Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings

Tomaž Erjavec[1] and Andrej Pančur[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute

[2] Institute for Contemporary History

Ljubljana, Slovenia

TEI 2019
Gradec, 2019-09-19

Overview of the talk

Introduction

Characteristics of parliamentary data

- Parliamentary data is interesting for a wide range of disciplines, e.g. political sciences, sociology, history, ...
- Typically no copyright or personal data protection issues
- Often available on-line, also in XML
- Rich metadata, audio, video, multilingual
- Many researchers have already produced corpora/datasets of parliamentary data, e.g. EuroParl, EPIC, Talk of Europe; Croatian, Czech, Danish, Dutch, French, German, Greek, Latvian, Norwegian, Polish, Romanian, Slovenian, UK, etc.

CLARIN background to the proposal

European Research Infrastructure for Language Resources and Technology: CLARIN ERIC + 20 national nodes:

- 2014 + 2015: CLARIN Traveling Campus 'Talk of Europe' (3 "Creative Camps" which used the proceedings of the European Parliament)
- 2017 workshop "Working with parliamentary records", Sofia
- 2018 workshop "ParlaCLARIN", LREC
- 2018–2019 reports "CLARIN Resource Families: Parliamentary corpora"

The covered corpora encoded in a variety of different annotation schemes, limiting their interchange and re-use.

CLARIN ParlaFormat Workshop

We wished to develop an encoding recommendation that could cover all (most) needs of researchers interested in parliamentary proceedings corpora (PPC):

- CLARIN "Type B" workshop, May 23-24, 2019, Amersfoort
- Proposed by the Jan Odijk & the CLARIN Interoperability Committee
- Developers: Tomaž Erjavec and Andrej Pančur
- Schedule:
  - Introduction to the recommendation
  - 15 talks by the participants presenting their experiences with PPCs with comments on the recommendation
  - Responses to the comments
- Blog post by Jan Odijk with link to the slides:
  `https://www.clarin.eu/blog/clarin-parlaformat-workshop`

Encoding schemas by 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	OASIS standard
Dargis	Corpus of the Saeima	RDF	+ CoNLL-U
Molnár	Hungarian Legislative Corpus	DB	CSV

Principles

Workshop take home messages

- 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
- and by computational linguists, not aware of the subtle points of the proceeding

The recommendations should allow very different types and depths of annotation: TEI.

What needs to be taken into account

- Structure: legislative periods, sessions, topics, speeches
- Metadata: titles, parliamentary body, location, date and time
- Speakers: age etc., links, party membership (time dependent!)
- Political parties: name, alternative name, abbreviation, history
- Speeches: speaker, role, transcription, interruptions
- Text versions: verbatim / redacted records
- Linguistic annotation: word-level, segmental, linking
- Multimedia: audio and video, facsimile, alignment
- Legislative aspects: voting results, laws

Interchange or interoperability?

- Interchange: documents can be accessed and understood by other humans
- Interoperability: documents can be accessed and understood by other programs

We aim for **interchange**, as interoperability would mean:

- aiming for the least common denominator: loss of information
- much more complex conversion process
- CLARIN is oriented towards humanities scholars, not computer scientists: room to experiment

Descriptive vs. prescriptive

- Two options:
  - bare: allow only elements/attributes that we know we need
  - all: allow all elements/attributes that we might need
- In the first stage leaning towards "All"
  - we don't understand the whole variety of the parliamentary debates in all languages / countries
  - tighten up the proposal as we go along (acquire examples & use cases)

Parla-CLARIN

Parla-CLARIN recommendations

Scope

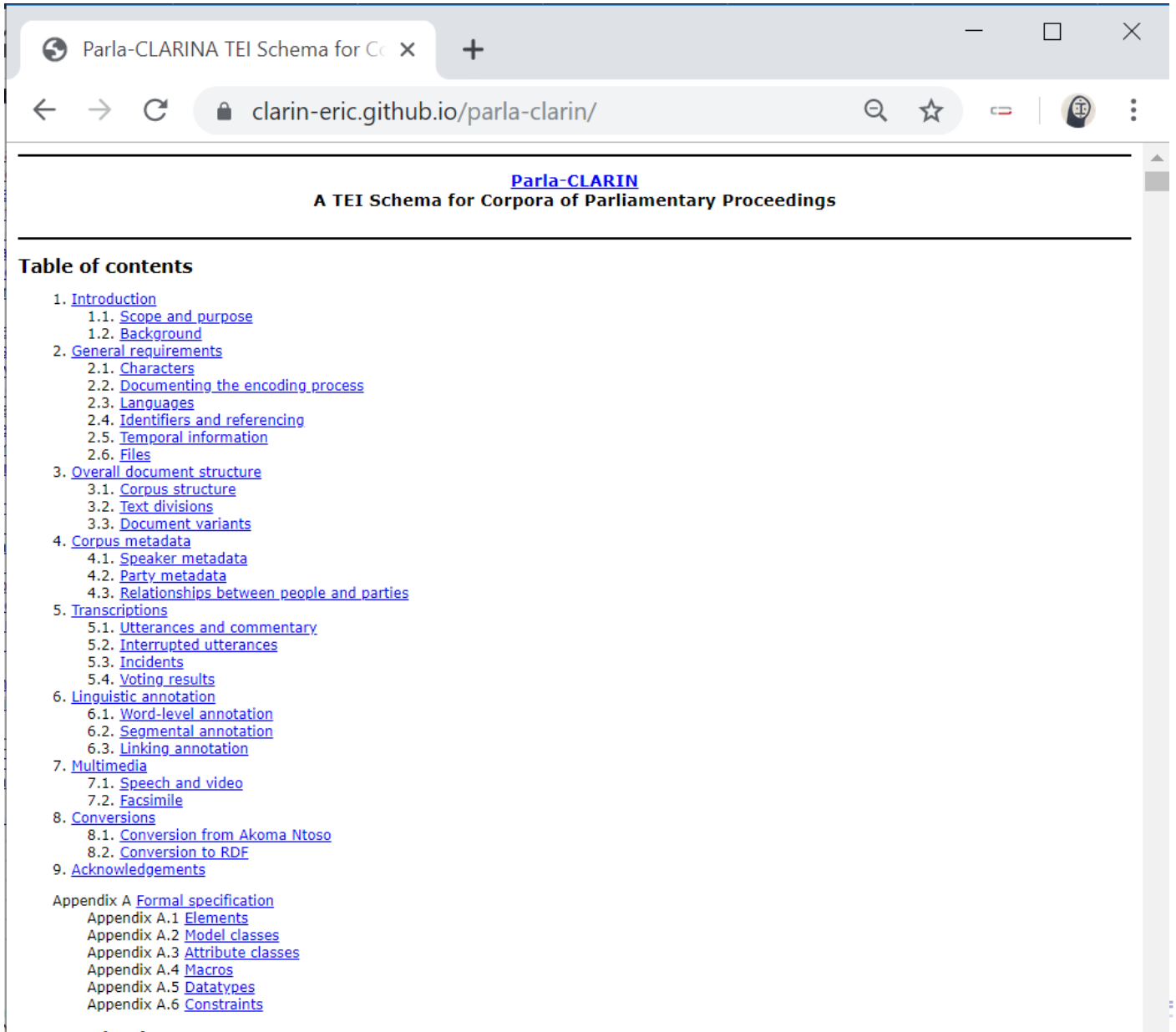
For encoding of PPCs, regardless of the language or country of origin, for the purposes of scholarly investigations, be they from the field of linguistics, political science, history or other humanities and social sciences disciplines.

Parla-CLARIN ODD modules

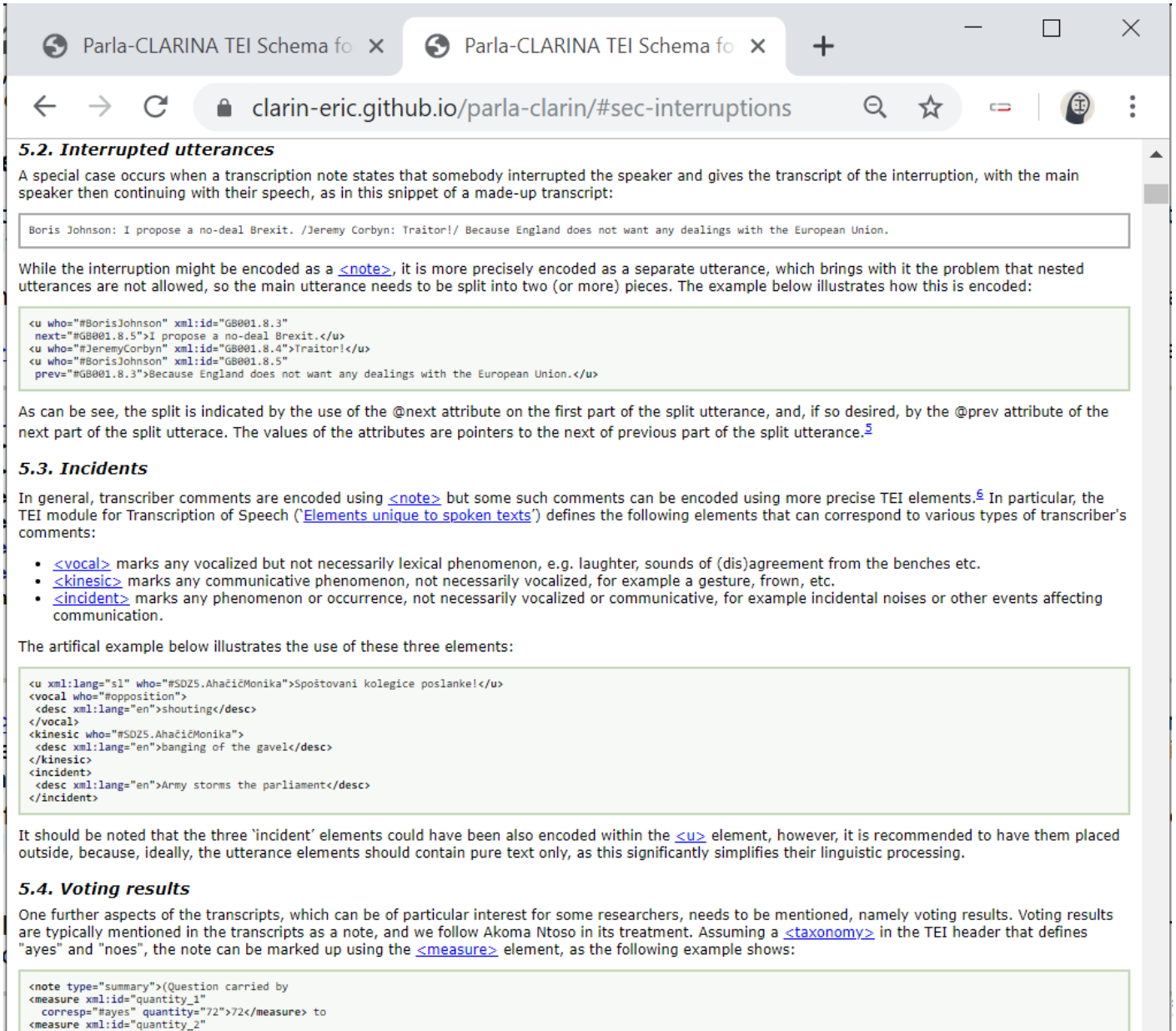
- Basic text type: TEI transcriptions of speech
- Overall structure and extended teiHeader: TEI corpus
- Details of speakers: TEI person
- Complex references: TEI linking
- Simple linguistic analysis: TEI analysis
- Complex (linguistic) analysis: TEI feature structures

+ Documentation

Parla-CLARIN documentation



Explanations & examples



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 see, 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 utterace. The values of the attributes are pointers to the next of previous part of the split utterance.[5]

5.3. Incidents

In general, transcriber comments are encoded using <note> but some such comments can be encoded using more precise TEI elements.[6] 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.
- <kinesic> 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 artifical example below illustrates the use of these three elements:

```
<u xml:lang="sl" who="#SDZ5.AhačičMonika">Spoštovani kolegice poslanke!</u>
<vocal who="#opposition">
 <desc xml:lang="en">shouting</desc>
</vocal>
<kinesic who="#SDZ5.AhačičMonika">
 <desc xml:lang="en">banging of the gavel</desc>
</kinesic>
<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_1"
  corresp="#ayes" quantity="72">72</measure> to
<measure xml:id="quantity_2"
```

Git

- Inspirations: TEI in libraries, TEI Lex0, ElTeC
- Parla-CLARIN can be found on GitHub
- Version control
- Collaborative development
- Support for (derived) static HTML pages on github.io
- Usage instructions on the project Wiki

Alternate solutions

Akoma Ntoso

- XML Schema explicitly developed to model legislative documents
- OASIS standard
- Already adopted by e.g. European Parliament, European Commission, EU Publication Office, UK & Scottish Parliament, Italian Senate, Parliament of Uruguay, several UN agencies
- Referencing and metadata compliance levels
- Use of FRBR (Functional Requirements for Bibliographical Records) for document metadata
- Extension mechanisms not very well specified (probably use different namespace)

Akoma Ntoso vs. TEI

Akoma Ntoso:

- Focus more on logical and legal structure and content of the documents than on the scholarly (linguistic) investigations of the transcriptions
- Lack of elements to define detailed speaker metadata (relegated to external data sources, e.g. FOAF)
- Lack of elements for linguistic annotation (and somewhat difficult to add elements from another namespace)
- Unfamiliarity of corpus compilers with AKN

Akoma Ntoso and TEI

- We don't see the two in competition, rather as synergy:
  - AKN can be the official format of the parliaments
  - TEI as the corpus storage format for scholarly analyses
  - developed (partial) AKN2TEI XSLT, tries to preserve all the information
  - TEI2AKN would be also nice
- AKN served as the model for to inspire the TEI recommendation:
  - FRBR
  - sample values of div/@type
  - voting results
- However, many solutions don't seem optimal (at least for Slovene PPC)

RDF / LOD

- Meant for use by machines, not humans:
  could be problematic in a HSS context
- Parliamentary debates have already been encoded in RDF:
  European Parliament LOD schema
- However, relatively shallow encoding
- Some support for linguistic annotation of texts:
  Linguistic Linked Open Data
- TEI (theoretically) allows linking with RDF
  (with RDFa, cf. issue 1860 of the TEI GitHub project)

Conclusions

Conclusions

- We consider TEI (Speech) as a good basis for an XML schema for corpora of parliamentary debates
- Currently a very general ODD
  but reasonably complete(?) documentation and examples
- The Parla-CLARIN recommendations can be found on:
  `https://github.com/clarin-eric/parla-clarin`
  `https://clarin-eric.github.io/parla-clarin`

Further work

- GitHub collaborative development of those willing: issues, pull requests, commits
- The project should include samples of existing (validating) corpora and conversion from source to Parla-CLARIN
- First corpora:
  - siParl 1.0 (1990-2018), `http://hdl.handle.net/11356/1236`.
  - SlovParl 2.0 (1990-1992), `http://hdl.handle.net/11356/1167`.
  - & something in English
- Examples and descriptions into `p5subset.xml`?
- Recommended taxonomies
- Recommended (required?) attribute values (`@type`)

Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings

Tomaž Erjavec[1] and Andrej Pančur[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute

[2] Institute for Contemporary History

Ljubljana, Slovenia

TEI 2019
Gradec, 2019-09-19