

Compiling parliamentary corpora

Tomaž Erjavec¹ and Andrej Pančur²

¹ Department of Knowledge Technologies, Jožef Stefan Institute

² Institute for Contemporary History

Ljubljana, Slovenia

PARTHENOS Workshop for CEE countries
Sofia, October 7–9, 2019

Introduction

Our background

- Tomaž Erjavec, Jožef Stefan Institute:
 - CLARIN.SI national coordinator
 - Language corpora and other language resources
 - Language technologies (for Slovene & other Slavic languages)
 - Encoding standards: MULTEXT-East, TEI, ISO TC 37 SC4
 - Digital Humanities
- Andrej Pančur, Institute for Contemporary History:
 - DARIAH-SI
 - Historical data: digital catalogues & libraries:
 - TEI and Web technologies
 - Work on parliamentary records in cooperation with the documentation center of the Slovenian National Assembly

Our work on Parliamentary corpora

- Pančur, A., Šorn, M., Erjavec, T. 2018. SloVParl 2.0: The Collection of Slovene Parliamentary Debates from the Period of Secession. *LREC 2018*.
 - Slovenian parliamentary corpus SloVParl 2.0 (1991-1992), <http://hdl.handle.net/11356/1167>, 2017.
- Ljubešič, N., Fišer, D., Erjavec, T. Dobranič, F. 2018. The ParlaMeter corpus of contemporary Slovene parliamentary proceedings. 2018. *Conf. on Language Technologies & Digital Humanities 2018*, Ljubljana, Slovenia.
 - Slovenian parliamentary corpus ParlaMeter-sl 1.0 (2014–2018), <http://hdl.handle.net/11356/1208>, 2019.
 - Croatian parliamentary corpus ParlaMeter-hr 1.0 (2016–2018), <http://hdl.handle.net/11356/1209>, 2019.
- Slovenian parliamentary corpus siParl 1.0 (1990-2018), <http://hdl.handle.net/11356/1236>, 2019. (200 million words)
- Parla-CLARIN format

Motivation

Parliamentary data is (or could be) interesting for a wide range of Social Sciences and Humanities disciplines:

- linguistics
- political sciences
- sociology
- history
- psychology
- ...

Language corpora

How to make parliamentary data available for analysis?

Compile them into *language corpora*:

- A *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.
- A *computer corpus* is a corpus which is encoded in a standardised and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance.

How do you make a corpus of parliamentary debates?

Mostly as any other text corpus:

- Get source data
- Up-convert this particular source it into a common encoding
- Add metadata
- Annotate texts with linguistic annotations
- Maybe add multimedia

And then use it:

- Concordancers (Linguists)
- Excel (Sociologists)
- SparQL (Computer Scientists)

Advantages of corpora of parliamentary debates

Compiling a text/speech corpus is usually also accompanied by many non-technical problems, but parliamentary data isn't:

- no copyright over source texts
- no personal data protection issues (GDPR)
- typically already in digital form
- no problems of acquisition: often available on-line or directly from the government
- metadata readily available

Problems with parliamentary data

- Different countries have different rules for parliamentary proceedings
- The digital sources are in many different formats, and structured quite differently
- Corpora of parliamentary proceedings are often compiled with a limited budget and time
- They are mostly compiled by computational linguists, not aware of the subtle points of the proceedings: the corpus might not be able to answer your research questions

CLARIN

CLARIN work on Parliamentary corpora

CLARIN has organised a number of initiatives and events that deal with parliamentary corpora:

- CLARIN Traveling Campus "Talk of Europe": 3 "Creative Camps" (2014–2015) used the proceedings of the European Parliament, curated as linked open data
- CLARIN-PLUS cross-disciplinary workshop Working with parliamentary records, Sofia 2017
- CLARIN Resource Families Parliamentary corpora, 2018-2019
- ParlaCLARIN workshop at LREC 2018

Parla-CLARIN

Latest event is work on a common encoding format, called Parla-CLARIN:

- Developers: us
- "CLARIN ParlaFormat" workshop (May 23-24, 2019, Amersfoort) with selected participants:
 - The idea of a standard format Parla-CLARIN was introduced
 - Participants presented their own experiences with encoding parliamentary corpora and gave their comments to the draft proposal
 - Response by the developers
 - Slides of the workshop
- Since the workshop, V0.1 of the proposal has been developed:
 - <https://github.com/clarin-eric/parla-clarin>
 - <https://clarin-eric.github.io/parla-clarin>
 - Wiki for technical instructions

Corpora of the Workshop participants

Participant	Title	Format	Description
Ogrodniczuk	Polish Parliamentary Corpus	TEI	stand-off
Banski	Spoken interaction data	TEI	ISO-TEI
Luxardo	TAPS-fr	TEI	XML-TXM
Hansen	Danish Parliament Corpus	TEI	drama module
Wissik	ParlAT	XML	moving to TEI
Marx	PoliticalMashup	XML	TEI inspired
Blätte	GermaParl	XML	TEI inspired
Morkevičius	Lithuanian Parliamentary Data	XML	TEI inspired
Barbaresi	German political speeches	XML	TEI inspired
Osenova	Bulgarian Corpus	XML	TEI for metadata
Eide	Swedish Parliamentary Data	XML	custom
Baranovsky	Knesset Corpus	XML	custom
Hessen	Spreek2Schrijf	XML	VLOS, CXML
Palmirani	Akoma Ntoso	AKN	Standard
Dargis	Corpus of the Saeima	multiple	RDF, CoNLL-U
Molnár	Hungarian Legislative Corpus	CSV	

Making a corpus

Important

Think well before starting work:

- Less noise = more time
- Richer encoding = more time
- Manual editing = much more time

The toolchest:

- Python: general purpose programming; fast; regular expressions
- XSLT: transforming XML to XML or to other formats, e.g. HTML
- Oxygen: XML editor
- Word (with automatic DOCX to XML conversion)

Sources

How to get source data:

- Focused web crawl
- Directly from the parliamentary office!

Source data formats:

- PDF :(
extract plain text; quality dependent on xxx2pdf engine; can be much noise in texts (character sets, ligatures, spacing, doubled chars, columns)
- HTML :|
can be ill-formed; visual rather than semantic tags
- XML :)
can be validated and converted with XSLT directly to (basic) target encoding

Cleaning source data

Characters:

- Character sets ok? Can be fixed?
- Bad chars: soft hyphen, nobreak-space
- Joining hyphenated words

Structures:

- Removing unprocesssable or uninteresting structures, e.g. tables
- Marking-up that data has been removed (<gap/>)

Documenting the encoding process

E.g. in the <teiHeader> of a TEI document:

```
<editorialDecl>
  <correction>
    <p>Found typos in the source have been silently
      corrected.</p>
  </correction>
  <normalisation>
    <p>Tables have been omitted from the corpus. Spacing
      has been normalised to single space. Soft
      hyphens have been removed.</p>
  </normalisation>
  <hyphenation>
    <p>End-of-line hyphens have been removed.</p>
  </hyphenation>
  <quotation>
    <p>Quotation marks have been left in the text and
      are not explicitly marked up.</p>
  </quotation>
</editorialDecl>
```

Extracting structure

- In the source texts important information is often given in typography, e.g.

Boris Johnson: I propose a no-deal Brexit. /Jeremy Corbyn: Traitor!/ Because England does not want any dealings with the European Union.

- Regular expression matching can be used to identify and explicitly mark-up such structures:

```
<u who="#BorisJohnson">I propose a no-deal  
  Brexit.</u>
```

```
<u who="#JeremyCorbyn">Traitor!</u>
```

```
<u who="#BorisJohnson">Because England does not  
  want any dealings with the European Union.</u>
```

Metadata

- Useful to know more about a speaker than just their name
- This information can often be found in external resources, e.g. other parliamentary data, Wikipedia
- Sex, date of birth, education, party affiliation, role in parliament, etc.
- Party information: name, lifetime, coalition, etc.
- Note that much of this information is time-dependent

Linguistic annotation

- Almost invariably automatic: annotation tools for your language; mistakes (precision, recall)
- Basic annotation levels: tokenisation & sentence segmentation, part-of-speech (morphosyntactic) tagging, lemmatisation
- Further annotations: named entities, linking to external resources, sentiment, translation

Multimedia

- For some parliaments, audio (video) is also available
- Alignment of speech signal with transcription
- Note that the officially published records are often redacted, i.e. quite different from the actual speech
- For older parliamentary debates: facsimile, i.e. images of pages → manuscript studies

Encoding formats

Purpose

How to encode the data and metadata?

- Database
- Resource Description Framework (RDF):
Semantic Web / Linked Open Data
- XML: the most common encoding format
- XML is a meta-annotation language:
we need an XML schema to defined what elements and
attributes are allowed and how they nest:
 - Self-developed XML schema:
can express exactly what you want but not interchangeable
 - A national “standard”:
used by the national NLP community, but not by others
 - Akoma Ntoso:
OASIS standard, not oriented towards corpora
 - Text Encoding Initiative:
very broad, used by many, not perscriptive

Corpora of the Parla-Format workshop participants

Participant	Title	Format	Description
Ogrodniczuk	Polish Parliamentary Corpus	TEI	stand-off
Banski	Spoken interaction data	TEI	ISO-TEI
Luxardo	TAPS-fr	TEI	XML-TXM
Hansen	Danish Parliament Corpus	TEI	drama module
Wissik	ParlAT	XML	moving to TEI
Marx	PoliticalMashup	XML	TEI inspired
Blätte	GermaParl	XML	TEI inspired
Morkevičius	Lithuanian Parliamentary Data	XML	TEI inspired
Barbaresi	German political speeches	XML	TEI inspired
Osenova	Bulgarian Corpus	XML	TEI for metadata
Eide	Swedish Parliamentary Data	XML	custom
Baranovsky	Knesset Corpus	XML	custom
Hessen	Spreek2Schrijf	XML	VLOS, CXML
Palmirani	Akoma Ntoso	AKN	Standard
Dargis	Corpus of the Saeima	multiple	RDF, CoNLL-U
Molnár	Hungarian Legislative Corpus	CSV	

Parla-CLARIN

- <https://github.com/clarin-eric/parla-clarin>
- <https://clarin-eric.github.io/parla-clarin/>
- Git: Version control, collaborative development, "Social media" support (issues), commit validation, support for (derived) viewable static HTML pages
- The proposed format is centered on storing and interchanging linguistically annotated corpora of parliamentary data of any country and language to be used in scholarly research
- Allows for different types and depths of annotation
- A parameterisation of the Text Encoding Initiative Guidelines

What needs to be taken into account

- Structure: legislative periods, sessions, topics, speeches
- General metadata: titles, parliamentary body, location, date and time
- Speaker metadata: age, party membership (time dependent!), links
- Metadata on political parties: name, alternative name, abbreviation, history
- Speeches: speaker, text, verbal and non-verbal interruptions
- Text versions: verbatim or redacted records
- Linguistic annotation: PoS tagging, named entities, syntax
- Multimedia: audio and video, alignment with transcription
- Legislative aspects: specification of laws, roll-calls

Text Encoding Initiative



< Text Encoding Initiative >

- Aim: enabling annotation of digital documents or any type and in any language for the purposes of scholarly analysis
- The TEI Guidelines define and name several hundred useful textual distinctions
- The TEI provides a framework for the definition of multiple schemas
- Probably the oldest still active standardisation effort for text
- TEI Consortium, tei-l mailing list
- Converters to and from TEI: Word, HTML, etc.

The ODD TEI schema

- "One Document Does it all"
- TEI schema, which is itself a TEI document
- A TEI ODD includes TEI modules (obligatory and optional) & possible constraints and modifications
- = formal specification of the schema, then converted to XML schema (W3C, DTD, RelaxNG, Schematron) with TEI XSLT stylesheets
- ODD also includes the documentation of the schema, i.e. the Guidelines

Parla-CLARIN modules

- Obligatory modules: TEI core, TEI header, TEI structure
- Basic text type: TEI transcriptions of speech
cf. ISO 24624:2016 Language resource management –
Transcription of spoken language
- Overall structure and extended TEI header: TEI corpus
- Details of speakers: TEI person
- Complex references: TEI linking
- Simple linguistic analysis: TEI analysis
- Complex (linguistic) analysis: TEI feature structures

= a rather general schema

+ **documentation**

Parla-CLARIN documentation

Parla-CLARIN
A TEI Schema for Corpora of Parliamentary Proceedings

Table of contents

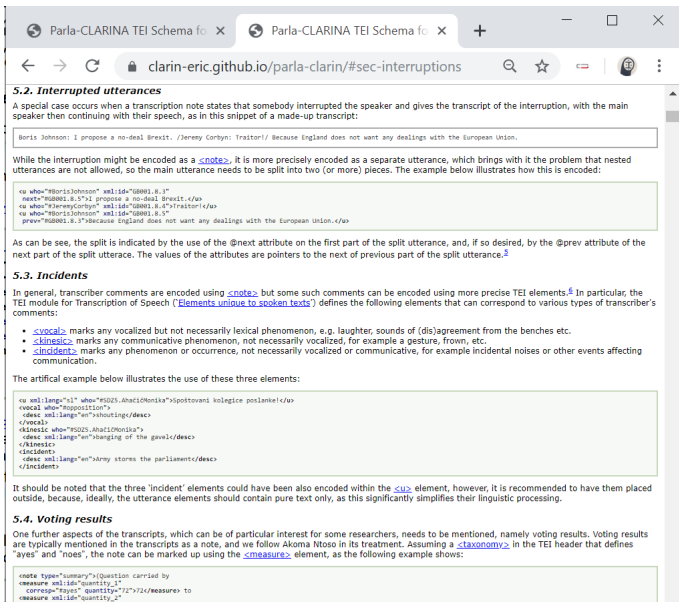
- [1. Introduction](#)
 - [1.1. Scope and purpose](#)
 - [1.2. Background](#)
- [2. General requirements](#)
 - [2.1. Characters](#)
 - [2.2. Documenting the encoding process](#)
 - [2.3. Languages](#)
 - [2.4. Identifiers and referencing](#)
 - [2.5. Temporal information](#)
 - [2.6. Files](#)
- [3. Overall document structure](#)
 - [3.1. Corpus structure](#)
 - [3.2. Text divisions](#)
 - [3.3. Document variants](#)
- [4. Corpus metadata](#)
 - [4.1. Speaker metadata](#)
 - [4.2. Party metadata](#)
 - [4.3. Relationships between people and parties](#)
- [5. Transcriptions](#)
 - [5.1. Utterances and commentary](#)
 - [5.2. Interrupted utterances](#)
 - [5.3. Incidents](#)
 - [5.4. Voting results](#)
- [6. Linguistic annotation](#)
 - [6.1. Word-level annotation](#)
 - [6.2. Segmental annotation](#)
 - [6.3. Linguo annotation](#)
- [7. Multimedia](#)
 - [7.1. Speech and video](#)
 - [7.2. Facsimile](#)
- [8. Conversions](#)
 - [8.1. Conversion from Akoma Ntoso](#)
 - [8.2. Conversion to RDF](#)
- [9. Acknowledgements](#)

Appendix A [Formal specification](#)

- [Appendix A.1 Elements](#)
- [Appendix A.2 Model classes](#)
- [Appendix A.3 Attribute classes](#)
- [Appendix A.4 Macros](#)
- [Appendix A.5 Datatypes](#)
- [Appendix A.6 Constraints](#)

1. Introduction

Explanations & examples



Parla-CLARINA TEI Schema fo x Parla-CLARINA TEI Schema fo x + - □ X

clarin-eric.github.io/parla-clarin/#sec-interruptions

5.2. Interrupted utterances

A special case occurs when a transcription note states that somebody interrupted the speaker and gives the transcript of the interruption, with the main speaker then continuing with their speech, as in this snippet of a made-up transcript:

```
Boris Johnson: I propose a no-deal Brexit. /Jeremy Corbyn: Traitor!/ Because England does not want any dealings with the European Union.
```

While the interruption might be encoded as a `<note>`, it is more precisely encoded as a separate utterance, which brings with it the problem that nested utterances are not allowed, so the main utterance needs to be split into two (or more) pieces. The example below illustrates how this is encoded:

```
<u who="#BorisJohnson" xml:id="GB001.8.3"
  next="#GB001.8.5">I propose a no-deal Brexit.</u>
<u who="#JeremyCorbyn" xml:id="GB001.8.4">Traitor!</u>
<u who="#BorisJohnson" xml:id="GB001.8.5"
  prev="#GB001.8.3">Because England does not want any dealings with the European Union.</u>
```

As can be seen, the split is indicated by the use of the `@next` attribute on the first part of the split utterance, and, if so desired, by the `@prev` attribute of the next part of the split utterance. The values of the attributes are pointers to the next of previous part of the split utterance.³

5.3. Incidents

In general, transcriber comments are encoded using `<note>` but some such comments can be encoded using more precise TEI elements.⁴ In particular, the TEI module for Transcription of Speech (`'Elements unique to spoken texts'`) defines the following elements that can correspond to various types of transcriber's comments:

- `<vocal>` marks any vocalized but not necessarily lexical phenomenon, e.g. laughter, sounds of (dis)agreement from the benches etc.
- `<kinetic>` marks any communicative phenomenon, not necessarily vocalized, for example a gesture, frown, etc.
- `<incident>` marks any phenomenon or occurrence, not necessarily vocalized or communicative, for example incidental noises or other events affecting communication.

The artificial example below illustrates the use of these three elements:

```
<u xml:lang="sl" who="#S02S.AhaCljMonika">Sploštovani kolegice poslanelc/u>
<vocal who="#Opposition">
<desc xml:lang="en">shouting</desc>
</vocal>
<kinetic who="#S02S.AhaCljMonika">
<desc xml:lang="en"> banging of the gavel</desc>
</kinetic>
<incident>
<desc xml:lang="en">Army storms the parliament</desc>
</incident>
```

It should be noted that the three 'incident' elements could have been also encoded within the `<u>` element, however, it is recommended to have them placed outside, because, ideally, the utterance elements should contain pure text only, as this significantly simplifies their linguistic processing.

5.4. Voting results

One further aspects of the transcripts, which can be of particular interest for some researchers, needs to be mentioned, namely voting results. Voting results are typically mentioned in the transcripts as a note, and we follow Akoma Ntoso in its treatment. Assuming a `<taxonomy>` in the TEI header that defines "ayes" and "noes", the note can be marked up using the `<measure>` element, as the following example shows:

```
<note type="summary">(Question carried by
<measure xml:id="quantity_3"
  corresp="#ayes" quantity="72">72</measure> to
<measure xml:id="quantity_2"
  corresp="#noes" quantity="56">56</measure>.
```


General structure of a Parla-CLARIN document

```
<teiCorpus xmlns="http://www.tei-c.org/ns/1.0"
           xml:lang="xx">
  <teiHeader>
    <!-- metadata for the entire corpus -->
  </teiHeader>
  <TEI>
    <teiHeader>
      <!-- metadata for one session or one sitting -->
    </teiHeader>
    <text>
      <!-- content of the parliamentary debate -->
    </text>
  </TEI>
  <!-- more TEI elements here -->
</teiCorpus>
```

Encoding person data

```
<person xml:id="HalbJanko1957">
  <persName>
    <surname>Halb</surname>
    <forename>Janko</forename>
  </persName>
  <sex value="M"/>
  <birth>
    <date when="1957-07-13">13. 7. 1957</date>
    <placeName ref="https://www.geonames.org/3193481">Pertotča</placeName>
  </birth>
  <education>economist</education>
  <trait type="ethnicity">
    <desc>Slovenian</desc>
  </trait>
  <affiliation ref="#parliament" role="#grp.member"
    from="1990-05-08" to="1992-12-23"/>
  <affiliation ref="#SKZ" role="#grp.member"
    notBefore="1988-05-12" notAfter="1992-06-27"/>
  <affiliation ref="#SLS" role="#grp.member"
    notBefore="1992-06-27" notAfter="2000-04-15"/>
  <idno type="URI">https://sl.wikipedia.org/wiki/Janko_Halb</idno>
</person>
```

- Time and place of birth, gender, party membership and role, official functions, constituency, education, biography, external links (e.g. Wikipedia), etc.

Encoding organization data

```
<listOrg>
  <org xml:id="SKZ">
    <orgName full="yes" from="1988-05-12" to="1990-12-18">Slovenska kmečka zveza
      </orgName>
    <orgName full="init" from="1988-05-12" to="1990-12-18">SKZ</orgName>
    <orgName full="yes" from="1990-12-18" to="1992-06-27">Slovenska kmečka zveza
      - Ljudska stranka</orgName>
    <orgName full="init" from="1990-12-18" to="1992-06-27">SKZ-LS</orgName>
  </org>
  <org xml:id="SLS">
    <orgName full="yes" from="1992-06-27" to="2000-04-15">Slovenska Ljudska
      stranka</orgName>
    <orgName full="init" from="1992-06-27" to="2000-04-15">SLS</orgName>
  </org>
  <listRelation>
    <relation name="successor" active="#SLS" passive="#SKZ" when="1992-06-27"/>
    <relation name="coalition" mutual="#pp.SDZ_#pp.SDSS_#pp.SKD_#pp.SKZ_#pp.SOS_
      #pp.ZS" from="1990-05-16" to="1992-05-14"/>
  </listRelation>
</listOrg>
```

Encoding speeches

```
<note type="speaker">PRESEDNIK JOŽE ZUPANČIČ:</note>
<u who="#ZupancicJoze1936" decls="#chair">
  <seg>In kako boš ti glasoval?</seg>
</u>
<note type="speaker">Jaklič:</note>
<u who="#JakicRoman1967" decls="#unauthorized">
  <seg>Glasoval bom seveda za.</seg>
</u>
<u who="#ZupancicJoze1936" decls="#chair">
  <seg>Gospod Andrej Verlič.</seg>
</u>
<incident>
  <desc>Aplavz.</desc>
</incident>
<note type="speaker">ANDREJ VERLIČ:</note>
<u who="#VerlicAndrej" decls="#regular">
  <seg>Spoštovane poslanke, spoštovani poslanci!</seg>
</u>
```

Linguistic Annotation

- Can be simple, or extremely complex
- TEI allows various methods of linguistic mark-up
- Almost as many ways as are practitioners
- Parla-CLARIN gives a recommendation for various types of linguistic annotation, e.g.:

```
<s>
  <w ana="#Pd-nsg" lemma="ta">Tega</w>
  <w ana="#Px-----y" lemma="se">se</w>
  <w ana="#Q" lemma="sploh">sploh</w>
  <w ana="#Va-ris-y" lemma="biti">nisem</w>
  <w ana="#Vmep-sm" lemma="zavesti"
    join="right">zavedel</w>
  <pc ana="#Z">.</pc>
</s>
```

Using the corpus

- Spreadsheets, R, Graph visualisations, etc.
- Concordancers:
 - Sketch Engine:
commercial, but free for use until 2022 for ELEXIS countries
 - noSketch Engine:
freely available part of Sketch Engine software
 - CLARIN.SI:
 - noSketch Engine
 - Kontext
 - more than 50 corpora

Conclusions

Conclusions

- Gave a general overview of how to compile a corpus of parliamentary debates
- Need (somebody with) knowledge of programming
- Should carefully study the source data and think about the intended use
- Then decide which aspects of the data to encode: more detail, less noise: more effort
- First develop a small pilot corpus to estimate the cost
- Document the corpus compilation process
- Deposit the corpus in one of the CLARIN repositories

Tutorial

- Will be lead by Andrej Pančur
- You should have received an email with instructions for the software you need
- Task: to encode a small example in Parla-CLARIN, to get a feeling for the data
- For those that already have experience: can talk with me

Compiling parliamentary corpora

Tomaž Erjavec¹ and Andrej Pančur²

¹ Department of Knowledge Technologies, Jožef Stefan Institute

² Institute for Contemporary History

Ljubljana, Slovenia

PARTHENOS Workshop for CEE countries
Sofia, October 7–9, 2019