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

Alexander von Lünen, Lesley Jeffries, Fransina Stradling, Hugo Sanjurjo González, and Paul Crossley

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 – 2021) 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, and gives example uses of the system.

Keywords: Hansard, parliamentary debates, corpus tools, 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 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 to 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 and add to public understanding of the way parliament addresses issues of social, political and historical interest.[1]

This article introduces the Hansard at Huddersfield (HaH) web application which combines corpus linguistic methods with interactive visualizations to make such analysis of linguistic choices in Hansard accessible to researchers from non-linguistic fields (e.g., History, Sociology, Political Science, single issue lobbyists and campaigners) as well as the general public. Corpus Linguistics is a methodology that examines large amounts of linguistic data, known as corpora, to reveal patterns of language choice. Corpora (*sg.* corpus) are representative datasets of language use which are subjected to quantitative and qualitative analysis using specialised computer software. Corpus Linguistics is comparable to text-mining, a computational method of text analysis regularly used in Digital Humanities research. However, while text mining is chiefly concerned with the technicalities of automatically extracting information from large datasets, corpus linguistics offers researchers computerised tools to examine how and what language use

communicates.[2] To that end, corpus linguists use their tools to inform both quantitative and qualitative analysis of text: the software can identify quantifiable patterns in corpora rapidly and accurately as well as aiding qualitative analysis by providing displays that allow quick surveys of the data. Because of its commitment to exploring language in use, corpus linguistic methods have recently been productively combined with practices of discourse analysis.[3] A corpus-based analysis of discourse computes the language patterns of texts to subsequently interpret how its topics are discursively constructed. In doing so, it does not just provide entry points into large datasets otherwise impractical to study qualitatively, but also provides ways to achieve both precision and richness in analyses.

Many researchers 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. However, the currently most prominent public-facing online versions of Hansard – the official Hansard website (hansard.parliament.uk) and TheyWorkForYou (theyworkforyou.com) – provide a limited number of search types which constrain the pattern analysis possible. On these 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.

In principle, the Hansard data is Open Data, offered by the Parliamentary Digital Service of the UK Parliament; we wanted to sustain this idea of openness and democratization of government data. However, just because data is open, does not mean it is accessible. To improve access to effective corpus linguistic searches of parliamentary records, we decided to design a user-friendly interface that simplifies the search process for those unable to undertake their own data preparation or learn how to use complex corpus software. The resulting 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.

In section 2 we will describe our approach to creating the HaH web application, outlining previous projects that offered a point of departure, and discussing design decisions. HaH started as a follow-on to the SAMUELS project (see next section), and was inspired by several digital humanities projects that the authors were involved in. These provided experience with, and insight into promising avenues for a simplified, graphical user interface to a large text corpus to allow users not familiar with corpus linguistics to harness some of the latter's analytical capabilities. The architecture and functionalities are detailed in section 3, using two case studies to illustrate possible HaH uses. As will be demonstrated, HaH has more to offer than just visualizations, and is of use to laypersons as well as experts analysing parliamentary discourse. Section 4, finally, will set out our conclusions.

## 2. Inspirations

### *2.1 Similar Projects*

HaH is not the first project to work on making UK Hansard more amenable to in-depth scrutiny. Apart from the inclusion of a small amount of Hansard data in the British National Corpus (BNC; 1,167,351 words), 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 and the Digging into Linked Parliamentary Data (DiLiPaD) project.[4] 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.[5] 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. However, its search functions require a high-level understanding of linguistics and statistics, making the corpus useful only for those well-versed in the use of corpus linguistic software. The Hansard corpus served as the launchpad for the HaH project out of a desire to provide a more user-friendly interface for a wider audience.

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.[6] Although some research has been published using the text-mining strategies made possible by the DiLiPaD project, unfortunately their website is no longer available.[7]

Similar platforms exist for the legislative systems of other countries, such as the Italian Law-Making Archive (ILMA) Web Portal 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.[8] 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 interactively. 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 of using interactive visualizations as part of the epistemic process to allow further explorations of the data as opposed to mere static presentations of data is rare within corpus linguistics, and is unique in the analysis of parliamentary records.[9] There are other applications of corpus methods which share with Hansard at Huddersfield the ambition to bring CL methods to groups outside corpus linguistics. For example, CliC provides search features for literary corpora, so far focusing on the works of Dickens with a range of other 19th century literature for comparison.[10] This project doesn't use interactive visualization as the entry point into the data, but like HaH it provides a user-friendly way to produce the kinds of results normally produced by corpus software, but with less difficulty. Like HaH, it also allows the user to click through to the fuller data if more context is needed. Other projects, such as

WordWanderer and Kaleidographic are also interested in bringing CL techniques to a wider audience, but are more experimental than HaH or CliC and prioritise visualization without necessarily linking through to raw data or standard corpus linguistic data presentation formats, such as concordances.[11] Our approach differs from these projects in using off-the-peg visualisations as a way to provide an interactive and intuitive search facility for the Hansard data which will appeal not only to those interested in tracing and researching linguistic patterns per se, but also to those wanting to locate and study parliamentary discussion of topics of importance to them.

The approach taken for HaH was further inspired by the the *Vision of Britain* (VoB) website, which allows for searching geographical and demographic data about Britain between 1801 and 2001.[12] The VoB data model, which facilitates the storage of concurrent historical-geographical information, helped the QVIZ project 2007—2009 (co-created by the VoB team, under leadership by the Humlab at the University of Umeå, Sweden) to develop a prototype for a faceted browser that uses digital maps as vehicle to explore archival collections across three countries with very distinct histories and geographies: Britain, Sweden and Estonia.[13] This approach in turn inspired HaH's use of a faceted browser in the form of a one-screen interface to initiate and direct searches across the Hansard corpus, albeit with charts rather than maps. 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.[14] EMOTIVE's website allows users to interact with the visualizations in order to filter data.[15] HaH combined these two ideas – faceted browsers and interactive visualizations – to simplify corpus linguistic methods for exploring the Hansard corpus.

## **2.2** Corpus Tools

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 (see below for explanation of these) rather than developing new tools or experimenting with cutting edge, but not yet widely accepted, approaches. This concurs with our aim to make corpus tools accessible to non-linguists instead of providing new tools for corpus linguists.

The 'frequency list' tool is self-explanatory: it is a list of the frequency of occurrence for a word or cluster of words. Frequency counts are often represented as raw frequencies and/or as standardized frequencies against the total number of unique words in the corpus. Frequency lists tend to be used in almost all corpus studies; they may be used to explore the style of a corpus as compared to another corpus of another language variety, but may also, importantly for Hansard, give an idea of what the corpus is about. They may be used to determine which words in the Hansard corpus warrant further exploration, but also require some statistical expertise to interpret.

Another frequency-related tool, namely dispersion plots, is used to show the spread of a word across the corpus, thereby indicating a pattern of usage. Dispersion data may lead to further investigation of a particular part of the Hansard dataset to explore the reasons for a high (or rather, low) frequency of use in that specific section. Frequency lists or dispersion plots alone may, however, not be maximally informative. Contextual analysis of frequent words and their dispersion is needed to identify in what way these words may be used to construct particular discourses. A common presentation format of corpus software, concordance lines, provides for an easy overview of words in their direct context. Concordances simply list all the occurrences of a word presented with its immediate corpus co-text on both the left and right of the term. They are sometimes also

referred to ‘key word in context’ (KWIC) displays, referring to the word under investigation as the keyword. Concordances can be sorted on both sides to reveal patterns of co-occurring words and provide for easy qualitative extrapolation of the patterns the context of a word reveal. Analysing Hansard through concordances can inform us of the kinds of discourse used in relation to particular topics or debates.

Whilst concordances may help identify patterns of co-occurrence (and thus traces of discourses) through close qualitative analysis, this may also be done automatically. When automatic analysis reveals that a word regularly appears close to another and this relationship is statistically significant, then patterns of co-occurrence are referred to as collocation with the words in this relationship referred to as collocates. Collocation is a way of finding co-occurrence patterns in the dataset that may be hard to spot without computerised software. Collocational patterns are derived using different (complex) statistical techniques to confirm how often words across a span of context to both the left and right of the word under analysis co-occur with it. For the purposes of searching Hansard, collocates may be grouped manually into thematic or semantic categories to find patterns that illuminate the content of the section of Hansard under consideration.

The final corpus tool introduced here that may provide relevant to a pattern analysis of Hansard is the measure of keyness. Keyness measures rely on statistical tests to compare frequency patterns between one corpus and another, a so-called reference corpus. Resulting keyword lists highlight saliency across the corpus rather than frequency. Instead of relying on impressionistic selection of frequent words from the frequency list, a saliency measure provides a more systematic approach to selecting words for further examination. Keyness measures first require statistical interpretation before supplementary concordance and collocational analysis may contextualise how these keywords function in the discourse.

As will be demonstrated in section 3.3.3, our project has so far exploited frequency lists, dispersion plots and keyness measures, with collocation remaining on the list of developments for the next stage. We were keen that users should not see the results of searches as definitive answers to (research) questions, but rather 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[1]. 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 entry point to the data and an indicator of patterns of usage 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. Furthermore, users can download a user guide and sample case studies to better understand how to use and interpret the site's search features.

## 3. Architecture and possible use

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 from 2005 to 2021 into one database of Hansard data from 1803 to 2021.[16] The HaH project converted this database (8,976,836 speeches over 218 years with around 50,973 speakers, totalling 1,798,662,121 words in the corpus at the time of this writing) into a relational database.[17] This conversion was necessary to enhance the performance of the searches the HaH front-end offers its users; a relational database favours 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 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.[18] 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 in our current system. 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, perhaps to apply to different datasets, 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 was used for NLP tasks.[19] Visualizations were rendered using D3.js, a JavaScript library for visualizing data by means of HTML,

CSS and SVG. For the front-end design the Bootstrap CSS framework created by Twitter was used.[20]

[Figure 1 near here]

Figure 1. General architecture of the web application.

### **3.3** Functionalities

The HaH web application allows users to interact with Hansard debates through three entry points: a line graph, a word cloud and a keyword bubble. 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 Line Graphs*

The web application's default screen provides access to Hansard through a search box with several filters: search terms, Houses of Parliament (Commons, Lords or both) and time period. Any search will produce a line graph (also called distribution graph) of search word frequencies covering the time period specified by the user. A search for *Brexit* (House of Commons, 2000-2021), for example, produces a flat line followed by a steep rise in usage between 2015 and 2019 (figure 2) showing a rise from 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. The line graph then dips again in 2020 and even further in 2021, given the European Withdrawal Agreement could not be debated any longer after it received Royal Assent on 23rd January 2020.

[Figure 2 near here]

Figure 2. Frequency of *Brexit* usage per year (House of Commons).

Searches can be delimited using additional parameters by clicking the 'Advanced Options' button. This allows for searching within specific dates (DD-MM-YYYY) rather than years, and for specifying a speaker (Member of Parliament or Peer) and a Debate Title. Having the 'Advanced Options' button selected when searching for a date range below 5 years will furthermore produce a distribution graph that displays word frequencies by month rather than year. Using this function to search for *Brexit* between 2016-2021, we can nuance the picture in figure 2 and see a variable rise in usage to the end of 2019 that dips again in 2020 (figure 3).[21] frequency of up to 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*).

[Figure 3 near here]

Figure 3. Frequency of *Brexit* usage per month (House of Commons).

[Figure 4 near here]

Figure 4. Comparison frequency of usage *Remainer**, *Remoaner**, *Brexite** (House of Commons).

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, which will display by contribution and indicate the number of hits in that contribution (see figure 5 for an example) which can be found by opening the full contribution and scrolling down to the highlighted hits. Instead, users can select Keyword in Context (KWIC) format with a default context of 10 words each side which will display each hit separately, indicating those that occur in the same contribution by the numbers in the left-hand column. 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 6) shows that it was a countable noun in 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 7), and by November 2019 the proper noun became so embedded that nothing in the context of these examples queries what *Brexit* is (figure 8).

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.

[Figure 5 near here]

Figure 5. Sample concordance lines Remainer* in document format (2019, House of Commons).

[Figure 6 near here]

Figure 6. Sample of concordance lines that show earliest uses *Brexit* as a countable noun in November 2015 (House of Commons).

[Figure 7 near here]

Figure 7. Sample of concordance lines that show mixed usage of *Brexit* as a countable noun and proper noun in February 2016 (House of Commons).

[Figure 8 near here]

Figure 8. Sample of concordance lines that show *Brexit* as a proper noun firmly embedded in Hansard discourse in November 2019 (House of Commons).

### 3.3.2 Word Clouds and Keyword Bubble Charts

The HaH site also allows users to explore Hansard without pre-determined search terms and based on either word frequency or keywords. The programming architecture of the search screen is linked to word clouds and keywords; a search using the search screen will produce clickable word clouds and keyword bubble charts on the right-hand side of the search screen. When either of these is clicked, the word cloud or keyword bubble chart exchanges places on the screen with the line graph and can be used interactively to produce contribution lists in the same formats as the default search function. It is also possible to adjust the parameters of a word cloud or keyword search to create an entirely new interactive visualisation.

The frequency-based feature uses word clouds to represent the most frequent words in a specific time period, defined by year for one or both Houses of Parliament. The font size of words in the word cloud reflects the frequency of a particular word in that period, i.e. the font size is proportional to the rank of the respective word in the frequency distribution. 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 9) and 2019 (figure 10), 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.

[Figure 9 near here]

Figure 9. Word cloud for 2016 showing most frequent words used in 2016 with the presence of *Brexit*, *EU* and *deal* highlighted (House of Commons).

[Figure 10 near here]

Figure 10. Word cloud for 2019 showing the most frequent words used in 2019 with the presence of *Brexit*, *EU*, *deal* and *agreement* highlighted (House of Commons).

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 (i.e., a reference or comparison corpus).[22] HaH defines keywords as in corpus linguistics, namely words which occur relatively more frequently in one corpus compared to another.[23] Standard corpus software, such as WordSmith, provides a keyword search facility for users, but our aim is to allow non-expert users the same functionality as corpus linguists without the need to assimilate the terminology and statistical understanding

required by specialist software. What our users need to know is that the keywords feature differs from the straightforward frequency feature in reflecting the saliency of words rather than simple frequency. Users first select their target corpus and then a comparison (also called 'reference') corpus. We have provided some pre-set corpora (Prime Ministers in Office and Administrations) 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 per cent certainty differences are not by chance, using a log-likelihood statistical test.[24] 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 2019 as compared with the 18 months (from Jan 1st 2015) prior to the referendum. Unsurprisingly, this bubble chart is dominated by *Brexit* (figure 11), but if we compare these periods in reverse, we find a different picture (figure 12). 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*.

[Figure 11 near here]

Figure 11. Bubble chart of keywords in the 18 months before the Brexit referendum (House of Commons).

[Figure 12 near here]

Figure 12. Bubble chart of keywords in the 3.5 years after the Brexit referendum (House of Commons).

**3.4** Applications of Hansard at Huddersfield searches

Most CL practitioners are concerned with finding out more about how language works or characterising the overall style of a particular corpus or sub-corpus. However, the kinds of results available from established CL techniques (e.g., concordances, word lists, collocations, keywords) are only a stepping-stone towards understanding the character and style of a body of texts, and there is almost always qualitative work to be done to follow up on automated results. That qualitative work will often be linguistic, of course, but it may also be historical, political, social etc. As we show in supplementary materials in the form of case studies on the HaH site, finding out that parliament has moved from talking more about *thrift* to talking about *austerity*, for example, is an observation not simply about language, but about the shift from personal virtue to state-imposed impoverishment which reflects policies of different governments as well as the moral and social climate of different periods.[25] This observation requires the kind of detailed scrutiny of the concordance lines that is the strength of much historical and socio-political research and is facilitated by HaH. A second case study shows that another option is to use a single event, such as the Peterloo massacre of 1819 as a starting point for a consideration of parliament's changing view of popular protest, since such events are often cited in parliamentary debate and can lead to investigation of other related terminology such as *(un)lawful assembly*, *protest* and *demonstration* (which is, of course, polysemous).[26]

As we saw in section 3.3, the site can be used to consider topics such as Brexit in focussed time periods, but we can also try to set such historical events in context by searching related terms across a wider timescale. For example, searching from 1970 to 2021 using terms such as EEC, EU and Europe can provide a sense of how the rise of Brexit is contextualised. A line graph (see figure 13) using these three terms alongside *Brexit* shows how there are two peaks of occurrence in Commons data, one around the 1975 referendum on the UK's membership of the EEC and a much larger one relating to the referendum in 2016 on the UK's departure from the EU. Perhaps a more surprising finding of this search is that the highest usage of the word 'Europe' comes in the period from 1988 – 2000 when the issue of the UK's membership had a lower profile in political discussions. Nevertheless, a closer look at the concordance lines containing the word Europe in this period (from 1992) shows that there is variation in the way that this word is used, sometimes referring to Europe as excluding the UK: '...is a fundamental difference between our practices and those of Europe' or 'Whether from Europe or domestically...'. At other times, the UK is subsumed into the referent of Europe: '...the fact that we are top of the league in Europe...' or '...as it will have a knock-on effect throughout the rest of Europe.' A quick comparison with concordance lines from the height of the Brexit debate in 2015 appears to show the UK more often being considered distinct from Europe than being part of it: 'British people can go and live in Europe' or '...their head-in-the-sand approach to what is going on in Europe.' These examples appear to demonstrate a growing, if at times subconscious, tendency amongst parliamentarians at this time to see the UK as separate from Europe.

[Figure 13 near here]

Figure 13. Comparison frequency of usage *EEC*, *EU*, *Europe*, *Brexit* (House of Commons).

This impressionistic overview of a search on terms relating to Europe shows that such general searches can produce testable hypotheses. However, further qualitative work on a properly chosen subset of the concordance lines would be needed to confirm or discard them. Such work may involve more consideration of the wider textual context of the search terms in case the limited window provided by the KWIC format is misleading. It may also require some consideration of party affiliation of the speakers and/or role in government to establish whether there is a pattern of usage that differs according to speaker loyalty or ideology. This requires users to bring knowledge beyond HaH to bear on their interpretations.

The sample searches in this article illustrate that the variety and extent of types of searches that could be relevant to not just linguistic but historical, political, social and policy researchers is unlimited. For that reason, our aim has been to provide a user-friendly but flexible search tool for all those interested in this rich dataset, including those whose research needs stretch beyond the visualisations and presentation formats HaH offers. A study by Griffiths and Lünen showcases this versatility in the way they extracted Hansard contributions relevant to their analysis.[27] This study aimed to compare the political rhetoric of eminent British nineteenth century politician Benjamin Disraeli (1804—1881), who played an important role in modernising the Conservative Party. He made his first steps in this direction in the 1840s by publishing a trilogy of political novels – *Coningsby* (1844), *Sybil* (1845) and *Tancred* (1847) – in which he sets out his political vision through the words of his fictional characters. At the time of the publication of these novels, Disraeli was but a young back-bencher with very limited political clout. Griffiths

and Lünen were therefore interested in assessing how the language and themes in these three novels compared to the contributions of Disraeli in parliament in the 1840s, and what the use of language could tell historians and others about the process of public, political discourse in the mid-nineteenth century. HaH itself would not allow such comparison, but the efficiency of the search interface of HaH enabled the quick extraction of the relevant contributions through its download feature from the results list, as described in section 3.3.1. The downloaded data could then be processed further in AntConc, a desktop tool for corpus analysis, and with custom-made scripts in Python.

The official Hansard website would have rendered the above approach futile.[28] The official website allows relatively simple searches, but offers no download facility for more than one debate at a time. Historic data is also available through the Hansard Millbank systems site, a depository of digitised Hansard volumes between 1803-2005 that was incorporated into the official Hansard site in 2018.[29] The search functionality on the Millbank system is very limited, as it uses the Google API to find search terms within the Hansard proceedings; if contributions from a particular MP are sought, the MP's page might be a better starting point.[30] No download facility exists on the Millbank site, other than navigating to each individual contribution and using copy & paste. There is, however, the possibility to download whole volumes of the Hansard proceedings as XML file, and a REST API exists. The latter does not allow for searching; rather, the dates for sittings or name of a person need to be specified in the URL to retrieve webpages in JSON format.

By using the HaH download interface to extract the relevant speeches in parliament and carrying out further analyses in third-party software, however, Griffiths and Lünen were able to find that there was very little overlap in word choices as well as topics being discussed between the novels and Disraeli's parliamentary contributions;

corroborating the authors' hypothesis that Disraeli used the novels to make the speeches he could not make in parliament, and hinting at an early form of populism where politicians use the media to circumvent the parliamentary process.

## 4. Conclusions and further work

This article described how the Hansard at Huddersfield web application recast common analytical tools from corpus linguistics into a user-friendly format for non-linguists. As discussed, the site overcomes technical challenges related to performance, user interaction and data presentation caused by the size and complexity of the Hansard dataset. The resulting site makes detailed study of the language used in UK Parliament democratically accessible, as its searchability does not require users to be trained in corpus methods or to have extensive knowledge of linguistics. Interested users are enabled to observe patterns in word usage through graphs and visualisations, with the additional ability to access the Hansard data in full and with provision of supporting materials to aid the interpretation of the site's automated results. This combination of easy-to-use tools and complementary interpretation advice empowers users to uncover discussion of topics in parliamentary proceedings in unprecedented ways. Non-linguists have started to reflect on, and make use of, the HaH web application in their research. One historian has noted how the site's innovative use of corpus linguistics helps historians and the public exploit the Hansard database better: 'The [Hansard at Huddersfield] project is attempting to address the issue of inaccurate or inappropriate search results and thus reduce the likelihood that the end user's research will be skewed by an inability to understand or properly exploit the data available'.[31] Other historians, as well as political scientists, have so far used the site in publications to make claims about the content of

Hansard contributions. For example, a political scientist used a HaH line graph that showed a nearly six-fold increase in relative frequency of the words constituency or constituent(s) in Commons Hansard between 1950 and 2019 as a 'high-quality proxy' for determining MPs' constituency focus.[32] Based upon this graph, this political scientist concluded that this reflected 'a long-term trend towards constituency focus'. Other researchers have utilised visualisations generated from the site to communicate their research with wider, non-academic audiences. Examples of such HaH visualisation use includes a historian's presentation 'Remembering Peterloo: protest, satire and reform' hosted at the Palace of Westminster by the History of Parliament Trust, The Parliamentary Archives and the Citizens Project in July 2019 and a blog by a different historian on the historic debtor sanctuaries of London.[33]

The site has not only proved popular with non-linguists, but linguists have also used it in their research. A prime example is a large-scale linguistic study on the representation of heart failure in English discourse published line graphs showing the compared frequency of use of the terms 'heart failure', 'cancer' and 'dementia' in UK House of Commons and House of Lords debates between 1945 and 2021.[34] Taking frequencies as an indicator of importance, they concluded that these comparisons indicated that politicians discuss cancer in debates disproportionally more often than heart failure. Their use of the HaH web application seems to be an effect of the other main database of Hansard proceedings exploitable with corpus tools, the Hansard Corpus, only includes Hansard data until 2005, compared to 2021 for HaH. Additionally, since their study was published in a cardiology journal with a non-linguist audience, the line graphs provided a way to present their results without readers needing to have any understanding of corpus tools.

To ensure site posterity, 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 Digital Services). With sustainability in mind, we are also working with Parliament to build some of the HaH functionality into the official Hansard website as it is developed within their data platform redesign. Future aspirations include adding more detailed metadata, such as political parties, to allow for more nuanced searches. The aforementioned case studies covering diverse subject areas (history, political science, cardiology), as well as the cited study on the value of visualizations done by the EMOTIVE team, show that there is some potential to harness simplified, visual tools for turning data truly 'open' and allow users with little to no background in corpus linguistics the powerful tools the latter provides. HaH will hopefully aid in a more robust analysis of parliamentary debate that allows to form an argument beyond mere opinion, but with 'hard' data.

**Funding**

This work was supported by the AHRC under Grant AH/R0007136/1 and funded further by Parliamentary Digital Services.

---

[1] L. Jeffries, *Critical Stylistics: The Power of English* (Basingstoke & New York, 2010).

[2] T. McEnery and A. Hardie, *Corpus Linguistics: Method, Theory and Practice* (Cambridge, 2012). Cited here at 228.

[3] P. Baker, 'Querying keywords: Questions of difference, frequency, and sense in keywords analysis', *Journal of English Linguistics*, 32 (2004), 4, 346–359; C. Taylor and A. Marchi, *Corpus Approaches to Discourse: A Critical Review* (London, 2018).

[4] For the BNC: D. Lee, 'Genres, Registers, Text Types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC jungle', *Language Learning* & Technology, 5 (2001), 3, 32—72. The Hansard Corpus: https://www.english-corpora.org/hansard/. DiLiPaD: https://blog.history.ac.uk/tag/digging-into-linked-parliamentary-data/.

[5] Cf. A. Dawn, 'Negotiating difference in political contexts: An exploration of Hansard', *Language Sciences*, 68 (2018), 22—41.

[6] http://search.politicalmashup.nl, last accessed 2 October 2022.

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

[8] M. Mesiti, A. Pellegata and P. Perlasca, 'Making the analysis of the Italian legislative system easy: The ILMA web portal', *Journal of Information Technology & Politics* 12 (2015), 1, 88–102.

[9] Though see P. E. Rayson, J. A. Mariani; B. Anderson-Cooper, A. Baron, D. S. Gullick, A. Moore, and S. Wattam, 'Towards Interactive Multidimensional Visualisations for Corpus Linguistics', *Journal for Language Technology and Computational Linguistics* 31 (2017), 1, 27–49, for a discussion of this option.

[10] https://clic.bham.ac.uk/, last accessed 11 November 2022. For more details on this project, see M. Mahlberg, P. Stockwell, J. de Joode, C. Smith and M. Brook O'Donnell, 'CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics', *Corpora* 11 (2016), 3, 433–463.

[11] WordWanderer: https://wordwanderer.org,last accessed 2 October 2022. For more details on this project, see M. Dörk and D. Knight, 'WordWanderer: a navigational approach to text visualisation', *Corpora* 10 (2015), 1, 83–94. Kaleidographic:

https://www.kaleidographic.org, last accessed 2 October 2022. For more details on this project, see H. Caple, A. Laurence and M. Bednarek, 'Kaleidographic', *International Journal of Corpus Linguistics*, 24 (2019), 2, 245–261.

[12] https://www.visionofbritain.org.uk/, last accessed 10 November 2022. See H. Southall, A. von Lünen and P. Aucott, 'On the organisation of geographical knowledge: Data models for gazetteers and historical GIS', *Proceedings of the 5th IEEE International Conference on e-Science Workshops*, (Oxford, 2009), 162–166.

[13] P. Aucott, A. von Lünen and H. Southall, '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 (2009), 4, 270–286. A 'faceted browser' is a data browsing tool in which interaction with one element will induce changes in the other elements presented in the same interface, such as using checkboxes to apply filters to search results for archive searches, showing the value of interactive visualization for explorative data analysis.

[14] M. D. Sykora, T. W. Jackson, A. O'Brien, S. Elayan and A. von Lünen, 'Twitter-based analysis of public, fine-grained emotional reactions to significant events', in A. Rospigliosi, & S. Greener, ed., *ECSM 2014. Proceedings of the European Conference on Social Media* (Brighton, 2014), 540–548.

[15] M. D. Sykora, T. W. Jackson, A. von Lünen, S. Elayan and A. O'Brien, 'The role of visualizations in social media monitoring systems', in A. Mesquita and P. Peres, ed., *ECSM2015: Proceedings of the 2nd European Conference on Social Media* (Porto, 2015), 437–444.

[16] http://www.data.parliament.uk/dataset, last accessed 2 October 2022.

[17] See M. Davies, 'The advantage of using relational databases for large corpora. Speed, advanced queries, and unlimited annotation', *International Journal of Corpus Linguistics* 10 (2005), 3, 307–334. for a discussion of the advantages of relational databases for corpus linguistics. The 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.

[18] 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.

[19] R. Rehurek and P. Sojka, 'Software framework for topic modelling with large corpora', in R. Witte, H. Cunningham, J. Patrick, E. Beisswanger, E. Buyko, U. Hahn, K. Verspoor, and A.R. Coden, ed., *LREC 2010. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, (Malta, 2010), 46—50.

[20] D3: https://d3js.org/. Bootstrap: https://getbootstrap.com/.

[21] 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.

[22] See C. Gabrielatos, 'Keyness analysis: nature, metrics and techniques', in C. Taylor and A. Marchi, ed., *Corpus Approaches to Discourse: A critical review*, (London, 2018), 225—258.

[23] Baker, 'Querying keywords', 346.

[24] P. Rayson, D. Berridge and B. Francis, 'Extending the Cochran rule for the comparison of word frequencies between corpora', in *Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004)*, vol. II, Louvain-la-Neuve, Belgium, March 10-12, 2004, 926—936.

[25] https://hansard.hud.ac.uk/site/pdf/Austerity.pdf, last accessed 2 October 2022.

[26] https://hansard.hud.ac.uk/site/pdf/Peterloo.pdf, last accessed 2 October 2022.

[27] S. Griffiths and A. von Lünen, 'Parliament and the language of political agency in Disraeli's "Young England" trilogy; a corpus-linguistic approach', in Claudia Schrag-Sternberg and Uta Staiger, ed., *Parliament Buildings: The architecture of politics in Europe* (London, forthcoming).

[28] https://hansard.parliament.uk/, last accessed 1 November 2022.

[29] http://hansard.millbanksystems.com, last accessed 3 November 2022.

[30] E.g., http://hansard.millbanksystems.com/people/mr-benjamin-disraeli, last accessed 2 October 2022.

[31] L. Brinkley, 'Innovation versus Tradition in Historical Research Methods: The "Digital Turn"', *Emergence* X (2018), 34-48. Cited here at 45.

[32] L. McKay, 'Does constituency focus improve attitudes to MPs? A test for the UK', *The Journal of Legislative Studies* 26 (2020), 1, 1-26. Cited here at 4.

[33] K. Navickas, '"Right of Public Meeting" in Hansard', unpublished paper presented at 'Remembering Peterloo: protest, satire and reform', London, 11 July 2019; J. Levin, *Resource: Cobbett's Parliamentary History*. Online journal *Alsatia*, created 8 October 2019, http://alsatia.org.uk/site/2019/10/resource-cobbetts-parliamentary-history/, last accessed 2 October 2022.

[34] J. Demmen, N. Hartshorne-Evans, E. Semino and R. Sankaranarayanan, 'Language matters: Representations of 'heart failure' in English discourse—a large-scale linguistic study', *Open Heart* 9 (2022), 1, e001988. https://doi.org/10.1136/openhrt-2022-001988.