German Political Speeches - Corpus and Visualization

Adrien Barbaresi

German Political Speeches Corpus and Visualization

2nd release

http://purl.org/corpus/german-speeches

Adrien Barbaresi

03/05/2012

This second version is released on the occasion of the DGfS-CL Poster-Session in Frankfurt, where I present a poster.

Contents

1 Interest

1.1 Corpus

To my knowledge, no corpus of this kind has so far been made publicly available for German. There are corpora containing political campaign speeches that were partly developed by commercial companies as well as different sources that gather political texts classified as important, but not systematically with a common reference.

Another main interest of this corpus is that most speeches could not be found on Google until I put them online, and the Chancellery speeches from before 2011 are not to be found on its website anymore.

Last, there is no copyright on this corpus, which is quite rare for German texts. As they were given in public, all the speeches can be freely republished as stated by German copyright law[1]. Nonetheless, the law indicates that a republication must not target a particular author.

1.2 Visualization

I provide raw data for researchers that are able to use it and a simple visualization interface for those who want to get a glimpse of what is in the corpus before downloading it or thinking about using more complete tools[2].

The output is in valid CSS/XHTML format, it uses tabbed navigation and takes advantage of recent standards. It is light both in size and in client-side computation needs, using just a little JavaScript.

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. A glimpse of the co-text is also available.

2 Contents

I chose to group the texts into two subcorpora, since the roles of President and Chancellor are quite different.

2.1 Presidency subcorpus (Bundespräsidenten)

2.1.1 Sources

The speeches were crawled from the online archive of the German Presidency (bundespraesident.de). Here are the available lists of speeches for each president :

- Richard von Weizsäcker (1984-1994)
- Roman Herzog (1994-1999)
- Johannes Rau (1999-2004)
- Horst Köhler (2004-2010)
- Christian Wulff (2010-2012)

The collection of speeches by Richard von Weizsäcker is far from complete. Still, it was added to provide the original texts.

[1] § 48 UrhG, Öffentliche Reden.
http://bundesrecht.juris.de/urhg/__48.html

[2] such as the TXM project.

2.1.2 Contents

The corpus contains a total of 1 442 texts comprising 2 392 074 tokens, covering a period extending from July 1, 1984 to February 17, 2012.

President	Texts	Tokens
Johannes Rau	568	961 538
Horst Köhler	527	774 563
Christian Wulff	202	285 893
Roman Herzog	131	322 468
Richard von Weizsäcker	14	47 612

2.2 Chancellery subcorpus (Bundesregierung)

2.2.1 Sources

There are many speeches available on the official website of the German Chancellery (bundesregierung.de) but no real classification for those before 2005 (Angela Merkel's time as chancellor). Not all the speeches in the corpus can be found on this website anymore due to a change of design.

Documents from three archives were used :

- Gerhard Schröder's terms (1998-2005)
- Angela Merkel's 1st term (2005-2009)
- Angela Merkel's 2nd term (2009-2011)

2.2.2 Contents

The corpus contains a total of 1 831 texts comprising 3 891 588 tokens, covering a period extending from the December 11, 1998 to the December 6, 2011.

There are not only speeches by the chancellors, but also a number of other state ministers that are linked to the head of the government and a few unrelated speeches from other politicians (which are not well-sorted).

Politician	Texts	Tokens
Angela Merkel	607	1 640 176
Gerhard Schröder	420	984 365
Bernd Neumann	248	281 636
Christina Weiss	206	299 175
k.A.	91	203 675
Michael Naumann	61	120 542
Julian Nida-Rümelin	48	92 663
Thomas de Maizière	43	88 837
Hans Martin Bury	42	73 502
Joschka Fischer	32	55 507
Rolf Schwanitz	23	28 066
Frank-Walter Steinmeier	10	23 444

2.3 Remarks and caveats

This ordering was made using regular expressions in both titles and excerpts. It seems to work properly but does not guarantee a perfect classification.

I made an effort to exclude all the interviews, the speeches that were held by foreign guests as well as speeches held in languages other than German.

An automaton stripped out the salutatory addresses of the speeches using regular expressions, with good accuracy, although not perfect due to the extreme variation among speakers.

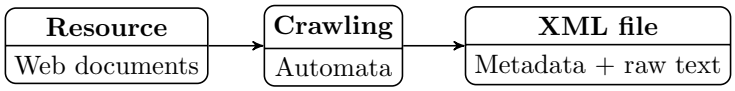
Chancellery subcorpus There are numerous authors and among them ministers whose archives should be elsewhere. As a consequence, the relative frequencies are not always significant. Speeches from two presidents are included, which I have not yet removed, they appear among others invited speakers in the 'no name found' *k.A.*[3] category.

In the texts from before 2005 the encoding is deficient, mostly affecting the punctuation marks and the spaces.

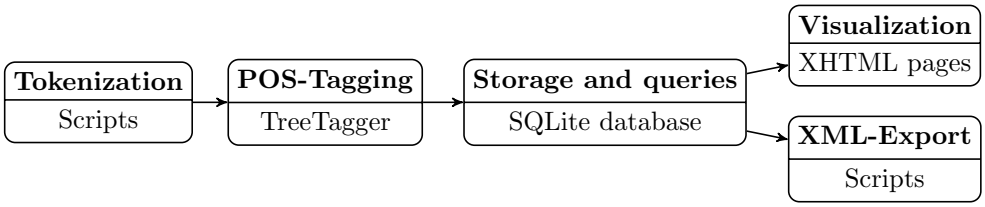
3 Storage and format

3.1 Workflow

3.1.1 From the web pages to the raw XML file



3.1.2 From the raw XML file to the visualization



3.2 XML format

The corpus is released in XML and Unicode format. There is one XML file grouping all the texts of each subcorpus, since I thought it was easier to manipulate that way (its size stays reasonable). The files have their own DTD, inspired by the TEI guidelines. The corpus is not fully TEI-compliant yet, but it is closer to it than the first release.

Raw XML file The metadata are properly encoded, the texts are given as they were crawled, with no enrichment whatsoever.

XML file Tokenisation[4], POS-Tags and Lemmas[5] are included.

3.3 Possible improvements

Major Among the major improvements in sight :

- Finer POS-Tags (i.e. using the RFTagger[6]).
- Better compliance with the XML TEI guidelines.

Minor Among the minor adjustments that could be done :

- Updates on a regular basis.
- Expanding to other available speeches archives.

[3] for 'keine Angabe' in German.

[4] Partly using a Perl script developed by Stefanie Dipper, which can be found on her website : http://www.linguistics.ruhr-uni-bochum.de/ dipper/

[5] Provided by the TreeTagger : http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/

3.4 Licence

When asked, the relevant services either did not reply at all or granted me the full rights to republish the corpus.

I chose to publish my work under the CC BY-SA (attribution and share-alike)[7] licence.

4 Interface

4.1 Description

This is the second release, most of the processes have been stabilized but a few tools are still under development.

By now there are static web pages : a series of queries were performed to generate web pages describing word frequencies. A list serves as a menu, it contains selected relevant words.

It is not only about counting words, the purpose is to give an insight on the topics developed by a government official and on the evolution in the use of general concepts (like security, Europe, freedom or war) – a sort of Zeitgeist.

The interface also provides the user with the context – more specifically the co-text, five words before, five words after and a link to the text.

JavaScript is used to ensure tabbed navigation, to complete the pages on the fly and to highlight words in the texts.

4.2 Determination of keywords

I designed an algorithm to try to assign relevant keywords to each text. It is based on a shallow parsing which uses the POS-tags. The goal is to look for frequent lexical heads as well as important verbs. I used a stoplist to filter out very common words like 'Nation', 'Deutschland', 'Europa', 'Mensch' and verbs like 'werden', 'können' or 'wollen'.

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.

4.3 Hypothetical development plan

- Dynamic version to enable users to perform their own queries (postponed due to infrastructure problems).
- Refine queries using lemmas and/or morpho-grammatical information and maybe regular expressions.
- Open-source the code once it is stabilized.

[7] https://creativecommons.org/licenses/by-sa/3.0/deed.en