Historical Parliamentary Corpora Viewer

Alenka Kavčič, Martin Stojanoski, Matija Marolt
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
alenka.kavcic@fri.uni-lj.si, ms7072@student.uni-lj.si, matija.marolt@fri.uni-lj.si

Abstract

Historical parliamentary debates offer a window into the past and provide valuable insights for academic research and historical analysis. This paper presents a novel web application tailored to the exploration of historical parliamentary corpora in the context of Slovenian national identity. The developed web viewer enables advanced search functions within collections of historical parliamentary records and has an intuitive and user-friendly interface. Users can enter search terms and apply filters to refine their search results. The search function allows keyword and phrase searching, including the ability to search by delegate and place names. It is also possible to search for translations of the text by selecting the desired languages. The search results are displayed with a preview of the proceedings and highlighted phrases that match the search query. To review a specific record, the full PDF document can be displayed in a separate view, allowing the user to scroll through the PDF document and search the content. In addition, the two corpora of Slovenian historical records integrated into the viewer—the Carniolan Provincial Assembly Corpus and the Parliamentary Corpus of the First Yugoslavia—are described and an insight into the corresponding preparation processes is provided.

Keywords: parliamentary corpora, web application, Parla-CLARIN, Carniolan Provincial Assembly, National Representation of the First Yugoslavia

1. Introduction

Parliamentary debates have long been a valuable resource of research data, as they are systematically prepared and reflect the state of society at the time. They cover a wide range of topics and are an important source for historians as well as for scholars from diverse fields such as political science, sociology, economics and linguistics.

While contemporary parliamentary corpora, such as those produced in projects like ParlaMint (Erjavec et al., 2022), are widely available and well-structured, historical records are often available in less accessible forms, such as PDFs or unstructured text. This poses a major challenge for effective queries and large-scale analysis and limits the use of these resources.

Although digitization has made it easier for researchers to access the historical records by enabling keyword searches and remote access to the data, advanced search and analysis tools are needed to enhance research capabilities and facilitate exploration of the data. These tools include keyword search, advanced search filters, text mining algorithms, sentiment analysis, topic modeling, and visualization techniques.

2. Historical Parliamentary Corpora

Historical parliamentary records in digital format are still a rarity and online access and exploration even rarer. However, there are notable exceptions that provide better access to transcriptions of past parliamentary debates. A good example of this is the digitized collection of lower house parliamentary debates in the French parliament from 1881 to 1940, accessible via the digital repository of the France National Library (Gallica[1]). This corpus, carefully curated as part of the AGODA project (Puren et al., 2022), has undergone extensive processing, including OCR, annotation and semantic enrichment, making it easily accessible for scholarly research.

Another notable example is the Historical Hansard[2] corpus, which contains transcriptions of speeches and debates in the UK Houses of Lords and the Commons from 1803 to the present day (Coole et al., 2020). Hansard has traditionally been published in printed form since the early 19th century. The older volumes of the collection have been digitized and enhanced with metadata, including tokenization, part-of-speech tagging and semantic annotations, to form a comprehensive corpus. In addition, a web-based search interface has been developed that allows linguistic queries to be performed on the corpus while offering visualizations that provide a deeper understanding of the evolution of political discourse over time.

Additional example is Congress.gov[3], which offers numerous ways to access the US Congressional Records from 1873 (43rd Congress) to the present day. Users can select records by specific

[1] https://gallica.bnf.fr/ark:/12148/cb328020951/date.item
[2] https://hansard.parliament.uk/
[3] https://www.congress.gov/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 127–132

dates and examine complete issues and all sections. The platform also facilitates keyword and phrase searches, with the option to refine results using various filters.

2.1. Slovenian Parliamentary History

The beginnings of Slovenian parliamentary history date back to the mid-19th century, and a large part of the parliamentary debates from this period have been digitized and made available in PDF format. While these digitized archives are invaluable repositories of historical knowledge, they also present significant challenges in terms of readability (due to archaic language or poor reproduction quality), completeness (due to gaps or omissions in the records), and biases associated with the reporting and recording processes, so a critical eye is required when interpreting these materials.

Beyond mere digital preservation, the efficiency of data exploration depends on the enrichment of resources with metadata and the provision of structured, annotated content. Enriching these digitized archives with comprehensive metadata facilitates efficient retrieval and categorization, while structured and annotated content improves interpretability and enables a more complex and systematic study of Slovenian parliamentary history.

In this section, we present two newly created historical corpora from the present-day territory of Slovenia.

2.2. Carniolan Provincial Assembly

The Carniolan Provincial Assembly (*Kranjski deželni zbor* in Slovenian or *Krainer Landtag* in German) was the highest legislative body of Duchy of Carniola, which was a hereditary land of the Habsburg monarchy and a part of Austrian Empire (from 1867 Austro-Hungarian Empire). The Carniolan Provincial Assembly was introduced with the February patent, a constitution of the Austrian Empire proclaimed on 26 February 1861. After 12 parliamentary terms, it ended with the onset of the First World War. A unicameral assembly consisted of 37 members (in 1908 the number was increased to 50) and was chaired by the provincial governor (*deželni glavar* or *Landeshauptmann*) who was appointed by the Emperor from among the members. The Carniolan Provincial Assembly passed laws that were within the province's jurisdiction, including educational, municipal, ecclesiastical and military matters, and issues of provincial importance (e.g. agriculture, culture, public buildings, public construction works, various economic matters, charity institutes).

The parliamentary meeting proceedings from 1861 to 1913 are available in the Carniolan Provincial Assembly corpus Kranjska 1.0 (Kavčič et al., 2023a). The corpus covers 694 sessions, with two documents for each parliamentary session: one in Parla-CLARIN compliant TEI XML format (see section 2.4) and a corresponding facsimile in PDF format (an example of the PDF facsimile is shown in Figure 2).

The documents are mostly bilingual; 58% of sentences are in Slovenian and 42% in German language. The XML documents together include over 44 thousand utterances, over 540 thousand sentences and approximately 10 million words that are also linguistically annotated (tokenisation, part-of-speech tagging and lemmatisation were used).

2.3. National Representation of the First Yugoslavia

First Yugoslavia refers to the Yugoslav state between the two world wars: the Kingdom of Serbs, Croats, and Slovenes, established after the collapse of the Austro-Hungarian Empire in 1918, and renamed Kingdom of Yugoslavia in 1929. In the newly formed Kingdom of Serbs, Croats, and Slovenes, a joint government and a Temporary National Representation were established. The latter performed the functions of Parliament from March 1919 to October 1920, when the elections to the Constituent Assembly were called. Its 296 delegates were not elected, but appointed. First parliamentary elections were held in November 1920, when 419 delegates were elected to the Constituent Assembly. There have been seven elections altogether in that time (i.e. between the two world wars): besides 1920, also in 1923, 1925 and 1927 (because of political instability every 2 years, although the terms lasted 4 years), and in 1931, 1935 and 1938. There were no elections during the dictatorship from January 1929 to September 1931, as the National Assembly was abolished at that time. In 1931, a bicameral system was introduced with the new constitution: National Representation consisted of Senate and National Assembly. Parliament's autonomy was very limited due to the great authority of the King, who according to the constitution controlled all three branches of government, including the legislative. The parliament had general passive and active suffrage, decided on laws, and on amendments to the constitution, while the King had the power, among the others, to suspend the law, and to summon and dissolve the Parliament.

The parliamentary meeting proceedings from 1919 to 1939 are available in the Parliamentary corpus of first Yugoslavia yu1Parl 1.0 (Kavčič et al., 2023b), covering proceedings in the three periods:

- Temporary National Representation of the Kingdom of Serbs, Croats, and Slovenes (1919-1920);

- Legislative Committee of National Assembly of the Kingdom of Serbs, Croats, and Slovenes (1921-1922);
- National Representation (National Assembly and Senate) of the Kingdom of Yugoslavia (1931-1939).

The meeting proceedings of the National Assembly of the Kingdom of Serbs, Croats, and Slovenes between years 1923 and 1928 are not (yet) available in digital form and therefore not included in the corpus.

The corpus comprises 714 sessions, where each session is available in two documents of different formats: Parla-CLARIN compliant TEI XML and a corresponding facsimile in PDF.

The documents are multilingual, in Slovenian (3% of sentences) and Serbo-Croatian. The latter is typeset in the Cyrillic (Serbian, 59% of sentences) or in the Latin (Croatian, 38% of sentences) alphabet. The XML documents together include over 34 thousand utterances, 578 thousand sentences and approximately 13 million linguistically annotated words, where words in Cyrillic script (Serbian) have lemmas in Latin script.

2.4. Parla-CLARIN TEI Format

There are various ways to annotate the content, but the TEI guidelines are the de facto standard for encoding text in the digital humanities (TEI Consortium, 2019). For parliamentary corpora, the Parla-CLARIN Guidelines (Erjavec and Pančur, 2021) were developed as a common TEI-based annotation scheme.

The preparation of the corpus started with the scanned images of the meeting proceedings. The scanned documents were OCR processed and automatically parsed with rule-based Python scripts to extract the metadata and annotate the speeches.

For structuring the session data, we used a subset of the Parla-CLARIN tags (Erjavec and Pančur, 2022) that were best suited for our content, the multilingual historical parliamentary debates.

Each meeting proceeding has a unique identifier based on the content of the proceeding and the date of the session. An XML file consists of a header and a body. The header contains the metadata of the file: title in several languages (English, Slovenian and other main languages used in the proceedings), information about the publisher, publication date, link to the related PDF files, etc. The body is parsed from the content: it starts with the title of the session, information about the delegates present, the agenda and the starting time of the session. This is followed by the individual sections dealing with the speeches of the individual delegates. Each section is labelled with the name of the speaker, followed by the content of the speech, which may also include comments or events during the speech (i.e. described events in the room such as "laughter from the left" or "reads"). The speech is divided into sentences, and these in turn are divided into words and punctuation. As the transcripts are multilingual, the language is marked for each sentence. The words are also linguistically annotated. The end time of the session is noted at the end.

The linguistic annotation in both corpora included tokenisation, MSD tagging and lemmatisation. Since the corpora contain different languages, different tools were used for the linguistic annotation in each corpus. The languages of the Carniolan Provincial Assembly were German and Slovene, so Trankit[4] was used as it works well for both languages. The National Representation of the First Yugoslavia, on the other hand, includes Slavic languages (Slovenian and Serbo-Croatian), which is why we opted for CLASSLA[5] (Ljubešić and Dobrovoljc, 2019; Terčon and Ljubešić, 2023).

3. Historical Parliamentary Corpora Viewer

To make the historical parliamentary session proceedings accessible to a wider audience that may not be proficient in parsing TEI encoded files, we developed the Historical Parliamentary Corpora Viewer. The Viewer is a web application that supports searching over collections of multilingual proceedings of historical parliamentary sessions. It allows the user to search for texts across languages and limit the results to certain speakers or place names. It is also possible to filter the results by language, date of sessions and to sort the results by date or relevance.

The two corpora described in sections 2.2 and 2.3 are currently included in the Viewer. If further parliamentary proceedings are scanned and prepared so that they are available as PDF and TEI XML documents, they can be added to the Viewer as an additional corpus.

3.1. Technical Details

The web application consists of three parts: frontend, backend and database. The frontend runs on a client device and implements the user interface. It sends requests to the backend and receives the data to be displayed to the user. The frontend was developed in the Vue.js JavaScript framework and uses the HTTP protocol for communication between the client and the server.

[4] https://github.com/nlp-uoregon/trankit
[5] https://github.com/clarinsi/classla

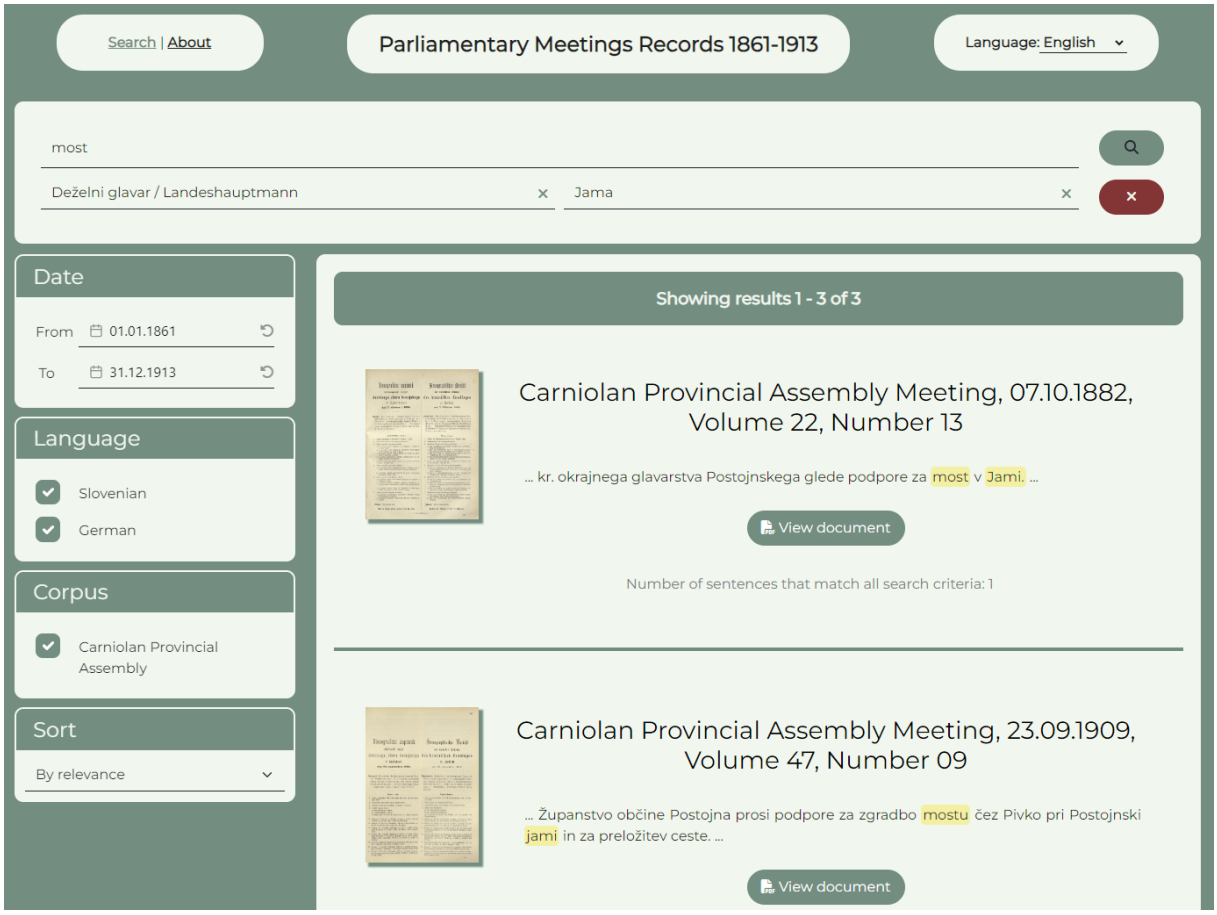

Figure 1: Searching the corpora. The search fields are at the top, the filters for narrowing down the results are on the left, while the search results are displayed in the main part of the page.

The backend runs on a Node.js server was implemented with Express.js. It offers a RESTful API and thus complies with the specifications of the REST architecture. The backend communicates with the database via the HTTP protocol. Elasticsearch[6], a RESTful search and analytics engine was used as the database, enabling fast unstructured text searches.

3.2. User Interface

The user interface is minimalist, intuitive and easy to use. It is divided into two parts: a page for searching and browsing the parliamentary proceedings and a page that displays a facsimile of the selected proceedings and allows search within the proceedings, as well as the display of PDF (OCRed) text transcription and its translations into target languages.

3.3. Searching the Corpora

The page for searching in the corpora is shown in Figure 1. The user can enter search terms and/or set specific filters for the search results.

If no search parameters are entered, all the proceedings in the corpus are displayed, so that the user can browse the collection.

By default, the keyword search finds all proceedings that contain all the words in the search query. The search terms can also be separated with OR operator to search for documents that contain at least one of the specified search words. It is also possible to search for a phrase by enclosing the phrase in quotation marks. The search in translations of the text can be activated by selecting the desired languages in the filters. If the search word is entered in its basic form (lemma), the search will also find all other forms of the word in the text.

Several filters are available to limit the search results: date, language and corpus (shown on the left in Figure 1). Restricting the search results by date is an important filter for historical documents, as it limits the search results to the parliamentary sessions within a selected time period. Without this filter, the search is applied to all documents contained in the selected corpora. As the documents are multilingual, it is also possible to use a language filter and search for keywords in all languages (i.e.

[6] https://www.elastic.co/

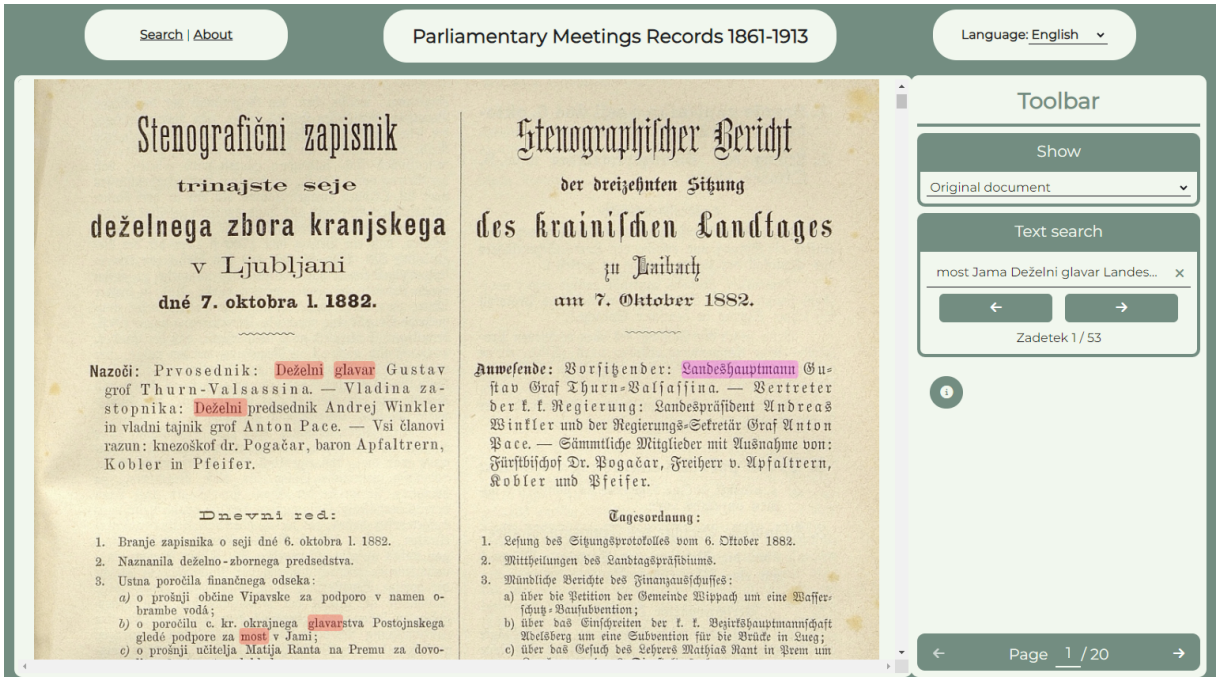

Figure 2: The viewer for PDF documents, showing an excerpt from the PDF facsimile of the Carniolan Provincial Assembly session dated 7.10.1882. Words from the query are highlighted.

also in the translations of the texts).

The search results are displayed with the proceedings preview and some sentences with the highlighted words corresponding to the query. The "Show document" button displays the PDF facsimile of the proceedings in an integrated PDF viewer.

3.4. Searching the PDFs and Transcriptions

If users want to inspect specific parliamentary proceedings, they can display the document in a separate view that allows them to scroll through the PDF and search within the PDF content. Figure 2 shows the PDF viewer with the search bar on the right.

Instead of a facsimile of the document, the user can display a transcript in all the main languages used in the parliamentary proceedings of a particular corpus (e.g. German and Slovene in the case of the records of the Carniolan Provincial Assembly), which also supports the content search.

4. Conclusion

The development of the web viewer for exploring the Slovenian historical parliamentary corpora represents a significant step forward in terms of the availability of historical resources for researchers and enthusiasts alike. The intuitive user interface caters to diverse users, from students to experienced scholars, and enables seamless navigation and exploration of the invaluable historical documents. By bridging the gap between academia and the public, this application not only enhances scholarly research, but also promotes a deeper understanding and appreciation of Slovenian parliamentary history.

Since the web viewer is designed to display corpora in a Parla-CLARIN TEI compatible format, the integration of new corpora is a straightforward process. In the future, we plan to gradually expand our dataset by integrating additional corpora. We are aiming for comprehensive coverage of parliamentary debates from the mid-19th century to the present day.

5. Acknowledgements

The work was supported by the Slovenian Research Agency research programme P6-0436: Digital Humanities: resources, tools and methods (2022–2027).

6. Bibliographical References

Matthew Coole, Paul Rayson, and John Mariani. 2020. Unfinished business: Construction and maintenance of a semantically tagged historical parliamentary corpus, UK Hansard from 1803 to the present day. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 23–27, Mar-

seille, France. European Language Resources Association.

Tomaž Erjavec and Andrej Pančur. 2021. The parla-clarin recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative*, 14.

Tomaž Erjavec and Andrej Pančur. 2022. Parla-clarin: A tei schema for corpora of parliamentary proceedings. https://clarin-eric.github.io/parla-clarin/. Accessed: 2024-02-08.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darǵis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Marie Puren, Pierre Vernus, Aurélien Pellet, Nicolas Bourgeois, and Fanny Lebreton. 2022. Extracting and providing online access to annotated and semantically enriched historical data. The AGODA project. In *DH Benelux 2022*, Luxembourg, Luxembourg.

TEI Consortium. 2019. TEI P5: Guidelines for electronic text encoding and interchange. http://www.tei-c.org/Guidelines/P5/. Accessed: 2024-02-08.

Luka Terčon and Nikola Ljubešić. 2023. Classla-stanza: The next step for linguistic processing of south slavic languages.

7. Language Resource References

Kavčič, Alenka and Mundjar, Aleksander and Marolt, Matija. 2023a. *Carniolan Provincial Assembly corpus Kranjska 1.0*. Faculty of Computer and Information Science, University of Ljubljana. PID http://hdl.handle.net/11356/1824. Slovenian language resource repository CLARIN.SI.

Kavčič, Alenka and Mundjar, Aleksander and Marolt, Matija. 2023b. *Parliamentary corpus of first Yugoslavia (1919-1939) yu1Parl 1.0*. Faculty of Computer and Information Science, University of Ljubljana. PID http://hdl.handle.net/11356/1845. Slovenian language resource repository CLARIN.SI.