TEI By Example

TEI by Example. Module 1

Edward Vanhoutte
Ron Van den Branden
Melissa Terras

Centre for Scholarly Editing and Document Studies (CTB), Royal Academy of Dutch Language and Literature, Belgium, Gent, 9 July 2010

Last modified on: 2018-04-28
Licensed under a Creative Commons Attribution ShareAlike 3.0 License
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Textual Phenomena</td>
<td>1</td>
</tr>
<tr>
<td>3.1 Structural Features</td>
<td>1</td>
</tr>
<tr>
<td>3.1.1 General</td>
<td>1</td>
</tr>
<tr>
<td>3.1.2 Title Pages</td>
<td>2</td>
</tr>
<tr>
<td>3.2 Renditional Features</td>
<td>3</td>
</tr>
<tr>
<td>3.3 Logical and Semantic Features</td>
<td>4</td>
</tr>
<tr>
<td>3.3.1 Referring Strings</td>
<td>4</td>
</tr>
<tr>
<td>3.3.2 Dates and Time</td>
<td>5</td>
</tr>
<tr>
<td>3.3.3 Numbers and Measures</td>
<td>6</td>
</tr>
<tr>
<td>3.3.4 Addresses</td>
<td>7</td>
</tr>
<tr>
<td>3.3.5 Abbreviations and Expansions</td>
<td>8</td>
</tr>
<tr>
<td>3.4 Analytical Features</td>
<td>9</td>
</tr>
<tr>
<td>3.4.1 Notes and Annotations</td>
<td>9</td>
</tr>
<tr>
<td>3.4.2 Index Entries</td>
<td>9</td>
</tr>
<tr>
<td>3.4.3 Apparent Errors</td>
<td>10</td>
</tr>
<tr>
<td>3.4.4 Regularisation and Normalisation</td>
<td>11</td>
</tr>
<tr>
<td>3.4.5 Additions, Deletions, and Omissions</td>
<td>12</td>
</tr>
</tbody>
</table>
3. Textual Phenomena

The TEI Guidelines define a set of rules to mark up the phenomena in a wide range of texts and textual objects in a descriptive fashion. Generally speaking, there are four classes of textual phenomena that can be described:

1. Structural
2. Renditional
3. Logical & Semantic
4. Analytic

Structural and renditional features are best understood because they concern a natural kind of textual, though culturally defined, organisation. Books mainly consist of chapters, sections, and paragraphs; poetry is mostly organised in poems, stanzas, and lines; whereas scenes, acts, and parts of speech are structural features of performance texts. In these texts, linguistic units are highlighted by the use of distinct fonts, colours, alignments, italics, underlinings, font weight etc. These textual codes signal underlying logical and semantic features and functions such as names of organisations, titles of books, distinctive languages, emphatic language use, etc. However, semantic and logical features don't need to be highlighted by means of typographic codes and can occur in texts unsuspiciously. It needs a thorough understanding of the text and the language to identify them. Semantic and syntactic interpretations added to a text or part of a text that together constitute a new text, we call analytical features. Examples are linguistic (wordclass, morpheme,...) and narrative (theme, motive,...) categorisations.

3.1 Structural Features

3.1.1 General

Challenge

Which structural features can commonly be found in prose, verse, and drama?

- Prose: paragraphs `<p>`, divisions `<div>`, headings `<head>`, lists `<list>`, listitem `<item>`, quotations `<q>`, page breaks `<pb>`, segments `<seg>`, figures `<figure>`, and tables `<table>`. See TBE Module 3: Prose.
- Verse: linegroups `<lg>` and lines `<l>`. TBE Module 4: Verse.
- Drama: divisions `<div>`, speeches `<sp>`, paragraphs `<p>`, linegroups `<lg>`, lines `<l>`, and segments `<seg>`. TBE Module 5: Drama.

The following example demonstrates a simple use of markup for the encoding of structural features in prose text:
For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren't even going to notice his absence.

Or worse, what would happen when another Academy member had decided to go for a stroll in the park instead? He quickly thought up several possible plans:

- hide behind a tree and duck
- catch the duck as subject material for a speech on the annual meeting
- be frank, meet his colleague, and:
  - 1. pat him on the shoulder
  - 2. tell a joke
  - 3. hand him the duck
  - 4. offer him a sip from his 2.5 l bottle of coke
  - 5. pull his beard

Or maybe he could still announce his absence from the meeting by sending an antedated letter of apology to Professor M. Orkelidius, Royal Academy of Whoopledywhaa, Queenstreet 81, TB90 00E Whoopledywhaa.

Plenty of options, he thought, sat on a bench and opened the book he had taken from the Whoopliedywhaaian National Library. It was titled 'While thou art here', by Sir Edmund Peckwood. While reading the first sentence, his placid expression turned to a certain je ne sais quoi:

For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa.

---

3.1.2 Title Pages

Title pages may be encoded within `<front>` or `<back>` by using the element `<titlePage>`. A title page commonly contains the title of the work `<docTitle>` which can consist of several subsections or divisions `<titlePart>` with an `<@type>` attribute documenting their role. The name of the author of the document `<docAuthor>` often occurs inside a byline `<byline>` which contains the primary statement of responsibility given for a work. Other components of `<titlePage>` may be the edition statement `<docEdition>`, the date of a document
<docDate>, and the imprint statement <docImprint> which may further contain the place of publication <pubPlace>, a date <date> or <docDate>, and names <name> of, e.g. the publisher <publisher>. Besides this information, a <titlePage> may also contain an anonymous or attributed quotation <epigraph>, a formal statement authorizing the publication of a work <imprimatur>, and/or an inline graphic, illustration, or figure <graphic>.

3.2 Renditional Features

Some textual features are commonly rendered in a text using some kind of highlighting. The TEI Guidelines define highlighting as 'the use of any combination of typographic features (font, size, hue, etc.) in a printed or written text in order to distinguish some passage of a text from its surroundings.' If the encoder prefers only to signal this highlighting, and not the underlying reason, the generic element <hi> can be used with a @rend or @rendition attribute describing its appearance in the text. There are no formally defined values for these attributes which may need to express a very large range of typographic features. Encoders, however, commonly prefer to indicate the reason underlying the highlighting by documenting logical or semantic information about the highlighted word or phrase.

......

1 <titlePage> must not be confused with <fileDesc> which may contain <titleStmt> and <publicationStmt>. Whereas <titlePage> is used for the transcription and encoding of the physical title page in <text>, <fileDesc> provides a bibliographic description of the electronic file in <teiHeader>.
For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren’t even going to notice his absence.

3.3 Logical and Semantic Features

Highlighted words or phrases in a text are commonly distinguished from their surroundings for a reason. Only a thorough understanding of the text and the language can lead to a correct identification and interpretation. The underlying semantics may be encoded with far more specific elements than the generic <hi>. Highlighting is commonly used to render the following logical and semantic features:

- Emphasis <emph>, foreign words <foreign> and other linguistically distinct uses of highlighting.
- The use of quotation marks in the representation of speech and thought <said>, quotation <quote>, cited quotation <cit>, words or phrases mentioned <mentioned> and words or phrases for which the author or narrator indicates a disclaiming of responsibility <soCalled>. See TBE Module 3: Prose -- Quotation.
- Technical terms <term>, glosses <gloss> or documentation of XML elements, attributes and classes with <altIdent>, <desc>, <equiv/> See TBE Module 3: Prose -- Quotation, and TBE Module 8: Customising TEI, ODD, Roma.

However, words or phrases carrying semantic and logical information don’t need to be highlighted by means of typographic codes and can occur in texts unsuspiciously. Think about titles <title>, names <name>, numbers <num>, measures <measure>, dates <date>, addresses <address>, abbreviations <abbr> and expansions <expan>.

3.3.1 Referring Strings

Proper nouns name people, places, and objects and are easily traceable in a text. This may be encoded with <name> carrying a @type attribute specifying the kind of object referred to.
'Plenty of options', he thought, sat on a bench and opened the book he had taken from the Whoopledywhaaian National Library. It was titled 'While thou art here', by Sir Edmund Peckwood. While reading the first sentence, his placid expression turned to a certain je ne sais quoi: 'For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa.'</p>

However, people, places, and objects may also be referred to with common nouns, for which the element <rs> may be used. This element may also carry a @type attribute specifying the kind of object referred to.

<p>'Plenty of options', <rs type="person">he</rs> thought, sat on a bench and opened the book <rs type="person">he</rs> had taken from the <rs type="organisation">Whoopledywhaaian National Library</rs>. It was titled 'While thou art here', by Sir Edmund Peckwood. While reading the first sentence, <rs type="person">his</rs> placid expression turned to a certain je ne sais quoi: 'For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the <name type="organisation">Royal Academy of Whoopledywhaa</name>.'</p>

### 3.3.2 Dates and Time

Any expression defining a date or time may be encoded with the corresponding elements <date> and <time>. The system or calendar to which the date belongs may be documented using a @calendar attribute. The @when attribute supplies the value of a date or time in a standard form, which is useful for text processing.

The normalised representation of the content of the <date> element should conform to a valid W3C schema datatype for expressing temporal data:

- <date when="2009" calendar="Gregorian">2009</date>
- <date when="2009-12">December 2009</date>
- <date when="2009-12-31">31 Dec 2009</date>
- <date when="2009-12-31">New Year's Eve 2009</date>
- <date when="2009-12-31" calendar="Persian">Panjshanbeh 10 Dey 1388</date>
- <date when="--12-31">last day of December</date>
- <date when="--12">December</date>

......

$rs$ may be used for any reference to a person, place or object in the form of a proper noun, a noun phrase or a common noun. <name> may be used synonymously with the $rs$ element in the special cases of referencing strings which consist only of proper nouns. The choice between $rs$ or <name> in these cases is the encoder's. <name> may also nest inside $rs$ where a proper name is part of a larger referring string.
• `<date when="--31">thirty-first of the month</date>`

The same counts for the normalized representation of the content of `<time>`:

• `<time when="23:55:00">11:55 pm</time>`
• `<time when="23:55:00">five minutes before midnight</time>`
• `<time when="2009-12-31T23:55:00">five minutes before the new year 2010</time>`

The `<date>` element can also be used to mark a period of time using `@from` and `@to` attributes or `@notBefore` and `@notAfter` attributes.

```
For the first time in `<date from="1935" to="1960">twenty-five years</date>`, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It was a sunny day in `<date notBefore="1960-09-15" notAfter="1960-09-30">late September 1960</date>` bang on `<time when="12:00">noontime</time>` and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren't even going to notice his absence.<p>
```

### 3.3.3 Numbers and Measures

Numbers and measures may be encoded using `<num>` and `<measure>` respectively.

`<num>` may contain numbers, written in any form and uses the attribute `@type` to indicate the type of numeric value and `@value` to supply the value of the number in standard form.

```
For the first time in `<num type="cardinal" value="25">twenty-five</num>` years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa.<p>
```

Further examples of the standardisation of the representation of numbers are:

• `<num value="25">xxv</num>`
• `<num type="percentage" value="25">twenty-five percent</num>`
• `<num type="percentage" value="25">25%</num>`
• `<num type="ordinal" value="25">25th</num>`

In their fullest form, a measure consists of a number, a phrase expressing units of measure, and a phrase expressing the commodity being measured, though not all of these components need to be present in every case. These three components may be encoded by using the attributes `@quantity`, `@unit`, and `@commodity` with `<measure>`.

```
In their fullest form, `<measure quantity="25">twenty-five</measure>` units of commodity `<measure unit="25">xxv</measure>` are needed.<p>
```

......

3 The last example also includes a date string and can equally well be tagged as `<date when="2009-12-31T23:55:00">`. 

---

3.3.3 Numbers and Measures
Or worse, what would happen when another Academy member had decided to go for a stroll in the park instead? He quickly thought up several possible plans:

- hide behind a tree and duck
- catch the duck as subject material for a speech on the annual meeting
- be frank, meet his colleague, and
  - pat him on the shoulder
  - tell a joke
  - hand him the duck
- offer him a sip from his 2.5 litre bottle of coke
- pull his beard

Or maybe he could still announce his absence from the meeting by sending an antedated letter of apology to

- Professor M. Orkelidius
- Royal Academy of Whoopledywhaa
- Queenstreet 81
- TB90 00E Whoopledywhaa

An alternative method of encoding can be applied using some more semantically rich elements such as <street>, <postCode> and <postBox>. Names of people, organisations, companies etc. may be encoded using <name> with a @type attribute indicating the type of object which is being named by the element content.
Or maybe he could still announce his absence from the meeting by sending an antedated letter of apology to Professor M. Orkelidius, Royal Academy of Whoopledywhaa, Queenstreet 81, TB90 00E, Whoopledywhaa.

### 3.3.5 Abbreviations and Expansions

It is sometimes useful to encode abbreviations and their expansions in texts. This facilitates special processing, regularisation by the full form of an abbreviation, or the rendering of different possible expansions of an abbreviation. Abbreviations may be marked using `<abbr>`. The `@type` attribute may be used to distinguish types of abbreviations by their function:

For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren't even going to notice his absence.

Alternatively, and depending on the encoder's preference, the expansion of an abbreviation may be encoded with `<expan>`. This is often done when the editor or encoder of a text has silently expanded the abbreviation for whatever reason. This will commonly be combined with the `<abbr>` element inside a `<choice>` element to record the relationship between the abbreviation and its expansion:

For the first time in twenty-five years, Dr Doctor Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW).
3.4 Analytical Features

The analysis of texts can generate information which may be added to the text and encoded as metadata or as part of the text. Explicit notes are the most common example of the latter while editorial statements like the correction of errors, the regularisation of or the marking of the text for indexing purposes are examples of the former. The creation of index entries also enhances further analysis of the text.

3.4.1 Notes and Annotations

The most explicit form of textual annotation is the addition of notes to the text using `<note>`.

All notes should be marked using the same tag `<note>`, whether they are already present in the text or supplied by the editor, whether they appear as block notes in the main text area, at the foot of the page, at the end of the chapter or volume, in the margin, or in some other place. The `@type` attribute distinguishes the different types of annotations in use in a text. In a `@resp` attribute, the responsible subject for a note can be documented. Where possible, a note can be inserted in the text at the point at which its identifier or mark first appears. The location of the note may be documented using a `@place` attribute.

```xml
<p>'Plenty of options', he thought, sat on a bench and opened the book he had taken from the Whoopledywhaian National Library <note n="1" place="foot" type="authorial">The National Library of Whoopledywhaa was found in 1886 with the acquisition of the library of the late King Anthony.</note>. It was titled 'While thou art here', by Sir Edmund Peckwood <note type="editorial" resp="EV">The manuscript reads 'Petwood'. </note>. While reading the first sentence, his placid expression turned to a certain je ne sais quoi: 'For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywha.'</p>
```

3.4.2 Index Entries

Pre-existing indexes may be encoded as a `<list>` inside `<div>` in `<front>` or `<back>`, for example. On the other hand, new indexes can be generated from machine readable text when the location to be indexed is marked with `<index>` with a headword encoded as `<term>`.

```
......
```

4 See section 3.6. Simple Links and Cross-References in the TEI Guidelines for a full discussion of notes which are encoded not at the point of attachment but at the point of appearance, e.g. at the end of a chapter or a volume. See chapter 16. Linking, Segmentation, and Alignment for mechanisms to encode multiple views of larger or heterogeneous spans of text. See section 17.3. Spans and Interpretation for a discussion of advanced interpretive annotations.
'Plenty of options', he thought, sat on a bench and opened the book he had taken from the Whoopledywhaian National Library. It was titled 'While thou art here', by Sir Edmund Peckwood. While reading the first sentence, his placid expression turned to a certain je ne sais quoi: 'For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaaw.

The effect of this will be to generate an index entry for the terms 'Library' and 'Academy', referencing the location of the original element.

### 3.4.3 Apparent Errors

Apparent errors in the text may be indicated using `<sic>` or corrected inside `<corr>`.

```
<p>It was titled 'While thou art here', by Sir Edmund Petwood</p>
```

```
<p>It was titled 'While thou art here', by Sir Edmund Peckwood</p>
```

Alternatively, the encoder may both record the original source text and provide a correction by using both `<sic>` and `<corr>` in either order wrapped in a `<choice>`.

```
<p>It was titled 'While thou art here', by Sir Edmund

  <corr>Peckwood</corr>

  <sic>Petwood</sic>

</choice>
```

The encoder may encode the degree of certainty associated with the intervention or interpretation using a `@cert` attribute and indicate the agency responsible for the intervention or interpretation, for instance an editor or transcriber, using `@resp`. The value of `@resp` is a pointer to an element in the document header that is associated with a person responsible for the intervention.

......

5 See section 3.8.3 Index Entries in the TEI Guidelines for a full discussion of the TEI encoding strategies applied to indexes.
It was titled 'While thou art here', by Sir Edmund
<choice>
<corr cert="high" resp="#EV">Peckwood</corr>
<sic>Petwood</sic>
</choice>
</p>

The attribute value '#EV' points to a <name> element in the <teiHeader>, for example in the <respStmt>.

<respStmt>
<resp>editor</resp>
<name xml:id="EV">Edward Vanhoutte</name>
</respStmt>

6

3.4.4 Regularisation and Normalisation

Standard or regularised forms for variant forms or non-standard spelling may be provided for a number of reasons. This is called regularisation or normalisation. The original, non-normalized form may be flagged using <orig>.

<p>It was titled 'While <orig>thou</orig> <orig>art</orig> here', by Sir Edmund Peckwood</p>

If the encoder wants to indicate that certain words have been normalised, which means modernisation of spelling in this example, <reg> may be used.

<p>It was titled 'While <reg>y</reg>ou <reg>are</reg> here', by Sir Edmund Peckwood</p>

Alternatively the encoder may decide to record both the original form <orig> and the regularised form <reg> wrapped inside a <choice>. In the case of the modernisation of spelling, an electronic text could thus serve as the basis of an old- or new-spelling edition.

......

6 See Section 6.4 Editorial Interventions of TBE Module 6: Primary Sources for a fuller treatment of editorial interventions.
It was titled 'While thou art here', by Sir Edmund Peckwood.

The @resp attribute may be used to specify the agency responsible for the regularisation or normalisation.

3.4.5 Additions, Deletions, and Omissions

Another editorial intervention in the text may be the documentation and creation of additions, deletions and omissions. When transcribing a source document, <gap> may be used to indicate a point where material has been omitted both because the material is illegible, invisible or inaudible in the source and because the editor or transcriber has decided to omit material for editorial reasons or as part of sampling practice. The reason for omission may be given in a @reason attribute. Sample values include sampling, illegible, inaudible, irrelevant, cancelled. Additional attributes like @extent and @unit may document the amount of characters, words, lines or any other unit omitted.

For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). <gap reason="irrelevant" unit="words" extent="32">It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren't even going to notice his absence.</gap>

If the omission is an editorial policy decision, e.g. the systematic exclusion of marginal commentaries from an encoding, the full details of the policy should be documented in <editorialDecl> inside the <encodingDesc> of the TEI Header. See section 3.2. The Encoding Description in TBE Module 2: The TEI Header.
The `<gap>` element may appear as an empty element, but may also contain a description of the material omitted using `<desc>`.

For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). `<gap reason="irrelevant" unit="words" extent="32">Commentary on the founding charter of the RAW</gap>` It was a sunny day in late September 1960 bang on noontime and Dr Burt was looking forward to a stroll in the park instead. He hoped his fellow members of the RAW weren’t even going to notice his absence.

Where words or phrases of moderate lengths have been added or deleted in the copy text, this may be recorded using `<add>` and `<del>`. As with all TEI elements, information on the actual rendition of the additions and deletions can be provided in the global `@rend` attribute. Additionally, the place of the addition may also be recorded using `@place`. See section **3.1.1. Simple additions and deletions** for a detailed discussion of these elements and their attributes.

Additions and deletions with a causal relationship may be grouped by the `<subst>` element.

For the first time in twenty-five years, Dr Burt Diddledygook decided not to turn up to the annual meeting of the Royal Academy of Whoopledywhaa (RAW). It was a sunny day in `<add place="supralinear">late</add>` September 1960 bang on noontime and Dr Burt was looking forward to a `<del rend="overstrike">walk</del>` `<add place="infralinear">stroll</add>` in the park instead. He hoped his fellow members of the RAW weren’t even going to notice his absence.

When an editor wants to mark his or her own additions as editorial interventions in the text `<corr>` or `<supplied>` should be used, not `<add>`. See Section **4. Editorial interventions** in TBE Module 6: Primary Sources.

For longer passages `<addSpan/>` and `<delSpan/>` may be used. See section **3.1.2. Complex additions and Deletions** in TBE Module 6: Primary Sources.

3. Textual Phenomena
Where deletions in the copy text cannot be read with confidence, `<unclear>` should be used with the `@reason` attribute indicating that the difficulty of transcription is due to deletion. See Section 4.1, *Unclear, supplied, omitted text* in *TBE Module 6: Primary Sources*. 