SlovParl 2.0:
The Collection of Slovene Parliamentary Debates from the Period of Secession

Andrej Pančur[1], Mojca Šorn[2], Tomaž Erjavec[3]
Institute of Contemporary History[1,2], Jožef Stefan Institute[3]
Ljubljana Slovenia
{andrej.pancur,mojca.sorn}@inz.si, tomaz.erjavec@ijs.si

Abstract

The paper describes the process of acquisition, up-translation, encoding, and annotation of the collection of the parliamentary debates from the Assembly of the Republic of Slovenia from 1990-1992, covering the period before, during, and after Slovenia became an independent country in 1991. The entire collection, comprising 232 sessions, 58,813 speeches and 10.8 million words was uniformly encoded in accordance with the Text Encoding Initiative (TEI) Guidelines, using the TEI module for drama texts. The corpus contains extensive meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations. The corpus was also converted to use the spoken corpus module of the TEI, and from this encoding automatically part-of-speech tagged and lemmatised. The corpus is maintained on GitHub and its major versions archived in the CLARIN.SI repository and is available for analysis under the CLARIN.SI concordancers, offering an invaluable resource for historians studying this watershed period of Slovenian history.

Keywords: Slovenia's independence, Text Encoding Initiative, Open Science

1. Introduction

Parliamentary papers are a rich source of data used by different academic disciplines, among others, historiography, sociology, political science, linguistics, economic and economic history. Parliamentary papers include transcriptions of parliamentary debates, debate reports, session papers, petitions, legal documents, amendments, statements, written questions, committee reports, transcription of committee debates, etc. In some European countries, a large part of parliamentary papers is already accessible in digital form, but mostly in PDF only (Benardou, et al., 2015).

This mostly also applies to Slovenia as transcriptions of parliamentary debates in PDF or HTML are available for different historical regional parliaments[1] and various national parliaments of the countries to which Slovenia belonged.[2] Only for the National Assembly of the Republic of Slovenia, all transcription of parliamentary debates after 1990 are available in the HTML format.[3] Other parliamentary papers, especially session papers,[4] are also increasingly available in digital form. Under a conservative estimate, the session papers and the transcriptions of parliamentary debates since 1945 alone have more than 170 million words (Pančur & Šorn, 2016).

It is clear that no researcher is able to read that much text in its entirety. Most researchers in the humanities understand such digital materials only in terms of easier and quicker access to desired information (Spiro, 2014). They have no intention of reading the whole text, but only to find what they are looking for in the text and typically use search engines to identify sources and do a full text search on the results. This also applies to researchers of Slovenian parliamentary history. However, this research method has its limitations. Inasmuch as the researcher does not carefully examine every search result, the results are always lacking their proper context (Robertson, 2016).

For these reason, a small group of historians decided to build a corpus of parliamentary debates that will capture as much contextual information as possible. We chose a relatively short but historically very interesting period of the National Assembly of the Republic of Slovenia between 1990 and 1992, covering the period before, during, and after Slovenia became an independent country in 1991. From 1945 to 1991 Slovenia was part of Yugoslavia, and the parliament reflected its socialist system. The first multi-party elections took place in April 1990. In the next two years, the socialist assembly served the democratically elected members as a framework that enabled them to change and adapt the legislation for the Slovene Republic, which then led to Slovenia's independence in June 1991. In 1992, the members of the assembly passed the new constitution, which formally ended the era of the Socialist Assembly of Slovenia and established the new classical parliament.

The first, pilot version of this corpus spanning 1990 to 1992 and containing 2.7 million words was released in 2016 (Pančur et al., 2016), and on its basis we have made experiments on the possible use of such corpora in historical research (Pančur & Šorn, 2016).

Building of annotated corpora of (historical) parliamentary debates has already been undertaken for a number of countries, e.g., United Kingdom from 1803 on (Alexander, et al., 2016), Netherlands from 1814 on (Marx & Schuth, 2010) and Canada from 1901 on (Beelen et al., 2017). The Dutch corpus has already been successfully used in historical research (Piersma et al., 2014). From the Slavic-speaking countries, we are aware of only one other available corpus of Parliament

[1] Representative assemblies of the Austrian crown lands Carniola (1861-1874) http://hdl.handle.net/11686/menu719 and Styria (1848-1914) http://www.landesarchiv.steiermark.at/cms/ziel/111284715. Assembly of the Yugoslav federal republic of Slovenia (1947-1990) http://hdl.handle.net/11686/menu407, http://hdl.handle.net/11686/menu407.

[2] Habsburg Monarchy (1861-1918), http://alex.onb.ac.at/sachlichegliederung.htm. Yugoslavia (1919-1939, 1942-1953), http://hdl.handle.net/11686/menu233, http://hdl.handle.net/11686/menu825, http://hdl.handle.net/11686/menu822.

[3] https://www.dz-rs.si.

[4] http://hdl.handle.net/11686/menu828.

Meetings, from the Czech republic (Pražák & Šmidl, 2012).

This paper documents the making, annotation and availability of the second, comprehensive (10.8 million words and 58,813 speeches) version of the SlovParl corpus and is structured as follows: Section 2 explains the process of compilation, Section 3 details its annotation, Section 4 focuses on its availability, Section 5 gives some possibilities of a quantitative analysis of the corpus, and Section 6 gives the conclusions and directions for further research.

2. Building the Corpus

2.1 Basic Principles

In the design of SlovParl corpus, we followed these basic principles:

1. **Multidisciplinary**: The corpus must be useful not only for historians, but also for other disciplines. That is why SlovParl corpus (and also this paper) was created in close cooperation between the Slovenian DARIAH[5] and CLARIN[6] communities.
2. **All-inclusive**: In addition to parliamentary debates, other types of parliamentary papers will eventually be included.
3. **Long-term**: Since such large-scale plans can't be realized during the period of a short-term research project, these activities should be financed as part of long-term research infrastructures.
4. **Open science**: All previous principles can be optimally realized only in accordance with the principles of open science.

2.2 Document structure

Parliamentary debates are typically published in a uniform format, which fluctuates very little in time (Marx, 2009). This also applies to Slovenian parliamentary debates. By analysing representative samples, we found the following structure of parliamentary proceedings (with minimal and maximal occurrences of structural elements):

- Document (1, n)
  - Table of contents (0, 1)
  - List of speakers (0, 1)
  - Index (0, 1)
  - Annex (0, n)
  - Meeting (1, n)
    - Non-verbal content (0, n)
    - Topic (1, n)
      - Non-verbal content (0, n)
      - Speech (1, n)
        - Non-verbal content (0,n)
        - Paragraph (1, n)
          - Non-verbal content (0, n)

The structure of individual documents is very flexible. They might contain all meetings of all parliamentary chambers in one year, one meeting that lasts for several days, or only one day of the meeting. The document may contain the table of contents, the list of speakers, the topic index and annexes (session papers, legislation), or these might be present in separate documents. Non-verbal content of parliamentary debates (information about the meeting and the chairperson, description of the outcome of a vote, description of actions like applause, etc.) can be present anywhere in the structure of the meeting. Transition from one topic to another can occur during the chairman's speech.

2.3 Source Files

Transcriptions of parliamentary debates are available as PDF or HTML files on the web portal SIstory – History of Slovenia and on the Web pages of the Slovenian parliament. PDFs contain either images or OCR scanned text, while HTML files contain the digitized analogue of paper transcripts or born-digital text. Furthermore, OCR produced at times high-quality results but also quite low-quality transcriptions, due to the low print quality of the original. The following conversion, transcription and annotation procedures have been developed for these different source file formats: PDF → DOCX → XML, HTML → XML (Pančur, 2016).

To build the SlovParl corpus we only needed the HTML → XML conversion path, as the transcriptions of parliamentary debates of the National Assembly of the Republic of Slovenia are available on their web portal in HTML. We originally scraped the wanted data from their website, but after 2016 the links to the HTML files, together with metadata, are openly accessible as XML files.[7] The information (such as transcriptions of parliamentary debates) from this web portal is regarded as information of public character, with the disclamer that it can be always altered.[8]

2.4 Transcription

Transcriptions of Slovene parliamentary debates from the period of secession (1990-1992) were initially published as analogue publications and were digitized a few years ago by the National Assembly. OCR errors have been in most cases corrected.

The uniform structure of documents with parliamentary debates is very well suited for automatic annotation. But because HTML files for the period 1990-1992 do not contain born-digital text, the document structure is not clearly marked. The layout and other typographical aspects of source text (bold, italic, underline, indent, uppercase, punctuation, spacing) are not always consistently applied. Therefore, when converting from HTML to XML, semi-automatic annotation was performed in several steps. Each step contained:

1. using an XSL stylesheet for automatic annotation;
2. searching for annotation errors (XPath and regular expression search);
3. additional manual annotation.

[5] http://www.dariah.si/en/.
[6] http://www.clarin.si/info/about/.
[7] https://www.dz-rs.si/wps/portal/Home/OpenData.
[8] https://www.dz-rs.si/wps/portal/en/Home/pravnoObvestilo.

3. Annotation

The SlovParl 2.0 corpus is encoded as one XML document. Ten years ago there was no special XML schema for parliamentary proceedings (Marx, 2009). Today, the situation is completely different, and one can choose between Political Mashups (Gielissen & Marx, 2009),[9] Parliamentary Metadata Language (PML) (Gartner, 2014), and, last but not least, the Akomo Ntoso[10] schema.

Despite these options esp. developed for annotating parliamentary proceedings, we decided to use the Text Encoding Initiative Guidelines (TEI Consortium, 2016) for encoding SlovParl. This decision was based on our first two basic principles: multidisciplinary and all-inclusive corpus design. TEI is not only the *de facto* standard for annotating electronic text in the humanities, but is also widely used in the Slovenian CLARIN community (Erjavec et al., 2016). TEI has community-based maintenance, extensive documentation and a number of supporting tools. A central aspect of TEI usage is customization and the TEI Guidelines are designed with customization in mind. Unlike Political Mashups and PML, TEI can be used not only for the annotation of parliamentary proceedings, but also for all other types of parliamentary papers. In this respect, only Akoma Ntoso is comparable with TEI, as it is specially designed for parliamentary, legislative and judiciary documents. It also allows customization. However, the TEI ODD (One Document Does it all) specification language can also be used as a powerful technical platform for customization, as it offers project and data specific customisations and documentation, comparison of TEI-based project through their ODDs and even ODD chaining.[11]

3.1 TEI drama and TEI speech

Each TEI document is rooted in the <TEI> elements, which first contains a <teiHeader> element with metadata (title, date, time period, parliamentary organization, licence, source, automatic annotation and revision description). The TEI header is followed by the <text>, which in our case contains the document structure described in Sec. 2.2 above. The table of contents, the list of speakers, Index and Annexes can be found as elements in <front> or <back>, while meetings are located in the <body> element. Topics are encoded as elements inside <body>. They bear the @corresp attribute with references to the table of contents.

Scenes, acts and speeches are structural features of performance text (Marx, 2009). We used the TEI module for Performance texts for implementing the analysis of the materials. These include elements for encoding the list of speakers as a cast list (<castList>), a speech (<sp>), the name of the speaker (<speaker>) and the "stage directions" (<stage>). Each speech element bears a @who attribute with a local reference to the <actor> element in the cast list. Different types of non-verbal content (<stage>) are annotated with the @type attribute, which can have the following values: location, time, vote, quorum, debate, comment, gap, vocal, kinesic, and incident.[12] The <timeline> element provides a set of ordered points in time which are linked to the <stage> element with information about the time of the beginning and end of the debate.

In the next phase, these TEI documents are included in the <teiCorpus> element. We made a common list of speakers and the index of topics for the entire corpus. In both cases, we encoded this data in separate TEI documents. In this way, we created a list of all MPs and other speakers (<listPerson>) and a list of all organizations (<listOrg>) whose members were these speakers. We used the TEI module for encoding persons (<person>), places and organizations (<org>). We took into account any changes to the names and structure of the organization. Through the attributes @ref and @ana of the <affiliation> element, persons are associated with the organizations (parliamentary chamber, political party, government institution) to which they belonged over different time periods. In the <speech> element, the local reference to the element <actor> was moved from @who to @corresp. Attribute @who now contains relative URI reference to a local document with <listPerson>.

For the next phase, we intended to carry out the linguistic annotation of the corpus. But within paragraphs (<p>) the speeches were very often interrupted by non-verbal <stage> elements. Therefore, we decided to break the existing paragraphs into verbal (utterance <u>) and non-verbal elements (<note>, <vocal>, <kinesic>, <incident> and <writing>) and these elements are defined in the TEI module for spoken corpora. An XSLT stylesheet was used to convert the source TEI drama-encoded corpus to the target TEI speech-encoded corpus. Local documents for the list of persons and the topic index have been included in <teiHeader> of the speech <teiCorpus>.

3.2 Linguistic annotation

The TEI-speech encoded corpus was tokenized, sentence segmented, tagged with morphosyntactic descriptions (MSDs) and lemmatised with the ReLDI tagger (Ljubešić & Erjavec 2016). The resulting corpus is encoded identically to the source one, but, as illustrated in Figure 1, with the added sentence (<s>) word (<w>), punctuation (<pc>) and whitespace (<c>) elements. The word elements also bear the @lemma attribute, while both word and punctuation elements are annotated with @ana, which gives the MSD of the token.

```
<s>
  <w lemma="2." ana="msd:Mdo">2.</w><c> </c>
  <w lemma="verifikacija" ana="msd:Ncfsn">Verifikacija</w>
  <c> </c>
  <w lemma="mandat" ana="msd:Ncmsg">mandata</w>
  <c> </c>
  <w lemma="v" ana="msd:Sl">v</w><c> </c>
  <w lemma="zbor" ana="msd:Ncmsl">zboru</w>
  <pc ana="msd:Z">.</pc>
</s>
```

Figure 1. Linguistic annotation of the corpus.

[9] http://schema.politicalmashup.nl/schemas.html.
[10] http://www.akomantoso.org/.
[11] https://wiki.tei-c.org/index.php/ODD_chaining.

[12] Those familiar with TEI will notice that the last four value are in fact also names of TEI elements. We used them as the values of stage/@type in order to have a uniform encoding of the "stage directions" as present in the original transcripts.

It should be noted that the MSDs are given using the <prefixDef> element in the TEI header, which defines the prefixing scheme used, showing how abbreviated URIs using the scheme may be expanded into full URIs. In the case of the SlovParl 2.0 corpus the "msd:" prefix is simply expanded to local reference (i.e. "#") with the definitions of the MSDs included in the <back> element of linguistically annotated corpus – there, each MSD is defined as a feature-structure giving the decomposition of the MSD into its features. It is thus a simple matter, using just the TEI encoded corpus, to move from "msd:Mdo" to "Category = Numeral, Form = digit, Type = ordinal".

4. Availability and maintenance

4.1 GitHub

In accordance with our fourth basic principle (open science), all TEI annotated versions of the corpus are accessible and maintained in GitHub repositories:

- https://github.com/SIstory/Sejni_zapiski (DOCX → TEI drama – Phase 1)
- https://github.com/SIstory/Seje_DZ (HTML → TEI drama – Phase 1)
- https://github.com/SIstory/SlovParl (TEI drama – Phase 2)
- https://github.com/DARIAH-SI/CLARIN.SI (TEI speech)

4.2 CLARIN.SI repository

The corpus from the last GitHub repository is made available under the Creative Commons CC BY licence in the CLARIN.SI repository, comprising 231 sessions, 58,813 speeches and 10.8 million words (Pančur et al., 2017).

This repository item comprises four datasets:

- the corpus in TEI (module Transcription of speech);[13]
- the corpus in TEI with added automatic linguistic annotation;
- the corpus in CSV for statistical analysis software;
- the corpus in vertical format used by various concordancers.

4.3 Concordancers

The linguistically annotated version of the SlovParl 2.0 corpus has also been mounted under the two concordancers recently installed at CLARIN.SI, namely KonText[14] and noSketch Engine[15], enabling on-line exploration of this and other corpora.

The two concordancers are open source[16] and both use the same Manatee back-end (Rychlý, 2007) and set of indexed corpora, but provide different front-ends. Apart from visual differences, KonText supports log-in via the authentication and authorization infrastructure (AAI), and, in fact, allows only basic functionality without logging in. However, log-in enables the user to personalise the visual appearance of the concordancer, save sub-corpora and the query history. On the other hand, noSketch Engine, does not support log-in, so all its functionality is available to anonymous users, however, this also has the disadvantage of not allowing personalisation of the interface etc. As both concordancers use the same back-end, they also support querying via the powerful CQL query language, enabling searching via logical combinations of annotations, using regular expression, etc.

In order for a corpus to be indexed by the concordancers it needs to be fist converted to the so called vertical file format. We down-converted the linguistically annotated TEI encoded corpus to this format, also flattening the structure of the original, so that the vertical file is structured into texts (corresponding to one session) and paragraphs (corresponding to one speech) and with non-verbal parts omitted. Both structures carry metadata on e.g. the title of the session and its date, the speaker name and sex, and the type and topic of the speech. As mentioned above, this encoding of the corpus is also available for download from the CLARIN.SI repository.

5. Quantitative analysis

As mentioned above, the original reason for building a corpus was its use in historical research. In order to obtain the desired statistical information from TEI documents, we used the XML Query Language (XQuery) and XSLT. As a programming language for transforming XML documents, XSLT is not really intended for use in quantitative analysis. On the other hand, as a group of digital humanist, we have a good knowledge of XSLT, which enabled us to quickly find interesting information in the corpus. (Pančur & Šorn, 2016) For example, at the longest session the total duration of speeches was more than 56 hours and 256,692 words were spoken. From the beginning to the end of this session three months passed, it lasted 13 days and was interrupted 36 times. On the other hand, the total duration of speeches at the briefest session was only 10 minutes (643 words).

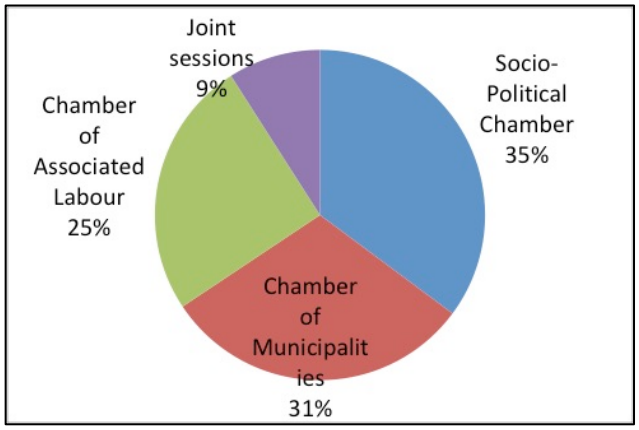


Figure 2: Number of words spoken in the chambers of the Assembly of the Republic of Slovenia (1990/92).

The Socialist Assembly of Slovenia comprised three chambers: the Socio-political chamber, the Chamber of Municipalities and the Chamber of Associated Labour.

[13] For researchers without XML knowledge this dataset is also available as a teiPublisher application. http://exist.sistory.si/exist/apps/parla/

[14] https://www.clarin.si/kontext/

[15] https://www.clarin.si/noske/

[16] The branch of KonText we use is available from https://github.com/ufal/lindat-kontext, while noSketch Engine can be downloaded via https://nlp.fi.muni.cz/trac/noske.

Joint sessions of all chambers represented less than a tenth of all parliamentary speeches (Figure 2). However, in previous research historians devoted almost exclusive attention only to some Joint sessions (Pesek, 2007). These researchers only read those small parts of the text that they considered relevant. Of course, such methods often yield useful results and a number of good studies have been created in such a manner using only pre-selected parts of parliamentary speeches. Why then would you need to build a corpus, if historians can still do well without it? Or the similar historian's question to the authors of an interdisciplinary book (corpus linguistics and historiography): "[…] what any quantification would actually show – it was clear that the corpus could quantify, but what was the purpose of that?" (McEnery an Baker, 2017, 200). We believe that it is the best to answer such a question with a concrete example:

After first multi-party elections (May 16, 1990 – May 14, 1992) the government consistent of newly established parties. Opposition parties stemmed from the former communist party and various socialist organizations. According to historians, because of its political inexperience, the coalition was relatively more silent compared to the opposition. (Pesek, 2007, p. 550) This finding was based on reading the speeches from some selected sessions (Gašparič, 2017). But corpus data show the opposite is true (Figure 3). The opposition numbered 36% of the MPs, who only had 20% of the speeches in which 32% of all words were spoken.

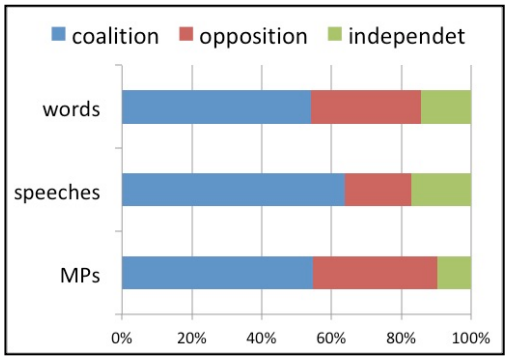


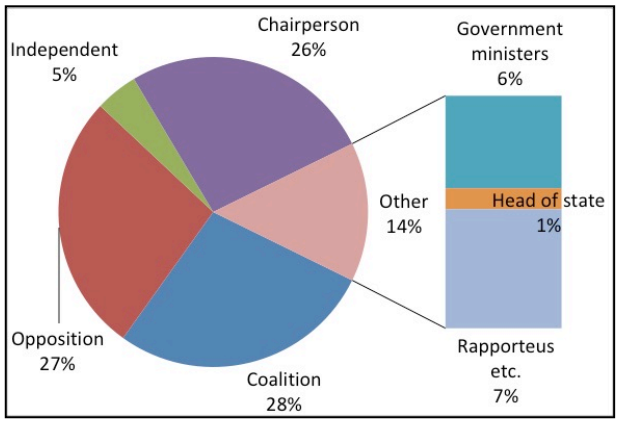
Figure 3: The percentage of members of parliament in coalition or opposition, the number of speeches and the number of spoken words; May 8, 1990 – May 14, 1992

But this simple analysis can also be misleading. All chairpersons were either coalition or independent MPs, and these chairpersons spoke as much as 26% of all words (Figure 4). If we exclude the speeches of chairmen, we find that the opposition and coalition MPs actually spoke about the same number of words.

On average, opposition MPs had more speeches than coalition MPs, which were also slightly longer. But this does not mean that the coalition as a whole was more silent than opposition. Both groups had outstanding speakers. Similarly, both groups had MPs who were almost completely silent (Figure 5).

The main question therefore is why the opposition on average had more MPs who were willing to speak more than coalition MPs? In addition to "political experience", an adequate answer to this question can only be given if other personal (gender, age etc.) and social factors (education, occupation, affiliation etc.) are taken into account.

Figure 4: Number of words spoken in the Assembly of the



Republic of Slovenia (May 8, 1990 – May 14, 1992) by organization membership.

We also made a set of CSV files containing various metadata from the corpus, appropriate for use with statistics-oriented software, such as R. This makes it easier for us to test new research hypotheses, as before, using only XSLT. At the same time, according to specific research needs, we can also easily add new metadata about persons and organizations.

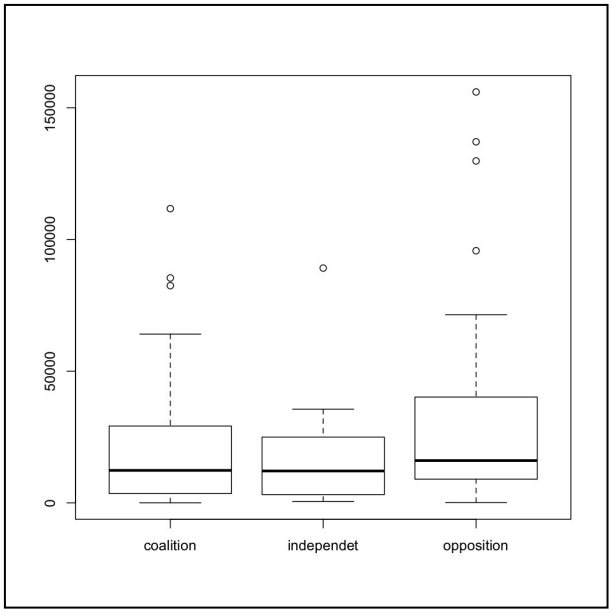


Figure 5: Number of words spoken in the Assembly of the Republic of Slovenia (May 8, 1990 – May 14, 1992) by organization membership.

6. Conclusions

The creation of SlovParl was in many aspects in accordance with good practices in the production of scholarly digital edition. This approach is particularly valuable in this regard:
"when digital editions are designed so that their textual data is captured using standards like TEI, this opens up

important opportunities for alternative deployments of the data." (MLA Commons, 2015)

In accordance with our basic principles, in the next years the corpus will not only be complemented with new parliamentary papers, but we will also pay special attention to research data reuse in different academic disciplines.

We hope that in this way we will be able to help other academic disciplines in tackling the shortage of not only these, but also related resources. At the moment, there are some larger projects aiming to collect and annotate similar political text resources. The Manifesto Project analyses parties' election manifestos[17] and the Comparative Agendas Project collects and organize data from archived sources to track policy outcomes across countries.[18] The results of these projects are of course also interesting for us. This is especially true for automatic topic classification of related language like Croatian (Karan et al., 2016). However, on the other hand, we believe that the topic classification from SlovParl can also provide a good basis for extension of contents analysis from only document titles to full text.

7. Acknowledgements

The work presented in this paper was supported by the Slovenian historiography research infrastructure (I0-0013), and the Slovenian ESFRI infrastructures DARIAH-SI and CLARIN.SI which are financially supported by the Slovenian Research Agency.

8. Bibliographical References

Beelen, K., Thijm, T.A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., and Whyte, T. (2017). Digitization of the Canadian Parliamentary Debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3): 849-864.

Benardou, A., Dunning, A., Schaller, M., and Chatzidiakou, N. (2015). Research Themes for Aggregating Digital Content: Parliamentary Papers in Europa. Europeana Cloud – Work Package 1.

Erjavec, T., Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Fišer, D., Laskowski, C., and Zupan, K. (2016) Annotating CLARIN.SI TEI corpora with WebAnno. In *Proceedings*, pp. 1-5. https://www.clarin.eu/sites/default/files/erjavec-etal-CLARIN2016_paper_17.pdf.

Gartner, R. (2014). A metadata infrastructure for the analysis of parliamentary proceedings. In *Big Humanities Data, The Second IEEE Big Data 2014 Workshop*. Bethesda, Maryland, USA.

Gašparič, J. (2017). Parlament im Übergang: Versammlung der Republik Slowenien zum Zeitpunkt des Verfalls des Sozialismus und Jugoslawiens als Gegenstand einer historischen Analyse. Lecture, Kommission für Geschichte des Parlamentarismus und der politischen Parteien, Berlin, Germany, September 21.

Gielissen, T. and Marx, M. (2009). Exemplification of Parliamentary Debates. In *Proceedings of the 9th Dutch-Belgian Information Retrieval Workshop*, DIR 2009, pp. 19-25.

Karan, M., Šnajder, J., Širinić, D. and Glavaš, G. (2016) Analysis of Policy Agendas: Lessons Learned from Automatic Topic Classification of Croatian Political Text. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Science, and Humanities (LaTeCH)*, pp. 12-21, Berlin, Germany, August 11.

Ljubešić, N. and Erjavec, T. (2016). Corpus vs. Lexicon Supervision in Morphosyntactic Tagging: the Case of Slovene. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Marx, M. (2009). Long, often quite boring, notes of meetings. In *ESAIR '09 Proceedings of the WSDM '09 Workshop on Exploiting Semantic Annotations in Information Retrieval*, pp. 46-53, Barcelona, Spain, February.

Marx, M. and Schuth, A. (2010). DutchParl: The Parliamentary Documents in Dutch. In *LREC 2010, Seventh International Conference on Language Resources and Evaluation*, pp. 3670–3677, Valleta, Malta, may. European Language Resource Association (ELRA).

McEnery, A. and Baker, H. (2017). *Corpus Linguistics and 17th-Century Prostitution: Computational Linguistics and History*. London and New York: Bloomsbury Academic.

MLA Commons (2015). Considering the Scholarly Edition in the Digital Age: A White Paper of the Modern Language Association's Committee on Scholarly Editions. https://scholarlyeditions.mla.hcommons.org/2015/09/02/cse-white-paper/.

Pančur, A. (2016) Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI = Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines. In *Proceedings of the Conference on Language Technologies & Digital Humanities*. Ljubljana: Ljubljana University Press, pp. 142-148.

Pančur, A. and Šorn, M. (2016) Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History. *Prispevki za novejšo zgodovino*. 56 (3): 130-146. http://ojs.inz.si/pnz/article/view/193.

Pesek, R. (2007). *Osamosvojitev Slovenije: 'Ali naj Republika Slovenija postane samostojna in neodvisna država?'*. Ljubljana: Nova revija.

Pierska, H., Tames, I., Buitinck, L., Doornik, J., and Marx, M. (2014). War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History.Digital Humanities Quarterly, 8(1). .

Robertson, S. (2016). The Differences between Digital Humanities and Digital History. In *Debates in Digital Humanities 2016*. Minneapolis and London: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/76.

Rychlý, P. (2007). Manatee/Bonito – A Modular Corpus Manager. Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing. 65–70. Brno, Masaryk University.

Spiro, L. (2014) Access, Explore, Converse: The Impact (and Potential Impact) of the Digital Humanities on

[17] https://manifesto-project.wzb.eu/.
[18] https://www.comparativeagendas.net/.

Scholarship. In *Keys for architectural history research in the digital era*. https://inha.revues.org/4925.
TEI Consortium. (2016). *Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/P5/.

9. Language Resource References

Alexander, M., Anderson, J., Archer, D., Baron, A., Davies, M., Hope, J., Jeffries, L., Kay, C., Rayson, P., and Walker, S. (2016). The Hansard Corpus: British Parliament. https://www.hansard-corpus.org.
Pančur, A., Šorn, M. and Erjavec, T. (2016). *Slovenian parliamentary corpus SlovParl 1.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1075.
Pančur, A., Šorn, M. and Erjavec, T. (2017). *Slovenian parliamentary corpus SlovParl 2.0*, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1167.
Pražák, A. & Šmídl, L. (2012). Czech Parliament Meetings, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.