

# Getting a Handle on Hansard with Python and NLTK, or How to Tame the Linguistic Picture of British Politics with NLP<sup>1</sup>

Sergey N. Gagarin

MGIMO UNIVERSITY

**Abstract.** This article proposes an optimised starter's set of basic Python and NLTK (Natural Language Toolkit) methods that are essential to 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 visualization 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 customize 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:** pcorpus linguistics, natural language processing, big data, cognitive linguistics, parliamentary discourse

---

1 English translation from the Russian text: Gagarin S. N. 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). P. 125–140. DOI: <https://doi.org/10.24833/2410-2423-2024-2-39-125-140>

Every country has its own unique linguistics, and Russia is no exception. For decades, Russian linguistics has been dominated, among other areas, by the cognitive approach. A key role in this branch of linguistics is traditionally accorded to the study of the verbalization of concepts. Since these mental constructs, which allow human consciousness to organize how we perceive reality, cannot be observed directly due to natural reasons, their study is largely reduced to the analysis and categorization of vocabulary that reflects them in languages, and of the linguistic pictures that are formed from this vocabulary.

On the one hand, we can say with some certainty that Russian linguists are as interested as ever in developing a systematic understanding of the specific verbalization of concepts. On the other hand, it is clear that the immensity of the empirical material means that such studies cannot be as effective or detailed as one might like. We thus find ourselves, somewhat ironically, at the mercy of a kind of “lexicologist’s curse,” where a comprehensive and representative study would require such time and effort that it beggars common sense to even embark upon one. The material available is so vast that almost any study on the verbalization of concepts, where hypotheses are put forward about the patterns that are characteristic of a particular language as a whole, would immediately be in a vulnerable position due to the unrealistic nature of fully implementing a research task of this level. In other words, researchers of linguistic pictures of the world often find themselves biting off far more than they can chew.

One obvious way to solve the problem of ensuring that a study is sufficiently representative is to narrow the empirical base of the study to a more sensible, and manageable, level. In this regard, the question about the optimal limits of this narrowing arises. The problem is that, attractive as compact material may be, this might not be the best option. After all, it is the convenient choice, and, as such, it may entice far too many researchers and thus be overstudied at the present juncture. This could make it unsuitable for linguists who value novelty in their research over the opportunity to have it published. We can thus assume that the optimal narrowing of the empirical base for studying the verbal representation of concepts should be large enough that most researchers would not want to tackle it, and at the same time not so large that someone might risk studying at the current stage of scientific and technological development. If implemented successfully, this concept could provide an important advantage: a research comfort zone that is outside the comfort zone of most other researchers (“competitors”) and located in the realm of Big Data, for example, in large text corpora containing hundreds of millions of words. Analysing such volumes of unstructured Big Data requires machine processing, something that is impossible without knowledge of programming languages and specialized libraries. For tasks that involve processing massive corpora, the Python programming language and the Natural Language Toolkit (NLTK) library can be considered the optimal combination.

We can state with confidence that the ability or inability to use these two tools is one of the factors that largely define the nature of the study of the same corpus data in

different countries. As evidence of this, we need look no further than the procedure for carrying out corpus studies of official reports on meetings of the British Parliament, also known as Hansard reports.

There are currently three main areas of Hansard research abroad: analysis of the linguistic representation of concepts and constructs (Coutto 2021; Huysmans, Buonfino 2008; Ihalainen, Sahala 2020; Jeffries, Walker 2019; Kettell, Kerr 2020; Leduc 2021; Mair 2023; Riihimäki 2019; Van Dijk 2010; Willis 2017); analysis of the tonality and emotional colouring of the text (Abercrombie, Batista-Navarro 2018a; Abercrombie, Batista-Navarro 2020 ; Abercrombie, Batista-Navarro 2018 b; Abercrombie, Batista-Navarro 2018c ; Duthie, Budzyńska 2018; Onyimadu et al. 2014; and analysis of the types of rhetoric and communication strategies employed (Bischof, Ilie 2018; Ilie 2003, 2010; McKenzie-McHarg, Fredheim 2017). Due to the scale of their goals, objectives, and empirical base, the overwhelming majority of these studies would not be possible without the use of quantitative analytical methods based on the capabilities of Python and NLTK. At the same time, many hi-tech research areas that rely on the use of Python and NLTK and are firmly embedded in the foreign practice of Hansard research are not yet found in Russian linguistic research. We are talking primarily about analyses of the accuracy of the transcriptions of parliamentary sessions (Cribb, Rochford 2018; Mollin 2007), and of the technical aspects of collecting and processing Big Data represented by textual material of parliamentary debates (Labat, Kotze, Szmrecsanyi 2023; McGill, Saggion 2023; Thundyill, et al. 2023).

The most common method of studying Hansard reports in Russia does not require reliance on programming languages and specialized libraries. Researchers approach the study of parliamentary transcripts as historians, and the materials are thus treated as historical sources (Aizenshtat 2016a, 2016b; Bykova, Sigova 2023; Kornilov, Lobanova, Egorov 2023; Kornilov, Lobanova, Zhernovaia 2022; Lobanova 2023; Mikhailov 2022; Khakhalkina 2022).

We should note in particular that the Russian practice of studying Hansard aligns with the foreign practice in three main areas: analyses of types of rhetoric and communication strategies (Golovina 2021; Ziubina, Maslova 2023); studies on the linguistic representation of concepts (Kovaliov, Ches 2017; Ches 2020); and lexicological studies (Zakharova 2015; Koretskaia 2021; Lobanova 2021; Aspinall 2020; Charteris-Black 2009; Hiltunen, et al. 2020). The very existence of these areas of study unites Russian and foreign research, but at the same time they are separated by the fact that Russian researchers have likely never used Python and NLTK in their work, which can be concluded from the lack of information about such studies in scientific publications.

The above example of the differences between the Russian and foreign approaches to corpus research indicates that the potential of the Russian schools in this area has not yet been fully realized. As such, we can state that Russian science lags behind in this regard. That said, bridging this gap appears entirely possible if we introduce a culture of using advanced text processing technologies. In this regard, the question about how exactly this should be done and where we should start naturally arises.

## Aim of the Study

The aim of this work is to define a range of basic methods for analysing large corpora of English-language discourse, which would be optimal for identifying the features and patterns of the verbalization of concepts, using Python and NLTK. The materials subject to machine processing are the end-to-end diachronic parallel corpora of reports on meetings of the House of Commons and the House of Lords of the British Parliament compiled by the author of this paper<sup>2</sup>. In terms of quality, these reports are as close as possible to transcripts, recording everything said in both houses of parliament during each day of the session almost verbatim. As of today, both corpora cover the period from July 2006 to July 2023, inclusive. This is because the Hansard archive, which is available on the official website of the British Parliament, is not complete. Some materials for the period prior to July 2006 are missing or unavailable due to internet resource failures. This is why the period under consideration is the only and most up-to-date one that is fully reflected, without any gaps. in the accessible, all-encompassing, and parallel discourse of both houses of the British Parliament.

The House of Commons corpus contains a total of 194,731,646 tokens, while the House of Lords corpus contains 150,866,492 tokens. The total volume is over a third of a billion tokens. Without machine processing, there is no practical use for analysing such material, as no single human life would be enough to do it. At the same time, working with these corpora helped to identify in practice the basic set of methods, based on a combination of Python and NLTK, that fully corresponds to the tasks of studying large volumes of unstructured language data facing modern Russian lexicology and cognitive linguistics.

## Basic Python Toolkit and NLTK

### Tools for Analysing Basic Vocabulary Parameters

At the initial stage of corpus research, there is often a need to assess its quantitative parameters, for example, vocabulary size and level of lexical diversity. These tasks are solved using such basic Python language functions as *len* and *set* (Table 1).

<sup>2</sup> House of Commons Hansard. URL: <https://hansard.parliament.uk/commons> (accessed: 12.09.2023); House of Lords Hansard. URL: <https://hansard.parliament.uk/lords> (accessed: 12.09.2023).

Table 1. Assessment of basic vocabulary parameters (House of Lords Corpus)

```

jupyter House_of_Lords_lexical_diversity (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Not Trusted Python 3

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

```

The *len* function outputs the total number of tokens, and the *set* function collapses duplicates and represents the entire vocabulary of the corpus as a set of unique tokens. In the example above, the lexical diversity index is accomplished using the *lords\_lexical\_diversity* function, which divides the total number of tokens in the corpus by the number of unique tokens.

When carrying out a comprehensive analysis of large parliamentary corpora, a parameter such as the level of lexical diversity can be particularly useful in the context of comparative studies involving a quantitative comparison of the vocabulary of parliamentary discourse in different countries. This parameter may also be of interest in a comparative analysis of more specific segments of the linguistic picture of politics. Specifically, it can be used to carry out a comparative analysis of the degree of diversity of the entire vocabulary used in parliamentary debates by prime ministers, ministers of foreign affairs, finance, and defence, and other representatives of the political leadership of one or more countries throughout their political careers. This will allow us to form an objective idea about the comparative diversity of their vocabularies and, consequently, about the comparative complexity of their individual linguistic worldviews.

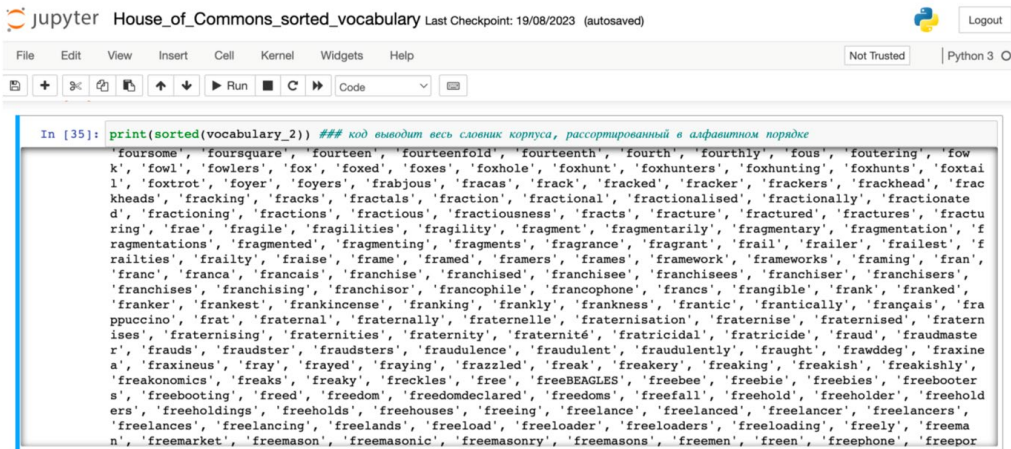
At first glance, such information may be of purely theoretical and very niche interest, but it has an obvious practical value. It is related directly to didactic goal-setting in the teaching of foreign languages at the university level. It is important to note here that quality standards for foreign language programmes at Russian universities ignore such an objective quantitative indicator as the lexical diversity coefficient of the vocabulary that needs to be learned. Including this indicator could provide a clearer idea of what lexical goals should be set for students. This parameter would allow for a more objective assessment of how close the vocabulary of, say, at postgraduate student at Moscow State Institute of International Relations (MGIMO) studying English is in terms of lexical diversity to that of a given prime minister or foreign secretary of the United Kingdom, or to the average indicator of lexical diversity for all those who have held these positions over the past 20 or 30 years, or any other period that corresponds

most closely to the established goals and objectives. In the long term, this approach could add a new dimension to the methods used to assess the quality of foreign language programmes at educational institutions. In addition to the advantages of this approach, it is important to point out one very significant drawback – the fact that such a concept can hardly be implemented in practice without having to introduce, on a mass scale, speech recognition systems and the automated compilation of individual corpora of all the words spoken by each student in foreign language classes into the technical side of the educational process. Such systems require the extensive use of artificial intelligence, and, unfortunately, developing and maintaining them would be prohibitively expensive.

### Terminology Extraction Tools

One of the most important stages in the work of a corpus lexicologist and lexicographer is the formation of a complete vocabulary of the corpus under study. This has always been a complex and cumbersome task, but NLTK simplifies the process greatly. With NLTK, vocabulary elements are extracted as dictionary keys from key-value pairs, where the value is lexical frequency (Table 2).

**Table 2. Extracting the complete vocabulary of the House of Commons Corpus for the last 17 Years (excerpt of results)**



The screenshot shows a Jupyter Notebook window titled 'jupyter House\_of\_Commons\_sorted\_vocabulary Last Checkpoint: 19/08/2023 (autosaved)'. The interface includes a menu bar (File, Edit, View, Insert, Cell, Kernel, Widgets, Help) and a toolbar with icons for file operations, running code, and viewing output. The code cell contains the following Python code:

```
In [35]: print(sorted(vocabulary_2)) ### код выводит весь словарь корпуса, рассортированный в алфавитном порядке
```

The output of the code is a long list of words, including: 'foursome', 'foursquare', 'fourteen', 'fourteenfold', 'fourteenth', 'fourth', 'fourthly', 'fous', 'fouting', 'fow', 'fowl', 'fowlers', 'fox', 'foxed', 'foxes', 'foxhole', 'foxhunt', 'foxhunters', 'foxhunting', 'foxhunts', 'foxtail', 'foxtrot', 'foyer', 'foyers', 'frabjous', 'fracas', 'frack', 'fracked', 'fracker', 'frackers', 'frackhead', 'frackheads', 'fracking', 'fracks', 'fractals', 'fraction', 'fractional', 'fractionalised', 'fractionally', 'fractionated', 'fractioning', 'fractions', 'fractious', 'fractiousness', 'fracts', 'fracture', 'fractured', 'fractures', 'fracturing', 'frae', 'fragile', 'fragilities', 'fragility', 'fragment', 'fragmentarily', 'fragmentary', 'fragmentation', 'fragmentations', 'fragmented', 'fragmenting', 'fragments', 'fragrance', 'fragrant', 'frail', 'frailer', 'frailest', 'frailties', 'frailty', 'fraise', 'frame', 'framed', 'framers', 'frames', 'framework', 'frameworks', 'framing', 'fran', 'franc', 'franca', 'francais', 'franchise', 'franchised', 'franchisee', 'franchisees', 'franchiser', 'franchisers', 'franchises', 'franchising', 'franchisor', 'francophile', 'francophone', 'francs', 'frangible', 'frank', 'franked', 'franker', 'frankest', 'frankincense', 'franking', 'frankly', 'frankness', 'frantic', 'frantically', 'français', 'frappuccino', 'frat', 'fraternal', 'fraternally', 'fraternelle', 'fraternisation', 'fraternise', 'fraternised', 'fraternises', 'fraternising', 'fraternities', 'fraternity', 'fraternité', 'fratricidal', 'fratricide', 'fraud', 'fraudmate', 'frauds', 'fraudster', 'fraudsters', 'fraudulence', 'fraudulent', 'fraudulently', 'fraught', 'frawddeg', 'fraxinea', 'fraxineus', 'fray', 'frayed', 'fraying', 'frazzled', 'freak', 'freakery', 'freaking', 'freakish', 'freakishly', 'freakonomics', 'freaks', 'freaky', 'freckles', 'free', 'freeBEAGLES', 'freebee', 'freebie', 'freebies', 'freebooter', 'freebooting', 'freed', 'freedom', 'freedomdeclared', 'freedom', 'freefall', 'freehold', 'freeholder', 'freeholders', 'freeholdings', 'freeholds', 'freehouses', 'freeing', 'freelance', 'freelanced', 'freelancer', 'freelancers', 'freelances', 'freelancing', 'freelands', 'freeload', 'freeloader', 'freeloaders', 'freeloading', 'freely', 'freema', 'freemarket', 'freemason', 'freemasonic', 'freemasonry', 'freemasons', 'freemen', 'free', 'freephone', 'freepor'.

One of the most significant advantages of being able to readily access the complete vocabulary of the entire corpus is that it makes work with foreign-language borrowings, for instance, so much easier. This is particularly important if the source of borrowing is a language that the researcher does not speak. Typical examples include borrowings from Irish Gaelic used in English-language debates in the Assembly of Ireland, and borrowings from Māori (so-called Maorisms) in parliamentary discourse in New Zealand. The number of Russian specialists in these languages is extremely small,

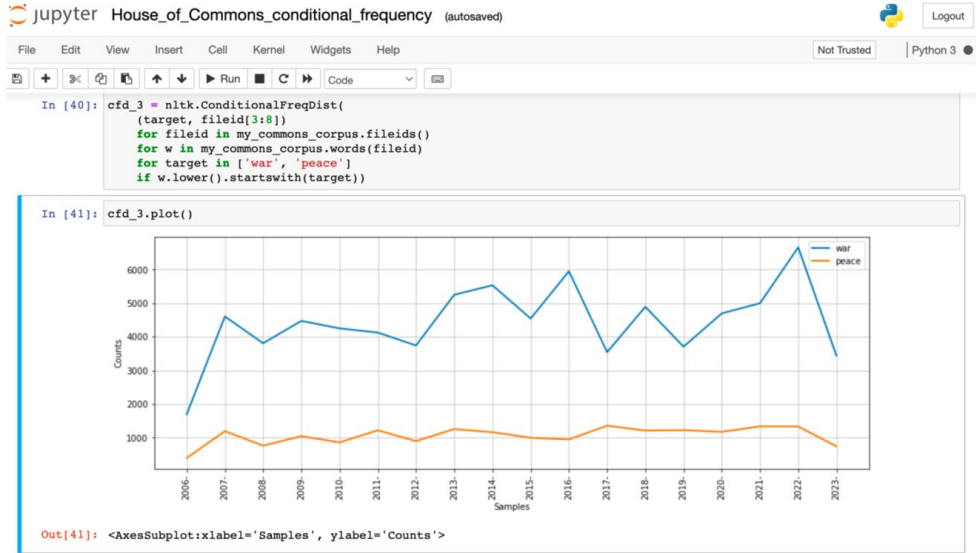


and the languages themselves are rightly considered extremely rare in this country. Even so, the lack of knowledge of Māori cannot be considered a serious obstacle to extracting Māori words from the English-language corpus of transcripts of New Zealand parliamentary proceedings. To accomplish this, the entire vocabulary of the corresponding corpus is output as an alphabetically sorted list, and then processed using WordNet or similar lexical databases. This is a preliminary stage of the process, allowing for the majority of English words to be automatically removed from the lexicon. The final part involves further manual processing. After this, the resulting vocabulary is sorted into lexical-semantic groups using a translation dictionary and an explanatory dictionary. This provides an idea of which concepts are verbalized with its help and what gaps in English-language discourse it fills. This is a simple yet rather effective approach that has been successfully applied to the study of Maorisms in New Zealand parliamentary discourse as part of the work being carried out under the supervision of the author of this paper at English Department No. 1 of MGIMO University of the Ministry of Foreign Affairs of the Russian Federation.

### Word Frequency Analysis Tools

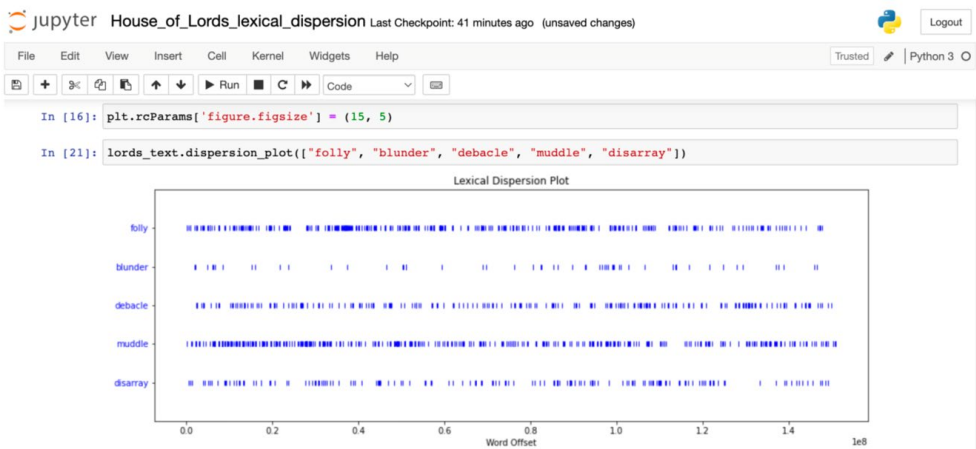
After extracting a vocabulary from a corpus, it may be necessary to understand the frequency distribution of the words that make it up in the material under study. Tasks of this nature are effectively solved using the *FreqDist* and *ConditionalFreqDist* functions. Both are effective for non-concatenated words. That is, preliminary end-to-end assembly of all materials in chronological order is not needed in order to use them, which, in turn, saves the researcher time. The *ConditionalFreqDist* function is used to analyse conditional frequency distribution, which is especially convenient in the end-to-end diachronic approach, when the distribution condition is the year or month of publication of the text, which is indicated in the name of each individual file in the format *yyyy\_mm\_dd* (Table 3). This method is also quite useful for visualizing language data, which can then be used to formulate preliminary hypotheses for research in cognitive linguistics or predictive analytics. The example below illustrates this feature using the conditional frequency distribution for the words *war* and *peace* in the House of Commons Corpus. Two graphs that clearly illustrate the changes in the frequency of their use over the years can provide a basis for hypotheses about the correlation of their peak values with extra-linguistic events, as well as for hypotheses about the strengthening of the tendency towards metaphorical use, which these meanings reflect. In any case, these extremes can be seen as indicative of the most informative segments of the corpus, which can provide the greatest amount of lexical material illustrating the features of the verbalization of two concepts of political discourse, which are traditionally among the most important and are thus of great interest to researchers.

**Table 3. Conditional distribution of the frequencies of the words war and peace in the House of Commons Corpus by year**



Often, when there are too many graphs on the same coordinate axes, it can become difficult to make sense of them. This is why alternative methods of visually representing frequencies are needed. In such cases, it may be worth resorting to the visualization of lexical dispersion using a concatenated corpus (that is, one that has been assembled as a single file). The *dispersion\_plot* function is used for this purpose (Table 4).

**Table 4. Visualization of the lexical dispersion of the words *folly*, *blunder*, *debacle*, *muddle* and *disarray* in the House of Lords Corpus**





This method of visualization is especially convenient for visually displaying the features of concept verbalization or the comparative frequency of elements of lexical-semantic groups. While less informative and accurate in terms of representing frequency extremes than a traditional line diagram, it nevertheless allows the researcher to create an accurate representation of the relative density of distribution of a given vocabulary item in the text.

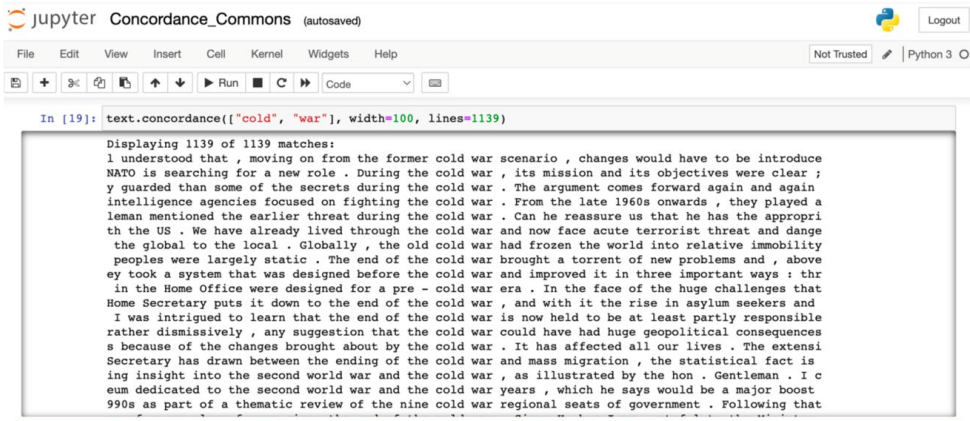
The example given above illustrates the features of lexical dispersion for a group of nouns united by a common pragmatic function – to exert psychological pressure on the listener through harsh criticism of their actions or performance. The dictionary definitions of these nouns give no indication of how popular they might be in contemporary usage in the UK Parliament. However, a lexical dispersion plot clearly shows that the most frequent lexemes in this group throughout the period of observation were *folly* (objectifying the concept of foolishness) and *muddle* (objectifying the concept of disorder). This information can serve as a basis for formulating a working hypothesis that lexical-semantic groups that verbalize the concepts of foolishness and disorder can be considered objects of priority study for determining the vocabulary that is best suited to the task of exerting pressure on political opponents.

### Tools for Working with Lexical Compatibility

Lexical frequency patterns play a key role in the comprehensive analysis of corpora. However, for some areas of research, lexical concordance patterns prove to be no less – and often far more – significant. At the basic level, they are extracted as concordances of words or phrases (Table 5), which can then be saved as separate files and used as mini-corpora of sorts for studying the immediate contextual environment of key words.

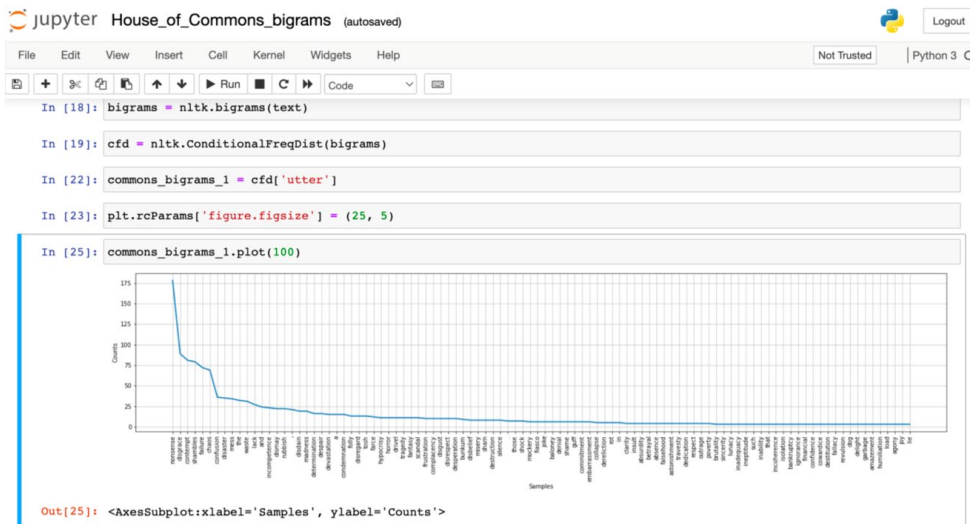
The *concordance* function can be particularly useful when it is important to discover whether a given lexeme is used exclusively as an element of an idiom. It can also help determine whether a particular word is used in a non-dominant language of the corpus outside of fixed and repetitive formulations. A typical example here is the discourse of the New Zealand Parliament, where prayers are read in Māori before each sitting, and where there are cases of the oath of office being recited in Māori and official condolences on the death of members of parliament also containing Māori farewell formulations. From the historical and cultural point of view, they are all of great interest. But for someone who is researching borrowings from the Māori language into the New Zealand variant of English, they are more of a hindrance, since the vocabulary used in them is invariably set Māori phrases, which are foreign-language inclusions and do not belong to the category of borrowings proper. Accordingly, before determining the spectrum of full-fledged lexical borrowings, we must identify these formulas and their components, and the *concordance* function allows us to solve this problem effectively.

Table 5. Using the *concordance* function to construct concordance for the collocation *cold war*, 1139 lines (excerpt)



Another effective tool for analysing lexical compatibility is the bigram (the ***bigrams*** function), especially when it is used in combination with frequency graphs visualizing them. In the example below (Table 6), the bigrams function is used to determine the hundred most commonly used bigrams with the initial word *utter*.

Table 6. Visualization of bigram frequency. The 100 most commonly used combinations with the word utter, House of Commons Corpus



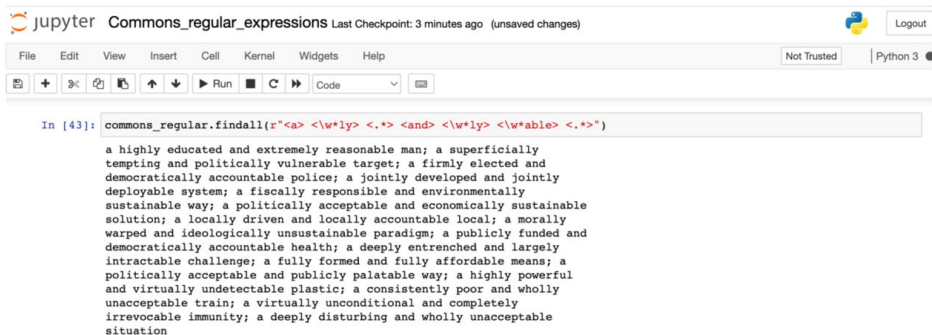
The second components of the bigrams are located along the x-axis in descending order of frequency. It is especially important to note that bigrams are useful when searching for set phrases and ranking them according to the degree of collocative stability. It should be especially noted that such a tool as bigrams is very convenient for searching for stable phrases and ranking them according to the degree of collocative

stability. At the same time, as the example below clearly shows, the picture of this stability will be far more useful for a researcher of the language of politics than the information obtained about the adjective *utter* from a dictionary entry or set of dictionary entries, since it says nothing about the generally accepted idea of the compatibility of the word *utter* for the English language, but rather corresponds to an objective picture of its contextual use in the sociolect of English that is characteristic of representatives of the UK Parliament.

In the context of studying the language of politics and parliamentary discourse, the use of the **bigrams** function can be especially important for clarifying and systematizing poorly researched features of compatibility of both terminological and non-terminological vocabulary.

When bigrams no longer satisfy the criteria of flexibility and universality required for optimal word searchers, the corpus can be explored using regular expressions. This tool uses metacharacters and makes the detection of almost any n-gram possible, since it allows the researcher to search for sequences of tokens according to a repeating pattern, which can simultaneously contain known, unknown, and partially known elements (Table 7).

**Table 7. A regular expression for finding n-grams with two known, two unknown, and three partially known elements. House of Commons Corpus**



```

In [43]: commons_regular.findall(r"<a> <\w+ly> <.*> <and> <\w+ly> <\w+able> <.*>")

a highly educated and extremely reasonable man; a superficially
tempting and politically vulnerable target; a firmly elected and
democratically accountable police; a jointly developed and jointly
deployable system; a fiscally responsible and environmentally
sustainable way; a politically acceptable and economically sustainable
solution; a locally driven and locally accountable local; a morally
warped and ideologically unsustainable paradigm; a publicly funded and
democratically accountable health; a deeply entrenched and largely
intractable challenge; a fully formed and fully affordable means; a
politically acceptable and publicly palatable way; a highly powerful
and virtually undetectable plastic; a consistently poor and wholly
unacceptable train; a virtually unconditional and completely
irrevocable immunity; a deeply disturbing and wholly unacceptable
situation

```

This tool can be especially useful when researching the style of speech of individual politicians, or comparative studies of the speech behaviour of political leaders who held the same or various posts. The use of regular expressions makes it possible to conduct comparative studies of the speech patterns of foreign ministers and, if necessary, compare them with speech patterns that are characteristic of defence ministers (or any other members of the cabinet) in order to analyse the influence of government post on the composition and structure of figures of speech.

It is particular important to stress that regular expressions can be considered as a tool for corpus analysis, allowing the researcher to identify the n-grams that are most characteristic in terms of their typology for a particular type of discourse. The next step after this could be a systematic analysis of the vocabulary used in their composition to see if it belongs to the lexical-semantic groups that verbalize the key concepts

of individual segments of the linguistic picture of the world. Thus, our understanding of the patterns and trends of their verbalization can be transferred to a higher level of systemic perception.

## Conclusions

Objective conditions demand that modern Russian linguists will most likely have to include programming languages and specialized vocabularies in their research toolkits. At the same time, the scope of their functionality is extremely broad, and it is not always obvious what needs to be studied first for the successful practical application of the research in a given area. There are many areas of linguistics where we can point to the very real absence of clear research protocols using the Python programming language and the NLTK library. In this paper, we attempted to put forward a basic set of tools that could be used to form, flexibly and adaptively, such protocols in the near future for scientific schools and areas of cognitive linguistics that involve carrying out corpus studies of the features of the verbalization of concepts in discourse. The tools we proposed have proven highly effective when working with large text data. The researcher who has mastered this basic set of applied methods of corpus analysis will inevitably have to make a choice: either to stop at the technical level that has been achieved, or to continue to develop this knowledge. In the first case, the researcher is left with a technical toolkit that, despite its limited nature, nevertheless gives them a tangible advantage over their competitors. In the second case, the researcher moves to a fundamentally new level, as they concentrate their efforts on mastering programming areas related to the automatic marking of text corpora using tags, automatic text classification, predictive analysis, training neural networks, etc. As such, the work of this researcher will move away from classical linguistics, and one of their greatest contributions to the profession could be quality of the large language models (LLMs) they create. No matter how committed the researcher is to integrating programming into their work, and no matter which of the two paths of further professional development they are prepared to follow, they can rest assured that this is a win-win situation, at least in the short term.

### **About the Author:**

**Sergey N. Gagarin** – Candidate of Philology, a Senior Lecturer at English Language Department No. 1, MGIMO University, Moscow, Russia. Research interests: corpus linguistics, natural language processing, Big Data, cognitive linguistics. Email: [lexicogr@mail.ru](mailto:lexicogr@mail.ru). ORCID: 0000-0002-8893-5083

### **Conflicts of interest**

The author declares no conflicts of interest.

## References:

- Aizenshtat M. P. 2016a. Novatsii v parlamentskoi praktike Britanii XVIII stoletia [Innovations in Britain's Parliamentary practice of the 18th and 19th centuries]. In *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]. P. 7–13.
- Aizenshtat M. P. 2016b. Parlamentskie materialy Britanii XVII–XIX vekov. Zaprety i preodoleniia. [Britain's parliamentary materials of the 18<sup>th</sup>–19<sup>th</sup> centuries. Prohibitions and how they were overcome]. *Novaia i noveishaia istoriia* [Modern and contemporary history]. 5. P. 16–25.
- Bykova E. A., Sigova A. A. 2023. 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]. In A. D. Matlin (Ed.), *Sbornik materialov Vtoroj Vserossiiskoi nauchnoi konferentsii* [Collected articles of the second national scientific conference] (otvetstvennyi redaktor). P. 22–27.
- Golovina N. M. 2021. «Neparlamentskie vyrazheniia» i rechevaia agressiia v britanskom parlamente: ritoricheskaia strategii ili institutsional'naia norma? [Unparliamentary language and verbal aggression in the British Parliament: Rhetorical strategy or institutional norm?]. In Myskin S.V. (Ed.) *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]. P. 37–39.
- Zakharova O. V. 2015. 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]. P. 93–96.
- Ziubina I. A., Maslova V. A. 2023. Realizatsiia kommunikativnykh strategii v britanskom parlamente [The implementation of communication strategies in the British Parliament]. *Ural'skii nauchnyi vestnik* [The Urals Science Bulletin]. 6(6). P. 53–60.
- Kovaliov N. A., Ches N.A. 2017. «SVOI» versus «CHUZHIE»: dinamika razvitiia i manipulativnyi potentsial kontsepta KHOLODNAIA VOINA v angliazychnom 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 družby narodov. Seriia: Teoriia iazyka. Semiotika. Semantika* [RUDN Journal of Language Studies, Semiotics and Semantics]. 8(4). P. 1171–1178.
- Koretskaia O. V. 2021. O nekotorykh politicheskikh evfemizmakh v epokhu postpravdy (na primere angliiskogo iazyka) [On some political euphemisms in the post-truth era (a case study of the English language)]. *Filologicheskie nauki v MGIMO* [Linguistics & Polyglot Studies]. 7(3). P. 16–23.
- Kornilov A. A., Lobanova N. S., Egorov A. I. 2023. 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]. 12(2). P. 363–384.
- Kornilov A. A., Lobanova N. S., Zhernovaia O. R. 2022. 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]. 11(2). P. 437–462.
- Lobanova N. S. 2021. Kliuchevye terminy dokumentov britanskogo parlamenta v oblasti blizhnovostochnoi politiki: etimologiya, politicheskoe znachenie i primery ispol'zovaniia [Key terms of the Middle East policy employed by the British Parliament: etymology, political significance and usage]. In *Regiony mira: problemy istorii, kul'tury i politiki. Sbornik nauchnykh statei* [The world's regions: historical, cultural and political problems. Collected articles]. P. 107–112.



Lobanova N. S. 2023. 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]. 6(36). P. 7–18.

Mikhailov V. V. 2022. Vkhodzenie 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* [Scientific Notes of V.I. Vernadsky Crimean Federal University. Historical Science]. 8(2). P. 3–87.

Khakhalkina E. V. 2022. «Pokolenie Vindrush» 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]. 6. P. 180–191.

Ches N. A. 2020. *Kontseptual'naia metafora v politicheskom mediadiskurse (na materiale angliiskogo iazyka): monografiia* [Conceptual Metaphor in English-Language Political Media Discourse]. MGIMO University.

Abercrombie G., Batista-Navarro R. 2018a. A sentiment-labelled corpus of Hansard parliamentary debate speeches. *Proceedings of ParlaCLARIN. Common Language Resources and Technology Infrastructure (CLARIN)*. P. 43–48.

Abercrombie G., Batista-Navarro R. 2020. Sentiment and position-taking analysis of parliamentary debates: A systematic literature review. *Journal of Computational Social Science*. 3(1). P. 245–270.

Abercrombie G., Batista-Navarro R. 2018b. '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)*. P. 4173–4180.

Abercrombie G., Batista-Navarro R. 2018c. Identifying opinion-topics and polarity of parliamentary debate motions. *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*. P. 280–285.

Aspinall P. 2020. Ethnic/racial terminology as a form of representation: A critical review of the lexicon of collective and specific terms in use in Britain. *Genealogy*. 4(3). P. 87–100.

Bischof K., Ilie C. 2018. Democracy and discriminatory strategies in parliamentary discourse. *Journal of Language and Politics*. 17(5). P. 585–593.

Charteris-Black J. 2009. *Metaphor and gender in British parliamentary debates*. Palgrave Macmillan.

Coutto T. 2021. Half-full or half-empty? Framing of UK–EU relations during the Brexit referendum campaign. In Voltolini B., Natorski M., Hay C. (Eds.) *Crisis and Politicisation* Routledge. P. 85–103.

Cribb M., Rochford S. 2018. The transcription and representation of spoken political discourse in the UK House of Commons. *International Journal of English Linguistics*. 8(2). P. 1–14.

Duthie R., Budzyńska K. 2018. Classifying types of ethos support and attack. *7th International Conference on Computational Models of Argument*. IOS Press. P. 161–168.

Hiltunen T. et al. 2020. Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century. *Language Sciences*. DOI: <https://doi.org/10.1016/j.langsci.2020.101270>

Huysmans J., Buonfino A. 2008. Politics of exception and unease: Immigration, asylum and terrorism in parliamentary debates in the UK. *Political Studies*. 56(4). P. 766–788.

Ihalainen P., Sahala A. 2020. Evolving conceptualisations of internationalism in the UK parliament: Collocation analyses from the League to Brexit. In Oiva M., Fridlund M., Paju P. (Eds.) *Digital Histories: Emergent Approaches within the New Digital History*. Helsinki University Press. P. 199–219.



- Ilie C. 2003. Parenthetically speaking: Parliamentary parentheticals as rhetorical strategies. *Dialogue Analysis 2000: Selected Papers from the 10th IADA Anniversary Conference*. P. 253–264.
- Ilie C. 2010. Strategic uses of parliamentary forms of address: The case of the UK Parliament and the Swedish Riksdag. *Journal of Pragmatics*. Vol. 42(4). P. 885–911.
- Jeffries L., Walker B. 2019. Austerity in the Commons: A corpus critical analysis of austerity and its surrounding grammatical context in Hansard (1803–2015). In Power K., Ali T., Lebdušková E. (Eds.) *Discourse Analysis and Austerity*. Routledge. P. 53–79.
- Kettell S., Kerr P. 2020. From eating cake to crashing out: Constructing the myth of a no-deal Brexit. *Comparative European Politics*. 18. P. 590–608.
- Labat S., Kotze H., Szmrecsanyi B. 2023. Processing and prescriptivism as constraints on language variation and change: Relative clauses in British and Australian English parliamentary debates. In Korhonen M., Kotze H., and Tyrkkö J. (Eds.) *Exploring Language and Society with Big Data: Parliamentary discourse across time and space*. John Benjamins. P. 250–276.
- Leduc R. 2021. The ontological threat of foreign fighters. *European Journal of International Relations*. 27(1). P.127–149.
- Mair C. 2023. Empire, migration and race in the British parliament (1803–2005). In Korhonen M., Kotze H., Tyrkkö J. (Eds.) *Exploring Language and Society with Big Data: Parliamentary discourse across time and space*. John Benjamins. P. 111–118.
- McGill E., Saggion H. 2023. 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*. P. 38–43.
- McKenzie-McHarg A., Fredheim R. 2017. 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*. 50(3). P. 156–169.
- Mollin S. 2007. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora*. 2(2). P. 187–210.
- Onyimadu O. et al. 2014. *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. P. 48–50.
- Riihimäki J. 2019. 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*. 30(4). P. 412–431.
- Thundyill S. et al. 2023. 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*. P. 459–479.
- Van Dijk T. 2010. Political identities in parliamentary debates. In C. Ilie (Ed.), *European parliaments under scrutiny: Discourse strategies and interaction practices*. John Benjamins. P. 29–56.
- Willis R. 2017. Taming the climate? Corpus analysis of politicians' speech on climate change. *Environmental Politics*. 26(2). P. 212–231.