In search of comity: TEI for distant reading

Lou Burnard, Christof Schöch and Carolin Odebrecht

Searching for comity

https://www.lexico.com/en/definition/comity

> *Those participating in conversational encounters have to have a care for the preservation of good relations by promoting the other's positive self-image, by avoiding offence, encouraging comity, and so on. The negotiation of meaning is also a negotiation of social relations.*

(Widdowson, 1990, p. 110)

Our project is to establish a comity which can bridge the gap (identified by Jan Rybicki in his plenary) between the two most interesting and longest-established subfields of the Digital Humanities ...

The TEI was born interdisciplinary

This is not such a new idea...

Participants at the Poughkeepsie conference came from many disciplines, including computer science and computer services, NLP, theoretical linguistics but also literary studies, classics, historical, and medieval studies.

Its advisory board and formative working groups sought varied representation, both geographically and by discipline.

There is ample anecdotal evidence of cross disciplinary fertilisation

What text really is...

- ”Text” is in the mind of the reader: a construction of and for a particular community
- A generic abstract model summarizing the *significant properties* of all texts is conceivable
- Such properties may usefully be considered independently of
    - their expression in a particular document
    - their use in a particular discipline

TEI has always had to mediate two orthodoxies

- text is no less and no more than the documents which instantiate it
- text is no less and no more than a linguistic phenomenon, a bag of words whose statistical properties suffice to describe it

Is this model of text compatible with the model underlying the fields of stylometry, stylistics, textual analytics, aka distant reading?

COST Action 16204 “Distant Reading for European Literary History”

Distant Reading

- see `https://www.distant-reading.net`
- European network with 35 members from 23 countries bringing together researchers from different disciplines and scientific backgrounds
- Like TEI, COST is a community initiative fostering collaboration, interoperability, and mutual understanding
- Four working groups, reporting to a management committee comprising two national representatives from each participating country

COST contd.

- We report here on WG1 ("Scholarly resources") which is charged with design and construction of the the European Literary Text Collection (ELTeC).
    - a set of comparable corpora for each of at least a dozen European languages,
    - a balanced selection of 100 novels from the 19th century
    - metadata situating them in their contexts of production and of reception.

ELTeC is of course a TEI application...

ELTeC Encoding Requirements

- support computational approaches to literary text analysis (authorship attribution, topic modelling, stylistic analysis ...)
- enrich corpora with metadata and impose only a minimal structure
- editorial issues of lesser interest
- markup should offer the encoder very little choice, and the software developer very few surprises
- aim to facilitate uniform and consistent access across multiple corpora

Traditional TEI, by contrast, rejoices in variety, which makes comparative work harder

ELTeC encoding scheme/s

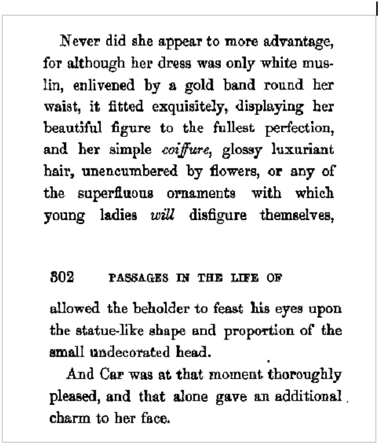
- level 0: minimal encoding scheme for texts produced manually or by OCR from print originals
- level 1 : somewhat richer format derivable automatically from texts encoded in other formats (Word, HTML TEI ...)
- level 2 : lingistically annotated and segmented

and a tightly constrained Header common to each level

level 0 : minimal

- discard non authorial front or back matter
- distinguish titlepage from other front matter
- mark chapter divisions and headings, but no substructure
- mark paragraphs (MLE blocks of text)
- reassemble words broken across lines
- discard paratext, illustrations, notes, corrections
- (optionally) mark pagebreaks and highlighting
- text may or may not be normalized: no indication either way

for example

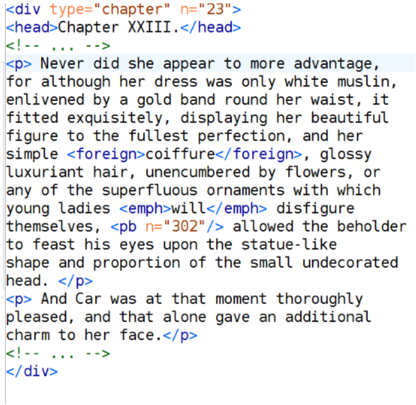


Never did she appear to more advantage, for although her dress was only white muslin, enlivened by a gold band round her waist, it fitted exquisitely, displaying her beautiful figure to the fullest perfection, and her simple *coiffure*, glossy luxuriant hair, unencumbered by flowers, or any of the superfluous ornaments with which young ladies *will* disfigure themselves,

302 PASSAGES IN THE LIFE OF

allowed the beholder to feast his eyes upon the statue-like shape and proportion of the small undecorated head.

And Car was at that moment thoroughly pleased, and that alone gave an additional charm to her face.



```
<div type="chapter" n="23">
<head>Chapter XXIII.</head>
<!-- ... -->
<p> Never did she appear to more advantage,
for although her dress was only white muslin,
enlivened by a gold band round her waist, it
fitted exquisitely, displaying her beautiful
figure to the fullest perfection, and her
simple <foreign>coiffure</foreign>, glossy
luxuriant hair, unencumbered by flowers, or
any of the superfluous ornaments with which
young ladies <emph>will</emph> disfigure
themselves, <pb n="302"/> allowed the beholder
to feast his eyes upon the statue-like
shape and proportion of the small undecorated
head. </p>
<p> And Car was at that moment thoroughly
pleased, and that alone gave an additional
charm to her face.</p>
<!-- ... -->
</div>
```

level 1 : some enrichment

- mark chapter substructure with <milestone> and <label> elements
- interpret highlighting, where possible, using <emph> <foreign> or <title>
- record authorial notes (gathered together into back)
- record graphics as <gap>
- normalized forms are explicit but original forms are lost

level 2 : basic linguistic annotation

- end to end segmentation using <s>
- tokenisation using <w>, with att.linguistic attributes *@pos*, *@lemma*, and *@join*
- (probably) mark named entities with <rs>
- inline annotation is possible because tagging is minimal

... a work in progress

Header checklist

All headers provide, in consistent format:

- Identification of title and language (*@xml:id* and *@xml:lang* on <TEI>)
- Title of the work, followed by the phrase ': ELTeC edition' (<title>)
- Author, in format 'Surname, Forename/s (birthYear-deathYear)' (<author>)
- Statements of responsibility in format 'encoded by Name' (<respStmt>)
- Date of publication in ELTeC (<publicationStmt>)
- Source description containing one or more <bibl> elements
- Encoding level : use the *@n* attribute of <encodingDesc>
- Profile description specifying languages used and sampling criteria values (<langUsage>, <textDesc>)
- Revision description containing at least one <change>

Most requirements are enforced by the schema: others by schematron rules

Metadata -1

A novel, for us, is an original, continuous fictional text, of over 10k words, published as a single work.

- Non-opportunistic design, aiming for a balanced representation of pre-defined features common across european tradition
- specifically:
  - date of first publication (one of five time slots between 1840 and 1920)
  - longevity/canonicity (high/low, as indicated by reprint count)
  - author sex (one of 3 values as perceived in 19c)
  - size (short/medium/long)
- Achieving a balance of these features is NOT EASY
- These features are encoded in the header using a customized <textDesc> element

```
<textDesc>
  <e:authorGender  key="M"/>
  <e:size  key="short"/>
  <e:canonicity  key="low"/>
  <e:timeSlot  key="T1"/>
</textDesc>
```

Metadata - 2

We aim to record the complete pedigree of our encoded texts in the header

- multiple types of <bibl> may be found in the <sourceDesc>
  - printEdition: print edition from which our text was derived (e.g. by OCR)
  - digitalEdition: published online edition from which our text was derived
  - firstEdition: details of first edition, irrespective of whether this was our source
- multiple <respStmt>s acknowledge encoders, editors, etc of previous editions

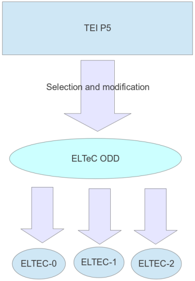
Publication issues

Only out of copyright works are included, and all texts are published under CC-BY licence.

- maintenance and development in publically accessible github repositories at `https://github.com/COST-ELTeC`
- Releases archived under Zenodo (first one due next month)

FAIR Guiding Principles: Findable on Zenodo; Accessible on Github; using TEI makes it Interoperable and Re-usable

ODD chaining



- base ELTeC ODD selects all and only elements required for each of the three schemas, and supplies generic constraints
- it is processed to create a TEI library, analogous to the "p5subset" supplied with TEI P5
- each ELTeC level is defined by a separate ODD, which selects a subset from that library

Promotes consistency of documentation and applicat

State of play and future work

Current state is visible at
`https://distantreading.github.io/ELTeC`

- Building the next (and subsequent!) releases
- Improving metadata, e.g. operationalizing canonicity counts
- Testing level2 proposals

News at `https://www.distant-reading.net/news/`