Advanced Digital Scholarly Editing

Analytic and Interpretive Encoding

Christopher Ohge

christopher.ohge@sas.ac.uk

ATNU ANIMATING TEXT NEWCASTLE UNIVERSITY
IES INSTITUTE OF ENGLISH STUDIES | SCHOOL OF ADVANCED STUDY UNIVERSITY OF LONDON
Newcastle University

This session

- Purposes of analytic encoding
- Basic analytical elements and attributes
- Creating a linking infrastructure for organising your analyses
- Putting it all together
- Stand-off approach

Purposes of analytic encoding

- Encoding your **semantic** or **syntactic** interpretations of the edited text
- A means for structured “annotating” of a non-edited text for literary critical analysis
- Creating the foundation for linguistic analyses (i.e., the basis for further text analysis)

Linguistic segmentation

- Generic use of <seg> applies to any text segment, but in analysis module there are more precise linguistic segments:
    - **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.
- These elements also allow for a seg-associated @function attribute, as well as a global @type and @subtype attributes.
- These elements can appear anywhere in the text transcription.

Linguistic segmentation

- <w> and <pc> (for punctuation) can also have attributes for linguistic annotation:
  - **@lemma** provides a lemma (base form) for the word, typically uninflected and serving both as an identifier (e.g. in dictionary contexts, as a headword), and as a basis for potential inflections.
  - **@lemmaRef** provides a pointer to a definition of the lemma for the word, for example in an online lexicon.
  - **@pos** (part of speech) indicates the part of speech assigned to a token (i.e. information on whether it is a noun, adjective, or verb), usually according to some official reference vocabulary (e.g. for German: STTS, for English: CLAWS, for Polish: NKJP, etc.).
  - **@msd** (morphosyntactic description) supplies morphosyntactic information for a token, usually according to some official reference vocabulary (e.g. for German: STTS-large tagset; for a feature description system designed as (pragmatically) universal, see Universal Features).
  - **@join** when present, it provides information on whether the token in question is adjacent to another, and if so, on which side. The definition of this attribute is adapted from ISO MAF (Morpho-syntactic Annotation Framework), ISO 24611:2012.

Linguistic segmentation, below the word level

- **m** (morpheme) represents a grammatical morpheme.
- **c** (character) represents a character.
- **pc** (punctuation character) contains a character or string of characters regarded as constituting a single punctuation mark.
- g (glyph) is for any non-standard character; it is like <pc> but can be nested within it for more precision.

```
<w type="verb" lemma="hit"
lemmaRef="http://www.example.com/lexicon/hitvb.xml">hitt
<m type="suffix">ing</m>
</w>
```

Linguistic annotation

To talk of many things....

IT hath been said of old that Playes are feasts ....?

```
<phr type="verb"

 function="extraposted_modifier">To talk

<phr type="preposition"

  function="complement">of

 <phr type="noun" function="object">many things</phr>

 </phr>

</phr>


????
```

Linguistic annotation

IT hath been said of old, that Playes are feasts, ...

(Adapted from the Folger Library's Early Modern English Drama version of The Wits: a Comedy by William Davenant.)

```
<l>
<w lemma="it" pos="pn"
xml:id="A19883-003-a-0100">IT</w>
<w lemma="have" pos="vvz"
xml:id="A19883-003-a-0110">hath</w>
 <w lemma="be" pos="vvn"
xml:id="A19883-003-a-0120">been</w>
 <w lemma="say" pos="vvn"
xml:id="A19883-003-a-0130">said</w>
 <w lemma="of" pos="acp-p"
xml:id="A19883-003-a-0140">of</w>
 <w lemma="old" pos="j"
xml:id="A19883-003-a-0150">old</w>
 <pc xml:id="A19883-003-a-0160">,</pc>
<w lemma="that" pos="cs"
xml:id="A19883-003-a-0170">that</w>
 <w lemma="play" pos="vvz"
xml:id="A19883-003-a-0180">
<choice><orig>Playes</orig><reg>Plays</reg>
</choice></w>
<w lemma="be" pos="vvb"
xml:id="A19883-003-a-0190">are</w>
<w lemma="feast" pos="n2"
xml:id="A19883-003-a-0200">Feasts</w> <pc
xml:id="A19883-003-a-0210">,</pc>
</l>
```

Other phrase-level elements: <mentioned>, <gloss>, and <soCalled>

```
<p>Although Chomsky's decision that all NL sentences are finite objects
was never justified by arguments from the attested properties of NLs, it
did have a certain <soCalled>social</soCalled> justification. It was
commonly assumed in works on logic until fairly recently that the notion
<mentioned>language</mentioned> is necessarily restricted to finite
strings.</p>

<nym xml:id="XYZ">
<form>Bogomil</form>
 <etym>Means <gloss>favoured by God</gloss> from the <lang>Slavic</lang>
elements <mentioned xml:lang="ru">bog</mentioned>
  <gloss>God</gloss> and <mentioned xml:lang="ru">mil</mentioned>
  <gloss>favour</gloss>
 </etym>
</nym>
```

Caveat, on <mentioned>: what's the difference between these three options?

```
<mentioned>
<phr>grandiloquent speech</phr>
</mentioned>

<mentioned>
<w>grandiloquent</w> <w>speech</w>
</mentioned>

<w>
<mentioned>grandiloquent</mentioned>
</w>
```

Interpretive elements

- **Span** versus **interpretation** method.
  - and <spanGroup>
  - <interp> and <interpGroup>
- Either way, you are using a structured vocabulary for representing your interpretations.
- These elements contain interpretive attributes @type (for the kind of textual phenomenon, e.g. “image,” “character,” “theme,” “allusion”) and @inst (instances that points to the analysis or interpretation represented by the current element). They can also use global attributes @cert (certainty) or @resp (responsibility for the intervention or interpretation, e.g., an editor or transcriber).

Interpretive elements:

Ch 108 of *Moby-Dick*

CARPENTER (*resuming his work*).

Well, well, well! Stubb knows him best of all, and Stubb always says he's queer; says nothing but that one sufficient little word queer; he's queer, says Stubb; he's queer—queer, queer; and keeps dinning it into Mr. Starbuck all the time—queer, sir—queer, queer, very queer. And here's his leg! Yes, now that I think of it, here's his bedfellow! has a stick of whale's jaw-bone for a wife! And this is his leg; he'll stand on this. What was that now about one leg standing in three places, and all three places standing in one hell—how was that? Oh! I don't wonder he looked so scornful at me! I'm a sort of strange-thoughted sometimes, they say; but

```
<stage>CARPENTER (resuming his
work).</stage>
<sp who="#carpenter">
<p><s xml:id="s1">Well, well,
well!</s>
<s xml:id="s2">Stubb knows him best of
all, and Stubb always says he's queer;
says nothing but that one sufficient
little word queer; he's queer, says
Stubb; he's queer—queer, queer; and
keeps dinning it into Mr. Starbuck all
the time—queer,sir—queer, queer, very
queer.</s>
<span from="#s1" to="#s2">Surprise
(cf. anticipation.</span>
<s xml:id="s3">And here's his leg!</s>
<s xml:id="s4">Yes, now that I think
of it, here's his bedfellow! has a
stick of whale's jaw-bone for a
wife!</s>
<span from="#s3" to="#s4">Ahab's
mutilation; marital metaphor.</span>
… </p>
</sp>
```

Interpretive elements: <interp>

Ch 108 of *Moby-Dick*

CARPENTER (*resuming his work*).

Well, well, well! Stubb knows him best of all, and Stubb always says he's queer; says nothing but that one sufficient little word queer; he's queer, says Stubb; he's queer—queer, queer; and keeps dinning it into Mr. Starbuck all the time—queer, sir—queer, queer, very queer. And here's his leg! Yes, now that I think of it, here's his bedfellow! has a stick of whale's jaw-bone for a wife! And this is his leg; he'll stand on this. What was that now about one leg standing in three places, and all three places standing in one hell—how was that? Oh! I don't wonder he looked so scornful at me! I'm a sort of strange-thoughted sometimes, they say; but

```
<p>Oh! <phr ana="#revenge">I don’t wonder he looked so scornful at me!</phr> <phr ana="#diffidence">I’m a sort of strange-thoughted sometimes, they say; but that’s only haphazard-like. Then, a short, little old body like me, should never undertake to wade out into deep waters with tall</phr>, <phr ana="#metaphor-animal">heron-built captains; the water chucks you under the chin pretty quick, and there’s a great cry for life-boats. And here’s the heron’s leg! long and slim, sure enough!</phr></p>
```

Organising your analyses

- In the previous slide, we added @ana attributes to the phrase elements containing our interpretations.
- The @ana specifically indicates a phrase (or other linguistic unit) that has an interpretive element, but of course we need to connect that @ana attribute value to an @xml:id.
- There are multiple ways to do this (in-text or stand-off), but the most efficient way is to create a separate section (in this case, in the <back>) for storing our <interp>s.

Organising your analyses

Within <text>

```
and all three places standing in one hell—how was that? Oh! <phr ana="#revenge">I
   don't wonder he looked so scornful at me!</phr>
<phr ana="#diffidence">I'm a sort of strange-thoughted sometimes, they say; but
   that's only haphazard-like. Then, a short, little old body like me, should never
   undertake to wade out into deep waters with tall</phr>, <phr
   ana="#metaphor-animal">heron-built captains; the water chucks you under the chin
   pretty quick, and there's a great cry for life-boats. And here's the heron's leg!
   long and slim, sure enough!</phr> Now, for most folks one pair of legs lasts a
```

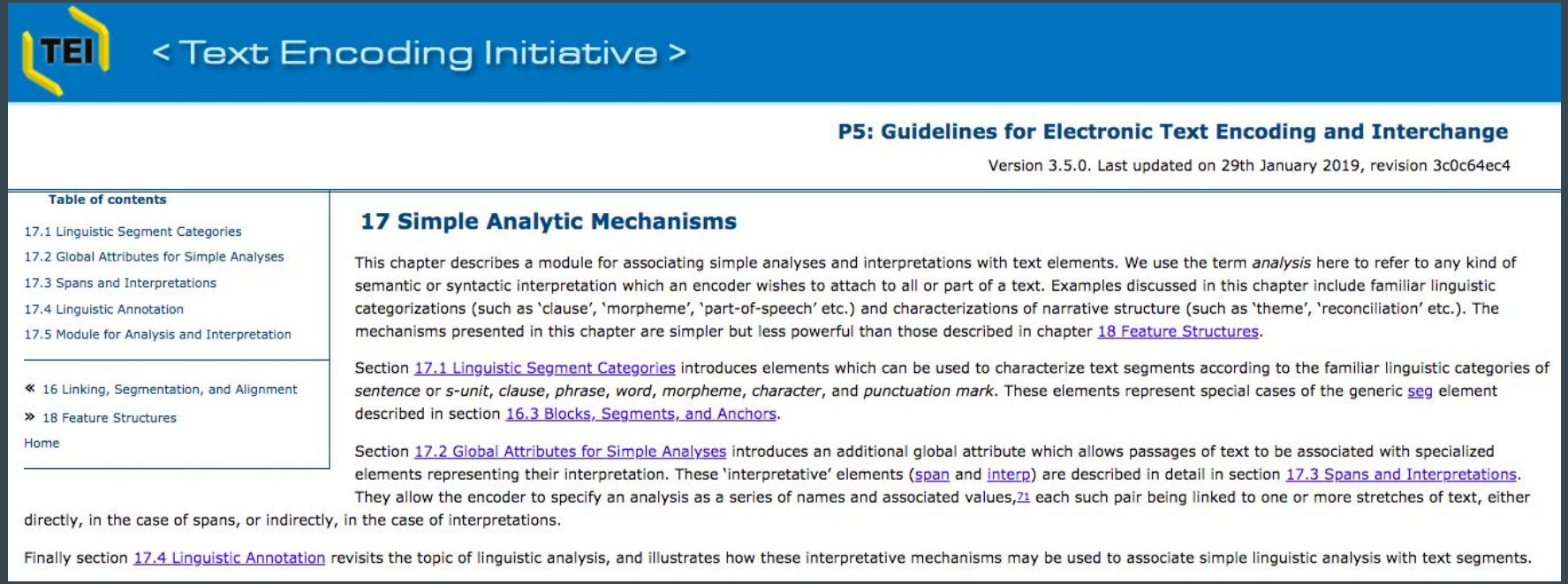
Connected to

<back>

```
<back>
   <div type="interpretive">
      <interpGrp>
         <interp xml:id="revenge">Revenge, mostly dealing with Ahab's rage.</interp>
         <interp xml:id="diffidence">Diffidence, usually being the uncertainty or mistrust of
            oneself in the ship-mates.</interp>
         <interp xml:id="metaphor-animal">Comparisons to animals.</interp>
      </interpGrp>
   </div>
</back>
```

I want to know more!

- Chapter 17 of the TEI guidelines for everything text structure



< Text Encoding Initiative >

P5: Guidelines for Electronic Text Encoding and Interchange

Version 3.5.0. Last updated on 29th January 2019, revision 3c0c64ec4

Table of contents

17 Simple Analytic Mechanisms

This chapter describes a module for associating simple analyses and interpretations with text elements. We use the term *analysis* here to refer to any kind of semantic or syntactic interpretation which an encoder wishes to attach to all or part of a text. Examples discussed in this chapter include familiar linguistic categorizations (such as 'clause', 'morpheme', 'part-of-speech' etc.) and characterizations of narrative structure (such as 'theme', 'reconciliation' etc.). The mechanisms presented in this chapter are simpler but less powerful than those described in chapter 18 Feature Structures.

Section 17.1 Linguistic Segment Categories introduces elements which can be used to characterize text segments according to the familiar linguistic categories of *sentence* or *s-unit*, *clause*, *phrase*, *word*, *morpheme*, *character*, and *punctuation mark*. These elements represent special cases of the generic seg element described in section 16.3 Blocks, Segments, and Anchors.

Section 17.2 Global Attributes for Simple Analyses introduces an additional global attribute which allows passages of text to be associated with specialized elements representing their interpretation. These 'interpretative' elements (span and interp) are described in detail in section 17.3 Spans and Interpretations. They allow the encoder to specify an analysis as a series of names and associated values,[71] each such pair being linked to one or more stretches of text, either directly, in the case of spans, or indirectly, in the case of interpretations.

Finally section 17.4 Linguistic Annotation revisits the topic of linguistic analysis, and illustrates how these interpretative mechanisms may be used to associate simple linguistic analysis with text segments.

Let’s practice!