TEI Texts

The Text Encoding Initiative (TEI - http://www.tei-c.org) is a consortium that develops guidelines for representation of texts in digital form, mainly for libraries, humanities research, social sciences, and linguistics. They have developed an XML format, which you can use to produce texts that can be uploaded into LaBB-CAT. At present, only texts that are *not* synchronised with audio or video are supported. However, the TEI P5 guidelines do specify methods for linking text to media, and support for this will be added to LaBB-CAT in the future.

Tags in the text

The P5 guidelines for TEI specify a dazzling array of tags for capturing all kinds of information about texts, only a subset of which will work well with LaBB-CAT. There is explicit support for the following TEI tags:

- `<p>`, `<div>`, and `<ab>` are interpreted as starting a new line
- `<w>` tags (for marking up words) are used for word tokenization if they are present (if they are absent, LaBB-CAT's standard whitespace-based tokenization is used). Attributes of the `<w>` tag like `lemma` or `type` can be mapped to word layers for capturing such tagging in the text.
- the `<choice><orig>…</orig><reg>…</reg></choice>` construction for marking regularization of text is recognized, and the contents of the `<reg>…</reg>` tag can be extracted to the *lexical* layer (for single-word regularization) or a selected 'phrase' layer (if multi-word regularization is used).
- `<foreign>` is recognised as marking sections of the transcript as being in another language, and so its contents are annotated on the *language* layer, using value of the the `xml:lang` attribute as the annotation label.
- `<note>` is recognised as a commentary marker, and so its contents are put on to the *comment* layer instead of being inserted into the transcript text.
- `<unclear>` tags can create annotations on a selected layer, and the `reason` and `cert` attributes recognised and used in the resulting annotation label if present.

Other tags are by default mapped on to the *entities* LaBB-CAT layer (in which case their tag name, and its type attribute if present, will be used for the entity label).

Alternatively, if there is a LaBB-CAT layer that is named after the TEI tag name, then tags will be mapped to that layer by default during upload. For example, to have all `<sic>` tags extracted to their own layer (instead of the *entities* layer) by default, create a new 'phrase' layer called *sic*.

The TEI header and Meta-data

LaBB-CAT recognises and imports certain constructions in the `<teiHeader>` section of a TEI file:

- in the `<fileDesc>` subsection:
  - the text in `<titleStmt><title>`... is taken to be the value of the *title* transcript attribute
  - the text in `<titleStmt><respStmt><name>`... is taken to be the value of the *scribe* transcript attribute (i.e. the name of the transcriber)
  - the text in `<publicationStmt><distributor>`... is taken to be the value of the *distributor* transcript attribute
  - the text in `<publicationStmt><publisher>`... is taken to be the value of the *publisher* transcript attribute
  - the text in `<publicationStmt><availability><p>`... is taken to be the value of the *availability* transcript attribute
  - the text in `<publicationStmt><date>`... is taken to be the value of the *airDate* transcript attribute
  - the text in `<publicationStmt><distributor>`... is taken to be the value of the *distributor* transcript attribute
  - the text in `<sourceDesc><bibleStruct><monogr><author>`... is taken to be the name of the author of the text (who is created as the sole 'participant' of the transcript)
- in the `<profileDesc>` subsection:
  - the text in `<creation><date>`... is taken to be the value of the *creationDate* transcript attribute
  - the value of the `<langUsage><language ident="...">` attribute is taken to be the value of the *creationDate* transcript attribute
  - the `<particDesc><person>` tags are taken to be participants, whose `<idno>` tag specifies the participant's identifier, and whose other tags specify the participant's attributes named after the tag name (or optionally are added as transcript attributes). The content of the `<person>` tag's `<age>` tag is converted to a single number (in years) if the text is formatted as `y;m.d` or as `y years m months d days`
- in the `<revisionDesc>` subsection:
  - the text in `<change><date>`... is taken to be the value of the *versionDate* transcript attribute
  - the text in `<change><respStmt><name>`... is taken to be the value of the *scribe* transcript attribute

Arbitrary transcript/document attributes are implemented by including a <notesStmt> within the <fileDesc> header, containing one <note> tag per attribute, and using the type attribute as the attribute key and the tag content as the value - e.g.

```
<notesStmt>
 <note type="subreddit">StrangerThings</note>
 <note type="parent_id">t1_dd5f8en</note>
</notesStmt>
```

Participant/speaker attributes can be included in the <person> tag, as per the TEI specification.

Arbitrary participant/speaker attributes (i.e. custom attributes or others not foreseen by the TEI specification) can be processed by including one <note> tag per attribute within each participant's <person> tag, and using the type attribute as the attribute key and the tag content as the value - e.g.

```
<person>
 <idno>ABCD</idno>
 <age>46</age>
 <education>Secondary</education>
 <note type="first language">English</note>
 <note type="origin">Liverpool</note>
</person>
```

Special support for regularization is used; for a construction like this:

```
<choice><orig>color</orig><reg>colour</reg></choice>
```

This deserializer supports part of the schema for Representation of Computer-mediated Communication proposed by Michael Beißwenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer (2012), with the exception of the following:

- When <posting> tags are sychronised to a <when> tag inside a <timeline>, the time synchronisation is ignored.
- The <addressingTerm>, <addressMarker> and <addressee> tags supported, and mapped to "entities" layer by default, but the who attribute of <addressee> is ignored.
- The type attribute of the tag is ignored.
- The revisedWhen, revisedBy, and indentLevel attributes of the <posting> tag are ignored
- The <interactionTerm> tag is ignored.
- The <interactionTemplate>, <interactionWord>, and <emoticon> tags are not explicitly supported.
- The <autoSignature> and <signatureContent> tags are not explicitly supported.