Alexandre Arkhipov
Moscow State University

Towards a more general model of interlinear text

ICLDC'2013 – Honolulu, March 1
I  Interlinear text model in [BBH 2003]
II  Extensions
   1 Axes
   2 Alternatives
   3 Multi-speaker and multi-lingual texts
   4 Comments and versioning
   5 Non-linear markup (syntactic, semantic etc.)
III  RDF back-end and LLOD perspective
I. Interlinear text model in BBH 2003
Cats are hunters.

Cats

are

hunters

cat -s

are

hunt -er -s
Cats are hunters.
II.1 From a tree to a set of trees

- An *interlinear-text* object is a tree.
- An *interlinear-text* object is bound to an **axis** (its root annotation is obligatorily aligned to the axis, child annotations are possibly aligned too).
II.1 From a tree to a set of trees

Cats are hunters.
II.1 From a tree to a set of trees

Interlinear-text is bound to an axis

Cats are hunters.

Tree, simplified
II.1 From a tree to a set of trees

- An *interlinear-text* tree is bound to an **axis**
- More than one axis (and >1 tree) may be needed for a complete linguistic analysis of a text (speech event)
- Each **axis type** can have its own **units** for fragment identification
- Axis types (+units)
  - Normalized (abstract) text: Plain text (character, line)
  - Transcribed text: Timeline (ms, digital samples)
  - Document: Graphic media (page + area)
  - Document: Formatting objects (page+block+char,line)
  - ...
II.1 Axis types & units: Abstract text

Plain text axis (character, line)

Cats are hunters.
II.1 Axis types & units: Transcribed text

Timeline (ms, digital samples)

[dɔɡz mm | kʰæts əː (0.4) | hʌn= | hʌntəz]
II.1 Axis types & units: Document

Graphic media (page + area)
Inter-axis alignment
Annotations in one axis can be aligned to (annotations in) another axis, e.g.:

• (abstract) text segments aligned to transcription segments / timeline

• BOLD-style audio annotations: sound fragments of different duration aligned to each other

• sign languages aligned to spoken language translations

• retelling a movie/cartoon: segments aligned to corresponding video fragments
II.1 Bundle of axes

Inter-axis alignment
Annotations in one axis aligned to (annotations in) another axis

Cats are hunters.
II.2 Alternative analyses

Alternative analyses of all kinds, including root annotations (e.g. alternative transcriptions; lexical and morphological homonymy; syntactic ambiguity) need to be stored and displayed as such

- Each alternative creates a divergence point (alternative subtrees)

- Support for feature-labeling of alternatives
  Marking divergence points for user-specified «features» allows to select for review e.g. all open/close vowel alternatives, or all Perfect vs. Evidential alternatives in a corpus

- «Feature values» for consistent choice of alternatives
  Marking each subtree for the particular analysis choice yielding this subtree allows to simultaneously settle e.g. all open/close vowel alternatives to close in one action
II.2 Alternative analyses

Features for marking alternatives

An example for ambiguity in transcription: full vowel vs. schwa

```xml
  <item type="txt" lang="aqc__IPA">nədo</item>
</word>
  <alt-default>
    <alt-feature alt-fname="vowel reduction" alt-fvalue="schwa" agreeID="11"/>
  </alt-default>
  <alt-option>
    <alt-feature alt-fname="vowel reduction" alt-fvalue="full" agreeID="11"/>
      <item type="txt" lang="aqc__IPA">nodo</item>
    </word>
  </alt-option>
</alt>
```

«schwa» option (default)

«full vowel» option
II.3 Multi-speaker and multi-lingual texts

- Multi-speaker texts are easily accounted for by introducing a "participant" attribute on segments.
- Multi-lingual texts are easily accounted for by introducing a "language" attribute on segments.
  - Indispensable for correctly dealing with code-switching.
  - Also for quotations, borrowings etc.

Note: in current versions of SIL FLEEx this cannot be done since each project can only hold data for one language (even if the xml format allowed).

From the application point of view, texts are better stored independently of grammar/lexica.
II.4 Comments and versioning

- Every piece of data can have associated comments
- Every piece of data can have associated attributes like confidence levels, grammaticality judgements (esp. for elicited texts but not only), workflow stages and assignments («check sound», «check grammar», «for John to approve» etc.)
- Every piece of data can have metadata attributes (created/edited by, created/edited timestamp etc.) and thus allow tracking of changes and version control
II.5 Non-linear markup

The basic interlinear setup is designed principally for morphological annotation, most importantly for linear annotation.

A more general format must allow for non-linear kinds of markup as well (e.g. dependency trees, constituency trees) necessary for full-scale syntactic or semantic analysis.

Grouping of (non-)contiguous elements (e.g. periphrastic forms) should also be supported.

Thus the model must support annotations as relations between annotations, overlaid upon the «basic» interlinear tree.
A fully-detailed XML implementation is possible but extremely complex. Moreover, for any particular editing / management / analysis application only a part of the whole data structure would probably be relevant.

Thus one can envisage a complex system which uses different data formats for different purposes, cf. S. Moran's PHOIBLE project [Moran 2012] (relational DB + huge flat plain text file + RDF/OWL repository).

RDF is also a natural solution in the LLOD perspective (Linguistic Linked Open Data, see [Chiarcos et al. 2011]).

An RDF-like intermediary triple repository is also proposed for ELAN-FLEEx interoperability (see Nakhimovsky's presentation).
The aim is to design a system as outlined below:
References


