Parliamentary Speech: the case of Bulgarian Parliamentary Corpus

Petya Osenova
IICT-BAS and CLaDA-BG

ParlaFormat Workshop, 23-24 May 2019

Plan

- Background
- The status quo of the corpus
- Lessons learnt
- Standardization/Format Considerations

Background

- Political speech is in predominantly textual form (*political.webclark.org*):
  - Parliament debates transcriptions (Parliament Control Sessions from years 2006-2012)
  - Public institution websites with news sections where the important speeches are provided in transcribed form
  - Our idea on extension: not only a synchronic extension, but also diachronic - to digitize the available parliament sessions in 20s, 30s, 40s years of XX century.

Dimensions

- The attitude dimension (opinion, sentiment, authorship) – we tried to cover partially (pos, neg, sent)
- The domain dimension - topicality (law, economics, health) – not covered
- The modality dimension (speech, video, text) – not covered
- The way of communication (debate, parliamentary control (question-answer, replication) - not covered in detail

Status Quo of the Corpus

- Standardized XML encoding and TEI specification:
  - Annotation of document and paragraph level (TEI)
  - Annotation of Topics and Speakers (TEI)
  - Linguistic Annotation (Text Corpus Format (TCF))
  - The two annotation standards are interrelated via XPointer link from TCF documents as well as TEI documents

Topics and Speakers

- The topic annotation is performed on two levels: *document level* and *sentence level*
- We do not have an initial list of topics, but ask the annotators to formulate the topic as a list of key words, separated by a semicolon (a similar strategy is undertaken by CrowdFlower system for annotation of tweets)
- The sentence level topic annotation at the moment is performed only when the sentence is annotated as a sentiment or opinion statement.

<sp>ПРЕДСЕДАТЕЛ ГЕОРГИ ПИРИНСКИ</sp>

Заповядайте, господин Димитров.<sp>ПРЕДСЕДАТЕЛ ГЕОРГИ ПИРИНСКИ</sp>

<sp>ФИЛИП ДИМИТРОВ (ОДС)</sp>: Уважаеми господин председател, уважаеми дами и господа на
представители, уважаеми господин премиер! Въпросът ми, разбира се, е безнадеждно дезактуализиран, защ
което имаше да се направи по темата, вече е направено или по-скоро не е направено. Проблемът обаче е в
че Вашият отговор все пак би могъл да ни помогне поне постфактум да си обясним мотивите и смисъла на
ието на правителството.

Looking at TEI more closely

- The speaker element contains also the role or the party (but no linking was performed)
- Thus, maybe we should use the following annotation and also a direct URI to DBPedia:
- **<name role="**politician" **type="**person" **ref**= "URI-DBpedia" >Георги Пиринski**</name>**
- Also kinesic ?? for additional context of the utterances like: (звъни (rings), от място (from the seat), след почивката (after the break), тропане по банките (drumming on the desks), реплики (replication))

Speakers are distinguished only by their names

- The speaker roles might be:
  - The chair of the Parliament,
  - The representative of a certain party,
  - member of the Government, etc.
- Their role has been stored in their meta-profile (together with the social role of the politician, time history of his political positions, party membership, etc.)

- In the TEI annotation we rely on a standard XML element for speaker annotation.
- In TCF the annotation of the speaker is done by an attribute (@sp) added to the sentence element. Additionally, we put element <ns2:speakers> with children elements <ns2:speaker> with the same attribute and the same value of the attribute, and a textual description of the speaker, if any.

Questions

- Can the parliamentary speech be classified into sub-genres:
  - Law (since they are mainly law makers: a) the legal document itself and b) the debates on it. The legal issue might refer to other domains as well.)
  - Health
  - Education
  - Economics…
- The scheme should capture some common and some specific features for annotation. They might profit from the standards developed about other related domains like: law, finances, etc.
- At the moment our annotation is amorphous: the speaker, his/her role, party, etc. are captured in one element only.