Encoding Guidelines for the ELTeC: documentaton only

Cost Action CA16204 – WG1 *2018-03-16*

This reference document defines the encoding scheme to be used for the European Literary Text Collection (ELTeC) which will be a major deliverable of COST Action 16204, *Distant Reading*. This draft reflects decisions taken at the WG1 meeting in Prague in February 2018, but has not yet been formally reviewed by the Work Group. Particular topics on which policy remains to be defined are signalled below with the label 'Open Question'.

1 Principles

The MoU for the project points out that 'Distant Reading methods cover a wide range of computational methods for literary text analysis, such as authorship attribution, topic modelling, character network analysis, or stylistic analysis.' The focus of the ELTeC encoding scheme is thus not to represent texts in all their original complexity of structure or appearance, but rather to facilitate a richer and better-informed distant reading than a transcription of its lexical content alone would permit. For example, it seems useful to distinguish headings and annotations from the rest of the text, and to be able to locate stretches of text within gross structural features such as pages, chapters, or paragraphs. Although it may be useful to distinguish passages belonging to different narrative levels (for example, direct speech versus narrative or quotation versus narrative), it is difficult to do so automatically with any degree of consistency. It is certainly less useful to record exact nuances of rendition or spelling in a particular version of a text. Our goal is thus not to duplicate the work of scholarly editors or to produce (yet another) digital edition of a specific source document. Rather it is to ensure that the ELTeC texts can be processed by very simple minded (but XML-aware) systems primarily concerned with lexis and to make life easier for the developers of such systems.

In selecting features for inclusion in the markup scheme, we have been guided, but not limited, by existing practice as far as possible. Our main goal has been to identify a small core set of textual features which can be readily (preferably automatically) identified in existing digital transcriptions, or easily and consistently provided by new transcriptions.

We distinguish three 'levels' of encoding, referred to below as level zero, level one and level two. All ELTeC texts are made available at level zero, the basic encoding format. Some texts may additionally be made available at levels one or two, which provide a richer set of encoded features. For example: a level one text will include information about rendition features missing from a level zero text; a level two text will include tokenization information missing from a level one text. As far as possible conversion between levels will be automatically scripted, but this is not possible in the general case.

This document lists all the textual features which are to be distinguished in an ELTeC conformant transcription at one of these three levels. Whenever a given feature exists in a text, it will be marked up as indicated here. No other features will be captured by the markup: if some textual feature not provided for here is identified by a marked up source text, that markup will be removed (though it may be retained in a version of the text encoded at a different level).

All ELTeC documents are TEI conformant, and therefore include a TEI Header, as discussed in section *4. Metadata in the TEI Header* below.

2 Basic Transcription Guidelines (all levels)

The basic unit of the ELTeC corpus is the text of a single novel, represented by a TEI `<text>` element. We propose no mechanism (other than metadata) to encode units larger than a single

novel, such as multipart novel series like Proust's *A la recherche du temps perdu* or Balzac's *Les Rougon-Macquart*.

To facilitate checking of a transcription against its source during production, the `<pb>` element must be provided to mark the point in a transcript where a new page begins. If a page begins with the second part of a hyphenated word, the `<pb>` tag may appear after that, but otherwise its position should be the same in transcription and source. The `<pb>` element has an attribute *@n* which should be used to number the pages. A level 1 text may also provide a *@facs* attribute to point to a page image of the corresponding source page.

As well as a titlepage or a table of contents, a published novel often includes material such as forewords or appendixes in addition to the text of the novel itself. This *liminal* matter is included in an ELTeC text only if it is believed to be authorial. Material before the body of the text begins is collected within a `<front>` element, and material following the body in a `<back>` element. In either case, distinct sections of the material, if encoded, are represented by a `<div>` with its *@type* attribute set to `liminal`.

At level zero, titlepages and tables of contents are omitted. At level one, they are replaced by a `<gap>` element. Non-authorial liminal material is silently omitted at all levels.

The Prague decision list says that we decided to exclude titlepages, tables of contents, errata list etc, but to include prefaces, introductions, afterwords, and appendixes, provided these are contemporary with the text. It also says to include footnotes and commentary, but does not specify whether these should also be contemporary, nor how the encoder can easily determine whether or not something is "contemporary".

Within the body of a text, major structural divisions (parts, sections, chapters etc.) will be captured using the generic `<div>` element, with attributes *@type*, *@xml:lang*, *@xml:id* and *@n* used as further detailed below.

The names used for hierarchic structural divisions of a novel above the chapter are arbitrary, culture-specific, and often inconsistent : in some novels things called 'part' contain things called 'book' and in others the reverse. We propose to follow TEI in using a single element (`<div>`) for every hierarchical structural division, down to the level of 'chapter'.

Open Question Is it useful to retain the name used for each level in the original source (the type of div) ?

- Yes: it is easy to keep and may help referencing : use the *@type* attribute to hold the name used for each level of div in the work in question
- No : this name adds no useful information beyond the level indicated by the XML structure
- No : it would be more useful to provide an explicit and normalised indication of the hierarchic level for the benefit of non-XML-aware processors (e.g. `level1`, `level2` etc.)

This issue was not discussed in Prague. Proposal is to use (and enforce) a predefined list of specific values.

The (human) language in which a text is expressed is indicated explicitly by the *@xml:lang* attribute which supplies the ISO 2 letter code for the language concerned. This attribute will always be supplied on the `<text>` element to specify a default, and may also appear on other elements to indicate passages where the language changes. The various different languages used in a given text will be itemized in its metadata (see `<langUsage>` element in the header).

Open question Should passages exhibiting regional or dialectal variation be specially signalled?

- No : this is too fine grained and controversial a distinction to be made with reliable consistency
- Yes : treat this in the same way as any other kind of code switching and define a set of appropriate language codes for the project

- Maybe : just use the <distinct> element to indicate the kind of variation concerned

In Prague there was some support for using either <foreign> or <distinct>, but no decision to do so. Proposal is not to do so.

A single reference scheme will be defined for the whole corpus, with the following components:

- text identifier : every text will have an identifier consisting of its two letter language code and a three digit serial number, for example FR042
- chapter identifier: each chapter or equivalent will have an identifier concatenating the text identifier and a three digit serial number, for example FR042012 is the twelfth chapter of the 42nd French novel.
- If sub-chapter segmentation (see below) is implemented, then the segments will append a further four digit serial number.

The identifier will be supplied as the value of an *@xml:id* attribute on each <text>, or <s> element as appropriate. Adding this identifier is an easily automated task which can be built into the workflow for accession to the ELTeC.

Note that these identifiers will not necessarily correspond with the numbering used in a particular source text. In a work where the first twelve chapters are considered to form part one, and the next twelve constitute part two, the first chapter of the second part will have an identifier ending 013, even though it may be numbered 1 in a source text.

No dissent from this proposal in Prague

Open question is it important to preserve the original numbering, particularly for deeply structured texts?

- Yes : the original numbering is widely used to reference the text: it should be supplied as using the *@n* attribute on the .
- No : the original numbering and referencing scheme are of no use in our intended applications, introduce unnecessary complexity, and may be a source of confusion.

Not explicitly addressed in Prague. Proposal is not to retain original numbering.

The chapters of a novel mostly consist of prose, arranged in paragraphs, for which we will use the TEI <p> element. It is not unusual to find other structures however, specifically verse, or passages of dialogue presented as if in a play, with speaker labels and even stage directions. Less frequently, novels may contain material presented in list or tabular formats. Graphics with their own associated heading or other text are also frequent.

Open Question how should material other than running prose and dialogue be encoded?

1. Use the appropriate TEI elements for verse or drama (<lg>, <l>, <sp>, <stage>)
2. Use the appropriate TEI elements for lists and tables (<list>, <label>, <item>, <table>, <cell>, <row>)
3. Use the appropriate TEI elements for embedded graphics (<figure>, <graphic>, <head>)
4. Suppress all non-prose material, replacing it by <gap>

In Prague, we decided to suppress annotation of linebreaks, lists, tables, figures, captions of figures, typographic information, pagebreaks, and quotation (i.e. direct/indirect speech). We explicitly agreed to annotate only paragraphs, divisions, and headings. Other features would be represented either by a <gap>, if they have been entirely suppressed, or by a <p> if they have textual content.We also agreed that hyphenation, like other typographic features, would not be preserved. Verse and drama were not explicitly addressed.

Novels are also full of direct speech, represented using various different conventions, but almost always distinguished from the narrative voice. The first person narrative is also common, but may be regarded as a special case. How exactly different narrative strands are articulated in a novel, and the extent to which they may be characterised by their lexis has been a preoccupation of many 'distant reading' style analyses. It might therefore be helpful to distinguish material purporting to be direct speech from material purporting to be narrative in our basic encoding, though to do so consistently and accurately may occasionally be problematic.

Open Question Should passages presented as direct speech in a novel be distinguished from passages presented as narrative?

- Yes : use `<q>` and avoid nesting problems by always nesting it within `<p>`
- Yes : use a `<milestone>` to mark the beginning and end of each passage of direct speech
- Sort of : provide an attribute on `<p>` to indicate whether or not the paragraph contains direct speech
- No : rely on (or normalise) typographic conventions such as quote marks or dashes to distinguish direct speech only.

In Prague the majority view was not to attempt to do more than preserve existing punctuation.

Printed texts typically deploy a number of conventions which can cause problems for linguistic analyses of even the most basic kind. Changes of font or style (italicization or use of superscript, for example) can have particular lexical significance which should be taken into account. End-of-line hyphenation can make it harder to identify the exact form of a token. Non-standard (i.e. non-modern) spellings can mislead parsers. Our proposed encoding aims above all for consistency and transparency in what is reliably achievable, leaving more difficult and problematic issues to be addressed by linguistic annotations.

We do not preserve the lineation of running prose in our source texts, since this is always purely an artefact of the source edition. For the same reason we will reassemble words broken across a line break, silently removing any hyphen present. (This will make it impossible to use our texts for hyphenation studies. So be it.)

Open Question : Should page breaks in the source text be preserved ?

- Yes : this is useful information (e.g. to determine words-per-page, or to anchor links to an image of the source text) which is usually available at no-cost in existing digital texts
- No : the proposed uses don't justify the cost of providing the information if it is missing. And pagination is inherently copy-specific.

Prague decision (as noted above) was to suppress page numbering; however the proposal is to retain it, since it will always be available for OCR texts, where it is essential information during text validation, and is usually available in other digital versions. The discussion in Prague concerned only its lack of utility during the analysis stage, but it is very useful during the transcription and validation stage.

Font and style variations in the source text usually signal something. Italics may signal emphasis, quotation, foreign language terms etc. Superscripts almost always signal abbreviation. The visual salience of these variations is of considerably less interest to distant readers than the intended function they signal. However, it is not always easy to determine that function reliably and consistently by algorithm. Some simple cases could however be addressed. A possibly strategy is outlined below. It assumes the existence of a digital version of the text in which visual features are explicit, whether by means of TEI-style markup or styling information such as that provided by Word.

- if possible, replace indications of highlighting by an appropriate TEI element, chosen from the following list : `<foreign>`, `<title>`, `<emph>`
- otherwise, replace all indications of highlighting by the TEI `<hi>` element
- indications of superscript characters (such as French '14 ') should be removed. Instead, the TEI element `<abbr>` should be used to indicate the presence of an abbreviated word: `<abbr>14e</abbr>`

Prague decision (as noted above) : was to suppress all encoding of renditional features.
Open Question: Is it feasible or useful to recode highlighted spans of text in this way?

- Yes : in many cases this can be an automatic process and the results justify investing the effort
- No : there are likely to be too many borderline or debatable cases to do this automatically so this would have to be done as part of a major proof reading exercise

Whichever solution is adopted, it should be applied uniformly across the ELTeC. A collection in which some texts make distinctions ignored by others is unsatisfactory.

3 TEI Elements used

This section will provide a checklist of TEI elements used in the body of each ELTeC text, with descriptions and examples of their intended applications.

4 Metadata in the TEI Header

This section describes the metadata associated with each text (title, authorship, date etc.) and with the collection as a whole. The intention is to provide this in a standardised way to facilitate subsetting of the collection, using (for example) coded values for the descriptive selection criteria associated with the text. As far as possible, our text should represent the first complete printed edition of each novel selected.

The TEI Header provides a very large number of possibilities for encoding such metadata. We will provide a checklist of the TEI Header elements which are always to be provided for each text, possibly in the form of a template. As in the body of the text, the intention is to provide a guaranteed minimal level of information, consistent across all parts of the ELTeC.

Note that metadata may be supplied at (at least) two levels: the level of the ELTeC as a whole, and that of individual texts within it. Information which applies uniformly to all parts of the collection should be supplied in the ELTeC header; information specific to a particular document in the text header.

5 Text-level metadata

Here is an example template for an individual text header

```
<teiHeader type="novelHeader">
 <fileDesc>
  <titleStmt>
   <title>
<!-- standard title of work -->
   </title>
   <author>
<!-- information about the author -->
   </author>
  </titleStmt>
  <extent>
<!-- size of the text, in pages and words -->
```

```
  </extent>
  <publicationStmt>
<!-- boilerplate statement about status as part of ELTeC -->
  </publicationStmt>
  <sourceDesc>
   <bibl>
<!-- bibliographic description of the printed source -->
   </bibl>
  </sourceDesc>
 </fileDesc>
 <profileDesc>
<!-- additional descriptive information -->
 </profileDesc>
 <revisionDesc>
<!-- revision information -->
 </revisionDesc>
</teiHeader>
```

Within the <teiHeader>, a <fileDesc>, a <profileDesc>, and a <revisionDesc> are all required. The <encodingDesc> may be supplied in (hopefully unlikely) event that some aspect of this document's encoding is anomalous.

5.1 Components of the file description

The <fileDesc> contains the following mandatory elements:

Taking these in turn, the <titleStmt> contains the title, author, and encoder of the document. For novels with multiple authors, titles, or encoders the element concerned is simply repeated. The <title >should be taken from an authoritative bibliographic source, and should include a phrase such as 'ELTeC edition'. The <author> may contain one or more of the following descriptive elements:

In addition to one or more <author> elements, a <titleStmt> should contain at least one <respStmt> element indicating the person responsible for the ELTeC encoded version, using the following elements

Here is an example :

```
<titleStmt>
 <title>Howards End : ELTeC edition</title>
 <author>
  <persName>
   <forename>Edward</forename>
   <forename>Morgan</forename>
   <surname>Forster</surname>
  </persName>
  <persName>E.M. Forster</persName>
  <birth when="1879"/>
  <death when="1970"/>
  <sex value="M"/>
  <idno type="viaf">https://viaf.org/viaf/31996364</idno>
  <idno type="wiki">https://www.wikidata.org/wiki/Q189119</idno>
 </author>
 <respStmt>
  <resp>ELTeC encoding</resp>
  <name>Lou Burnard</name>
 </respStmt>
</titleStmt>
```

The <extent> provides information about the size of the document, given by means of the following elements Exactly which measurements will be most useful and easily incorporated is yet to be determined: probably a count of words and pages will suffice.

```
<extent>
 <measure unit="words" quantity="20010"/>
 <measure unit="pages" quantity="245"/>
</extent>
```

The <publicationStmt> is required for TEI conformance: in individual text headers it will contain some standard boiler plate text referring to the fuller statement which will be furnished by the collection-level header.

```
<publicationStmt>
 <p>Incorporated into the ELTeC <date>2018-02-12</date>
 </p>
</publicationStmt>
```

The <sourceDesc> element is also required for TEI conformance. It will contain a bibliographic description of the source text against which the digital text has been validated, typically the first published edition of the work concerned. Where the ELTeC version derives from a pre-existing digital version of this work, a reference to that source will also be provided. The following elements are used to record this information:

```
<sourceDesc>
 <bibl>
  <author>E.M. Forster</author>
  <title>Howards End</title>
  <pubPlace>London</pubPlace>
  <publisher>Edward Arnold</publisher>
  <date>1910</date>
  <idno type="wiki">https://www.wikidata.org/wiki/Q1146642</idno>
 </bibl>
 <bibl>
  <title>The Project Gutenberg Etext of Howards End, by E. M. Forster</title>
  <ref target="http://www.gutenberg.org/files/2891/2891-h/2891-h.htm">HTML
     version downloaded on <date>2017-12-26</date>
  </ref>
 </bibl>
 <note type="editions" source="worldcat"> Worldcat lists 484 print editions in
   English</note>
</sourceDesc>
```

5.2 Components of the profile description

The <profileDesc> of an ELTeC text has the following mandatory components:

The <langUsage> element contains one or more <language> elements, one for each language, dialect, sublanguage etc. explicitly identified in the body of the text, indicating roughly how much of the text uses this language. For example, a text which is almost entirely in British English, but also contains some parts in US English would have an entry like this:

```
<langUsage>
 <language ident="en-GB" usage="90">British English</language>
 <language ident="en-US" usage="10">North American English</language>
</langUsage>
```

The TEI <textClass> element can contain one or more of the following elements: These three methods for classifying texts can be used in parallel. It is an **open question** which we should use for the ELTeC collection: the schema proposed here permits any combination.

The <keywords> option allows us to supply one or more <term> elements to categorise a text in some way. If the values are taken from a known closed list or authority file, that file should be specified using the *@source* attribute.

```
<textClass>
 <keywords source="http://wikidata.org">
  <term>social class</term>
  <term>social convention</term>
  <term>modernity</term>
  <term>family drama</term>
 </keywords>
</textClass>
```

Open Question : should we invent our own taxonomy, use a pre-existing one, make no attempt to constrain or predefine terms used here?

The <classCode> option allows us to use classification codes used or defined by existing authorities, such as library catalogue schemes, while the <catRef> option allows us to specify such codes using our own classification scheme.

```
<catRef target="#author_m #reprint_3"/>
<classCode source="UDC">8231.111</classCode>
```

Since our selection and descriptive criteria are likely to be specific to the project, we will probably have to define them in the corpus header using the following elements:

```
<taxonomy>
 <category xml:id="author_m">
  <catDesc>male authorship</catDesc>
 </category>
 <category xml:id="author_f">
  <catDesc>female authorship</catDesc>
 </category>
 <category xml:id="author_u">
  <catDesc>author gender unknown</catDesc>
 </category>
 <category xml:id="reprint_0">
  <catDesc>no reprints found</catDesc>
 </category>
 <category xml:id="reprint_1">
  <catDesc>1 to 50 editions</catDesc>
 </category>
 <category xml:id="reprint_2">
  <catDesc>50 to 100 editions</catDesc>
 </category>
 <category xml:id="reprint_3">
  <catDesc>Over 100 reprints</catDesc>
 </category>
</taxonomy>
```

5.3 Components of the Revision Description

The <revisionDesc> element is used to document significant points in the version history of the document. At least one entry should be provided for an ELTeC document, specifying when it was first added to the collection. The following elements can be used:

```
<revisionDesc>
 <change when="2018-02-21" who="ELTeC:LB">Added new linguistic
   classifications</change>
 <change when="2018-01-29" who="ELTeC:LB">Added to the ELTeC</change>
</revisionDesc>
```

5.4 Encoding description

The TEI allows for the specification of encoding practice, by which is meant documentation of the specific editorial policies followed during transcription (treatment of printed hyphens, lexical normalisation, sampling procedures, features included, ignored, or normalised, etc.). Such specification may be supplied at the individual document level, or once for all across the whole of a corpus. It is even possible to specify that different parts of a document follow different policies, provided that all the available policies are defined somewhere.

Open Question : We propose as far as possible not to allow for any variation in encoding policies applied within the ELTeC. We will still need to determine our encoding policies, of course, and to document them appropriately in the ELTeC corpus header, but there should be no need for separate specifications at the document level.

6 Linguistic and semantic annotation (level 2)

Additional markup facilities will be needed to represent more sophisticated annotations, which may be motivated linguistically (for example, to provide a normalised form, part of speech, etc.) or semantically (for example to distinguish proper names, names of people, places, events, etc.).

Sources consulted

[1] An introduction to TEI Simple Print <http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_simplePrint.doc.html>.

[2] Burnard, Lou 2005 'Metadata for corpus work' in *Developing Linguistic Corpora: A guide to good practice* ed. Martin Wynne. Oxford: Oxbow Books, pp 30-46.

[3] Odebrecht, Carolin. (2017). Metadata for Historical Corpora. Realization of the Metamodel for Corpus Metadata with the help of TEI Customization [Data set]. Zenodo. http://doi.org/10.5281/zenodo.267999

[4] <github.com/cligs/textbox>.