A TEI-encoding of the Danish Parliament Corpus

Dorte Haltrup Hansen &
Costanza Navarretta
CLARIN-DK
University of Copenhagen, Denmark

UNIVERSITY OF COPENHAGEN

The Danish Parliamentary Corpus – version 1

- The **first Danish Parliamentary Corpus** 2009 - 2017 was distributed in 2018 in agreement with the Danish parliamentary administration.

- The corpus, consisting of app. **41 million running words** and 182.200 speeches, was distributed in the same format as received.

- We have **re-structured the corpus** to do statistics and added annotations to the data comprising information about the MPs (gender, age, origin, education), subject areas of the agenda items and linguistic information (lemma and POS-tags).

The Danish Parliamentary Corpus – version 2

- This year the Danish Parliament has released all Hansards since 2009 as **Open Data**. The format has changed since we received the first database dump.

- Because of the new xml-format from the Danish Parliament and the newly added information, we are **in need of a new format** for the next version of the Danish Parliament Corpus. As the TEI standard is well established in the area of Humanities and is a format we normally recommend, we will use that as one of the distribution formats.

View debates as transcribed speech or as meetings

- The transcriptions of the parliamentary debates are revised by the Parliament administration to conform to correct **written** Danish, therefore they cannot be considered as spoken language since they **lack characteristics such as pauses, interruptions and self-corrections**.

- Consequently, we do not use the TEI module *Transcriptions of Speech* including elements for recording, pause, utterance, vocal, shifts and prosody.

- But rather the **<meeting>**-element

The proposed structure

We are currently following the logical structure of the meetings by using:

<meeting> [meeting related information]

<list> [meeting agenda]

<item> [agenda item information]

<sp> [speech related information]

<speaker> [speaker related information]

<p> [the actual speech]

Important for us:

- The logical structure of the agendas
- Detailed information about the speaker
- Elements for classification
- Elements for topics
- The timing
- Elements for text annotation.

Structuring the meeting, speech and speaker elements

- **<speaker>-element** *persName*, *orgName* (party) and *roleName* (role in Parliament), but information about *gender, age, birth, education and origin* must be added in **<note>-elements** with @type attributes describing their semantics.
- Another way of implementing person information is**<personList><person>** instead of <speaker>.
- Elements such as *case type, case number* and *case stage* at the agenda item level must be implemented as a **<label>-element** with the attributes *n, type* and *ana* which might not be **semantically transparent.**
- Classification codes at **agenda item level** is implemented as:

  <classSpec type="model" ident="SocialAffairs"/>
- and topics at **speech level** is implemented as:

  <interp type="topic">migration</interp>

An example using the <meeting>and <sp>elements:

```
<text><body><div>
        <meeting n="20161-2">
        <list>
          <item n="1">
            <label n="7 " ana="BEH1"/>
            <title type="agendatitle">1. behandling af L 7: Om Haagerkonventionen af 2007. </title>
            <title type="longAgendatitle">Det første punkt på dagsordenen er: 1) 1. behandling af lovforslag nr. L 7: Forslag til lov om Haagerkonventionen
                        af 2007. Af social- og indenrigsministeren (Karen Ellemann). (Fremsættelse 05.10.2016). </title>
            <date> 2016-10-05</date>
            <classSpec type="model" ident="SocialAffairs"/>
            <note type="rubrica">Forhandling</note>
       ->     <sp xml:id="sp2016211" >
              <speaker>
                <persName>Pernille Rosenkrantz-Theil</persName>
                <orgName> S</orgName>
                <roleName> medlem </roleName>
                <note type="gender">Female</note>
                <note type="age">40</note>
                <note type="birth">1977-01-17</note>
                <note type="education">Bachelor i statskundskab, Københavns Universitet</note>
                <note type="origin">Skælskør</note>
              </speaker>
              <timeline origin="start"><when absolute="13:00:05"/> </timeline>
              <timeline origin="end" unit="s" interval="120"><when absolute="13:02:05"/> </timeline>
              <interp type="topic">migration</interp>
              <p> Der sker jo det med globaliseringen, at flere og flere bliver gift på tværs af landegrænser…. </p>
```

OR:

```
            <p>     <w xml:id="w20162111" lemma="der" ana="UNIK">Der</w>
                    <w xml:id="w20162112" lemma="ske" ana="V_PRES">sker</w>
                    <w xml:id="w20162113" lemma="jo" ana="ADV">jo</w>
                    <w xml:id="w20162114" lemma="det" ana="PRON_PERS">det</w>
             …</p>
       ->    </sp>
          </item> </list></meeting> </div></body></text></TEI>
```

Comparison of TEI/speech/speaker and TEI/utterance

- It is possible to exchange the **<sp>-element** with the **<u>-element**. The <speaker> will then represented as <listPerson><person>:

<item>:	**<item>:**
<sp>	<u>
<speaker>	<listPerson><person>
<timeline>	<timeline>
<interp>	<interp>
<p>	<s><w>

- And the person elements do even have clearer semantics without the need of <note>-elements with type-attributes.

The same example using the <u>-element:

```
<text><body><div>
        <meeting n="20161-2">
        <list>
          <item n="1">
            <label n="7 " ana="BEH1"/>
            <title type="agendatitle">1. behandling af L 7: Om Haagerkonventionen af 2007. </title>
            <title type="longAgendatitle">Det første punkt på dagsordenen er: 1) 1. behandling af lovforslag nr. L 7: Forslag til lov om Haagerkonventionen
                        af 2007. Af social- og indenrigsministeren (Karen Ellemann). (Fremsættelse 05.10.2016). </title>
            <date> 2016-10-05</date>
            <classSpec type="model" ident="SocialAffairs"/>
            <note type="rubrica">Forhandling</note>
        ->      <u xml:id="u2016211" >
            <listPerson>
               <person xml:id="I27" role="medlem">
                        <persName> Pernille Rosenkrantz-Theil </persName>
                        <sex> F</sex>
                        <age> 42</age>
                        <birth> 1977-01-17 </birth>
                        <education> Bachelor i statskundskab, Københavns Universitet </education>
                        <affiliation> S</affiliation>
                        <residence> Skælskør </residence>
                </person>
            </listPerson>
            <timeline origin="start"><when absolute="13:00:05"/> </timeline>
            <timeline origin="end" unit="s" interval="120"><when absolute="13:02:05"/> </timeline>
            <interp type="topic">migration</interp>
            <s> Der sker jo det med globaliseringen, at flere og flere bliver gift på tværs af landegrænser…. </s>
                                                                                                    OR:

            <s>     <w xml:id="w20162111" lemma="der" ana="UNIK">Der</w>
                    <w xml:id="w20162112" lemma="ske" ana="V_PRES">sker</w>
                    <w xml:id="w20162113" lemma="jo" ana="ADV">jo</w>
                    <w xml:id="w20162114" lemma="det" ana="PRON_PERS">det</w>
                      … </s>
      ->    </u> </item> </list></meeting> </div></body></text></TEI>
```

Conclusion

- A common TEI format for parliamentary data might be very useful for concrete projects and applications across languages. e.g. allowing for comparison of subject areas and topics discussed in the parliaments or the distribution of parties, gender and speaking time.

- The point of departure could be **<meeting>** or *Transcriptions of Speech*

- When working with concrete problems or projects people tend to use the format that is most useful for their task.

- We therefore opt for a common **minimal TEI exchange format** constructed with specific purposes in mind e.g. inter-linking of subject areas, topics and/or timing.