EMELD Language Digitization Project
Conference 2003
Workshop on
Digitizing & Annotating Texts & Field Recordings
LSA Institute, Michigan State University, July 11th-13th

Wayne State University - Eastern Michigan University

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)

Lobhi Kukur (Lo)[1]

(1)
yo yoTaa lobhi kukur-ko kathaa ho.
PROX.L one greedy dog-GEN story be1.3smL
This is a story of a greedy dog.

(2)
yo kathaa-ko sirsak lobhi kukur ho.
PROX.L story-GEN title greedy dog be1.3smL
The title of this story is 'Greedy Dog'.

(3)
ek coTi yoTaa kukur khaanaa-ko khoji-maa baahira nisk-yo.
one instance one dog food-GEN search-LOC outside come.out-3smL.PST
Once, one dog went out in search of food.

(4)
u hi~D-day gar-eko thi-yo.
3L walk-SP do-PP be.PST-3smL.PST
He was walking.

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

Survey – standard 3 levels (Ainu)

1. *I-resu yupi i-resu sapo i-res-pa hine oka-an ike:-*
   1SG/O-raise brother foster sister 1SG/O-raise-PL and be (PL)-1SG then
2. *Kamuy kat casi casi-upsor a-i-o-resu.*
   god build castle castle-inside PASS-1SG/O-APPL-raise
3. *Tapan inuma ran-pes kunne cirikinka, enkasike nispa-mut-pe*
   such treasure cliff like rise high over there master-wear-thing
   otu-santuka o-uka-uyru otu-pusa-kur suypa kane asso-kotor mike
   many-hilt APPL-REC-exist many-knot-shadow sway gold wall glitter
   kane anramasu auwesuye.
   gold pleasing interesting

Translation:
(1) My foster brother and foster sister raising me, we lived then. (2) The god-built mountain castle, inside the mountain castle, I was raised. (3) The pile of treasure was heaped like a cliff, and above it the master's swords were crossing their hilts, and when the shadows of the sword knots swayed, the walls glittered in gold. How beautiful and how interesting! (4) In front of

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

Survey – different alignment (Nivkh)

p' -at'ɨk -xe p' -nanak -xe pań -d'. atik
REFL younger-brother AND REFL elder-sister AND grow-up FIN younger-
maika -d'. k'u -ye puńd-ye ḷbo -ror p'u -r
brother be-small FIN arrow AND bow AND take GER-3SG go-out GER-3SG
ɨevrq ya -d'. iy -ror ƒak -ŋan nanak ɨevrq tupƒ
bird shoot FIN kill GER-3SG bring GER elder-sister bird feather
ƒavrk -r'. t'uur-tox ƒa -tot iń -d' -yu.

(Extracted from V. Z. Panfilov, *Grammatika nivxskogo jazyka*, vol. 2. Moscow-Leningrad, 1965.)

Notes
For the syntax of sentences consisting of more than one clause, in particular for the function of verb forms glossed 'GER' and 'AND', see the discussion on pp. 270–2.
A younger brother and an elder sister grew up. The brother was small. He took his arrows and bow, went out, and shot birds. When he killed them, he brought them and the elder sister plucked the birds' feathers. When they had cooked them on the fire, they ate them.

- **Text and gloss aligned by morpheme**
- **Metadata beneath text**
- **Separate section of notes**
- **Separate FT of entire text**

Survey – extra information (Garrwa)

\sp	G
\ft	nanyi??/nganyi?? ngaranjan jilajbaya ngabayan yanba *cont.*
\fg	my mother go white man talk *cont.*
\ft	ngangangi
\fg	want
\ncr	
\ncft	It is difficult to determine word order here, and which sentences words belong to.
\ncfg	"white man" refers to a "Mr Haely??", whose name is written beneath *ngabayan*.
\fft	
\ncfft	
\sp	S
\ft	ngayu junkuyi nanyina namukeya yanbalajba
\fg	I been up here I sitting down here paint
\ncfg	This gloss (and also the free translation) are quite tentative, as it is difficult to work them out from the original transcription.
\fft	I been up here. I couldn't go down there [paint]

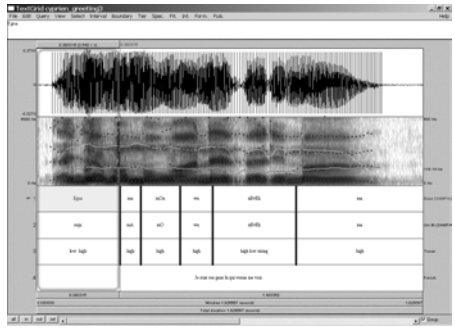
- **Two speakers**
- **Extra layers of notes by different analysts**
- **Incomplete data / annotation**
- **Problems of alignment**

Survey – extra info (South Efate)

```
\_sh v3.0  485  SE Text
\itm kalsrap.mov
\nt Story from tape 20001bx told by Kalsarap Namaf. Transcribed and translated
into Bislama by Manuel Wayane. The story concerns a natopu or spirit called
Litrapong, also known in Bislama as a Lisepsep. Story is also told on video.

\aud kalsrap.mov
\as 0
\ae 13.0002
\tx  Akit    tumaui              tae   esan  ipi,       go
\mr  akit    tu-       mau       tae   esan  i     - pi go
\mg  1plincS 1plincRS- all       know  place 3sgRS - be and
\POS pron    pron-     quantifier vambi n     pron  - v  conj

\fg We all know that place, and this Litrapong, I want to tell you about
Litrapong. She is of grandfather's clan. Those two, grandfather and Litrapong,
would talk every now and then.
\fgb Yumi evriwan isave ples ia. Mo Litrapong (Lisepsep) ia. Mi wantem talem
long yufala abaot Litrapong ia. Hemi naflak blong olfala. Hem mo olfala (apu)
tufala istap storian samtaem.
```

- **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)

	8414	1961	776	776	853	8064	853	430	1254	7218
1.	תֹהוּ	הָיְתָה	וְהָאָרֶץ	הָאָרֶץ׃	וְאֵת	הַשָּׁמָיִם	אֵת	אֱלֹהִים	בָּרָא	בְּרֵאשִׁית
2.	tohu	hoytah	veha'arets	ha'arets	we'et	hashamayim	'et	'Elohim	bara'	bere'shit
	desert (formless)	proved to be	and the earth (to be firm)	the earth	and	the heavens		God (plural of excellence)	created (cut)	in the beginning (head)

- **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**

98. *garu* *ŋapaŋ* *yaymi:lna*
by-and-by I-O ask-PURP
[I'll tell this to Guyala when] by-and-by [he] asks me [where I was in the battle, Damari thought to himself.]

Yidinj
- **no separation of words into morphemes**

4. *kaṭi* *ṭana* *wiri-ṇḍa* *pudi-yi*
clothing-ABS 3pls wear-PART AUX-PRES
They used to wear clothes

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**

p' *-nanak* *-xe* *pań* *-d'.*
REFL elder-sister AND grow-up FIN

gar-eko thi-yo.
do-PP be.PST-3smL.PST
kane asso-kotor
gold wall

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

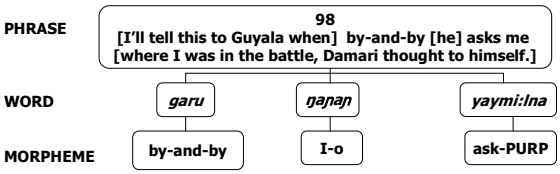
3. *Tapan inuma ran-pes kunne cirikinka, enkasike nispa-mut-pe*
such treasure cliff like rise high over there master-wear-thing
otu-santuka o-uka-uyru otu-pusa-kur suypa kane asso-kotor mike
many-hilt APPL-REC-exist many-knot-shadow sway gold wall glitter
kane anramasu auwesuye.
gold pleasing interesting

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

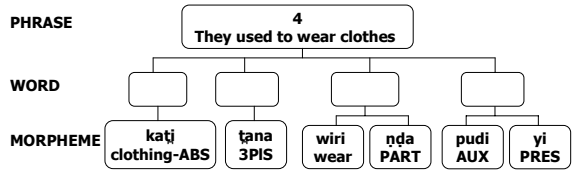
Building a Model - Yidinj

98. *garu ŋaɲaɲ yaymi:lna/*
by-and-by I-o ask-PURP
[I'll tell this to Guyala when] by-and-by [he] asks me [where I was in the battle, Damari thought to himself.]

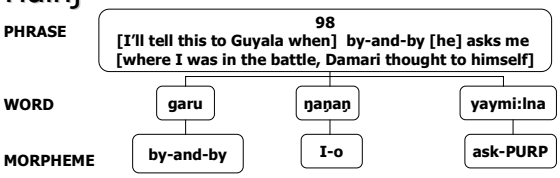


Building a Model - Diyari

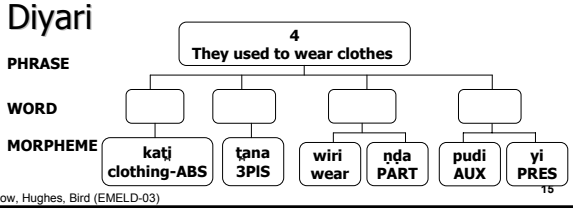
4. *kaṭi ṭana wiri-ɳḍa pudi-yi*
clothing-ABS 3PlS wear-PART AUX-PRES
They used to wear clothes



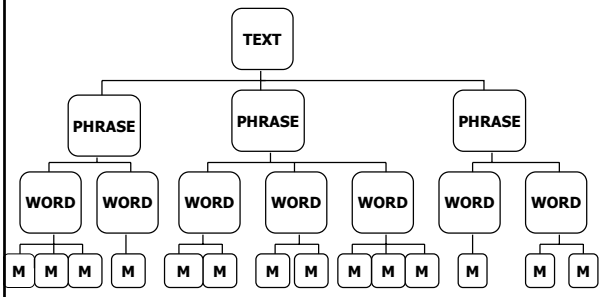
Yidinj



Diyari



Proposed Four-level Model



XML Representation

```
<interlinear-text>
  <item type="user-defined">
    Content at the text level, such as metadata,
    or an unaligned transcription of the entire
  text,
    or a pointer to an unaligned audio file
  </item>
  <phrases>
    Nested XML content to represent the phrasal
    constituents of the text
  </phrases>
</interlinear-text>
```


XML Representation (cont)

```
<interlinear-text>
  <item type="title">The Title</item>
  <phrases>
    <phrase>
      <item type="gls">A phrasal translation</item>
      <words>
        <word>
          <item type="txt">word</item>
          <morphemes>
            <morph>
              <item type="txt">Morph</item>
              <item type="gls">Gloss</item>
            </morph>
            <morph>
              <item type="txt">Morph</item>
              <item type="gls">Gloss</item>
            </morph>
          </morphemes>
        </word>
      </words>
    <phrase>
  </phrases>
</interlinear-text>
```

XML Representation (example)

```
<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>
  <phrases>
    <phrase>
      <item type="gls1">We all know that place ...</item>
      <item type="gls2">Yumi evriwan isave ples ia ...</item>
      <words>
        <word>
          <item type="txt">Akit</item>
          <morphemes>
            <morph>
              <item type="txt">akit</item>
              <item type="gls">1plincS</item>
              <item type="POS">pron</item>
            </morph>
          </morphemes>
        </word>
        ...
```

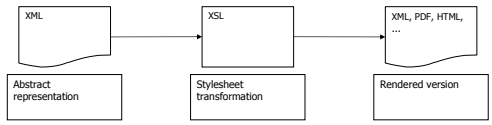

Simple XML DTD

```
<!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*)>

<!ELEMENT item (#PCDATA)>

<!ATTLIST item
      type CDATA #IMPLIED
>
```

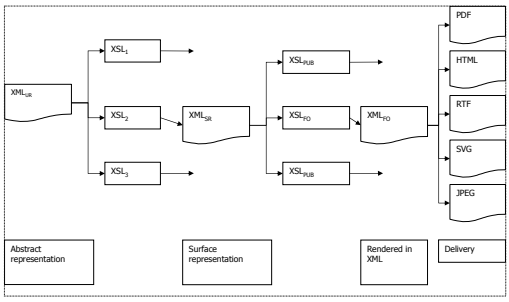

Rendering Issues

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

XSL Implementation



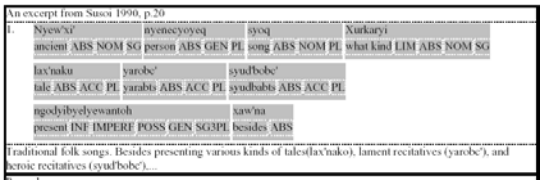
XSL Implementation (cont)



XSL Example

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

Example: Nenets



Example: Nenets

An excerpt from Susoi 1990, p.20

1. Nyew'xi' nyenecyoyeq
ancient ABS NOM SG person ABS GEN PL

syoq Xurkaryi
song ABS NOM PL what kind LIM ABS NOM SG

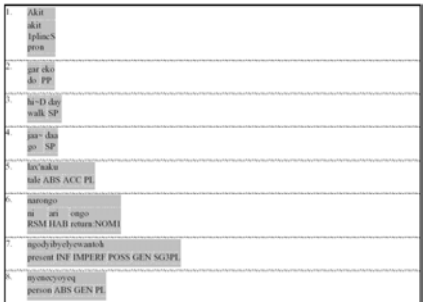
lax'naku yarobe'
tale ABS ACC PL yarabts ABS ACC PL

syud'bobe'
syudbabts ABS ACC PL

ngodyibyelyewantoh xaw'na
present INF IMPERF POSS GEN SG3PL besides ABS

Traditional folk songs. Besides presenting various kinds of tales(lax'nako), lament recitatives (yarobe'), and heroic recitatives (syud'bobe'),...

Example: Document level



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>
```

ComplexWord's
Complex- Word +s
PrefixGloss- RootGloss +CliticGloss

ComplexWord's
Complex- Word +s
PrefixGloss RootGloss CliticGloss

Extension: 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>
  <phrases>
    <phrase>
      <item type="gls">A phrasal translation</item>
      <words>
        ...
      </words>
    <phrase>
  </phrases>
</interlinear-text>
```


Topics for Further Research

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