Digital Scholarship in the Humanities, 2025, **00**, 1–11
https://doi.org/10.1093/llc/fqaf068
DSH 2023 Conference Supplement

Historical parliamentary corpora: the Carniolan provincial assembly records

Alenka Kavčič [1,*], Darja Fišer [2], Andrej Pančur [2], Matija Marolt [1]

[1]Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia
[2]Institute of Contemporary History, Privoz 11, 1000 Ljubljana, Slovenia

*Corresponding author. Faculty of Computer and Information Science, University of Ljubljana, Večna pot 113, 1000 Ljubljana, Slovenia.
E-mail: alenka.kavcic@fri.uni-lj.si

Abstract

Parliamentary debates provide an invaluable source of information for research in various fields, as they reflect the prevailing ideas and beliefs of the time. We have developed a corpus of Slovenian parliamentary records from 1861–1913, covering 694 sessions of the Carniolan Provincial Assembly and containing about 10 million words. This article describes the process of creating the first version of the corpus, which consisted of obtaining OCR processed stenographic notes of the parliamentary sessions, analysing the content to extract metadata, and developing a linguistically processed, metadata-rich and structurally encoded corpus in the Parla-CLARIN format. Transforming these records into more accessible and analysable resources will facilitate in-depth study and analysis of the historical development and change of political concepts and ideas. In addition, a brief analysis of the corpus vocabulary is provided, focusing on the use of Slovenian and German language. As language played a key role in shaping ethnic identity during this period, the results are also discussed in their historical context. To make the corpus accessible to the public, we have developed a web application that facilitates exploration of the corpus and enables efficient searching in an intuitive and user-friendly way.

Keywords: parliamentary corpora; Carniolan Provincial Assembly; Parla-CLARIN; multilingual parliamentary debates; quantitative linguistic analysis.

1. Introduction

Parliamentary debates reflect the political, societal, and cultural ideas and beliefs of their time and represent an invaluable source of information for a wide range of research questions in the digital humanities (Blaxill and Beelen 2016; Blätte, Gehlhar, and Leonhardt 2020; Müller-Hansen *et al.* 2021). While contemporary parliamentary corpora for a number of languages have been compiled, annotated, and made available through projects such as ParlaMint (Erjavec *et al.* 2023), similar resources with historical data are still rare (Puren *et al.* 2022). However, they are indispensable as in the last two decades the focus of research on parliamentary debates has shifted from the history of politics to the history of the political. Conceptual history or *Begriffsgeschichte*, which examines the historical development and transformation of concepts and ideas over time, has become increasingly important for understanding society and its processes. It explains how certain political concepts were introduced, used, and understood within their specific historical contexts (Koselleck 2006). Such in-depth historical analysis, especially in conjunction with speech act theory, is of great importance for political history. In this framework, language is not only a tool for describing the world, but also an essential part of the world it helps to create (Hampsher-Monk 1998).

The study of Slovenian parliamentary history has been notably constrained by the lack of comprehensive corpora, hindering in-depth research and analysis. While contemporary parliamentary records are well documented and readily available in corpora such as siParl (Pančur *et al.* 2024), which cover the period 1990–2022, there is a pressing need to develop and expand resources with historical parliamentary data to enable comprehensive research. To address this deficit, we have developed a new corpus that includes the older parliamentary records from the period 1861–1913. In doing so, we are filling this critical gap and providing a valuable resource for future research on Slovenian political history.

1.1 Related work

Historical parliamentary records have predominantly been available as unstructured documents, which

considerably limits their usability and poses a challenge for extensive queries and analyses. In recent years, however, there have been concerted efforts to create structured, standardized datasets to overcome these challenges. We focus here exclusively on historical records, particularly those dating back more than 100 years. All of these structured datasets are available in machine-readable formats and represent an important step towards making historical parliamentary records more accessible and analysable for researchers.

The Swedish Parliament Corpus (Yrjänäinen *et al.* 2024) is an excellent example, compiling and structuring records from 1867 to 2022 in a unified corpus with a standardized data format, including comprehensive metadata about the members of parliament. Similarly, the Italian parliamentary historical documents from 1848 to the present (Frasnelli and Palmero Aprosio 2024) have been compiled into a digitally readable, structured corpus containing speeches, reports, and legislative proposals, providing a detailed overview of 175 years of Italian legislative history.

Other advances include the Finnish Parliamentary Corpus (1907–2022) (Drobac, Sinikallio, and Hyvönen 2023), which contains bilingual debates (Finnish and Swedish). The Polish Parliamentary Corpus (Ogrodniczuk and Nitoń 2020) offers a comprehensive collection of transcripts of Polish parliamentary proceedings dating back to 1919. In addition, the German Parliamentary Corpus (GerParCor; Abrami, Bagci, and Mehler 2024) is the largest German-language corpus for parliamentary transcripts, comprising historical transcripts from Austria, Germany, Liechtenstein, and Switzerland, with documents dating back to 1797.

The collection Minutes of the Council of Ministers of Austria and the Austro-Hungarian Monarchy 1848–1918 is closely related to our work and is available as a high-quality full text in the TEI-based digital edition (Kurz 2024) in addition to the printed volumes. In particular, the five volumes of Series 3, The Minutes of the Cisleithanian Council of Ministers 1867–1918, which comprise meetings of the central governing body of the Austrian part of the Dual Monarchy, can be regarded as a supplementary data source that enables a more comprehensive understanding of the events of that period.

Another exemplary project is the digitized collection of parliamentary debates in the French Parliament from 1881 to 1940, which is available through the National Library of France's digital repository[1] and was compiled, annotated, and semantically enriched as part of the AGODA project (Puren *et al.* 2022). The Historical Hansard[2] corpus (Coole, Rayson, and Mariani 2020) is also an important resource, which contains transcripts of speeches and debates in the British House of Lords and House of Commons from 1803 to this day. The older volumes of the Hansard, originally published in print since the early 19th century, have been digitized and metadata added. Similarly, Congress.gov[3] provides access to the records of the US Congress dating back to 1873.

2. Slovenian historical parliamentary corpus

2.1 Carniolan provincial assembly

The Duchy of Carniola (*Vojvodina Kranjska* in Slovenian, *Herzogtum Krain* in German) was a historical region in Central Europe, located in present-day Slovenia. It was founded in 1364 as a duchy within the Holy Roman Empire under the control of the Habsburg dynasty, became part of the Austrian Empire and later, in 1867, was incorporated into the Cisleithanian territories of the Austro-Hungarian Empire. The February Patent, a new constitution of the Austrian Empire promulgated on 26 February 1861, introduced the Carniolan Provincial Assembly (*Kranjski deželni zbor* in Slovenian, *Krainer Landtag* in German), which was the highest legislative body of the Duchy of Carniola (Vilfan 1961).

The Carniolan Provincial Assembly was active for twelve legislative periods (terms) and ceased its activities with the dissolution of the Austro-Hungarian Empire at the end of the First World War. The mandate of the provincial assembly lasted six years, but there were several shorter mandates in the first decade (see Table 1), due to the invalid elections and political instability (Vilfan 1961).

Initially, this unicameral assembly consisted of thirty-seven members elected by four separate voting groups on the basis of social or economic status: ten delegates from large landowners (the curia of great landowners), eight delegates from representatives of towns and markets (urban curia), sixteen delegates from other municipalities (rural curia), and two delegates from the chambers of commerce (Vilfan 1961; Taylor 1976). The Bishop of Ljubljana was also a member, as a virilist. The assembly was headed by the provincial governor (*deželni glavar* in Slovenian or *Landeshauptmann* in German), a position appointed by the Emperor from among the members of the assembly. After the reform of 1908, two more delegates were added to the urban curia and the general curia was introduced with eleven delegates, increasing the number of members to fifty (Vilfan 1961).

The electoral system was complex and heavily favoured property, wealth, and the towns. Its unique feature was its manipulation in favour of the Germans (Taylor 1976; Štih, Simoniti, and Vodopivec 2008). Although the Duchy of Carniola was predominantly

Table 1. Number of documents, pages, sentences, words (tokens), and different words by lemma (types) in the corpus. The numbers are provided separately for each legislative period (term) and in total for the entire corpus. The start of the mandate and its duration are given for each legislative period. The number of sessions is presented as the number of documents with the number of pages. The number of sentences is given separately for Slovenian (sl) and German (de) sentences. The same applies to the number of tokens (count of all words) and the number of types (count of different words by lemma).

Term	Start date	Duration	#Docs	#Pages	#Sentences		#Tokens		#Types	
					sl	de	sl	de	sl	de
1	April 1861	6 years	113	1,981	2,493	67,781	45,076	1,602,451	6,343	70,461
2	February 1867	15 days	7	66	996	1,509	21,133	28,575	3,170	4,167
3	April 1867	3 years	44	871	10,136	18,125	211,755	381,128	12,813	29,754
4	July 1870	18 months	17	148	2,800	1,654	55,385	36,455	6,266	5,608
5	December 1871	6 years	78	1,180	16,417	20,608	329,851	509,828	18,155	29,955
6	July 1877	6 years	55	1,200	12,771	23,092	249,730	588,265	15,192	33,755
7	June 1883	6 years	103	1,929	38,183	20,170	824,716	499,271	30,521	53,109
8	July 1889	6 years	92	2,467	61,343	19,045	1,308,750	470,040	38,407	26,604
9	November 1895	6 years	103	2,800	63,431	24,349	1,347,117	572,390	38,813	33,951
10	September 1901	6 years	30	616	17,640	3,402	341,274	61,095	17,306	8,765
11	March 1908	6 years	52	2,095	62,714	13,572	1,163,028	236,306	36,033	18,655
Total			694	15,353	288,924	213,307	5,897,220	4,985,178	108,368	185,904

Slovene—92 per cent of the population according to the 1846 census, albeit mainly in rural areas (Zwitter 1967)—the curia of great landowners, which made up a quarter of the delegates, remained entirely German. The chambers of commerce were, by nature, in German hands, as was the urban curia, since the Germans were the wealthy and urban nationality (Taylor 1976).

In its capacity as legislator, the Carniolan Provincial Assembly was vested with the power to enact laws that were not reserved for the Imperial Council (*Reichsrat* in German), the central parliament in Vienna. Until 1873, it appointed delegates to the Imperial Council in Vienna (Cvirn 2006). As an electoral body, it was more significant than ever before (Taylor 1976).

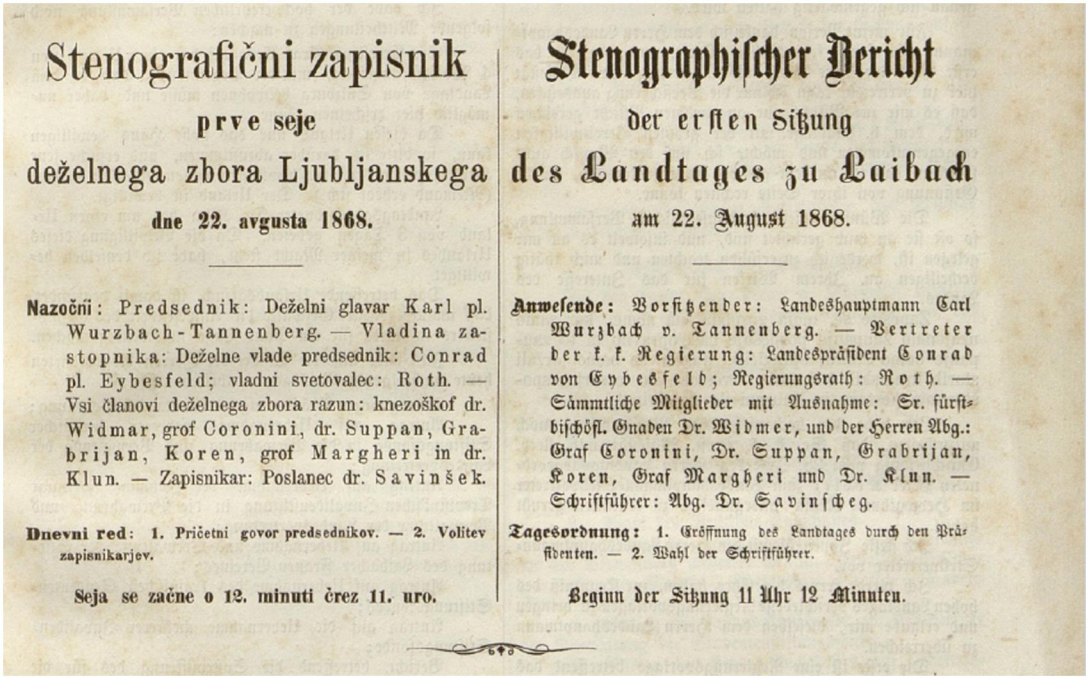
Although the provincial assembly was the highest legislative body of the province, its powers were limited and of a more administrative nature. It only concerned matters of provincial importance, such as agriculture, public buildings (management and control of buildings financed from provincial funds), and charitable institutions. In addition, the assembly passed laws on the provincial budget and various economic matters. It was able to enact special laws for the province, including for education, municipal administration, ecclesiastical matters, and military affairs within the province.

In practice, however, provincial legislation enjoyed considerably less independence under the Austrian system than stated in the constitution, as every bill passed by the provincial assembly required the approval of the government. Proposals that conflicted with government policy were routinely rejected. As a result, many of the bills proposed by the Carniolan Provincial Assembly in the 1860s remained unconfirmed, such as the 1868 proposal on the regulation of the language of instruction in primary and secondary schools (Melik 1969).

2.2 Source data description and preparation

The printed volumes of the Stenographic Records of the Carniolan Assembly Meetings, which cover the period from January 1863 to October 1913, have already been scanned and OCR processed. The PDF documents are available in the Digital Library of Slovenia.[4] We have collected the documents (903 PDFs, 42,746 pages), which contain both the minutes of the sessions and supplementary materials such as laws, budgets, etc. We have manually separated the meeting minutes from the supplementary materials, as the latter often contain complex layouts (e.g. tables), which we will deal with in our future work. In addition, the scanned images of the stenographic notes from 1861, which were missing from the collected documents, were acquired, OCRed with ABBYY FineReader,[5] and added to the collection, resulting in additional nine meeting notes. Our source data thus comprised 694 meeting notes (from the first to the eleventh parliamentary term), each in a separate PDF file, totalling 15,353 pages.

The documents are bilingual, in Slovenian and German. In most of the documents, Latin script is



Stenografični zapisnik

prve seje

deželnega zbora Ljubljanskega

dne 22. avgusta 1868.

Nazočni: Predsednik: Deželni glavar Karl pl. Wurzbach-Tannenberg. — Vladina zastopnika: Deželne vlade predsednik: Conrad pl. Eybesfeld; vladni svetovalec: Roth. — Vsi članovi deželnega zbora razun: knezoškof dr. Widmar, grof Coronini, dr. Suppan, Grabrijan, Koren, grof Margheri in dr. Klun. — Zapisnikar: Poslanec dr. Savinšek.

Dnevni red: 1. Pričetni govor predsednikov. — 2. Volitev zapisnikarjev.

Seja se začne o 12. minuti črez 11. uro.

Stenographischer Bericht

der ersten Sitzung

des Landtages zu Laibach

am 22. August 1868.

Anwesende: Vorsitzender: Landeshauptmann Carl Wurzbach v. Tannenberg. — Vertreter der k. k. Regierung: Landespräsident Conrad von Eybesfeld; Regierungsrath: Roth. — Sämmtliche Mitglieder mit Ausnahme: Sr. fürstbischöfl. Gnaden Dr. Widmer, und der Herren Abg.: Graf Coronini, Dr. Suppan, Grabrijan, Koren, Graf Margheri und Dr. Klun. — Schriftführer: Abg. Dr. Savinscheg.

Tagesordnung: 1. Eröffnung des Landtages durch den Präsidenten. — 2. Wahl der Schriftführer.

Beginn der Sitzung 11 Uhr 12 Minuten.

Figure 1. The PDF facsimile of the proceedings of 22 August 1868, showing an excerpt from the first page, on which the title of the session, the participants, the agenda, and the beginning of the session are transcribed in two languages, Slovenian in the left-hand column and German in the right-hand column. Gothic script is used for German. An ornament at the bottom separates the title section from the transcriptions of the debates.

used, although in the earlier documents the German was typeset in Gothic script. The documents are divided into two main sections: the title section and the debate section. The documents begin with the title section, which contains the name and date of the session, the participants, the agenda, and the time of the start of the session. The title section in the early stenographic records was monolingual and in a single column, predominantly in German, with Slovenian appearing from 1867 onwards. This change corresponded to the implementation of the articles on the equality of nations, which made it possible to recognize Slovenian as an official language in all regions of the Austro-Hungarian Empire in which the Slovenian population predominated. In 1868, the title section was changed to a bilingual format with two columns, in which both languages (German and Slovenian) were used in parallel. Figure 1 shows an example of a bilingual title section.

The title section usually ends with an ornament and is followed by the debate section. The debate section is in two columns (also in earlier documents) and contains the transcription of the parliamentary debates (mainly speeches by the delegates). The language of the delegate's speech depends on the speaker and is only transcribed in the original language (it is not translated in the transcriptions). The content of the earlier stenographic recordings was exclusively in German. Although the Carniolan Provincial Assembly stipulated in its rules of procedure in 1861 that delegates' speeches should be recorded in the language in which they were delivered, this rule was not consistently enforced (Novak 2007). Besides, the delegates mostly used the German language in the first electoral term. From 1867 onwards, the provincial assembly took on a more Slovene character as the Slovene delegates began to routinely speak in Slovenian (Melik 1969).

2.3 Corpus preparation and metadata extraction

First, we assessed the quality of the OCR. We randomly selected forty-five pages from the stenographic records in different time periods (about 0.3 per cent of all pages) and manually corrected the optically recognized text. We then calculated the character error rate (CER) for these documents and found that it was around 2 per cent for most documents (but slightly higher for the Gothic script), which is acceptable. However, there were some outlier pages where the OCR performed poorly, especially when characters from neighbouring pages were visible in the scans. Nevertheless, the CER for the outliers was below 5 per

cent, with the exception of two pages with values of 6.08 per cent and 25.81 per cent. The latter had an even worse value of 34.11 per cent for the German text in Gothic script. Manual examination of the problematic page revealed that the poor OCR performance was due to old, yellowed and stained paper, unclear printing (possibly due to scanning resolution), and, most importantly, text showing through from the flip side of the page. The page was also slightly creased during scanning. Despite this anomaly, we considered all of these outlier pages for further processing. In the future, however, we plan to automatically detect such outliers and perform OCR again on these pages with suitable pre-processing.

To process the documents, we started with a small subset of fifteen years of debates. We developed a set of rules, coded in a Python script, to detect the layout and extract the metadata.

We faced several challenges when creating the rule-based scripts. First, the quality of the OCR was variable, especially for the Gothic script, where some letters (e.g. *s*, *f*, *t*, and *k*) were difficult to distinguish and often misinterpreted. Text presented in tabular form was generally not recognized as a table and was also frequently misinterpreted. In addition, some ornaments on the pages were recognized as a series of different characters that were difficult to identify using the established rules. We have deferred these challenges for future work to improve OCR. Then there were the changes in layout, from single column to double column, which also varied over the years. The design of the agenda and attendance list also changed over the years, as did the use of language in the header (from a single language, usually German, to both languages in parallel).

After fine-tuning the rule set, we applied it to all the documents and made adjustments to account for the changes in the formatting of the documents that occurred over the years.

The process was carried out in four stages. First, the layout of the title section was detected, and the content extracted separately for each language, resolving inconsistencies in layout across different transcriptions (early stenographic records had a single-language, single-column format, while the format later changed to a dual-language, two-column layout). Subsequent parsing involved identifying session metadata, such as the session date, the time the session started, the agenda, and a list of participants with their titles, taking into account different styles and formats (e.g. the way agenda items were separated and the use of Roman or Arabic numerals).

The second stage involved parsing the text in the debate section, which included removing headers and page numbers, detecting headings and text segments, managing hyphenated words, identifying speakers and assigning the text to the speakers, and extracting the end time of the session.

The third stage focused on parsing individual segments into sentences, detecting the language of each sentence using the Lingua Python library,[6] and distinguishing events occurring during the sessions (e.g. *reads*, *applause on the left*) to ensure accurate annotation of the speeches. Each segment was also linguistically annotated (tokenization, part-of-speech tagging, and lemmatization). For lemmatization and part-of-speech tagging, we used Trankit[7] (Nguyen *et al.* 2021) for both languages. Initially, we also experimented with the CLASSLA Slovenian fork[8] (Ljubešić and Dobrovoljc 2019; Ljubešić, Terčon, and Dobrovoljc 2024) for lemmatization of Slovenian words and assessed the PoS tagging accuracy of both lemmatizers. On a sample of eight pages of text (3,400 words), CLASSLA outperformed Trankit by only 0.5 per cent. The lemmatization F1 scores were 94.4 per cent for CLASSLA and 93.9 per cent for Trankit. These results are about 5 per cent lower than for modern Slovenian, reflecting the challenges posed by the more archaic language used in the texts, which differs from the modern Slovenian on which the NLP tools were trained. As CLASSLA is limited to the Slovenian language and Trankit offers nearly equivalent performance while supporting multiple languages, we decided in favour of Trankit.

The final stage involved preparing the XML files in the Parla-CLARIN compliant format, ensuring that the data was properly annotated.

2.4 Parla-CLARIN compliant XML

The XML files are coded in the Parla-CLARIN compliant format (Erjavec and Pančur 2021; Erjavec and Pančur 2022), which follows the Text Encoding Initiative (TEI) guidelines (TEI Consortium 2023) and was specially designed for annotation of parliamentary debates.

The corpus XML document has a `<teiCorpus>` root element, which contains `<teiHeader>` element with the corpus metadata (i.e. for the corpus as a whole) and a series of links to `<TEI>` elements, one for each corpus component (one session, i.e. one PDF document) that is described in a separate XML document. The metadata of the corpus applies to the entire corpus and includes a bibliographic description of the file (e.g. title, edition, publisher, licence, the sources from which the corpus was generated, and the number of documents, utterances, sentences, and words), a description of the encoding (including the applications used), and the languages represented in the text.

Each parliamentary session (PDF document) is described in a separate XML document with a <TEI> root element and two sub-elements: <teiHeader> with the metadata of this specific session and <text> with the text of the document. The metadata in the header contains the session title (i.e. session, date, proceedings volume, and number), the source description (i.e. the corresponding PDF facsimile), the publication data, and the licence. The text part contains only the mandatory <body> element, which contains the transcription of the session, divided into sections (elements). The main section of the debate (element) starts with the title of the meeting as parsed from the stenographic records (<meeting> element), the list of participating delegates (<list type="attendance"> element), the list of agenda items (<list type="agenda"> element), and the time at which the meeting started (<note type="time"> tag) (see Fig. 2a). A similar time tag marks the time at which the meeting ended and is inserted after the last speech of the respective meeting.

Identified speakers are marked with the tag <note type="speaker"> and their speech is labelled as an utterance <u> comprising a segment <seg>. The segments are divided into sentences (<s>), which in turn are divided into words (<w>) and punctuation marks (<pc>). All elements have a unique identifier (id) that extends the document identifier assigned to each document based on the date of the session. Sentences also contain a language attribute. Words are assigned properties such as morphosyntactic description (msd), part-of-speech tag (pos), and the canonical form of the word (lemma). An example of an annotation at word level can be found in Fig. 2b.

Parts of stenographic transcripts that are not assigned to a specific speaker (such as an introductory text) are labelled as a separate paragraph (<p> element). However, this does not apply to comments and descriptions of events that took place during the speech, such as 'Bravo!', 'Unrest in the hall'. or 'reads'. These are marked within the speeches as <note> tags.

3. Corpus overview

The Stenographic Records of the Carniolan Assembly Meetings from 1861 to 1913, comprising 694 sessions, have been compiled into the Carniolan Provincial Assembly corpus Kranjska 1.0 (Kavčič, Mundjar, and Marolt 2023), which is available in the CLARIN.SI[9] repository. Each parliamentary session is represented by two documents: a TEI XML file conforming to the Parla-CLARIN format (as described in Section 2.4) and a corresponding facsimile of stenographic records in PDF format.

In total, the processed XML documents contain over 48 thousand utterances, over 502 thousand sentences and approximately 10 million words. Earlier debates were predominantly in German (see the number of sentences for both languages in Table 1), while language use shifted towards Slovenian in the seventh term, which then became the predominant language in the following terms. This change was not only a linguistic shift, but also an expression of a broader socio-political dynamic and the formation of a national identity. It clearly illustrates the effectiveness of language policy and its impact on the prestige of a language and the status of its speakers (Stergar 2019). Overall, about 58 per cent of the sentences in the entire corpus are in Slovenian and 42 per cent in German.

The great dominance of German in the parliamentary debates of the first legislative period is not surprising, as Slovenian was not yet recognized as an official language in administrative matters at that time. Although the imposed constitution of 1849 stipulated the equality of all languages in the Austrian Empire and recognized Slovenian as the official language, the constitution was abolished in 1851 and German was again declared the internal official language. The

(a)

```
<list type="agenda" xml:lang="sl">
   <head>Dnevni red</head>
   <item n="1">1. Pričetni govor predsednikov.</item>
   <item n="2">2. Volitev zapisnikarjev.</item>
</list>
<list type="agenda" xml:lang="de">
   <head>Tagesordnung</head>
   <item n="1">1. Eröffnung des Landtages durch den
   Präsidenten.</item>
   <item n="2">2. Wahl der Schriftführer.</item>
</list>
<note type="time" xml:lang="sl">Seja se začne o 12.
minuti črez 11. uro.</note>
<note type="time" xml:lang="de">Lcginn der Sitzung 11
Uhr 12 Minuten.</note>
```

(b)

```
<note type="speaker">Arafident:</note>
<u>
   <seg xml:id="DZK_1868-08-22_08_01_seg1" n="1">
      <s xml:lang="de" xml:id="DZK_1868-08-22_08_01_seg1.s1">
         <w xml:id="DZK_1868-08-22_08_01_seg1.s1.w1" msd=
         "UPosTag=ADJ|Case=Nom|Gender=Masc|Number=Sing" pos=
         "ADJA" lemma="hoch">Hoher</w>
         <w xml:id="DZK_1868-08-22_08_01_seg1.s1.w2" msd=
         "UPosTag=NOUN|Case=Nom|Gender=Masc|Number=Sing" pos=
         "NN" lemma="Landtag" join="right">Landtag</w>
         <pc xml:id="DZK_1868-08-22_08_01_seg1.s1.w3" msd=
         "UPosTag=PUNCT" pos="$." lemma="!">!</pc>
      </s>
```

Figure 2. The data format of the stenographic records in the Parla-CLARIN XML file, showing an excerpt from the proceedings of 22 August 1868. The agenda and the time of the start of the session are shown on the left (a), while a speaker and the first sentence of his speech are shown on the right (b). The sentence is broken up into words, followed by punctuation.

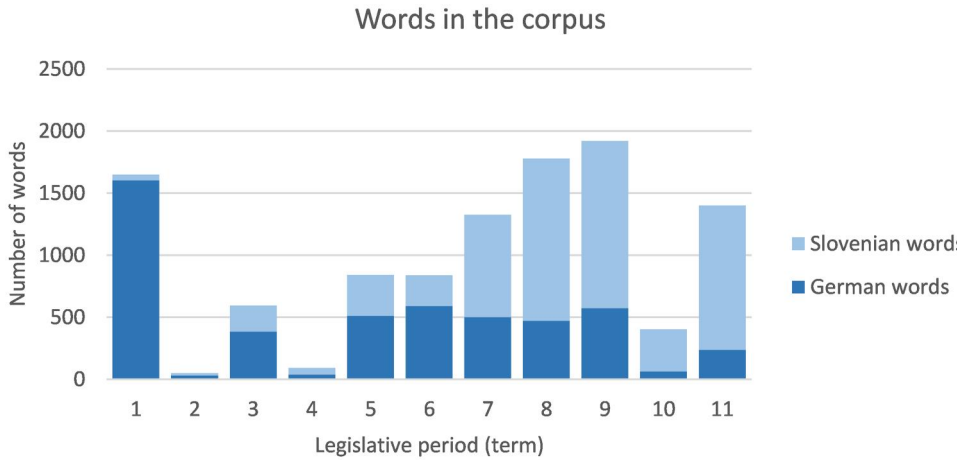


Figure 3. The number of all words per legislative period (term) by language. The numbers also depend on the duration of each term.

predominance of German is therefore also the result of the unresolved status of Slovenian as an official language (Novak 2007).

Furthermore, German delegates occupied at least 28 per cent of the seats in the Carniolan Provincial Assembly, although the proportion of the German population in Carniola was around 5 per cent. This disproportionate representation was primarily due to the great landowners' curia, which held 27 per cent of the parliamentary seats (Cvirn 2006).

The increased use of the Slovenian language in parliamentary debates can be attributed to the political awareness and national awakening of the Slovenes in the 19th century. Language is not only a means of communication, but also defines ethnic affiliation and is an expression of national identity. For the Slovenes, the Slovenian language was one of the most important foundations of personal and national identity as well as national consciousness (Mikolič 2000). At the time when national issues came to the fore, the Slovenes had neither a social upper class nor a developed high culture in their own language, and were only just beginning to develop socially, politically, and culturally. The German language dominated the official, administrative, and elite circles, while the use of Slovenian was limited to oral communication with the common people and occasionally to written documents intended to be understood by them (Zwitter 1967).

Slovenian national politics in the 1880s focussed mainly on language rights in schools and offices, and access to jobs in the civil service (Zwitter 1967). The growth of the national movement and efforts to achieve cultural and linguistic autonomy influenced the use of the mother tongue in public affairs and debates. The increasing use of Slovenian in parliamentary debates thus reflects the general social and political changes of the time.

The number of words in each language by term is shown in Fig. 3. The second and fourth terms were significantly shorter (elections were repeated after several months) and had a lower number of sessions, which is reflected in the word volume produced in those periods. Interestingly, the average number of words per stenographic record is also lower for these two prematurely ended terms (approximately 7,100 and 5,400, respectively), while the average for the whole corpus is approximately 15,700.

Although the tenth term completed the entire six-year mandate, the number of words in this term is relatively low (approx. 402 thousand), even lower than in the third term (approx. 593 thousand), which ended after three years. However, the number of words per stenographic record for the tenth parliamentary term is only just below average at around 13,400. This can be attributed to the small number of sessions in this term—only thirty were held, the fewest in any term apart from the short-lived second and fourth terms. This low number probably reflects a combination of political, institutional, and historical circumstances rather than a simple lack of activity. It is noteworthy that the average number of pages per stenographic record and the average word count are consistent with those of other completed terms.

In the eleventh term, the number of delegates in the assembly increased from thirty-seven to fifty. This change is also clearly recognizable in the stenographic records of the debates. The total number of words in this mandate is not the highest, although it is above average, as the total number of sessions in this period is well below average. However, the debates were longer, as indicated by the number of words per stenographic record, which is the highest of all terms at almost 27 thousand. Consequently, the number of pages per

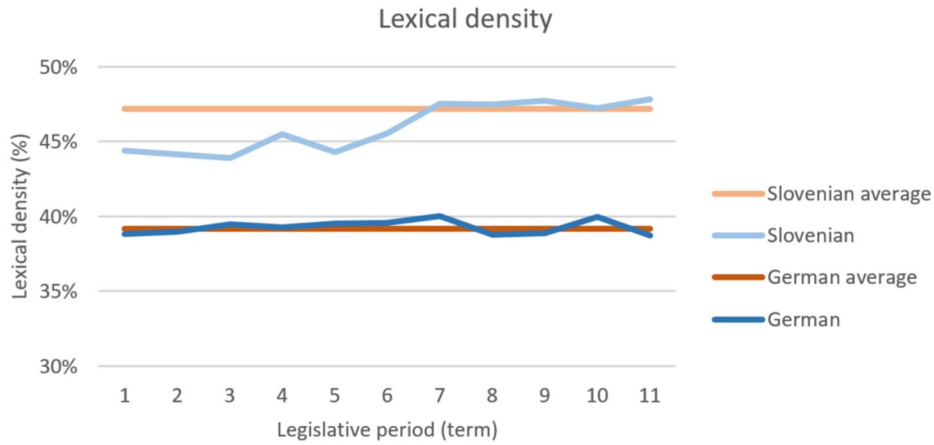


Figure 4. Lexical density per legislative period (term) by language. Calculated as the ratio of content words (nouns, verbs, adjectives, and adverbs) to the total number of words. The average value for the entire corpus is shown in orange.

document is also the highest at an extreme 40.29 pages (the average for all terms is 22.12).

Lexical diversity of the corpus, expressed as a type-token ratio (TTR), is extremely low and varies between 3 per cent and 15 per cent for both languages across the terms. It is also very low for the whole corpus, 4 per cent for German and only 2 per cent for Slovenian. A low TTR is not a consequence of the delegates' limited vocabulary, but is rather due to the nature of the specialized parliamentary discourse, which is highly institutionalized and highly procedural and formulaic. In parliamentary debates, repeated terminology is often used and the texts are lengthy. A higher TTR is typically observed in shorter texts. In comparison, the TTR of the siParl corpus (Pančur *et al.* 2024) is even lower with 0.12 per cent as the corpus contains over 230 million words.

We also calculated the lexical density of the corpus as the ratio of lexical items (content words, which include nouns, verbs, adjectives, and adverbs) to the total number of words. The results, presented in Fig. 4, show that the lexical density varies slightly over the terms, ranging from 43.9 per cent to 47.8 per cent for Slovenian and from 38.7 per cent to 40.0 per cent for German, which is not a significant difference. The overall diversity for the entire corpus is 47.2 per cent for Slovenian and 39.2 per cent for German. The differences in lexical density between the languages can be attributed to their structural and inherent linguistic features rather than to the structure of the debates. However, the increase in lexical density in Slovenian is probably due to the national awakening and cultural emphasis on the language and its use in formal situations.

Table 2 shows the lists of the ten most frequently occurring words in the corpus for each language. To account for the diversity of word forms in the language, we counted the word lemmas and restricted the count to nouns with at least three letters. Unsurprisingly, the most common words in both languages match and are closely related to the workings of the regional parliament. Furthermore, the lists are almost identical in content and differ by only one word, with the differences mainly being in the order.

The most common word in both languages, however, is the word *gospod* or *Herr* (meaning *Sir* or *Mister*), a title of honour for men, which is usually preceding their surname. This can be seen as a formal address of the delegates in their debates.

Further frequent words in the corpus can be found in the word clouds in Fig. 5. Here, too, we have restricted the words to lemmas of nouns that have at least three letters. The cloud for Slovenian (left) contains similar words to those in the cloud of German words (right).

4. Accessing and querying the corpus

The corpus is prepared in a standardized TEI format that is machine-readable and suitable for automatic processing. Although it can be opened as plain text in any text viewer, it is not intended for direct reading by humans and requires special tools for effective use. To facilitate analysis, the corpus is integrated into the concordancers NoSketch Engine[10] and KonText (Machálek 2014), both of which are available on the CLARIN.SI platform. Both are useful for the statistical analysis of the corpus, as they enable corpus queries and analyses without programming knowledge, and contain various visualizations.

To make the historical parliamentary records accessible to a wider audience—especially those without experience of interpreting TEI-encoded files or using

Table 2. The ten most frequent words in the entire corpus, separately for each language. Only the nouns are counted and the lower case lemma of the word is used as a distinguishing feature. The English translation of the word is given in brackets. The words that do not appear in both lists are printed in bold.

Rank	Slovenian	German
1	gospod (sir/mister)	herr (sir/mister)
2	odbor (committee)	antrag (proposal/application)
3	**odsek** (committee/section)	jahr (year)
4	predlog (proposal/application)	land (country/state/province)
5	leto (year)	landesausschuß (state/province committee)
6	zbor (parliament)	landtag (parliament)
7	poslanec (delegate/member of parliament)	**bericht** (report)
8	občina (municipality/community)	wort (word)
9	beseda (word)	gemeinde (municipality/community)
10	dežela (country/state/province)	abgeordnete (delegate/member of parliament)

Figure 5. The 100 most frequent words in the corpus for Slovenian (a) and German (b) words. Only the lemmas of nouns (at least three letters long) are counted.

concordancers—we have developed ParlaVis (Kavčič, Stojanoski, and Marolt 2024),[11] a web-based application that facilitates the exploration of PDF facsimiles of stenographic records. The main advantage of ParlaVis lies in its robust, multilingual search function, which provides a flexible and user-friendly interface for navigating through complex historical parliamentary data. The application supports keyword-based searching, including exact phrases in quotation marks and canonical word forms (lemmas), with the search algorithm incorporating all morphological variants. Users can further refine their search queries by filtering results by speaker, place name, language, or session date.

The search results provide a preview of the first page of the facsimile and display the queried terms in their immediate textual context. The full documents can be accessed in the PDF format of the facsimile, with the search terms highlighted and navigable via integrated controls. Alternatively, users can switch to a transcription view that displays the original text and provides ready-made translations between Slovenian and German created with the LibreTranslate[12] Python library.

5. Conclusion

The article presents the first version of the corpus of the Carniolan Provincial Assembly and its creation process. The stenographic notes of the sessions from 1861–1913 were automatically processed and not manually checked or corrected. Due to the varying quality of the OCR, which had to cope with partially worn documents, archaic language, Gothic script as well as tables and ornaments in the text, the quality of the results is limited and in need of improvement. However, as the first corpus of Slovenian parliamentary debates from the Austro-Hungarian Empire, it is a welcome and much-needed resource.

The developed corpus will enable us to identify types of changes that occur in political language in a particular political, national, and partisan context (Ihalainen, Ilie, and Palonen 2018), thus providing new and innovative source-based fundamentals to re-think the political life in the multinational Habsburg empire. The corpus is available in the CLARIN.SI repository and makes an important contribution to the availability of comparable, interoperable collections of parliamentary

debates for diachronic, multilingual, and transnational research. To facilitate the exploration of the corpus, we have also developed a web application with advanced search capabilities. The search results can be explored in a facsimile PDF of the stenographic notes or in its transcript.

Our future work will focus on several challenges. One of them is to improve the OCR quality by preprocessing the scanned raw images and reducing the image noise to enable more accurate OCR. In addition, post-processing the text with machine learning and natural language processing techniques could help to correct the errors in the texts. Another task is to extract additional metadata from the texts, such as named entities (e.g. delegates, persons, places, and countries), and link these to external data sources (e.g. for delegates the name of their party, active years in parliament, and political orientation). As part of the next corpus update, we intend to refactor the code used to generate the TEI documents and publish the documented version under an open licence in a publicly accessible code repository. We also plan to add more historical parliamentary debates to the collection to fill the gap and provide records from 1861 to the present day in a standardized, structured format.

Funding

This work was supported by the Slovenian Research and Innovation Agency research programmes Digital Humanities: resources, tools and methods (2022–2027) [grant number P6-0436]; Political History (2022–2027) [grant number P6-0281]; and the DARIAH-SI research infrastructure (2022–27).

Notes

1. https://gallica.bnf.fr/ark:/12148/cb328020951/date.item
2. https://hansard.parliament.uk/
3. https://www.congress.gov/
4. http://dlib.si/
5. https://pdf.abbyy.com/
6. https://github.com/pemistahl/lingua-py
7. https://github.com/nlp-uoregon/trankit
8. https://github.com/clarinsi/classla
9. https://www.clarin.si
10. https://www.sketchengine.eu/nosketch-engine
11. https://parlavis.si
12. https://github.com/LibreTranslate/LibreTranslate

Data availability

The data underlying this article are available in Slovenian language resource repository CLARIN.SI (ISSN 2820-4042), at http://hdl.handle.net/11356/1824.

References

Abrami, G., Bagci, M., and Mehler, A. (2024) 'German parliamentary corpus (GerParCor) reloaded', in N. Calzolari *et al.* (eds) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7707–16. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.681/

Blätte, A., Gehlhar, S., and Leonhardt, C. (2020) 'The Europeanization of parliamentary debates on migration in Austria, France, Germany, and the Netherlands', in D. Fišer, M. Eskevich, and F. de Jong (eds) *Proceedings of the Second ParlaCLARIN Workshop*, pp. 66–74. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.parlaclarin-1.12/

Blaxill, L., and Beelen, K. (2016) 'A Feminized Language of Democracy? The Representation of Women at Westminster since 1945', *Twentieth Century British History*, **27**: 412–49.

Coole, M., Rayson, P., and Mariani, J. (2020) 'Unfinished business: construction and maintenance of a semantically tagged historical parliamentary corpus, UK Hansard from 1803 to the present day', in D. Fišer, M. Eskevich, and F. de Jong (eds) *Proceedings of the Second ParlaCLARIN Workshop*, pp. 23–27. Marseille, France: European Language Resources Association. https://aclanthology.org/2020.parlaclarin-1.5/

Cvirn, J. (2006) *Razvoj Ustavnosti in Parlamentarizma v Habsburški Monarhiji: Dunajski Državni Zbor in Slovenci (1848–1918)*. Ljubljana: Filozofska fakulteta, Oddelek za zgodovino.

Drobac, S., Sinikallio, L., and Hyvönen, E. (2023) 'An OCR Pipeline for Transforming Parliamentary Debates into Linked Data: Case ParliamentSampo—Parliament of Finland on the Semantic Web', *Digital Humanities in the Nordic and Baltic Countries Publications*, **5**: 287–96.

Erjavec, T., and Pančur, A. (2021) 'The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings', *Journal of the Text Encoding Initiative*, **14**: 1–20. https://doi.org/10.4000/jtei.4133

Erjavec, T. and Pančur, A. (2022) 'Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings', https://clarin-eric.github.io/parla-clarin/, accessed 12 Oct. 2022.

Erjavec, T. *et al.* (2023) 'The ParlaMint Corpora of Parliamentary Proceedings', *Language Resources and Evaluation*, **57**: 415–48. https://doi.org/10.1007/s10579-021-09574-0

Frasnelli, V., and Palmero Aprosio, A. (2024) 'There's something new about the Italian parliament: the IPSA Corpus', in N. Calzolari *et al.* (eds) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16037–46. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1394/

Hampsher-Monk, J. (1998) 'Speech Act, Languages or Conceptual History', in I. Hami'siier-Monk, K. Tilmans, and F. Van Vree (eds) *History of Concepts: Comparative Perspectives*, pp. 37–50. Amsterdam: Amsterdam University Press.

Ihalainen, P., Ilie, C., and Palonen, K. eds (2018) *Parliament and Parliamentarism: A Comparative History of a European Concept*. New York, NY: Berghahn Books. https://doi.org/10.2307/j.ctvgs0b7n

Kavčič, A., Mundjar, A., and Marolt, M. (2023) 'Carniolan Provincial Assembly corpus Kranjska 1.0', Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1824

Kavčič, A., Stojanoski, M., and Marolt, M. (2024) 'Historical parliamentary corpora viewer', in D. Fiser, M. Eskevich, and D. Bordon (eds) *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pp. 127–32. Torino, Italia: ELRA and ICCL. https://aclanthology.org/2024.parlaclarin-1.19/

Koselleck, R. (2006) *Begriffsgeschichten: Studien zur Semantik und Pragmatik der Politischen und Sozialen Sprache*. Frankfurt am Main: Suhrkamp Verlag.

Kurz, S. (2024) 'Oeaw-ministerratsprotokolle/mp-edition-data: v. 1.5 Including CMR Calendar Data 1872–1914 (v.1.5) (Zenodo; pubd online 5 June 2024)', https://doi.org/10.5281/zenodo.11484662, accessed 2 Apr. 2025.

Ljubešić, N., and Dobrovoljc, K. (2019) 'What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian', in T. Erjavec *et al.* (eds) *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pp. 29–34. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Machálek, T. (2014) 'KonText—Application for Working with Language Corpora (FF UK, Praha)', http://kontext.korpus.cz, accessed 2 Apr. 2025.

Melik, V. (1969) 'Nekaj Značilnosti Razvoja na Kranjskem 1867–1871', *Zgodovinski časopis*, **23**: 65–74. https://hdl.handle.net/11686/file76

Mikolič, V. (2000) 'Povezanost Narodne in Jezikovne Zavesti', *Jezik in Slovstvo*, **45**: 173–85. http://www.dlib.si/details/URN: NBN: SI: doc-5BD79SLS

Müller-Hansen, F. *et al.* (2021) 'Who Cares about Coal? Analyzing 70 Years of German Parliamentary Debates on Coal with Dynamic Topic Modeling', *Energy Research & Social Science*, **72**: 101869.

Novak, N. (2007) 'Oblikovanje Prvih Vzorcev Pravniških Besedil v Slovenščini in Njihova Raba v Praksi', *Philological Studies*, **5**: 187–94.

Nguyen, M. V. *et al.* (2021) 'Trankit: a light-weight transformer-based toolkit for multilingual natural language processing', in D. Gkatzia and D. Seddah (eds) *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 80–90. Stroudsburg, PA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-demos.10

Ogrodniczuk, M., and Nitoń, B. (2020) 'New developments in the Polish parliamentary corpus', in D. Fišer, M. Eskevich, and F. de Jong (eds) *Proceedings of the Second ParlaCLARIN Workshop*, pp. 1–4. Marseille, France: European Language Resources Association.

Pančur, A. *et al.* (2024) 'Slovenian parliamentary corpus (1990-2022) siParl 4.0', *Slovenian Language Resource Repository CLARIN.SI*, ISSN 2820-4042, http://hdl.handle.net/11356/1936

Puren, M.A. *et al.* (2022) 'Extracting and Providing Online Access to Annotated and Semantically Enriched Historical Data. The AGODA project'. *Digital Humanities Conference 2022 Book of abstracts*, pp. 540–3. Tokyo, Japan: The University of Tokyo.

Stergar, R. (2019) 'Chapter 4: The Evolution of Linguistic Policies and Practices of the Austro-Hungarian Armed Forces in the Era of Ethnic Nationalisms: The Case of Ljubljana-Laibach', in M. Prokopovych, C. Bethke, and T. Scheer (eds) *Language Diversity in the Late Habsburg Empire*, pp. 50–71. Leiden, The Netherlands: Brill. https://doi.org/10.1163/9789004407978_005

Štih, P., Simoniti, V., and Vodopivec, P. (2008) *Slovenska zgodovina: Družba—Politika—Kultura*. Ljubljana: Inštitut za Novejšo Zgodovino: Sistory. https://hdl.handle.net/11686/file448

TEI Consortium. (2023) 'TEI: Guidelines for Electronic Text Encoding and Interchange, P5', https://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html, accessed 27 May 2024.

Taylor, A. J. P. (1976) *The Habsburg Monarchy, 1809–1918: A History of the Austrian Empire and Austria-Hungary*. Chicago, IL: University of Chicago Press.

Ljubešić, N., Terčon, L., and Dobrovoljc, K. (2024) 'CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages', in Š. Arhar Holdt, and T. Erjavec (eds) Proceedings of the Conference on Language Technologies and Digital Humanities (JT-DH-2024), pp. 251–274. Ljubljana, Slovenia: Institute of Contemporary History. https://doi.org/10.5281/zenodo.13936406

Vilfan, S. (1961) *Pravna zgodovina Slovencev: Od Naselitve do Zloma Stare Jugoslavije*. Ljubljana: Slovenska Matica.

Yrjänäinen, V. A. *et al.* (2024) 'The Swedish Parliament Corpus 1867–2022', in N. Calzolari (ed.) *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16100–12. Torino, Italia: ELRA and ICCL.

Zwitter, F. (1967) 'Slovenci in Habsburška Monarhija', *Zgodovinski Casopis*, **21**: 49–67. https://hdl.handle.net/11686/file73

Digital Scholarship in the Humanities, 2025, 00, 1–11
https://doi.org/10.1093/llc/fqaf068
DSH 2023 Conference Supplement