PRE-PRINT

Hansard at Huddersfield: Adapting Corpus Linguistic Methods for Non-Specialist Use

Alexander von Lünen[a]*, Hugo Sanjurjo González[b], Lesley Jeffries[c], and Fransina Stradling[d]

[a]History, University of Huddersfield, Huddersfield, United Kingdom, [b]Computing, Electronics and Communications Technologies, University of Deusto, Bilbao, Spain, [c and d]Linguistics and Modern Languages, University of Huddersfield, Huddersfield, United Kingdom.

This article introduces the Hansard at Huddersfield web application, which allows users to make a range of different searches of the Hansard data (from 1803 – 2020) and which draws on the insights of corpus linguistics and visualization techniques to appeal to researchers from backgrounds where these approaches are relatively underused. The web application aims to be accessible and includes advice on interpreting the results of searches as well as always allowing the user to access the original Hansard text. This article explains the scope and functionality of the site as well as its architecture, using the search term *Brexit* to illustrate.

Keywords: Hansard, parliamentary debates, corpus applications, data visualization

1. Introduction

The UK's official, substantially verbatim, written report of spoken parliamentary proceedings from both Houses of Parliament is called Hansard. As the authoritative account of what is said in parliament, its central function is to facilitate public scrutiny of the parliamentary decision-making process. While the public may nowadays access parliamentary debates in other formats (e.g. television and social media), Hansard's written records still provide the most accurate and comprehensive access to the language spoken in parliamentary debate. Full machine-readable reports of debates since 1803 are

available online, and since 2010 Hansard has been publishing draft transcripts of daily chamber debates within 3 hours of them finishing. Hansard thus provides an important source of evidence the electorate may use to hold parliament to account without the needing to be present.

Though mostly used by MPs, Peers and the media to find and reference individual contributions to specific debates, Hansard contains a wealth of socio-political, historical and linguistic information regarding parliamentary responses to topics of societal interest. It is therefore ideally suited studying parliament's treatment of societal issues at any one moment (synchronically) or across time (diachronically). In either case, the study of topics and themes in Hansard benefits from investigating language choices to reveal the underlying ideas and ideologies of the speakers (Jeffries, 2010) and add to public understanding of the way parliament addresses issues of social, political and historical interest.

This article introduces the Hansard at Huddersfield (HaH) web application which combines corpus linguistic methods with interactive visualizations to make quantitative and qualitative pattern analysis in Hansard accessible to researchers from non-linguistic fields and the general public. Underlying this approach is the conviction that the many people from outside linguistics who are interested in parliamentary debate could benefit from using corpus analysis techniques to explore the huge amounts of text in Hansard are hampered by the investment of time needed to master corpus software. In section 2 we will describe our approach to creating the HaH web application. We will describe its architecture and functionalities in section 3 using the search term *Brexit* to illustrate the HaH interface. Section 4 will set out our conclusions.

2. Contextualisation

Currently, the most prominent public-facing online versions of Hansard are the official Hansard website (hansard.parliament.uk), and TheyWorkForYou (theyworkforyou.com). These sites provide a limited number of search types which constrain the pattern analysis possible. On both sites, online transcripts can be searched by date, speaker and search word, with raw frequencies computed for hits of individual words but the context is not easily available, being only accessible via clickable debate titles. The official Hansard

website does not present results chronologically, and users cannot order them by speaker or delimit searches for date, debate, speaker and search word together. TheyWorkForYou offers these more sophisticated search options, but like the official Hansard site does not allow comparison of multiple search terms in their immediate contexts or searching for key topics within or across debates without using pre-determined search terms. Both sites offer download of debate files that may then be searched using other software, but this leaves researchers with the task of cleaning the data, integrating it into their software of choice and organizing it for their own purposes. Hansard at Huddersfield, therefore, improves access to effective searches of parliamentary records and simplifies the search process for those unable to undertake their own data preparation.

2.1 Taking a Corpus Linguistic Approach

The HaH web application provides user-friendly searches by adopting aspects of the analytical methodology and presentational features of corpus linguistics and pairing them with open-source visualizations. Our aim in developing the Hansard at Huddersfield interface was to use well-established and well-understood corpus tools such as word frequency lists; dispersion plots; keyword lists and collocation patterns rather than developing new tools. So far, we have exploited the first three of these, with collocation remaining on the list of developments for the next stage of the project.

We were keen that users should not see the results of searches as definitive answers to (research) questions, but as an entry point to the data. There are many potential stumbling blocks, familiar to corpus linguists, requiring such caution, including the assumption that all occurrences of a word represent the same meaning, a danger exacerbated when the corpus extends over a long historical period. Patterns identified by the tools also rely on complex statistical techniques in both computation and presentation so that interpreting the results they produce may require some statistical understanding. Our emphasis, therefore, has been to encourage users to consider the visualizations to be provisional results, best seen as an indicator of patterns of usage in the data which need confirmation by close qualitative investigation of the contextualised search terms. Each of our tools, therefore, allows the user to link through to the context of the search term occurrence. For convenience at the close analysis stage of an investigation, our site

provides alternative presentation formats for this task, including both standard document and optional concordance line format for the immediate context and the option to click through to the whole contribution if more context is needed.

2.2 Related Work

To our knowledge, there are two other major projects that have attempted to make the UK parliamentary record more accessible for analysis: the Hansard Corpus[1] and the Digging into Linked Parliamentary Data (DiLiPaD) project.[2] The Hansard corpus was created as part of the SAMUELS project (2014-2016) (grant reference AH/L010062/1) and contains the UK parliament's speeches between 1803 and 2005. It is searchable using a corpus linguistic front-end hosted by Brigham Young University. Its main advantages are the potential to create personalized corpora containing speeches selected by speaker, time period or topic and the option to search the corpus thematically based on categories from the Historical Thesaurus of English. The corpus is most useful for those well-versed in the use of corpus linguistic software as its search functions require a high-level understanding of linguistics and statistics. This corpus was the launchpad for the Huddersfield project.

The DiLiPad project (2014-2016), led by researchers at universities in the UK, the Netherlands and Canada, aimed to enrich the UK and Canada Hansards with extra metadata and develop tools to study them. The UK arm of the project added party, portfolio and gender metadata to parliamentary proceedings of the House of Commons (from 1935) and developed a beta interface to allow full-text searching of this dataset and visualize its results[3]. Although some research has been published using the text-mining strategies made possible by the DiLiPaD project (see Blaxill & Beelen, 2016), unfortunately their website is no longer available.

Similar platforms exist for the legislative systems of other countries, such as the Italian Law-Making Archive (ILMA) Web Portal (Mesiti et al., 2015) which aims to support scholars in their analysis of the Italian law-making process by providing a system to search Italian legislative data between 1987-2008 along parameters such as type of legislature, date and place of approval, year, number of the law and topics. It also allows the researcher to download sections of the dataset, provides the option to analyse the data

statistically and has options for data visualization. The ILMA Web Portal took inspiration from and improved upon similar applications devised for U.S. Congress and the Italian parliament.

The approach to parliamentary proceedings that HaH takes is therefore not unique in using data visualization or corpus linguistics techniques, but it is unusual in combining the two. Given its aim of providing non-linguist researchers with insights into language choices and patterns across Hansard, the data visualizations on the HaH web application are used to simplify the process of pattern interpretation. Furthermore, they provide entry-points into the actual Hansard data with access only through visualizations and concordances. This approach is not only unique within corpus linguistics, but also in the analysis of parliamentary records.

The approach taken was partly inspired by the *Vision of Britain* website[4] which the first author of this article (von Lünen) was technical lead for between 2007—2012. Like Hansard, it holds substantial amounts of data, in this case geographical and demographic data about Britain between 1801 and 2001 (Southall et al., 2009), and it has been used as a point of departure to create a graphical interface for archive searches (Aucott et al., 2009). Another major influence was the EMOTIVE project, a sentiment analysis tool that produces a fine-grained analysis of emotive words used in texts, particularly in social media messages (Sykora et al., 2014). While working for the EMOTIVE project, von Lünen designed and implemented a visualization system that eventually became the inspiration for the HaH project (Sykora et al., 2015).

3. Architecture and technology

In this section, we explain the technical basis of the interface provided by HaH and demonstrate its functions, using the search term *Brexit* to illustrate its strengths and

drawbacks.

3.1 Database

The current version of the HaH database collates data from the Hansard Corpus, released by the SAMUELS project, comprising debates from 1803 to 2005 with Hansard data from the official UK Hansard API[5] from 2005 to 2020 into one database of Hansard data 1803 to 2020. The HaH project converted this database (8,808,168 speeches over 217 years with around 50,577 speakers[6]) into a relational database. This conversion was necessary to enhance the performance of the searches the HaH front-end offers its users; a relational database favors future database integration with alternative front-ends, makes retrieval of data more efficient and provides a good basis on which to build the site's sustainability.

The HaH relational database was built using PostgreSQL, a well-known open-source object-relational database management system (DBMS) that has superior Full Text indexing (FTI) features compared to other open-source DBMSs. The database is composed of three main schemas: (1) House of Commons, (2) House of Lords and (3) pre-calculated data. Schemas (1) and (2) hold identifying information about all the contributions[7] made in both Houses of Parliament; for each contribution, the date the contribution was made, the member of parliament who made it and the title of the debate it appeared in are recorded. These three parameters were included to allow users to filter queries along these three parameters rather than only using just one search term. Schema (3) holds information about word-frequencies for distributions and other visualizations.

Several additional data structures were added to improve the efficiency of searches, such as various FTI tailored to the queries performed by the system. Furthermore, the database features a table of pre-calculated frequencies of word occurrence across time, to improve the performance time for generating a distribution graph of word frequencies over the whole period covered by the data. At the start, this front-end function required analysis of the complete database for every query, meaning the distribution graph computation was slow. Pre-calculating frequencies of word occurrence per year improved the processing time for single search terms. The high number of variations for multi-word units, however, precludes this approach for this type

of query. We use full-text searches together with indexes to improve performance for multi-word queries which nevertheless remain slower than single word queries.

3.2 Software architecture

The web application has been developed using only well-established open source components to ensure the sustainability of the project. Proprietary software would have made the system harder to replicate for other researchers; it was anticipated that other parties might wish to learn about the techniques used, or the system would be migrated to a different server. Figure 1 shows the different components of the application. In summary, PHP coding was employed for retrieving data from the database and to prepare the data for the visualization pipeline, while the open-source Python library Gensim (Rehurek & Sojka, 2010) was used for NLP tasks. Visualizations were rendered using D3.js, a JavaScript library for visualizing data by means of HTML, CSS and SVG[8]. For the front-end design the Bootstrap CSS framework created by Twitter was used[9].

3.3 Functionalities

The HaH web application allows users to interact with Hansard debates through two main entry points: *Search* and *Explore*. This section will use the example of a frequent word in British politics recently, *Brexit*, to demonstrate the functions, visualizations and potential uses afforded by these entry points.

3.3.1 Search

The *Search* function is the web application's default screen which provides access to Hansard through two search boxes: *Basic Search* and *Advanced Search.* Both options use search terms to find debate contributions which include them. *Basic Search* allows for filtering queries by House of Parliament (Commons and Lords or both) and full years, whereas *Advanced Search* allows for filtering by House, date, speaker and debate title. The Search box in the *Basic Screen* also includes a *Related Terms* button that, when

clicked, provides words that are commonly employed in the same context as the search term. The *Related Terms* feature relies on the fastText library (Bojanowski et al., 2016).

Any search from these screens will produce a distribution graph covering the time period specified by the user. A basic search for *Brexit*, for example, produces a flat line followed by a steep rise in usage between 2015 and 2019 (figure 2) showinga risefrom zero to 932.32 occurrences per million words in five years consistent with an increasing focus in parliament on the referendum about the UK's membership of the European Union.

In the *Advanced Screen,* selecting a date range below 5 years will produce a distribution graph displaying word frequencies by month rather than year. Using this function to search for *Brexit* between 2015-2019*,* we can nuance the picture in figure 2 and see a variable rise in usage to the end of 2019[10] (figure 3). In both search screens, frequency of upto four search terms can be compared and displayed simultaneously on the distribution graph (see, for example, figure 4 for distribution of nouns relating to people involved in *Brexit*: *Brexiteers*, *Remainers* and *Remoaners*).

Selecting a particular date range from graphs like figures 2 or 3 by clicking on two data points will produce a list of all contributions within that date range. The default display of the list of contributions is in a document format, but users can select Keyword in Context (KWIC) format with a default context of 10 words each side. Both the document format and KWIC format display date, speaker, contribution and (optionally) the title of the debate. Clicking on a contribution will produce the complete text of that contribution. For example, the context of *Brexit* (figure 5) shows that it was a countable noun early usages (i.e. *a Brexit*), implying that there could be more than one (type of) Brexit. By February 2016, usage developed into a mixture of countable and proper nouns, (figure 6), and by November 2019 the proper noun became so embedded that nothing in the context of these examples queries what *Brexit* is (figure 7).

Users can download distribution graphs as .png files or as raw data in .csv and Excel files. Contribution lists can also be downloaded, in full or part, as .txt, .csv or Excel files.

3.3.2 Explore

The *Explore* function of HaH allows users to explore Hansard without pre-determined search terms and based on either word frequency or keywords. Both features display

results using an interactive visualization which produces contribution lists in the same formats as the *Search* function, since their programming architecture is linked.

The frequency-based feature uses word clouds to represent the most frequent words in a specific time period, defined by year using a slider, for one or both Houses of Parliament. The font size of words in the word cloud reflects their relative frequency and the algorithm fits in as many words as possible, given differing word lengths and sizes, up to a maximum of 500 most frequent words. Excluded from the visualization are grammatical words (e.g. articles and prepositions) and commonly-used parliamentary words (e.g. Hon, Rt Hon). Users can select up to four words from the word cloud to display in a distribution graph, after which they can access contribution lists in the same way as for the *Search* function. Word clouds provide an entry point into Hansard data for users unsure which search terms to use or to discover and comparing word frequencies. Comparing word clouds for 2016 (figure 8) and 2019 (figure 9), we see the small font size and thus low frequency of *Brexit*, *deal* and *EU* in 2016 compared with 2019, as well as the high frequency of *agreement* in 2019 which does not feature in 2016. Upon inspection of contribution lists, we find that in 2019 *deal* almost exclusively refers to a Brexit deal, while in 2016 it appeared in other contexts.

The keyword-based feature is an adapted corpus linguistic feature allowing users to compute the keywords of a Hansard sub-corpus of choice against another sub-corpus. HaH defines keywords as in corpus linguistics, namely words which occur relatively more frequently in one corpus compared to another (Baker, 2004: 346). The keywords feature differs from the frequency feature in reflecting the saliency of words rather than simple frequency. Users first select their target corpus and then a comparison corpus. We have provided some pre-set corpora (wars and decades) but users can also define their own corpora by selecting the period, House of Parliament, and (optionally) speaker or search term. The keyword function in HaH uses a bubble chart visualization to show keywords for which there is a 99.99% certainty this difference is not by chance, using a log-likelihood statistical test. The user may choose up to four of the keywords on the bubble chart for their contexts to be listed in the same way as before. For example, we can consider the keywords of the period from the referendum on June 23rd 2016 to the end of the data (here November 5th 2019) as compared with the 18 months (from Jan 1st 2015) prior to the referendum. Unsurprisingly, this bubble chart is dominated by *Brexit* (figure 10), but if we compare these periods in reverse, we find a different picture

(figure 11). Though the earlier period hints at Brexit by the presence of *Greek* and *Greece* (Greece was thought to be on a *Grexit* path before Brexit happened), there is a much wider range of topics evident, including Islamic radicalism (*ISIL* and *Daesh*), war in *Syria*, trade deals (*TTIP*), the steel industry crisis (*steel*, *Tata*), power generation (*oil*, *wind*, *onshore*), the doctors' dispute (*bma*, *junior*, *doctors*), transport (*tfl*), schooling (*academies*), the balance of power between the parliaments of the UK (*devolution*) and the normal (financial) business of government represented by *fiscal* and *deficit*.

4. Conclusions and further work

This article describes how the Hansard at Huddersfield web application recast common analytical tools from corpus linguistics into a user-friendly format for non-linguists. To do so, the team addressed technical challenges related to performance, user interaction and data presentation caused by the size and complexity of the Hansard dataset. We aim to update the data set with regular additions and to add further corpus linguistic search tools in the next stage of the project (funded by Parliamentary Data Services). With sustainability in mind, we also aim to explore the potential for building some of the HaH functionality into the official Hansard website. Future aspirations include adding more detailed metadata, such as political parties, to allow for more nuanced searches. So far, the HaH web application has received positive responses from users close to parliament, has been used in teaching A-level students, and has supported arguments in a journal article about attitudes to MPs (McKay, 2020).

Funding

This work was supported by the AHRC under Grant AH/R0007136/1.

Notes

1. https://www.english-corpora.org/hansard/

2. https://blog.history.ac.uk/tag/digging-into-linked-parliamentary-data/

3. http://search.politicalmashup.nl

4. https://www.visionofbritain.org.uk/

5. http://www.data.parliament.uk/dataset

6. These figures are based on counting each individual contribution separately. Some are very short interjections and some much longer speeches. The number of speakers is approximate because the naming conventions in Hansard historically have not always identified individual speakers uniquely, since they are labelled according to their roles, which change over time.

7. We use the neutral term 'contribution' to encompass any contribution to a debate by any MP or Peer, whether that is a speech, comment, or reply. Contributions do not include interruptions. Though Hansard records interruptions to debates with the tag "interruption", we deleted these from our database as Hansard does not specify any wording of interruptions and therefore do not add to understanding language patterns across Hansard.

8. https://d3js.org/

9. https://getbootstrap.com/

10. Note that, although this is not a consistent rise across the period, the troughs that go right down to zero can be discounted as they represent recess months where parliament was not sitting.

References

Aucott, P., von Lünen, A., & Southall, H. (2009). Exposing the history of Europe: The creation of a structure to enable time-spatial searching of historical resources within a European framework. *OCLC Systems & Services*, *25*(4), 270–286.

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. *Journal of English Linguistics*, *32*(4), 346–359.

Blaxill, P., & Beelen, K. (2016). A feminized language of democracy? The representation of women at Westminster since 1945. *Twentieth Century British History*, *27*(3), 412–449.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Jeffries, L. (2010). *Critical Stylistics: The Power of English*. Palgrave MacMillan.

Jeffries, L. and Walker, B. (2017). *Keywords in the Press: The New Labour Years.* Bloomsbury.

McKay, L. (2020). Does constituency focus improve attitudes to MPs? A test for the UK. *The Journal of Legislative Studies, 26*(1), 1-26.

Mesiti, M., Pellegata, A., & Perlasca, P. (2015). Making the analysis of the italian legislative system easy: The ILMA web portal. *Journal of Information Technology & Politics*, *12*(1), 88-102.

Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. In R. Witte, H. Cunningham, J. Patrick, E. Beisswanger, E. Buyko, U. Hahn, K. Verspoor, & A.R. Coden (Eds.), *LREC 2010. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (pp. 46-50). ELRA.

Southall, H., von Lünen, A., & Aucott, P. (2009, December). On the organisation of geographical knowledge: Data models for gazetteers and historical GIS. In e-science 2009. *Proceedings of the 5th IEEE International Conference on e-Science Workshops* (pp. 162–166). IEEE Computer Society.

Sykora, M. D., Jackson, T. W., O'Brien, A., Elayan, S., & von Lünen, A. (2014, July). Twitter-based analysis of public, fine-grained emotional reactions to significant events. In A. Rospigliosi, & S. Greener (Eds.), *ECSM 2014. Proceedings of the European Conference on Social Media* (pp. 540–548). Academic Conferences.

Sykora, M. D., Jackson, T. W., von Lünen, A., Elayan, S. & O'Brien, A. (2015, July). The role of visualizations in social media monitoring systems. In A. Mesquita, & P. Peres (Eds.), *ESM2015. Proceedings of the 2nd European Conference on Social Media* (pp. 437–444). Academic Conferences.

Address for correspondence

Alexander von Lünen

History (Music, Humanities and Media)

University of Huddersfield

Queensgate

Huddersfield, HD1 3DH

England

Email: a.f.vonlunen@hud.ac.uk

Co-author information

Hugo Sanjurjo-Gonzalez,

Computing, Electronics and Communications Technologies

University of Deusto

Lesley Jeffries

Linguistics and Modern Languages

University of Huddersfield

Fransina Stradling

Linguistics and Modern Languages

University of Huddersfield

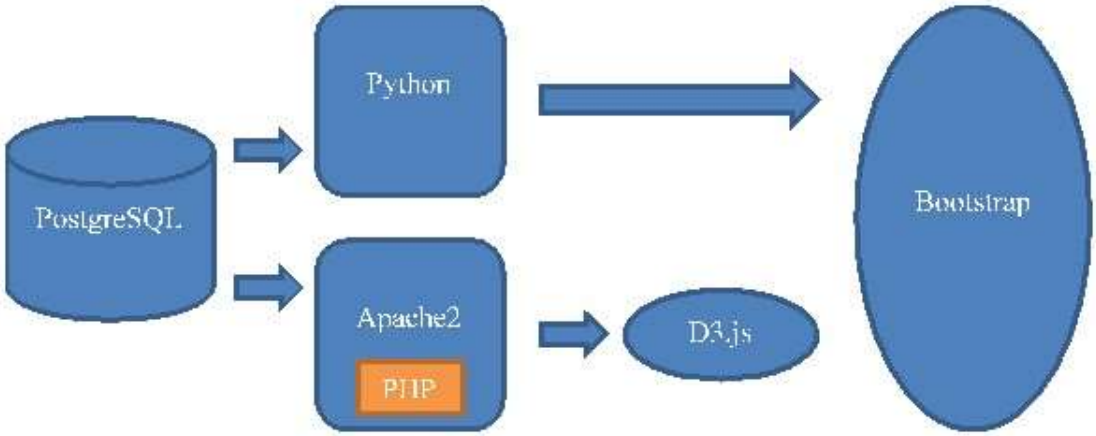

Figure 1. General architecture of the web application.

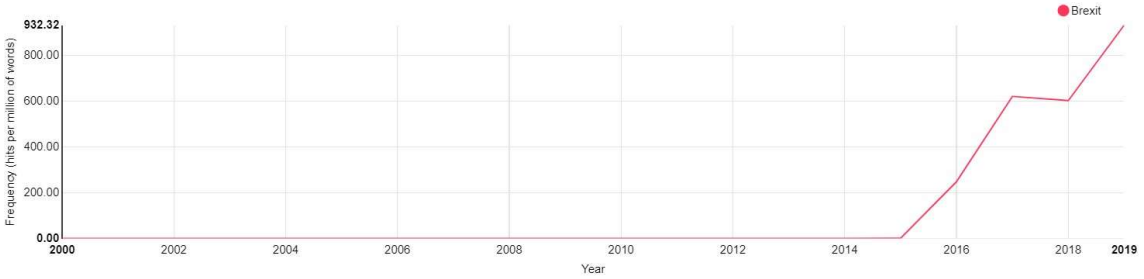

Figure 2. Frequency of *Brexit* usage per year.

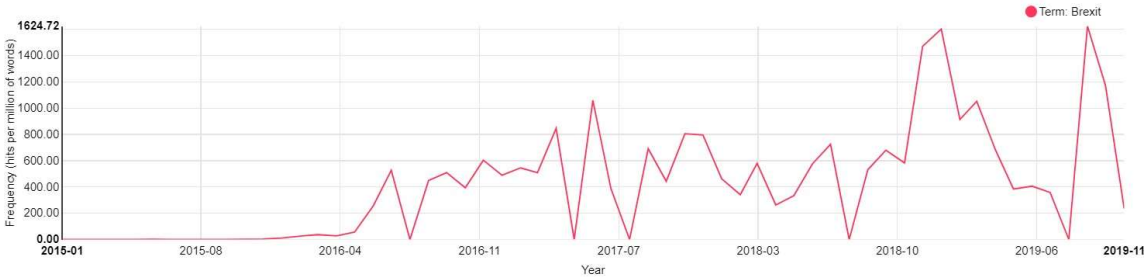

Figure 3. Frequency of *Brexit* usage per month.

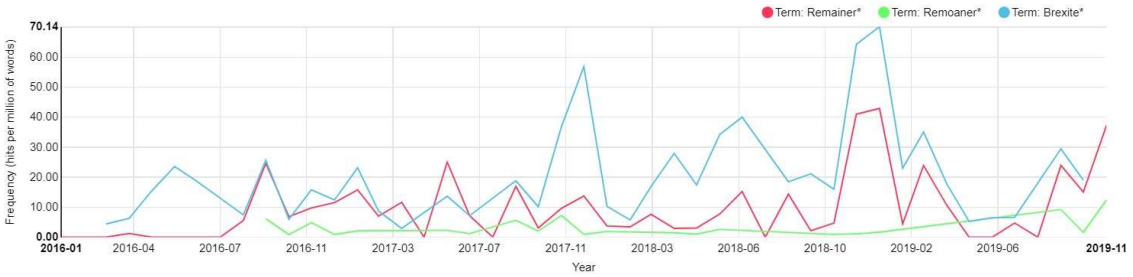

Figure 4. Comparison frequency of usage *Remainer**, *Remoaner**, *Brexite**.

	Date	Member	Contribution
☐	2015-11-10	Pat McFadden	**...** trength in trade with the rest of the world, and stands together to combat the urgent security problems that we face. We do not stand for the nationalism that says that we would be better off out, or for a **Brexit** that would see Britain weaker in power and influence, and diminished in the eyes of the world. In his speech this morning and in the letter to the President of the European Council, the Prime **...** *[0 more]*
☐	2015-11-05	Calum Kerr	**...** ltural policy payments make the difference between bankruptcy and continuing in business. The Secretary of State has been repeatedly asked to confirm whether those payments would continue in the event of a **Brexit**. Simply batting that question away is no longer acceptable. What will happen?**...** *[0 more]*

Figure 5. Sample of concordance lines that show earliest uses *Brexit* as a countable noun in November 2015.

☐	2016-02-25	Edward Leigh	**...** cretary is rightly doing a sort of cost-benefit analysis of this issue. Why do the Government not institute an independent study, by a genuinely independent body, to go in some detail into the effects of a **Brexit**, plus or minus, on, say, GNP? That would surely be very useful**....** *[0 more]*
☐	2016-02-25	Nick Herbert	**...** My right hon. Friend is rightly drawing attention to the potential impact of **Brexit** on our economy, but may I take him back to the issue of security? It was suggested earlier that there would be no adverse consequences for security from our leaving the European Union, because we would remain members of NATO. Did he hear the remarks this morning of the former Secretary-General of NATO, Anders Rasmussen,**...** *[0 more]*
☐	2016-02-25	Stuart Donaldson	**...** any other people my age. I have been contacted by a number of young people who are slightly worried about what will happen when they leave school or are in their university holidays. They fear that a **Brexit** might mean that they will not have the opportunity to jet off easily to Magaluf or Zante for the aforementioned holiday. Will they have to go through the hassle of getting visas just for a week or t**...** *[0 more]*
☐	2016-02-25	Stewart Jackson	**...** ers that we must stay at the heart of Europe, fight our corner and reform within. That has failed and it is a fool's errand to believe that it will not be a calamitous failure in the future. We know what **Brexit** will be like, as my hon. Friend the Member for Harwich and North Essex (Mr Jenkin) has said. In conclusion, all power is a leasehold given to us on trust, and it is not ours to give away. For to**...** *[0 more]*

Figure 6. Sample of concordance lines that show mixed usage of *Brexit* as a countable noun and proper noun in February 2016.

☐	2019-11-05	Dominic Raab	**...** hening the special relationship. I spoke to the NATO Parliamentary Assembly, affirming our leading role in NATO and our commitment to it. Above all, I am focused on supporting the Prime Minister in getting **Brexit** done so that this country can move forward as an open, outward-looking country with global reach and global ambition**....** *[0 more]*
☐	2019-11-05	Emily Thornberry	**...** mpted Russian interference in the 2016 referendum, whether that was successful or otherwise? I fear it is because it realises that the report will lead to other questions about the links between Russia and **Brexit**, and with the current leadership of the Tory party, that risk derailing its election campaign. There are questions about the relationship between the FSB-linked Sergey Nalobin and his "good fr**...** *[1 more]*
☐	2019-11-05	Stephen Gethins	**...** ons were damning, and I am not surprised he did not answer them. Given the threat Russia poses to elections, and given that his Government have wanted an election for months, why is this not a priority? **Brexit** has taught us that this Government like to hide unhelpful reports—lots of them—so prove me wrong and publish the report**....** *[0 more]*

Figure 7. Sample of concordance lines that show *Brexit* as a proper noun firmly embedded in Hansard discourse in November 2019.

Figure 8. Word cloud for 2016 showing most frequent words used in 2016 with the presence of *Brexit*, *EU* and *deal* highlighted.

Figure 9. Word cloud for 2019 showing the most frequent words used in 2019 with the presence of *Brexit, EU, deal* and *agreement* highlighted.

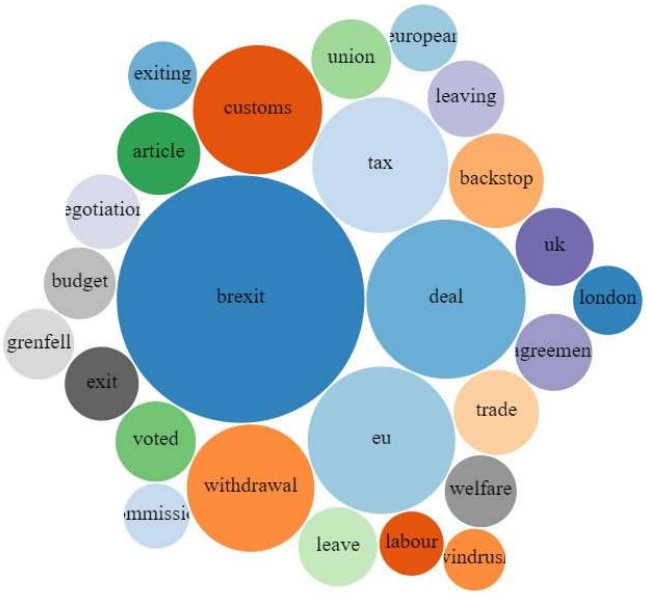

Figure 10. Bubble chart of keywords in the 18 months before the Brexit referendum.

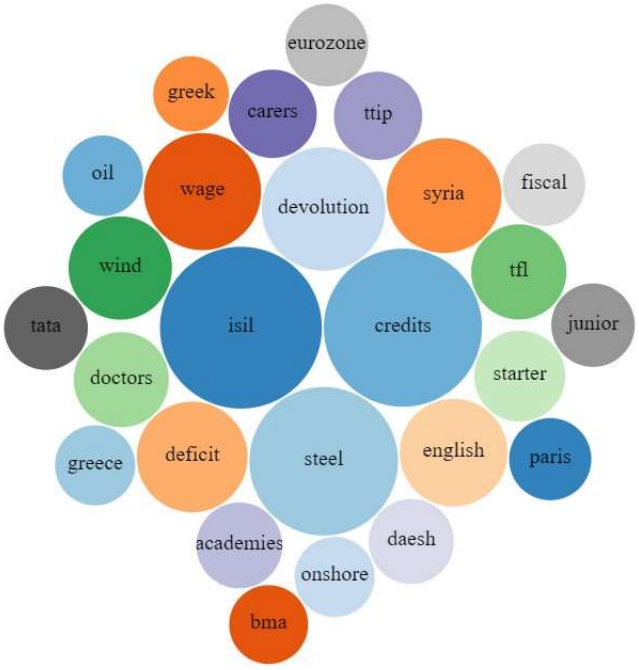

Figure 11. Bubble chart of keywords in the 3.5 years after the Brexit referendum.