Towards a balanced corpus of recent political speeches

Mittwoch,
04.03.2020
15:45–16:30
ESA1 W Foyer

Adrien Barbaresi
Berlin-Brandenburg Academy of Sciences, Center for Digital Lexicography (BBAW-ZDL)
barbaresi@bbaw.de

CLP

Speeches are widely researched upon in political science as they can be used as a data source to reveal important information about the policy positions of their authors (Laver et al. 2003). Improving coverage for this text type can be of particular interest for corpus and computational linguistics. Words, themes and phraseology found in speeches are likely to complement other contemporary texts. Additionally, their comparatively open copyright status makes them highly relevant for research, for instance to perform replication studies. Clean categorized data is necessary for most approaches and potential use cases so that a peculiar scrutiny is required for corpus construction (Graën et al. 2014).

Previous versions of the corpus – first presented at this poster session (Barbaresi 2012) – have already been used in various scientific publications and in different disciplinary contexts (Barbaresi 2018). Three main approaches can be distinguished overall: qualitative analysis, mostly in history and political science; quantitative uses, mostly in machine translation; and integration into reference corpora and corpus linguistics tools. The current focus lies on significant political figures from German-speaking countries and regions from 1990 to 2020. Corpus development strives towards a balance in terms of linguistic and political diversity by taking into account population data as well as party-based distinctions (e.g. leaders of the opposition in Parliament) and institutional roles (for example the presidents of the Austrian National Council). From a linguistic standpoint, the form and content of the speeches can be expected to diverge according to the conditions in which they were held and to the characteristics of the speaker. There are also qualitative differences from a political standpoint, as some speeches are part of the daily routine of state bodies whereas others are considered to be important because of a particular political situation or institutional relevance. In order to reflect and operate on these factors of diversity, the corpus now features enhanced metadata including the country, the role and the affiliation of each speaker, which makes it possible to build a specialized subcorpus or to target certain speeches using a faceted search.

The corpus is publicly available for download (https://purl.org/corpus/german-speeches) and for querying on the DWDS infrastructure (https://www.dwds.de/d/k-spezial#politische_reden). The poster will present recent changes (most notably thousands of new speeches for more diversity), feature examples of use, and discuss work on further completion and balancing.

References: Barbaresi A. (2012). German Political Speeches – Corpus and Visualization, DGfS-CL postersession, Frankfurt. Barbaresi, A. (2018). A corpus of German political speeches from the 21st century. Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018), ELRA, 792–797. Graën, J., Batinić, D. & Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. Proceedings of KONVENS 2014, University of Hildesheim, 222–227. Laver, M., Benoit, K. & Garry, J. (2003). Extracting policy positions from political texts using words as data. American Political Science Review, 97(2), 311–331.