Journal of the Text Encoding Initiative

Issue 14 | 2021
Selected Papers from the 2019 TEI Conference

The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings

Tomaž Erjavec and Andrej Pančur

Electronic version
URL: https://journals.openedition.org/jtei/4133
DOI: 10.4000/jtei.4133
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Tomaž Erjavec and Andrej Pančur, "The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings", *Journal of the Text Encoding Initiative* [Online], Issue 14 | 2021, Online since 28 April 2022, connection on 28 September 2022. URL: http://journals.openedition.org/jtei/4133 ; DOI: https://doi.org/10.4000/jtei.4133

The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings

Tomaž Erjavec and Andrej Pančur

ABSTRACT

Parliamentary proceedings are a rich source of data that can be used by scholars in various humanities and social sciences disciplines. Unlike the sources of most other language corpora, parliamentary proceedings are not subject to copyright or personal privacy protections, and are typically available online, thus making them ideal for compilation into corpora and for open distribution. For these reasons many countries have already produced corpora of parliamentary proceedings, but each typically in their own encoding, limiting their comparability and utilization in a multilingual setting. In this paper we propose an encoding schema which could serve as an interchange format for parliamentary corpora compiled for the purposes of scholarly investigations. The schema, called Parla-CLARIN, was developed within the CLARIN research infrastructure, and is written as a TEI ODD which includes a TEI customization and prose

guidelines with examples of use. We discuss the coverage and choices made in designing the recommendations, and give an overview of the guidelines. We also discuss two other standard schemas for encoding parliamentary data, Akoma Ntoso and RDF, and their relation to Parla-CLARIN. We conclude by presenting corpora already encoded in Parla-CLARIN and discussing further work, especially the provision of a set of example documents and of transformation scripts that would make the proposed encoding more usable.

INDEX

Keywords: parliamentary corpora, encoding recommendations, TEI ODD

ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers for their helpful comments and suggestions, Jan Odijk for his leadership in proposing and organizing the Amersfoort CLARIN ParlaFormat Workshop, and the workshop participants for their constructive criticisms on the draft Parla-CLARIN proposal. The work presented here was supported by CLARIN ERIC through the CLARIN Type B ParlaFormat Workshop, through the CLARIN ParlaMint project, and by the CLARIN.SI Slovenian node of CLARIN and the DARIAH-SI Slovenian node of the DARIAH research infrastructure.

1. Introduction

1 The unique content, structure, and language of records of parliamentary debates make them an important object of study in a wide range of disciplines in the digital humanities and social sciences, such as political science (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). With parliaments in Europe playing an increasingly decisive role in their constituents' lives and because of their rapidly changing relations with the public, mass media, executive branches, and international organizations, further empirical research and development of integrative analytical tools are necessary in order to achieve a better

understanding of parliamentary discourse as well as its wider societal impact, in particular with studies that represent diverse parts of society (women, minorities, and marginalized groups) and cross-cultural studies (Hughes et al. 2015).

2 The most distinguishing characteristic of records of parliamentary debates is that they are essentially transcriptions of spoken language produced in controlled and regulated circumstances and are rich in valuable (sociodemographic) metadata. 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. For these reasons, and because parliamentary proceedings are often available online, many researchers have already compiled corpora of parliamentary proceedings.[1] However, these corpora are encoded using a variety of different annotation schemes, limiting their interchange and reuse.

3 In order to overcome this problem, the CLARIN Research Infrastructure for Language Resources and Technologies organized the ParlaFormat Workshop (May 23–24, 2019, Amersfoort)[2] at which the idea and draft of a TEI-based 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 on the proposal. On this basis we developed the Parla-CLARIN recommendations for encoding parliamentary corpora which we present in this paper. These recommendations aim to provide community-based reference TEI customizations for a specific application, following similar prior initiatives, from EpiDoc[3] for epigraphic documents to TEI Lex-0[4] for dictionaries.

4 In the process of the developing the recommendations, we also produced the second version of the siParl corpus (Pančur and Erjavec 2020; Pančur et al. 2020), a carefully encoded and linguistically annotated collection of parliamentary debates from the Assembly of the Republic of Slovenia from 1990 to 2018, with over 1 million speeches and 200 million words. This was the first substantial corpus to be fully Parla-CLARIN encoded, where we had the good fortune that the first version of siParl was the basis for an extended tutorial on how corpora can be used to investigate language use and communication practices in the context of political discourse (Fišer and Pahor de Maiti

2020). The work on this tutorial revealed various shortcomings of the first version, which were addressed in the encoding and structure of siParl 2.0 and also informed the development of the Parla-CLARIN recommendations.

5 Section 2 gives an overview of the Parla-CLARIN proposal, section 3 presents related work, and section 4 gives conclusions and directions for further work.

2. Overview of the Parla-CLARIN Proposal

6 The Parla-CLARIN recommendations are implemented as a customization of the TEI Guidelines that allows a wide range of corpora of parliamentary proceedings to be encoded and makes explicit prose recommendations on the manner of encoding various phenomena. In particular, the 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: date of birth, sex, education, party membership, links to external resources
- Political parties: name(s), history, relations
- Speeches: speaker, text, transcriber comments, verbal and nonverbal interruptions
- Linguistic annotation: part-of-speech (PoS) tagging, named entity annotation, syntax, etc.
- Multimedia: audio and video, facsimile of original transcript

2.1 The Parla-CLARIN Schema

7 Parla-CLARIN is written as a TEI ODD document, consisting of the prose guidelines and the schema specification, on the basis of which it is possible, using the standard TEI XSLT stylesheets, to derive an XML schema expressed either as a RelaxNG schema, a DTD, or a W3C schema, which is then used for formal validations of a Parla-CLARIN parliamentary corpus.

8 While the proposal tries to cater for many encoding needs, it is possible that new users will have to use TEI elements or attributes that are not discussed in the prose guidelines. Since the recommendations are still under development, the formal schema specification has been left rather unconstrained, so it can accommodate encoding practices that have not yet been foreseen. The downside of this approach is that the schema allows constructs that are at odds with those

proposed in the prose of the recommendations. Currently the prose should therefore be taken as the definitive way of encoding the phenomena under discussion: that is, even if a corpus validates against the schema, it might still not be encoded according to the recommendations.

9 The Parla-CLARIN schema uses the following TEI modules:

- the obligatory `tei`, `header`, `core`, and `textstructure` modules defining the elements used in any TEI-encoded text;
- the `corpus` module to encode the root `<teiCorpus>` element and the details of various metadata of the corpus;
- the `spoken` module for encoding speech corpora, so that speeches are encoded as utterance elements;
- the `figures` module, which, among other things, defines the elements for encoding tables, as these can also appear in parliamentary proceedings;
- the `namesdates` module, which defines a wealth of elements and attributes to describe people and organizations;
- the `transcr` module, which is used for transcription of primary sources, such as manuscripts, and is useful for encoding historical parliamentary corpora where the facsimile is of interest;
- the `linking` module, which defines useful attributes for linking elements of the corpus;
- the `analysis` and `iso-fs` modules, the former for encoding linguistic analyses of the transcriptions and the latter for defining formal pieces of data, such as decomposing morphosyntactic descriptions into their features;
- the `gaiji` module for encoding special characters that are especially likely to appear in historical parliamentary proceedings.

10 The crucial choice above was to use the `spoken` module for encoding speech corpora, so that parliamentary debates are encoded, essentially, as a speech corpus.[5] The competing options would be to encode them either using the default text structure elements (`<div>` and `<p>` or `<ab>`), or to use the `drama` module, that is, treating the debates as a play. The first option gives TEI elements that are too generic, which could encode the semantics of various structures only with many type attributes, while the second, maybe surprisingly, maps quite well onto the structure of parliamentary debates and is, in fact, used as the base encoding of siParl. However, the semantics

of the elements made for performance texts are nevertheless rather far from the reality of parliamentary debates, and often require mixed content, which is best avoided if the text is also to be linguistically annotated. Importantly, the `speech` module also supports encoding the alignment of the utterances (or other contained elements) with the speech or video signal. That the choice of this module is the correct one is also supported by the overview of the encodings of other parliamentary corpora (see section 3), where the majority use TEI `speech`.

11 Unlike some other TEI customizations, such as the one for encoding Computer-Mediated Communication proposed by the TEI CMC SIG,[6] the Parla-CLARIN schema does not introduce new elements, but rather uses existing TEI ones. The advantages of this approach are that the documentation is already directly available and that standard TEI-aware tools can be used to process the elements. On the downside, the elements used might not have as specific semantics as would be desired in the context of parliamentary corpora, and their content models can be too complex or too general. For the modules used, however, the schema does disallow certain elements that we assume that would never be used in the context of parliamentary corpora, and the schema also introduces recommended values for certain attributes, such as for the `div/@type` attribute.

2.2 The Parla-CLARIN Guidelines

12 The Parla-CLARIN guidelines—that is, the prose recommendations—are, to an extent, based on "Best Practices for TEI in Libraries" (Hawkins et al., 2018) in that they also give general recommendations for such matters as encoding relating to characters and filenames and how to document the encoding process and languages used. After such introductory considerations, the guidelines comprise the following chapters:

- *overall document structure*, which explains how the corpus should be structured, what the text divisions look like and how they should be typed, and how to encode document variants, which are meant for cases where both versions of parliamentary proceedings are to be encoded—that is, the original, "raw" transcription and the edited, "redacted" transcription;
- *corpus metadata*, which can contain significant and valuable information about the proceedings, with descriptions and examples given on how to encode information about the speakers and parties, and the relationships between the speakers and parties;
- *transcriptions*, where the encoding of the utterances (speeches) and commentary (notes by the transcribers) is explained, including specifics such as the encoding of interrupted utterances, verbal and nonverbal incidents, and how to encode voting results;
- *linguistic annotation*, which details the encoding of linguistic features added to the utterances and is further divided into word-level annotation (e.g., part-of-speech tagging), segmental annotation (e.g., named entities), and linking annotation (e.g., dependency syntax);
- *multimedia*, explaining how to align speech or video recordings, as well as facsimiles, to the transcription.

13 Finally, the guidelines also contain a chapter on conversions to or from other formats. This chapter, and especially the accompanying code, should be significantly extended, as having good conversion procedures is crucial to ensure the viability of the proposal. Currently, only one conversion procedure has been developed, namely that of converting parliamentary debates encoded in Akoma Ntoso to Parla-CLARIN, a topic that we discuss further in section 3.

14 In the Parla-CLARIN guidelines, every mention of an element is linked to its definition, where examples of use are also given, and the text also makes frequent reference to the text of the TEI Guidelines. However, the TEI Guidelines give generic examples and explanations, which can be at odds with particular recommendations that are made for Parla-CLARIN, so the ones from the TEI Guidelines should in this context be taken with a grain of salt.

2.3 Presentation of Parla-CLARIN

15 Like the TEI Guidelines, the Parla-CLARIN recommendations are available on GitHub, as a project[7] of the CLARIN ERIC collection. The project contains a folder for the schema (i.e., the Parla-CLARIN ODD document and XML schemas derived from it), a folder for the programs that convert the ODD into the XML schemas and to the HTML of the prose and schema definitions, and a folder for examples, which contains an artificial but fully worked out example of a Parla-CLARIN document and subfolders with various example resources, where each should contain:

1. a sample of a corpus in its source encoding;
2. XSLT script to convert it into Parla-CLARIN; and
3. the output of the conversion.

16 Currently, the folder contains examples of Akoma Ntoso encodings and two Slovene corpora of parliamentary proceedings: SlovParl (Pančur, Šorn, and Erjavec 2017) and siParl version 1.0 (Pančur et al. 2019).

17 Furthermore, the project contains a document folder with the HTML of the recommendations, which can be read online via the github.io pages.[8] Finally, the project has a wiki, which gives the technical details on how to validate a corpus against the Parla-CLARIN XML schema, how to add example documents to the project, and how to change the ODD in case it turns out to be insufficient for the modeling needs of a particular corpus.

3. Related Work

18 Before embarking on designing the Parla-CLARIN recommendations, we studied which other formats are used for encoding parliamentary corpora to gauge whether TEI-based recommendations would make sense. First, as shown in table 1, and as evidenced by the overview of parliamentary corpora by Fišer and Lenardič (2018)[9] and a survey of those presented by the

participants of the ParlaFormat Workshop,[10] we found that the majority of such corpora already use TEI in some guise or another when we counted "TEI-inspired" encodings, which, even though they use their own schemas, use elements with names and semantics similar to the ones used by TEI.

Table 1. Overview of encodings of selected parliamentary corpora.

Corpus name	Format	Reference
Polish Parliamentary Corpus	TEI speech	(Ogrodniczuk 2018; Ogrodniczuk and Nitoń 2020)
ParlaMeter-sl and ParlaMeter-hr	TEI speech	(Ljubešić et al. 2018; Dobranić, Ljubešić, and Erjavec 2019)
Debates on Europe at the Bundestag	TEI speech	(Truan 2016)
TAPS-fr	TEI speech	(Diwersy and Luxardo 2020)
ParlAT	TEI-inspired	(Wissik and Pirker 2018)
DutchParl	TEI-inspired	(Marx and Schuth 2010)
GermaParl	TEI-inspired	(Blätte and Blessing 2018)
German Political Speeches	TEI-inspired	(Barbaresi 2018, 2019)
Danish Parliament Corpus	custom XML	(Hansen 2018)
Czech Parliamentary Meetings	TRS XML	(Pražák and Šmídl 2012)
Corpus of the Saeima	RDF, CoNLL-U	(Darģis et al. 2018)
The Debates of the European Parliament	RDF	(Aggelen et al. 2017)[11]
Swedish Parliamentary Data	Prolog	(Eide 2019)

19 This was, of course, encouraging, as the barriers to converting from such encodings into Parla-CLARIN are much lower than those to converting from completely unrelated encoding systems. Nevertheless, there are two other formats that each also have a significant following and are contenders for being or becoming a standard format for parliamentary corpora.

3.1 Akoma Ntoso

20 Akoma Ntoso[12] was explicitly developed as an XML format for encoding legislative and judiciary documents, including parliamentary proceedings, is an OASIS standard, and has already been used to encode various legal documents in a number of countries. Akoma Ntoso defines an XML schema (called AKN) for modeling legal documents, uses FRBR (Functional Requirements for Bibliographic Records) concepts, and has a built-in relationship to ontologies. While these are all good reasons for AKN to be the interchange format for parliamentary corpora, there are, we feel, nevertheless a number of reasons why TEI-based recommendations are preferable.

21 First, corpus compilers and users are typically unfamiliar with AKN but—as shown in table 1—relatively familiar with TEI. It should also be noted that for most corpus compilers, parliamentary corpora will be only one type of corpus they will be compiling, so it is somewhat unrealistic for them to learn AKN and develop conversion scripts from it for just one type of corpus; on the other hand, TEI can be used for practically any type of corpus.

22 Second, AKN makes no provisions for storing speaker metadata, which is instead accessed from external data sources, using an AKN-specific referencing system; on the other hand, TEI has a number of elements for recording details about persons. For reasons of completeness, uniform and easier processing, and experimental replicability it is better to include such data directly in the corpus.

23 Third, AKN has no built-in support for linguistic annotation (apart from named entities). And while it would be possible to add elements for such annotation via a different namespace to AKN, AKN has no provisions for extending its schema, while TEI already has such elements available. As Parla-CLARIN is focused on corpora of parliamentary proceedings, and it is difficult to imagine a corpus without any linguistic annotation, TEI seems a better choice.

24 Nevertheless, AKN is an important schema for modeling parliamentary proceedings, especially as the primary encoding standard used by various legislative bodies, so some of AKN's solutions were used in developing the Parla-CLARIN proposal, in particular the typology of divisions of a document. Also developed was a partial, but non-trivial, conversion from AKN to Parla-CLARIN, which covers several AKN example documents. As mentioned in section 2.3, the example documents and conversion script can be found in the `Examples` folder of the Parla-CLARIN Git repository. The `akn2tei.xsl` script attempts to preserve the IDs of the source AKN document, converts the AKN addressee, role, and questions and answers to Parla-CLARIN, and maps FRBR data (which distinguishes a "work" from its "expression" and its expression from its "manifestation") to the appropriate TEI elements and attributes. For the FRBR mapping, we used the already mentioned recommendations of Best Practices for TEI in Libraries (Hawkins et al. 2018), in particular those in section 4.1.5, "The TEI Header and FRBR,"[13] which recommends that FRBR information is encoded in a `<listRelation>` element within `<sourceDesc>`. Unlike the original AKN identification element, `<listRelation>` contains a simple list of relation elements, so these must also specify the relation between the particular piece of data and the fact that it belongs to the FRBR "work," "expression," or "manifestation." These and other formalized relations are taken from the formal vocabularies of W3C (for RDF and OWL) and (via purl.org) of Dublin Core and vocab.org.

3.2 RDF

25 The Resource Description Framework (RDF[14]) is a W3C specification and a standard model for computer-processable data interchange on the web. It is also the base format for modeling information in the context of Linked Open Data (LOD), an influential model for linking data on the web. While most LOD datasets do not focus on language data, there does exist the Linguistic Linked Open Data Cloud (LLOD[15]) initiative, which is explicitly aimed at language resources and is active, for example, in the scope of the COST Action NexusLinguarum, "European network for Web-centred linguistic data science."[16] RDF/LOD has also been used for modeling parliamentary debates, in particular the Debates of the European Parliament as Open Data (Aggelen et al. 2017).

26 Again, the question is, why not use RDF to encode parliamentary corpora, rather than developing a TEI-based solution? Partially it is the community addressed by Parla-CLARIN: while LOD is targeted at computer scientists and assumes a machine-processable and -linked World Wide Web, TEI addresses mostly researchers from the digital humanities and models internally complete resources, while XML is also more popular than RDF in the NLP community. In practical terms, RDF data are machine-processable but hardly human-readable, and highly connected to external data sources; TEI documents, in contrast, are relatively self-explanatory, especially after some exposure to the TEI Guidelines, can be edited in any XML editor, and are mostly self-contained. Furthermore, while LLOD is indeed focused on language data, it has so far been used mostly to encode lexical resources, and its treatment of corpora has so far been concerned only with encoding linguistic annotations (Declerck et al. 2020). Detailed encodings of particular types of language resources are thus still lacking, and the same can be said for RDF attempts at encoding parliamentary debates, which model only rather shallow aspects of such data.

27 Our position is that RDF could be a useful downstream encoding for exploiting such corpora, and it would be a worthwhile exercise to develop a TEI-to-RDF conversion. Some prior work does exist on this topic (Chiarcos and Ionov 2019) and the best way seems to be to use RDFa attributes `@about`, `@property`, and `@resource` on TEI elements.

4. Conclusions

28 In this paper we have given an overview and presented the motivation and justification for developing a proposal for a standard interchange encoding scheme for corpora of parliamentary proceedings meant for scholarly investigations. This scheme is currently a straightforward customization of the TEI Guidelines, with the majority of the effort having gone into the writing of the prose guidelines of the Parla-CLARIN recommendations and into developing the conversion from Akoma Ntoso to Parla-CLARIN. We have not included examples of the encoding, as these are readily available on the GitHub documentation page of the project, and large Parla-CLARIN encoded corpora are openly available.

29 Apart from the siParl 2.0 corpus mentioned above (sections 1 and 2), the recommendations have already been used to encode the Czech (Hladka, Kopp, and Straňák 2020) and Icelandic (Steingrímsson, Barkarson, and Örnólfsson 2020) parliamentary corpora. Furthermore, the

ParlaMint project[17] has so far compiled four comparable Parla-CLARIN encoded parliamentary corpora (Bulgarian, Croatian, Polish, and Slovenian) containing sessions from 2015 to 2020 with about 20 million words each (Erjavec et al. 2020). In all four cases, the corpora are based on previously compiled corpora (Osenova and Simov 2012; Dobranić, Ljubešić, and Erjavec 2019; Ogrodniczuk and Nitoń 2020; Pančur et al. 2020) but reduced to include only sessions from 2015 on, extended with the latest sessions, recoded in Parla-CLARIN, and linguistically annotated for Universal Dependencies[18] and named entities with state-of-the-art tools.

30 As we wanted to have corpora that are not only interchangeable but interoperable as well, we created a bespoke ParlaMint XML schema directly in RelaxNG – the schema is compatible with Parla-CLARIN as it validates a subset of documents that would be validated against Parla-CLARIN. We produced common scripts that can convert any of the four corpora to plain text, to CoNLL-U format as used by the Universal Dependencies project, and to vertical format as used by the CWB[19] and Sketch Engine[20] (Kilgarriff et al. 2014) concordancers, as well as to extract complete speech metadata into TSV files.

31 In order for Parla-CLARIN to achieve its goal of becoming a widely recognized encoding format for corpora of parliamentary proceedings, significant work remains to be done. On the basis of the lessons learned in creating ParlaMint, we plan to revise the prose recommendations and to write a more prescriptive TEI ODD that will allow, as much as possible, only elements, attributes, and content models that are in line with the recommendations. Furthermore, we plan to change the examples given in the schema specification from the default ones in the TEI Guidelines to ones taken or adapted from the collected parliamentary corpora.

32 Second, as we have already done for ParlaMint, we plan to add to the GitHub Parla-CLARIN project more down-conversion scripts with which we would increase the usability of the Parla-CLARIN corpora. As mentioned, work also needs to be done to develop a conversion to RDF.

33 Last, but not least, one of the great benefits of Git is the ability to support collaborative work, be it through posting issues, or through using pull requests to incorporate changes. While the community has not so far made use of these options, we hope that Parla-CLARIN will eventually become a collaborative effort of those interested in compiling parliamentary corpora.

BIBLIOGRAPHY

Aggelen, Astrid van, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. 2017. "The Debates of the European Parliament as Linked Open Data." *Semantic Web – Interoperability, Usability, Applicability.* Vol. 8, No. 2. IOS Press. Revised version submitted January 21, 2016. Preprint available at http://www.semantic-web-journal.net/system/files/swj1300.pdf.

Barbaresi, Adrien. 2018. "A Corpus of German Political Speeches from the 21st Century." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 792–97. Paris: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/324.html.

———. 2019. *German Political Speeches Corpus and Visualization*. 4th release, June 17, 2019. https://politische-reden.eu/.

Bayley, Paul. 2004. "Introduction: The Whys and Wherefores of Analyzing Parliamentary Discourse." In *Cross-Cultural Perspectives on Parliamentary Discourse*, edited by Paul Bayley, 1–44. Philadelphia: John Benjamins Publishing.

Blätte, Andreas, and Andre Blessing. 2018. "The GermaParl Corpus of Parliamentary Protocols." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, edited by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, et al., 810–16. Paris: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/1024.html.

Cheng, Jennifer E. 2015. "Islamophobia, Muslimophobia or Racism? Parliamentary Discourses on Islam and Muslims in Debates on the Minaret Ban in Switzerland." *Discourse & Society* 26 (5): 562–86. doi:10.1177/0957926515581157.

Chiarcos, Christian, and Max Ionov. 2019. "Linking the TEI: Approaches, Limitations, Use Cases." Paper presented at Digital Humanities Conference, Utrecht, July 9–12, 2019. https://dev.clariah.nl/files/dh2019/boa/0910.html.

Darģis, Roberts, Ilze Auziņa, Uldis Bojārs, Pēteris Paikens, and Artūrs Znotiņš. 2018. "Annotation of the Corpus of the Saeima with Multilingual Standards." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 39–42. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/21_W2.html.

Declerck, Thierry, John McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel, Philipp Cimiano, et al. 2020. *Recent Developments for the Linguistic Linked Open Data Infrastructure*. Geneva: Zenodo. doi:10.5281/zenodo.3934626.

Dijk, Teun A. van 2010. "Political Identities in Parliamentary Debates." In *European Parliaments under Scrutiny: Discourse Strategies and Interaction Practices*, edited by Cornelia Ilie, 29–56. Amsterdam: John Benjamins Publishing.

Diwersy, Sascha, and Giancarlo Luxardo. 2020. "Querying a Large Annotated Corpus of Parliamentary Debates." *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 75–79. Paris: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.parlaclarin-1.13.

Dobranić, Filip, Nikola Ljubešić, and Tomaž Erjavec. 2019. *Croatian Parliamentary Corpus ParlaMeter-hr 1.0*. Slovenian language resource repository CLARIN.SI. Ljubljana: Jožef Stefan Institute. http://hdl.handle.net/11356/1209.

Eide, Stian Rødven. 2019. "The Swedish PoliGraph: A Semantic Graph for Argument Mining of Swedish Parliamentary Data." In *Proceedings of the 6th Workshop on Argument Mining*, edited by Benno Stein and Henning Wachsmuth, 52–57. Florence: Association for Computational Linguistics. https://www.aclweb.org/anthology/W19-4506; doi:10.18653/v1/W19-4506.

Erjavec, Tomaž, Vladislava Grigorova, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Andrej Pančur, Michał Rudolf, and Kiril Simov. 2020. *Multilingual Comparable Corpora of Parliamentary Debates ParlaMint 1.0.* Slovenian language resource repository CLARIN.SI. CLARIN ERIC. https://hdl.handle.net/11356/1345.

Fišer, Darja, and Jakob Lenardič. 2018. "CLARIN Corpora for Parliamentary Discourse Research." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 2–7. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/14_W2.html.

Fišer, Darja, and Kristina Pahor de Maiti. 2020. "Voices of the Parliament." *Modern Languages Open*, 2020 issue 1: article 46. doi:10.3828/mlo.v0i0.295.

Hansen, Dorte Haltrup. 2018. *The Danish Parliament Corpus 2009–2017, v1*. CLARIN-DK-UCPH Centre Repository. Copenhagen: Centre for Language Technology, NorS, University of Copenhagen; The Danish Parliament. http://hdl.handle.net/20.500.12115/8.

Hawkins, Kevin, Michelle Dalmau, Elli Mylonas, and Syd Bauman, eds. 2018. "Best Practices for TEI in Libraries: A Guide for Mass Digitization, Automated Workflows, and Promotion of Interoperability with XML Using the TEI." Version 4.0.0. Last updated September 10, 2018. http://purl.org/TEI/teiinlibraries; https://tei-c.org/extra/teiinlibraries/4.0.0/bptl-driver.html.

Hirst, Graeme, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. 2014. "Argumentation, Ideology, and Issue Framing in Parliamentary Discourse." In *ArgNLP 2014: Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, edited by Elena Cabrio, Serena Villata, and Adam Wyner. https://dblp.org/rec/conf/argnlp/2014; http://ceur-ws.org/Vol-1341/paper6.pdf.

Hladka, Barbora, Matyáš Kopp, and Pavel Straňák. 2020. "Compiling Czech Parliamentary Stenographic Protocols into a Corpus." In *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, pp. 18–22. Paris: European Language Resources Association. https://www.aclweb.org/anthology/2020.parlaclarin-1.4.

Hughes, Lorna M., Paul S. Ell, Gareth A. G. Knight, and Milena Dobreva. 2015. "Assessing and Measuring Impact of a Digital Collection in the Humanities: An Analysis of the SPHERE (Stormont Parliamentary Hansards: Embedded in Research and Education) Project." *Digital Scholarship in the Humanities* 30 (2): 183–98. doi:10.1093/llc/fqt054.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine: Ten Years On." *Lexicography: Journal of ASIALEX* 1 (1): 7–36. doi:10.1007/s40607-014-0009-9.

Ljubešić, Nikola, Darja Fišer, Tomaž Erjavec, and Filip Dobranić. 2018. "The Parlameter Corpus of Contemporary Slovene Parliamentary Proceedings." In *Proceedings of the Conference on Language Technologies & Digital Humanities*, edited by Darja Fišer and Andrej Pančur, 162–67. Ljubljana, Slovenia: Ljubljana University Press. http://nl.ijs.si/jtdh18/proceedings-en.html; http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Ljubesic-et-al_The-Parlameter-corpus-of-contemporary-Slovene-parliamentary-proceedings.pdf.

Marx, Maarten, and Anne Schuth. 2010. "DutchParl: The Parliamentary Documents in Dutch." In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, 3670–77. Valletta, Malta: European Language Resource Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2010/.

Ogrodniczuk, Maciej. 2018. "Polish Parliamentary Corpus." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 15–19. Paris: European Language Resources Association. http://lrec-conf.org/workshops/lrec2018/W2/summaries/11_W2.html.

Ogrodniczuk, Maciej, and Bartłomiej Nitoń. 2020. "New Developments in the Polish Parliamentary Corpus." *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 1–4. Paris: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.parlaclarin-1.1.

Osenova, Petya, and Kiril Simov. 2012. "The Political Speech Corpus of Bulgarian." *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, et al., 1744–47. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/956_Paper.pdf.

Pančur, Andrej, and Tomaž Erjavec. 2020. "The siParl Corpus of Slovenian Parliamentary Proceedings." In *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 28–34. Paris: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/2020.parlaclarin-1.6.

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

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

Pančur, Andrej, and Mojca Šorn. 2016. "Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History." *Prispevki za novejšo zgodovino / Contributions to Contemporary History* 56 (3): 130–46. https://ojs.inz.si/pnz/article/view/193; doi:10.51663/pnz.56.3.09.

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

Pražák, Aleš, and Luboš Šmídl. 2012. *Czech Parliament Meetings*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLoS ONE* 11 (12): e0168843. https://doi.org/10.1371/journal.pone.0168843.

Steingrímsson, Steinþór, Starkaður Barkarson, and Gunnar Thor Örnólfsson. 2020. "IGC-Parl: Icelandic Corpus of Parliamentary Proceedings." In *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 11–17. Paris: European Language Resources Association. https://www.aclweb.org/anthology/2020.parlaclarin-1.3.

Truan, Naomi. 2016. *Parliamentary Debates on Europe at the Deutscher Bundestag (1998–2015)* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/de-parl.

Wissik, Tanja, and Hannes Pirker. 2018. "ParlAT Beta Corpus of Austrian Parliamentary Records." In *Proceedings of the LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, edited by Darja Fišer, Maria Eskevich, and Franciska de Jong, 20–23. Paris: European Language Resources Association (ELRA). http://lrec-conf.org/workshops/lrec2018/W2/summaries/2_W2.html.

NOTES

1 For an overview, see Fišer and Lenardič (2018).

2 Jan Odijk, "CLARIN ParlaFormat workshop" (blog post), CLARIN, May 29, 2019, https://www.clarin.eu/blog/clarin-parlaformat-workshop.

3 Tom Elliott, Gabriel Bodard, Hugh Cayless, et al., *EpiDoc: Epigraphic Documents in TEI XML*, v. 9.3 (November 2021), https://sourceforge.net/projects/epidoc/.

4 DARIAH-ERIC Working Group *Lexical Resources*, TEI Lex-0 v. 0.9.0 (September 26, 2021), https://github.com/DARIAH-ERIC/lexicalresources.

5 Similarly to Truan (2016).

6 Computer-Mediated Communication Special Interest Group, TEI Consortium, accessed January 13, 2022, https://tei-c.org/activities/sig/cmc/.

7 Tomaž Erjavec and Andrej Pančur, Parla-CLARIN project GitHub site, last updated March 17, 2021, https://github.com/clarin-eric/parla-clarin/.

8 Tomaž Erjavec and Andrej Pančur, Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings, v. 0.2 (December 8, 2020), https://clarin-eric.github.io/parla-clarin/.

9 See also "Parliamentary Corpora," CLARIN Language Resources: Resource Families, accessed January 13, 2022, https://www.clarin.eu/resource-families/parliamentary-corpora.

10 See Tomaž Erjavec and Andrej Pančur, "Response to the Presentations," slide 3, presented at ParlaFormat Workshop, Amersfoort, May 24, 2019, https://www.clarin.eu/sites/default/files/erjavec-pancur-teiparla_2019_response_.pdf.

11 This corpus was compiled as part of the CLARIN Traveling Campus "Talk of Europe," which encoded the proceedings of the European Parliament as linked open data: see http://www.talkofeurope.eu/.

12 Accessed January 13, 2022, http://www.akomantoso.org/; see also the slides by Monica Palmirani, "Akoma Ntoso for Parliamentary Documents," presented at the 2019 Parla-CLARIN workshop: https://www.clarin.eu/sites/default/files/palmirani-akn-clarin-parl2019.pdf.

13 https://tei-c.org/extra/teiinlibraries/4.0.0/bptl-driver.html#frbr.

14 Last updated February 25, 2014, https://www.w3.org/RDF/.

15 Open Linguistics Working Group, "The Linguistic Linked Open Data Cloud Diagram" (draft), accessed January 13, 2022, http://linguistic-lod.org/llod-cloud.

16 Nexus Linguarum website, accessed January 13, 2022, https://nexuslinguarum.eu/; see also "Description," CA18209 – European Network for Web-centred Linguistic Data Science (Brussels: COST: European Cooperation in Science & Technology), accessed January 13, 2022, https://www.cost.eu/actions/CA18209/.

17 See Maciej Ogrodniczuk and Petya Osenova, "ParlaMint: Towards Comparable Parliamentary Corpora," accessed January 13, 2022, https://www.clarin.eu/content/parlamint.

18 Accessed January 13, 2022, https://universaldependencies.org/.

19 The IMS Open Corpus Workbench (CWB), last modified March 30, 2021, http://cwb.sourceforge.net/.

20 Accessed January 13, 2022, http://www.sketchengine.eu/.

AUTHORS

TOMAŽ ERJAVEC

Tomaž Erjavec is a senior researcher at the Department of Knowledge Technologies, Jožef Stefan Institute, and at the Fran Ramovš Institute of the Slovenian Language at the Scientific Research Centre of the Slovenian Academy of Sciences and Arts. His work focuses on developing language resources, especially as regards their annotation and encoding, in the fields of language technologies and of digital humanities. He is the national coordinator of the Slovenian branch of the CLARIN research infrastructure for language resources and tools and a member of ISO/TC 37/SC 4 "Language resource management."

ANDREJ PANČUR

Andrej Pančur is a research fellow at the Institute of Contemporary History, a national coordination institution for the Slovenian branch of DARIAH. Since 2011 he has been working for the ICH's infrastructure program Research Infrastructure of Slovene Historiography, where he is responsible for technological development, research data, digital editions, and bringing digital humanities techniques and methods into the Slovenian research area. In 2018 he became the head of the research infrastructure.