ParlAT beta
Corpus of Austrian Parliamentary Records

Tanja Wissik, Hannes Pirker
Austrian Centre for Digital Humanities, Austrian Academy of Sciences
Sonnenfelsgasse 19, 1010 Vienna, Austria
{tanja.wissik, hannes.pirker}@oeaw.ac.at

Abstract

The paper presents the beta Version of the ParlAT Corpus, a corpus of Austrian parliamentary records and its current state. The ParlAT project aims to create a corpus of all digitally available parliamentary records from the National council – one of the two chambers of the Austrian parliament – starting with 1945, i.e. for the period of the so called "Second Republic". The ParlAT beta contains parliamentary records for the last 21 years (i.e. between 1996 - 2017), that is 36% of the relevant digitally available parliamentary records. This paper will describe the data collection and data processing and give an outlook on future work.

Keywords: parliamentary records, corpus building, annotation, linking to external sources

1 Introduction

Parliamentary records are an interesting resource for various fields in the Humanities and Social Sciences, such as linguistics, political science, history, as well as for fields in the Information Sciences such as NLP or information retrieval. As a consequence, there are many initiatives, on the national and international level, that aim at compiling and analysing parliamentary records. Examples for monolingual corpora are the Hansard Corpus, the collection of the parliamentary records of the British Parliament between 1803 and 2005 (Alexander and Davies, 2015) or the Talk of Norway, a collection of the Norwegian parliamentary data (Lapponi and Søyland, 2016), examples for multilingual corpora are the ECPC Corpus, the European Parliamentary Comparable and Parallel Corpora for Spanish and English (Calzada Pérez, Marín Cucala and Martínez Martínez, 2006). A recent survey on parliamentary data in the context of the research infrastructure CLARIN (Fišer and Lenardič, 2017) has identified over 20 corpora of parliamentary data. However, the available corpus for Austrian parliamentary records, listed in the survey only covered a short time period, from 2013 to 2015 (Sippl et al. 2016), and is therefore the corpus not suitable for all research questions, especially not for diachronic analysis.F

The aim of the ParlAT project is to fill this gap and create a corpus of all digitally available parliamentary records from the National Council ("Nationalrat"), one of the two chambers of the Austrian Parliament for the "Second Republic." i.e. for the historic period starting in 1945 until today.

The verbatim records, in German called "Stenographische Protokolle" ("shorthand record") are available from the website of the Austrian Parliament[1] starting from the V legislation period. Prior to this, for the legislation period I-IV ("First Republic"), only scans are available via the platform ALEX at the Austrian National Library[2]. For the legislation period V to XIX (from 1945 to 1995) the documents are only available in pdf format, for the legislation period XX – XXV (starting from 1996) the documents are also available in html format. While parliamentary records are highly standardised and structured texts, changes to the structure over time can be observed (see also Figure 1).

The so called "Geschäftsordnungsgesetz 1975" law dictates that all public sessions of the National Council are recorded verbatim and in their entirety. However, it is also true that these records are not live recordings, and that the speakers get the verbatim records prior to publication and can make, for example, stylistic changes. In case of doubt, the President of the National Council decides if a change is admissible or not.

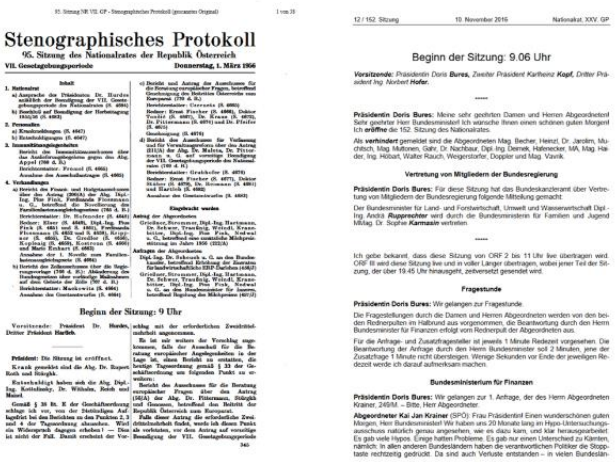

Figure 1: Comparison of verbatim records over time: left from 1950 in column style and right from 2016

All the texts are downloadable from the website of the Austrian Parliament and parliamentary documents can be searched by keyword or by identification number in the search interface. Besides the verbatim records, other documents, such as requests and inquiries, are available. However, the search interface does not support a full-text search or linguistically motivated queries nor is it possible to query texts across different legislative periods.

[1] https://www.parlament.gv.at/PAKT/STPROT/
[2] http://alex.onb.ac.at/spe.htm

Legislation period	Beginning	End	Nr. of documents	Format
V	19.12.1945	08.11.1949	117	pdf
VI	08.11.1949	18.03.1953	103	pdf
VII	18.03.1953	08.06.1956	95	pdf
VIII	08.06.1956	09.06.1959	84	pdf
IX	09.06.1959	14.12.1962	109	pdf
X	14.12.1962	30.03.1966	95	pdf
XI	30.03.1966	31.03.1970	175	pdf
XII	31.03.1970	04.11.1971	52	pdf
XIII	04.11.1971	04.11.1975	151	pdf
XIV	04.11.1975	04.06.1979	123	pdf
XV	05.06.1979	18.05.1983	149	pdf
XVI	19.05.1983	16.12.1968	175	pdf
XVII	17.12.1968	04.11.1990	52	pdf
XVIII	05.11.1990	06.11.1994	151	pdf
XIX	07.11.1994	14.01.1996	57	pdf
XX	15.01.1996	28.10.1999	183	pdf, html
XXI	29.10.1999	19.12.2002	118	pdf, html
XXII	20.12.2002	29.10.2006	164	pdf, html
XXIII	30.10.2006	27.10.2008	76	pdf, html
XXIV	28.10.2008	28.10.2013	220	pdf, html
XXV	29.10.2013	08.11.2017	199	pdf, html
Total number of documents			2648	

Table 1: Availability of Austrian parliamentary records for the time period 1945 – 2017.

2 ParlAT – Corpus of Austrian Parliamentary Records

The present version of the ParlAT corpus covers the parliamentary records from the XX to the XXV legislative period (1996 – 2017). The legislative period that has started in November 2017 is not yet included.

However, the ParlAT is planned as a monitor corpus and new material will be added over time. In this phase of the project, we focused on the documents that are available in html. However, we are testing and establishing a workflow to also include the documents in pdf format, once several OCR-related issues have been resolved.

2.1 Coverage and size of the ParlAT beta

The ParlAT beta contains the parliamentary records from 1996 – 2017. However, out of the 960 documents available for this period only 952 documents are included in the corpus, as eight documents are only preliminary verbatim records in pdf format which could not be processed for the first version of the corpus. Therefore, 36% of the available documents have already been processed and are available in the corpus query system. The corpus size is 75,222,970 tokens.

Number of documents	952
Number of tokens	75,222,970
Number of types	585,628
Number of lemmas	123,894

Table 2: Size of the ParlAT beta.

3 Data collection and processing

In the following section we will describe the data collection and data processing for this project. There are a lot of different workflow described in literature for example inter alia Marx 2009, Marx and Schuth 2010, Blessing et al 2015, Blätte 2016.

3.1 Data format

The html documents were scraped from the Austrian parliamentary website and transformed into so called "vertical" or "word-per-line (WPL)" text, because the corpus query system we are using (see section 3.3), requires this input format. In this format, words are written one word per line, so each line contains one word, number or punctuation mark. The "vertical" is a plain text file without any formatting. In this format, the part-of-speech tagging and lemmatization are provided in two additional columns, separated by tabs (Kilgarriff et al.,2004; SketchEngine 2017).

3.2 Metadata, annotation and markup

For the parliamentary records, we only used a reduced set of metadata: type of document, legislative period, date, year and where the original file is stored.

The parliamentary records were part-of-speech tagged and lemmatized using both the TreeTagger and the RFTagger.. Moreover, basic structural markup in form of xml tags were added. Structural information is recorded in the xml element <section> with a @type attribute that can take three different values: "preamble", "sitzung" and "postfix". The section type "preamble" contains general information on the parliamentary session such as legislation period, date, agenda items, notification of sickness or absence of delegates, request or inquires to be treated during the parliamentary session. The section type "sitzung" contains the actual parliamentary session, recording the speakers and the speeches, but also interjections and heckling. The section type "postfix" contains the imprint information.

Another structural element is the <comment> element. Text passages that were set in italics and between brackets in the original documents and contained information other than the utterances of the speakers, were annotated using this element. An example for such 'comments' would be notes on occurrences such as applause from a specific party "*(Beifall bei der SPÖ)*" or interjections from members of a specific party *"(Zwischenruf bei der SPÖ.)" or* information on procedural elements like the president taking the chair *"(Präsidentin Bures übernimmt den Vorsitz)*.

The original html files also contain references and links to other documents (such as motions, reports etc.). Furthermore, persons appearing in the documents, such as delegates, are annotated and linked to their profile on the website of the Austrian Parliament. In the transformation into a "vertical" text to be processed in a corpus query tool a lot of this linking information is lost because the corpus query tool cannot process complex markup (Kilgarriff 2004; SketchEngine 2017). However, we tried to keep information such as speaker identification by annotating it with the element <person>. Since the project is in its preliminary phase, we have not yet annotated information on turn-taking, and onset and endings of the utterances of speakers – so when one speaker starts and ends his speech in front of the parliament. However, this is planned as a next step as well as the enrichment with biographical metadata for the speakers.

3.3 Corpus query system and interface

At the moment, internally, we use the SketchEngine as corpus query system.

Figure 2: Concordance of "Antrag" (Motion) in the ParlAT beta within the SketchEngine interface.

However, it is not the most suitable tool when using source text with complex xml markup and integrated links.

4 Intertextuality and referencing

One of the characteristics of parliamentary records is the complex intertextuality with a high frequency of cross-references. These cross references can refer to other parliamentary documents such as motions, reports etc. or they can refer to legal texts and legislation. Therefore, the annotation of these cross-references and linking to the external document is one of the prospects of this project. It has already been mentioned that, while the complex linking system would already be in place in the html documents, it is lost when processing the text in a corpus query system. For this reason, we are testing different components to configure a system where both approaches can be used in parallel, to establish and maintain the links to external documents and to visualise networks, for instance, but also to conduct linguistic queries on the material like in a corpus environment at the same time.

5 Discussion and further work

As stated at the beginning, we are building a corpus of Austrian parliamentary records for different user scenarios within linguistics, political science or history. The reported work is in progress. After finishing the speaker annotation a first version of the corpus – including the documents from 1996 to 2017 – will be published in ARCHE[3] and will made available through the CLARIN infrastructure.

In the second phase, we will start processing the pdf files and we will expand our work on the semantic annotation. Furthermore, we will look into the issue of cross referencing to external documents and to combine the two approaches into one interface: the linking to external resources and the corpus query paradigm.

[3] ARCHE (A Resource Centre for Humanities) is the depositing service of the Austrian Centre for Digital Humanities.

6 Acknowledgments

This work has been partly funded by the Nationalstiftung für Forschung, Technologie und Entwicklung in Österreich.

7 Bibliographical References

Calzada Pérez, María; Marín Cucala, Noemí; Martínez Martínez, José Manuel (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos. *Procesamiento del Lenguaje Natural* 37, 349-350.

Blätte, Andreas (2016). Integrationspolitik im Bundesländervergleich: Die Analyse thematischer Verknüpfungen von Integration auf Basis der PolMine-Plenarprotokollkorpora. Presentation at the the Forum CA[3].CLARIN-D, Hamburg. Available at https://www.clarin-d.de/images/forumca3/4_5_blaette_clarin_hamburg.pdf

Blessing, André; Kliche, Fritz; Heid, Ulrich; Kantner, Cathleen; Kuhn, Jonas. (2015). Computerlinguistische Werkzeuge zur Erschließung und Exploration großer Textsammlungen aus der Perspektive fachspezifischer Theorien. In: Baum Constanze; Stäcker, Thomas (Hrsg.). Grenzen und Möglichkeiten der Digital Humanities. (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). text/html Format. DOI: 10.17175/sb001_013

Calzada Pérez, María; Marín Cucala, Noemí; Martínez Martínez, José Manuel (2006). ECPC: European Parliamentary Comparable and Parallel Corpora / Corpus Comparables y Paralelos de Discursos Parlamentarios Europeos. *Procesamiento del Lenguaje Natural* 37, 349-350.

Fišer, Darja and Lenardič, Jakob (2017). Overview of Parliamentary Data and Corpora. Available at https://office.clarin.eu/v/CE-2017-1019-Parliamentary-data-report-version-2.pdf

Kilgarriff, Adam, Rychlý, Pavel, Smrž, Pavel and Tugwell, David (2004). The sketch engine. *Information Technology*. 105-116. Available at https://www.sketchengine.co.uk/wp-content/uploads/The_Sketch_Engine_2004.pdf

Marx, Maarten (2010). Advanced Information Access to Parliamentary Debates. *Journal of Digital Information* Vol 10 Nr. 6, Available at https://journals.tdl.org/jodi/index.php/jodi/article/view/668

Marx, Maarten and Schuth Anne (2010). DutchParl: A corpus of parliamentary documents in Dutch. In: Calzolari , Nicoletta et al. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 3680-3677.

Available at http://www.lrec-conf.org/proceedings/lrec2010/pdf/263_Paper.pdf

SketEngine (2017). Documentation. Available at https://www.sketchengine.co.uk/documentation/

Sippl, Colin; Burghardt, Manuel; Wolff, Christian; Mielke, Bettina (2016). Korpusbasierte Analyse Österreichischer Parlamentsreden. In: Jusletter IT, 25. Februar 2016.

8 Language Resource References

Alexander, Marc and Mark Davies. (2015) Hansard Corpus 1803-2005. Available online at http://www.hansard-corpus.org.

Lapponi, Emanuele and Søyland, Martin G. (2016). Talk of Norway. Available at https://github.com/ltgoslo/talk-of-norway (2016-10-29).

Wissik, Tanja and Pirker, Hannes (2018). ParlAT Corpus. Austrian Centre for Digital Humanities. https://id.acdh.oeaw.ac.at/parlat