Proceedings of ParlaCLARIN II Workshop, pages 23–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Unfinished Business: Construction and Maintenance of a Semantically Tagged Historical Parliamentary Corpus, UK Hansard from 1803 to the present day

Matthew Coole, Paul Rayson, John Mariani
Lancaster University
Lancaster, UK
{m.coole, p.rayson, j.mariani}@lancaster.ac.uk

Abstract

Creating, curating and maintaining modern political corpora is becoming an ever more involved task. As interest from various social bodies and the general public in political discourse grows so too does the need to enrich such datasets with metadata and linguistic annotations. Beyond this, such corpora must be easy to browse and search for linguists, social scientists, digital humanists and the general public. We present our efforts to compile a linguistically annotated and semantically tagged version of the Hansard corpus from 1803 right up to the present day. This involves combining multiple sources of documents and transcripts. We describe our toolchain for tagging; using several existing tools that provide tokenisation, part-of-speech tagging and semantic annotations. We also provide an overview of our bespoke web-based search interface built on LexiDB. In conclusion, we examine the completed corpus by looking at four case studies making use of semantic categories made available by our toolchain.

Keywords: Corpus, Construction, Hansard, Semantic Annotation

1. Introduction

Parliamentary discourse is of concern not only to political and linguistic scholars but also social charities and community groups. The transcriptions of speeches and discussions in the UK Houses of Lords and the Commons are better known as Hansard. Recent reports from these proceedings are freely available online. Historical transcriptions are available in the form of the Historical Hansard corpus which includes the transcriptions from 1803-2005. Previously the SAMUELS[1] (Semantic Annotation and Mark-Up For Enhancing Lexical Searches) project has researched tokenising and tagging this corpus (Wattam et al., 2014). As political engagement grows daily alongside the Hansard corpus transcripts, bridging this gap from 2005 to the present and maintaining an up to date, fully tokenised and tagged version of this dataset becomes increasingly relevant and important to improve search functionality and timeliness.

This paper presents the process, tools and output of our efforts to build a complete corpus of Hansard that contains linguistic and semantic annotations and runs right up to the present day. We also describe how this corpus is made available through a bespoke search interface built on top of the corpus database LexiDB. Through the use of our framework, the latest data from Hansard is continually downloaded, tagged and indexed in the database daily, meaning we have a live version of the parliamentary proceedings that is always up to date.

There are other forms of the Hansard corpus available online. Mark Davies provides access to the historical portion of Hansard up to 2005 through an online interface [2]. The primary advantage over this is our corpus has data from the latest parliamentary debates right up to the present day. The Hansard at Huddersfield[3] project has been updated to include all contributions up to 2019. However, whilst this data is presented in an attractive interface it is not linguistically tagged with semantic tags or POS (part-of-speech) tags making it more difficult to search for linguistic features based on such tags.

2. Background

In recent years, more and more nations have had their parliamentary discourse curated into a corpus format. This has inevitably involved various methods of cleaning the source data and transcriptions. Sometimes this is a simple task when the original data is consistently formatted using XML or another easy to interpret form. This may be simply mapping from one format to another such as the case of SLovParl 2.0 (Pančur et al., 2017) converting between HTML and XML. Sometimes the process may be more involved, such as parsing PDF source documents that may even be scans of the original handwritten paper transcripts.

Transcriptions for UK Hansard are made available online daily[4] and are available in XML format. Although easy to parse, the data within this XML format is not clean and is very sparsely documented with no consistent schema to process many aspects of the documents, particularly regarding the metadata. Previous work on the Parliamentary Discourse[5] and SAMUELS[6] projects had provided a cleaned-up version of the data prior to 2005. This put the data into a single text file per member contribution (speech or similar). The metadata for each contribution was then recorded in a separate TSV (tab-separated values) file containing information such as the member name, date of the contribution, current parliament sitting, house session etc. This made for

[1] https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/

[2] https://www.english-corpora.org/hansard/

[3] https://hansard.hud.ac.uk/site/index.php

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

[5] https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/parliamentarydiscourse/

[6] https://www.gla.ac.uk/schools/critical/research/fundedresearchprojects/samuels/

easier consumption of the source texts for linguistic tagging and annotation whilst still retaining the metadata in a form that could be easily searched and cross-referenced.

Other efforts have been made to add semantic topics to British Parliamentary speeches. Research (Nanni et al., 2019) grouping all speeches from the houses into semantic topics based on the content of the contribution provides a means of searching within Hansard for speeches not only based on MP information but also based on the topic discussed. This work utilised the publication of the Hansard corpus from another initiative, TheyWorkForYou [7] (run by MySociety[8]) provides a version of the Hansard transcripts back to 1918 which is cleaned with disambiguated MPs names and affiliations allowing for easier searching of this metadata when compared to that provided in the Historic Hansard corpus from the SAMUELS project.

3. Data collection

The Historical Hansard corpus covers transcriptions from 1803 - 2005 in both the House of Lords and the House of Commons. This data is freely available online 4^0 to anyone who wishes to use it. Previously the historic portion of the Hansard corpus has been processed through Lancaster University's linguistic toolchain (described below). This historic section of the corpus consists of just under 1.7 billion words (when tokenised through CLAWS) in around 7.5 million files.

For post-2005 data, several sources are available. The parliamentary website provides Atom feeds to allow for daily transcripts to be downloaded, but their API is not particularly useful in retrieving individual speeches from specific dates. TheyWorkForYou provides a means of accessing raw scraped XML from speeches back to 1919 as well as an open-source parser for cleaning the source XML data. Using this as well as a script provided by the Hansard at Huddersfield project all missing data after 2005 was retrieved and added to the original historic data to create a complete corpus of Hansard from 1803 onwards. The additional data consisted of approximately 315 million additional tokens in 4,302 files bringing the total corpus to approximately two billion words (the modern data contains several member contributions per file as opposed to the historic data which was divided into a single member's contribution per file).

The post-2005 data was brought as close as possible to being in line with the format produced from the SAMUELS project for Historic Hansard. Each source XML file contains multiple contributions from a single sitting of one of the houses. Each contribution was split into a separate file (consistent with Historic Hansard). This created around 1.2 million additional files which brought the total number of files in the corpus up to around 8.8 million. Each contribution file is stored in the original plain text as a TXT file and a tagged version in TSV format. TSV files were used as opposed to other XML based formats such as TEI based ParlaClarin format[9] to remain consistent with the output of the SAMUELS project. From the source XML file metadata was extracted to produce a similar supplementary TSV as is available with Historic Hansard. From the source XML, the date, member name, current parliament and sitting were extracted. In addition to this, we also extracted the PimsId (Parliamentary Information Management System ID) as these allow us to link to the open parliament site [10] and can provide means of linking to resources on the semantic web.

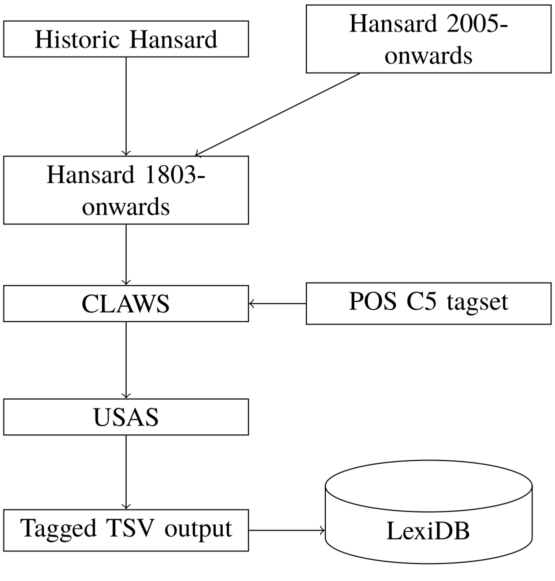


Figure 1: Annotation Processing pipeline

4. Tool Chain

4.1. Processing pipeline

4.1.1. CLAWS

CLAWS (Constituent Likelihood Automatic Word-tagging System) (Garside, 1987) is a part-of-speech (POS) tagger that also functions as a tokeniser. POS tagging is the most common form of linguistic annotation and CLAWS performs this operation on English text and has been used to tokenise and POS tag many different corpora in the past including the British National Corpus (BNC)[11](Leech et al., 1994). CLAWS outputs a vertical format where each line corresponds to a single token (the smallest meaningful unit of text) and includes a POS tag based on the C5 tagset[12]. This tagset consists of 62 tag codes e.g. NN1 (singular noun), PNX (reflexive noun) etc. CLAWS has an error-rate of only 1.5%, is the defacto standard for British corpora such as the BNC (Garside and Smith, 1997), Mark Davies' BYU English corpora and was the tagger used in the SAMUELS project.

4.1.2. USAS

USAS (UCREL Semantic Analysis System) (Rayson et al., 2004) semantically tags text using a semantic tagset[13] based on 21 main discourse fields. The major fields include categories such as; emotion, money & commerce, science & technology, food & farming etc. The tagset is tiered with

[7] http://parser.theyworkforyou.com/hansard.html

[8] https://www.mysociety.org/

[9] https://github.com/clarin-eric/parla-clarin

[10] https://api.parliament.uk/

[11] http://www.natcorp.ox.ac.uk/

[12] http://ucrel.lancs.ac.uk/claws5tags.html

[13] http://ucrel.lancs.ac.uk/usas/semtags.txt

each of these main 21 domains containing a number of subgroups[14]. In total there are 232 semantic tags. USAS can make use of CLAWS' vertical POS tagged output and produce output in various formats such as TSV. The English tagger is around 91% accurate, and it has been extended to multiple languages beyond English (Piao et al., 2016), and experiments are ongoing to incorporate neural and deep learning methods (Ezeani et al., 2019).

4.2. Corpus Interface

4.2.1. Overview

The data produced by the above pipeline is then indexed and stored in a LexiDB (Coole et al., 2016) instance. LexiDB is used as previous work (Coole et al., 2015) has shown other database technologies struggle to handle language corpora of the scale constructed here. LexiDB was specifically designed to handle corpus data in a way that allows it to both scale-out and be queried in a manner akin to other corpus data systems. The advantage of LexiDB is as further parliamentary data becomes available the database can easily be added to regularly, even as often as daily. This makes it feasible to run the processing pipeline whenever new data becomes available online[15] making for a truly "live", semantically tagged version of parliamentary debates available at all times.

4.2.2. Web Interface

A web interface[16] to the LexiDB instance hosting the compiled data was built to allow access to the full annotated corpus. This interface allows for several corpus queries to be run;

- Concordance search
- NGrams
- Word Lists
- Collocations
  - Log-likelihood
  - Mutual Information

Each of these query types has various options for filtering and sorting. Beyond this, a multitude of visualizations are available ranging from histograms for term occurrence over time to sunburst diagrams for exploring n-grams. Figure 4 shows the web interface.

The search bar allows for all of the annotation layers added to the data in the processing pipeline to be queried for. The query syntax takes the form of a regular expression over token stream and uses JSON query by example objects to represent tokens. A full in-depth guide to this syntax is available online[17]. The syntax will seem intuitive to corpus linguists and those already familiar with CQL (Corpus Query Language) used by CWB, CQPweb and SketchEngine, although the syntax differs from CQL in many ways, it is a result of combining JSON and regular expression syntax.

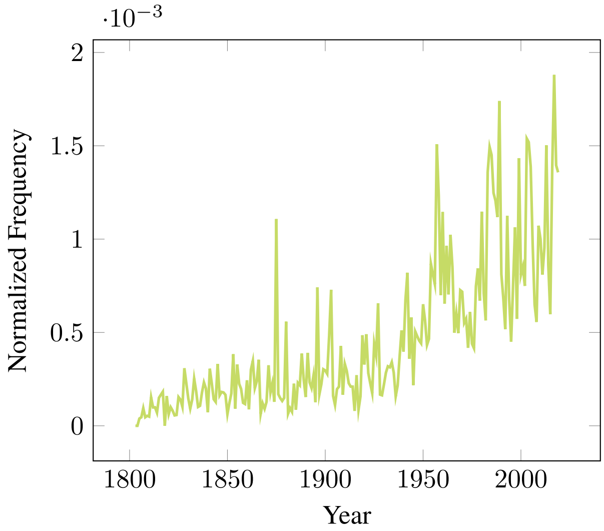


Figure 2: Semantic Category Y (Science & Technology) over time

5. Semantic Exploration

With the corpus complete and semantic tags available from 1803 onwards, we can examine various changes in the discourse based on the semantic categories available to us. In this vein, we look at four case studies examining the change in these semantic domains within both houses over time.

5.1. Science & Technology

The first semantic category examined is Science & Technology (Y*). This category includes two tags; Science & technology in general (Y1) and Information technology and computing (Y2). Figure 2 illustrates the change in this category over time. The plot is based on sub-sampling around 500 contributions per year and then the frequency is normalized as a proportion of all semantic tags that year. We can see a general trend that the discourse across both houses is becoming more and more frequently part of this semantic category. Digging beyond this we can observe how variations or spikes in the normalised frequency become bigger the later in the corpus we go. This could suggest Science & Technology can become hot topics of debate coinciding with major world incidence or advances in technology. Each of these spikes could be analysed in turn to examine what may have caused the discourse to shift towards this category at that time.

5.2. Numbers & Measurement

Numbers & Measurements (USAS tagged N*.*) contains various tags relating to maths and measurements; Mathematics (N2), Measurement: Distance (N3.3), Measurement: Area (N3.6) etc. Interestingly when comparing the semantic category of Science & Technology to that of Numbers & Measurement (Figure 3) we find that the trend of the normalised frequency of both generally increasing over time is true up until the late 20th century. Alarmingly at this point, the usage of numbers and measurements in both houses drops (proportionally to other semantic categories). One might expect as society moves towards greater scientific and technological understanding that there would be a continued increase in usage of specific measurements and statistics in

[14] http://ucrel.lancs.ac.uk/usas/Lancaster_visual/Frames_Lancaster.htm

[15] http://www.data.parliament.uk/

[16] http://ucrel-hansard-1.lancs.ac.uk/

[17] https://github.com/matthewcoole/lexidb/wiki/Query-Syntax

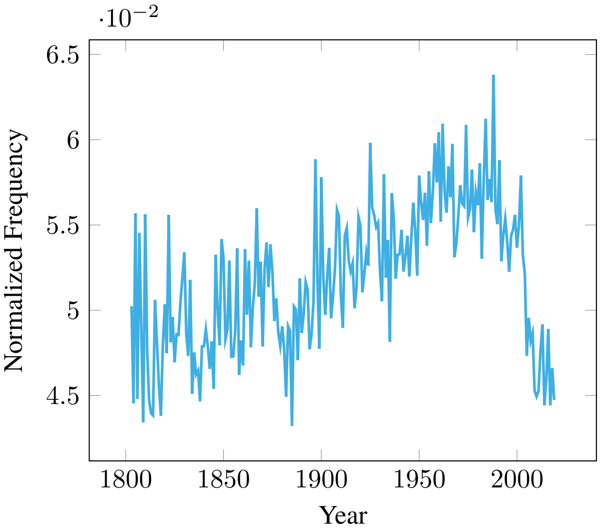


Figure 3: Semantic Category N (Numbers and Measurement) over time

discourse. This sudden decline entering the 21st century could indicate that politicians are becoming less precise when discussing issues and policies and not using precise figures, measurements or estimates and using vaguer language.

5.3. War & Warfare

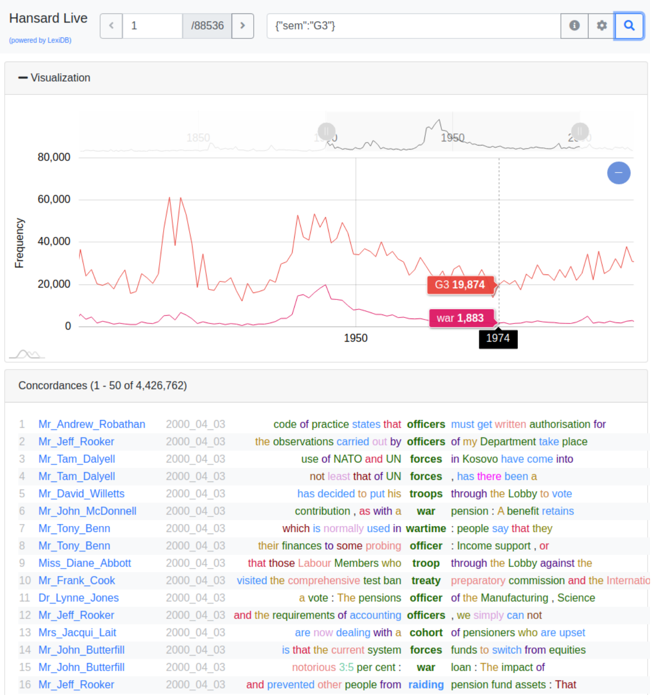
The semantic category G3 (Warfare, defence and the army; weapons) is shown in the interface's histogram visualization[18]. This category represents a wider semantic field compared to the word "war" and the trends between this term and the semantic category can be compared over time. As can be seen in Figure 4 both the semantic category and the term "war" have peaks around both world wars as would be expected. Interestingly though through the latter part of the 20th century the G3 semantic category noticeably rises to a greater extent than the term "war", this could be investigated further through a concordance analysis as to why politicians are speaking more about topics relating to war than specifically mentioning "war", or are using terms metaphorically. This is a good example of the kind of exploration of this corpus that is possible with our web interface.

5.4. COVID-19

Another visualization possible through our web interface is the ability to produce word clouds based on collocation metrics. Figure 5 shows a word cloud based on a general search for coronavirus; `(covid|COVID)-19|[Cc]oronavirus` and using the log-likelihood metric for collocations around this search term, the higher the metric the larger word appears. This includes the top 150 collocations of the search and is a good starting point for linguists to explore what is being said in this area before performing a more fine-grained analysis.

6. Conclusion

We have presented here our research to update the UK Hansard corpus. Our main contributions are: 1) a fully



Concordances (1 - 50 of 4,426,762)

1	Mr_Andrew_Robathan	2000_04_03	code of practice states that	officers	must get written authorisation for
2	Mr_Jeff_Rooker	2000_04_03	the observations carried out by	officers	of my Department take place
3	Mr_Tam_Dalyell	2000_04_03	use of NATO and UN	forces	in Kosovo have come into
4	Mr_Tam_Dalyell	2000_04_03	not least that of UN	forces	, has there been a
5	Mr_David_Willetts	2000_04_03	has decided to put his	troops	through the Lobby to vote
6	Mr_John_McDonnell	2000_04_03	contribution , as with a	war	pension : A benefit retains
7	Mr_Tony_Benn	2000_04_03	which is normally used in	wartime	: people say that they
8	Mr_Tony_Benn	2000_04_03	their finances to some probing	officer	: Income support , or
9	Miss_Diane_Abbott	2000_04_03	that those Labour Members who	troop	through the Lobby against the
10	Mr_Frank_Cook	2000_04_03	visited the comprehensive test ban	treaty	preparatory commission and the Internatio
11	Dr_Lynne_Jones	2000_04_03	a vote : The pensions	officer	of the Manufacturing , Science
12	Mr_Jeff_Rooker	2000_04_03	and the requirements of accounting	officers	, we simply can not
13	Mrs_Jacqui_Lait	2000_04_03	are now dealing with a	cohort	of pensioners who are upset
14	Mr_John_Butterfill	2000_04_03	is that the current system	forces	funds to switch from equities
15	Mr_John_Butterfill	2000_04_03	notorious 3:5 per cent :	war	loan : The impact of
16	Mr_Jeff_Rooker	2000_04_03	and prevented other people from	raiding	pension fund assets : That

Figure 4: Web Interface showing War & G3 Semantic Tag searches

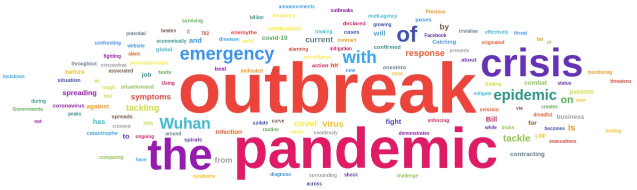


Figure 5: COVID-19 Wordcloud

tagged version of the Hansard corpus from 1803 up to the present data that is tokenised, POS tagged and semantically annotated 2) an NLP framework for annotation that can be fully automated and will be used going forward to keep our linguistically annotated version of the Hansard corpus up to date with new data as transcriptions from parliament become available daily 3) a search interface that allows for linguistic queries to be performed against the corpus as well as providing visualizations to better understand changes in political discourse over time. We will also make the entire semantically tagged corpus available for download through a link on our web interface page[19], observing the same licences as for the untagged data.

7. Acknowledgements

This research conducted by the lead author is funded at Lancaster University by an EPSRC Doctoral Training Grant. The SAMUELS project was funded by the Arts and Humanities Research Council in conjunction with the Economic and Social Research Council (grant reference AH/L010062/1) from January 2014 to April 2015.

[18] `https://www.amcharts.com/`

[19] `http://ucrel.hansard-1.lancs.ac.uk`

8. Bibliographical References

Coole, M., Rayson, P., and Mariani, J. (2015). Scaling out for extreme scale corpus data. In 2015 IEEE International Conference on Big Data, pages 1643–1649. IEEE.

Coole, M., Rayson, P., and Mariani, J. (2016). lexiDB: A scalable corpus database management system. In 2016 IEEE International Conference on Big Data (Big Data), pages 3880–3884. IEEE.

Ezeani, I., Piao, S., Neale, S., Rayson, P., and Knight, D. (2019). Leveraging pre-trained embeddings for welsh taggers. In Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 270–280, Florence, Italy, August. Association for Computational Linguistics.

Garside, R. and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 102–121.

Garside, R. (1987). The CLAWS word-tagging system. *The Computational analysis of English: A corpus-based approach. London: Longman*, pages 30–41.

Leech, G., Garside, R., and Bryant, M. (1994). CLAWS4: the tagging of the British National Corpus. In COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics.

Nanni, F., Menini, S., Tonelli, S., and Ponzetto, S. P. (2019). Semantifying the UK Hansard (1918-2018). In 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), pages 412–413. IEEE.

Pančur, A., Šorn, M., and Erjavec, T. (2017). Slovenian parliamentary corpus SlovParl 2.0.

Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Křen, M., Löfberg, L., Nawab, R. M. A., Shafi, J., Teh, P. L., and Mudraya, O. (2016). Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 2614–2619, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Rayson, P., Archer, D., Piao, S., and McEnery, T. (2004). The UCREL semantic analysis system. In Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 7–12.

Wattam, S., Rayson, P., Alexander, M., and Anderson, J. (2014). Experiences with parallelisation of an existing NLP pipeline: Tagging hansard. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 4093–4096, Reykjavik, Iceland, May. European Language Resources Association (ELRA).