A corpus of German political speeches from the 21st century

Adrien Barbaresi

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 a 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 symbolical 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 as 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 (Graën 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; Górajek, 2015; Pühringer, 2015a; Pühringer, 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-

tion (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 for the German Language (Lüngen, 2017), and soon among the resources of the Digital Dictionary of the German Language (Geyken et al., 2017) as a useful complement to existing newspaper corpora, since comparisons on lexical level show that it is significantly different from other written genres (Barbaresi and Würzner, 2014). Additionally, the resource is used by several tools for demonstration purposes, corpus exploration tools like the *Corpus Explorer*[1] (Dang-Anh and Rüdiger, 2015) and the *Leipzig Corpus Miner*[2] (Wiedemann and Niekler, 2016) as well as the citation format CTS (Tiepmar and Heyer, 2017).

Last, while website statistics indicate that the texts are regularly read by the public including politicians and political staff, links to the texts and the interface have been posted by politicians and newspapers (for example in tweets or blog posts), for single texts but also for topical visualizations, e.g. *Der Spiegel* with the notion of values (*Werte*).[3]

2. Contents of present release

2.1. Sources

The corpus gathers four different types of speakers, which currently correspond to the four highest ranked functions on German federal state level, it also includes speeches by affiliated state ministers and state secretaries (members of the Cabinet):

1. President (*Bundespräsident*)
2. President of the Bundestag (*Bundestagspräsident*)
3. Chancellor (*Bundeskanzler*) and corresponding state ministers/secretaries
4. Minister for Foreign Affairs (*Bundesminister des Auswärtigen*), which in recent times frequently carried the title of vice-chancellor (*Vizekanzler*), and corresponding state ministers/secretaries

The 2012 release featured speeches from the presidency and the chancellery only. The present corpus includes two further types as well as an update for a full legislative period. The main focus is on the 21st century, with a few speeches from the end of the 20th century, as shown in Table 1, the token counts are obtained using the tokenizer *SoMaJo* (Proisl and Uhrig, 2016). The collection is not necessarily exhaustive, the official sources from which the speeches have been downloaded are trustworthy but they have no legal obligation to make all the speeches available online. 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.

2.1.1. Presidency

The speeches were collected from the online archive of the German Presidency (*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

[3] https://web.archive.org/web/20160926185139/-http://www.spiegel.de/wissenschaft/mensch/afd-auf-eure-werte-kann-ich-verzichten-kolumne-a-1113649.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.[4] 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.

Speaker	Date	Texts	MTokens
Presidency			
R. von Weizsäcker	1984-1994	23	59
R. Herzog	1994-1999	135	326
J. Rau	1999-2004	571	902
H. Köhler	2004-2010	527	775
C. Wulff	2010-2012	204	290
J. Gauck	2012-2017	588	933
Presidency of the Assembly			
N. Lammert	2005-2017	220	$\sim$ 200
Chancellery (including members of the Cabinet)			
G. Schröder	1998-2005	420	984
J. Fischer	1998-2005	32	56
R. Schwanitz	1998-2005	23	28
H.-M. Bury	1999-2002	42	74
F.-W. Steinmeier	1999-2005	10	23
M. Naumann	1999-2000	61	121
J. Nida-Rümelin	2001-2003	48	93
C. Weiss	2002-2005	206	299
A. Merkel	2005-2017	1,030	2,694
T. de Maizière	2005-2009	43	89
B. Neumann	2005-2013	323	370
M. Grütters	2013-2017	162	259
others		93	207
Foreign Affairs (including members of the Cabinet)			
F.-W. Steinmeier	2005-2009 2013-2017	552	912
G. Erler	2005-2009	81	116
G. Gloser	2005-2009	48	75
G. Westerwelle	2009-2013	254	277
C. Pieper	2009-2013	84	90
W. Hoyer	2009-2011	42	45
M.-G. Link	2012-2013	19	21
M. Roth	2013-2017	220	248
M. Böhmer	2013-2017	44	39
M. Ederer	2014-2017	12	5
S. Steinlein	2014-2017	34	38
S. Gabriel	2017	42	88
others		121	150

Table 1: Synoptic view of the corpus

2.4. License

As they were given in public, all the speeches can in principle be freely republished as stated by German copyright law[5], 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 not 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 (attribution and share-alike)[6] license.

[4] http://www.tei-c.org

[5] § 48 UrhG, Öffentliche Reden: http://bundesrecht.juris.de/urhg/__48.html

[6] https://creativecommons.org/licenses/by-sa/4.0/

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-

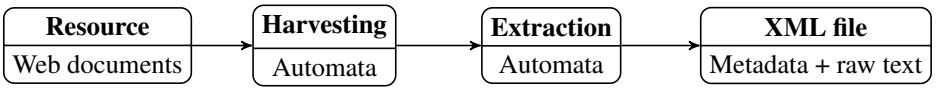


Figure 1: From the web pages to the XML archive

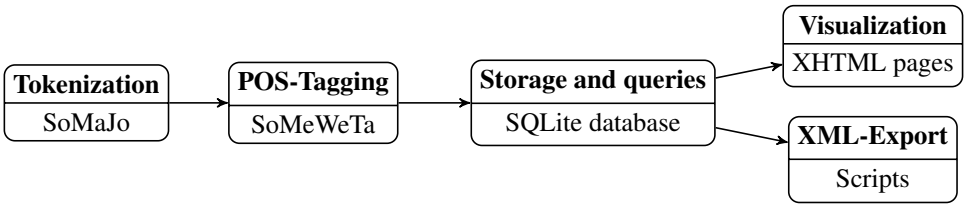


Figure 2: From the XML archive to export and visualization

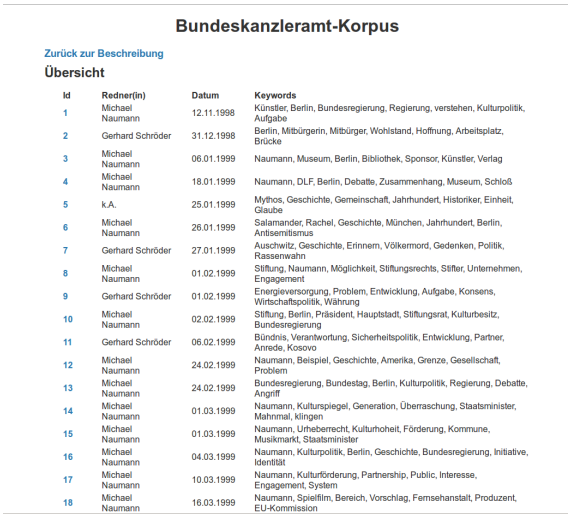
guistic model implemented using finite-state automata. The transducer takes part of speech tags as input and prints as output assumptions about the composition of the phrases and about the position of the verb. The grouping into possibly relevant chunks enables a valency detection for each verb based on topological fields, the goal is to look for frequent lexical heads as well as important verbs. Nominal and prepositional phrases are in focus, the results are often 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 syntactical 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 standards. 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 visualizations, 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 numbers 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-



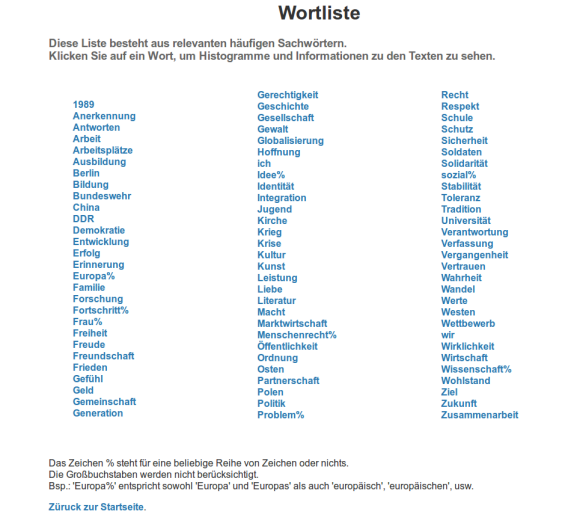
Bundeskanzleramt-Korpus

Zurück zur Beschreibung

Übersicht

Id	Redner(in)	Datum	Keywords
1	Michael Naumann	12.11.1998	Künstler, Berlin, Bundesregierung, Regierung, verstehen, Kulturpolitik, Aufgabe
2	Gerhard Schröder	31.12.1998	Berlin, Mitbürgerin, Mitbürger, Wohlstand, Hoffnung, Arbeitsplatz, Brücke
3	Michael Naumann	06.01.1999	Naumann, Museum, Berlin, Bibliothek, Sponsor, Künstler, Verlag
4	Michael Naumann	18.01.1999	Naumann, DLF, Berlin, Debatte, Zusammenhang, Museum, Schloß
5	k.A.	25.01.1999	Mythos, Geschichte, Gemeinschaft, Jahrhundert, Historiker, Einheit, Glaube
6	Michael Naumann	26.01.1999	Salamander, Rachel, Geschichte, München, Jahrhundert, Berlin, Antisemitismus
7	Gerhard Schröder	27.01.1999	Auschwitz, Geschichte, Erinnern, Völkermord, Gedenken, Politik, Rassenwahn
8	Michael Naumann	01.02.1999	Stiftung, Naumann, Möglichkeit, Stiftungsrechts, Stifter, Unternehmen, Engagement
9	Gerhard Schröder	01.02.1999	Energieversorgung, Problem, Entwicklung, Aufgabe, Konsens, Wirtschaftspolitik, Währung
10	Michael Naumann	02.02.1999	Stiftung, Berlin, Präsident, Hauptstadt, Stiftungsrat, Kulturbesitz, Bundesregierung
11	Gerhard Schröder	06.02.1999	Bündnis, Verantwortung, Sicherheitspolitik, Entwicklung, Partner, Anrede, Kosovo
12	Michael Naumann	24.02.1999	Naumann, Beispiel, Geschichte, Amerika, Grenze, Gesellschaft, Problem
13	Michael Naumann	24.02.1999	Bundesregierung, Bundestag, Berlin, Kulturpolitik, Regierung, Debatte, Angriff
14	Michael Naumann	01.03.1999	Naumann, Kulturspiegel, Generation, Überraschung, Staatsminister, Mahnmal, klingen
15	Michael Naumann	01.03.1999	Naumann, Urheberrecht, Kulturhoheit, Förderung, Kommune, Musikmarkt, Staatsminister
16	Michael Naumann	04.03.1999	Naumann, Kulturpolitik, Berlin, Geschichte, Bundesregierung, Initiative, Identität
17	Michael Naumann	10.03.1999	Naumann, Kulturförderung, Partnership, Public, Interesse, Engagement, System
18	Michael Naumann	16.03.1999	Naumann, Spielfilm, Bereich, Vorschlag, Fernsehanstalt, Produzent, EU-Kommission

Figure 3: List of texts with metadata



Wortliste

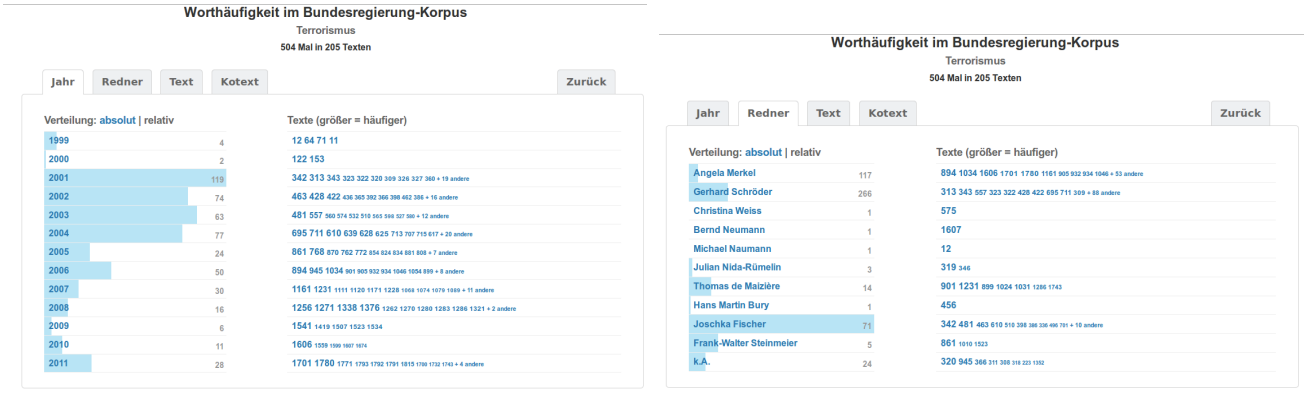
Diese Liste besteht aus relevanten häufigen Sachwörtern.
Klicken Sie auf ein Wort, um Histogramme und Informationen zu den Texten zu sehen.

1989
Anerkennung
Antworten
Arbeit
Arbeitsplätze
Ausbildung
Berlin
Bildung
Bundeswehr
China
DDR
Demokratie
Entwicklung
Erfolg
Erinnerung
Europa%
Familie
Forschung
Fortschritt%
Frau%
Freiheit
Freude
Freundschaft
Frieden
Gefühl
Geld
Gemeinschaft
Generation
Gerechtigkeit
Geschichte
Gesellschaft
Gewalt
Globalisierung
Hoffnung
ich
Idee%
Identität
Integration
Jugend
Kirche
Krieg
Krise
Kultur
Kunst
Leistung
Liebe
Literatur
Macht
Marktwirtschaft
Menschenrecht%
Öffentlichkeit
Ordnung
Osten
Partnerschaft
Polen
Politik
Problem%
Recht
Respekt
Schule
Schutz
Sicherheit
Soldaten
Solidarität
sozial%
Stabilität
Toleranz
Tradition
Universität
Verantwortung
Verfassung
Vergangenheit
Vertrauen
Wahrheit
Wandel
Werte
Westen
Wettbewerb
wir
Wirklichkeit
Wirtschaft
Wissenschaft%
Wohlstand
Ziel
Zukunft
Zusammenarbeit

Das Zeichen % steht für eine beliebige Reihe von Zeichen oder nichts.
Die Großbuchstaben werden nicht berücksichtigt.
Bsp.: 'Europa%' entspricht sowohl 'Europa' und 'Europas' als auch 'europäisch', 'europäischen', usw.

Zurück zur Startseite.

Figure 4: Selected keywords for the Chancellery corpus



(a) Diachronic view for keyword *terrorism*

(b) Overview sorted by speaker for *terrorism*

Figure 5: Aggregated views of the contents in the course of time or grouped by speaker in the Chancellery corpus

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

Barbaresi, A. and Würzner, K.-M. (2014). For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In *KONVENS 2014, NLP4CMC workshop*, pages 2–10. Hildesheim University Press.

Barbaresi, A. (2012). German Political Speeches – Corpus and Visualization. Technical report, École Normale Supérieure de Lyon.

Barbaresi, A. (2013). A one-pass valency-oriented chunker for German. In Hans Uszkoreit Zygmunt Vetulani, editor, *Language & Technology Conference*, Proceedings of the 6th Language & Technology Conference, pages 157–161, Poznan, Poland.

Birch, A., Durrani, N., and Koehn, P. (2013). Edinburgh SLT and MT system description for the IWSLT 2013 evaluation. In *Proceedings of the 10th International Workshop on Spoken Language Translation*, pages 40–48.

Dang-Anh, M. and Rüdiger, J. O. (2015). From Frequency to Sequence: How Quantitative Methods Can Inform Qualitative Analysis of Digital Media Discourse. *10plus1: Living Linguistics*, 1:57–73.

Ditfurth, J. (2012). *Zeit des Zorns: warum wir uns vom Kapitalismus befreien müssen*. Westend Verlag.

Freitag, M., Wuebker, J., Peitz, S., Ney, H., Huck, M., Birch, A., Durrani, N., Koehn, P., Mediani, M., Slawik, I., et al. (2014). Combined spoken language translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Geyken, A., Barbaresi, A., Didakowski, J., Jurish, B., Wiegand, F., and Lemnitzer, L. (2017). Die Korpusplattform des "Digitalen Wörterbuchs der deutschen Sprache" (DWDS). *Zeitschrift für germanistische Linguistik*, 45(2):327–344.

Górajek, A. (2015). Von der Mehrdimensionalität der Geschichte–Gerhard Schröder und seine Haltung gegenüber Polen. *Zeitschrift des Verbandes Polnischer Germanisten*, 4(4):273–280.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *Proceedings of KONVENS*, pages 222–227.

Guerini, M., Strapparava, C., and Stock, O. (2008). CORPS: A corpus of tagged political speeches for persuasive communication processing. *Journal of Information Technology & Politics*, 5(1):19–32.

Huck, M. and Birch, A. (2015). The Edinburgh machine translation systems for IWSLT 2015. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–38.

James, J. (2012). *Preservation and National Belonging in Eastern Germany: Heritage Fetishism and Redeeming Germanness*. Palgrave MacMillan.

Jehl, L., Simianer, P., Hitschler, J., and Riezler, S. (2015). The Heidelberg University English-German translation system for IWSLT 2015. *Proceedings of IWSLT*.

[7] http://purl.org/corpus/german-speeches

Kilgour, K., Mohr, C., Heck, M., Nguyen, Q. B., Nguyen, V. H., Shin, E., Tseyzer, I., Gehring, J., Müller, M., Sperber, M., et al. (2013). The 2013 KIT IWSLT Speech-to-Text Systems for German and English. In *International Workshop on Spoken Language Translation (IWSLT)*.

Kilgour, K., Heck, M., Müller, M., Sperber, M., Stüker, S., and Waibel, A. (2014). The 2014 KIT IWSLT speech-to-text systems for English, German and Italian. In *International Workshop on Spoken Language Translation (IWSLT)*.

Laver, M., Benoit, K., and Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331.

Lüngen, H. (2017). DeReKo – Das Deutsche Referenzkorpus. *Zeitschrift für germanistische Linguistik*, 45(1):161–170.

Müller, M., Nguyen, T.-S., Sperber, M., Kilgour, K., Stüker, S., and Waibel, A. (2015). The 2015 KIT IWSLT Speech-to-Text Systems for English and German. In *International Workshop on Spoken Language Translation (IWSLT)*.

Osenova, P. and Simov, K. I. (2012). The political speech corpus of bulgarian. In *Proceedings of LREC*, volume 2012, pages 1744–1747. ELRA.

Proisl, T. and Uhrig, P. (2016). SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop*, pages 57–62. Association for Computational Linguistics.

Proisl, T. (2018). SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In *Proceedings of LREC*. ELRA.

Pühringer, S. (2015a). Markets as "ultimate judges" of economic policies: Angela Merkel's discourse profile during the economic crisis and the European crisis policies. *On the Horizon*, 23(3):246–259.

Pühringer, S. (2015b). Marktmetaphoriken in Krisennarrativen von Angela Merkel. In Walter Otto Ötsch, et al., editors, *Markt! Welcher Markt?*, pages 229–252. Metropolis-Verlag.

Schax, A. (2012). Tracing Transformations-The development of Germany's Strategic Culture during the last two decades. Master's thesis, Utrecht University.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*.

Seewald, F. (2013). Die deutsche Außen-und Sicherheitspolitik von 2001 bis 2012 im Lichte des Zivilmachtkonzepts. Master's thesis, University of Hagen.

Simons, J. P. (2014). Discourse and the Shift in Social Democratic Ideology and Employment Policies: A Comparison of the PvdA and the SPD. Master's thesis, Leiden University.

Thonfeld, C. (2014). Cosmopolitan Normalisation? The Culture of Remembrance of World War II and the Holocaust in Unified Germany. *TáiDàlLìShǐXuéBào*, 53:181–227.

Tiepmar, J. and Heyer, G. (2017). An Overview of Canonical Text Services. *Linguistics and Literature Studies*, 5(2):132–148.

van de Rijt, L. (2015). Enabling and Constraining: A Study on Possibilities of Agents in the EU-Polity during the Turkish Accession Process from 1999 until 2013. B.S. thesis, Utrecht University.

Wiedemann, G. and Niekler, A. (2016). Analyse qualitativer Daten mit dem "Leipzig Corpus Miner". In *Text Mining in den Sozialwissenschaften*, pages 63–88. Springer.

Yu, T. (2015). The German revolution of 1918-1919 in modern studies and in public perception. *History Magazine: Researches*, 3:280–287.

Zhu, L., Kilgour, K., Stüker, S., and Waibel, A. (2015). Gaussian free cluster tree construction using deep neural network. In *Sixteenth Annual Conference of the International Speech Communication Association*.