From Project Data to Sustainable Archiving of Linguistic Corpora

Thomas Schmidt (University of Hamburg) & Andreas Witt (University of Tübingen) http://www.sfb441.uni-tuebingen.de/c2/index-engl.html

Deutsche Forschungsgemeinschaft DFG

Project "Sustainability of Linguistic Resources"

Project facts

- since 2006

A joint initiative by three German research centres:
- SFB 441 'Linguistic Data Structures' (Tübingen)
- SFB 538 'Multilingualism' (Hamburg)
- SFB 632 'Information Structure' (Potsdam/Berlin)
- over 40 individual research projects
- heterogeneous collection of linguistic corpora: written and spoken, many languages, different research paradigms, different annotation categories, different corpus tools and storage formats

Project aims

- consolidate different approaches to corpus encoding and processing
- pave the way for sustainable archiving of linguistic resources
- develop rules of best practice for sustainable data handling and sharing

Seven work packages

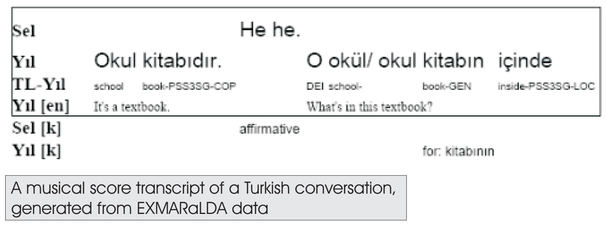
- Annotation frameworks
- Query tools
- Tools for data access
- Meta-data
- Integration of terminologies
- General rules of best practice
- Legal issues in data sharing and archiving

TUSNELDA, EXMARaLDA and PAULA

```
<figure id="s35b5" entity="belgiji/s35b5.bmp">
 <figtrans>
   <sp who="Obeliks">
     <spokenpar>
       Gde da na&dstrok;em belu zastavu ?
       <marked type="deic-loc">Ovde</marked>
       je sve pusto !
     </spokenpar>
     <situation>
       <keywords>
         <term>open hands <te
         <term>slightly bent<
       </keywords>
     </situation>
   </sp>
   <sp who="Asteriks">
     [...]
   </sp>
 </figtrans>
</figure>
```

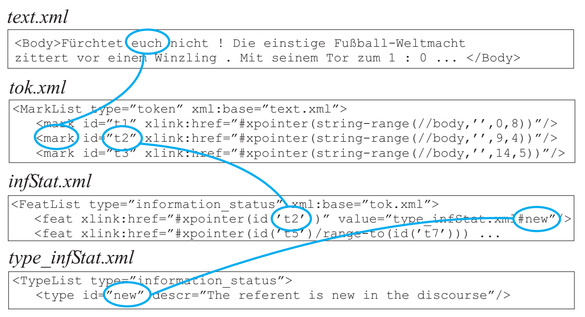
TUSNELDA (Tübingen)
- XML format, data repository, annotation guideline
- predominantly intended for corpora of *written* language
- *hierarchy*-centric conception of data
- *inline* annotation (single document)
- based on suggestions by the *TEI* and *XCES*

Figure on the left: an Asterix cartoon and its annotation in TUSNELDA



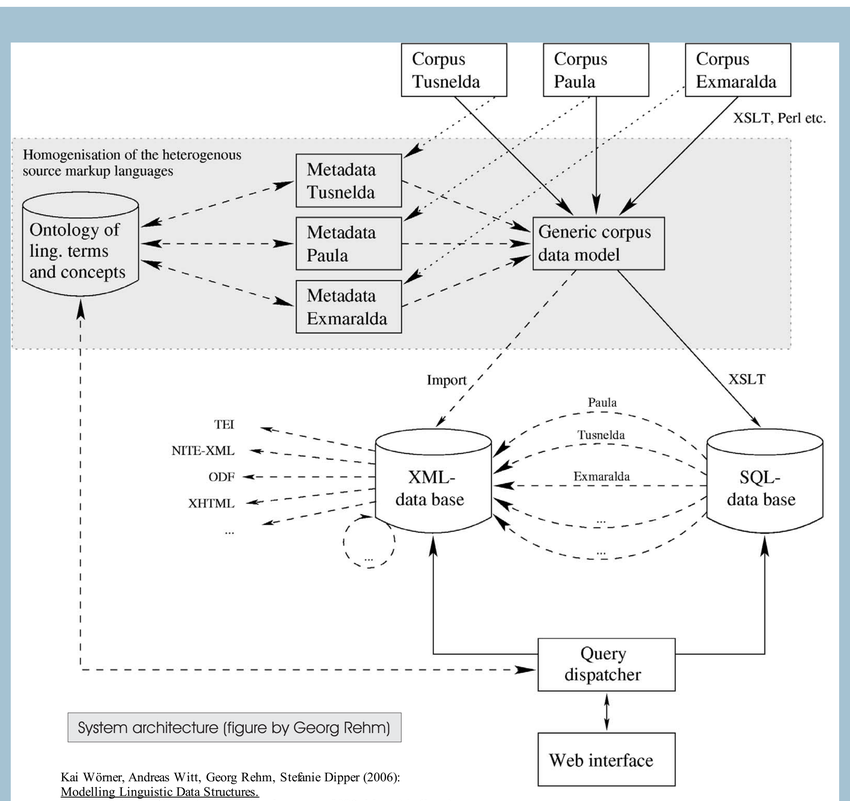
A musical score transcript of a Turkish conversation, generated from EXMARaLDA data

EXMARaLDA (Hamburg)
- XML format, tools for transcription, corpus management, query
- predominantly intended for corpora of *spoken* language
- *timeline*-centric conception of data
- *standoff* annotation (single document)
- based on the idea of *Annotation Graphs*
- EXMARaLDA for spoken, TEI/MENOTA for written data!



PAULA (Potsdam/Berlin)
- XML format, data repository, annotation guideline
- for corpora of *spoken and written* language
- hybrid conception of data (multiple hierarchies with timestamps)
- *standoff* annotation (multiple documents)
- based on Linguistic Annotation Framework (LAF)

Figure on the left: an example of PAULA standoff annotation of information structure



System architecture (figure by Georg Rehm)

Kai Wörner, Andreas Witt, Georg Rehm, Stefanie Dipper (2006):
Modelling Linguistic Data Structures.
In: Proceedings of the Extreme Markup Languages 2006, Montréal, Canada.

Approach: Generalised data model with different serialisations

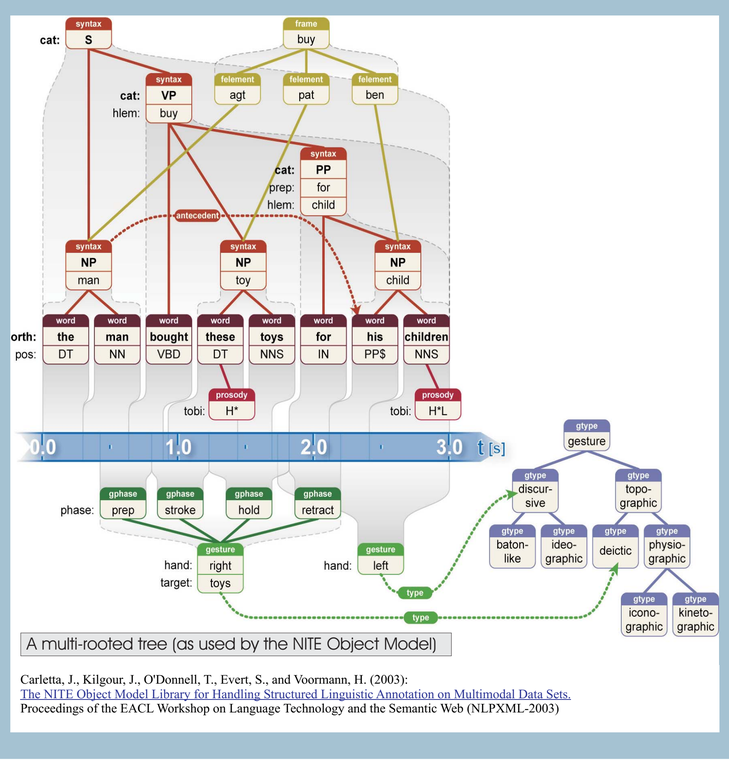
- generalise over existing frameworks
- abstract over physical file formats (serialisations)

→ **Generic Corpus Data Model (GCDM)**

- provide mappings from existing frameworks into GCDM
- provide mappings from GCDM to sustainable serialisations (e.g. TEI conformant XML)

Properties of the Generic Corpus Data Model

- a multi-rooted tree
- optional time-stamps
- trees relate to another via shared leafs
- similar to the NITE Object Model (Carletta et al. 2003), but: more flexible base level



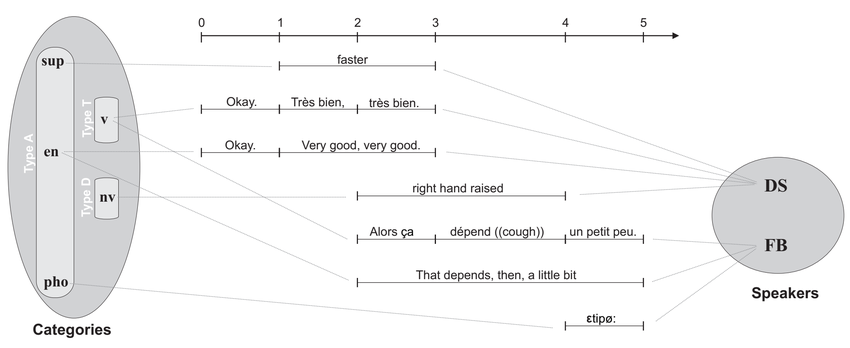
A multi-rooted tree (as used by the NITE Object Model)

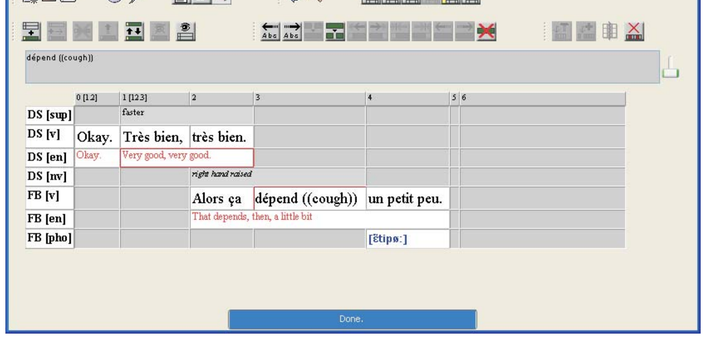
Carletta, J., Kilgour, J., O'Donnell, T., Evert, S., and Voormann, H. (2003):
The NITE Object Model Library for Handling Structured Linguistic Annotation on Multimodal Data Sets.
Proceedings of the EACL Workshop on Language Technology and the Semantic Web (NLPXML-2003)

Example: Time-based data models and the TEI's Guidelines for Transcriptions of Speech

The single timeline, multiple tiers data model

- one fully ordered timeline
- events, organised into tiers
- assign each event to a start and end point on the timeline
- no overlap of events within a tier
- assign tiers to speaker, type and category
- used by: PRAAT, EXMARaLDA, TASX-Annotator, ELAN, ANVIL, ...





Serialisation in an EXMARaLDA file (stratified)

```
<basic-transcription>
    <head>
        <speakertable>
            <speaker id="SPK0" abbreviation="DS"/>
            <speaker id="SPK1" abbreviation="FB"/>
        </speakertable>
    </head>
    <body>
        <common-timeline>
            <tli id="T1"/>
            <tli id="T2"/>
            <!-- [...] -->
            <tli id="T7"/>
        </common-timeline>
        <tier id="TIE1" speaker="SPK0" category="sup" type="a">
            <event start="T2" end="T4">faster</event>
        </tier>
        <tier id="TIE2" speaker="SPK0" category="v" type="t">
            <event start="T1" end="T2">Okay. </event>
            <event start="T2" end="T3">Très bien, </event>
            <event start="T3" end="T4">très bien. </event>
        </tier>
        <tier id="TIE3" speaker="SPK0" category="en" type="a">
            <event start="T1" end="T2">Okay. </event>
            <event start="T2" end="T4">Very good, very good.</event>
        </tier>
        <tier id="TIE4" speaker="SPK0" category="nv" type="d">
            <event start="T3" end="T6">right hand raised</event>
        </tier>
        <tier id="TIE5" speaker="SPK1" category="v" type="t">
            <event start="T3" end="T4">Alors ça </event>
            <event start="T4" end="T6">dépend ((cough)) </event>
            <event start="T6" end="T5">un petit peu. </event>
        </tier>
    </body>
</basic-transcription>
```

Compatibility between the TEI's guidelines for transcriptions of speech and timeline-centric data formats (e.g. Praat, EXMARaLDA, ELAN) is achieved via the reference to **a common underlying data model** - the "single timeline, mutliple tiers" data model.

Going from the EXMARaLDA format to a TEI conformant format is a **hierarchisation**:
- summarise contiguous sequences of events in tiers of type 't' into **<u>** elements
- use **start** and **end** attributes to assign **<u>** elements to the timeline
- use **<anchor>** elements for timeline assignment inside **<u>** elements
- map other elements accordingly

Going the other way is a **stratification:**
- decompose **<u>** elements into **<event>** elements (according to timestamps)
- distribute events onto tiers (one tier per speaker/category combination)

Benefits:
- clearly defined subset of TEI tags for transcriptions of speech
- clearly defined usage of these tags
- use existing transcription software to create or process TEI-conformant data

Serialisation in a TEI conformant file (hierarchised)

```
<TEI.2>
    <teiHeader>
        <fileDesc/>
        <profileDesc>
            <particDesc>
                <person id="DS"/>
                <person id="FB"/>
            </particDesc>
        </profileDesc>
    </teiHeader>
    <text>
        <timeline>
            <when id="T0"/>
            <when id="T1"/>
            <!-- [...] -->
            <when id="T5"/>
        </timeline>
        <u who="DS" start="T0" end="T3">
            <div type="segmental">
                Okay. <anchor synch="T1"/>
                Très bien, <anchor synch="T2"/>
                très bien.
            </div>
            <div type="prosody">
                <prosody feature="tempo" desc="getting faster" start="T1" end="T3"/>
            </div>
        </u>
        <event who="DS" desc="right hand raised" start="T2" end="T4"/>
        <u who="FB" start="T2" end="T5">
            <div type="segmental">
                Alors ça <anchor synch="T3"/>
                dépend <vocal desc="cough"/> <anchor synch="T4"/>
                un petit peu.
            </div>
        </u>
    </text>
</TEI.2>
```