Response to the presentations

Tomaž Erjavec\textsuperscript{1} and Andrej Pančur\textsuperscript{2}

\textsuperscript{1} Department of Knowledge Technologies, Jožef Stefan Institute
\textsuperscript{2} Institute for Contemporary History

Ljubljana, Slovenia

ParlaFormat Workshop
Amersfoort, 2019-05-24
Most seem to be in favour of pursuing the TEI proposal

Very useful feedback, some responded to below
## Participants’ formats

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

- Very important to define what the target application of the proposed annotation scheme is.
- Draft definition: an interchange format for encoding corpora of parliamentary speeches for the purpose of linguistic investigations.
- Would concentrate on the edited and non-edited transcripts, not on interpretation (at least in the data).
- Should allow for very simple (but also complex) encodings.
- Should be useful also for other, similar types of texts, e.g. non-parliamentary political speeches.
- Suggestion to use `<meeting>` rather than speech elements
- Because PDs are not really speech
- True, but they are very similar in structure to speech corpora
- We feel that the Speech module is still the best fit
Some details of elements used

- **Utterances** (<u>): a speech, possibly interrupted
- **Notes** (<note>): comments or parts of the transcript that have not been spoken (speaker, voting)
- **Some others possible elements**: <incident>, <kinesic>, <vocal>, <writing>
- **Paragraphs** (<p>): Running non-spoken text, possibly useful for transcripts that have not been split into utterances
- **Taxonomies** (<taxonomy>): teiHeader taxonomies defining categorisations of divisions, utterances, speaker roles etc.
In addition to the schema and examples it is important to include/develop down-conversion to other formats
Esp. formats that are needed by various analysis tools
Some feel that including person data does not make much sense, better to refer to it (e.g. Wikimedia, DBpedia, Schema.org, FOAF, GeoNames)

For reasons of completeness, uniform processing, experiment and reproducibility we think it is better to include in the corpus as much data as possible / is made use of

However, this is not a requirement!

Minimal data: name of speaker
Mistakes

- The proposal does not exactly follow ISO TEI Speech & latest TEI Analysis additions
- True (for the most part): our mistake / hurry to finish siParl, will be corrected
- Except for listPerson, listOrg, listNym, listEvent: when we want to use listRelation/relation, we need them.
- Except for @msd: it should be a pointer
Synchronisation with speech/video stream

- TEI does support it
- It is rather prolix, but that is the way it is.

```xml
<timeline unit="s"
    origin="#GosVL01_pravo.t0"
    corresp="#GosVL01_pravo.wav">
<when xml:id="GosVL01_pravo.t0"/>
<when xml:id="GosVL01_pravo.t1"
    interval="4.905"
    since="#GosVL01_pravo.t0"/>
<when xml:id="GosVL01_pravo.t2"
    interval="7.852"
    since="#GosVL01_pravo.t0"/>
</timeline>
```
For old PD it would be useful to include the facsimile of the original

TEI has the <facsimile> element that allows this

However, this element is defined in the module for Representation of Primary Sources; so far, we did not plan to include this module

This raises the problem of how to define the schema: with transcr module or without, or have several schemas / ODDs?

With 2 ODDs, problem of redundancy

Maybe use ODD chaining?

or just include the transcription module...
Akoma Ntoso

- Different focus from the CLARIN-Parla proposal
- Seems unrealistic to expect corpus compilers to use a completely different schema for one type of text
- However, Akoma a very good model for developing the CLARIN-Parla proposal
- AKN2TEI one of the priorities + TEI2AKN
- Downstream usefulness of TEI linguistic annotation for AKN
Further work (for you)

- Send samples of your PD to us (URL if openly available)
- Register on https://github.com/clarin-eric/parla-clarin to be notified of changes
- Post issues on aspects we should take into account (esp. once we start writing the Guidelines)
- Once we have more or less finished, pull requests
Response to the presentations

Tomaž Erjavec\textsuperscript{1} and Andrej Pančur\textsuperscript{2}

\textsuperscript{1} Department of Knowledge Technologies, Jožef Stefan Institute
\textsuperscript{2} Institute for Contemporary History

Ljubljana, Slovenia

ParlaFormat Workshop
Amersfoort, 2019-05-24