Language Sciences 79 (2020) 101270

Contents lists available at ScienceDirect

Language Sciences

journal homepage: www.elsevier.com/locate/langsci

Investigating colloquialization in the British parliamentary record in the late 19th and early 20th century

Turo Hiltunen [a,*], Jenni Räikkönen [b], Jukka Tyrkkö [c]

[a] University of Helsinki, Finland
[b] Tampere University, Finland
[c] Linnaeus University, Sweden

ARTICLE INFO

Article history:
Received 18 July 2019
Received in revised form 27 December 2019
Accepted 31 December 2019
Available online 6 February 2020

Keywords:
Parliamentary discourse
n-grams
Colloquialization
Democratization

ABSTRACT

In this paper, we explore how sociocultural changes were reflected in the parliamentary record, a genre that combines elements of spoken, written and written-to-be-spoken discourses. Our main interests are in the processes of linguistic colloquialization and democratization, understood broadly as tendencies towards greater informality and equality in language use. Previous diachronic studies have established that written language has increasingly adopted features associated with spoken language, although genre and register differences are considerable. Our starting point is that as Parliament has become more demographically representative and as prescriptive norms have loosened in society on the whole, the relative frequency of informal features in parliamentary language may have increased. At the same time, profound changes took place in the practices of recording parliamentary proceedings, most importantly the introduction of the official report in 1909. Our data on British parliamentary debates come from the *Hansard Corpus* (Alexander and Davies, 2015). We investigate the 60-year-period 1870–1930, which includes reports of parliamentary debates and, after 1909, verbatim reports (in total ca. 40 million words). Adopting a pattern-driven approach, we focus on n-gram frequencies. The analysis first identifies major shifts in the language of the reports using unsupervised grouping methods, and then investigates in more detail the frequency trends of individual n-grams associated with spoken language, as well as their function in parliamentary debates. The findings indicate that the introduction of the official report resulted in clear changes in n-gram frequencies, which can be linked to democratization and colloquialization.

1. Introduction

Parliamentary records can be invaluable sources of information to scholars in different fields, not only concerning the contents of the debates but also about how language is used to expound ideas, express arguments and rebut the claims of others. The language used in the British Parliament has increasingly begun to attract the interest of corpus linguists, as the parliamentary record, known as *Hansard*, is in many ways a "corpus linguist's dream" (Mollin, 2007: 187). In a sense, this dream has become reality with the release of the *Hansard Corpus* (Alexander and Davies, 2015), containing records of all the speeches given in the British Parliament, as represented in the Historic Hansard (https://api.parliament.uk/historic-hansard/index.html). The corpus enables investigations into a variety of discourse-analytical and linguistic topics ranging from representation of particular themes and groups of people (e.g. Blaxill and Beelen, 2016) and the choice of pragmatic strategies (e.g. Archer, 2017) to the progression of linguistic change (e.g. De Smet, 2016).

* Corresponding author
E-mail addresses: turo.hiltunen@helsinki.fi (T. Hiltunen), jenni.raikkonen@tuni.fi (J. Räikkönen), jukka.tyrkko@lnu.se (J. Tyrkkö).

https://doi.org/10.1016/j.langsci.2020.101270

In this paper, we investigate the idea that the language of the Parliament, as represented in the British Hansard, has been influenced by a discourse-pragmatic process referred to as "colloquialization". This research stems from recent work in corpus linguistics suggesting that as societies become more democratic, this is reflected in language among other things as the increasing acceptance of informal language in contexts that have traditionally been dominated by formal and regulated usage (e.g. Mair and Leech, 2006; Leech and Smith, 2009; Leech et al., 2009). In many ways, the parliamentary record is a good source of material for this question as the written records included in *Hansard* represent very specific kinds of speech events in a written format. These records enable us in principle to study changes in this situational variety of spoken language, as well as when and how the norms and conventions of reporting it change over time (Kruger and Smith, 2018: 4). However, the conventions of transcribing speech vary depending on the type of situation and the purpose of the activity (Cameron, 2003), and previous studies show that the contemporary *Hansard* cannot be treated as a fully accurate representation of spoken language (Chilton, 2004; Mollin, 2007; Slembrouck, 1992). This issue is particularly acute for the earlier decades, as the production context of *Hansard* has undergone dramatic changes. For our purposes, the key event was the introduction of the official report in 1909 (Vice and Farrell, 2017). Before this year, the parliamentary record consisted of selective third-person summaries of the debates, whereas in the official report that followed all contributions were included and reproduced in the first person (Alexander and Dallachy, 2019).

The introduction of the official report was an important transitional moment both societally and linguistically, marking a point at which two important sociocultural determinants of language change, democratization and colloquialization, intersect. In providing a fair and accurate account of the parliamentary debates to the general public, the official report is a concrete example of societal democratization. At the same time, by changing the norms of recording and representing parliamentary speech in writing, it is by definition linked to questions of colloquialization, which is defined in relation to these two modes.

2. Democratization, colloquialization and parliamentary language

The term "democratization" has been used in previous research in multiple ways to refer to processes involving societal, sociocultural or linguistic changes (Hiltunen and Loureiro-Porto, 2020). We take as our starting point Farrelly and Seoane (2012), for whom democratization consists of three aspects: *democratization* proper (the tendency of speakers to avoid unequal modes of interaction), *colloquialization* (the tendency to incorporate features typical of speech into written language), and *informalization* (the process of reducing the distance between the speaker/writer and the hearer/reader, resulting in an increasing acceptance of informal and interactional features).

Our focus in this paper is primarily on the second aspect, namely the process of colloquialization. The process of shifting to a more speech-like style in written genres has received a good deal of attention in diachronic corpus studies focusing on 20th-century English (e.g. Mair and Leech, 2006; Leech and Smith, 2009; Leech et al., 2009). For Mair, 1997a, colloquialization is a powerful explanatory factor accounting for many recent and on-going linguistic changes. As he puts it:

> some of the most drastic changes … are not directly due to grammatical innovation … Rather, they reflect a colloquialization of written English, that is a change in stylistic conventions which is due to a current of informalization and (pseudo-)democratization affecting advanced industrial societies (Mair, 1997a: 198).

According to this view, colloquialization is a discourse-pragmatic process where changes in the social realm are reflected in the changing frequencies of linguistic features with clear register associations, such that features characteristic of spoken language increase in written texts, and correspondingly features associated with formal and literary style decrease (Leech and Smith, 2009: 175). Leech et al. (2009, ch. 11) document the increase of several grammatical features in writing, including contracted verb forms (e.g. *don't*), semi-modals (e.g. *have to*), *not*-negation and questions, and attribute this increase to a general trend towards more spontaneous directness and immediacy in written communication (2009: 239). Colloquialization is also manifested in the decrease of features like *be*-passives and *wh*-relative clauses (Leech and Smith, 2009: 180–182). Typically, colloquialization seems to operate gradually over a long period of time, as shown for example by Biber and Finegan (1989), who have documented a general "drift" towards more oral styles in written genres from the 17th century onwards. On the other hand, some colloquialization processes apparently take place much more quickly, as demonstrated by Rühlemann and Hilpert (2017), who observe a dramatic increase of colloquial inserts (incl. response particles like *yeah* and discourse markers like *well*) in journalistic writing in the 1990s–2000s.

What is crucial to the analysis of colloquialization (and of other determinants of linguistic change) is the perspective of registers, since individual registers and genres differ considerably in terms of how receptive they are to linguistic and stylistic changes. This has been established by Hundt and Mair (1999), who show how academic prose is relatively unaffected by many recent changes observable in other written genres like newspaper writing. Based on these observations, Hundt and Mair propose a cline of openness to innovation ranging from "agile" to "uptight" genres. This idea is further elaborated on by Biber and Gray (2013), who note that while register differences are crucial to documenting grammatical changes in corpora, even individual sub-registers can exhibit substantial differences in the patterning of linguistic features. For this reason, Biber and Gray argue that sub-register differences should be considered in the research design, because ignoring them "would confound the description of linguistic change with patterns that in fact reflect register differences" (2013: 130).

In this study we focus on a single register, namely parliamentary discourse, which is a particularly interesting one from the point of view of colloquialization and corpus linguistic analysis. As we adopt a corpus linguistic perspective, our source of data are the written records of parliamentary sessions as recorded in *Hansard*. In other words, the material is drawn from a written report, which has its origins in a specific type of (typically planned) spoken language that comes close to, but cannot be defined as, language "written-to-be-spoken". Although Members of Parliament are generally not allowed to read prepared statements,

they may use notes, and a variety of conventions and restrictions govern parliamentary language use. As such, the parliamentary record provides an interesting testing ground for the colloquialization hypothesis: can we observe an increase of frequency for features of spoken language in *Hansard* due to their becoming increasingly accepted in written language over time? Another advantage of using *Hansard* for this type of study is that it provides uninterrupted temporal continuity over a 200-year period, which enables us to investigate changes in real-time data. That said, the long time period also means that the reporting conventions have changed considerably, which means that different subsections of the *Hansard Corpus* are not directly comparable, and this needs to be taken into account in the study design; these changes are discussed in more detail in Section 3 below.

There is considerable agreement on and evidence of a trend towards colloquialization (and informalization) especially in 20th-century English, and therefore it can be expected that the same processes are also observable as linguistic changes in the parliamentary record. Using evidence from readability metrics, Spirling (2016) has suggested that parliamentary speeches became less complex after the Second Reform Act in 1867, which doubled the electoral roll to include poorer and less educated voters. Some contemporary descriptions of parliamentary speech also support the idea of increasing colloquialization in the MPs' styles of speech in the time period in focus. For example, McDonagh's (1913) account of the difficulties faced by parliamentary reporters also makes reference to a colloquial and conversational style, although in a non-technical sense:

> The style of speaking that has become common in Parliament is what is called "the conversational" - - though it is not the speed of this new style that worries the reporter so much as its colloquial mode and subdued intonation. It is very clever talk in its way. It is simple, plain and natural, sometimes dropping, it may be, to a pedestrian and prosaic level - - it is, perhaps, more in consonance with actual life and reality than the old set speeches of lofty and studied eloquence. (1913: 20)

The colloquialization hypothesis is further supported by Kruger and Smith's (2018) study on the Australian Hansard in the 20th century, providing evidence of change in the frequencies of several pertinent grammatical features. For example, they identify a decrease in the use of agentless passives and relative clauses, and an increase in the frequency of contractions and emphatics, although evidence towards a densification trend and a more compressed style can also be observed, which makes it difficult to draw generalizations (see also Biber and Gray, 2012). A decline in the frequency of the *be*-passive has also been attested in the *Hansard Corpus* by Hou and Smith (2018), although their main focus is on explanatory factors other than colloquialization.

Indeed, as Mair and Leech (2006) point out, colloquialization processes are unpredictable and difficult to model with precision, because they involve different combinations of semantic and pragmatic factors depending on the context and the features in focus. A further difficulty in colloquialization research is that corpus-based approaches (in Tognini-Bonelli's, 2001 sense), which are followed in most studies, are not necessarily conducive to discovering new patterns in the data (Rühlemann and Hilpert, 2017: 105). In the present study, we adopt a pattern-driven approach using clustering and principal component analysis (described in Section 4), which enables us to identify features of interest without *a priori* assumptions. After features that are relevant to the topic of colloquialization have been identified, they can be analysed closely to obtain a more nuanced view of how colloquialization affects the parliamentary record.

3. Changes in parliamentary reporting

Currently, the official record of the parliamentary debates in the United Kingdom (as well as in many Commonwealth countries) is provided in *Hansard*, and it is also recognized as an important resource by Members of Parliament (MPs). It acquired an official status in 1909, having been established by Thomas Curson Hansard as a supplement to William Cobbett's *Annual Register* 1803 (Jordan, 1931: 439). T.C. Hansard started publishing the debates under his own name in 1812 after he had bought *Cobbett's Parliamentary Debates* (Rix, 2014: 456), and in 1829 the series was titled *Hansard's Parliamentary Debates* (MacDonagh, 1913: 428). After Hansard's death in 1833, his son continued publishing *Hansard* (Rix, 2014: 457). The history of *Hansard* offers a window into societal democratization and the history of institutions in the UK, and information about its composition is also crucial when using it for linguistic research in general, and in the analysis of colloquialization in particular.

From a societal perspective, having accurate and reliable reports of the parliamentary proceedings publicly available is an important aspect of a democratic society today, as it ensures that people know of the actions of the legislators they have voted into office. However, for a long time it was thought that parliamentary proceedings should remain private, because in that way MPs would not be influenced by the general public, and it would be easier to debate controversial issues (Aspinall, 1956; Eagles and Rix, 2015). It was also feared that, if the debates were published, some Members' opinions and actions could be misrepresented (MacDonagh, 1913: 131). Before the establishment of an official report, information about parliamentary debates were only available in newspapers and publications devoted to this purpose, and these reports were selective and incomplete, in part due to unfavourable conditions of reporting[1] but also due to the covert manipulation of politicians (Williams, 2010: 72). Many of the actions towards the full, official report were taken in the 19th century, as were many developments towards a more representative democracy, such as extending the franchise.

The early *Hansard* was not a complete or accurate source of what was actually spoken in Parliament either, but the reports were collated and summarized from newspaper accounts, especially from *The Times* (Port, 1990: 179; Rix, 2014: 457). Even after reporters were given the right to take notes on the back row of the Stranger's Gallery, this right was not indisputable, as a Member of Parliament could ask the Speaker to exclude them from the House by saying "I espy strangers" (MacDonagh, 1913:

[1] Note-taking was banned until 1783 in the House of Commons and until 1830 in the House of Lords (Maartens, 2019: 229).

308, 401).[2] Typically, speeches of Ministers and ex-Ministers were fully reported in the first person and other Members' speeches were condensed in the third person. In addition, Hansard gave Members the opportunity to revise the reports of their speeches before publishing them, or even send a complete manuscript for publication.[3] These practices naturally led to complaints about the incompleteness and bias of the reports (see Jordan, 1931 for some examples).

In 1878, the House of Commons decided to pay Hansard an annual subsidy to enable him to employ a staff of note-takers to the Chamber, especially to report the proceedings after midnight, which were often not reported by newspapers (MacDonagh, 1913: 429). In 1908, following a recommendation of a Select Committee, it was decided that there should be an official report of the debates and that the Commons should employ its own reporting staff and every speech should be reported in full and in the first person (MacDonagh, 1913: 442). These arrangements came into operation in the House of Commons in 1909.[4] The definition of the official report was adopted in 1907 by the Select Committee on Parliamentary Debates (HC 239 1907) as being one which:

> though not strictly verbatim, is substantially the verbatim report with repetitions and redundancies omitted and with obvious mistakes corrected but which, on the other hand, leaves out nothing that adds to the meaning of the speech or illustrates the arguments.

The introduction of the official report thus represents a major "environmental change in the textual habitat" (Szmrecsanyi, 2016: 153), which is expected to result in systematic changes in the frequencies of a variety of linguistic features when data is compared from the periods before and after the official report was introduced. Furthermore, these changes can be expected to reflect the increasing colloquialization of *Hansard*, given that the official report clearly represents spoken debates more accurately than third-person summaries do.

4. Data and methods

4.1. The Hansard Corpus

As our primary data, we use the full-text stand-alone version of the *Hansard Corpus* (Alexander and Davies, 2015).[5] We focus on the period 1870–1930, which contains 371 million words, the vast majority of which come from debates in the House of Commons, as shown in Fig. 1. In this paper, we will focus exclusively on the debates in the House of Commons.

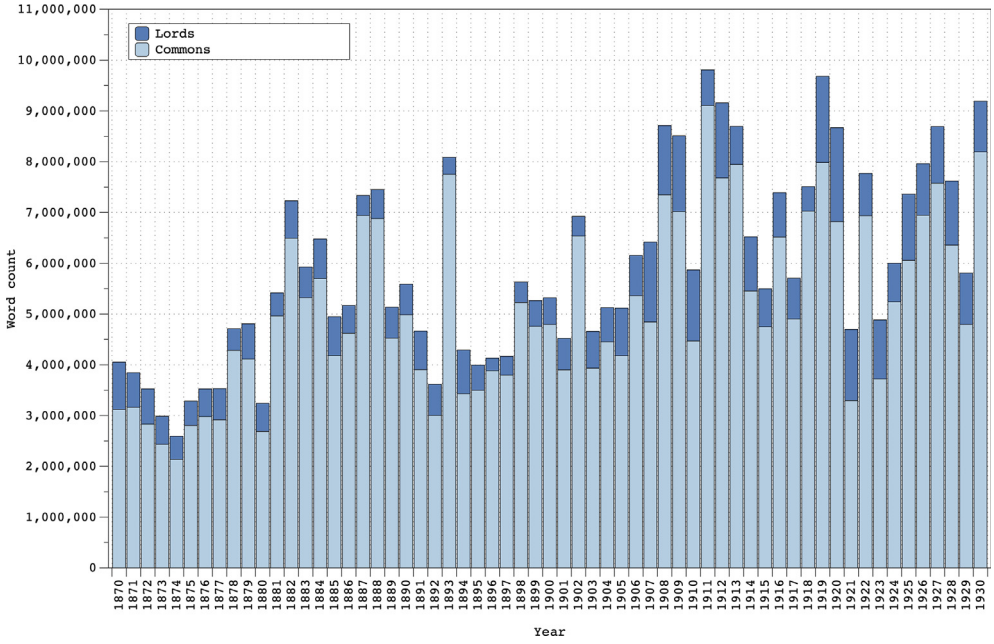

Fig. 1. Word counts per year for both houses of Parliament in the *Hansard Corpus* (1870–1930).

[2] After the Houses of the Parliament were burnt down in 1834, reporters were provided with their own place in the new buildings of both Chambers (Eagles and Rix, 2015). Lack of space in the press galleries remained a problem and a source of discord throughout the 19th century (Maartens, 2019).

[3] An asterisk (*) was added after the name of the speaker whose speech had been revised (MacDonagh, 1913: 431).

[4] In the House of Lords, an almost official system of reporting had existed since 1889 (Jordan, 1931: 442). Hansard did not have staff in the Upper House before that, for which reason the reports before 1889 are very imperfect (Jordan, 1931: 442).

[5] The corpus was originally compiled at the University of Glasgow by the JISC Parliamentary Discourse project in 2011 by Jean Anderson and Marc Alexander and developed further by the SAMUELS project. The corpus is freely available online on the English-Corpora.org website (https://www.english-corpora.org/hansard/) as well as on the Hansard at Huddersfield website (https://hansard.hud.ac.uk), in addition to which the texts of *Hansard* can be accessed and downloaded from the Parliament's own website (https://hansard.parliament.uk).

In the stand-alone version of the *Hansard Corpus*, each speech is represented as a separate TEI XML-annotated file that contains several items of metadata in addition to the written record of the debate itself. These files are at the bottom of a hierarchical directory structure in which years, months and days are represented by nested folders.

4.2. Methods

As noted earlier, most previous studies on colloquialization have adopted a corpus-based perspective (Tognini-Bonelli, 2001), taking as their starting point sets of linguistic features that have been found in previous research to index either colloquial and informal style, or formal and literary style. Features representing the former category (e.g. semi-modal verbs, contractions, and progressive verb forms) have been found to increase over time while those in the latter category (e.g. *be*-passives) have decreased, which in turn would support the hypothesis of written registers becoming increasingly colloquial (see e.g. Collins and Yao, 2013; Hundt and Mair, 1999; Leech et al., 2009; Leech and Smith, 2009; Mair and Leech, 2006; Smitterberg, 2008). While these studies are valuable and provide convincing evidence of the operation of colloquialization as a determinant of linguistic change, their methodological setup limits them to providing only one side of the story. As argued by Rühlemann and Hilpert (2017: 105), because corpus-based studies only investigate specific previously known features, they run the risk of overlooking changes that may have taken place in the frequencies of other features, which had not been investigated in earlier work. To tackle this issue, Rühlemann and Hilpert's study on the *TIME Magazine Corpus* employed a corpus-driven approach: they first used keyword analysis applied to the BNC to determine what features are characteristic of conversation, and then investigated how the frequencies of these features change over time in the *TIME Magazine corpus*.

Our data-driven investigation into the possible colloquialization of the Hansard record is conceptually similar to Rühlemann and Hilpert (2017) and has similar aims, but differs from it in how the method is implemented. In particular, we adopted a *pattern-driven approach*, which is described by Tyrkkö and Kopaczyk (2018) as a theoretical middle ground between corpus-based and corpus-driven analysis. While the first is based on extensive *a priori* assumptions and knowledge-based queries and the second is, conceptually at least, aligned with the idea of a fully theory-agnostic starting point, the objective of pattern-driven analysis is to focus on repetitive patterns such as n-grams, POS-grams and open-slot lexico-grammatical constructions, which rely on knowledge-based starting points but allow for freely occurring variation and analyses thereof. Our specific focus here is on n-grams, or recurrent sequences of word forms in a corpus.

The reason we focus on lexical n-grams is based on the idea that large-scale diachronic trends of high-frequency word sequences provide an efficient inroad into capturing register changes in a maximally informative manner. In particular, n-gram analysis reveals commonly occurring multi-word units in terms of formulaic features and lexical bundles. The former are strongly associated with formally managed registers, such as legal and religious language, and the latter with contexts where speakers and writers have a tendency to employ pre-established patterns without implicit external pressure to do so (Tyrkkö and Kopaczyk, 2018: 2–6).

In the case of the Hansard record, our two-fold premise is that parliamentary language is likely to have become more colloquial over time and that the methodological shift in the recording of the debates that took place in 1909 will be visible as changing frequencies of a broad range of lexical n-grams. As for specific (groups of) n-grams and their functions in the data, we further expect colloquialization to be visible as an increase in the use of items linked to spoken usage and affect, and a corresponding decline in features associated with formal written language, in accordance with previous work on colloquialization reviewed above.

Our approach, summarized in Fig. 2, consists of (1) identifying key n-grams from the point of view of colloquialization and democratization in a data-driven fashion from the parliamentary record 1871–1930 (covering ca. 30 years before and 20 years after the introduction of the official report), (2) using changes in the frequencies of these n-grams as indicators of linguistic change during the main time period and after, (3) analysing the use of these terms in the context of parliamentary debates with the help of concordances, in order to consider their status as markers of colloquialization, and describe their discourse functions in detail.

4.2.1. Retrieval of n-grams

N-grams (also referred to as *lexical bundles, multiword units* or *clusters*) are recurrent sequences of word forms in a corpus (Biber et al. 1999; Gray and Biber, 2015).[6] Given that n-grams are identified on the basis of sequential co-occurrence frequency and not the occurrence of specific pre-defined lexical items, this approach amounts to a "radical corpus-driven approach" (Biber, 2009). Furthermore, because n-grams do not necessarily form complete grammatical structures, the method enables

[6] The terminology concerning n-grams is not entirely consistent within the field of corpus linguistics. In general, the term n-gram is a neutral term that refers to a sequence of items (typically characters, words or tags) that is of interest for some reason, with the letter n indicating the length of the sequence. Lexical bundles are n-grams that also fulfil additional quantitative criteria to do with the minimum frequency of occurrence and occasionally also distributional thresholds. Multi-word units are typically understood to be lexical n-grams that make up units that are intuitively recognizable to human evaluators. Clusters are often understood as lexical n-grams that include one or more predetermined items, such as all 3-grams that include the word "good".

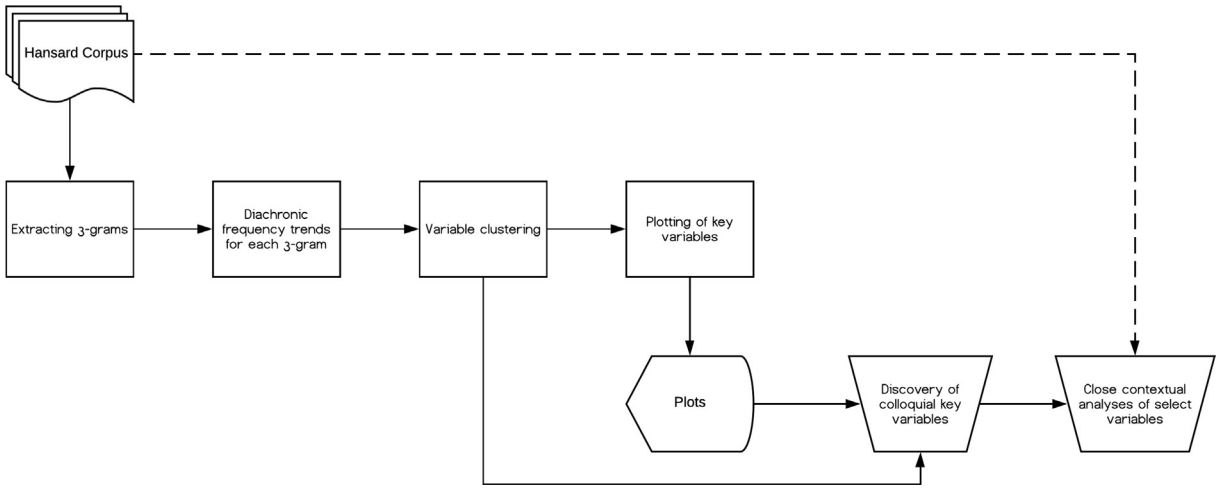

Fig. 2. Flowchart of stages of research process.

us to trace diachronic developments in the (reported) language of Parliament both as they pertain to conscious choices and to 'building blocks' of language that exist below the level of consciousness of speakers.

We first wrote an R script generating a comprehensive frequency-ranked list of all n-grams (n = {3,4}) for each year in focus,[7] merged these lists into one list and extracted from it the 1000 most frequent types and their frequencies. As our interest is in the possible colloquialization of parliamentary language, we wanted to focus on n-grams that are discourse-functional and potentially indicative of style, rather than those that would reflect the contents and substance of the texts ("aboutness"). For this reason, we removed all content-based n-grams such as *house of lords, house of commons* or *the right honourable*[8] from the list. After this, we were left with 339 n-grams for the House of Commons. We then determined the normalized frequencies of these n-grams and used this data as input for the statistical analyses. Importantly, because the 3-grams studied are all topically generic high-frequency items, we did not delve into speaker-specific usage. Although it is possible that a small number of individual MPs could overuse certain 3-grams to the extent that their frequencies appear abnormally high, this is very unlikely, and it is almost certain that such high frequencies would be sustained over a longer period of time. During the years under investigation here, the House of Commons had between 650 and 700 MPs at any given time, and altogether thousands of individual MPs.

4.2.2. Statistical analysis of n-gram frequencies

The starting point of the statistical analysis was that each of the 3-gram patterns discovered has its own unique frequency pattern over the timeline, typically showing either a positive or negative cline. Although individual items show year-by-year fluctuations in frequency and some may exhibit fairly substantial peaks and valleys in individual years, from the perspective of the present study the phenomenon of interest is the way in which the trendlines of the 3-grams reveal more general diachronic changes. Thus, while the patterns of individual 3-grams can be interesting in and of themselves, our focus was on discovering overall trends in the data, which requires analysis of the full repertoire of frequent 3-grams as a whole, rather than focusing on individual items. Our next step was therefore to make use of statistical grouping methods, also known as machine learning methods, which reveal the bigger picture. The methods used were *hierarchical clustering*, *principal component analysis* and *variable clustering*.

The overall methodological paradigm of statistical grouping involves the identification of similarities in the variable patterns. In the present study, we were interested in two main questions. Firstly, we wanted to discover whether years that are chronologically close to one another appear similar when it comes to the frequencies of the 339 3-grams identified in the first stages of the study, and secondly, whether the frequency trends of the individual 3-grams resemble one another, which would suggest that they may be related to a more general process of change, such as colloquialization. Although the two questions are closely linked, their implications are somewhat different.

Hierarchical cluster analysis was used for examining whether a large-scale change took place in the language of *Hansard* before and after 1909, when the sourcing of *Hansard* transitioned from press reports to essentially verbatim reports by parliamentary reporters; for details of the method, see, e.g., Desagulier (2017: 276–281). In short, two-way matrices were created of the frequencies of the 3-grams and individual years. These matrices were then standardized for each 3-grams by first calculating the mean frequency and standard deviation of each 3-gram over the timespan and then transforming the frequencies into z-scores. This is done so that the frequency differences between the different 3-gram are mitigated. Hierarchical cluster analysis was then carried out. We used Ward's method, in which clustering is approached in an agglomerative

[7] The bundles are case-insensitive. N-grams were retrieved with the help of the R package *ngram* (Schmidt and Heckendorf, 2017).
[8] See Archer (2018) on the use of modes of address in Hansard.

fashion. At the beginning of the clustering process, each year starts out as an independent observation, and the algorithm then works in a stepwise fashion to iteratively identify and group together the most similar years, or previously identified clusters of years, until every year has been included in the clustering tree. When reading the resulting dendrogram (Fig. 3), it is worth remembering that the order in which the years appear on the left-hand side of the dendrogram is based entirely on the frequencies of the 3-grams included in the model. The fact that years that are chronologically close to each other also appear close to each other in the graph signals that the 3-grams were used at similar frequencies during those years.

Principal Component Analysis (PCA) is a well-established statistical method that is widely used for identifying underlying patterns in multivariate datasets (see, e.g., Desagulier, 2017: 242–245). In short, the method is based on analysing the variables as a multidimensional space and finding linear trajectories through the data that maximize variance. The number of these so-called eigenvectors is relative to the number of variables. The eigenvector with the highest eigenvalue is called the first component, the second best is called the second component, and so on. The first two to three components typically explain the vast majority of the variance and the convention is to interpret the first few components based on the close analysis of the distribution of the variables and, if possible, to give them descriptive labels, such as 'time' or 'register'.

Variable clustering is a statistical variable reduction method that facilitates the discovery of similar patterns across a large number of variables, in the present case the lexical 3-grams. In short, we take the unique frequency pattern of each 3-gram over the timeline, and then compare all the frequency patterns with each other, identifying reasonably similar patterns and grouping the relevant words together. We employed the PROC VARCLUS method,[9] in which variables are grouped into clusters iteratively according to the first two principle components at each iterative step. In each cluster, one of the items is the most typical representative of the group, with the others being more or less similar to it; the degree of similarity can be quantified and a minimum threshold can be set. Once the key items have been identified, they can be used as proxies for their respective groups of items, with the obvious caveat that the similarity threshold for group membership ought to be fairly high (see Appendix A for more information).

4.2.3. Qualitative analysis

After observing the general trends in the data, we look more closely at specific n-gram clusters, as well as individual n-grams which are particularly relevant to the colloquialization hypothesis. This means looking at their frequencies not only in the period in focus (1870–1930) but over the entire period covered by the corpus (1803–2005). In addition, we are mindful of the fact that individual n-grams can be linked to different grammatical structures, which in turn may realize different discourse functions, and that taking these functions into account requires close reading of extracts from the corpus in context. We illustrate the importance of a detailed, contextual discourse-pragmatic analysis as a complement to a statistical pattern-driven approach by focusing on two 3-grams that are stylistically key from the perspective of colloquialization, namely *is going to* and *I think it*.

5. Results

This section is divided into two major parts. Section 5.1 reports the findings of the data-driven analysis based on the application of the statistical techniques (hierarchical cluster analysis, principal component analysis, and variable clustering), and Section 5.2 zooms in on two key 3-grams, *is going to* and *I think it*.

5.1. Cluster analysis

As explained in 4.2., we used hierarchical cluster analysis to determine whether the language of *Hansard* exhibits changes around 1909 following the introduction of the official report. Fig. 3 shows the results of the clustering. Each row represents a year and each column a unique 3-gram, such as *I think it* or *is going to*; we will return to these examples later. The colours of the heatmap go from green to white to red, with green representing a low standardized frequency and red a high standardized frequency. The order in which the years appear at the left-hand side of the image is based on cluster analysis. Thus, the first row represents the year 1871, the second row represents the year 1884, and so on. The fact that 1871 and 1884 are next to each another means that when the frequencies of all the 339 3-grams are considered together, these two years are very similar. The black horizontal lines have been added to highlight the major clusters in the data, also seen in the horizontal tree diagram on the right-hand side of the image. There is no space to display the labels of the individual 3-grams in the image, but the vertical tree diagram at the bottom of the image shows how they cluster together. We explore their distributional patterns below using principal component analysis and variable clustering.

Fig. 3 provides strong evidence that the introduction of the official report marks a major stylistic shift in the data, which is witnessed in the changing distributions of the 3-grams. As can be seen, there are four major clusters in the timeline. At the top, we have the oldest part of *Hansard*, and there appear to be relatively few differences between the years when it comes to the use of the most common 3-grams. At the bottom of the figure, we have nearly all the texts from 1909 onwards, again showing a very uniform distribution of 3-gram frequencies. This division fits perfectly with the prior knowledge that the practices of *Hansard* changed around this time: we can thus conclude with considerable certainty that the oldest records and the youngest records are linguistically and stylistically very dissimilar. As such, it also confirms that our chosen method is a

[9] For details, see *SAS/STAT User guide*, version 15.1. The VARCLUS method was developed by Warren Sarle at the SAS Institute.

suitable one for investigating how environmental changes may bring about changes in text frequencies and identifying which changes are key in specific datasets.

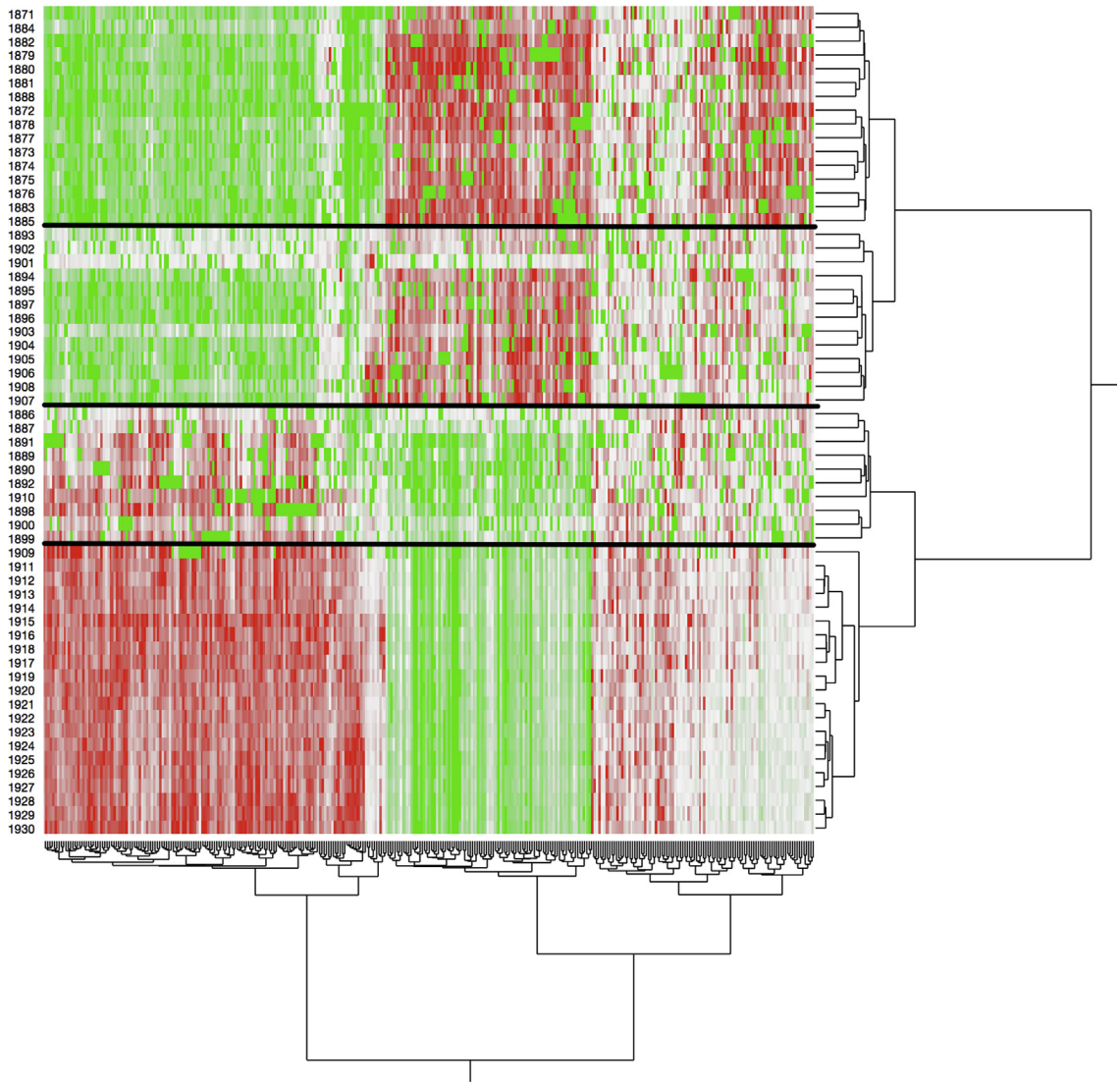

Fig. 3. Hierarchical cluster analysis of n-grams in the *Hansard Corpus*.

In the middle of the figure, we have two clusters that paint a more diverse picture. The order or grouping of the years is not simple to interpret, and we can only say that either the language used in Parliament, or the way it was reported and recorded in *Hansard*, were undergoing a change. Importantly, as these clusters cover a roughly 20-year period before 1909, this would suggest that the linguistic changes brought about by the official report were gradual rather than abrupt. This stylistic variation is likely to be linked to other changes in the extralinguistic context: the period coincides with the fourth series of *Hansard* (1892–1908), which Alexander and Dallachy (2019) describe as "chaotic" due to the fact that the Hansard record's printer changed several times.

To investigate the grouping of 3-grams in more depth, we carried out a Principal Component Analysis on the 3-grams, the results of which are shown in Fig. 4. As can be seen the first component explains 72.4 per cent of the variation, and close analysis of the distribution of items shows it relates strongly to time. On the left-hand side of the figure, we can identify features of third-person reporting (*that he had*, *in which the*), whereas features of spoken language (*I am not*, *to me that*) are concentrated on the right. In the middle of the plot, we see generic 3-grams that are common to both third-person reporting and the verbatim report (*and in the*, *of the great*).

To dig even deeper into the specific items that appear to have been affected by the introduction of the official report, we looked into the trends of specific n-grams. However, rather than looking at individual 3-grams, we first employed *variable clustering* to identify patterns of diachronic frequency change in the whole data set (see 4.2.4), and then focused on groups of 3-grams that exhibited particularly interesting trends.

The results of the variable clustering are shown as a circular dendrogram in Fig. 5, and Fig. 6 zooms in on one section of the same dendrogram. Each cluster of variables in the dendrogram is represented by the 3-gram that is its statistically most typical member. As can be seen, the number of variables differs widely between the clusters. For example, the most typical member of cluster 16 (the top left side of Fig. 5) is *the whole of*, and it includes eleven 3-grams (see Appendix A for more information). After identifying these key 3-grams, we plotted their frequencies diachronically and employed visual data exploration to identify items of potential interest from the perspective of colloquialization and the introduction of the official report.

Four examples of such 3-grams are shown in Fig. 7: *is going to*, *I will not*, *that it was* and *he did not*. For these items, we can observe interesting changes both in the frequencies and the amount of dispersion. First, *is going to* increases rapidly after 1909, while frequencies of *that it was* and *he did not* appear to drop significantly around the same time. For *that it was*, the trendline appears relatively stable at c. 500 hits per million words until about 1888, after which the year-by-year frequencies begin to exhibit considerable variation: first a few low-frequency years, then a hike from 1894 to 1898, then a drop followed by another steep positive cline to 1909. And then in 1909, the frequency drops to c. 200 hits per million words and remains level thereafter.

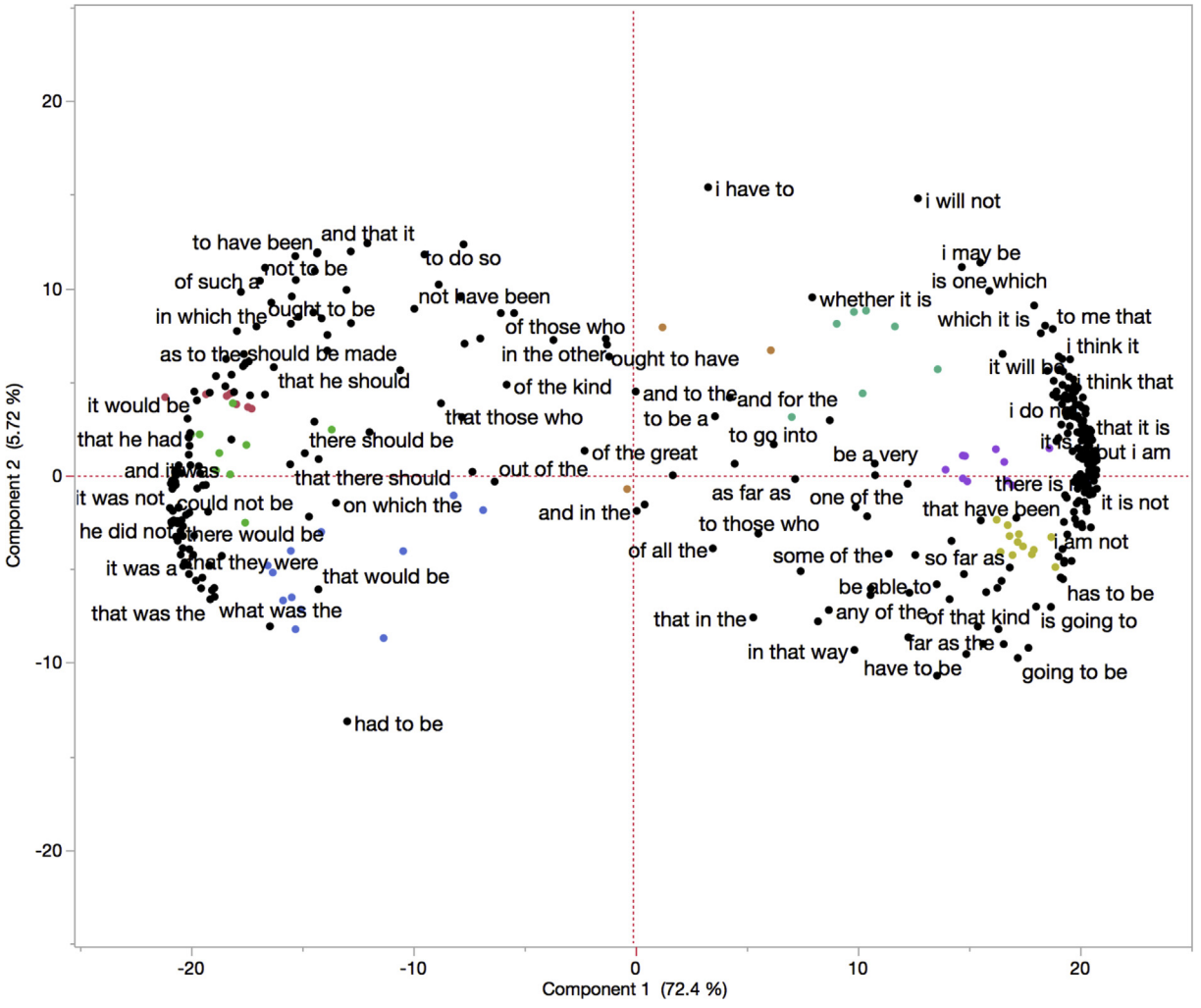

Fig. 4. Principal Component Analysis of n-gram frequencies in the *Hansard Corpus*.

This strongly indicates that the 3-gram *he did not* and *that it was* are stylistically linked to third-person reporting, whereas *is going to* is associated with near-verbatim first-person reports.

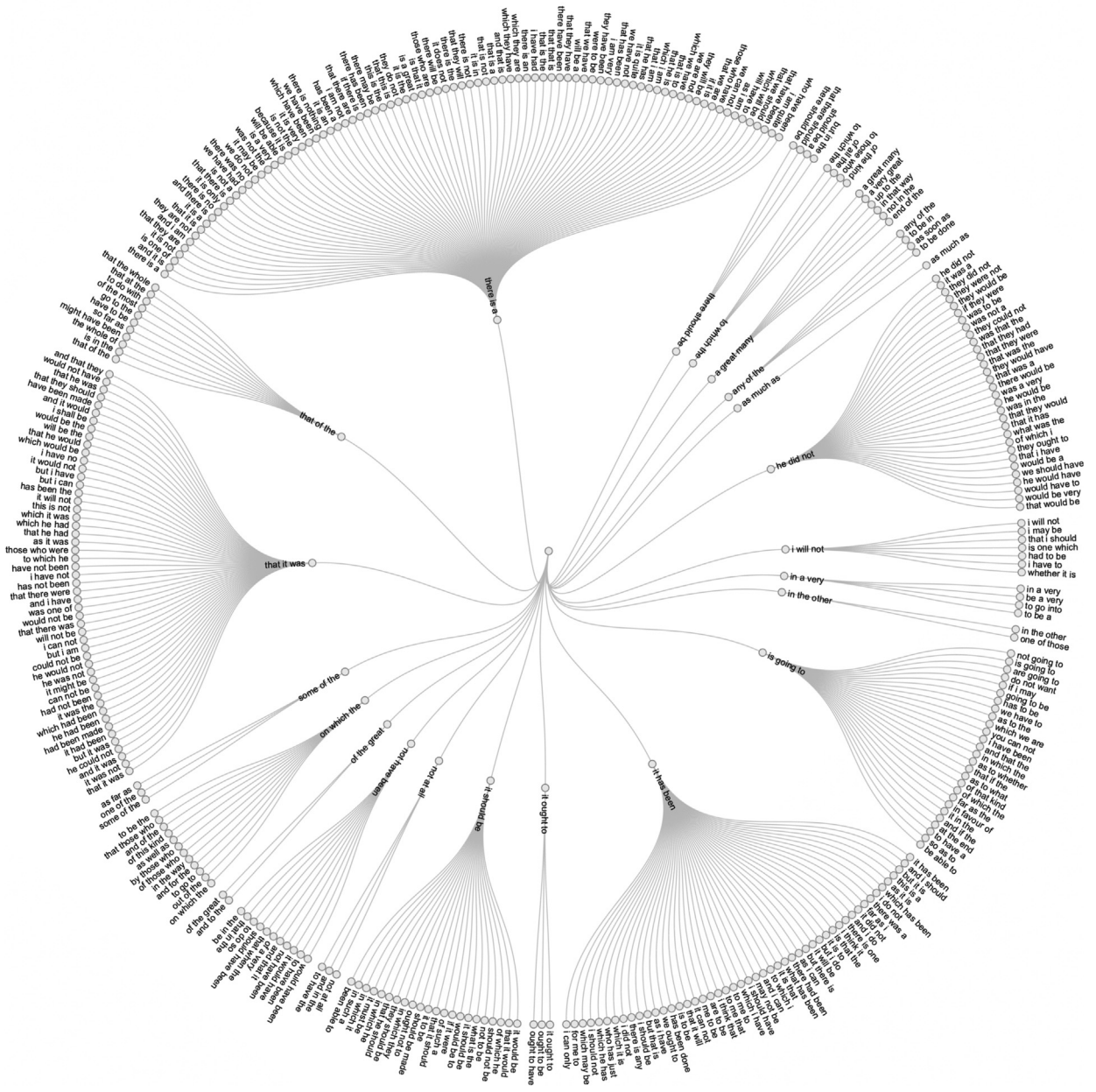

Fig. 5. Key n-grams and their corresponding cluster components in the *Hansard Corpus*.

This interpretation is also confirmed by concordance data: for *he did not*, the antecedent is typically the MP whose speech is being reported, as illustrated in example (1). The 3-gram *that it was* is typically a fragment of a content clause controlled by a noun or a verb, and is similarly associated with third-person reporting in the 19th-century Hansard (example 2).

(1) Mr. Norwood … said that the subject which he wished to bring before the House was so intimately connected with the one under discussion that they might advantageously be considered together … He regretted that **he did not** share the sanguine views which the hon. Member for Hastings (Mr. T. Brassey) took of the state of the Mercantile Marine, which from careful inquiry, as well as from personal experience, he believed to have greatly deteriorated of late years. (House of Commons, 17 April 1874)

(2) Mr: Bouverie contended that the procuring of these Returns would involve a great amount of labour, and necessitate an expenditure of considerable magnitude: … He spoke from experience when he stated **that it was** impossible ever to procure satisfactory Returns of this nature from unpaid officials: … (House of Commons, 27 April, 1860)

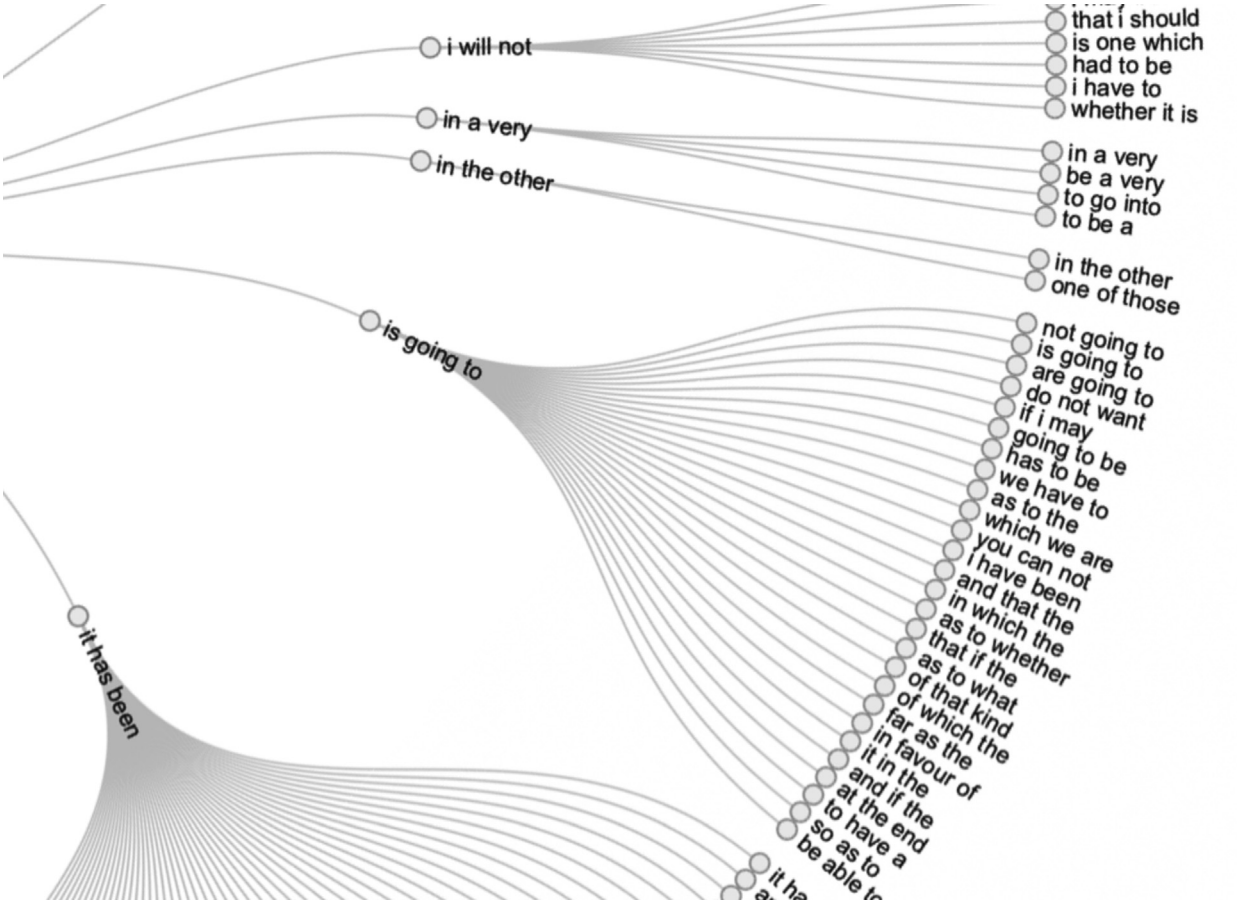

Fig. 6. Closer view of key n-grams and their corresponding cluster components in the *Hansard Corpus*.

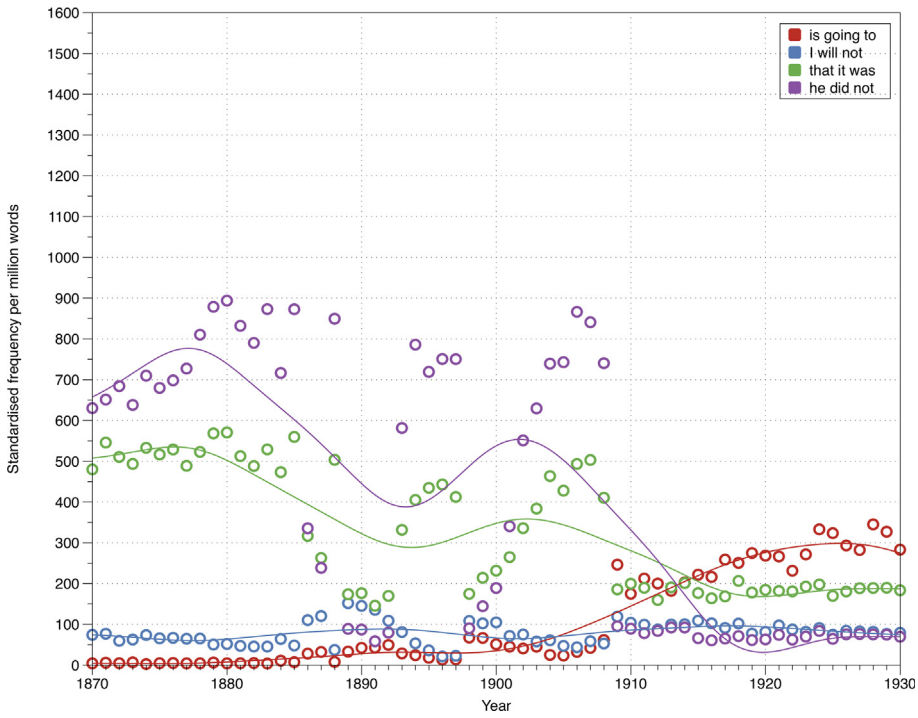

Fig. 7. Frequencies of specific n-grams in the *Hansard Corpus* 1870–1930.

Although the 3-gram *is going to* can in principle be also used in third-person reporting, it is more commonly found in first-person reports exemplified in (3), which accounts for the sharp increase in its frequency after the 1910s. We analyse the development of *is going to* more closely in 5.2.1 below.

(3) Something may be done, but I cannot believe it **is going to** have the effect to which my noble Friend has pinned his faith (House of Commons, 12 February 1920)

Interestingly, the year-by-year frequencies of *that it was* and *he did not* become remarkably steady after the near-verbatim report was introduced. This difference cannot be explained by a change in the MPs' collective language use, but is likely to be linked to stylistic differences in the reports from which the Hansard record was collated. The other two 3-gram *is going to* and *I will not* show a similar stability, though in their case the yearly frequency changes were minimal even before the threshold. Finally, we can observe that the frequency of *I will not* remains more or less level throughout the timeline. This observation supports the interpretation that the change in 3-gram frequencies is not simply due to an increase in the use of first-person pronouns.

Next, we shall look more closely at the use of two 3-grams, *is going to* and *I think it*, considering their frequency across the whole period represented by *Hansard* and paying attention to their functions and typical contexts of use.

5.2. Qualitative analysis

5.2.1. is going to

The 3-gram *is going to*, which emerged as important in the quantitative analysis, has also been linked to colloquialization in previous research. The 3-gram is clearly linked with the *going-to* future, a grammatical structure that is in many contexts interchangeable with the *will* future, and when this is the case, it is considered the more informal alternative (Mair, 1997b: 1538). Thus, if the parliamentary record changes due to colloquialization, we would expect the *going-to* future to become more frequent at the expense of *will* and also *shall*, which has undergone a considerable decline as a future marker in recent decades. If we look at Fig. 8 below, this seems to be the case, as the frequency of *is going to* shows a strongly increasing trend between 1909 and 1930.

In Fig. 8, we can see that *is going to* is fairly infrequent up until 1909, and the increase in the frequency of this 3-gram right after the establishment of the official report suggests that the language in the "substantially verbatim" Hansard was more colloquial than in the series where more indirect speech reporting was used. However, as can be seen in Fig. 9 below, which shows a longer diachronic trendline for the same n-gram, the diachronic development of the frequency of *is going to* does not in fact continue in the 1930s, and the decrease in the frequency since then would challenge the idea of a continuing trend towards an increasingly colloquial parliamentary record. This finding highlights the importance of paying close attention to extralinguistic changes in the norms and conventions of reporting in *Hansard.*

So how do we explain this observation? First, according to Alexander and Dallachy (2019), the editing practices in Parliament started to develop in the 1920s and 1930s towards a more formal style associated with the current Hansard, and it may be that the decrease in the frequency of *is going to* can be at least partly explained by reporters starting to avoid the informal future marker. However, the fact that the trend continues to decline until the 1990s suggests that there must be more to this than just a change in the editing style, and that the actual language used in the debates may have also changed.

To further explore the frequency changes of *is going to*, we categorized the items appearing in the subject position from 1870 to 1929 (see Fig. 10). We used a taxonomy of ten subject types (dummy (i.e. *it* or *there*), number, person (e.g. a name, a title), pronoun, *that*, *what*, *which* and *who*, passive, and other) and plotted the yearly distributions as a stacked bar chart. Keeping in mind the overall frequency changes over the time period, we can see somewhat surprisingly that although the subject type distribution is more stable than from 1910 onwards, the overall picture is relatively unchanged over the entire timeline. Minor, but nonetheless interesting increases are seen in the proportion of relativizers *what, which* and *that* in the subject position, while person subjects would appear to become somewhat less frequent. The latter observation, in particular, can be directly linked to the change of text type from narrative reports to near-verbatim direct speech.

The frequency trend of *is going to,* as well as the trends of all the items in the cluster of n-grams represented by *is going to*, suggests that the reporting practices of *Hansard* may play a much more significant role in the data than we previously suspected. Given that there has been a steady positive trend in favour of the increasing use of *going to* in British English over the last 100 years, it seems very unlikely that British MPs would have gone against this trend and decreased their use by two thirds over the same period.

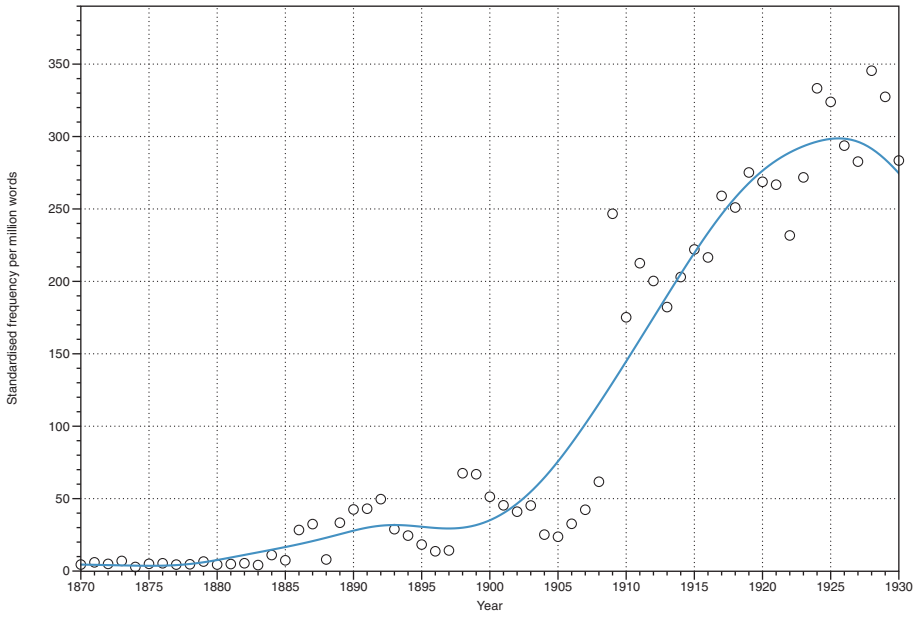

Fig. 8. Frequency of *is going to* in the *Hansard Corpus* 1870–1930.

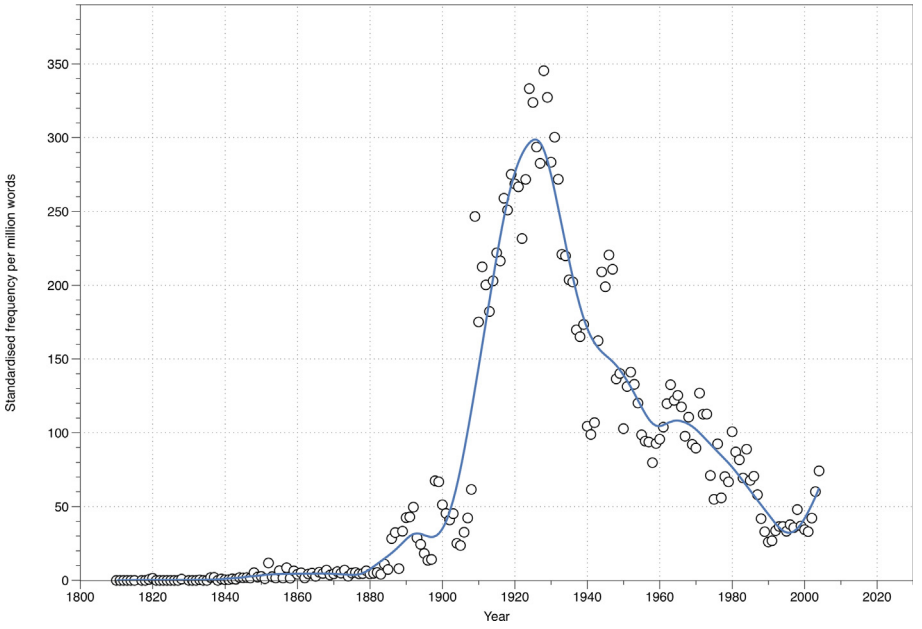

Fig. 9. Frequency of *is going to* in the *Hansard Corpus* 1803–2005.

5.2.2. I think it

I think it is another key 3-gram that shows interesting diachronic changes from the point of view of colloquialization and democratization. Typically, *I think it* is used as a hedge at the beginning of a clause to indicate uncertainty or mark the following statement as the speaker's personal opinion (example 5).

(4) Mr: HARMSWORTH **I think it** is only due to a clerical error in the office: Numbers, too, on the Blue Paper that is circulated, are not always the same as the numbers on the White Paper that we get. (House of Commons, 2 November 1930).

The frequency trends observed for *I think it* appear to be linked to this hedging function. This is indicated by the fact that many of the other 3-grams in the same cluster (top right side of Fig. 5 above) which show similar frequency trends are also

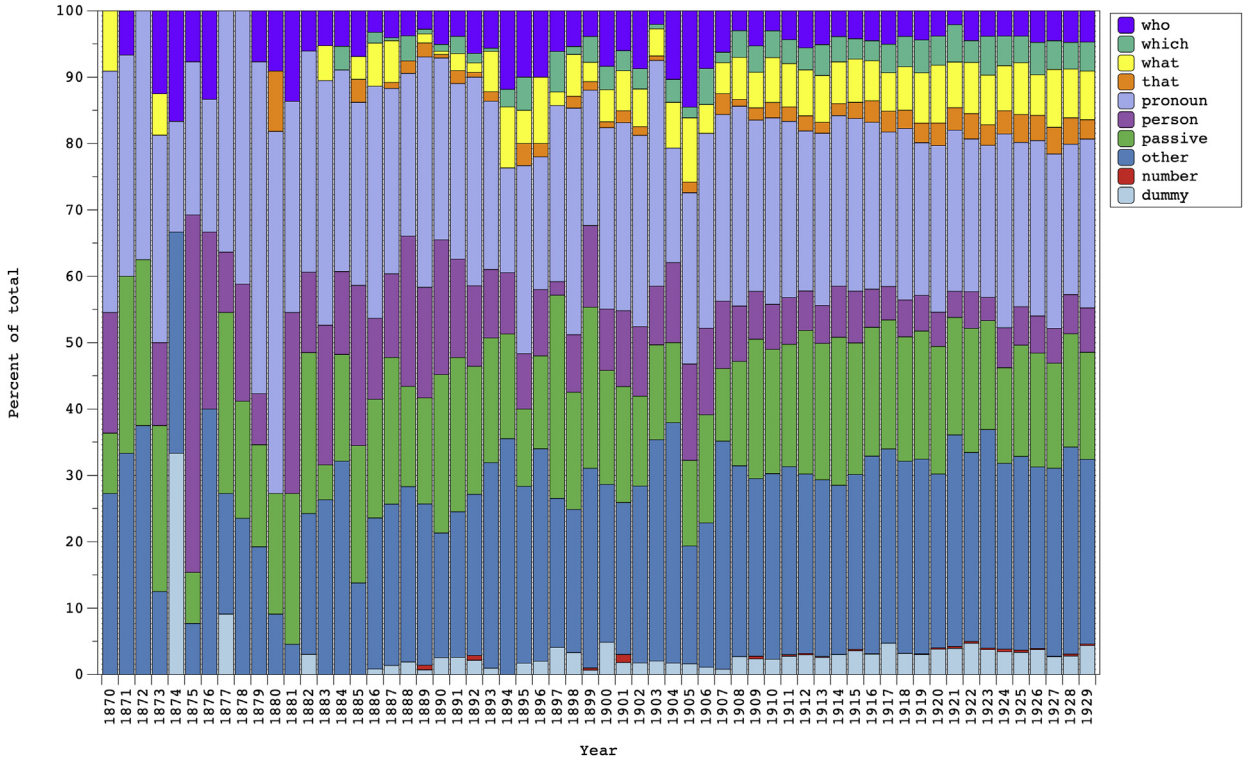

Fig. 10. Subjects of clauses containing *is going to* in the *Hansard Corpus* 1870–1930.

often used in hedging phrases: *to me that*, *far as I*, *as I can*, *I can only*. This cluster also includes 3-grams with the modal *may*, such as *which may be* and *may not be*, which also denote uncertainty.

In Fig. 11, we can see a peak in the frequency of *I think it* in 1909, which clearly shows the transition to the official report. However, the period between 1910 and 1930 is again interesting, because there is a great deal of fluctuation, which would seem to reflect the fact that the reporting conventions were still taking shape. Accordingly, we may surmise that the Hansard records from this period may represent the speeches more accurately than what we find in the later periods, when the editing process had become more uniform. During this period, the frequency of *I think it* decreases from the maximum of over 300 occurrences per million words to below 150 occurrences, and the current editorial guidelines in fact ask reporters to remove such otiose phrases as *I believe* and *I think* (Alexander and Dallachy, 2019).

Since the 1950s, the frequency of *I think it* has decreased quite clearly and by the 1980s, it has dropped below 50 per million words. However, similar to *is going to*, this drop in the frequency is not entirely explained by the editorial guidelines, as the decreasing trend has continued for at least three decades. Thus, it could be argued that the style of reporting in *Hansard* has affected the language used in the parliamentary debates and that MPs have adopted the style used in *Hansard* as the proper style of speaking in Parliament.

6. Conclusions

Methodologically, the main innovation of this mixed-methods study is the use of pattern-driven methods to identify the specific items that are stylistically relevant in the data, followed by a closer qualitative analysis of their frequencies and uses. By using this approach, the study provides evidence that colloquialization affects the parliamentary record across the board, thus converging with and confirming earlier observations about this register while also drawing attention to other features not discussed in previous studies on colloquialization.

The investigation of n-gram frequencies with the help of Hierarchical Cluster Analysis and Principal Component Analysis convincingly demonstrates that the introduction of the official report in 1909 represents a major stylistic divide in the parliamentary record, which is something all studies based on *Hansard* data ought to take into account in their research design and interpretation of findings. This finding can also be linked to democratization and colloquialization, albeit in complex ways. In terms of the former process, it shows how an environmental, language-external change principally motivated by societal democratization—the aim to provide a fair, accurate and comprehensive report of parliamentary debates to the public—can bring about a clear stylistic shift in the texts. As a "substantially verbatim" account of the debates, the official report clearly marks a shift towards an increasingly colloquial parliamentary report, one which incorporates elements

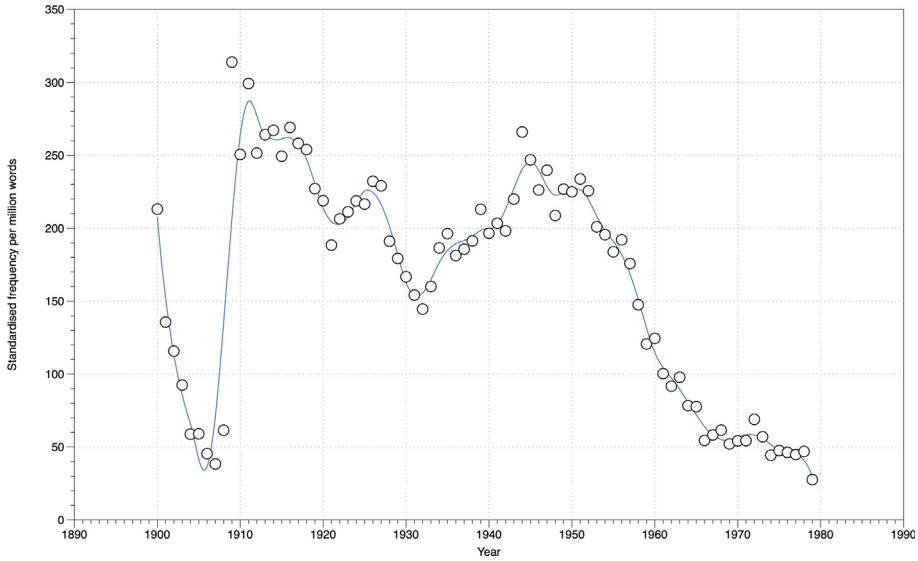

Fig. 11. Frequency of *I think it* in the *Hansard Corpus* 1803–2005.

of first-person reporting. Compared to the third-person summaries found in *Hansard* before 1909 it arguably represents a different "sub-register" (Biber and Gray, 2013) of the parliamentary record, which should be taken into consideration in any diachronic study on the *Hansard Corpus*.

A trend towards increasing colloquialization is also observable in the increase of informal linguistic variants as exemplified by the *going-to* future. Interestingly, the effect levels off quickly after the 1930s, along with the use of hedges such as *I think it*. This is at least partly due to an increasing editorial control of *Hansard*, which appears as the main determinant controlling and mediating the transfer of spoken-like linguistic features to the written report of parliamentary proceedings. The linkage between the conventions and the incidence of individual features, combined with the pragmatic analysis of the functional and contextual associations, emerges as a promising avenue for future research applying a pattern-driven approach to colloquialization.

Acknowledgements

This work was supported by the Academy of Finland (decision 258434). We would like to thank the SAMUELS project and especially Marc Alexander and Fraser Dallachy at the University of Glasgow for making the full-text version of the *Hansard Corpus* available to us. We are also grateful to the anonymous referees for their useful comments and suggestions.

Appendix A

The table shows the cluster members of cluster 16 in our data. The column "R^2 with own cluster" indicates how close the individual cluster members are to the cluster centroid, and the columns "R^2 with next closest cluster" indicates how close the items are to the next possible cluster. The column "1-R^2 ratio" is a measure of the relative distance between the cluster that the variable is a member of and the next most similar cluster. The lower the value in the column "R^2 with own cluster", the less well the frequency pattern of the item matches the cluster's central trendline.

Cluster	Cluster members	R^2 with own cluster	R^2 with next closest cluster	1-R^2 ratio
16	that of the	0.892	0.868	0.82
16	is in the	0.866	0.845	0.869
16	the whole of	0.829	0.644	0.479
16	might have been	0.824	0.783	0.812
16	so far as	0.804	0.685	0.622
16	have to be	0.779	0.631	0.598
16	go to the	0.698	0.64	0.837
16	of the most	0.656	0.537	0.744
16	to do with	0.532	0.399	0.778
16	that at the	0.521	0.434	0.848
16	that the whole	0.49	0.326	0.757

References

Alexander, Marc, Davies, Mark, 2015. Hansard Corpus 1803-2005. Available online at http://www.hansard-corpus.org.

Alexander, Marc, Dallachy, Fraser, 2019. Historic Hansard 1805-2005: two centuries of speech representation. Conference Presentation in the Workshop "Big Data and the Study of Language and Culture: Parliamentary Discourse across Time and Space" at ICAME40. Neuchatel, Switzerland, 1-5. June, 2019.

Archer, Dawn, 2017. Mapping hansard impression management strategies through time and space. Stud. Neophilol. 89 (1), 5–20. https://doi.org/10.1080/00393274.2017.1370981.

Archer, Dawn, 2018. Negotiating difference in political contexts: an exploration of Hansard. Lang. Sci. 68, 22–41.

Aspinall, A., 1956. The reporting and publishing of the House of Commons' debates 1771–1834. In: Pares, Richard, Taylor, A.J.P. (Eds.), Essays Presented to Sir Lewis Namier. Macmillian & Co Ltd, London, pp. 227–257.

Biber, Douglas, 2009. A corpus-driven approach to formulaic language in English. Multi-word patterns in speech and writing. Int. J. Corpus Linguist. 14 (3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib.

Biber, Douglas, Finegan, Edward, 1989. Drift and the evolution of English style: a history of three genres. Language 65 (3), 487–517.

Biber, Douglas, Gray, Bethany, 2012. The competing demands of popularization vs. economy. Written language in the age of mass literacy. In: Nevalainen, Terttu, Closs Traugott, Elizabeth (Eds.), The Oxford Handbook of the History of English. Oxford University Press, Oxford, pp. 314–328.

Biber, Douglas, Gray, Bethany, 2013. Being specific about historical change: the influence of sub-register. J. Engl. Linguist. 41 (2), 103–134. https://doi.org/10.1177/0075424212472509.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan, Finegan, Edward, 1999. Longman Grammar of Spoken and Written English. Longman, London.

Blaxill, Luke, Beelen, Kaspar, 2016. A feminized language of democracy? The representation of women at Westminster since 1945. Twentieth Century Br. Hist. 27 (3), 412–449. https://doi.org/10.1093/tcbh/hww028.

Cameron, Deborah, 2003. Working with Spoken Discourse. Sage, London.

Chilton, Paul, 2004. Analysing Political Discourse: Theory and Practice. Routledge, London.

Collins, Peter, Yao, Xinyue, 2013. Colloquial features in World Englishes. Int. J. Corpus Linguist. 18 (4), 479–505.

Desagulier, Guillaume, 2017. Corpus Linguistics and Statistics with R: Introduction to Quantitative Methods in Linguistics. Springer, New York.

De Smet, Hendrik, 2016. How gradual change progresses: the interaction between convention and innovation. Lang. Var. Change 28, 83–102.

Eagles, Robin, Rix, Kathryn, 2015, December 1. The 'Story of Parliament': Parliament and the Press [Blog Post]. Retrieved from: https://thehistoryofparliament.wordpress.com/2015/12/01/the-story-of-parliament-parliament-and-the-press/.

Farrelly, Michael, Seoane, Elena, 2012. Democratization. In: Nevalainen, Terttu, Closs Traugott, Elizabeth (Eds.), The Oxford Handbook of the History of English. Oxford University Press, Oxford, pp. 392–401.

Gray, Bethany, Biber, Douglas, 2015. Phraseology. In: Biber, Douglas, Reppen, Randi (Eds.), The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press, Cambridge, pp. 125–145.

Hiltunen, T., Loureiro Porto, L., 2020. (Accepted/In press). Democratization of Englishes: Synchronic and diachronic approaches. Lang. Sci. https://doi.org/10.1016/j.langsci.2020.101275.

Hou, Liwen, Smith, David A., 2018. Modeling the decline of English passivization. Proceedings of the Society for Computation in Linguistics (SCiL) 1 (5), 34–43.

Hundt, Marianne, Mair, Christian, 1999. "Agile" and "uptight" genres: the corpus-based approach to language change in progress. Int. J. Corpus Linguist. 4 (2), 221–242.

Jordan, H. Donaldson, 1931. The reports of parliamentary debates, 1803-1908. Economica 34, 437–449.

Kruger, Haidee, Smith, Adam, 2018. Colloquialization versus densification in Australian English: a multidimensional analysis of the Australian diachronic hansard corpus (ADHC). Aust. J. Ling. 38 (3), 293–328. https://doi.org/10.1080/07268602.2018.1470452.

Leech, Geoffrey, Hundt, Marianne, Mair, Christian, Smith, Nicholas, 2009. Change in Contemporary English. A Grammatical Study. Cambridge University Press, Cambridge.

Leech, Geoffrey, Smith, Nicholas, 2009. Change and constancy in linguistic change: how grammatical usage in written English evolved in the period 1931–1991. In: Renouf, Antoinette, Kehoe, Andrew (Eds.), Corpus Linguistics: Refinements and Reassessments. Rodopi, Amsterdam, pp. 171–200.

MacDonagh, Michael, 1913. The Reporters' Gallery. Hodder and Stoughton, London and New York.

Mair, Christian, 1997a. Parallel corpora. A real-time approach to the study of language change in progress. In: Ljung, Magnus (Ed.), Corpus-based Studies in English. Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora. Rodopi, Amsterdam, pp. 195–209.

Mair, Christian, 1997b. The spread of the going-to-future in written English. A corpus-based investigation into language change in progress. In: Hickey, Raymond, Puppel, Stanislav (Eds.), Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak on His 60th Birthday. Walter de Gruyter, Berlin, pp. 1537–1543.

Mair, Christian, Leech, Geoffrey, 2006. Current changes in English syntax. In: Aarts, Bas, McMahon, April (Eds.), The Handbook of English Linguistics. Blackwell Publishing Ltd., Oxford, pp. 318–342.

Maartens, Brendan, 2019. 'What the country wanted': the houses of parliament, the press and the origins of media management in Britain, c. 1780–1900. Public Relat. Rev. 45 (2), 227–235. https://doi.org/10.1016/j.pubrev.2018.03.005.

Mollin, Sandra, 2007. The Hansard Hazard. Gauging the accuracy of British parliamentary transcripts. Corpora 2 (2), 187–210.

Port, Michael Harry, 1990. The official record. Parliam. Hist. 9 (1), 175–183.

Rix, Kathryn, 2014. 'Whatever passed in parliament ought to be communicated to the public': reporting the proceedings of the reformed commons, 1833–55. Parliam. Hist. 33 (3), 453–474.

Rühlemann, Christoph, Hilpert, Martin, 2017. Colloquialization in journalistic writing: the case of inserts with a focus on *well*. J. Hist. Pragmat. 18 (1), 104–135. https://doi.org/10.1075/jhp.18.1.05ruh.

SAS/STAT User Guide, version 15.1. Available online at: <https://support.sas.com/en/software/sas-stat-support.html.

Schmidt, Drew, Heckendorf, Christian, 2017. "ngram: Fast N-Gram Tokenization." R Package Version 3.0.4. URL: https://cran.r-project.org/package=ngram.

Slembrouck, Stef, 1992. The parliamentary Hansard '"Verbatim"' report: the written construction of spoken discourse. Lang. Lit. 1, 101–119.

Smitterberg, Erik, 2008. The progressive and phrasal verbs: evidence of colloquialization in nineteenth-century English? In: Nevalainen, Terttu, Taavitsainen, Irma, Pahta, Päivi, Korhonen, Minna (Eds.), The Dynamics of Linguistic Variation: Corpus Evidence on English Past and Present. John Benjamins Publishing Company, Amsterdam, pp. 269–289.

Spirling, Arthur, 2016. Democratization and linguistic complexity: the effect of franchise extension on parliamentary discourse, 1832–1915. J. Politics 78 (1), 120–136.

Szmrecsanyi, Benedikt, 2016. About text frequencies in historical linguistics: Disentangling environmental and grammatical change. Corpus Linguistics and Corpus Linguist. Linguistic Theory 12 (1), 153–171. https://doi.org/10.1515/cllt-2015-0068.

Tognini-Bonelli, Elena, 2001. Corpus Linguistics at Work. John Benjamins Publishing Company, Amsterdam.

Tyrkkö, Jukka, Kopaczyk, Joanna, 2018. Present applications and future directions in pattern-driven approaches to corpus linguistics. In: Kopaczyk, Joanna, Tyrkkö, Jukka (Eds.), Applications of Pattern-Driven Methods in Corpus Linguistics, Studies in Corpus Linguistics 82. John Benjamins Publishing Company, Amsterdam, pp. 1–12.

Vice, John, Farrell, Stephen, 2017. The History of Hansard. House of Lords Hansard and the House of Lords Library, London.

Williams, Kevin, 2010. Read All about it. A History of the British Newspaper. Routledge, London.