Parliamentary corpora

Treasure trove for the Digital Humanities
Challenge for infrastructure development

Christian Mair (University of Freiburg & CLARIN-D F2)

CLARIN-D
SPONSORED BY THE
Federal Ministry of Education and Research

1. What makes parliamentary data attractive for research?

- natural database for **interdisciplinary** research, bringing together **(historical) linguistics**, **history**, **political science** and **cultural studies** under the **Digital Humanities** umbrella
- **self-replenishing** source of data with potentially considerable **historical time-depth**
- potential for **comparative research** across national parliamentary traditions
- potential (as yet largely unexplored) for research on **multimodality**
- significant amounts of **multilingual** and **translation data** (e.g. Belgium, European Parliament)

2. Case studies from the Hansard Corpus (CLARIN-UK): "Distant Reading" (Moretti 2013) of 1.6 billion words, 1803-2005

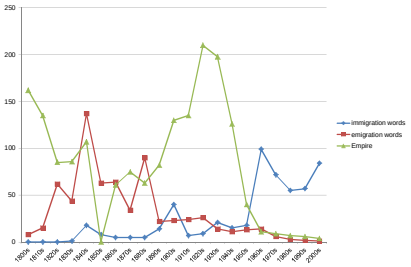


Fig. 1: Immigration, emigration, Empire (1803-2005, n/pmw)

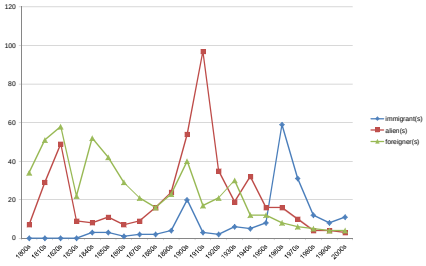


Fig. 2: Immigrants, aliens, foreigners (1803-2005, n/pmw)

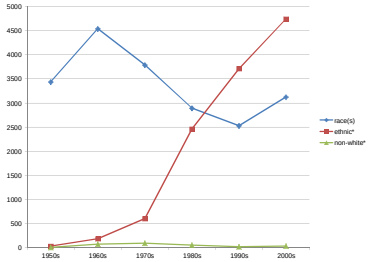
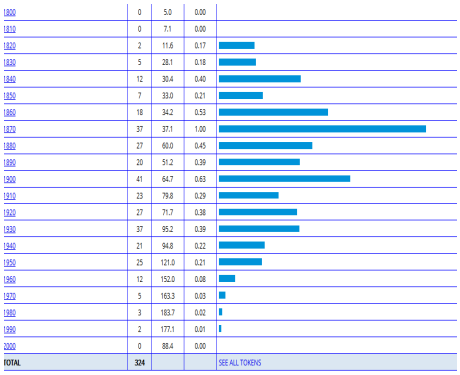
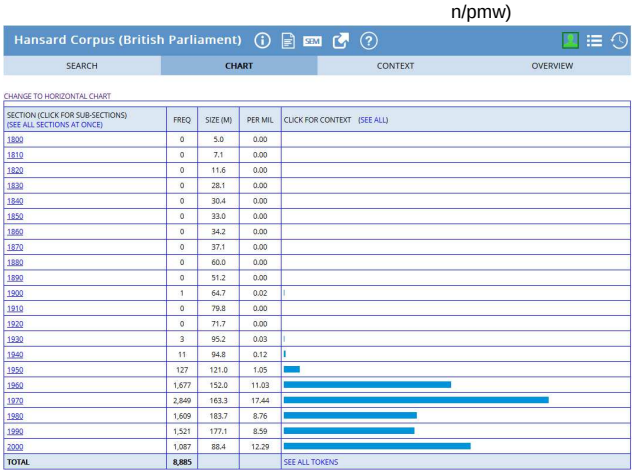


Fig. 3: From 'race' to 'ethnicity' (1951-2005, n/pmw)



1800	0	5.0	0.00	
1810	0	7.1	0.00	
1820	2	11.6	0.17	
1830	5	28.1	0.18	
1840	12	30.4	0.40	
1850	7	33.0	0.21	
1860	18	34.2	0.53	
1870	37	37.1	1.00	
1880	27	60.0	0.45	
1890	20	51.2	0.39	
1900	41	64.7	0.63	
1910	23	79.8	0.29	
1920	27	71.7	0.38	
1930	37	95.2	0.39	
1940	21	94.8	0.22	
1950	25	121.0	0.21	
1960	12	152.0	0.08	
1970	5	163.3	0.03	
1980	3	183.7	0.02	
1990	2	177.1	0.01	
2000	0	88.4	0.00	
TOTAL	324			SEE ALL TOKENS

↑**Fig. 4:** dominant|inferior|superior|backward|advanced RACE_n* (1803-2005)



Hansard Corpus (British Parliament)

SEARCH | CHART | CONTEXT | OVERVIEW

CHANGE TO HORIZONTAL CHART

SECTION (CLICK FOR SUB-SECTIONS) (SEE ALL SECTIONS AT ONCE)	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT (SEE ALL)
1800	0	5.0	0.00	
1810	0	7.1	0.00	
1820	0	11.6	0.00	
1830	0	28.1	0.00	
1840	0	30.4	0.00	
1850	0	33.0	0.00	
1860	0	34.2	0.00	
1870	0	37.1	0.00	
1880	0	60.0	0.00	
1890	0	51.2	0.00	
1900	1	64.7	0.02	
1910	0	79.8	0.00	
1920	0	71.7	0.00	
1930	3	95.2	0.03	
1940	11	94.8	0.12	
1950	127	121.0	1.05	
1960	1,677	152.0	11.03	
1970	2,849	163.3	17.44	
1980	1,609	183.7	8.76	
1990	1,521	177.1	8.59	
2000	1,087	88.4	12.29	
TOTAL	8,885			SEE ALL TOKENS

→**Fig. 5:** race relations|discrimination|equality (1900-2005)

3. Challenges for infrastructure development in the European CLARIN context

- existing parliamentary corpora (cf. Fišer & Lenardič 2017) extremely variable in size and historical time depth: e.g. Dutchparl 800m, 1814-2014; PTPARL 1m, 1970-2008
- enormous differences in type and depth of annotation

Perspectives for the future:

Given CLARIN's helpful but generally modest role in the various national projects, comprehensive standardisation of data and tools on CLARIN terms is unrealistic. What will work is a **one-stop easy-mode web interface**, hosted by CLARIN and providing access to European parliamentary corpora on the following terms:

- easy-to-use full-text search (with POS tagging, collocational statistics, and named entity recognition) for all corpora
- standardisation across corpora and tools, to attract large numbers of novice users across the disciplines
- long-term and systematic support to help regular users develop into a visible international research community.

4. Traditional "close reading"

1848: Other plans had been tried: European colonists had been sent out, but **European immigration** had, to a great extent, failed: **Coolie immigration** had been tried, and that had likewise to a considerable degree failed […] (context: post-Abolition West Indies)

1905: […] they must bear in mind that some of the undoubted evils which had fallen upon portions of the country from an **alien immigration** which was largely **Jewish**, gave those of them who, like the right hon: Baronet and himself, condemned nothing more strongly than the manifestation of the **anti-Semitic** spirit, some reason to fear that this country might be, at however great a distance, in danger of following the evil example set by some other countries […]

1967: The hon: Member for Barnsley (Mr: Mason), the Minister of Defence for Equipment, who is grinning at this, should go back to Barnsley and tell his voters that his policy is to have more **black people** in this country than **white** and see what they do to him: The Home Office says that by 1985 there will be 3½ million **coloured people** in this country: It should know more about it than the hon: Gentleman does: By doing nothing, the Government are betraying this country and compelling the English [to] **commit race suicide**: […] This is true; it is no good hon: Members tut-tutting: For all these reasons I find the Address bitterly disappointing and the Government a cowardly failure (Commons, Osborne, Labour)

Refererences

Alexander, Marc, & Andrew Struan. 2017. Digital Hansard: The politics of the uncivil. In *Digital Humanities* 2017, Montréal, Conference Abstracts. https://dh2017.adho.org/abstracts/DH2017-abstracts.pdf.

Baker, Helen, Vaclav Brezina, and Tony McEnery. 2017. Ireland in British parliamentary debates, 1803-2005. In Tanja Säily, Arja Nurmi, Minna Palander-Collin and Anita Auer, eds. *Exploring future paths for historical sociolinguistics*. Amsterdam: Benjamins. 83-108.

Fišer, Darja, and Jakob Lenardič. 2017. Parliamentary corpora in the CLARIN infrastructure. *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18 - 20 September 2017*. Conference Proceedings published by Linköping University Electronic Press at www.ep.liu.se/ecp/contents.asp?issue=147.

Gabrielatos, Costas, Tony McEnery, Peter Diggle, and Paul Baker. 2012. The peaks and troughs of corpus-based contextual analysis. *International Journal of Corpus Linguistics* 17: 151-175.

Mair, Christian. 2019. Brexitiness: The Ebbs and Flows of British Eurosceptic Rhetoric since 1945. *Open Library of Humanities*, 5(1): 50, 1-26. DOI: https://doi.org/10.16995/olh.424

Mollin, Sandra. 2007. The Hansard hazard: Gauging the accuracy of British parliamentary transcripts. *Corpora* 2: 187-210.

Moretti, Franco. 2013. *Distant reading*. London: Verso.