Page 1

Draft paper for presentation at "IPP 2012: Big Data, Big Challenges?" Please do not cite without permission.
All correspondence to: jonathan.bright@eui.eu

The Dynamics of Parliamentary Discourse in the UK: 1936 – 2011

Draft paper for presentation at "IPP 2012: Big Data, Big Challenges?" Please do not cite without permission.
 jonathan.bright@eui.eu

Parliamentary discourse is one of the most important mechanisms through which democracy functions in the UK. Speeches made before parliament by its members (MPs) fulfill a wide variety of roles: they allow government ministers to present and defend new legislation; allow opposition MPs to debate the merits of such legislation; and they also allow any member to raise questions about the current functioning of government, or propose new actions and initiatives. As Ilie puts it, "political speech and action are tightly intertwined".[1]

A crucial element of parliamentary discourse is its public nature. Speeches before parliament can be witnessed by members of the public, and are also frequently broadcast on television and via radio. Furthermore (more significantly for the present article), these speeches are all transcribed into the official journal of the UK parliament (Hansard). The public nature of parliamentary discourse is fundamentally about ensuring democratic accountability. But its recorded nature in Hansard, and especially recent digitisations of the Hansard archives, also open up significant possibilities to study the way such discourse takes place on a large scale. While in the field of informatics there is increasing recognition of these possibilities,[2] in the area of political science these vast tranches of data have remained largely unexploited. Current literature on parliamentary discourse is generally qualitative in nature, consisting of small scale studies which prioritise depth over breadth and which often select debates to study simply because they appear interesting or relevant[3]. Hence we still know little about the overall functioning of parliamentary debate.

This article seeks to remedy this deficit. On the basis of a dataset consisting of around 740 million words spoken in the UK's House of Commons in the period 1936-2011, I analyze the way in which the dynamics of parliamentary discourse have changed over the past 75 years. Two main lines of investigation are pursued. Firstly, I seek to describe the general dynamics of parliamentary debate, and how they have changed over time, looking at both the quantity of interventions and the types of topic being debated. Then this overall picture is broken down with an analysis of the differences between speakers on the basis of their personal characteristics, in order to assess the extent to which different members are treated differently. Throughout the article, I will also discuss various challenges encountered when coding data on such a large scale, together with some strategies used to try and ameliorate these difficulties.

Debate in the House of Commons: An Overview

I will begin with a brief overview of the institutional rules governing interventions in the UK's parliament. Members of the UK House of Commons, known as Members of Parliament (MPs), are the only

---

[1] Ilie, Cornelia. 2010. 'Analytical perspectives on parliamentary and extra-parliamentary discourses'. In: *Journal of Pragmatics*, 42(4), 879-1172.

[2] See e.g. Marx, Maarten. 2009. 'Advanced Information Access to Parliamentary Debates'. In: *Journal of Digital Information*.

[3] See e.g. Ilie, Cornelia. 2010. 'Strategic uses of parliamentary forms of address: The case of the U.K. Parliament and the Swedish Riksdag'. In: *Journal of Pragmatics*, 42(4), 885-911.

people allowed to speak before the commons.[4] The amount of MPs present in the commons has changed infrequently throughout its long history: major changes in size usually only resulting from a change in makeup of the United Kingdom itself. The last such change occurred in 1921, when Ireland separated from the UK, which reduced the house in size to 621 MPs[5]. Between 1921 and 2010 the number of MPs has increased slowly in response to periodical reviews of constituency sizes, themselves occasioned by growth in population (especially in urban areas). By the 2010 general election the UK House of Commons' membership had risen to 650[6] (though legislation currently before the house envisages a reduction to 600 at the next election).

The "speaker" of the House of Commons, with the support of two deputies, is the major actor who determines who is allowed to speak and when[7]. During debates MPs who wish to speak will rise at the end of any speech, theoretically in an attempt to catch the eye of the speaker. In practice, most MPs who wish to speak will have already made this known to the speaker, who will have largely worked out a timetable of interventions before the debate takes place[8]. The speaker should aim to balance evenly distribute speaking opportunities between the different parties, with senior members, ministers and 'shadow' ministers (members of opposition parties who are assigned to cover particular portfolios) given priority.

In principle, the subject of debates is decided by the house of commons itself via the agreement of timetables and programme motions, though in practice the government, which generally commands a parliamentary majority, is usually able to determine what the topic will be and the amount of time spent debating each item. Time is also usually limited for debates: in a typical week, the commons will be in session on Monday and Tuesday afternoon, and most of the day on Wednesday and Thursday. This time is generally filled up completely, thought debate has occasionally been known to dry up before the time limit is reached, in which case it is suspended.[9]

Measuring the Dynamics of Debate

The data for this study was obtained from the website of the 'UK Parliament Parser', which is itself part of the parliamentary monitoring organization 'They Work For You'.[10] The parliament parser consists of a project to collect information from the various websites run by the UK's houses of parliament, through a process known as 'scraping' (which essentially refers to the creation of bespoke programs which can understand the specific characteristics of individual websites). As part of this project, they have collected transcripts of 'Hansard', the official record of debates which took place in parliament, for debates taking place in the UK's houses of commons for the period 1935-2012 (this studied has removed the first and last year in order to only work with years where complete information is available). These transcripts, which are stored in XML format, amount to approximately 12 gigabytes of data. The dataset for this study was

---

[4] In rare circumstances members of the House of Lords have also spoken there, though not in the time period covered by this dataset. See: http://www.publications.parliament.uk/pa/cm200708/cmselect/cmberr/1116/111602.htm
[5] McLean, Ian. 1995. Are Scotland and Wales Over-Represented in the House of Commons? The Political Quarterly, 66: 250–268
[6] See: http://www.parliament.uk/about/faqs/house-of-commons-faqs/members-faq-page2/
[7] See Armitage, Faith. 2010. 'The Speaker, Parliamentary Ceremonies and Power'. In: *Journal of Legislative Studies*, 16:3, 325-337.
[8] See http://www.parliament.uk/documents/education/online-resources/printed-resources/debates.pdf
[9] http://blogs.telegraph.co.uk/news/jameskirkup/100024175/house-of-commons-suspended-because-mps-cant-think-of-anything-to-talk-about/
[10] For the Parliament Parser see: http://ukparse.kforge.net/parlparse/, for They Work for You see: http://www.theyworkforyou.com/

created using several further computer programs, which analyzed this mass of XML and measured is various relevant features.

Figures one to four present a basic overview of the data in the dataset. In total it contains approximately 740 million words spoken debates during in the period 1936 – 2011. As figure one shows, these words are, broadly speaking, evenly distributed across time. Especially after 1950, the amount of words spoken per year has fluctuated between 8 and 12 million, rising slightly towards the end of the data. But in general the amount of debate has been more or less consistent. Broad consistency in the amount of words spoken is related to a rough consistency in the amount of speeches made, as shown in figure two. As might be expected, the number of speeches (defined as an uninterrupted period of speech by one particular individual) has fluctuated in the same way as the number of words has, remaining roughly between 50,000 and 70,000.
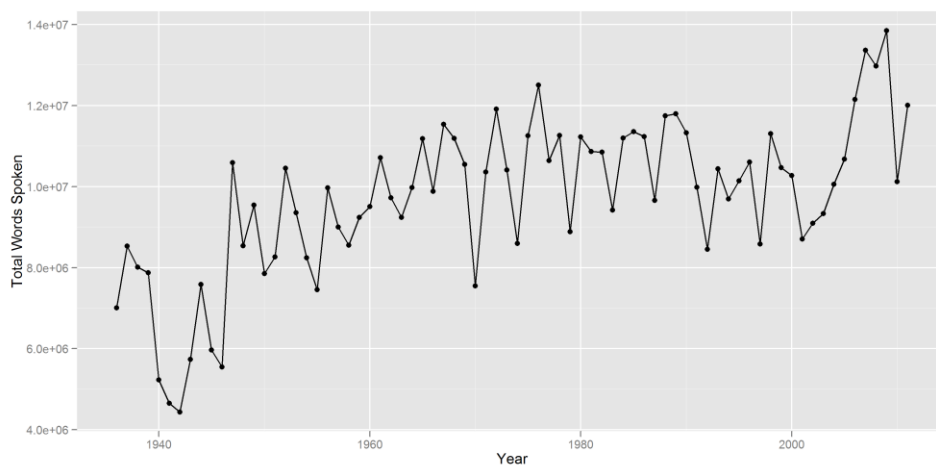
Figure 1: Total Words Spoken in the House of Commons
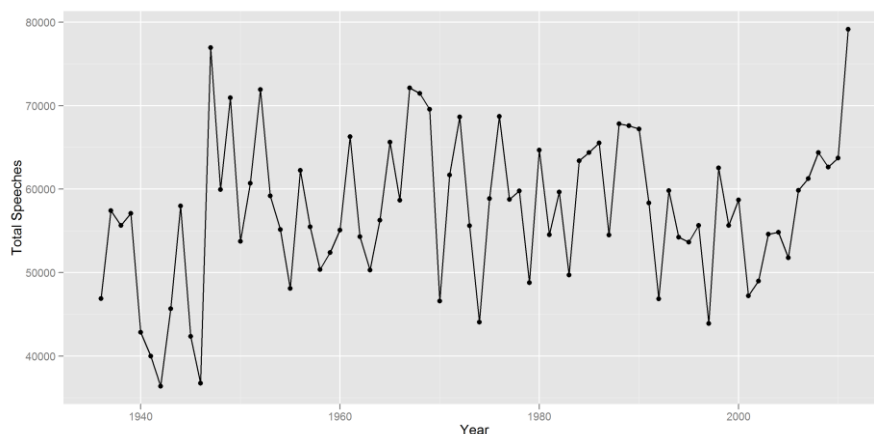
Figure 2: Total Speeches Made in the House of Commons

The number of speakers per year, which is shown in figure 3, proved to be a difficult variable to estimate. Overall there are records for around 10,700 different individual in the time period in question. However, numerous typographical errors in the way individual names are recorded in the data, coupled with different ways of referring to the same person (e.g. David Cameron, Mr. Cameron, The Prime Minister) meant that the same person was often split into multiple different records by the programme which generate the dataset. Two strategies were used to try and mitigate this problem. First, the data itself came

Page 4

Draft paper for presentation at "IPP 2012: Big Data, Big Challenges?" Please do not cite without permission. All correspondence to: jonathan.bright@eui.eu

with a number of other identifying variables created by the parliament parser project, allowing the accurate identification and disambiguation of approximately 3,000 speakers in the dataset. For the remaining 7,000 or so, a random sample of 200 was selected. 90 of these records were found to be duplicates of some variety. The total speakers for each year was then scaled to try and reflect this duplication. In particular, the total was reduced by 90/200 * X, where X is the amount of speakers for that year which did not have a pre-existing unique identifier.

Figure three displays the results of this coding. Two main conclusions can be drawn from it. Firstly, though relatively accurate, it still overestimates the number of speakers by at least 100 on any given year (the spikes in the graph are in the main election years where, of course, more than 650 MPs may speak during the course of the year as different individuals are replaced). This inaccuracy however diminishes over time, with the last 10 or so years in particular appearing to be almost entirely accurate. Secondly, despite these inaccuracies, we can still conclude with confidence that the vast majority of MPs, if not every single one, tend to speak at least once per year in parliament.
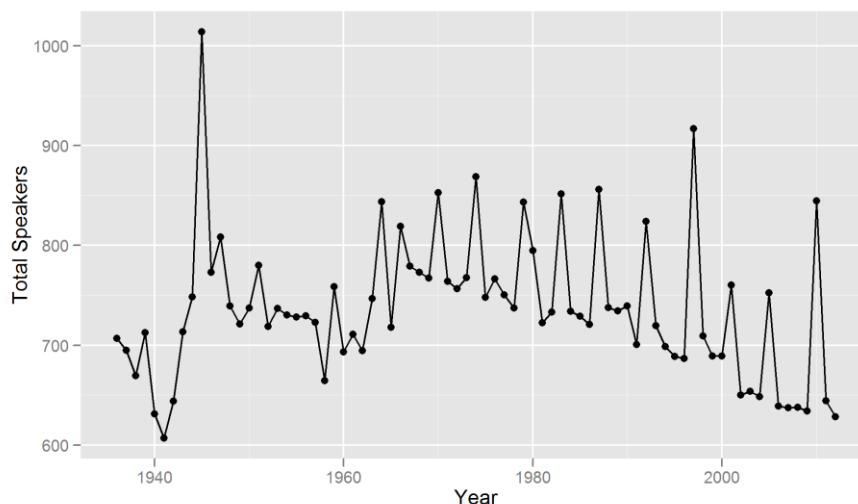

*Figure 3: Total Speakers in the House of Commons (scaled)*

One area in the data where there has been a relatively clear change is in the number of times speakers are interrupted in parliament, which is shown in figure four. Interruptions are caused by other members of parliament trying to shout down the current speaker when they disagree with what they are saying. This shouting, although theoretically not permitted, is a customary part of practice in the UK houses of parliament and typifies important debates in the commons.[11] Figure four shows that the amount of interruptions per year has approximately quadrupled, rising from less than 500 in 1936 to over 2,000 in 2011. Figure five scales this data to take into account the variation in words spoken per year. A broad separation can be made between the pre-1980 period, where the amount of words per interruption fluctuated quite widely, and the period after the appointment of Margaret Thatcher as Prime Minister, when the amount of words per interruption has both been much lower and much more consistent. From

---

[11] See Ilie, Cornelia, 2005b. Interruption patterns in British parliamentary debates and drama dialogue. In: Betten, A., Dannerer, M. (Eds.), Dialogue Analysis. IX. Dialogue in Literature and the Media. Selected Papers from the 9th IADA Conference, Salzburg 2003. Niemeyer, Tubingen, pp.415–430, and Ilie 2010, pp897-898

these data we can conclude that debate in the house of commons has been getting more contentious over time.
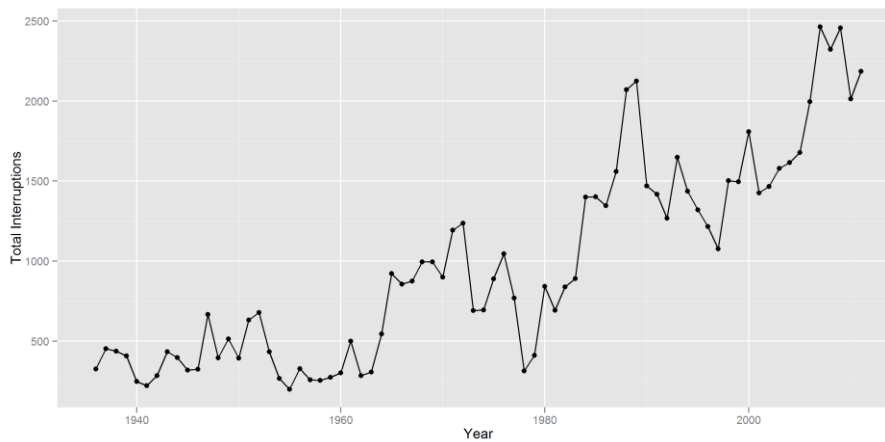
Figure 4: Interruptions per year in the House of Commons
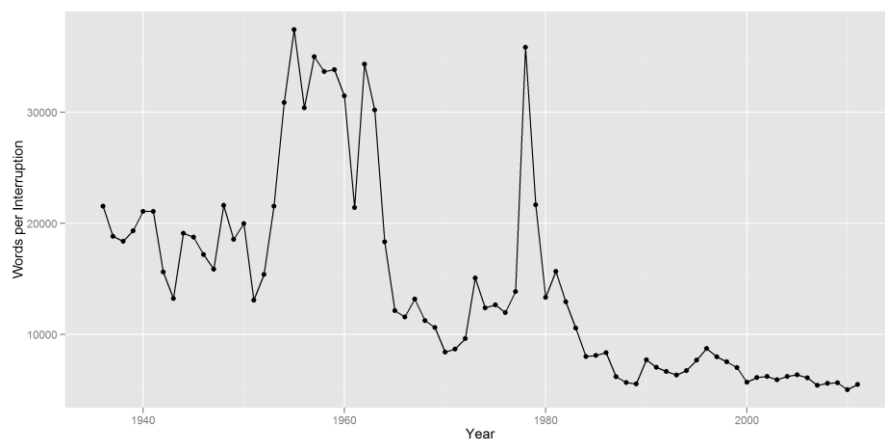
Figure 5: Average words spoken per interruption

I will now move on to look at the issues being discussed in the house of commons. The variety of different types of speaking opportunity within the house (from asking questions to debates on bills) mean that there is space for quite a wide range of issues to be discussed in any given year. Nevertheless, we might also expect that the house focuses on specific issues of importance at specific times, to the detriment of others, both because of wider changes in salience of particular topics and because of the agenda of the government, which is able to set the schedule for debate and hence control to an extent the topics discussed.[12]

For this part of the project, the data in the dataset was analysed to determine the frequency with which 10 different topics were discussed. These 10 topics were selected from the list developed by the

---

[12] This agenda has itself been tracked in Will Jennings, Shaun Bevan and Peter John. (2011). 'The British Government's Political Agenda: the Speech from the Throne, 1911-2008.' Political Studies, 59(1): 74-98

policy agendas project, which has developed a coding schema which aims to encompass most different types of policy discourse.[13] The coding is based on a simple count of the amount of individual speeches classified as referring to a particular topic (defined as an uninterrupted period of speech). The classification itself is achieved on the basis of keyword matching, with the keywords also drawn from the descriptions of different policy topics in the agendas project. While the internal validity of the topic coding was established by some manual checking, it is however difficult to establish the validity of the overall topic coding program (in particular, it is difficult to know whether differences in speech count represent genuine differences in attention of the topics before the house, or simply differences in the keywords chosen). It should be noted as well that many speeches remained unclassified at the end of the analysis. For this reason, my comments here are restricted to looking at changes in individual topics across time, rather than comparing the levels of different topics. Future research will involve more sophisticated topic classification based on machine learning.

Figure six tracks changes in the extent to which different topics are discussed in the house of commons. A number of conclusions can be drawn from this graphics. Some topics, such as healthcare, defence and education have remained relatively stable over time, year on year fluctuations taken into account. Others have shown notable trends. The only topic to have noticeably declined during the period in question is agriculture, which in the last 20 years (1991 – 2011) was discussed only about 20% as frequently as it was between 1935 and 1955. Several topics have grown in stature. From being almost ignored in the early part of the data, speeches relating to the environment increased 10 fold in the early 1970s, which probably reflects the start of the global environmental movement which took place in and around the 1972 UN Conference on the Human Environment. They remained at this new level until the early 1990s, when they started to decline, though they still remain way above the level of the 1950s and 1960s. The importance of 'Europe' as a topic of discussion doubled around this time, in the years leading up to in particular 1990, and has remained high ever since. This perhaps represents the renewed efforts at European integration which began with the Single European Act and the Maastricht treaty. Finally, references to the economy have increased markedly in importance in recent years, which presumably relates to the current period of financial crisis.

Two broad conclusions can be drawn from this analysis of the topics under discussion at the house of commons. Firstly, in the space of a year, the scope for discussion is large. Only the environment can genuinely claim to have received little attention, and then only in the period before 1970. Secondly, however, attention does change over time, and in particular does seem to be clearly linked to external events.

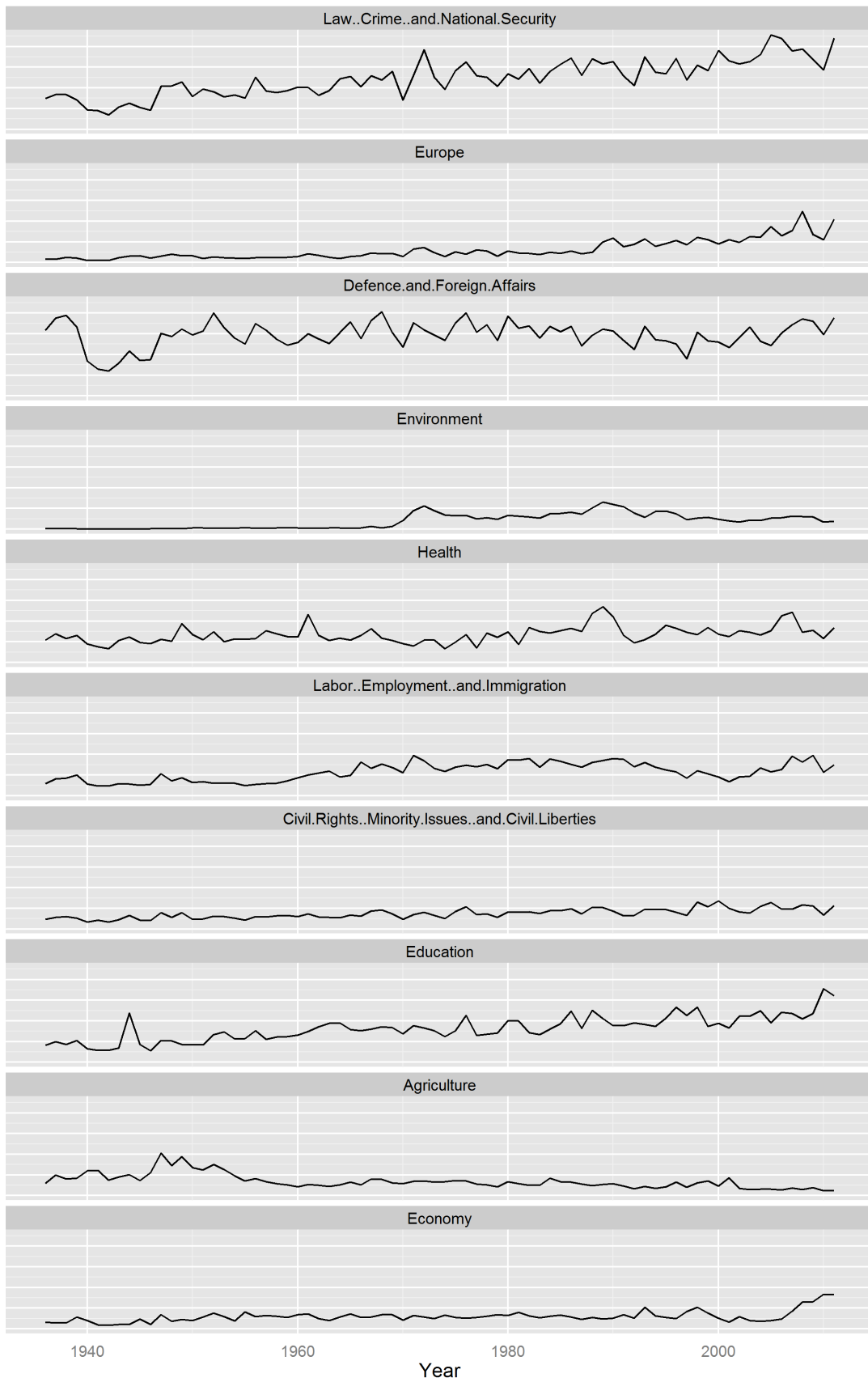[13] See: http://www.policyagendas.org/page/datasets-codebooks

Figure 6: Topics discussed in the commons

Diversity and Equality in the Commons

In this second section, I want to move on to look in more detail at the individuals speaking before the house of commons, analyzing the extent to which their experience of parliamentary debate and the length of their parliamentary career differs according to their personal characteristics, in particular their gender and whether they are a member of the nobility. Several authors have speculated that the particularly adversarial style of the British house of commons might lead to gender inequalities. For example, Shaw claims that it is "likely that the debating chamber will prove to be the setting for a speech event in which gender is a particularly salient variable".[14] As Marx notes, parliamentary informatics present a unique opportunity to measure the practical application of concepts such as gender equality and hence test this sort of claim.[15]

Automatic coding the gender of parliamentary speakers presented a further challenge for this project, which was tackled in two steps. First, gender specific titles (such as Mr., Lady and Marquess) were used where possible as a relatively unambiguous method of coding. Approximately half the dataset could be coded in such a fashion. Secondly, a dataset based on the US census was employed, which consisted of a list 6,000 or so name-gender pairings, based on the frequency which a given name is associated with a given gender.[16] This dataset was used in the first instance as a way of establishing a list of approximately 1,000 unique first names which were used in parliament during this period, relating to records where no gender specific title exists, and an 'estimated' gender based on the information in the US census. This much smaller list was then double checked by hand, with obviously incorrect assignments recoded (names which could be employed by either gender, such as George or Chris, were left as estimated).

In total this process left just 715 records without a coded gender (which is why the entries in the first column of table 1 do not sum to 10,733). The majority of these records contain name entries from which gender assignment is impossible (e.g. "the minister for health"). Coding of the entitled nobility was much more straightforward: anyone with a noble title such as 'Sir' or 'Lord' in their name was coded, though likely inaccuracies remain, as these titles are not always employed systematically.

	Number of Speakers	Average Speeches Made	Average Words Spoken	Average Words Per Interruption	Average Career Length
Male	9341	430	73027	10153	6.4
Female	683	574	89526	7831	10.9
Titled	1062	366	62260	8595	4.5
Non-Titled	9671	415	70021	10061	6.5
Total	**10733**	410	69253	9911	6.3

Table 1: Summary statistics for different speaker groups

---

[14] Shaw, Sylvia. 2000. 'Language, Gender and Floor Apportionment in Political Debates'. In: *Discourse and Society*, 11, 401-418. Citation from page 402.

[15] Marx, Maarten. 2010. 'Calculating the women friendliness of parliament'.

[16] The dataset used was adapted from Tigas, Panagiotis (2012) 'name2gender in Python'. Available from: http://ptigas.com/blog/2012/01/21/name2gender-in-python/

The results of these classifications are summarised in table one. As has already been outlined above, the absolute number of speakers suffers from measurement problems which means that it should be treated as an estimated figure, which is almost certainly inflated, as should other columns which depend on the amount of speakers. However, there is no reason to believe that these problems affect any of the sub-categories in this table disproportionately; hence valid inter-group comparisons can still be drawn.

The results for male and female MPs need first to be placed in the context of the very unequal distribution of seats in the house of commons during the time period covered in the dataset.[17] The first woman to be elected to the commons was the Countess de Markievicz, a member of Sinn Fein who was elected whilst in prison in 1918. Like all members of Sinn Fein, she declined to take up her seat, meaning that the first woman to actually sit in the Commons was Viscountess Nancy Astor, who was elected in 1919. By 1935 when the dataset for this study begins, the number of women had increased to 15. However, if the presence of women MPs is itself now almost 100 years old, it is only relatively recently that they have begun to form a significant part of the Commons: in 1983, almost 70 years after de Markievicz's selection, the number of women sat at only 23.

In 1992, the Labour party introduced a policy of 'all women' shortlists for 50% of what are known as 'inherited' seats (where a sitting MP either retires or passes away), which contributed to a significant rise in the amount of women parliamentarians in 1997[18]. Though this policy had been found to be unconstitutional in 1996, legislation passed in 2002 and 2008 made the practice legal; and all major parties have now substantially increased their female representation through similar types of selection policy. In 2010, 143 female MPs were elected to the commons: at less than a quarter of the house still way short of parity but also a rapid increase in a relatively short space of time.

Two major differences between genders is notable. Firstly, it appears that female MPs are interrupted significantly more frequently than their male counterparts. This appears to lend support to claims made by authors such as Shaw and Holmes[19] that women's experience of parliamentary discourse is different to that of men. However, it must be remembered that the house of commons as a whole has been getting more contentious over the last 30 or so years, which is when the vast majority of female MPs have occupied the house (see figure eight). Indeed, if we compare the average words per interruption only for the last 30 years they are much closer to parity, with if anything women being interrupted less frequently (7312 for women, 6550 for men). Figure seven shows how the amount of interruptions has fluctuated for male and female MPs across time, while figure nine shows the average interruptions per speaker. As can be seen, both in the early period of the dataset and after 1990, there appears to be little difference between the amount that men and women are interrupted. Only in the 1980s does it seem that women were interrupted significantly more frequently: a difference probably caused by the contentious premiership of Margaret Thatcher (the top two interrupted politicians in the dataset are, unsurprisingly, Britain's two longest serving prime ministers: Tony Blair and Margaret Thatcher). On the basis of this indicator, therefore, it seems that women and men are treated more or less equally in parliament.

---

[17] Lovenduski, Joni, and Pippa Norris. 2003. 'Westminster Women: the Politics of Presence', in: *Political Studies*, 51, 84-102.

[18] For a summary see Norris, Pippa. 2000. 'Breaking the Barriers: Positive Discrimination Policies for Women'. In: Jyette Clausen and Charles Maier, eds, *Has Liberalism Failed Women? Parity, Quotas and Political Representation.* New York: St Martin's Press.

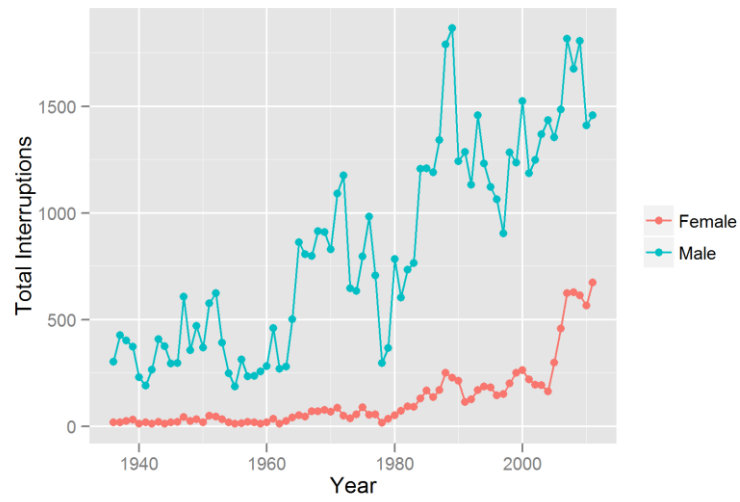[19] Shaw 2000, Holmes, J. 1995. *Women, Men and Politeness.* London: Longman.

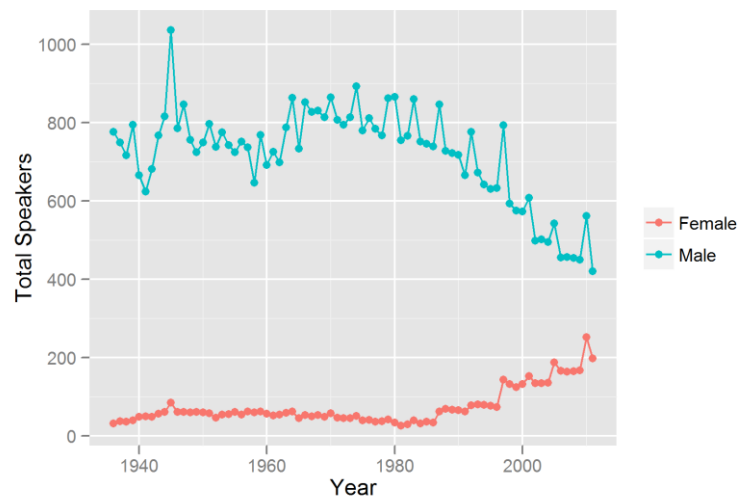*Figure 7: Total interruptions per year for male and female speakers*

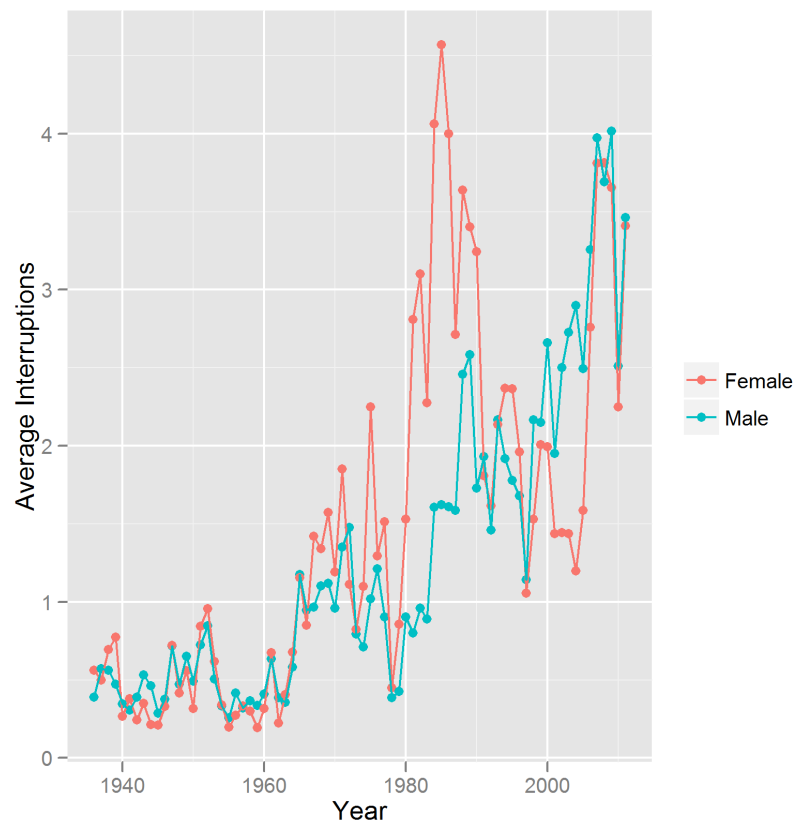*Figure 8: Total male and female speakers*

Figure 9: Average interruptions per speaker

The difference between career lengths is more difficult to explain. Career length as a whole is affected by the disambiguation problems mentioned above, with a tendency towards underestimation (as whenever a speaker is mistakenly divided into separate records their career length is also divided). We might hypothesise that, as these problems affect the front half of the database disproportionately, the male entries in the database which dominate this period are the most affected. However the way career length varies over time shows that, if anything, the reverse is true. As figure ten shows, average career length is overall on decline. The figures fluctuate quite widely in the dataset, which is partially a function of coding errors, and left and right censoring of course also plays a role. There is also a clear pattern of spikes around election years: understandably, MPs who start at an election are more likely to last for longer than MPs who enter in between elections, as they will have more time before the next election comes around. However even with these caveats in mind it is still possible to speak of a relatively clear trend towards a decline in average career lengths, especially between 1960 and 1980. All this means that the difference between male and female career lengths is especially surprising. It seems that, at a time when careers overall are shortening, women MPs entering parliament have nevertheless managed to stay in office significantly longer than their male counterparts.
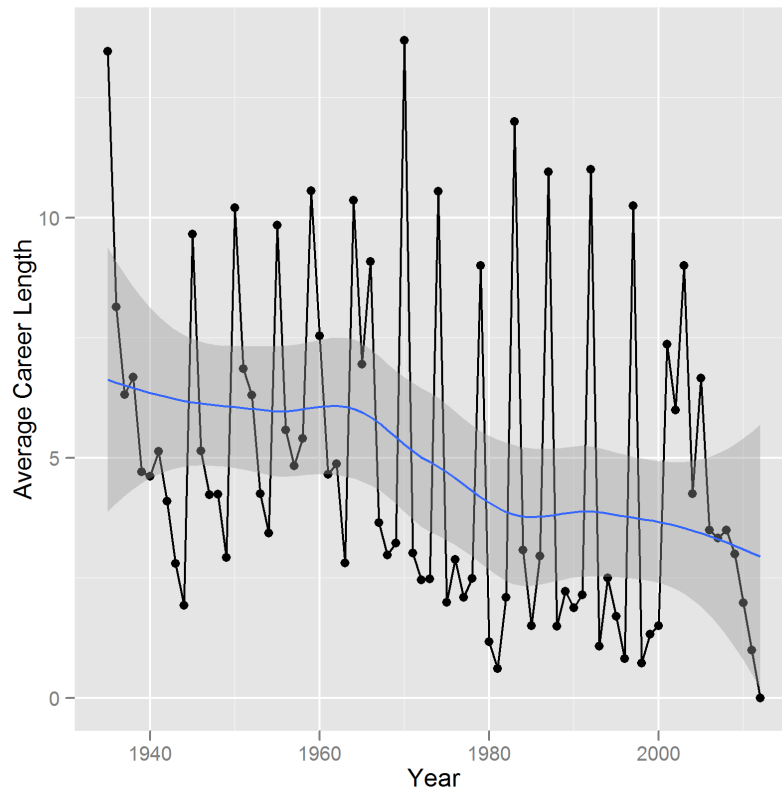
Figure 10: Average Career Length by Year, with trend line

The behaviour of those with titles and positions in the nobility presents a different picture. Their average words per interruption is almost the same as that of those without titles. However their average length of career is shorter: and this is reflected in both fewer speeches and fewer words spoken. It should be noted that this subgroup of MPs is also unevenly distributed across time. As women have been coming into parliament, those with titles have been going out at almost the same rate. As figure eleven shows, the majority of the entitled in the dataset began their careers in the period before 1980, with very few coming in afterwards. The impact of left censoring on this grouping (i.e., many of them would have began their career before 1936) probably has some bearing on their reduced career and word count figures, and certainly causes the large spike in 1936; they will also be more affected by the problems of disambiguation which are highest in the earlier part of the database. However it may be as well that the entitled also pursued different types of parliamentary career: in particular, they may have become members later in their life, or may have moved on more quickly to the house of lords.
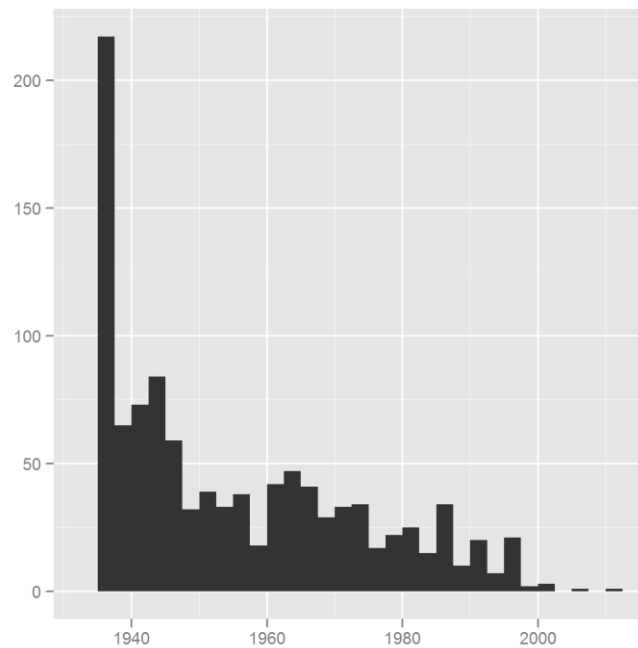
Figure 11: Distribution of Start Dates for Parliamentary Careers of the Entitled

Overall, the evidence presented in this section has not suggested major differences in the treatment of men and women or the titled and non-titled in terms of their access to debate. Though their career lengths may differ, the amount of times they speak and the amount of times they are interrupted remains comparable. This lends broad support to claims made by authors such as Celis and Wauters: that changing the makeup of parliamentary bodies does not have major effects on the way such bodies function (because, as they argue, new members are more likely to be socialised into existing behavioural norms than they are to change them).[20]

Conclusions

This article has attempted to tackle two major tasks. Firstly, it has highlighted some of the challenges involved in the automatic coding of political discourse data on a large scale. Minor typographical errors which could be easily corrected in smaller datasets can create major problems of disambiguation, leading to inflated estimations of speakers in the early half of the dataset. Gender recognition also proved a challenge: with approximately 10,000 speaker records, hand coding would have been unwieldy, and ways of automatic gender classification had to be developed. Various ways around these challenges have been discussed. In general, I can conclude that manual coding and checking remain a fundamental part of establishing accuracy in this type of large dataset: the challenge for the researcher is to find ways of keeping this time consuming task as small as possible whilst retaining validity.

---

[20] Karen Celis & Bram Wauters (2010): Pinning the Butterfly: Women, Blue-Collar and Ethnic Minority MPs vis-à-vis Parliamentary Norms and the Parliamentary Role of the Group Representative, The Journal of Legislative Studies, 16:3, 380-393

Secondly, and more importantly, it has highlighted both continuity and change in UK parliamentary discourse over a period of 85 years. Overall contributions to debate have remained roughly stable, with similar amounts of speakers and similar amounts of speeches. But within this continuity, there have also been some important changes. The tone of parliamentary debate, as measured by the amount of interruptions, has grown notably more hostile and contentious. The types of subjects being debated have shifted: the environment has become much more prominent, whilst topics such as agriculture have faded somewhat. Different subgroups within parliament, meanwhile, have different experiences, with titled MPs likely to have shorter than average careers, and female MPs likely to have longer ones.

Overall, the main aim has been to showcase the extraordinary potential parliamentary informatics has to improve our understanding of contemporary democratic systems, by allowing us to test and refine existing assumptions (such as the differing treatment of women in parliament) and make new discoveries, such as the clear impact of the 1972 UN Conference on the Human Environment. Further research and work in this area will likely contribute to making 'big data' driven approaches to parliamentary studies just as important as both existing quantitative work and in-depth qualitative studies.