Semantic Change in the Language of UK Parliamentary Debates

Gavin Abercrombie and **Riza Batista-Navarro**
School of Computer Science
University of Manchester
Kilburn Building, Manchester M13 9PL
gavin.abercrombie@postgrad.manchester.ac.uk
riza.batista@manchester.ac.uk

Abstract

We investigate changes in the meanings of words used in the UK Parliament across two different decade-long epochs. We use word embeddings to explore changes in the distribution of words of interest and uncover words that appear to have undergone semantic transformation in the intervening period. We explore different ways of obtaining target words for this purpose. We find that semantic changes are generally in line with those found in other corpora, and little evidence that parliamentary language is more static than general English. It also seems that words with senses that have been recorded in the dictionary as having fallen into disuse do not undergo semantic changes in this domain.

1 Introduction

Commonly known as *Hansard*, transcripts of debates held in the United Kingdom (UK) Parliament from 1802 to the present day are publicly and freely available in digitized format. These transcipts are important sources of historical and current information for many people including scholars in the political and social sciences, the media, politicians, and members of the public who wish to scrutinize the activities of elected representatives.

Natural languages (such as English) are known to be dynamic, with the meaning of many lexical items drifting over time (for example., *gay*: *cheerful* → *homosexual* (Wijaya and Yeniterzi, 2011)).

Knowledge of the level of such semantic change existing in a particular domain can assist in the design of systems for downstream natural language processing tasks such as sentiment analysis. For example, training and testing on in-domain data from different periods of time has been shown to negatively affect perfomance in named entity recognition (Fromreide et al., 2014) and sentiment analysis (Kapovciute-Dzikiene and Krupavicius, 2014). Successful analysis of such changes in Hansard could therefore be an important element in the development of civic technology applications for parliamentary analysis.

In this paper, we investigate to what extent diachronic semantic change occurs in the Hansard record by examining the contexts in which words appear during two different periods in the corpus.

2 Analysis

2.1 Data: The Hansard record

We collected the transcripts of debates in the House of Commons chamber from the parliamentary Hansard website[1] in `html` format and extracted the text elements that correspond to speaker utterances. These 'substantially verbatim'[2] transcripts are recorded by parliamentary reporters present at the debates.

Comparison across epochs Following Dubossarsky et al. (2017), we organised the transcripts into two decade-long epochs for comparison. We selected the periods 1909-18 and 2009-18 due to (a) the latter being the most recent period to comprise data from 10 complete years, and (b) the former consisting of transcripts from a full century prior, with the intervening period having seen a variety of significant changes, both in Parliament (for example, women's suffrage, the rise of the Labour party) and in wider society (two World Wars, the growth of technology). We considered the data for periods prior to the twentieth century to possibly be insufficiently complete for comparison with recent transcripts.[3]

Examination of the data available in the two

[1] https://hansard.parliament.uk

[2] https://hansard.parliament.uk/about?historic=false

[3] The Hansard record of debates from the 19th century includes only 92 days per year on average.

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 210–215
Florence, Italy, August 2, 2019.

periods (see Table 1) shows that, due to changes in Hansard transcription practice, the more recent epoch consists of a larger amount of data. Additionally, the large quantity of unique tokens (around 46 thousand items) that appear in only one of the epochs (the disjunctive union of the two sets) indicates that the vocabulary of the corpus changes considerably in this period.

	1909-18	2009-18	Total
Debate days	975	1455	2430
Utterances	448k	2.1M	2.5M
Tokens	33.7M	105.8M	124.9M
Unique tokens	95.6k	174.2k	222.8k

Table 1: Comparison of Hansard over two epochs. Each day's transcript typically includes several debates, which can be broken down into individual utterances (unbroken passages of text) and tokens (words).

2.2 Representation of the distributional space

In preproceesing, we stripped all utterances of punctuation, lowercased and tokenized them.

We extracted embedding vectors using `gensim`[4]'s `word2vec`[5] (Mikolov et al., 2013), with a context window of four tokens and vector dimensionality of size 300 (following settings used in previous work (Hamilton et al., 2016)).

As in Dubossarsky et al. (2017) and Hamilton et al. (2016), we trained word embeddings on each epoch, and aligned these using orthogonal procrustes transformation (Schönemann, 1966). We then compared word embedding vectors for each word of interest across the different time windows by calculating the cosine similarity of its embedding vectors in the two different periods. Assuming that lower similarity between these vectors indicates a higher degree of difference in the meaning and usage of a term, we use these calculations to identify which of these words has undergone semantic change in Hansard over time.

We calculate the cosine similarity between the word embedding vector of each word that appears in both epochs. The mean similarity across the entire vocabulary is only 0.154, indicating that the distributions of words in these two periods is quite different overall. We use this figure for comparison with our target words.

[4] Řehůřek and Sojka (2010).

[5] https://radimrehurek.com/gensim/models/word2vec.html

2.3 Target words of interest

We investigate instances of semantic change in the Hansard record from the two chosen epochs in four groups of lexical items: (1) words known from previous work to have undergone semantic change in the twentieth century; (2) words with senses that are no longer in use according to the Oxford English Dictionary (OED); (3) words from the parliamentary website's glossary;[6] and (4) words not appearing in the first three categories that demonstrate the greatest degree of distributional change across epochs. We consider words in the latter category to represent 'discovered' changes from this domain.

Known words from prior work Overall, it seems that words known to have undergone semantic change in English, have also done so in the Hansard record. Of the 21 known items (see Table 2), 18 have lower cosine similarity than the mean, suggesting that these semantic shifts also take place in Hansard. Observing the words with most similar embedding vectors in each epoch, we consider that 14 of these exhibit clear shifts in usage. The word with most dissimilar embedding vectors for the two periods is *checking*, which appears to undergo a similar shift in meaning as that described by Kulkarni et al. (2015) (see Figure 1).

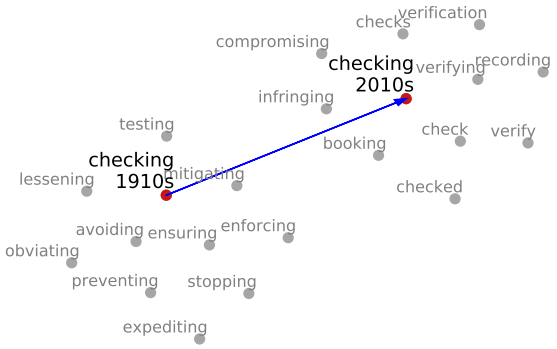

Figure 1: T-SNE visualisation (Maaten and Hinton, 2008) of the embedding space for 'checking' across both epochs, where it's sense appears to shift from *stopping* towards *verifying*.

While some items do appear to have undergone change in this data, this is not always of the form reported in the original literature. For example, while Hamilton et al. (2016) observe *broadcast* moving from being an agricultural term to the media and technology domain, in Hansard, it's earlier

[6] https://www.parliament.uk/site-information/glossary/

Word of interest	Earlier sense	Later sense	Source
actually	—	—	Hamilton et al. (2016)
broadcast $*$	cast out seeds	transmit signal	Hamilton et al. (2016)
calls	—	—	Hamilton et al. (2016)
check	—	—	Hamilton et al. (2016)
checking	stop doing	look at	Kulkarni et al. (2015)
diet	foods	weight-loss regime	Kulkarni et al. (2015)
gay $*$	happy	homosexual	Hamilton et al. (2016)
headed	top of body/entity	direction	Hamilton et al. (2016)
honey $*$	foodstuff	form of address	Kulkarni et al. (2015)
major	—	—	Hamilton et al. (2016)
monitor $*$	—	screen	Hamilton et al. (2016)
mouse $*$	rodent	device	Jatowt and Duh (2014)
plastic	flexible	synthetic polymer	Kulkarni et al. (2015)
propaganda	Papal committee	political information	Jatowt and Duh (2014)
record	—	album	Hamilton et al. (2016)
recording	set down in writing	stored copy	Kulkarni et al. (2015)
sex	biological gender	have intercourse	Kulkarni et al. (2015)
started	—	—	Hamilton et al. (2016)
starting	—	—	Hamilton et al. (2016)
transmitted $*$	pass	broadcast	Kulkarni et al. (2015)
wanting	lacking	wishing for	Hamilton et al. (2016)

Table 2: List of words of interest known to have undergone semantic change during the twentieth century, their sense shifts (if stated in the literature), and sources. Words we deem to have also undergone semantic change in Hansard are in bold. Those which appear to have shifted, but between different senses than those reported in the prior work, are marked with an asterisk ($*$). Note: Hamilton et al. (2016) compiled their original word list from Jatowt and Duh (2014); Jeffers and Lehiste (1979); Kulkarni et al. (2015); Simpson and Weiner (1989).

sense seems to be related to the *distributution* of printed material.

A number of observations seem to be artifacts of this particular dataset. A feature of the earlier epoch is that many of the MPs were ex-military officers, so in this period the most similar words to *major* are other rank titles such as *colonel* and *captain*, while this term later adopts the sense of *important* or *significant*. The word that appears to have changed the least according to vector similarity is *honey*. This is perhaps unsurprising, as the later sense recorded by Kulkarni et al. (2015) is both an informal term of address and associated with American English—and therefore unlikely to feature in UK parliamentary language. Given this, the fact that this item still has fairly low cosine similarity may be attributible to its frequent appearance as a surname in debates in the earlier epoch.

Disused words We obtained a list of words which have a least one sense that has fallen out of use and was last recorded by the OED between 1900 and the present day.[7] Of these, 39 appear in both epochs of Hansard. While we determine that most of these have not undergone semantic change in the corpus, even the three items that do seem to have shifted appear not to have been used in the disused sense listed in the dictionary (see Table 3).

Word of interest	Disused sense (OED)	1910s sense	2010s sense
slag	chain	coal bi-product	criticize
screen	banknote	barrier	electronic display
sky	enemy	? (unclear)	media organisation

Table 3: Words with a disused sense in the OED together with the senses in which thay are apparently used in the two epochs of the Hansard record.

[7]Downloaded from the API https://developer.oxforddictionaries.com/our-data#!/word/get_words.

↑	***racket***, ***levers***, ***balances***, *abet*, ***leans***, ***tailor***, *consensual*, ***implements***, ***riddle***, ***teen***, *invalidates*, *delivering*, ***honouring***, ***relay***, ***technological***, ***traverse***, *directs*, ***capitalise***, ***plurality***, ***disguised***
↓	*porcelain*, *whales*, *lesions*, *moat*, *professors*, *turnip*, *exceptionally*, *decreased*, *employ*, *suicides*, *insist*, *scaffold*, *assertions*, *daughters*, *murders*, *lasted*, *unfurnished*, *seeking*, *dams*, *fishes*

Table 4: Top 20 words that have undergone the most (↑) and least (↓) semantic change. Words the authors verify as clearly having undergone semantic change are in bold.

Parliamentary vocabulary Examination of the 51 single-word items in the parliamentary glossary reveals that only 56.9% of these have cross-epoch cosine similarity above the mean for the whole corpus, indicating that, as might perhaps be expected, these words have been semantically stable in Hansard through the last century. Among the most stable items are aristocratic titles such as *earl*, *bishop*, and *baron* that are used to refer to particular MPs and members of the House of Lords.

Discovered changes We examined the top 20 words from the whole vocabulary that are most and least similar according to cosine measurement across the two epochs of interest, excluding proper nouns, foreign words and numerals (see Table 4). Examining the words with most similar embedding vectors, we were able to confirm that most of the top changed words have indeed undergone semantic shifts, while none of those with the most similar embeddings across epochs appear to have done so. Examples, which may reflect societal changes between the epochs are *tailor* (*profession* $\rightarrow$ *adapt*) and *riddle* (*sieve* $\rightarrow$ *puzzle*).

3 Discussion

Words that have been shown to experience semantic change in English, in general seem to exhibit similar behaviour in parliamentary speeches. When compared to nearest neighbours from the distributional spaces in each epoch, it seems that the words that are least similar over time do indeed undergo semantic change in Hansard during this period. While it might be expected that words with senses specific to Parliament should not exhibit semantic change over time, they do not in fact seem to be much more stable than other items. This fact, combined with the overall low similarity across epochs for all words, may suggest that the differences in quantity and recording of the data in the two observed periods makes alignment of the word vectors problematic.

Additionally, words acknowleged to be in disuse in the OED tend to remain constant in this domain, and even those that do undergo change do not always seem to be used in the previously observed senses in this dataset. It would seem that, while these words had their last ever recorded uses in the period in question, they had already fallen out of use in Parliament.

In making the above observations, we acknowledge that it remains to be seen to what extent the observed changes are actually representative of diachronic change and how many of these are simply artifacts of the changing topics of dicussion in Parliament and the extent and manner of their recording in the Hansard record over the two epochs. We leave exploration of these questions for future work.

4 Related work

The phenomenon of language change has long been recognised (Sapir, 1921), and various social, cultural and cognitive factors have been proposed to explain it (Labov, 2011).

In recent years, efforts have been made to perform computational analyses of semantic change in dichronic corpora, and a number of methods have been proposed.[8] For example, (Wijaya and Yeniterzi, 2011) used topic modelling and clustering to investigate changes in the meanings of words in the Google Books corpus, while Frermann and Lapata (2016) proposed a Bayesian sense modelling approach to uncover gradual changes in meaning.

Much work in this area has focused on the use of vector space models such as Latent Semantic Analysis (LSA). Sagi et al. (2011) use this approach to track differences in the use of target words in historical texts, and Jatowt and Duh (2014) compare LSA with other distributional measurements.

A popular approach, which we adopt for this paper, is to use word embeddings. Kulkarni et al. (2015) compare distributional with

[8] For a recent overview, see Tang (2018).

frequency-based and syntactic analyses for diachronic change investigation, while Hamilton et al. (2016) and Dubossarsky et al. (2017) use embeddings to test hypotheses about the causes of such changes.

While Bamler and Mandt (2017) and Rudolph and Blei (2018) explore semantic change in the political domain on US State of the Union and Senate speeches repectively, we are unaware of any similar work on UK parliamentary debate transcripts.

5 Conclusion

We have explored four ways of obtaining target words of interest for diachronic semantic change analysis and conducted an initial study of this task in the domain of parliamentary debate transcripts. We found that using similarity measurement of word embedding vectors trained on two different epochs of the data, we are able to verify shifts in meaning in words that are known to have undergone this process in general English, and that we are also able to identify previously unknown changes in this data. We also observe that words with senses specific to parliamentary language do not appear to be particularly stable across time.

Future work will focus on conducting more comprehensive and systematic analyses of semantic change throughout the whole Hansard corpus in an effort to track senses and identify changes. We would also like to explore the possibility of using dynamic embeddings (e.g., Bamler and Mandt, 2017; Rudolph and Blei, 2018; Yao et al., 2018) to jointly train on different subsets of the data.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful and insightful comments.

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 380–389. JMLR.org.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark. Association for Computational Linguistics.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating NER for Twitter #drift. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Adam Jatowt and Kevin Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 229–238, Piscataway, NJ, USA. IEEE Press.

Robert Jeffers and Ilse Lehiste. 1979. *Principles and Methods for Historical Linguistics*. MIT Press.

Jurgita Kapovciute-Dzikiene and Algis Krupavicius. 2014. Predicting party group from the Lithuanian parliamentary speeches. *Information Technology And Control*, 43(3):321–332.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 625–635, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

William Labov. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 36. John Wiley & Sons.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 1003–1011, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2011. Tracing semantic change with latent semantic analysis. In Kathryn Allan and Justyna A. Robinson, editors, *Current methods in historical semantics*, pages 161–183. De Gruyter Mouton Berlin.

Edward Sapir. 1921. *Language: An Introduction to the study of speech*. NY: Harcourt, Brace & Co.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

John Simpson and Edmund Weiner. 1989. *The Oxford English Dictionary (20 Volume Set)*. Oxford University Press, USA.

Xuri Tang. 2018. A state-of-the-art of semantic change computation. *Natural Language Engineering*, 24(5):649–676.

Derry Tanti Wijaya and Reyyan Yeniterzi. 2011. Understanding semantic change of words over centuries. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, DETECT '11, pages 35–40, New York, NY, USA. ACM.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic word embeddings for evolving semantic discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, pages 673–681, New York, NY, USA. ACM.