1. Challenges

- Different types of annotations to query and visualize
  - (Semi-)automatic: multiple taggers, constituency/dependency parsers ...
  - Manual: coreference, information structure, rhetorical structure ...

- Different types of corpus data to support
  - Historical corpora: diplomatic and normalized text, manuscript structure
  - Multimodal corpora: aligned audio/video, multiple overlapping speakers
  - Parallel corpora: representing conflicting word & sentence alignment
  - Learner corpora: conflicting base text and target hypotheses
  - And other multilayer corpora: any and every annotation may repeat/conflict with other structures

Many corpora violate assumption of one continuous stream of segments (multiple languages, speakers, corrected texts...)
- Combinatorial explosion of types: unrealistic to design a system for each
- Reusing the architecture for unique search and visualization applications
- Simplifying the query language (AQL) to deal with new structures

2. Unified data model and query language

Dealing with multiple source formats
- Annotations come from multiple formats
- Convert multiple formats with SaltNPepper (Zipser & Romary 2010)
- Salt model represents merged data in ANNIS (Zeldes et al. 2009)
- Reconcile conflicting segmentations
- Archive data in PAULA XML (Dipper 2005), a standoff XML format for multi-layer corpora

Segmentations in the new ANNIS3 data model
- Deal with multiple alternative base texts: one segmentation each
- Any annotation layer can be a segmentation:
  - Diplomatic/normalized word forms
  - Broad and narrow phonetic transcription
  - Data from different speakers
- Segmentations can be selected as:
  - The base text for concordance KWIC views (Key-Word in Context)
  - The unit for defining the desired context (e.g. ±5 normalized word forms)
- Search criteria for proximity and adjacency in the ANNIS Query Language (AQL), using typed precedence operators:
  
  Search consecutive utterances of a speaker (even if others intervene):
  &quot;ja&quot; /the instructor says ja twice

Find differently spelled words within 10 diplomatic units in a manuscript:
/\S/.*/dipl.1,10 /f.*/ & //words in s- and f- in 1-10 dipl

3. Reusable, configurable visualizations

- Dedicated visualization are needed for many common data types:
  - Syntax trees, coreference, rhetorical trees, etc.

- But many corpora have unique data types:
  - Alternative views of the base text in digital editions of manuscripts
  - Alternating, non-overlapping text in subtitle corpora, film transcripts
  - Conflicting annotations for different layers of information structure ...
  - Impossible to foresee all necessary visualization types

- Approach: Use annotation triggered style sheets
  - Expressiveness of HTML5 with flexibility of CSS3
  - Short development cycles from corpus to visualization
  - Implementation as configuration file and CSS file:

<table>
<thead>
<tr>
<th>token</th>
<th>span</th>
<th>style:word value</th>
<th>spkr</th>
<th>div:spkr</th>
<th>style:spk value</th>
<th>lang</th>
<th>lang:lang value</th>
<th>info</th>
<th>t:title</th>
<th>style:info value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Applications and future directions

- A variety of dedicated visualizations can be developed with little code
- Visualization of Information Structure in PCC (Stede 2004)
- Digital manuscript editions for Coptic (Projects KOMeT/SCRIPTORIUM): http://coptic.pacific.edu/

Some planned extensions include:
- Adding matching javascript files for more interactive visualizers
- Visualizer-triggered searching (click on words, jump between linked results)
- Aggregate visualizers based on results from multiple documents/corpora

References


