Talk 8: Encoding Linguistic Analysis

James Cummings

28 January 2014

Encoding Linguistic Analysis

When transcribing, some people are more interested in the linguistic values of texts than their physical or semantic contexts.

Analysis and Linguistics

- associating simple analyses and interpretations with text elements
- semantic or syntactic interpretations which an encoder wishes to attach to all or part of a text
- mainly covering linguistic information
- as often in the TEI, you can do the same thing in many ways:
  - using generic <seg> elements with *@type* attributes
  - using the straightforward *canned* analyses described here
  - using the more powerful and general TEI Feature Structures

Linguistic units

To mark up text for linguistic purposes:

- <s> (s-unit) contains a sentence-like division of a text.
- <cl> (clause) represents a grammatical clause.
- <phr> (phrase) represents a grammatical phrase.
- <w> (word) represents a grammatical (not necessarily orthographic) word.
- <m> (morpheme) represents a grammatical morpheme.
- <c> (character) represents a character.

From the att.segLike class, these elements all have *@type* and *@function* attributes

Example of linguistic markup

Compare

```
<u>Like a suck of one of my sweets?</u>
<u>No I don't take sweets from strangers, oh God</u>
```

with....

linguistic markup

```
<u who="PS1K5">
 <s n="5963">
  <w type="AV0">Like</w>
  <w type="AT0">a</w>
  <w type="NN1">suck</w>
  <w type="PRF">of</w>
  <w type="CRD">one</w>
  <w type="PRF">of</w>
  <w type="DPS">my</w>
  <w type="NN2">sweets</w> ?</s>
</u>
<u trans="smooth" who="PS1BY">
 <s n="5964">
  <w type="ITJ">No </w>
  <w type="PNP">I </w>
  <w type="VDB">do</w>
  <w type="XX0">n't </w>
  <w type="VVI">take </w>
  <w type="NN2">sweets </w>
  <w type="PRP">from </w>
  <w type="NN2">strangers</w>
  <c type="PUN">, </c>
  <w type="ITJ">oh </w>
  <w type="NP0">God</w>
 </s>
</u>
```

(from British National Corpus, KSV 5963)

Mixing analysis with structure

Analytic units often cross structural boundaries. The <cl> (clause) elements here cross the verse lines (<l>). We can use the *@part* attribute to show how a <cl> can be assembled:

```
<div type="stanza">
  <l>
    <cl part="I">Tweedledum and Tweedledee</cl>
  </l>
  <l>
    <cl part="F">Agreed to have a battle;</cl>
  </l>
  <l>
    <cl part="I">For Tweedledum said Tweedledee</cl>
  </l>
  <l>
    <cl part="F">Had spoiled his nice new rattle.</cl>
  </l>
</div>
```

Or the *@next* attribute

```
<l>
  <cl next="#c5" xml:id="c3" part="I">For Tweedledum said
  <cl next="#c6" xml:id="c4" part="I">Tweedledee</cl>
  </cl>
</l>
<l>
  <cl prev="#c3" xml:id="c5" part="F">
    <cl prev="#c4" xml:id="c6" part="F">Had spoiled his nice new rattle.</cl>
  </cl>
</l>
```

Stand-off interpretation

When inline markup is inappropriate, the element can be used to make *ad hoc* remarks about bits of text, linked to by ID. As usual, <spanGrp> is available to group assertions together.

```
<sp>
  <speaker>CORNWALL</speaker>
  <ab  xml:id="eye_start">Lest it see more, prevent it.  Out, vile jelly!</ab>
  <ab>Where is thy lustre now?</ab>
</sp>
<sp>
  <speaker>GLOUCESTER</speaker>
  <ab>All dark and comfortless.  Where's my son Edmund?</ab>
  <ab>Edmund, enkindle all the sparks of nature,</ab>
  <ab  xml:id="eye_end">To quit this horrid act.</ab>
</sp>
<span  from="#eye_start"  to="#eye_end">the eye is pulled out</span>
```

Stand-off interpretation with <interp>

The <interp> element is used to encode an interpretation. The global *@ana* attribution can point from the text to such an interpretation:

```
<ab n="2Sam_12:14">
  <gap/>by this deed thou hast given great occasion to the enemies of the LORD
to blaspheme, the child also that is born unto thee shall surely die.
</ab>
<ab n="2Sam_12:15">
  <gap/>And the LORD struck the child that Uriah's wife bare unto David<gap/>
</ab>
<gap/>
<ab n="2Sam_12:18" ana="#infanticide">And it came to pass on the seventh day,
that the child died.</ab>
<!-- elsewhere in document -->
<interp resp="#SAB" xml:id="infanticide">Infanticide: God seems to like
killing children.</interp>
```

The <interpGrp> element is used to group interpretations together.

Interpretation example (1)

In this example:

- A set of possible interpretations is defined, using <interp> elements
- <seg> is used to markup distinct portions of a narrative
- <s> is used to mark sentences
- the *@ana* attribute links sections or milestones to appropriate interpretation

```
<interpGrp resp="#TMA" type="structuralUnit">
  <interp xml:id="INTRO">introduction</interp>
  <interp xml:id="CONFLICT">conflict</interp>
  <interp xml:id="CLIMAX">climax</interp>
  <interp xml:id="REVENGE">revenge</interp>
  <interp xml:id="RECONCIL">reconciliation</interp>
  <interp xml:id="AFTERM">aftermath</interp>
</interpGrp>
```

Interpretation example (2)

```
<p xml:id="PP1">
  <seg xml:id="SS1-SS3" ana="#INTRO">
    <s xml:id="SS1">Sigmund ... was a king in Frankish country.</s>
    <s xml:id="SS2">Sinfiotli was the eldest of his sons.</s>
    <s xml:id="SS3">Borghild, Sigmund's wife, had a brother ... </s>
  </seg>
  <s xml:id="SS4A" ana="#CONFLICT">But Sinfiotli ... wooed the same woman</s>
  <s xml:id="SS4B" ana="#I3">and Sinfiotli killed him over it.</s>
  <seg xml:id="SS5-SS17" ana="#CLIMAX">
    <s xml:id="SS5">And when he came home, ... she was obliged to accept
it.</s>
    <s xml:id="SS6">At the funeral feast Borghild was serving beer.</s>
    <s xml:id="SS17">Sinfiotli drank it off and at once fell dead.</s>
  </seg>
</p>
<anchor xml:id="NIL1" ana="#RECONCIL"/>
<p xml:id="PP2">Sigmund carried him a long way in his arms ... </p>
```

Phrase segmentation

```
<s>
  <cl type="finite-declarative" function="independent">
    <phr type="NP" function="subject">It</phr>
    <phr type="VP" function="predicate">
      <phr type="V" function="verb-main">was</phr>
        also
    <phr type="NP" function="predicate-nom.">a crucial year for me</phr>
    </phr>
  </cl>
</s>
```

Words with lemmas and morphemes with types

```
<s xml:lang="la">
  <w lemma="timeo">timeo</w>
  <w lemma="danaii">Danaos</w>
  <w lemma="et">et</w>
  <w lemma="donum">dona</w>
  <w lemma="fero">ferentes</w>
</s>
```

or

```
<w type="adjective">
  <m type="prefix" baseForm="con">com</m>
  <m type="root">fort</m>
  <m type="suffix">able</m>
</w>
```

Nested <w>

```
<s>
  <w>I</w>
  <w>
    <w>did</w>
    <m>n't</m>
  </w>
  <w>do</w>
  <w>it</w>
  <c>.</c>
</s>
```

Word analysis

```
<s>
  <w ana="#AT0">The</w>
  <w ana="#NN1">victim</w>
  <w ana="#POS">'s</w>
  <w ana="#NN2">friends</w>
  <w ana="#VVD">told</w>
  <w ana="#NN2">police</w>
  <w ana="#CJT">that</w>
  <w ana="#NP0">Kruger</w>
  <w ana="#VVD">drove</w>
  <w ana="#PRP">into</w>
  <w ana="#AT0">the</w>
  <w ana="#NN1">quarry</w>
  <w ana="#CJC">and</w>
  <w ana="#AV0">never</w>
  <w ana="#VVD">surfaced</w>
</s>
```

Interpretation

```
<interpGrp type="POS">
  <interp xml:id="AT0">Definite article</interp>
  <interp xml:id="AV0">Adverb</interp>
  <interp xml:id="CJC">Conjunction</interp>
  <interp xml:id="CJT">Relative that</interp>
  <interp xml:id="NN1">Noun singular</interp>
  <interp xml:id="NN2">Noun plural</interp>
  <interp xml:id="NP0">Proper noun</interp>
  <interp xml:id="POS">Genitive marker</interp>
  <interp xml:id="PRP">Preposition</interp>
  <interp xml:id="VVD">Verb past tense</interp>
</interpGrp>
```

More interpretation

```
<u xml:id="u1">Can I have ten oranges and a kilo of bananas please?</u>
<u xml:id="u2">Yes, anything else?</u>
<u xml:id="u3">No thanks.</u>
<u xml:id="u4">That'll be dollar forty.</u>
<u xml:id="u5">Two dollars</u>
<u xml:id="u6">Sixty, eighty, two dollars. Thank you.</u>
<spanGrp type="transactions">
  <span from="#u1">sale request</span>
  <span from="#u2" to="#u3">sale compliance</span>
  <span from="#u4">sale</span>
  <span from="#u5">purchase</span>
  <span from="#u6">purchase closure</span>
</spanGrp>
```

British National Corpus

- a snapshot of British English, taken at the end of the 20th century
- 100 million words in approx 4000 different text samples, both spoken (10%) and written (90%)
- synchronic (1990-4), sampled, general purpose corpus
- available under licence; latest edition is BNC-XML (13 March 2007)
- Part-of-speech and lemma tagging
- Uses a variant of TEI XML originally called CDIF

BNC XML

```
<div level="1" n="1" type="leaflet">
 <head type="MAIN">
  <s n="1">
   <w c5="NN1" hw="factsheet" pos="SUBST">FACTSHEET</w>
   <w c5="DTQ" hw="what" pos="PRON">WHAT</w>
   <w c5="VBZ" hw="be" pos="VERB">IS</w>
   <w c5="NN1" hw="aids" pos="SUBST">AIDS</w>
   <c c5="PUN">?</c>
  </s> </head>
 <p>
  <s n="2">
   <hi rend="bo"> <w c5="NN1" hw="aids" pos="SUBST">AIDS</w>
    <c c5="PUL">(</c>
    <w c5="VVN-AJ0" hw="acquire" pos="VERB">Acquired</w>
    <w c5="AJ0" hw="immune" pos="ADJ">Immune</w>
    <w c5="NN1" hw="deficiency" pos="SUBST">Deficiency</w>
    <w c5="NN1" hw="syndrome" pos="SUBST">Syndrome</w>
    <c c5="PUR">)</c>
   </hi>
   <w c5="VBZ" hw="be" pos="VERB">is</w>
   <w c5="AT0" hw="a" pos="ART">a</w>
   <w c5="NN1" hw="condition" pos="SUBST">condition</w>
<!-- ... -->
  </s>
 </p>
</div>
```

Next

Any Questions? Next, the timetable says we're going to .