Introduction to the proposed annotation scheme

Tomaž Erjavec[1] and Andrej Pančur[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute

[2] Institute for Contemporary History

Ljubljana, Slovenia

ParlaFormat Workshop
Amersfoort, 2019-05-23

Introduction

Motivation

- 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
- Rich metadata, audio, video, multilingual
- Therefore many researchers have produced corpora of parliamentary data, e.g. EuroParl, EPIC, Talk of Europe; UK, Dutch, Czech, Norwegian, German, Polish, Greek, Danish, French, Latvian, Romanian, Slovenian, Croatian etc.
- However, they are encoded in a variety of different annotation schemes, limiting their interchange and re-use

Participants: Different formats

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	

Purpose

- The proposed format is centered on storing and interchanging linguistically annotated corpora of parliamentary data to be used in scholarly research
- Facts of life:
  - 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 compiled by computational linguists, not aware of the subtle points of the proceedings
- Therefore the proposed format must allow for different types and depths of annotation

What needs to be taken into account

- Structure: legislative periods, sessions, topics, speeches
- Metadata: titles, parliamentary body, location, date and time
- Speakers: age, party membership (time dependent!), links
- 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, normalisation, syntax
- Multimedia: audio and video, alignment with transcription
- Legislative aspects: specification of laws, roll-calls

CLARIN background to the proposal

- European Research Infrastructure for Language Resources and Technology: CLARIN ERIC + 20 national nodes
- CLARIN-PLUS cross-disciplinary workshop "Working with parliamentary records", Sofia 2017
- CLARIN Resource Families: Parliamentary corpora, 2018–2019
- ParlaCLARIN workshop at LREC 2018
- CLARIN Annual Conference, Pisa 2018: Work on proposal for COST Action Parlant, "Parliamentary Data and Language Technology" (submitted, proposers Maciej Ogrodniczuk, Petya Osenova)

Developers' background

- Tomaž Erjavec, JSI:
  - CLARIN.SI national coordinator
  - Language corpora
  - Language technologies (for Slovene & other Slavic languages)
  - Encoding standards: MULTEXT-East, TEI, ISO TC 37 SC4
- Andrej Pančur, INZ:
  - DARIAH-SI
  - Digital catalogues & libraries: historical data
  - 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.
- Ljubešić, N., Fišer, D., Erjavec, T. Dobranić, 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)

TEI

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
- ODDs can be chained

Current 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 teiHeader: TEI corpus
- Details of speakers: TEI person
- Complex references: TEI linking
- Simple linguistic analysis: TEI analysis
- Complex (linguistic) analysis: TEI feature structures

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 parliamentary debate -->
    </text>
  </TEI>
  <!-- more TEI elements here: other sessions,
       morphosyntactic specifications etc. -->
</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">Pertoča</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">PREDSEDNIK 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
- In-place vs. stand-off
- Almost as many ways as are practitioners: Slovenia, Poland cf. contribution by Piotr Banski

Developing the proposal

Related project: eLexis

- H2020 Infrastructure project, Lead: JSI (Simon Krek)
- Harmonising & Integrating Dictionaries: TEI and RDF
- DARIAH: "TEI Lex0" (Laurent Romary, Toma Tasovac)
- `https://github.com/DARIAH-ERIC/lexicalresources`
- `https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html`
- `https://gitlab.clarin.si/et/tei-lex0-sl`
- Aiming for TEI Lex0 to become an OASIS standard

Related project: ELTeC

- COST Action "Distant Reading for European Literary History" (Christof Schöch, University of Trier)
- Creating a diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written. The Action will create a shared theoretical and practical framework to enable innovative, sophisticated, data-driven, computational methods of literary text analysis across at least 10 European languages.
- `https://www.distant-reading.net/`
- ElTeC corpus: 100 novels (1850 – 1920) per language chosen to explicit criteria
- Common TEI schema: Level0, Level1, Level2

Parla-CLARIN platform: GitHub

- Version control
- Collaborative development
- "Social media" support (issues)
- Commit validation
- Support for (derived) viewable static HTML pages

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: lack of time
  - CLARIN is oriented towards humanities scholars, not computer scientists: room to experiment
- We therefore aim to produce rather general Guidelines with examples of best/current practice
- Tighten up the proposal as we go along

Approach

- Two options:
  - bare: allow only elements that we know we need
  - all: allow all elements 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)

Alternate schemas

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
- Allows inclusion of other schemas via namespaces

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

TEI:

- Familiarity of the corpus compilers with TEI
- Can use (generally) the same TEI schema for all corpora
- Availability of ready-made TEI aware tools for processing

Akoma Ntoso and TEI

- We don't see the two in competition, rather as synergy:
  - AKN as official format of the parliaments
  - TEI as the corpus storage format for scholarly analyses
  - develop (initial / partial) XSLT for AKN2TEI and TEI2AKN
- AKN will serve to inspire the TEI schema (e.g. div/@type)
- Maybe develop an AKN ODD, cf.
  Laurent Romary, Charles Riondet. EAD-ODD: A solution for project-specific EAD schemes. ArchivalScience, Springer Verlag, 2018, 10.1007/s10502-018-9290-y. hal-01737568v2

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
- Some support for linguistic annotation of texts: Linguistic Linked Open Data
- TEI (theoretically) allows linking with RDF

Conclusions

Conclusions

- We consider TEI (Speech) as a good basis on which to develop a XML schema for corpora of parliamentary debates
- The proof-of-concept proposal / guidelines are on `https://github.com/clarin-eric/parla-clarin` `https://clarin-eric.github.io/parla-clarin`
- Currently no Guidelines and a very general ODD

Further steps

- Discussion tomorrow
- GitHub collaborative development of those willing (issues, pull requests, commits)
- The project must include samples of existing (validating) corpora, from siParl 1.1 for start (& something in English)
- Automatic generation of examples in p5subset.xml?

Introduction to the proposed annotation scheme

Tomaž Erjavec[1] and Andrej Pančur[2]

[1] Department of Knowledge Technologies, Jožef Stefan Institute

[2] Institute for Contemporary History

Ljubljana, Slovenia

ParlaFormat Workshop
Amersfoort, 2019-05-23