Proceedings of ParlaCLARIN II Workshop, pages 28–34
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

The siParl corpus of Slovenian parliamentary proceedings

Andrej Pančur, Tomaž Erjavec
Institute of Contemporary History, Department of Knowledge Technologies, Jožef Stefan Institute
Privoz 11, SI-1000 Ljubljana, Slovenia, Jamova cesta 39, SI-1000 Ljubljana, Slovenia
andrej.pancur@inz.si, tomaz.erjavec@ijs.si

Abstract

The paper describes the process of acquisition, up-translation, encoding, annotation, and distribution of siParl, a collection of the parliamentary debates from the Assembly of the Republic of Slovenia 1990–2018, covering the period from just before Slovenia became an independent country in 1991, and almost up to the present. The entire corpus, comprising over 8 thousand sessions, 1 million speeches and 200 million words was uniformly encoded in accordance with the TEI-based Parla-CLARIN schema for encoding corpora of parliamentary debates, and contains extensive meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations. The corpus was also linguistically annotated using state-of-the-art tools. siParl is open source and maintained on GitHub with its major versions archived in the CLARIN.SI repository. It is also available for linguistic and content analysis through the on-line CLARIN.SI concordancers, thus offering an invaluable resource for scholars studying Slovenian political history.

Keywords: Slovenian parliamentary corpus, Text Encoding Initiative, StanfordNLP

1. Introduction

The unique content, structure and language of records of parliamentary debates are all factors make them an important object of study in a wide range disciplines in digital humanities and social sciences, such as political science (Van Dijk, 2010), sociology (Cheng, 2015), history (Pančur and Šorn, 2016), discourse analysis (Hirst et al., 2014), sociolinguistics (Rheault et al., 2016), and multilinguality (Bayley, 2004). Despite the fact that parliamentary discourse has become an increasingly important research topic in various fields of digital humanities and social sciences in the past 50 years (Chester and Bowring, 1962; Franklin and Norton, 1993), it has only recently started to acquire a truly interdisciplinary scope (Bayley, 2004). Recent developments enable cross-fertilization of linguistic studies with other disciplines and in-depth exploration of institutional uses of language, interpersonal behavior patterns, interplay between language-shaped facts, and reality-prompted language ritualisation and change (Ihalainen et al., 2016).

The most distinguishing characteristic of records of parliamentary debates is that they are essentially transcriptions of spoken language produced in controlled and regulated circumstances. For this reason, they are rich in invaluable (sociodemographic) meta-data. They are also easily available under various Freedom of Information Acts set in place to enable informed participation by the public and to improve effective functioning of democratic systems, making the datasets even more valuable for researchers with heterogeneous backgrounds.

This has motivated a number of national as well as international initiatives (for an overview, see Fišer and Lenardič (2018)) to compile, process and analyze parliamentary corpora. They are available for most European countries, with the UK's Hansard Corpus being the largest (1.6 billion tokens) and spanning the longest time period (1803–2005) while corpora from other countries are significantly smaller (most comprise between 10 and 100 million tokens) and cover significantly shorter periods (mostly from the 1970s onward).

There have also been two parliamentary corpora compiled in Slovenia, in particular the SlovParl corpus (Pančur et al., 2018) with about 10 million words, which contains minutes of the Assembly of the Republic of Slovenia for the legislative period 1990–1992 when Slovenia became an independent country, and the Parlameter corpus (Fišer et al., 2019) with about 40 million words, which covers the seventh mandate (2014–2018) of the Slovenian parliament. Both corpora are available under CC BY licences via the CLARIN.SI repository of language resources (Pančur et al., 2017; Dobranić et al., 2019).

In this paper we present the siParl corpus, which was made using the same basic workflow as for the SlovParl corpus, but encompasses, in addition to SlovParl, i.e. the period 1990–1992, also all the years up to 2018, leading to an (almost) comprehensive Slovenian parliamentary corpus, with close to 200 million words, making it one of the larger available parliamentary corpora.

Version 1.0 of this corpus was compiled and made available in the CLARIN.SI repository in 2019 (Pančur et al., 2019), while in this paper we present siParl version 2.0, which is also available in the repository (Pančur et al., 2020) and includes the same text but, in contrast to version 1.0, has much improved manually checked metadata, has been re-encoded to comply with the Parla-CLARIN recommendation for encoding of parliamentary corpora (Erjavec and Pančur, 2019), and has been newly linguistically annotated using state-of-the-art tools.

The rest of the paper is structured as follows: Section 2. overviews the compilation of the corpus and gives information on the corpus structure and size, Section 3. introduces the Parla-CLARIN encoding of the corpus, Section 4. explains its linguistic annotation, Section 5. describes how the corpus is distributed, and Section 6. gives some conclusions and directions for further research.

2. Corpus compilation and structure

In the design of siParl corpus, we attempted to satisfy the following desiderata:

1. **Multidisciplinary:** The corpus must be useful for as many disciplines as possible. To attain this goal, the siParl corpus (as well as this paper) was created in close cooperation between the Slovenian DARIAH and CLARIN infrastructures.
2. **All-inclusive:** In addition to parliamentary debates, other types of parliamentary papers are also planned to 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 the work of long-term research infrastructures.
4. **Open science:** All previous principles can be optimally realized in accordance with the principles of open science, i.e. the corpus should be made available under FAIR principles[1].

The transcriptions of parliamentary debates of the National Assembly of the Republic of Slovenia are available as HTML files on the web pages of National Assembly[2]. With the help of BeautifulSoup[3] and Python we scraped the wanted data from their website.

The uniform structure of documents with parliamentary debates is, in principle, well suited for automatic annotation. However, for the case of our source documents, it turned out that the HTML files for the period 1990–1996 do not contain born-digital text, thus being differently structured from the rest and had to be processed separately. The later documents also have problematic HTML markup, as 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, a rather complex and very time-consuming (the effort estimated at about 1.2 FTE) semi-automatic annotation needed to be performed in several steps, where each step contained:

1. developing XSLT stylesheets for automatic annotation;
2. developing XPath and regular expressions to search for annotation errors;
3. manual correction of identified errors.

It should also be noted that significant effort was invested in obtaining speaker metadata, such as their place and date of birth, the chronology of their party membership, link to their Wikipedia article etc., and that this information is also included in the corpus.

Parliamentary debates for a particular nation typically have a quite uniform structure, which fluctuates very little in time (Marx, 2009) and this also applies to Slovenian parliamentary debates. By analysing representative samples we arrived at the following general structure of parliamentary proceedings, given below with the minimal and maximal occurrences of structural elements:

[1] https://www.go-fair.org/fair-principles/

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

[3] https://www.crummy.com/software/BeautifulSoup/

Level	Count
Legislative periods	8
Sessions	8,571
Days	11,351
MPs	660
Speakers	8,418
Speeches	1,083,233
Words	200,406,464
Sentences	11,019,550
Words	195,296,618
Tokens	228,152,632

Table 1: Basic statistics of the corpus

- 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)

However, inside this general structure, that 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 an interrupted meeting. A 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 (i.e. metadata about the transcription, such as information about the meeting and chairperson, outcome of a vote, 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.

On the basis of the encoded corpus we computed some basic statistics over siParl, which are given in Table 1. The first part of the table contains a summary of the main aspects of the corpus, with the number of words encompassing the complete corpus. The second part concerns the summary of the automatic linguistic annotation (further detailed in Section 4.), where it should be noted that this annotation was preformed only on the transcription proper, i.e. the non-verbal content was omitted from this annotation.

Table 2 gives similar information but separately for each of the eight legislative periods, where we also make the split between the sittings of the National Assembly and that of other working bodies, such as various commissions; for these we give the number of such bodies for each period. It should be noted that the "Words" column here contains all the text, including that of non-verbal items.

Legislative period	Organisation	No. of organisations	Sessions	Days	Speakers	MPs	Items of agenda	Speeches	Non-verbal items	Words
1990—1992	National Assembly	3	234	608	521	242	5,154	59,062	120,824	11,131,908
1992—1996	National Assembly	1	94	462	315	101	2,864	66,555	152,698	11,698,884
1996—2000	National Assembly	1	76	430	359	105	3,256	55,852	133,164	10,763,395
	Working Bodies	27	1,274	1,711	1,268	108	6,327	155,514	216,135	17,953,666
2000—2004	National Assembly	1	89	303	296	105	2,636	56,157	91,089	10,358,962
	Working Bodies	27	1,129	1,405	1,291	102	4,855	126,014	187,105	16,408,922
2004—2008	National Assembly	1	82	283	237	102	2,770	63,443	79,326	13,312,828
	Working Bodies	25	1,211	1,300	1,886	101	4,779	115,260	175,972	18,323,287
2008—2011	National Assembly	1	84	233	191	103	2,340	51,381	62,262	11,712,103
	Working Bodies	25	1,204	1,308	2,144	102	4,508	82,380	123,984	17,031,791
2011—2014	National Assembly	1	96	196	202	114	1,912	37,073	43,925	8,205,794
	Working Bodies	28	1,136	1,125	1,994	111	3,831	68,777	99,783	14,784,252
2014—2018	National Assembly	1	104	288	207	101	2,809	52,268	61,837	14,516,417
	Working Bodies	22	1,758	1,699	2,697	100	5,606	93,497	139,834	24,204,255

Table 2: Basic statistics regarding different legislative periods

3. Corpus encoding

As mentioned, many researchers have already compiled corpora of parliamentary proceedings. However, these corpora are encoded in a variety of different annotation schemes, limiting their interchange and re-use. In order to overcome this problem, the CLARIN research infrastructure organised a workshop in 2019[4] at which the idea and draft of a common annotation scheme for encoding corpora of parliamentary proceedings was introduced, the participants presented their own experiences with encoding parliamentary corpora and gave their comments to the draft proposal. On this basis, guidelines and an XML schema, called Parla-CLARIN, was developed (Erjavec and Pančur, 2019), which is meant for encoding of parliamentary corpora for the purposes of scholarly investigations, and that could serve as a common storage and interchange format for such corpora. These recommendations attempt to take into account the following aspects of parliamentary corpora:

- Structure: legislative periods, sessions, topics, speeches, transcription variants
- Metadata: mandates, titles, parliamentary bodies, locations, dates and times
- Speakers: sex, date of birth, education, party membership, links to external resources
- Political parties: name(s), history, relations
- Speeches: speaker, text, comments, verbal and non-verbal interruptions
- Linguistic annotation: PoS tagging, word normalisation, named entity tagging, syntactic parsing etc.
- Multimedia: audio and video, facsimile of original transcript

The Parla-CLARIN recommendations are implemented as a parameterisation of the TEI Guidelines (TEI Consortium, 2011), which are XML-based recommendations for encoding texts for scholarly purposes. As opposed to most other such recommendations, the TEI Guidelines have the ambition to be applicable to texts in any language, of any date, and without restriction on form or content. There are a number of advantages of taking the TEI as the foundation of Parla-CLARIN. The recommendation does not need to specify and document a large number of elements, but only narrow down the choices offered by TEI and exemplify their use on concrete examples. As the CLARIN workshop showed, a number of existing parliamentary corpora are already encoded in some variant of TEI, making the conversion into a common TEI based-format much easier. The TEI parameterisation proposed for Parla-CLARIN allows a wide range of parliamentary proceedings to be encoded, while making explicit recommendations on the manner of encoding various phenomena.

Parla-CLARIN is written as a TEI ODD document, i.e. as a TEI document that contains both explanatory prose and the definition of the schema in the TEI ODD language. This document can be automatically converted either to a HTML view of the prose and schema parts, and to an XML schema. The recommendations are maintained on GitHub[5] from where they can be cloned, or read on the equivalent github.io pages.

[4] https://www.clarin.eu/blog/clarin-parlaformat-workshop

[5] https://github.com/clarin-eric/parla-clarin/

The presented siParl corpus is the first complete corpus that has been encoded according to the Parla-CLARIN recommendation, also in the hope that it will serve as a best-practice exemplar. Of course, siParl does not contain all the encoding supported by the Parla-CLARIN recommendation (e.g. multimedia, verbal interruptions), however, it does have rich metadata and linguistic annotation, so it can serve as a good example for the encoding recommendation. We next give some examples from the Parla-CLARIN encoded corpus[6]. First, complex metadata is encoded in various elements available in the teiHeader element, such as taxonomies and various types of lists offered by TEI. Figure 1 gives the example of the start of the event list element that contains the Slovene and English names of the eight legislative periods, treated as events, and their start and end dates; crucially, each period is also given its ID, which can then be referred to by other elements.

```
<listEvent>
 <head>Legislative periods</head>
 <event xml:id="DZ.1"
  from="1992-12-23" to="1996-11-27">
  <label xml:lang="sl">1. mandat</label>
  <label xml:lang="en">Term 1</label>
 </event>
 <event xml:id="DZ.2"
   from="1996-11-28" to="2000-10-26">
  <label xml:lang="sl">2. mandat</label>
  <label xml:lang="en">Term 2</label>
 </event>
 ...
```

Figure 1: Encoding of legislative periods.

Figure 2 illustrates the encoding of speakers, where, again, each person is given their ID that can then be referred to in the speeches, and then contains the person's basic metadata and their chronologically marked role(s) in the parliament with — where relevant — party affiliation(s), followed by link(s) to external resources, in particular, Wikipedia articles in Slovene, and, where available, also in English.

```
<person xml:id="ŠpiletičBogomir">
 <persName>
  <surname>Špiletič</surname>
  <forename>Bogomir</forename>
 </persName>
 <sex value="M"/>
 <birth when="1961-11-01"/>
 <death when="2013-06-17"/>
 <affiliation role="MP" ana="#DZ.2" ref="#DZ"
  from="1996-11-28" to="2000-10-26"/>
 <affiliation role="member" ana="#DZ.2"
  ref="#party.SDS.1"
  from="1996-11-28" to="2000-10-26"/>
 <idno type="wikimedia"
   xml:lang="sl">https:...</idno>
</person>
```

Figure 2: Encoding of person metadata.

[6]Note that for illustrative purposes some details of the encoding have been omitted, and the lines split, sometimes in places that would lead to ill-formed data.

Next, Figure 3 gives the encoding of one political party with its ID, name in Slovene and English, its acronym, the period of its existence, and link(s) to Wikipedia articles.

```
<org xml:id="party.SDS.1"
 role="political_party">
 <orgName full="yes"
  xml:lang="sl">Socialdemokratska
  stranka Slovenije</orgName>
 <orgName full="yes" xml:lang="en">Social
  Democratic Union of Slovenia</orgName>
 <orgName full="init">SDS</orgName>
 <event from="1989-02-16" to="2003-09-19">
  <label xml:lang="en">existence</label>
 </event>
 <idno type="wikimedia"
   xml:lang="sl">https:...</idno>
 <idno type="wikimedia"
   xml:lang="en">https:...</idno>
</org>
```

Figure 3: Encoding of party metadata.

Finally, Figure 4 gives the start of the body of one parliament session. The non-verbal events, i.e. transcription metadata is encoded in the note element, which are of various types, e.g. the first one giving the time when the session came to order and the second introducing the speaker. Each speech is marked by the utterance element, which gives the reference to the ID of the speaker and, in the analysis attribute, the reference to the role of the speaker inside the session.
Each speech is divided into segments, i.e. paragraphs as distinguished in the source transcriptions. These, in siParl, then have pure textual content with the exception of the gap element, which indicates that a part of the speech is missing, also giving the reason.

```
<body>
 <div>
  <note type="time">Seja se je pričela ob
  9.30 uri.</note>
  <note type="speaker">PREDSEDNIK RAFAEL
   KUŽNIK:</note>
   <u who="#KužnikRafael" ana="#chair">
   <seg>
    <gap reason="inaudible"/> in potem
    še točko razno.
    ...
```

Figure 4: Encoding of the transcription.

4. Linguistic annotation

The siParl corpus is available in two variants. The first, as introduced in the previous section, and manitaned on GitHub, contains meta-data, structural annotations, non-verbal items and speeches, and the plain-text of their segments. The second version is identical to the first, except that the segments have been linguistically annotated: each is tokenised, sentence segmented, part-of-speech tagged, lemmatised, parsed, and tagged with named entities. Such annotations significantly expand the possibilities of corpus analysis, in particular they allow to mount the corpus into

web-based concordancers, such as those of the CLARIN.SI infrastructure, which then support complex queries over sequences of token annotations and displaying their concordancers, frequency lists, keyword lists of selected parts of the corpus based on the metadata etc.

The main linguistic annotation of the corpus was performed by CLASSLA-StanfordNLP[7] (Ljubešić and Dobrovoljc, 2019), a fork of the well-known StanfordNLP library[8] (Qi et al., 2018), which, inter alia, supports part-of-speech tagging, lemmatisation and dependency parsing. As opposed to StanfordNLP, the CLASSLA-StanfordNLP fork introduces some extensions, such as using an external dictionary while performing lemmatisation, and training the tagger and lemmatiser on more data than available in Universal Dependencies treebanks.

The CLASSLA-StanfordNLP pipeline obtains significantly better performance than previous tools for Slovenian, e.g. the accuracy of predicting the fine-grained PoS tags was improved from the previous 94.21% (using a CRF tagger trained on the same resources) to 97.06%.

We have annotated the corpus for fine-grained part-of-speech, i.e. morphosyntactic descriptions (MSD) using the MULTEXT-East[9] (Erjavec, 2012) schema for Slovenian, as well as with the part-of-speech and morphological features in the Universal Dependencies formalism for Slovenian (Dobrovoljc et al., 2017). The corpus was also lemmatised, important for Slovenian, as it is a highly inflecting language. Finally, the tool also parsed the corpus using the Universal Dependencies formalism. The CLASSLA-StanfordNLP models used for morphosyntactic annotation, lemmatisation and parsing are available from the CLARIN.SI repository (Ljubešić, 2020c; Ljubešić, 2020b; Ljubešić, 2020a).

The corpus was also annotated for named entities, using the Janes-NER tool [10], which is CRF-based and uses a rather standard feature set relevant for identifying named entities, as well as distributional information in form of Brown clusters (Brown et al., 1992). The evaluation of the tool (Fišer et al., 2018) showed that it has an average F1 score of 0.69, with the "other" class having the lowest F1 = 0.30, followed by organisations with F1 = 0.56, locations with F1 = 0.80, and the person class having the highest F1 = 0.92.https://rdcu.be/7RX4

Figure 5 illustrates the encoding of segments with added linguistic analyses. Each segment is composed of sentences, and these of words and punctuation symbols; the fact that adjacent tokens are not separated by a space is indicated by the join attribute. Each token is then annotated by its MULTEXT-East MSD as the value of the ana attribute and using the extended pointer syntax offered by TEI; in short, an MSD with the "mte" prefix in effect points to the definition of the MSD which gives its decomposition into a feature structure containing its attributes and their values. The Universal Dependencies annotation is given as the value of the msd attribute. The syntactic dependencies are stored in the link group element, which contains links that connect the head and argument of the dependency relation, itself given in the ana attribute; again, the extended pointer syntax is used, to point to the full name of of each relation.

```
<seg xml:id="...seg8">
 <s xml:id="...seg8.s1">
  <gap reason="inaudible"/>
  <w xml:id="...seg8.s1.t1" ana="mte:Pr-nsa"
   msd="UposTag=PRON|Case=Acc|..."
   lemma="kar">Kar</w>
  <w xml:id="...seg8.s1.t2" ana="mte:Vmbm2p"
   join="right" msd="UposTag=VERB|.."
   lemma="izvoliti">izvolite</w>
  <pc xml:id="...seg8.s1.t3" ana="mte:Z"
   msd="UposTag=PUNCT">,</pc>
  <w xml:id="...seg8.s1.t4" ana="mte:Cs"
   join="right" msd="UposTag=SCONJ"
   lemma="da">da</w>
  <pc xml:id="...seg8.s1.t5" ana="mte:Z"
   msd="UposTag=PUNCT">.</pc>
  <linkGrp corresp="#...seg8.s1"
   targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:obj"
     target="#...seg8.s1.t2 #...seg8.s1.t1"/>
    <link ana="ud-syn:root"
     target="#...seg8.s1 #...seg8.s1.t2"/>
    <link ana="ud-syn:punct"
     target="#...seg8.s1.t4 #...seg8.s1.t3"/>
    <link ana="ud-syn:discourse"
     target="#...seg8.s1.t2 #...seg8.s1.t4"/>
    <link ana="ud-syn:punct"
     target="#...seg8.s1.t2 #...seg8.s1.t5"/>
  </linkGrp>
 </s>
</seg>
```

Figure 5: Linguistic annotation of the siParl corpus segment "*Kar izvolite, da.*"

5. Availability and maintenance

In accordance with our fourth basic principle (open science), we have made sure that the corpus is openly available, can be further developed in a collaborative fashion, has been converted into several immediately usable formats and, for the purposes of digital humanities and social sciences, also available through web applications.

As mentioned, the plain text version of the Parla-CLARIN encoded corpus is accessible and maintained on the DARIAH-SI GitHub repository[11]. This is also the place for users to post issues about the corpus or even send pull requests. It should be noted that while this project does not, due to its size, contain the linguistically annotated corpus, it does contain a folder with example documents and the scripts for annotation and conversion.

As mentioned, the major 2.0 version of the corpus is also available via the CLARIN.SI repository[12] (Pančur et al.,

[7] https://github.com/clarinsi/classla-stanfordnlp

[8] https://stanfordnlp.github.io/stanfordnlp/

[9] http://nl.ijs.si/ME/

[10] https://www.github.com/clarinsi/janes-ner

[11] https://github.com/DARIAH-SI/siParl

[12] http://hdl.handle.net/11356/1300

2020) under the Creative Commons CC BY licence. This repository item comprises six datasets:

1. the Parla-CLARIN encoded plain-text corpus (essentially a copy of the corpus from GitHub);
2. the Parla-CLARIN encoded linguistically analysed corpus;
3. the linguistically analysed corpus in the so called vertical format, used by various concordancers (this is a much simpler format to use than the source TEI, but does not contain all the information from the source);
4. the text of the linguistically analysed corpus in the CONLL-U format, used by the Universal Dependencies project;
5. the plain text of the linguistically analysed corpus;
6. TSV files giving the metadata of all sessions / speeches in the corpus.

The linguistically annotated version of siParl has also been mounted under the two concordancers available at CLARIN.SI, namely KonText and noSketch Engine, enabling on-line exploration of this and other corpora. The two concordancers are open source 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.

6. Conclusions

The paper presented siParl, a corpus of Slovenian parliamentary debates spanning the complete history of Slovenia as an independent country up to 2018. The corpus is fairly large with about 200 million words, and is, given the large amount of manual editing, fairly error-free. siParl 2.0 has been encoded according to the TEI-based Parla-CLARIN annotation scheme, is linguistically annotated with state-of-the-art tools for Slovenian, and is openly available via the CLARIN.SI repository and concordancers.

The corpus development is a good example of the possibilities of cooperation between the two distinct, but related research infrastructures, namely the Slovenian DARIAH and CLARIN, which is esp. obvious in the various distribution modes and formats of the corpus; here it should be noted that we also plan to mount the corpus in a form amiable for reading and browsing in the DARIAH-SI digital library and interconnect the corpus there with the one available via the concordancers.

In further work we plan to extend the corpus in three directions: include in it, in addition to the parliamentary debates, also other types of parliamentary papers, such as voting results, legislation, and summary records of meetings; extend it to include the materials from 2019 and 2020; and to include materials from before 1991, i.e. from the time when Slovenia was a part of the Socialist Republic of Yugoslavia. We also plan to further refine the encoding in combination with updating the Parla-CLARIN recommendation, which might become necessary when we consider other corpora of parliamentary debates to be included as exemplars for the proposed encoding.

Finally, we plan to make some effort in popularising the corpus among potential users from the fields of history, political science and linguistics, using the DARIAH-SI and CLARIN.SI dissemination networks and various local events.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful suggestions on how to improve the paper. The work presented here was funded by the Slovenian research infrastructures DARAH-SI and CLARIN.SI and by the Slovenian Research Agency within the research program P2-0103 “Knowledge Technologies”, and research infrastructure program I0-0013 “Slovenian historiography research infrastructure“.

7. Bibliographical References

Bayley, P. (2004). *Cross-cultural perspectives on parliamentary discourse*, volume 10. John Benjamins Publishing.

Brown, P. F., Desouza, P. V., Mercer, R. L., Pietra, V. J. D., and Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–479.

Cheng, J. E. (2015). Islamophobia, muslimophobia or racism? parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5):562–586.

Chester, D. N. and Bowring, N. (1962). *Questions in parliament*. Clarendon Press.

Dobrovoljc, K., Erjavec, T., and Krek, S. (2017). The universal dependencies treebank for Slovenian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 33–38, Valencia, Spain, April. Association for Computational Linguistics.

Erjavec, T. and Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings. In *TEI members meeting: What is text, really? TEI and beyond. Book of Abstracts*. University of Graz, September. https://gams.uni-graz.at/o:tei2019.157.

Erjavec, T. (2012). MULTEXT-East: morphosyntactic resources for Central and Eastern European languages. *Language Resources and Evaluation*, 46(1):35–57. https://doi.org/10.1007/s10579-011-9174-8.

Fišer, D., Ljubešić, N., and Erjavec, T. (2018). The Janes project: language resources and tools for Slovene user

generated content. *Language Resources and Evaluation*. https://rdcu.be/7RX4.

Fišer, D. and Lenardič, J. (2018). Parliamentary corpora in the CLARIN infrastructure. In *Selected papers from the CLARIN Annual Conference 2017, Budapest, 18–20 September 2017*, pages 75–85. Linköping University Electronic Press.

Fišer, D., Ljubešić, N., and Erjavec, T. (2019). Parlameter – a corpus of contemporary Slovene parliamentary proceedings. *Prispevki za novejšo zgodovino*, 59(1):70–98. http://ojs.inz.si/pnz/article/view/327/615.

Franklin, M. N. and Norton, P. (1993). *Parliamentary Questions: For the Study of Parliament Group*. Oxford University Press, USA.

Hirst, G., Feng, V. W., Cochrane, C., and Naderi, N. (2014). Argumentation, ideology, and issue framing in parliamentary discourse. In *ArgNLP*.

Ihalainen, P., Ilie, C., and Palonen, K. (2016). *Parliament and Parliamentarism: A Comparative History of a European Concept*. Berghahn Books.

Ljubešić, N. and Dobrovoljc, K. (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, August. Association for Computational Linguistics.

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*, pages 46–53, Barcelona, Spain, February.

Pančur, A. and Šorn, M. (2016). Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History. *Prispevki za novejšo zgodovino/Contributions to Contemporary History*, 56(3):130–146.

Pančur, A., Šorn, M., and Erjavec, T. (2018). SlovParl 2.0 : The collection of Slovene parliamentary debates from the period of secession. In Darja Fišer, et al., editors, *Proceedings of LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, pages 8–14, Paris, France, May. European Language Resources Association (ELRA). http://lrec-conf.org/workshops/lrec2018/W2/summaries/4_W2.html.

Qi, P., Dozat, T., Zhang, Y., and Manning, C. D. (2018). Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Rheault, L., Beelen, K., Cochrane, C., and Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, 11(12):e0168843.

Rychlý, P. (2007). Manatee/Bonito - A Modular Corpus Manager. In *1st Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 65–70, Brno. Masarykova univerzita.

TEI Consortium, e. (2011). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/P5/.

Van Dijk, T. A. (2010). Political identities in parliamentary debates. *European Parliaments under Scrutiny. Discourse strategies and interaction practices*, pages 29–56.

8. Language Resource References

Dobranić, Filip and Ljubešić, Nikola and Erjavec, Tomaž. (2019). *Slovenian parliamentary corpus ParlaMeter-sl 1.0*. Jožef Stefan Institute, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1208.

Ljubešić, Nikola. (2020a). *The CLASSLA-StanfordNLP model for for UD dependency parsing of standard Slovenian*. Jožef Stefan Institute, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1258.

Ljubešić, Nikola. (2020b). *The CLASSLA-StanfordNLP model for lemmatisation of standard Slovenian 1.1*. Jožef Stefan Institute, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1286.

Ljubešić, Nikola. (2020c). *The CLASSLA-StanfordNLP model for morphosyntactic annotation of standard Slovenian*. Jožef Stefan Institute, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1251.

Pančur, Andrej and Šorn, Mojca and Erjavec, Tomaž. (2017). *Slovenian parliamentary corpus SlovParl 2.0*. Institute of Contemporary History, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1167.

Pančur, Andrej and Erjavec, Tomaž and Ojsteršek, Mihael and Šorn, Mojca and Blaj Hribar, Neja. (2019). *Slovenian parliamentary corpus siParl 1.0 (1990-2018)*. Institute of Contemporary History, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1236.

Pančur, Andrej and Erjavec, Tomaž and Ojsteršek, Mihael and Šorn, Mojca and Blaj Hribar, Neja. (2020). *Slovenian parliamentary corpus siParl 2.0 (1990-2018)*. Institute of Contemporary History, distributed via the Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1300.