

Turning the Carniolan regional assembly proceedings into an enriched historical corpus

Marolt, Matija

matija.marolt@fri.uni-lj.si
University of Ljubljana, Slovenia

Gašparič, Jure

Jure.Gasparic@inz.si
Institute of Contemporary History, Slovenia

Mundjar, Aleksander

am9973@student.uni-lj.si
University of Ljubljana, Slovenia

Kavčič, Alenka

alenka.kavcic@fri.uni-lj.si
University of Ljubljana, Slovenia

Fišer, Darja

Darja.Fiser@inz.si
Institute of Contemporary History, Slovenia

Pančur, Andrej

andrej.pancur@inz.si
Institute of Contemporary History, Slovenia

Introduction

Parliamentary proceedings reflect the political, societal, and cultural Zeitgeist of a certain period and are as such an invaluable source for a broad range of DH research questions (Blaxhill and Beelen 2016, Blätte et al. 2020, Müller-Hansen et al. 2021). While contemporary parliamentary corpora have been compiled, annotated and made available for a number of languages through projects such as ParlaMint (Erjavec et al., 2022), similar resources with historical data remain rare (Puren et al., 2022). They are essential, however, as in the past two decades the focus of research on parliamentary discussions has shifted from the history of politics to the history of the political. Conceptual history (Begriffsgeschichte) has become increasingly relevant for the understanding of the society and its processes: how certain political concepts have been introduced, used, and understood (Koselleck, 2006). Such in-depth historical analysis, connected with the speech act theory, is of major importance for political history. In this framework, language is not just a tool used for describing the world, but also a significant part of the world that it co-creates (Hampsher-Monk, 1998).

In this abstract, we present the development of a richly annotated corpus of historical proceedings of the Krainer Landtag (The Carniolan Regional Assembly), which was the highest legislative body of the autonomy of the Carniola Herzogtum of the Habsburg Empire. It was introduced with the February patent (February 26, 1861) when the Austrian part of the Empire entered the second constitutional period and ended with the onset of the first world war, comprising altogether 12 parliamentary terms. The Assembly, which was initially composed of 37 delegates and was in 1908 extended to 50 delegates, adopted laws that were within its jurisdiction, which included agriculture, public buildings, charitable institutions, municipal, church and school matters, matters related to the provision of harness, provision and accommodation of the military, the regional budget, and until 1873 appointed delegates to the central parliament in Vienna (Reichsrat).

Procedure

We collected the scanned and OCR processed pdf documents from the The Digital Library of Slovenia - dLib.si. The documents span the period from 1861 to 1913 (903 pdfs, 42746 pages) and include the assembly meeting proceedings (694 pdfs, 15353 pages), as well as supplementary materials (laws, budgets etc.). The documents are bilingual, in Slovenian and German, depending on the speaker. German was first typeset in the Gothic script and later on in Latin.

We first separated the meeting notes from the supplementary materials, as the latter often contain complex layouts (tables etc.), which will be addressed in our future work. We assessed the quality of OCR and found it to be acceptable on most documents (around 2% character error rate, a bit worse for German written in Gothic). Although there were some outlier pages where OCR performed poorly, especially when characters from neighbouring pages were visible in the scans, these pages were also included for further processing, but we plan to automatically detect such cases and perform OCR on them again with appropriate preprocessing.

On a subset of debates spanning 15 years, we devised a set of rules coded in a Python script to detect layout, find headings and text segments, deal with hyphenated words, assign text to speakers, detect events and extract meeting metadata including the date and hour of the beginning/end of each meeting, its agenda and a list of attendants with their titles. We used the Lingua Python library for language detection at sentence level. For lemmatization and part-of-speech tagging, we used Trankit¹ for both languages. We first tried the CLASSLA Slovenian fork² for lemmatization of Slovenian words and assessed PoS tagging accuracy of both lemmatizers. On a small sample of 3400 words (8 pages of text), CLASSLA performed only 0.5% better than Trankit. The lemmatization F1 scores were at 94.4% and 93.9%, respectively. The results are around 5% worse than for modern-day Slovenian, as the used Slovenian language is quite different (more archaic) to modern-day Slovenian on which the NLP tools were trained.

After fine-tuning the set of rules, we applied them to the entire corpus and made adjustments to handle the changes in document formatting that occurred during the years. Altogether, the processed documents contain over 44.000 segments, over 540.000 sentences (roughly 58% in Slovenian and 42% in German) and approximately 10M words. Earlier debates were predominantly in German, while the language shifted towards Slovenian in the latest years.

We output the documents in the modified TEI XML format according to the schema compliant with Parla-CLARIN (Erjavec

