German political speeches from the 21st c.

ParlaFormat Workshop – Amersfoort

Adrien Barbaresi

May 24, 2019

Berlin-Brandenburg Academy of Sciences

Timeline

- *21st century since most speeches have been written after 2001*
- 2012: Initial release (Presidency and Chancellery)
- Become components of German reference corpora, IDS and BBAW (Digital Dictionary of German – *dwds.de*)
- Preserved web pages which have now partly moved
- 2018: updated release of the corpus, extension to other speakers

Content

1. President (*Bundespräsident*)
   2,048 texts – 1984-2017
2. President of the Bundestag (*Bundestagspräsident*)
   220 speeches – 2005-2017
3. Chancellor (*Bundeskanzler*) and corresponding state ministers/secretaries
   1,831 texts – 1998-2017
4. Minister for Foreign Affairs (*Bundesminister des Auswärtigen*) and corresponding state ministers/secretaries
   1,552 speeches – 2006-2017
5. *Others to come: Austria, Germany and Switzerland*

Uses so far

- Scientific publications
  - qualitative analysis, mostly in history and political science
    $\rightarrow$ in several countries, detailed analyses and uses comparisons
  - quantitative uses, mostly in machine translation
    $\rightarrow$ for inclusion in shared tasks or system development: build language models, provide in-domain texts for statistical machine translation and a clean text source for backtranslations
  - integration into reference corpora and corpus linguistics tools
    $\rightarrow$ blueprint for other corpora targeting political speeches, corpus comparison, demonstration purposes

Uses so far

- Scientific publications
  - qualitative analysis, mostly in history and political science
    $\rightarrow$ in several countries, detailed analyses and uses comparisons
  - quantitative uses, mostly in machine translation
    $\rightarrow$ for inclusion in shared tasks or system development: build language models, provide in-domain texts for statistical machine translation and a clean text source for backtranslations
  - integration into reference corpora and corpus linguistics tools
    $\rightarrow$ blueprint for other corpora targeting political speeches, corpus comparison, demonstration purposes
- Occasional uses
  - Newspapers
  - Tweets (politicians)
  - Website statistics $\rightarrow$ regularly read by political staff

Release

Currently available

- Full text archive (XML format)
- All speeches (accessible through lists and visualizations) online
- Statistics for selected keywords with access to the texts

Further work on extension and format ahead
$\rightarrow$ XML TEI as coming milestone

Formal considerations

Available metadata

Title(s), speaker, date, place, source, excerpt, salutations, keywords

- Precise and TEI-conform `sourceDesc`
- Speaker (main role, guest speeches) + institution/event type
  `affiliation` / `occupation` / `roleName` ?
- Forms of salutations
  `name` / `persName` + `role="..."` ?

Draft: Metadata

```
<sourceDesc>
  <biblFull>
    <titleStmt>
      <title type="main">...</title>
      <title type="sub">...</title>
      <speaker>N.N.</speaker>
    </titleStmt>
    <publicationStmt>
      <publisher>Government agency</publisher>
      <date type="publication">YYYY-MM-DD</date>
      <date type="archive">YYYY-MM-DD</date>
      <idno type="URL">http://...</idno>
    </publicationStmt>
    <seriesStmt>
        [additional info]
    </seriesStmt>
  </biblFull>
</sourceDesc>
```

Draft: Content

```
<group>
  <text type="speech">
    <body>
      <p>...</p>
      <p>...</p>
    </body>
  </text>
</group>
```

Thank you for your attention!

purl.org/corpus/german-speeches

barbaresi@bbaw.de

@adbarbaresi