1 Abstract

Any expansion of the TEI beyond its traditional user-base involves a recognition of differing answers to the traditional question ‘What is text, really?’ [[8]; [7]; [13]], and hence a rethinking of some aspects of TEI praxis. We report on work carried out in the context of the COST Action CA16204 "Distant Reading", in particular on the TEI-conformant schemas developed for one of the Action's principal deliverables: the European Literary Text Collection (ELTeC).

The ELTeC will contain comparable corpora for each of a dozen European languages, each being a balanced sample of 100 novels from the 19th century, together with metadata situating them in their contexts of production and of reception. We hope that it will become a reliable basis for comparative work in data-driven textual analytics, enabling researchers to go beyond a simple ‘bag of words’ approach, while respecting views of "what text really is" currently dominant in such fields as statistically-derived authorship attribution, topic modelling, character network analysis, and stylistic analysis in general.

The focus of the ELTeC encoding scheme is not to represent texts in all their original complexity, nor to duplicate the work of scholarly editors. Instead, we aim to facilitate a richer and better-informed distant reading than a transcription of lexical content alone would permit. Where the TEI permits diversity, we enforce consistency, by defining encodings which permit only a specific and quite small set of textual features, both structural and lexical. We also define a single TEI-conformant way of representing the results of textual analyses such as named entity recognition or morphological parsing, and a specific set of metadata features. These constraints are expressed by a master TEI ODD, from which we derive three different schemas by ODD-chaining, each associated with appropriate documentation.

Lou Burnard is an independent consultant in TEI XML. He was for many years Associate Director of Oxford University Computing Services, and was one of the original editors of the TEI.

Christof Schöch is Professor of Digital Humanities at the University of Trier, Germany, and Co-Director of the Trier Centre for Digital Humanities (TCDH). He chairs the COST Action "Distant Reading for European Literary History" (CA16204).

Carolin Odebrecht is a corpus linguist at Humboldt-Universität zu Berlin. Her research fields are modelling, creating, archiving of historical corpora and corpus metadata.

In search of comity: TEI for distant reading

Lou Burnard, Christof Schöch and Carolin Odebrecht

2 Introduction

Comity is a term from theology or political studies, where it is used to describe the formal recognition by different religions, nation states, or cultures that other such entities have as much right to existence as themselves. In applied linguistics, the term has also been used by such writers as Widdowson or Aston [[2]] seeking to demonstrate how the establishment of comity can facilitate successful inter-cultural communication, even in the absence of linguistic competence[1]. We appropriate the term here in this latter sense, as a means of re-asserting the inter-disciplinary roots of the TEI.

Recent histories of the TEI (e.g. [9]) have a tendency to under-emphasize the multiplicity of disciplines gathered at its birth, preferring to focus on those disciplines which can be plausibly framed as prefiguring our current configuration of the ‘digital humanities’ in some way. Yet the Poughkeepsie conference and the process of designing the Guidelines which followed alike were kickstarted by input from corpus linguists and computer scientists as well as traditional philologically-minded editors and source-driven historians. The TEI belongs to a multiplicity of research communities, dating as it does from a period when computational linguists and traditional philologists alike were beginning to wake up to the implications of the advent of massive amounts of digital text for their disciplines. The steering committee which oversaw its development and the TEI editors alike conscientiously attempted to ensure that the Guidelines should reflect a view of text which was generally shared and generic, rather than specific to any discipline or to any particular usage model.

The TEI's radical proposition that there was such a thing as a single abstract model of textual components, which might usefully be considered independently of its expression in a particular source or output, or its use in any particular discipline, was necessarily at odds with at least two prevailing orthodoxies: on the one hand, the view that a text is no less and no more than the physical documents which instantiate it, and can be adequately described and represented by its salient visual properties alone; on the other hand, the view

[1] ‘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.’ [14]

that a text is solely a linguistic phenomenon, comprising a bag of words, the statistical properties of which are adequate to describe it. But the TEI tried very hard to prefer comity over conflict, not only in its organization, which brought together an extraordinarily heterogeneous group of experts, but also in its chief outputs: a set of encoding Guidelines which while supporting specialization did not require any particular specialisation to prevail.

Old orthodoxies do not die easily, and it is interesting to hear how some of the same arguments are still being played out in the somewhat different context of today's DH theorizers. But in our present paper, we simply want to explore the extent to which the TEI's model of text can be adapted to conform to the model of text characterising such fields as stylometry, stylistics, textual analytics, or (to use the current term) ‘distant reading’. We hope also to explore the claim that by so doing we may facilitate the enrichment of that model, and thus facilitate more sophisticated research into textual phenomena across different corpora. And we hope to demonstrate that this is best done by cultivating mutual respect for the widely differing scientific, cultural, and linguistic traditions characterising this cross-European and cross-disciplinary project, that is, by acknowledging a comity of methods as well as languages.

3 The COST Action “Distant Reading for European Literary History”

The context for this work is the EU-funded COST Action “Distant Reading for European Literary History” (CA 16204) a principal deliverable of which will be the European Literary Text Collection (ELTeC). This is a set of comparable corpora for each of at least a dozen European languages, each corpus being a balanced selection of 100 novels from the 19th century, together with metadata situating them in their contexts of production and of reception. It is hoped that the ELTeC will become a reliable basis for comparative work in cross-linguistic data-driven textual analytics, eventually providing an accessible benchmark for a particular written genre of considerable cultural importance across Europe during the period between 1840 and 1920.

Two significant decisions made early on in the planning of the COST Action underlie the work reported here. Firstly, it was agreed that the ELTeC should be delivered in a TEI-encoded format, using a schema developed specifically for the project. Secondly, the design of that encoding scheme, in particular the textual features it makes explicit by means of markup, should be defined as far as possible by the needs of the distant reading research community, rather than any pre-existing notion of textual ontology, to the extent that the needs of that community could be determined. The target audience envisaged includes experts in computational stylistics, in corpus linguistics, and in traditional literary studies as well as more general digital humanists, but is probably best characterized as having major enthusiasm and expertise in the application of statistical methods to literary and linguistic analysis, and only minor interest in the kinds of textual features most TEI projects have tended to focus on.

The work of the Action [2] is carried out in four Working Groups (WGs), whose activities are subject to endorsement and acceptance by a Management Committee, composed of two national representatives from each of the 29 countries currently participating in the Action. The Working Group heads are also members of a smaller ‘core’ group responsible for day to day management of the Action. WG1 Scholarly Resources is responsible for the work described in this paper; WG2 Methods and Tools is concerned with text analytic techniques and tools; WG3 Literary Theory and History is concerned with applications and implications of those methods and for literary theory ; WG4 Dissemination is responsible for outreach and communication.

The design and construction of the ELTeC is the responsibility of WG1, as noted above. Initially, this work was split into three distinct tasks: First, defining selection criteria (corpus design); second, developing basic encoding methods (both for data and for metadata); and third, defining a suitable workflow for preparation of the corpus. Working papers on each of these topics plus a fourth on theoretical issues of sampling and balance were prepared for discussion and approval by the 30 members of WG1, and remain available from the Working Group's website. Their proposals were ratified by the Management Committee after discussion at two meetings during 2018.

4 The ELTeC Encoding Scheme/s

The encoding requirements for ELTeC were perceived by WG1 to be somewhat different from those of many other TEI projects. Distant Reading methods cover a wide range of computational approaches to literary text analysis, such as authorship attribution, topic modelling, character network analysis, or stylistic analysis but they are rarely concerned with editorial matters such as textual variation, the establishment of an authoritative text, or production of print or online versions of a text. Consequently, the ELTeC encoding scheme was deliberately not intended to represent source documents in all their original complexity of structure or appearance, but rather to make it as simple as possible to access the words of which texts are composed in an informed and predictable way. The goal was neither to duplicate the work of scholarly editors nor to produce (yet another) digital edition of a specific source document. Rather, the encoding scheme was designed in such a way as to ensure that ELTeC texts could be processed by simple minded (but XML-aware) systems primarily concerned with lexis and to make life easier for the developers of such systems.

An important principle following from this latter goal is that ELTeC markup should offer the encoder very little choice, and the software developer very few surprises: the number of tags available is greatly reduced, and their application is tightly constrained. It facilitates processing greatly if access to each part of the

[2] Further information about the Action is available from its website at https://www.distant-reading.net

XML tree can be provided in a uniform and consistent way across multiple ELTeC corpora.

By default, the TEI provides a very rich vocabulary, and many subtly different ways of doing more or less the same thing. TEI encoders have frequently taken full advantage of that to produce texts which vary enormously, both in the subset of XML tags used and in the range of attribute values associated with them. It is tempting, but entirely mistaken, to assume that the allegedly TEI-conformant deliverables from project A will necessarily be marked up in the same way as the allegedly TEI-conformant deliverables from project B [3]. On the contrary, all that 'TEI conformance' really guarantees is that the intended semantics of the markup used by the two projects should be recoverable by reference to a published standard, and are not entirely ad hoc or sui generis. (This may not seem much of an advance, though it is: see further [[5]]).

Following this No Surprises principle, the simplest ELTeC schema (the 'level zero ' schema) provides the bare minimum of tags needed to mark up the typical structure and content of a nineteenth century novel. All preliminary matter other than the title-page and any authorial preface or introduction is discarded; the remainder is marked as a <`div`> of *@type*`titlepage` or `liminal`, within a <`front`> element. Within the <`body`> of a text, the <`div`> element is also used to make explicit its structural organization, with *@type* attribute values `part`, `chapter`, or `letter` only[4]. For our purposes, a 'chapter' is considered to be the smallest subsection of a novel within which paragraphs of text appear directly. Further subdivisions within a chapter (often indicated conventionally by ellipses, dashes, stars etc.) are marked using the <`milestone`> element; larger groupings of <`div`> elements are indicated by <`div`> elements, always of type `part`, whatever their hierarchic level. Headings, at whatever level, are always marked using the <`head`> element when appearing at the start of a <`div`>, and the <`trailer`> element when appearing at the end. Within the <`div`> element, only a very limited number of elements is permitted: specifically, in addition to those already mentioned, <`p`> or <`l`> (verse line). Within these elements we find either plain text, <`hi`> (highlighted), <`pb`> (page break) or <`milestone`> elements. After some debate, the Action's Management Committee agreed that it would be practical to require only this tiny subset of the TEI for all ELTeC texts.

It should be noted that the texts included in an ELTeC corpus may come from different kinds of source. For some language collections, no digital texts of any kind exist: the encoder must start from page images, manually transcribe or put them through OCR, and introduce ELTeC markup from scratch. For others, existing digital texts may already be available: the encoder must research the format used and find a way of converting it to ELTeC. In some cases, a TEI

[3] A large-scale project called MONK (Metadata Offer New Knowledge) demonstrated some of the technical consequences of this for integrated searching of TEI resources: see further http://monk.library.illinois.edu

[4] An exception is made for epistolary novels which contain only the representation of a sequence of letters, with no other significant content: these may be marked as

version may already exist; in others a project Gutenberg HTML version; in yet others the text may be stored in a database of some kind. Whichever is the case, if it is possible to retain distinctions which the ELTeC scheme permits, this is clearly desirable; perhaps less obviously, it is also necessary to remove distinctions made by the original format which the ELTeC scheme does not permit. This diversity of source material was one motivation for permitting multiple encoding levels in the ELTeC scheme: at level zero, only a bare minimum of markup is required or permitted, while at level 1 a slightly richer (but still minimalist) encoding is also defined. At level 2, further tags again are introduced to support linguistic processing of various kinds, as discussed further below. Down-conversion from a higher to a lower level is always automatically possible, but up-conversion from a lower to a higher level generally requires human intervention or additional processing.

At level 1, the following additional distinctions may be made in an encoding:

- the <label> element may be used for heading-like titles appearing in the middle of a division;
- the <quote> element may be used to distinguish passages such as quotations, epigraphs, stretches of verse, letters etc. which seem to 'float' within the running text;
- the <corr> element may be used to indicate a passage (typically a word or phrase) which is clearly erroneous in the original and which has been editorially corrected;
- the elements <foreign>, <emph>, or <title> are available and should be used in preference to <hi> for passages rendered in a different font or otherwise made visually salient in the source, where an encoder can do so with confidence;
- the element <gap> may be used to indicate where some component of a source (typically an illustration) has been left out of the encoding;
- the elements <note> and <ref> may be used to capture the location and content of authorially supplied footnotes or end-notes; wherever they occur in the source, notes must be collected together in a within a <back> element.

For those already familiar with the TEI, this list of elements may seem distressingly small. It lacks entirely some elements which every TEI introductory course regards as indispensable (no <list> or <item>; no <choice> or <abbr>; no <name> or <date>...) and tolerates some practices bordering on tag abuse. For example, all the components of a title page are marked as <p> since no specialised elements (<titlePage>, <docImprint> etc.) are available. In the absence of specialised but culture-specific features (for example, publisher name, imprint, imprimatur, etc.) the encoding identifies only fundamental textual features common to every kind of text. Nevertheless, we believe that the set of concepts it supports overlaps well with the set of textual features which almost any existing digital transcription will seek to preserve in some form or another. This may explain both why the majority of the texts so far collected in the ELTeC

have been encoded at level 1 rather than level0, and also the speed with which the collection is growing.

ELTeC level 1 is intended to facilitate a richer and better-informed distant reading of a text than a transcription of its verbal content alone would permit. ELTeC level 2 is partly intended to provide a consistent and TEI-conformant way of representing the results of such readings, in partiocular those concerned with linguistic annotation. Its primary goal is to represent in a standard way additional layers of annotation of particular importance to distant reading applications such as stylometry or topic modelling. Enrichment of each lexical token to indicate its morpho-syntactic category (POS) or its lemma, and identification of tokens which refer to named entities are both well within the scope of existing text processing techniques, and are also routinely used in distant reading applications. The challenge is that the input and the output formats typically used by such tools are rarely XML-based, and seem superficially to have a model of text quite different from that of the ‘ordered hierarchy of content objects’ in terms of which the TEI community traditionally operates. For many in the distant reading community (it seems) a text is little more than a sequence of tokens, mostly corresponding with orthographically-defined words, though there is some variability in the principles underlying the process of tokenisation, for example in the modelling of clitics, compound forms, etc. Each token has a number of properties, which might include such attributes as its part of speech, its lemma, or its position in the sequence of tokens making up the document. Information about a token which in an XML model would be properties of some higher level construct such as its status as dialogue, quoted matter, emphasis, etc. is occasionally considered as well, but is typically modelled as an additional property of the token.

If a community is defined by its tools, it would appear therefore that the distant reading community has not fully embraced the notion of XML as anything other than a rather verbose archival format. However, communities are not defined solely by their tools : by seeking a way of reconciling these differing views of what text really is in a spirit of comity we hope to demonstrate that there are advantages both for the distant reader or stylometrician and for the literary analyst or textual editor.

At ELTeC level2, all existing elements are retained and two new elements <s> and <w> are introduced to support segmentation of running text into sentence-like and word-like sequences respectively. Individual tokens are marked using the <w> element, and decorated with one or more of the TEI-defined linguistic attributes *@pos*, *@lemma*, and *@join*. Both words and punctuation marks are considered to be ‘tokens’ in this sense, although the TEI suggests distinguishing the two cases using <w> and <pc> respectively. The <s> (segment) element is used to provide an end-to-end tessellating segmentation of the whole sequence of <w> elements, based on orthographic form. This provides a convenient extension of the existing text-body-div hierarchy within which tokens are located.

The elements <p>, <head>, and <l> (which contain just text at levels 0 and 1) at level 2 can contain a sequence of <s> elements. Empty elements <gap>, <milestone>, <pb> or <ref> are also permitted within text content at any point, but these are disregarded when segmentation is carried out. Each <s> element can contain a sequence of <w> elements, either directly, or wrapped in one of the sub-paragraph elements <corr>, <emph>, <foreign>, <hi>, <label>, <title>. To this list we might add the element <rs> (referring string), provided by the TEI for the encoding of any form of entity name, such as a Named Entity Recognition procedure might produce.

This approach implies that <w> elements may appear at two levels in the hierarchy which may upset some software; it also implies that <w> elements must be properly contained within one of these elements, without overlap. If either issue proves to be a major stumbling block, an alternative would be to remove the tags demarcating these sub-paragraph elements, indicating their semantics instead by additional attribute values on the <w> elements they contain.

This TEI XML format is equally applicable to the production of training data for applications using machine learning techniques and to the outputs of such systems. However, since such machine learning applications typically operate on text content in a tabular format only, XSLT filters which transform (or generate) the XML markup discussed here from such tabular formats without loss of information are envisaged. At the time of writing, however, Working Group 2 has yet to put this proposed architecture to the test.

5 ELTeC metadata and corpus design

Like every other TEI document, every ELTeC text has a TEI Header, though its organization and content alike are constrained much more tightly than is common TEI praxis, for the reasons already mentioned. The structure of an ELTeC Header is the same no matter what level of encoding applies to the text. It provides minimal bibliographic information about the encoded text and its source, sufficient to identify the text and its author, in a fixed and consistent format. It is assumed that if more detailed bibliographic information is required, for example about the author or work encoded, this is better obtained from standard authority files; to that end a VIAF code may be associated with them.

As noted above, ELTeC texts may be derived from many sources, each of which should be documented correctly in the header's <sourceDesc> element. After some debate, a common set of practices has been identified to distinguish (for example) ELTeC texts derived directly from a print source from those derived from a digital source, itself derived from a known print source, and to provide information about each source. In the following example, the source of the ELTeC version is a pre-existing digital edition provided by Project Gutenberg , but the source description also provides information about the first print edition of the work concerned.

```
<bibl type="digitalSource">
 <title>Project Gutenberg EBook A engomadeira de Almada Negreiros</title>
 <ref target="http://www.gutenberg.org/ebooks/23879"/>
</bibl>
<bibl type="firstEdition">
 <title>A engomadeira</title>
 <author>José de Almada Negreiros</author>
 <publisher>Typographia Monteiro &amp;amp; Cardoso</publisher>
 <date>1917</date>
</bibl>
```

In most cases, the ELTeC text will correspond with the first edition of a work in book form; but even where this is not the case, or where information about the precise source used is not available, minimal information about that first edition should also be provided in order to place the work in its original temporal context.

As with other TEI conformant documents, beside the mandatory file description, the TEI Header of every ELTeC text contains a publication statement which specifies its licensing conditions (all ELTeC documents are licensed CC-BY); an encoding statement specifying the level of encoding used; and a revision description containing versioning information. The TEI Header is also used to provide metadata describing the associated text in a standardized form; this is held in the <`profileDesc`> element which must specify the languages used by the text, may optionally include a <`textClass`> element containing any culture-specific keywords considered useful to describe the text, and must contain a <`textDesc`> element which documents the text's status with respect to selection criteria discussed below.

One of the knottier problems or (to be positive) more distinctive features of an ELTeC language collection is that it is not intended to be an *ad hoc* accidentally constructed corpus but a designed one. Its composition is determined not by the happenstance of whatever we can get our hands on, but is instead defensible, at least in theory, as a principled and representative selection.

The big question is, of course, representative of *what*.

It would be nice to say that it represents the production of novels in a specific language in 19th century Europe. WG1 has working definitions for both "novels" and "Europe" which we do not discuss further here, though both are clearly problematic terms. It is hoped that the ELTeC will provide data for an empirical discussion of such terms, feeding into the work of WG3 on literary theory and terminology.

But we cannot make that claim without any data about the population we are claiming to represent -- which is hard to come by for many of the languages concerned. We know about the novels which we know about, which tend to be the ones that national libraries or equivalent cultural heritage institutions have

chosen to preserve, which publishers over time have been able to sell, and which lecturers in literary studies have chosen to teach. More ephemeral titles may have been collected (for example by a copyright library); but equally well may have been discarded or even suppressed as unworthy of inclusion in the national patrimony. Titles and authors alike can go in and out of fashion. But how can we express opinions about changes in the nature of the published novel if the sample on which we base those opinions is wildly different in composition from the actual population? If our data leads us to assert that novels in a given language are never written by women, or are never of fewer than 100,000 words is this simply because no female authors happen to have been preserved, or because short novels were routinely discarded from the collection? Or, on the other hand, does this actually indicate something fundamental, a characteristic of the population we are investigating? This matters particularly for ELTeC, one of the goals of which is precisely to facilitate cross-language comparisons.

This problem of representativeness is of course one which every corpus linguist has to face, and discussions of its implications are easy to find in the literature [5]

Our approach is to sidestep the impossibility of representing an unknown (and sometimes unknowable) population by attempting instead to represent the range of possible variation in the values of a predefined set of variables, each corresponding with a more or less objective category of information available for all members of the population. To take a trivial example, every novel can be characterised as short, medium, or long; there is no possible fourth value for this category unless we revise our definition of length (elastic? unknown? instantaneous?). So, as a working hypothesis, we might say that a corpus in which roughly a third of the titles are short, a third are long, and a third are medium will represent the variation possible for this category. If we apply this principle uniformly across all our corpora, we can reliably investigate (for example) cross language variation in some other observable phenomenon (say a fondness for syntactically complex sentences) with respect to length. But note that we have made absolutely no claim about whether novel length in the underlying population is also divided in this way.

The decade in which a novel first appears in book form is a similarly objectively characteristic, which in principle we can determine for every member of the population. We can also classify every title according to the actual sex of their author (with values such as female, male, mixed, unknown). And we can likewise classify a title in terms of its staying power or persistence by looking at the number of times it has been reprinted since its first appearance. We suggest that texts which have been frequently reprinted over a long period may reasonably be considered ‘canonical’ in some sense of that vexed term. The goal of our corpus balancing exercise is to ensure more or less equal time for each possible value for each of these four categories -- size, decade, authorSex, and canonicity.

[5] Some notable examples include [3]; [11]; [4]

Ideally, each corpus should have equivalent numbers not just for each value, but for each combination of values: so, for example, looking at the third of all titles which are characterised as "short", there should be roughly equal numbers for each decade of first appearance, roughly equal numbers by male and female authors, and so on. This may however be a council of perfection. It is already apparent that for some languages, it is very difficult to find any texts at all within some time periods, or by female authors. Similarly, our definition of "short" (10-50 thousand words), "medium" (50-100 thousand words) and "long" (over 100 thousand words) though objective and easy to validate, assumes that there will be enough novels of a given length in the underlying population for us to extract a balanced sample; but in some languages it may be that the distribution of lengths across the population is entirely different. We cannot tell whether (for example) the absence of any "long" novels at all in Czech, Serbian, or Norwegian is characteristic of those languages, or an artefact of the selection process so far. Another difficulty is that our corpus design deliberately seeks to include some forgotten or marginal works along with well-known canonical texts: this is relatively easy for traditions such as English, French, or German where copyright laws have led to the maintenance and documentation of large national collections, but less so for other less well documented languages. (For some initial data, see the summary page at https://distantreading.github.io/ELTeC/)

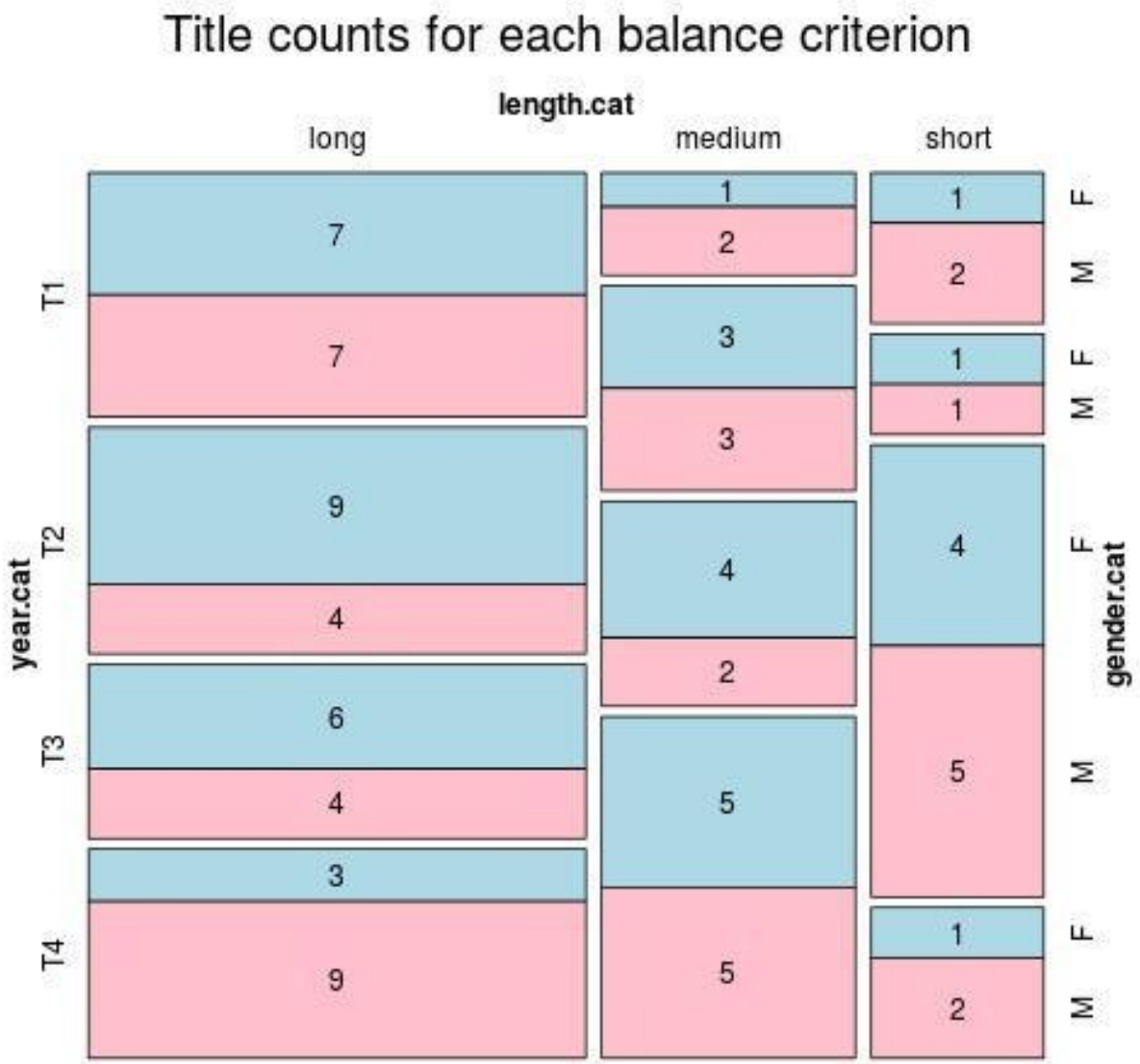
To encode these balance criteria in the TEI Header in as direct and accessible a manner as possible, we have chosen to re-purpose the little-used <textDesc> element, originally provided by the TEI as a wrapper for a set of so-called situational parameters proposed by corpus linguists as a way of objectively characterizing linguistic production [6] In our case, we replace the TEI's suggested vocabulary for these parameters with a vocabulary representing our four criteria, expressed as new non-TEI elements in the ELTeC namespace. These elements (<eltec:sex>, <eltec:size>, <eltec:canonicity>, and <eltec:timeSlot>) are required by the ELTeC schemas and have an attribute *@key* which supplies a coded value for the criterion concerned taken from a predefined closed list. So, for example, a long (over 100,000 words) novel by a female author first published between 1881 and 1900 but only infrequently reprinted thereafter might have a text description like the following:

```
<textDesc
   xmlns:eltec="http://distant-reading.net/ns">
 <eltec:authorGender key="F"/>
 <eltec:canonicity key="low"/>
 <eltec:size key="long"/>
 <eltec:timeSlot key="T3"/>
</textDesc>
```

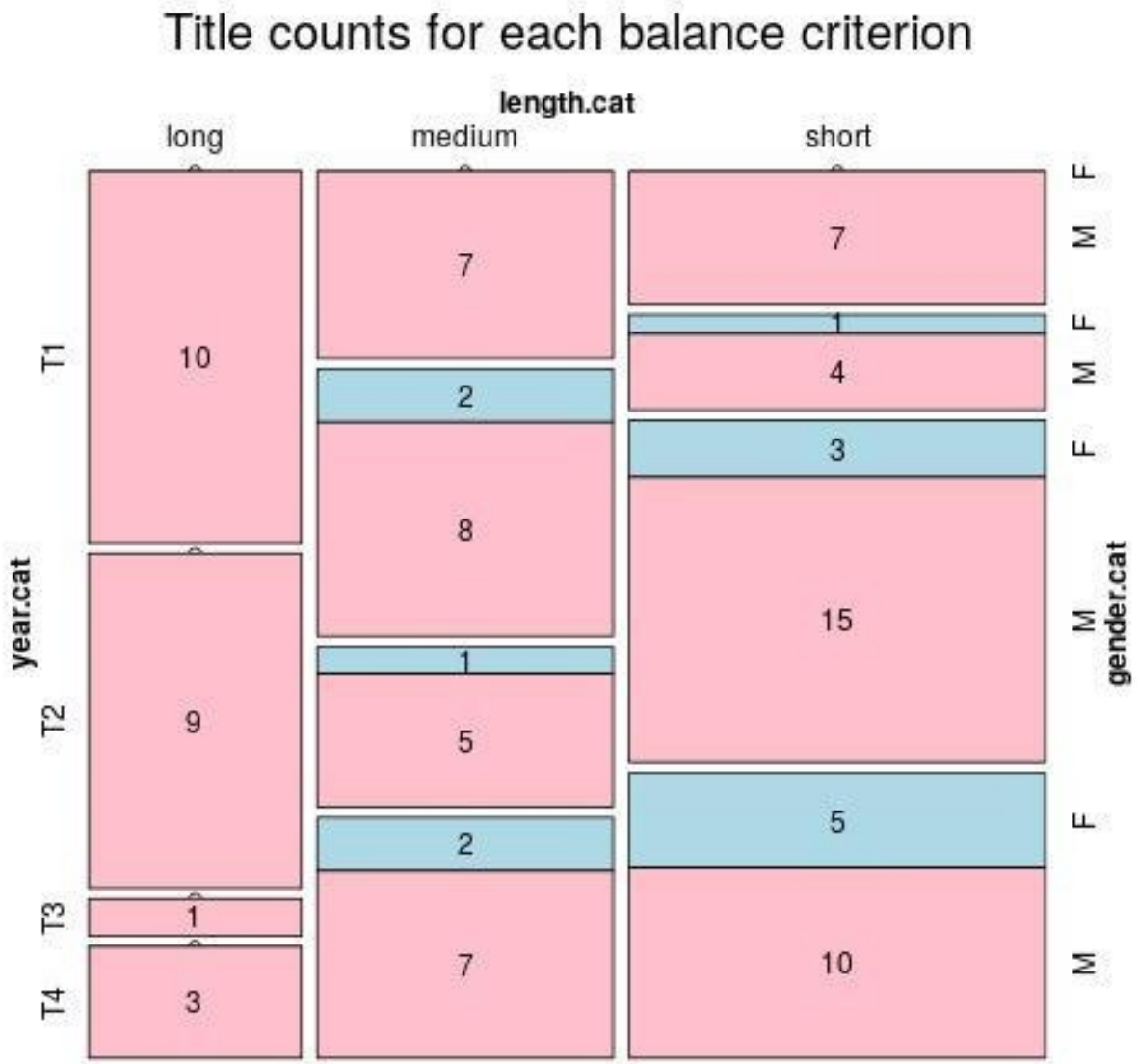
When complete, this information can be used to select subcorpora from the collection as a whole, thus permitting more delicate cross-linguistic comparisons: for example between the lexis of male and female writers, or between the

[6] The <textDesc> element is discussed in section 15.2.1 of the TEI *Guidelines* (https://tei-c.org/release/doc/tei-p5-doc/en/html/CC.html#CCAHTD).

stylistic features typically associated with long or short texts. During the construction phase, these coded values also make it easy to monitor the emerging composition of the corpus, for example to detect whether or not the ratio of male to female writers is consistent across different time periods, by means of a simple visualisation like the following



This ‘mosaic plot’ for the current state of the English corpus (90 texts) shows that there are roughly as many female (blue) as male (pink) writers across the board, but that there is a preponderance of long texts and of titles published in time slot 3.



For comparison, the same plot for the current state of the Hungarian corpus (100 texts) shows significantly fewer female writers, and a higher proportion of short texts. Whether these variations are an artefact of the sampling process or represent differences in the underlying population is precisely one of the research questions which our approach requires us to address.

6 Chaining ODDs

The TEI's ODD (One Document Does it all) system [[12]] is widely used as a means of customizing the TEI and documenting the customization in a standard way. When only a single ODD customization is used across a project, there is a natural tendency to produce broadly permissive schemas, to allow for the inevitable variation of requirements when material of different kinds are to be processed in an integrated collection. But this prevents the encoder from taking full advantage of the ability of an XML schema to check that particular

documents conform to predefined rules, unless they are willing greatly to increase the complexity of their work flow. A better approach, pioneered by the Deutsch Textarchiv [10], has been the use of a technique known as ODD chaining [5] Here, a project first defines a base ODD which selects all the TEI components considered to be useful anywhere and then uses this as the basis for smaller, more constraining, ODDs which select from the base only the components (or other rules) specific to a subset of the project's documentary universe. For example, an archive may have identified a common set of metadata it wishes to document across all of its holdings but also have particular metadata requirements for print and manuscript sources respectively. Simply defining two different ODDs, one for print and one for manuscript when many other components apply to either kind of source opens the door to redundant duplication and the risk of inconsistency. The ODD chaining approach requires definition of a base ODD which contains the union of the components needed for these two different ODDs, constructed as an appropriate selection from the full range of TEI components. The ODDs for print and manuscript are then defined as further specialisations or customizations of the base, ensuring thereby that the common components are used in a consistent manner, but preserving comity by allowing equal status to the two specialised schemas.

In the ELTeC project, we begin by defining an ODD which selects from the TEI all the components used by any ELTeC schema at any level. This ODD also contains documentation and specifies usage constraints applicable across every schema. This base ODD is then processed using the TEI standard `odd2odd` stylesheet to produce a standalone set of TEI specifications which we call `eltec-library`. Three different ODDs, eltec-0, eltec-1, and eltec-2 then derive specific schemas and documentation for each of the three ELTeC levels, using this library of specifications as a base rather than the whole of the TEI. This enables us to customize the TEI across the whole project, while at the same time respecting three different views of the resulting encoding standard. As with other ODDs, we are then able to produce documentation and formal schemas which reflect exactly the scope of each encoding level.

The ODD sources and their outputs are maintained on GitHub and are also being published on Zenodo along with the ELTeC texts. [7]

7 State of play and future work

The ELTeC is still very much a work in progress, and hence we cannot report that our design goals have been achieved with any plausibility. An initial release of the collection is due on Zenodo in September 2019, and we expect several future releases before the target of 100 texts per language is reached. The corpora are also maintained as a collection of publicly visible GitHub repositories, as noted above.

[7] The GitHub repository for the ELTeC collection is found at https://github.com/COST-ELTeC/ : the Zenodo community within which it is being published lives at: https://zenodo.org/communities/eltec,

As well as continuing to expand the collection, and continuing to fine-tune its composition, we hope to improve the consistency and reliability of the metadata associated with each text, as far as possible automatically. For example, we have developed two complementary methods of automatically counting the number of reprints for each title, one by screen scraping from WorldCat, and the other by processing data from a Z39.50 server where this is available. These methods should provide more reliable data than has hitherto been available for the 'canonicity' criterion mentioned above.

The main area of future work we anticipate is however in the testing of the proposed ELTeC level 2 encoding and an evaluation of its usefulness. At a technical level, this may necessitate some changes in the existing markup scheme, but of perhaps more interest is the extent to which its availability will exemplify the virtue of striving for comity amongst the many ways in which TEI XML markup can be applied.

[1] References

[2] Aston, Guy (1988) Learning Comity: An Approach to the Description and Pedagogy of Interactional Speech (Testi e discorsi: Strumenti linguistici e letterari, vol 9) Bologna: CLUEB

[3] Biber, Douglas (1993). "Representativeness in Corpus Design". In: Literary and Linguistic Computing (8), pp. 243–257.

[4] Bode, Katherine (2018). A World of Fiction - Digital Collections and the Future of Literary History. eng. University of Michigan Press.

[5] Burnard, Lou (2016) ODD Chaining for Beginners. Available from http://teic.github.io/PDF/howtoChain.pdf

[6] Burnard, Lou (2019) "What is TEI Conformance, and why should you care?". In: Journal of the Text Encoding Initiative, Issue 12. https://journals.openedition.org/jtei/1777

[7] Caton, Paul (2013). "On the term text in digital humanities". In: Literary and Linguistic Computing 28.2, pp. 209–220.

[8] De Rose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear (2002). "What is Text, Really?" In: Journal of Computing in Higher Education I(2), pp. 3–26.

[9] Gavin, Michael (2019) "How to think about EEBO". In: Textual Cultures Vol 11, no 1-2 (2017). https://doi.org/10.14434/textual.v11i1-2.23570

[10] Haaf, Susanne and Christian Thomas (2016) "Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the

Format” In: Journal of the Text Encoding Initiative, Issue 10. https://journals.openedition.org/jtei/1650

[11] Lüdeling, Anke (2011). “Corpora in Linguistics. Sampling and Annotation”. In: Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Ed. by Karl Grandin. Vol. 147. Nobel Symposium 147. New York:

[12] Rahtz, Sebastian, and Lou Burnard (2013) “Reviewing the TEI ODD System”. In Proceedings of the 2013 ACM Symposium on Document Engineering. DocEng ’13. ACM, 2013. http://doi.acm.org/10.1145/2494266.2494321

[13] van Zundert, Joris and Tara L. Andrews (2017). “Qu’est-ce qu’un texte numérique? A new rationale for the digital representation of text”. In: Digital Scholarship in the Humanities 32, pp. 78–88.

[14] Widdowson, Henry (1990) Aspects of Language Teaching. OUP.