ParlaCLARIN II:
Session 1

11 May 2020
Virtual Valley
https://www.clarinet.eu/ParlaCLARIN-II
Session 1: Creation of parliamentary corpora

Jan Odijk

- New Developments in the Polish Parliamentary Corpus. Maciej Ogrodniczuk and Bartłomiej Nitoń
- Anföranden: Annotated and Augmented Parliamentary Debates from Sweden. Stian Rødven Eide
  * Not present at the virtual event.
- IGC-Parl: Icelandic Corpus of Parliamentary Proceedings. Steinþór Steingrímsson, Starkaður Barkarson and Gunnar Thor Örnólfsson
- Compiling Czech Parliamentary Stenographic Protocols into a Corpus. Barbora Hladka, Matyáš Kopp and Pavel Straňák
- Unfinished Business: Construction and Maintenance of a Semantically Tagged Historical Parliamentary Corpus, UK Hansard from 1803 to the present day. Matthew Coole, Paul Rayson and John Mariani
- The siParl Corpus of Slovene Parliamentary Proceedings. Andrej Pančur and Tomaž Erjavec
Observations

- All papers describe richly annotated formalised versions of the Parliamentary Corpora, and the work done is excellent.
- Parla-CLARIN guidelines are mentioned by 5/6, used by some (3/6), and some desired extensions/modifications described.
- Availability for download: all but not all links resolve
- Availability via a web interface: all, but all for searching for linguistic properties
- Persistent Identifiers: given by some (2/6, one of which does not resolve)
- License explicit: 3/6
- External metadata (CMDI): 2/6
Questions

1. Parliamentary data are useful for many different users, as mentioned in most submissions. But do the corpora provide the information, annotations, links to other sources, etc, that are needed by these different users, or are they just suited for (computational) linguists?

2. Continuous Process in two directions:
   a. Continuous extensions: can we fully automate this, or if not, should we not work with the parliamentary services to make this possible?
   b. All the enrichments, improvements, formalisations: do they flow back to the original providers, and if so how, and if not shouldn’t that be desirable? And if so how should that be organized and implemented technically?

3. Semantic Tagging (UK), Topics in Icelandic (Iceland): shouldn’t we work towards language-independent codes for this and for other tags needed?
What's new in the corpus (as compared to 2018)?

1. more data: 340k documents and 750M tokens
2. new MTAS-based query engine offering search across annotation layers
3. better metadata, standardized document identifiers
4. all content and metadata maintained in the database
5. utterances linked to the ontology describing politicians
6. corpus texts linked to source data
7. historical documents (from before the 1936 orthography reform) transcribed to facilitate processing with modern NLP tools
The corpus contains metadata that should interest others than only linguists:

- The speaker’s age, gender, role in the government or in the parliament, and his or her party affiliation.
- Did the speaker’s party belong to the coalition?
- Categories and topics.

This offers various kinds of quantitative analysis:

- Comparison between the genders, different age groups, political parties ...
IGC-Parl: Icelandic Corpus of Parliamentary Proceedings. Steinþór Steingrímsson, Starkaður Barkarson and Gunnar Thor Örnólfsson

The corpus will be extended yearly in almost fully automated way.

Few python scripts gather the necessary information from the parliamentary website.

- This automated process could easily be broken by minor changes by the website administrator.
- The website’s structure does not facilitate scraping.

Both us and the website’s administrator would benefit from a collaboration.
Categories and topics will be translated into English in future versions.

Language-independent codes:

- Categories (33) could be encoded in a straightforward way.
- Topics amount to more than 10,000 and new ones are added every year.
Compiling Czech Parliamentary Stenographic Protocols into a Corpus.

*Barbora Hladká, Matyáš Kopp and Pavel Straňák*

ParCzech is a project on compiling Czech parliamentary data into annotated corpora.

**ParCzech PS7 1.0**

- Stenographic protocols that record the Chamber of Deputies’ meetings held in the 7th term 2013—2017
- LINDAT repository, PID [http://hdl.handle.net/11234/1-3174](http://hdl.handle.net/11234/1-3174)
- KonText, TEITOK
- Metadata about protocols: meeting, sitting date, agenda item, authorization,
- Audio recordings
- POS tags + named-entities automatically
Compiling Czech Parliamentary Stenographic Protocols into a Corpus.

Barbora Hladká, Matyáš Kopp and Pavel Straňák

1. Parliamentary data are useful for many different users, as mentioned in most submissions. But do the corpora provide the information, annotations, links to other sources, etc, that are needed by these different users, or are they just suited for (computational) linguists?

External sources

- Ideas
  - enrichment
  - SPARQL endpoint
  - e.g. gazetteer of the members of the Parliament -> their speeches in the corpus
  - e.g. speakers in the sittings -> more info about them (party affiliation, voting etc.)
2. Continuous Process in two directions:
   a. Continuous extensions: can we fully automate this, or if not, should we not work with the parliamentary services to make this possible?
   b. All the enrichments, improvements, formalisations: do they flow back to the original providers, and if so how, and if not shouldn’t that be desirable? And if so how should that be organized and implemented technically?

Ideas
- **stable** corpora, i.e. traditional corpora
- **live** corpora, i.e. data that keeps changing
  - like source code in a version control system with its revisions, diffs, and releases
  - how to version them? An automated script that makes a monthly snapshot and preserves it in the repository, gives it a PID, etc.?
  - does it require more than one instance of searching services?

We have not talked to the original data providers yet. We plan to organize a tutorial in June 2020 and invite them.
Unfinished Business: Construction and Maintenance of a Semantically Tagged Historical Parliamentary Corpus, UK Hansard from 1803 to the present day. *Matthew Coole, Paul Rayson and John Mariani*

- Intended for corpus and computational linguists but can also be useful for political scientists, historians and content/discourse analysts.
  - Visualisations
  - Metadata
  - Future plans for semantic web links
- Toolchain is fully automated
  - Latest data available from atom feeds
  - Tagging in CLAWS and USAS
  - LexiDB can handle live updates
  - Export to TEI format
- SAMUELS project and future work collaborating with UK Parliament Hansard team
USAS semantic tagging:
- Bootstrapping expansion to six more (Chinese, Dutch, Italian, Portuguese, Spanish, Malay) (2015)
- 12+ languages at various levels (Welsh, French, Indonesian, Swedish, Turkish, Urdu) (2016)
- Common language-independent semantic field taxonomy
- Mapping to other taxonomies
- Depends on major word class

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>general and abstract terms</td>
<td>the body and the individual</td>
<td>arts and crafts</td>
<td>emotion</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>food and farming</td>
<td>government and public</td>
<td>architecture, housing and the home</td>
<td>money and commerce in industry</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>entertainment, sports and games</td>
<td>life and living things</td>
<td>movement, location, travel and transport</td>
<td>numbers and measurement</td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>O</td>
<td>P</td>
<td>Q</td>
<td></td>
</tr>
<tr>
<td>substances, materials, objects and equipment</td>
<td>education</td>
<td>language and communication</td>
<td>social actions, states and processes</td>
<td></td>
</tr>
<tr>
<td>R</td>
<td>S</td>
<td>T</td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>world and environment</td>
<td>psychological actions, states and processes</td>
<td>science and technology</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>W</td>
<td>X</td>
<td>Y</td>
<td></td>
</tr>
<tr>
<td>Z</td>
<td>names and grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The siParl Corpus of Slovene Parliamentary Proceedings. Andrej Pančur and Tomaž Erjavec

A brief account of the corpus:

- The paper describes the acquisition, up-translation, encoding, annotation, and distribution of siParl 2.0
- siParl: Parliamentary debates from the Assembly of the Republic of Slovenia 1990–2018: 8,000 sessions, 1 million speeches and 200 million words
- Extensive meta-data about the speakers, a typology of sessions etc. and structural and editorial annotations
- Tokenised, sentence segmented, part-of-speech tagged, lemmatised, parsed, and tagged with named entities
- First complete corpus to be encoded in the TEI-based Parla-CLARIN schema
- CC-BY http://hdl.handle.net/11356/1300, available in CLARIN.SI concordancers
Does the corpus provide the information, annotations, links to other sources, etc., that are needed by these different users, or are they just suited for (computational) linguists?

- We think so: the corpus has lots of metadata on speakers and political parties
- However, it is true that the corpus under the concordancers is best suited to linguists
- On the other hand, the corpus, incl. metadata TSV files on people and speeches, are also available for download
- We plan to mount the corpus in a form amiable for searching, reading and browsing in digital library: teiPublisher application in eXist-db and Interconnect the corpus in digital library with the one available via the concordancers
- Of course, further down-conversions or annotations could be made, if we knew what exactly these “different users” would require (and we had financing), or others can do it
Continuous extensions: can we fully automate this, or if not, should we not work with the parliamentary services to make this possible?

- They can be *semi*-automated, at least if we want to keep the current level of quality: the sources contain errors and omissions (names of speakers, duplicates)
- We are in touch with them and they are happy that siParl exists, however, they do not have the extra capacities that would be needed to change their workflow (they also do not have good support from their IT dept.)
All the enrichments, improvements, formalisations: do they flow back to the original providers, and if so how, and if not shouldn’t that be desirable? And if so how should that be organized and implemented technically?

- The previous slides already answers this questions: they do not flow back, this would be desirable, but is currently unrealistic
- So, also no thoughts on how this would be technically implemented