Towards a general model of interlinear text

Cathy Bow, Baden Hughes, Steven Bird
University of Melbourne

E-MELD objectives:
- Best practice methods for annotating text
- Developing appropriate archiving methods

Overview of presentation
1. Survey of interlinear texts
2. Discussion of issues raised in survey
3. Proposed model
4. XML Representation
5. Topics for further research

Survey – standard 3 levels (Nepali)

- 3 lines: text, gloss, FT
- Phonetic transcription aligned word-by-word to gloss
- Numbered phrases with FT
- Portmanteau morphemes

Survey – standard 3 levels (Ainu)

- Phonetic transcription aligned word-by-word to gloss
- Numbered phrases also with FT – separate on page
- Gaps in glossing
- Wrapping of text

Survey – different alignment (Nivkh)

- Text and gloss aligned by morpheme
- Metadata beneath text

Survey – extra information (Garrwa)

- Two speakers
- Extra layers of notes by different analysts
- Incomplete data / annotation
- Problems of alignment
Survey – extra info (South Efate)

- Four lines of metadata
- Links to audio files
- Transcription of word & morpheme
- Gloss and POS label
- FT in two languages

Survey – phonetic data (Ega)

- Praat text grid
- Layers of phonetic analysis
- Segmentation of audio file

Survey – complex examples (Hebrew)

- Non-Roman characters transliterated
- Text written R-L, gloss L-R
- No morpheme breakdown
- No FT
- Ostensibly uses canonical versification
- Numbers refer to concordance

Survey – Yidinj & Diyari

- Numbered phrases of text, gloss, FT
- Yidinj: no separation of words into morphemes
- Diyari: words separated into morphemes by hyphens

Discussion – Mapping & Alignment

Mapping
- One-to-one from word or morpheme to gloss
- One-to-many – e.g. portmanteau morphemes
- One-to-zero – missing information
- Many-to-one, zero-to-one – rare but possible

Alignment
- Vertical – gloss aligned to morpheme or word
- Horizontal – can lead to wrapping
- Some lines of information wrap as units

Discussion – presentation issues

- Page presentation
- Text above gloss
- FT can be within text or separately below
- Notes within text or separately below
- Use of line numbers
- Position of metadata

- Typographical issues
- Use of punctuation, indentation
- Use of typeface to distinguish information
- Bold, italic, font size
By-and-by he asks me [where I was in the battle, Damari thought to himself.]

They used to wear clothes

Proposed Four-level Model

XML Representation

XML Representation (cont)
**XML Representation (example)**

```xml
<interlinear-text>
  <item type="title">SE Text</item>
  <item type="media">kalsrap.mov</item>
  <item type="comment">Story from tape 20001bx told by Kalsarap Namaf. Transcribed and translated into Bislama by ...</item>
</interlinear-text>
```

**Simple XML DTD**

```xml
<!ELEMENT document (interlinear-text*)>
<!ELEMENT interlinear-text (item*, phrases)>  
<!ELEMENT phrases (phrase*)>  
<!ELEMENT phrase (item*, words)>  
<!ELEMENT words (word*)>  
<!ELEMENT word (item*, morphemes)>  
<!ELEMENT morphemes (morph*)>  
<!ELEMENT morph (item*)>  
<!ATTLIST item type CDATA #IMPLIED>
```

**Rendering Issues**

- content grouping
- rows to be displayed
- row styles
- row ordering

**XSL Implementation**

```
<xsl:template match="phrase">
  <phrase>
    <xsl:apply-templates select="words"/>
    <xsl:apply-templates select="item"/>
  </phrase>
</xsl:template>
```

**XSL Example**

```xml
<xsl:template match="phrase">
  <phrase>
    <xsl:apply-templates select="words"/>
    <xsl:apply-templates select="item"/>
  </phrase>
</xsl:template>
```
Example: Nenets

An excerpt from Nead 1990, p.29.

- Nenets
  - ancient ABS NOM SG person ABS GEN PL
  - song ABS NOM PL what kind LIM ABS NOM SG
  - tale ABS ACC PL yunboc ABS ACC PL
  - yunboc ABS ACC PL

Traditional folk songs. Besides presenting various kinds of iñiñik, lament recreatives (yuni), and heroic
recreatives (yunboc),...

Example: Document level

Example: OLAC Metadata

<interlinear-text>
<olac>
<dc:title>Title</dc:title>
<dc:language xsi:type="language" code="x-sil-BAN">
<dc:author>Lastname, Firstname</dc:author>
...
</olac>
</interlinear-text>

Extension: Punctuation

- Morpheme types:
  - prefix, suffix, proclitic, enclitic
  - punctuation on certain rows only (",", "+")
- Add type attribute to <morph>
  
  
  <ATTLIST morph type (prefix | suffix | proclitic | enclitic) #REQUIRED>
  
  
  <word>
  <item type="txt">ComplexWord's</item>
  <morphemes>
  <morph type="prefix">
  <item type="txt">Complex</item>
  <item type="gls">PrefixGloss</item>
  </morph>
  <morph>
  <item type="txt">Word</item>
  <item type="gls">RootGloss</item>
  </morph>
  <morph type="enclitic">
  <item type="txt">s</item>
  <item type="gls">CliticGloss</item>
  </morph>
  </morphemes>
  </word>

Topics for Further Research

- Editing operations
- Audio alignment
- Incorporating OLAC metadata
- Application programming interface
- Implementation
- Format conversion tools