A corpus of German political speeches from the 21st century
Adrien Barbaresi

To cite this version:
Adrien Barbaresi. A corpus of German political speeches from the 21st century. 11th Language Resources and Evaluation Conference (LREC 2018), May 2018, Miyazaki, Japan. pp.792-797. hal-01798703

HAL Id: hal-01798703
https://hal.archives-ouvertes.fr/hal-01798703
Submitted on 23 May 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial License 4.0 International License
A corpus of German political speeches from the 21st century

Adrien Barbaresi
Berlin-Brandenburg Academy of Sciences
Jägerstr. 22/23 – 10117 Berlin – Germany
barbaresi@bbaw.de

Abstract
The present German political speeches corpus follows from an initial release which has been used in various research contexts. This article documents an updated and extended version: as 2017 marks the end of a legislative period, the corpus now includes the four highest ranked functions on federal state level. Besides providing a citable reference for this resource, the main contributions are (1) an extensive description of the corpus to be released and (2) the description of an interface to navigate through the texts, designed for researchers beyond the corpus and computational linguistics communities as well as for the general public. The corpus can be considered to be from the 21st century since most speeches have been written after 2001 and also because it includes a visualization interface providing synoptic overviews ordered chronologically, by speaker or by keyword as well as consequent accesses to the texts.

Keywords: Web corpus construction, Visualization, Keyword extraction

1. Introduction

1.1. Context
In the Western world, political speeches held by state representatives are often the focus of people’s attention and they are endowed with a high symbolic value. Words, themes and phraseology found in speeches are likely to be found in a wide array of other texts as speeches either give an impulse or entail a reaction to events or trends. Speech writing as a genre follows a series of rhetorical rules which make it distinct from other written texts, while the reading in front of an audience also shapes their style toward proper speech. First and foremost, it is necessary to discriminate between political speeches and speeches held by religious dignitaries or other kinds of public addresses. Another distinction can be made between electoral speeches and speeches made by politicians vested with a state function. From a linguistic standpoint, a further difference regards the flow of the speech, as there are speeches which are meant to be read without interruption and others which may not be delivered fluently, for example in a parliament address with immediate reactions. Finally, there are genres marginally related to speeches such as press conferences and interviews. From a political standpoint, there are qualitative differences 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, for example after a new government has taken office.

A major goal of speech collections is to gather information on government work, since political texts are the concrete by-product of strategic political activity and have a widely recognized potential to reveal important information about the policy positions of their authors (Laver et al., 2003). Linguists are also particularly interested in building corpora and improving coverage for this particular genre, while its versatility paves the way towards the use as corpus (Guerini et al., 2008) and its inclusion into reference corpora. Last but not least, since speeches are held in public and part of official debate, they are very often not subject to copyright issues. Consequently, their copyright status makes them highly relevant for replication studies as well as a wide range of purposes. Experience shows that a peculiar scrutiny is required for corpus construction, since clean data is necessary for most approaches, which can be seen for example in the case of Europarl (Graëen et al., 2014).

The present endeavor follows from a preliminary corpus of German political speeches released in 2012 (Barbaresi, 2012), which was at that time the first corpus of its kind for German to be made publicly available. The speeches as a whole cannot be found online, they are only partly stored by search engines or other sources. As such, the corpus has preserved texts which have since vanished from official web pages only to be accessible on paper at the German state archives. The year 2017 marks the end of the legislative period in Germany (September) as well as a succession of presidency (April). Both changes prompt for a updated release of the corpus, which has also been extended to other speakers. The present article provides an extensive, citable reference for this resource, beyond the technical documentation from the first release. The main contributions include a description of the corpus to be released and the presentation of the interface used to navigate through the texts, designed for researchers beyond the corpus or computational linguistics communities as well as for the general public.

1.2. Uses so far
The corpus has already been cited in various scientific publications and in different disciplinary contexts. Three main approaches can be distinguished overall: qualitative analysis, mostly in history and political science; quantitative uses, mostly in machine translation; and the integration into reference corpora and corpus linguistics tools. First, the most frequent use seems to be in history and political science in several countries other than Germany, with detailed analyses of the speeches and uses as a basis for comparison (Ditfurth, 2012; James, 2012; Thonfeld, 2014; Görajek, 2015; Pühninger, 2015a; Pühninger, 2015b; Yu, 2015) including research work by Bachelor, Master or PhD students (Schax, 2012; Seewald, 2013; Simons, 2014; van de Rijt, 2015). Examples for quantitative uses are mostly to be found in machine translation studies, for inclusion in shared tasks or system development, for example in the context of the International Workshop on Spoken Language Transla-
ation (Birch et al., 2013; Kilgour et al., 2013; Freitag et al., 2014; Kilgour et al., 2014; Huck and Birch, 2015; Jehl et al., 2015; Müller et al., 2015). Most notably, it has been employed in order to build language models, to provide in-domain texts for statistical machine translation and a clean text source for backtranslations. Since the corpus is freely available, it also has been included in machine learning studies as well (Zhu et al., 2015).

Integration into reference corpora, uses in tools as well as corpus linguistics studies have been the main use cases so far in a linguistic perspective. The corpus has been cited as blueprint for other corpora targeting political speeches (Osenova and Simov, 2012). Parts of it have become components of German reference corpora, first at the Institute (Osenova and Simov, 2012). The peculiarities of each subcorpus are described separately below. An effort has been made to exclude interviews, speeches that were held by foreign guests as well as speeches held in languages other than German.

2.1. Presidency

The speeches were collected from the official website of the President of the German Republic (bundespraesident.de). The speeches anterior to 1999 are much less numerous and seem to be only a selection. The collection of speeches by Richard von Weizsäcker is far from being complete. Still, it was added to provide the original texts, as such it is the oldest part of the corpus (starting from 1984).

2.1.2. President of the Bundestag

The speeches were gathered from the professional website of the last president of the Bundestag, Norbert Lammert (norbert-lammert.de), who has been in office from 2005 to 2017. The advantage of this source is that a selection has been made, both from speeches held in the Bundestag and on other occasions, so that only highly significant speeches are available on the website. There is however less text to be found than in the archives of the Bundestag. Especially for routine interventions, it would be necessary to filter the plenary protocols, which in their current form (PDF format) is impractical.

2.1.3. Chancellery

The speeches are available from the official website of the German Chancellery (bundesregierung.de). This source also includes speeches from Ministers of State (Staatsminister), who are members of the Cabinet working at the Chancellery. Since these prominent members of the Cabinet are directly linked to representative functions, they have been included in the corpus. The Chancellery speeches represent the largest part of the corpus both in terms of texts and tokens. Documents from four different archives were used: Gerhard Schröder’s terms (1998-2005), Angela Merkel’s 1st (2005-2009), 2nd (2009-2013), and 3rd term (2013-2017). Consequently, the online archives are not homogeneous, there was no real classification for texts anterior to 2005, where a few unrelated speeches from other politicians are to be found. They appear among others invited speakers in the others category. The encoding is sometimes deficient, mostly affecting the punctuation marks and the spaces, which have been partly restored for the corpus. Most speeches from 1998 to 2009 are not available online anymore, some others cannot be found anymore because of a change of website design.

2.1.4. Foreign Affairs

The speeches were collected from the website of the German Ministry of Foreign Affairs (auswaertiges-amt.de). A larger proportion of speeches in languages other than German were to be found, these texts were removed, which

1 http://notes.jan-oliver-ruediger.de/korpora/
2 http://lcm.informatik.uni-leipzig.de/download.html
partly explains the skewed distribution of speeches between the different ministers and state ministers.

## 2.2. Format and metadata

The corpus is made available as a downloadable archive as well as through a series of visualizations and HTML pages. The full text archive is in XML format and Unicode encoding, it follows the guidelines of the Text Encoding Initiative. There is one XML file grouping all the texts of each subcorpus, the files have their own DTD, inspired by the TEI guidelines. Text and metadata have been extracted automatically. In some cases, an automaton has been designed to strip out the salutatory addresses of the speeches using regular expressions, with good accuracy, although not perfect due to the extreme variation among speakers. The following metadata are available (the ones that are not available for all texts are in italics): title(s), speaker, date, place, source, excerpt, salutations, keywords.

For a schematic view of the steps needed to build the corpus, see Figure 1 and Figure 2 which focus on the operations performed from the archive to the corpus in XML format and the visualization. The tokenizer SoMaJo (Proisl and Uhrig, 2016) and the part-of-speech tagger SoMeWeTa (Proisl, 2018) have been used as they achieve state-of-the-art accuracies on web data for German. The older parts of the corpus have been tagged using the TreeTagger (Schmid, 1995).

## 2.3. Texts

The presidency corpus contains a total of 2,048 texts comprising about 3.3 million tokens, covering a period ranging from July 1, 1984 to March 12, 2017. The presidency of the Bundestag subpart features 220 speeches, from October 15, 2005 to September 6, 2017, and a total of about 200,000 tokens. The chancellery subcorpus covers a period extending from the December 11, 1998 to the September 21, 2017. It contains a total of 1,831 texts comprising about 5.3 million tokens. Last, the Foreign Affairs part includes 1,552 speeches from January 16, 2006 to September 17, 2017 for a total of approximately 2.1 million tokens. The total amounts to approximately 10.9 million tokens. Table 1 gives a synoptic view of the contents.

### 2.4. License

As they were given in public, all the speeches can in principle be freely republished as stated by German copyright law, so that there is theoretically no copyright restrictions on this corpus, which is quite rare for German texts. Nonetheless, the law indicates that a republication must partly target a particular author. Although the situation de jure is not clear-cut, the corpus now has been online for more than five years without receiving a warning or a take-down notice, which makes it usable de facto. The corpus as a whole is released under the CC BY-SA (attribute and share-alike) license.

## 3. Interface

In order to provide access to the texts, the texts are also published online separately with a specially designed interface. The texts can be listed, explored, and navigated in two different ways: first a chronological list including metadata such as speaker and extracted keywords, and second a diachronic visualization for selected keywords along with statistics. The purpose is to give insights on the evolution in the use of general concepts (e.g. security, Europe, freedom or war) in a kind of Zeitgeist.

### 3.1. Determination of keywords

The selection of relevant keywords is performed manually after morpho-syntactic linguistic annotation: a shallow parser uses the information after tokenization and POS-tagging to group the tokens in chunks. The valency-oriented chunker (Barbaresi, 2013) uses a bottom-up lin-

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Date</th>
<th>Texts</th>
<th>MTokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>R. Herzog</td>
<td>1994-1999</td>
<td>135</td>
<td>326</td>
</tr>
<tr>
<td>J. Rau</td>
<td>1999-2004</td>
<td>571</td>
<td>902</td>
</tr>
<tr>
<td>H. Köhler</td>
<td>2004-2010</td>
<td>527</td>
<td>775</td>
</tr>
<tr>
<td>C. Wulff</td>
<td>2010-2012</td>
<td>204</td>
<td>290</td>
</tr>
<tr>
<td>J. Gauck</td>
<td>2012-2017</td>
<td>588</td>
<td>933</td>
</tr>
</tbody>
</table>

**Presidency**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Date</th>
<th>Texts</th>
<th>MTokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>N. Lammert</td>
<td>2005-2017</td>
<td>220</td>
<td>~200</td>
</tr>
</tbody>
</table>

**Chancellery (including members of the Cabinet)**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Date</th>
<th>Texts</th>
<th>MTokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>G. Schröder</td>
<td>1998-2005</td>
<td>420</td>
<td>984</td>
</tr>
<tr>
<td>J. Fischer</td>
<td>1998-2005</td>
<td>32</td>
<td>56</td>
</tr>
<tr>
<td>R. Schwantitz</td>
<td>1998-2005</td>
<td>23</td>
<td>28</td>
</tr>
<tr>
<td>H.-M. Bury</td>
<td>1999-2002</td>
<td>42</td>
<td>74</td>
</tr>
<tr>
<td>F.-W. Steinmeier</td>
<td>1999-2005</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>M. Naumann</td>
<td>1999-2000</td>
<td>61</td>
<td>121</td>
</tr>
<tr>
<td>J. Nida-Rümelin</td>
<td>2001-2003</td>
<td>48</td>
<td>93</td>
</tr>
<tr>
<td>C. Weiss</td>
<td>2002-2005</td>
<td>206</td>
<td>299</td>
</tr>
<tr>
<td>A. Merkel</td>
<td>2005-2017</td>
<td>1,030</td>
<td>2,694</td>
</tr>
<tr>
<td>T. de Maizière</td>
<td>2005-2009</td>
<td>43</td>
<td>89</td>
</tr>
<tr>
<td>B. Neumann</td>
<td>2005-2013</td>
<td>323</td>
<td>370</td>
</tr>
<tr>
<td>M. Grütters</td>
<td>2013-2017</td>
<td>162</td>
<td>259</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td>93</td>
<td>207</td>
</tr>
</tbody>
</table>

**Foreign Affairs (including members of the Cabinet)**

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Date</th>
<th>Texts</th>
<th>MTokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>F.-W. Steinmeier</td>
<td>2005-2009</td>
<td>552</td>
<td>912</td>
</tr>
<tr>
<td>G. Erler</td>
<td>2005-2009</td>
<td>81</td>
<td>116</td>
</tr>
<tr>
<td>G. Gloser</td>
<td>2005-2009</td>
<td>48</td>
<td>75</td>
</tr>
<tr>
<td>G. Westerwelle</td>
<td>2009-2013</td>
<td>254</td>
<td>277</td>
</tr>
<tr>
<td>C. Pieper</td>
<td>2009-2013</td>
<td>84</td>
<td>90</td>
</tr>
<tr>
<td>W. Hoyer</td>
<td>2009-2011</td>
<td>42</td>
<td>45</td>
</tr>
<tr>
<td>M.-G. Link</td>
<td>2012-2013</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>M. Roth</td>
<td>2013-2017</td>
<td>220</td>
<td>248</td>
</tr>
<tr>
<td>M. Böhmer</td>
<td>2013-2017</td>
<td>44</td>
<td>39</td>
</tr>
<tr>
<td>M. Ederer</td>
<td>2014-2017</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>S. Steinlein</td>
<td>2014-2017</td>
<td>34</td>
<td>38</td>
</tr>
<tr>
<td>S. Gabriel</td>
<td>2017</td>
<td>42</td>
<td>88</td>
</tr>
<tr>
<td>others</td>
<td></td>
<td>121</td>
<td>150</td>
</tr>
</tbody>
</table>

Table 1: Synoptic view of the corpus
guistic model implemented using finite-state automata. The
transducer takes part of speech tags as input and prints as
output assumptions about the composition of the phrases
and about the position of the verb. The grouping into pos-
sibly relevant chunks enables a valency detection for each
verb based on topological fields, the goal is to look for fre-
quent lexical heads as well as important verbs. Nominal
and prepositional phrases are in focus, the results are of-
ten comparable to text chunks, but the approach is closer to
grammatical rules and to the linguistic understanding of a
phrase.
Thus, the potential keywords are head nouns of nominal
and prepositional phrases, they are then counted and filtered
using term frequencies and inverse document frequency
(TF-IDF). That way, the selection process combines syntac-
tical and statistical indicators of relevance. The keywords
are added as supplementary metadata for the texts. The first
eight words by order of frequency (and relevance) appear in
the general overview of the texts, whereas the first five ones
can be found in the representation of the query by texts.
Additionally, they are scrutinized manually and a list of 50
to 100 keywords is used to explore each subcorpus.

3.2. Visualizations
For each selected keyword, visualizations are generated to
provide an access to and an overview of the corpus in the
form of static web pages, which so far proved easier to
maintain. A series of queries are performed on a SQLite
database to generate web pages, in a diachronic way to see
the keyword evolve in the course of time but also classified
by speaker. The data can be sorted by year, name or
text. The word frequency is displayed using histograms.
This process makes it easier to look for distinctive and/or
relevant keywords. The interface also provides the user with
the context, more specifically the co-text, five words before,
five words after and a link to each text.
The information is put together in CSS/XHTML format, it
uses tabbed navigation and takes advantage of web stan-
dards. It is light both in size and in client-side computation
needs. JavaScript is used to ensure tabbed navigation, to
complete the pages on the fly and to highlight words in the
texts. A chronological overview with metadata is available,
as shown in Figure 3. The list of selected keywords serves
as a menu and can be used to browse the corpus, as shown
in Figure 4. Clicking on a keyword then leads to the visu-
alizations, most notably a diachronic view as in Figure 5a
and an overview sorted by speaker as shown in Figure 5b.
The numbers on the right side stand for the texts, the num-
bers are larger if the keywords are more frequent. Clicking
on the numbers then leads to the texts where the keyword is
highlighted. It is also possible to browse the texts sequen-
tially in chronological order.

4. Conclusion

The main contributions of this article include a scientific reference for the corpus to be released and the description of an interface to navigate through the texts, designed for researchers beyond the corpus or computational linguistics communities as well as for the general public. Indeed, the corpus has been used in various disciplinary contexts, three main approaches can be distinguished overall: qualitative analysis, quantitative uses, and integration into reference corpora and corpus linguistics tools.

The corpus can be considered to be from the 21st century since most speeches have been written after 2001 and also because it includes a modern visualization interface providing both synoptic overviews ordered chronologically, by speaker or by keyword as well as consequent accesses to the texts. An updated and extended version of the corpus is described, it features the four highest ranked functions on federal state level up to the year 2017. The corpus is made available as an archive as well as through a series of visualizations and HTML pages. Data and visualization are both accessible online.7

Acknowledgements

Parts of these work have been supported by CLARIN-D.

5. Bibliographical References


