
Out[14]: 999.1753945599406
```

Таблица 1. Оценка базовых параметров вокабуляра (корпус Палаты лордов)

Функция ***len*** выводит общее число токенов, а функция ***set*** схлопывает дубликаты и представляет весь вокабуляр корпуса в виде набора уникальных токенов. В приводимом примере вывод показателя лексического разнообразия осуществляется с помощью функции ***lords_lexical_diversity***, которая делит общее число токенов в корпусе на число уникальных токенов.

При сплошном анализе крупных парламентских корпусов такой параметр как уровень лексического разнообразия может быть особенно полезен в контексте компаративных исследований, в рамках которых осуществляется количественное сопоставление лексики парламентского дискурса разных стран. Также данный параметр может быть интересен при сравнительном анализе более частных сегментов языковой картины политики. В частности, с его помощью можно осуществлять сопоставительный анализ уровня разнообразия всего вокабуляра, использованного в рамках парламентских дебатов премьер-министрами, министрами иностранных дел, финансов, обороны и прочими представителями политического руководства одной или нескольких стран

на протяжении всей их политической карьеры. Таким образом, можно составить объективное представление о сравнительном разнообразии их вокабуляров и, следовательно, о сравнительной сложности их индивидуальных языковых картин мира.

На первый взгляд, подобная информация может представлять сугубо теоретический и весьма нишевой интерес, однако у неё есть неочевидная практическая ценность. Она имеет непосредственное отношение к дидактическому целеполаганию при обучении иностранным языкам в вузе. Здесь необходимо сделать особенный акцент на том факте, что в нашей стране в настоящее время при формулировании стандартов качества образовательных программ по иностранным языкам не используется такой объективный количественный показатель как коэффициент лексического разнообразия обязательного к освоению вокабуляра. Вместе с тем, его введение могло бы дать более чёткое представление о том, какие цели лексического характера следует ставить перед студентами. Этот параметр позволил бы более объективно оценивать, насколько вокабуляр, к примеру, изучающего английский язык аспиранта МГИМО, близок в плане разнообразия к вокабуляру того или иного премьер-министра или министра иностранных дел Великобритании, или к усреднённому показателю лексического разнообразия, выведенному для всех, кто занимал эту должность за последние двадцать, тридцать лет или какой-либо другой период, определённый в качестве наиболее точно соответствующего устанавливаемым целям и задачам. В перспективе такой подход мог бы добавить новое измерение к методикам оценки качества образовательных программ по иностранным языкам. Наряду с достоинствами данного подхода, необходимо особо отметить имеющийся у него существенный недостаток. Он связан с тем, что практическая реализация подобной концепции едва ли представляется возможной без массового внедрения в техническую сторону образовательного процесса систем распознавания речи и автоматической компиляции индивидуальных корпусов, состоящих из всех слов, произнесённых каждым студентом на занятиях по иностранному языку. Такие системы требуют широкого применения искусственного интеллекта, а их разработка и обслуживание, к сожалению, могут быть сопряжены со значительными материальными затратами.

3.2 Инструменты для извлечения лексики

Одним из важнейших этапов в работе корпусного лексиколога и лексикографа является формирование полного словника исследуемого корпуса. Эта задача традиционно являлась сложной и громоздкой, но с помощью NLTK её решение возможно с предельной простотой При этом элементы словника извлекаются как ключи словаря из пар ключ-значение, где в качестве значения выступает лексическая частота (Таблица 2).

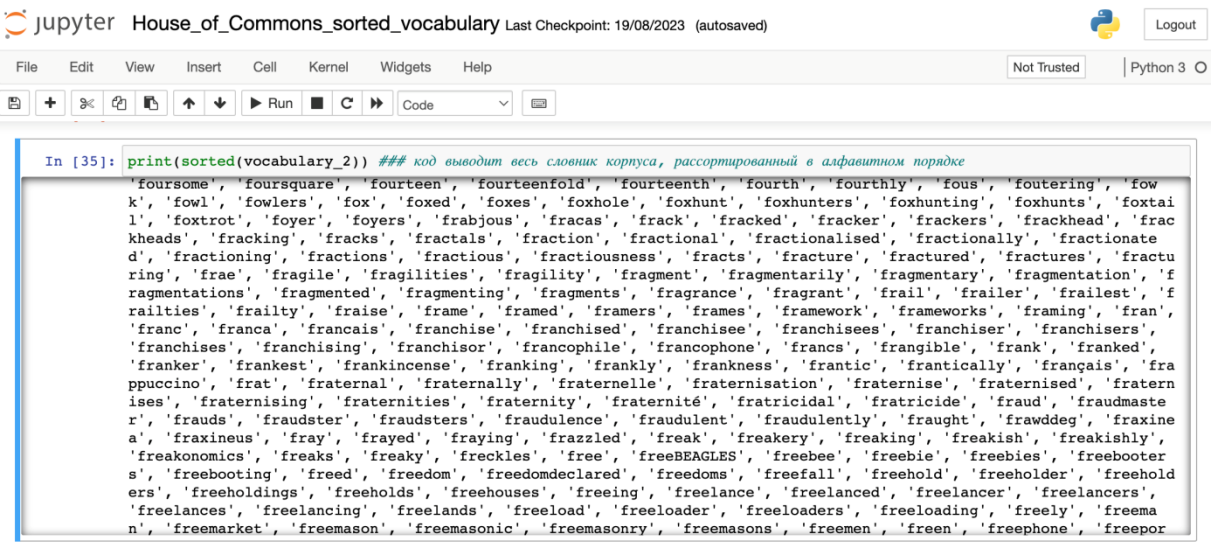

Таблица 2. Извлечение полного словника корпуса Палаты общин за последние 17 лет (фрагмент результата)

В числе наиболее значимых преимуществ быстрого доступа к полному словнику целого корпуса следует упомянуть тот факт, что с его помощью существенным образом упрощается работа с таким лексическим пластом как иноязычные заимствования. Особую актуальность это приобретает в том случае, если в качестве источника заимствований выступает язык, которым сам исследователь не владеет. В качестве характерных примеров здесь можно привести заимствования из ирландского гэльского языка, используемые в англоязычных дебатах парламента Республики Ирландия, а также заимствования из языка маори (маоризмы) в парламентском дискурсе Новой Зеландии. В Российской Федерации на сегодняшний день количество специалистов по данным языкам крайне невелико, и они справедливо относятся к категории чрезвычайно редких для нашей страны. Тем не менее, даже отсутствие знания языка маори нельзя рассматривать в качестве серьёзного препятствия для извлечения его слов из англоязычного корпуса транскриптов заседаний новозеландского парламента. Для этого весь словник соответствующего корпуса выводится в виде рассортированного по алфавиту списка, а затем к нему применяется функционал WordNet или аналогичных лексических баз данных. Такая обработка представляет собой черновой этап процесса, поскольку позволяет автоматически удалить из словника большую часть английских слов. Затем следует чистовая часть процесса, и словник дообрабатывается вручную. После этого полученная лексика рассортировывается с помощью переводного и толкового словаря по лексико-семантическим группам, что даёт представление о том, какие концепты с её помощью вербализуются и какие лакуны англоязычного дискурса она заполняет. При всей своей простоте данный подход довольно эффективен и с успехом применяется для исследования маоризмов в парламентском дискурсе Новой Зеландии в рамках работ, которые в настоящее время ведутся под руководством автора на кафедре английского языка №1 МГИМО МИД России.

3.3 Инструменты для работы с лексической частотностью

После извлечения из корпуса словника может возникнуть потребность в осмыслении картины частотного распределения его лексики в исследуемом материале. Задачи подобного характера эффективно решаются с помощью функций ***FreqDist*** и ***ConditionalFreqDist***. Обе они эффективно работают на неконкатенированных корпусах, то есть, для того, чтобы ими воспользоваться, нет необходимости предварительного монтажа встык в хронологическом порядке всех материалов, а это, в свою очередь, существенно экономит время исследователя. Функция ***ConditionalFreqDist*** используется для анализа условного распределения частот, что особенно удобно при сквозном диахроническом подходе, когда в качестве условия распределения выступает год или месяц публикации текстового материала, указанные в названии каждого отдельно взятого файла в формате гггг_мм_дд (Таблица 3). Также этот метод довольно удобен для визуализации языковых данных, на основании которой можно формулировать предварительные гипотезы исследований в области когнитивной лингвистики или предиктивной аналитики. Приведённый ниже пример иллюстрирует эту особенность на примере условного распределения частот для слов war и peace в корпусе дебатов Палаты общин. Два графика, наглядно иллюстрирующих изменения частоты их употребления по годам, могут дать основания как для гипотез о соотнесённости их пиковых значений с событиями внеязыковой реальности, так и для гипотез об усилении тенденций их метафорического употребления, которые эти значения отражают. В любом случае, данные экстремумы можно рассматривать как указание на наиболее информативные сегменты корпуса, которые могут дать наибольшее количество лексического материала, иллюстрирующего особенности вербализации двух концептов политического дискурса, которые традиционно входят в число наиболее ключевых, а потому представляющих для исследователей неизменно высокий интерес.

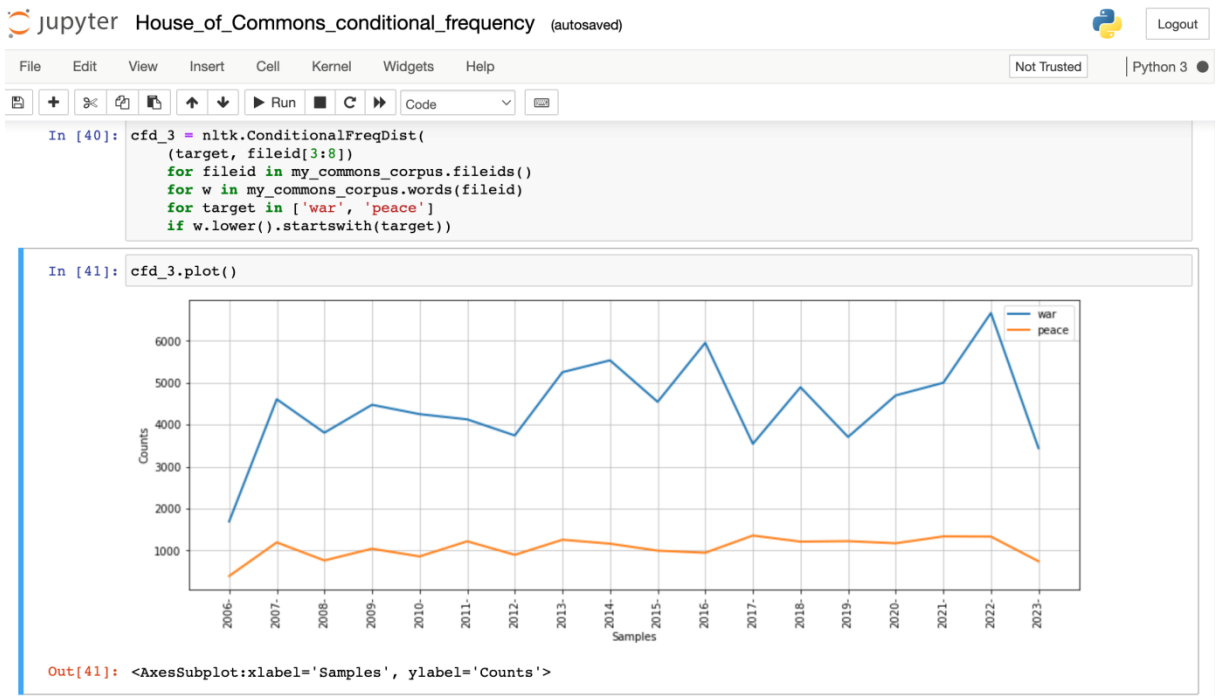

Таблица 3. Условное распределение частот слов war и peace в корпусе Палаты общин по годам

Зачастую, когда графиков в одних и тех же координатных осях становится слишком много, их восприятие может существенно затрудняться, и в результате возникает необходимость в альтернативных методиках визуального представления частотности. В таких случаях может быть целесообразным прибегнуть к визуализации лексической дисперсии на материале конкатенированного, то есть смонтированного в виде единого файла, корпуса. Для этих целей предусмотрена функция ***dispersion_plot*** (Таблица 4).

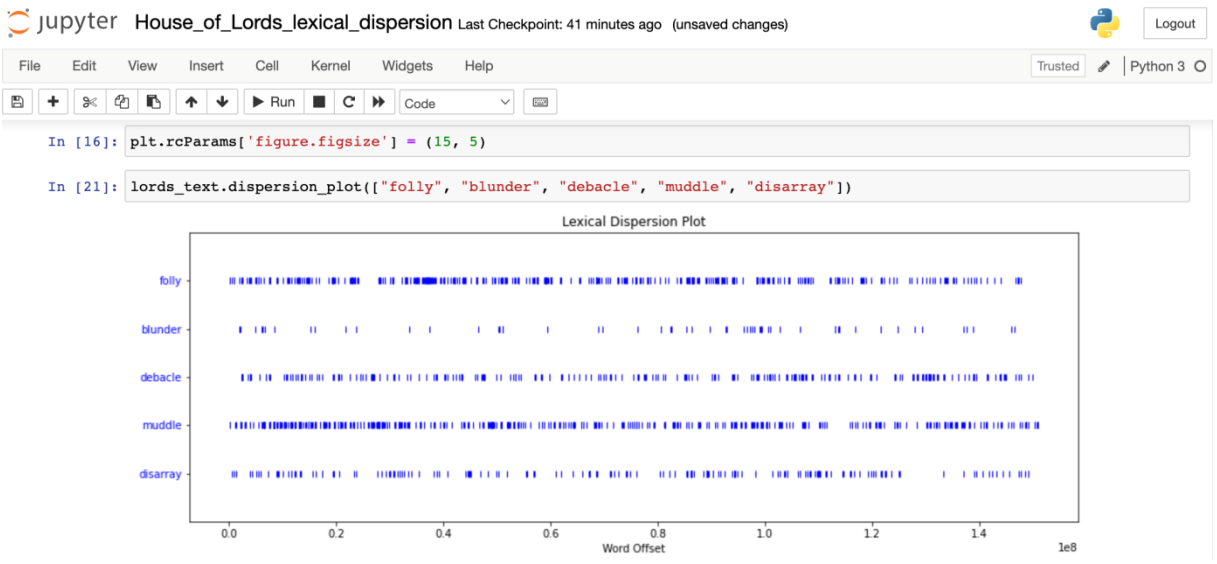

Таблица 4. Визуализация лексической дисперсии слов ***folly, blunder, debacle, muddle*** и ***disarray*** в корпусе Палаты лордов

Этот метод визуализации особенно удобен для наглядного отображения особенностей вербализации концептов или сравнительной частотности элементов лексико-семантических групп. Будучи менее информативным и точным в плане репрезентации частотных экстремумов, чем традиционная линейная диаграмма, он, тем не менее, позволяет создать весьма точное представления об относительной плотности распределения в тексте искомой лексики.

Приводимый выше пример иллюстрирует особенности лексической дисперсии группы существительных, объединённых общей прагматической функцией. Все они используются для оказания психологического давления на адресата через суровую критику его действий или их результатов. Их словарные дефиниции не могут дать никакого представления о том, насколько популярным они могут быть в современном речевом узусе британских парламентариев. Гистограмма лексической дисперсии, напротив, даёт чёткое указание на то, что из означенной группы наиболее частотными на протяжении всего периода наблюдения являются лексемы *folly* (объективирующая концепт ***foolishness***) и лексема *muddle* (объективирующая концепт ***disorder***). Эта информация может служить основанием для формулирования рабочей гипотезы о том, что лексико-семантические группы, вербализующие концепты ***foolishness*** и ***disorder***, могут рассматриваться как объекты приоритетного изучения в целях определения лексики, оптимально соответствующей задачам оказания давления на политических оппонентов.

3.4 Инструменты для работы с лексической сочетаемостью

Картины лексической частотности играют ключевую роль в комплексном анализе корпусов, но для некоторых направлений исследований не менее, а зачастую и гораздо более значимыми являются картины лексической сочетаемости. На базовом уровне они извлекаются в виде конкордансов слов или словосочетаний (Таблица 5), которые затем можно сохранять в виде отдельных файлов и использовать в качестве своеобразных мини-корпусов для исследования ближайшего контекстного окружения узловых слов.

Функция ***concordance*** может быть особенно полезной, когда необходимо узнать, употребляется ли искомая лексема исключительно в качестве элемента идиомы. Также она может помочь в определении того, употребляется ли то или иное слово на неосновном языке корпуса за пределами устойчивых и повторяющихся формулировок. В качестве характерного примера здесь можно привести дискурс парламента Новой Зеландии, где фигурируют молитвы на языке маори перед началом каждого заседания, а также случаи произносимой на этом языке присяги при вступлении в должность и прощальных формул в рамках официальных соболезнований в связи с кончиной парламентариев. С историко-культурной точки зрения все они представляют однозначный интерес, однако для исследователя заимствований из языка маори в новозеландский вариант английского языка они скорее представляют помеху, поскольку используемая в них лексика употребляется в рамках устойчивых маорийских формул, которые представляют из себя крупные иноязычные вкрапления и к собственно заимствованиям не относятся. Соответственно, перед определением спектра полноправных лексических заимствований эти формулы и их составляющие необходимо идентифицировать, и функция ***concordance*** позволяет эффективно решить данную задачу.

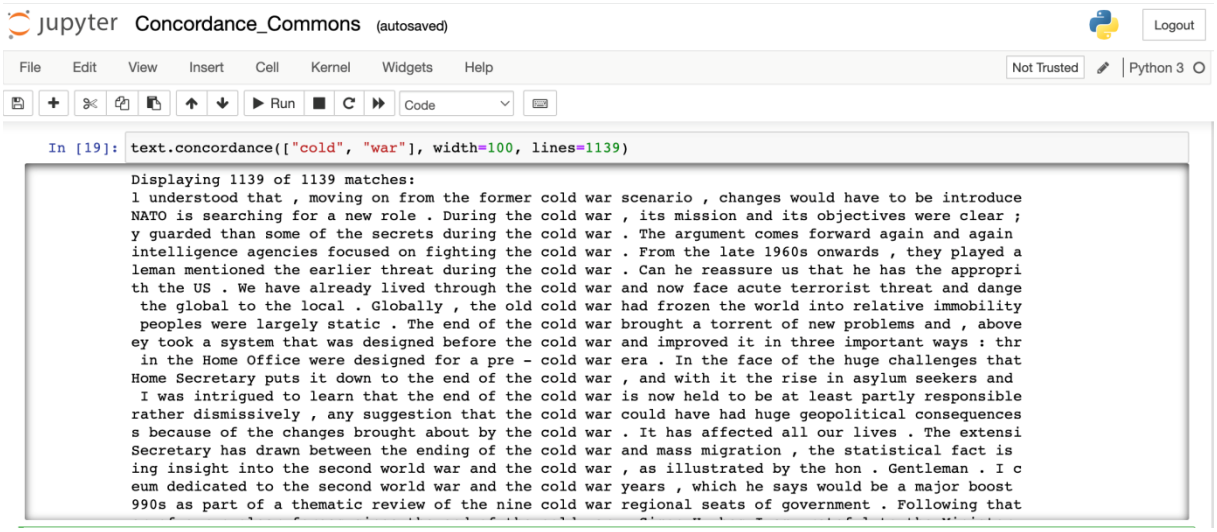

Таблица 5. Построение конкорданса коллокации ***cold war*** с помощью функции ***concordance***. Корпус Палаты общин, 1139 строк (фрагмент)

Эффективным инструментом анализа лексической сочетаемости также являются биграммы (функция ***bigrams***), особенно в сочетании с визуализирующими их графиками частотности. В приводимом ниже примере (Таблица 6) на материале корпуса Палаты общин с помощью данной функции демонстрируется частотность ста наиболее употребимых биграмм, первым компонентом которых является слово *utter*.

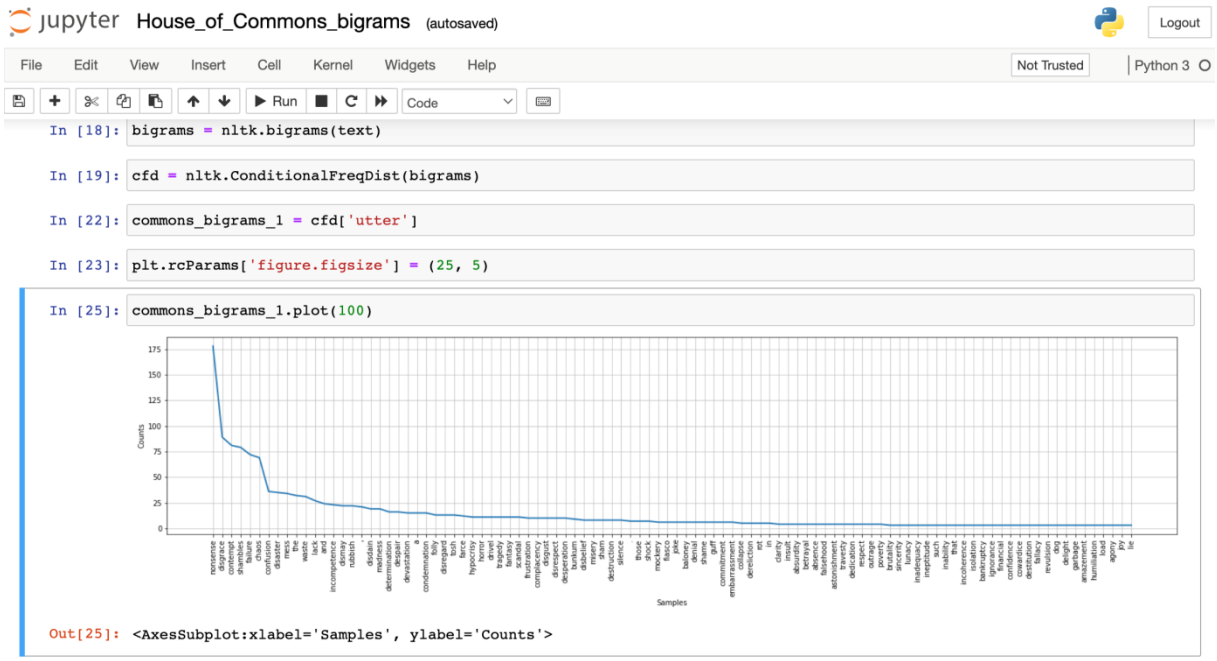

Таблица 6. Визуализация частотности биграмм.
Сто наиболее употребимых сочетаний со словом ***utter***, корпус Палаты общин

По оси ***x*** расположены вторые компоненты биграмм в порядке убывания частоты. Необходимо особо отметить, что такой инструмент как биграммы весьма удобен для поиска устойчивых словосочетаний и ранжирования их по степени коллокативной устойчивости. При этом, как явствует из приведённого ниже примера, картина этой устойчивости будет гораздо более информативной для исследователя языка политики, чем та, которую можно получить о прилагательном

utter из словарной статьи или из совокупности таких статей, поскольку она отражает не усреднённое представление о сочетаемости заголовочного слова для английского языка в целом, а соответствует объективной картине его контекстуального употребления именно в том социолектном варианте данного языка, который характерен для представителей британского парламента.

В контексте изучения языка политики и парламентского дискурса, использование такой функции как ***bigrams*** может иметь особое значение для уточнения и систематизации недоисследованных сочетаемостных особенностей как нетерминологической, так и терминологической лексики.

В тех случаях, когда биграммы перестают удовлетворять тем критериям гибкости и универсальности, которые необходимы для оптимального поиска словосочетаний, корпус можно исследовать с помощью регулярных выражений. Этот инструмент использует метасимволы и позволяет обнаруживать практически любые n-граммы, поскольку даёт возможность поиска последовательностей токенов по повторяющемуся шаблону, в котором могут одновременно присутствовать известные, неизвестные и частично известные элементы (Таблица 7).

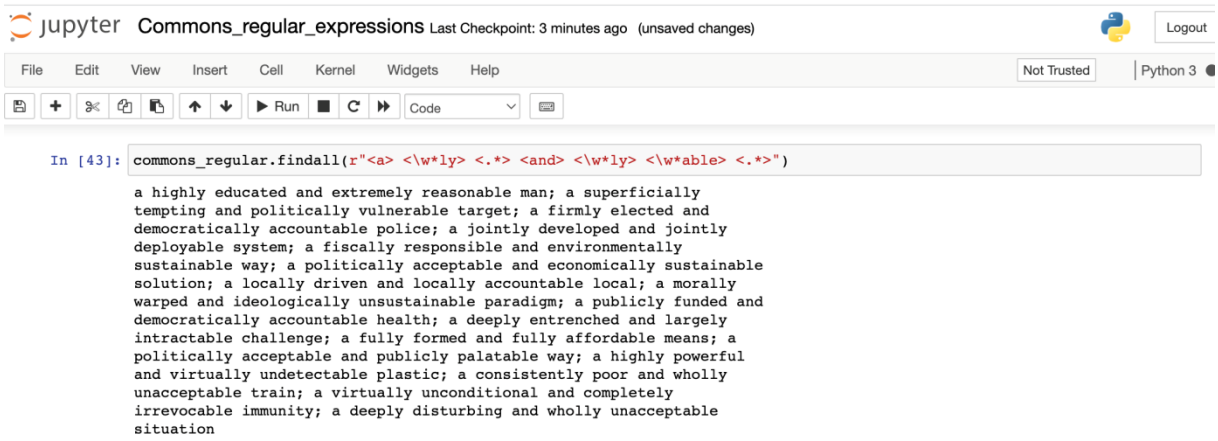

Таблица 7. Регулярное выражение для поиска n-грамм с двумя известными, двумя неизвестными и тремя частично известными элементами. Корпус Палаты общин

Данный инструмент может быть особенно полезен для исследований особенностей речевого стиля отдельных политиков или компаративных исследований речевого поведения представителей политического руководства, занимавших один и тот же или разные посты. Так, с помощью регулярных выражений можно осуществлять сравнительные исследования речевых паттернов министров иностранных дел и при необходимости сравнивать их с паттернами, характерными для министров обороны (или любых других членов кабинета министров), с целью анализа влияния занимаемой должности на состав и структуру фигур речи.

Следует особо выделить тот факт, что регулярные выражения можно рассматривать как инструмент корпусного анализа, позволяющий выделять наиболее типологически характерные для того или иного типа дискурса n-граммы. Следующим шагом после их идентификации может стать системный анализ той лексики, которая употребляется в их составе, на предмет её принадлежности к лексико-семантическим группам, вербализующим ключевые концепты отдельных сегментов языковой картины мира. Таким образом, наше представление о закономерностях и тенденциях их вербализации может быть переведено на более высокий уровень системного восприятия.

4. Выводы

В силу объективно сложившихся условий, современному отечественному лингвисту с большой долей вероятности придётся включить в свой исследовательский инструментарий языки программирования и специализированные библиотеки. При этом спектр их функционала чрезвычайно обширен, и далеко не всегда оказывается очевидным, что именно необходимо изучить

в первую очередь для успешного применения на практике в своей отдельно взятой области. Для ряда направлений лингвистики на сегодняшний день можно констатировать фактическое отсутствие чётких исследовательских протоколов с использованием языка Python и библиотеки NLTK. В рамках данной статьи была сделана попытка предложить базовый набор инструментов, на основе которого было бы возможным гибкое и адаптивное формирование подобных протоколов в ближайшем будущем для тех научных школ и направлений когнитивной лингвистики, в рамках которых ведутся корпусные исследования особенностей вербализации концептов в дискурсе. Предлагаемые инструменты показали высокую эффективность при работе на больших текстовых данных. Перед исследователем, освоившим эту базовую совокупность прикладных методов корпусного анализа, неизбежно встанет выбор: остановиться на достигнутом техническом уровне или развиваться дальше. В первом случае он остаётся с техническим инструментарием, который, невзирая на всю свою лаконичность, даёт ему ощутимое преимущество перед конкурентами. Во втором случае он переходит на принципиально новый уровень, поскольку сосредотачивает свои усилия на освоении направлений программирования, связанных с автоматической разметкой корпусов текста с помощью теггеров, автоматической классификацией текстов, предиктивной аналитикой, обучением нейросетей и т.д. Таким образом, его работа будет продвигаться в русле, отличном от классической лингвистики, а одним из основных мерил его профессионального успеха можно будет считать качество создаваемых им больших языковых моделей (LLM). Вне зависимости от того, насколько глубоко исследователь готов интегрировать программирование в свою работу и какому из двух означенных путей дальнейшего профессионального развития он готов следовать дальше, он может быть уверен в том, что оба из них можно считать относительно беспроигрышными, по крайней мере в ближайшей перспективе.

Список литературы

1. Айзенштат М.П. Новации в парламентской практике Британии XVIII столетия//Honoris causa. Сборник научных статей, посвящённый 70-летию профессора Виктора Владимировича Сергеева. Санкт-Петербург, 2016. С. 7–13.
2. Айзенштат М.П. Парламентские материалы Британии XVII-XIX веков// Запреты и преодоления. Новая и новейшая история. 2016. № 5. С. 16–25.
3. Быкова Е.А. Вопрос признания советского государства в политической дискуссии британского парламента/ Е.А. Быкова, А.А. Сигова. Ветер Перестройки – 2022. // Сборник материалов Второй Всероссийской научной конференции/ отв. ред. А. Д. Матлин. Санкт-Петербург, 2023. С. 22–27.
4. Головина Н.М. «Непарламентские выражения» и речевая агрессия в британском парламенте: риторическая стратегия или институциональная норма? // Речь и языки общения в конфликтогенном мире. Материалы международной научно-практической конференции/ отв. ред. С.В. Мыскин. Москва, 2021. С. 37–39.
5. Захарова О.В. Обсуждение миграционной политики в британском парламенте// Человек, образ, слово в контексте исторического времени и пространства. Материалы Всероссийской научно-практической конференции. Москва, 2015. С. 93–96.
6. Зюбина И.А. Реализация коммуникативных стратегий в британском парламенте/ И.А. Зюбина, В.А. Маслова. // Уральский научный вестник. 2023. Т. 6, № 6. С. 53–60.
7. Ковалёв Н.А. «СВОИ» versus «ЧУЖИЕ»: динамика развития и манипулятивный потенциал концепта ХОЛОДНАЯ ВОЙНА в англоязычном политическом дискурсе/ Н.А. Ковалёв, Н.А. Чес. // Вестник Российского университета дружбы народов. Серия: Теория языка. Семиотика. Семантика. 2017. Т. 8, №4. С. 1171–1178.
8. Корецкая О.В. О некоторых политических эвфемизмах в эпоху постправды (на примере английского языка)// Филологические науки в МГИМО. 2021. Т. 7, № 3 (27). С. 16–23.
9. Корнилов А.А. Британский парламент как центр выработки внешнеполитических решений в период сирийского кризиса (2011-2015 годы)/ А.А. Корнилов, Н.С. Лобанова, А.И. Егоров. // Научный диалог. 2023. Т. 12, № 2. С. 363–384.
10. Корнилов А.А. Обсуждение палестино-израильского конфликта в комитете британского парламента по иностранным делам (2014 год)/ А.А. Корнилов, Н.С. Лобанова, О.Р. Жерновая. // Научный диалог. 2022. Т. 11, № 2. С. 437–462.
11. Лобанова Н.С. Ключевые термины документов британского парламента в области ближневосточной политики: этимология, политическое значение и примеры использования // Регионы мира: проблемы истории, культуры и политики. Сборник научных статей. Нижний Новгород, 2021. С. 107–112.
12. Лобанова Н.С. Подход комитета по иностранным делам британского парламента к кризису на Украине// Научно-аналитический вестник Института Европы РАН. 2023. № 6 (36). С. 7–18.
13. Михайлов В.В. Вхождение Азербайджана в состав советского государства и политика Великобритании в отношении Закавказья в 1918-1920 гг.: политический и социально-экономический аспекты// Учёные записки Крымского федерального университета имени В.И. Вернадского. Исторические науки. 2022. Т. 8, № 2. С. 73–87.

14. Хахалкина Е.В. “Поколение Виндраш” в контексте современного развития мультирасовой Великобритании (по материалам британского парламента)// Новая и новейшая история. 2022. № 6. С. 180–191.
15. Чес Н.А. Концептуальная метафора в политическом медиадискурсе (на материале английского языка): монография/ Н.А. Чес. Москва: МГИМО-Университет, 2020. 190 с.
16. Abercrombie G., Batista-Navarro R. A sentiment-labelled corpus of Hansard parliamentary debate speeches// Proceedings of ParlaCLARIN. Common Language Resources and Technology Infrastructure (CLARIN). 2018. P. 43–48.
17. Abercrombie G., Batista-Navarro R. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review// Journal of Computational Social Science. Vol 3, №1. 2020. P. 245–270.
18. Abercrombie G., Batista-Navarro R. ‘Aye’or ‘no’? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). 2018. P. 4173–4180.
19. Abercrombie G., Batista-Navarro R. Identifying opinion-topics and polarity of parliamentary debate motions// Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis. 2018. P. 280–285.
20. Aspinall P. Ethnic/racial terminology as a form of representation: A critical review of the lexicon of collective and specific terms in use in Britain// Genealogy. Vol. 4, № 3. 2020. P. 87–100.
21. Bischof K., Ilie C. Democracy and discriminatory strategies in parliamentary discourse// Journal of Language and Politics. Vol. 17, № 5. 2018. P. 585–593.
22. Charteris-Black J. Metaphor and gender in British parliamentary debates/ J. Charteris-Black. Palgrave Macmillan UK, 2009.
23. Coutto T. Half-full or half-empty? Framing of UK–EU relations during the Brexit referendum campaign// Crisis and Politicisation. Routledge, 2021. P. 85–103.
24. Cribb M., Rochford S. The transcription and representation of spoken political discourse in the UK House of Commons// International Journal of English Linguistics. Vol. 8, № 2. 2018. P1–14.
25. Duthie R., Budzyńska K. Classifying types of ethos support and attack// 7th International Conference on Computational Models of Argument. IOS Press, 2018. P. 161–168.
26. Hiltunen T. et al. Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century// Language Sciences. 2020 [Электронный ресурс]. – URL: https://doi.org/10.1016/j.langsci.2020.101270 (дата доступа 04.03.2024).
27. House of Commons Hansard. [Электронный ресурс] – URL: https://hansard.parliament.uk/commons Avialable from: https://hansard.parliament.uk/commons (дата доступа 12.09.2023).
28. House of Lords Hansard. [Электронный ресурс] – URL: https://hansard.parliament.uk/lords. Avialable from: https://hansard.parliament.uk/lords (дата доступа 12.09.2023).
29. Huysmans J., Alessandra Buonfino A. Politics of exception and unease: Immigration, asylum and terrorism in parliamentary debates in the UK// Political studies. Vol. 56, № 4. 2008. P. 766–788.
30. Ihalainen P., Sahala A. Evolving conceptualisations of internationalism in the UK parliament: Collocation analyses from the League to Brexit// Digital Histories: Emergent Approaches within the New Digital History. 2020. P. 199–219.
31. Ilie C. Parenthetically speaking: Parliamentary parentheticals as rhetorical strategies// Dialogue Analysis 2000: Selected Papers from the 10th IADA Anniversary Conference. Tübingen: Niemeyer, 2003. P. 253–264.
32. Ilie C. Strategic uses of parliamentary forms of address: The case of the UK Parliament and the Swedish Riksdag// Journal of pragmatics. Vol. 42, № 4. 2010. P. 885–911.
33. Jeffries L., Walker B. Austerity in the Commons: A corpus critical analysis of austerity and its surrounding grammatical context in Hansard (1803–2015)// Discourse Analysis and Austerity. Routledge, 2019. P. 53–79.
34. Kettell S., Kerr P. From eating cake to crashing out: constructing the myth of a no-deal Brexit// Comparative European Politics. 2020. Vol. 18. P. 590–608.
35. Labat S., Kotze H., Szmrecsanyi B. Processing and prescriptivism as constraints on language variation and change: Relative clauses in British and Australian English parliamentary debates// Exploring Language and Society with Big Data: Parliamentary discourse across time and space. 2023. P. 250–276.
36. Leduc R. The ontological threat of foreign fighters// European Journal of International Relations. 2021. Vol. 27, № 1. P. 127–149.
37. Mair C. Empire, migration and race in the British parliament (1803–2005)// Exploring Language and Society with Big Data: Parliamentary discourse across time and space. 2023. P. 111–118.
38. McGill E., Saggion H. BSL-Hansard: A parallel, multimodal corpus of English and interpreted British Sign Language data from parliamentary proceedings// Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages. 2023. P. 38-43.
39. McKenzie-McHarg A., Fredheim R. Cock-ups and slap-downs: A quantitative analysis of conspiracy rhetoric in the British Parliament 1916–2015// Historical Methods: A Journal of Quantitative and Interdisciplinary History. 2017. Vol. 50, № 3. P. 156–169.
40. Mollin S. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts// *Corpora*. 2007. Vol. 2, № 2. P. 187–210.
41. Onyimadu O. et al. Towards sentiment analysis on parliamentary debates in Hansard// Semantic Technology: Third Joint International Conference, JIST 2013, Seoul, South Korea, November 28-30, 2013. Revised Selected Papers. Vol. 3. Springer International Publishing, 2014. P. 48–50.
42. Riihimäki J. At the heart and in the margins: Discursive construction of British national identity in relation to the EU in British parliamentary debates from 1973 to 2015// Discourse & Society. 2019. Vol. 30, № 4. P. 412–431.
43. Thundyill S. et al. Moving Fingers Write History and Having Writ Become Digital: Towards a Big Data Framework for the Analysis of Parliamentary Proceedings// Future of Information and Communication Conference. Cham: Springer Nature Switzerland. 2023. P. 459–479.
44. Van Dijk T. Political identities in parliamentary debates// European parliaments under scrutiny: Discourse strategies and interaction practices. 2010. P. 29–56.

45. Willis R. Taming the climate? Corpus analysis of politicians' speech on climate change// Environmental Politics. Vol. 26, № 2. 2017. P. 212–231.

References

1. Aizenshtat, M.P. Novatsii v parlamentskoi praktike Britanii XVIII stoletiia [Innovations in Britain's Parliamentary practice of the 18th and 19th centuries] // *Honoris causa. Sbornik nauchnykh statei, posviashchennyi 70-letiiu professora Viktora Vladimirovicha Sergeeva [Honoris causa. Collected Articles of the scientific conference celebrating the 70th anniversary of Professor Viktor Sergeev]*. Sankt-Peterburg, 2016. P. 7–13.
2. Aizenshtat, M.P. Parlamentskie materialy Britanii XVII-XIX vekov. Zaprety i preodoleniia. [Britain's parliamentary materials of the 18-19 centuries. Prohibitions and their overcoming] // *Novaia i noveishaia istoriia [Modern and contemporary history]*. 2016. № 5. P. 16–25.
3. Bykova, E.A., Sigova, A.A. Vopros priznaniia sovetskogo gosudarstva v politicheskoi diskussii britanskogo parlamenta [The recognition of the Soviet state in the political debate of the British Parliament] // *Veter Perestroiki – 2022 [The Wind of Perestroika – 2022]. Sbornik materialov Vtoroj Vserossiiskoi nauchnoi konferentsii [Collected articles of the second national scientific conference]/ A. D. Matlin (otvetstvennyi redaktor) [ed.-in-chief A. D. Matlin]*. Sankt-Peterburg, 2023. P. 22–27.
4. Golovina, N.M. «Neparlamentskie vyrazheniia» i rechevaia agressiia v britanskom parlamente: ritoricheskaia strategiia ili institutsional'naia norma? [Unparliamentary language and verbal aggression in the British Parliament: rhetorical strategy or Institutional norm?] // *Rech' i iazyki obshcheniia v konfliktogennom mire. Materialy mezhdunarodnoi nauchno-prakticheskoi konferentsii. [Speech and languages of communication in a conflict-prone world. Proceedings of an international research-to-practice conference]* / S.V. Myskin (otv. red.) [ed.-in-chief S.V. Myskin]. Moskva, 2021. P. 37–39.
5. Zakharova, O.V. Obsuzhdenie migratsionnoi politiki v britanskom parlamente. [Debates on Migration Policy in the British Parliament] // *Chelovek, obraz, slovo v kontekste istoricheskogo vremeni i prostranstva. Materialy Vserossiiskoi nauchno-prakticheskoi konferentsii [Man, image and word in the context of historical time and space. Proceedings of an international research-to-practice conference]*. 2015. P. 93–96.
6. Ziubina, I.A., Maslova, V.A. Realizatsiia kommunikativnykh strategii v britanskom parlamente [The implementation of communication strategies in the British Parliament] // *Ural'skii nauchnyi vestnik [The Urals Science Bulletin]*. 2023. Vol. 6. № 6. P. 53–60.
7. Kovaliov, N.A., Ches, N.A. «SVOI» versus «CHUZHIE»: dinamika razvitiia i manipuljativnyi potentsial kontsepta KHOLODNAIA VOINA v angloiazychnom politicheskom diskurse [Us vs Them: the Development Dynamics and Manipulative Potential of the Concept "Cold War" in Russian and English-Language Political Discourse] // *Vestnik Rossiiskogo universiteta druzhby narodov. Seriia: Teoriia iazyka. Semiotika. Semantika [Peoples' Friendship University of Russia Bulletin. Language theory, semiotics and semantics]*. 2017. Vol 8, №4. P. 1171–1178.
8. Koretskaia, O.V. O nekotorykh politicheskikh evfemizmakh v epokhu postpravdy (na primere angliiskogo iazyka) [On select English political euphemisms in an age of post-truth] // *Filologicheskie nauki v MGIMO [Linguistics & Polyglot Studies]*. 2021. Vol. 7, № 3 (27). P. 16–23.
9. Kornilov, A.A., Lobanova, N.S., Egorov, A.I. Britanskii parlament kak tsentr vyrabotki vneshnepoliticheskikh reshenii v period siriiskogo krizisa (2011-2015 gody) [The Role of the British Parliament in foreign policymaking during the Syria crisis of 2011-2015] // *Nauchnyi dialog [Scientific Dialogue]*. 2023. Vol. 12. № 2. P. 363–384.
10. Kornilov, A.A., Lobanova, N.S., Zhernovaia, O.R. Obsuzhdenie palestino-izrail'skogo konflikta v komitete britanskogo parlamenta po inostrannym delam (2014 god) [The Israeli-Palestinian conflict as debated by the Foreign Affairs Committee of the British Parliament in 2014] // *Nauchnyi dialog [Scientific Dialogue]*. 2022. Vol. 11. № 2. P. 437–462.
11. Lobanova, N.S. Kliuchevye terminy dokumentov britanskogo parlamenta v oblasti blizhnevostochnoi politiki: etimologiia, politicheskoe znachenie i primery ispol'zovaniia [Key terms of the Middle East policy employed by the British Parliament: etymology, political significance and usage] // *Regiony mira: problemy istorii, kul'tury i politiki. Sbornik nauchnykh statei. [The world's regions: historical, cultural and political problems. Collected articles]*. Nizhnii Novgorod, 2021. P. 107–112.
12. Lobanova, N.S. Podkhod komiteta po inostrannym delam britanskogo parlamenta k krizisu na Ukraine [The Ukraine crisis as seen by the Foreign Affairs Committee of the British Parliament] // *Nauchno-analiticheskii vestnik Instituta Evropy RAN [The Scientific and Analytical Bulletin of the Institute for Europe of the Russian Academy of Sciences]*. 2023. № 6 (36). P. 7–18.
13. Mikhailov, V.V. Vkhozhdenie Azerbaidzhana v sostav sovetskogo gosudarstva i politika Velikobritanii v otnoshenii Zakavkaz'ia v 1918-1920 gg.: politicheskii i sotsial'no-ekonomicheskii aspekty [Azerbaijan's accession to the USSR and the UK Transcaucasia policy in 1918-1920: Political and socio-economic aspects] // *Uchenye zapiski Krymskogo federal'nogo universiteta imeni V.I. Vernadskogo. Istoricheskie nauki [Proceedings of Vernadsky Crimea Federal University. History Section]*. 2022. Vol. 8. № 2. P. 73–87.
14. Khakhalkina, E.V. "Pokolenie Vindrash» v kontekste sovremennogo razvitiia mul'tirasovoi Velikobritanii (po materialam britanskogo parlamenta) [Windrush generation in the context of the modern development of multiracial Great Britain (based on the materials of the British Parliament)] // *Novaia i noveishaia istoriia [Modern and contemporary history]*. 2022. № 6. P. 180–191.
15. Ches, N.A. Kontseptual'naia metafora v politicheskom mediadiskurse (na materiale angliiskogo iazyka): monografiia [Conceptual Metaphor in English-Language Political Media Discourse]/ N.A. Ches. Moskva: MGIMO-Universitet, [Moscow, MGIMO University] 2020. 190 p.
16. Abercrombie, G., Batista-Navarro, R. A sentiment-labelled corpus of Hansard parliamentary debate speeches // *Proceedings of ParlaCLARIN. Common Language Resources and Technology Infrastructure (CLARIN)*. 2018. P. 43–48.
17. Abercrombie, G., Batista-Navarro, R. Sentiment and position-taking analysis of parliamentary debates: a systematic literature review // *Journal of Computational Social Science.* Vol 3, №1. 2020. P. 245–270.

18. Abercrombie, G., Batista-Navarro, R. 'Aye'or 'no'? Speech-level sentiment analysis of Hansard UK parliamentary debate transcripts // *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* 2018. P. 4173–4180.
19. Abercrombie, G., Batista-Navarro, R. Identifying opinion-topics and polarity of parliamentary debate motions // *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis.* 2018. P. 280–285.
20. Aspinall, P. Ethnic/racial terminology as a form of representation: A critical review of the lexicon of collective and specific terms in use in Britain // *Genealogy*. Vol. 4, № 3. 2020. P. 87–100.
21. Bischof, K., Ilie, C. Democracy and discriminatory strategies in parliamentary discourse // *Journal of Language and Politics.* Vol. 17, № 5. 2018. P. 585–593.
22. Charteris-Black, J. *Metaphor and gender in British parliamentary debates* / J. Charteris-Black. Palgrave Macmillan UK, 2009.
23. Coutto, T. Half-full or half-empty? Framing of UK–EU relations during the Brexit referendum campaign // *Crisis and Politicisation.* Routledge, 2021. P. 85–103.
24. Cribb, M., Rochford, S. The transcription and representation of spoken political discourse in the UK House of Commons // *International Journal of English Linguistics.* Vol. 8, № 2. 2018. P. 1–14.
25. Duthie, R., Budzyńska, K. Classifying types of ethos support and attack // *7th International Conference on Computational Models of Argument*. IOS Press, 2018. P. 161–168.
26. Hiltunen, T. et al. Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century // *Language Sciences.* 2020, https://doi.org/10.1016/j.langsci.2020.101270 (дата доступа 04.03.2024).
27. House of Commons Hansard, https://hansard.parliament.uk/commons Avialable from: https://hansard.parliament.uk/commons (дата доступа 12.09.2023).
28. House of Lords Hansard, https://hansard.parliament.uk/lords. Avialable from: https://hansard.parliament.uk/lords (дата доступа 12.09.2023).
29. Huysmans, J., Alessandra Buonfino A. Politics of exception and unease: Immigration, asylum and terrorism in parliamentary debates in the UK // *Political studies.* Vol. 56, № 4. 2008. P. 766–788.
30. Ihalainen, P., Sahala, A. Evolving conceptualisations of internationalism in the UK parliament: Collocation analyses from the League to Brexit // *Digital Histories: Emergent Approaches within the New Digital History.* 2020. P. 199–219.
31. Ilie, C. Parenthetically speaking: Parliamentary parentheticals as rhetorical strategies // *Dialogue Analysis 2000: Selected Papers from the 10th IADA Anniversary Conference.* Tübingen: Niemeyer, 2003. P. 253–264.
32. Ilie, C. Strategic uses of parliamentary forms of address: The case of the UK Parliament and the Swedish Riksdag // *Journal of pragmatics.* Vol. 42, № 4. 2010. P. 885–911.
33. Jeffries, L., Walker, B. Austerity in the Commons: A corpus critical analysis of austerity and its surrounding grammatical context in Hansard (1803–2015) // *Discourse Analysis and Austerity.* Routledge, 2019. P. 53–79.
34. Kettell, S., Kerr, P. From eating cake to crashing out: constructing the myth of a no-deal Brexit // *Comparative European Politics.* 2020. Vol. 18. P. 590–608.
35. Labat, S., Kotze, H., Szmrecsanyi, B. Processing and prescriptivism as constraints on language variation and change: Relative clauses in British and Australian English parliamentary debates // *Exploring Language and Society with Big Data: Parliamentary discourse across time and space.* 2023. P. 250–276.
36. Leduc, R. The ontological threat of foreign fighters // *European Journal of International Relations*. 2021. Vol. 27, № 1. P. 127–149.
37. Mair, C. Empire, migration and race in the British parliament (1803–2005) // *Exploring Language and Society with Big Data: Parliamentary discourse across time and space.* 2023. P. 111–118.
38. McGill, E., Saggion, H. BSL-Hansard: A parallel, multimodal corpus of English and interpreted British Sign Language data from parliamentary proceedings // *Proceedings of the Second International Workshop on Automatic Translation for Signed and Spoken Languages*. 2023. P. 38–43.
39. McKenzie-McHarg, A., Fredheim, R. Cock-ups and slap-downs: A quantitative analysis of conspiracy rhetoric in the British Parliament 1916–2015 // *Historical Methods: A Journal of Quantitative and Interdisciplinary History.* 2017. Vol. 50, № 3. P. 156–169.
40. Mollin, S. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts // *Corpora*. 2007. Vol. 2, № 2. P. 187–210.
41. Onyimadu, O. et al. Towards sentiment analysis on parliamentary debates in Hansard // *Semantic Technology: Third Joint International Conference, JIST 2013, Seoul, South Korea, November 28-30, 2013. Revised Selected Papers.* Vol. 3. Springer International Publishing, 2014. P. 48–50.
42. Riihimäki, J. At the heart and in the margins: Discursive construction of British national identity in relation to the EU in British parliamentary debates from 1973 to 2015 // *Discourse & Society.* 2019. Vol. 30, № 4. P. 412–431.
43. Thundyill, S. et al. Moving Fingers Write History and Having Writ Become Digital: Towards a Big Data Framework for the Analysis of Parliamentary Proceedings // *Future of Information and Communication Conference.* Cham: Springer Nature Switzerland. 2023. P. 459–479.
44. Van Dijk, T. Political identities in parliamentary debates // *European parliaments under scrutiny: Discourse strategies and interaction practices*. 2010. P. 29–56.
45. Willis, R. Taming the climate? Corpus analysis of politicians' speech on climate change // *Environmental Politics.* Vol. 26, № 2. 2017. P. 212–231.

Сведения об авторе:

Гагарин Сергей Николаевич − кандидат филологических наук, старший преподаватель кафедры английского языка №1 МГИМО МИД России (Россия, Москва). Сфера научных интересов: корпусная лингвистика, обработка искусственного языка, большие данные, когнитивная лингвистика.
Email: lexicogr@mail.ru. ORCID: 0000-0002-8893-5083

About the author:
Sergey N. Gagarin, Candidate of Philology, is Senior Lecturer of the English Language Department №1 at MGIMO University (Moscow, Russia). Research interests: corpus linguistics, natural language processing, big data, cognitive linguistics. Email: lexicogr@mail.ru. ORCID: 0000-0002-8893-5083

Автор заявляет об отсутствии конфликта интересов.
The author declares no conflicts of interest.

* * *