The European Literary Text Collection ELTeC

Distant Reading

Carolin Odebrecht; Lou Burnard; Borja Navarro Colorado; Maciej Eder; Christof Schöch

COST Action Distant Reading and Working Groups

Network: European Cooperation in Science and Technology Action "Distant Reading for European Literary History"(CA16204)

Goals:

- create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written
- contribute to the development and distribution of methods, competencies, data, best practices, standards and tools relevant to Distant Reading research

Data: create an open source multi-lingual benchmark corpus for European literature

Methods: coordinate activities related to sharing, evaluating and improving methods and tools for Distant Reading research

Theory: exploring theoretical concerns that stem from the application of Distant Reading methods to literary history

Members of the Action: red dots indicate the home instition locations of our members.

Corpus Data – ELTeC

Design: collect 100 texts per language, follow a non-normative but metadata-based approach (not canon-based, see (1)) and aim to represent the variety of a population (2)

Text candidates (3):

- **language**: European languages, no translations
- **sources**: narrative fictional prose
- **period**: 1840–1920
- **length**: min. 10.000 words
- **publication**: prefer books over novels published in serial publications
- **access**: only freely available digitizations

Languages: Spanish, English, French, German, Greek, Portuguese, Dutch, Serbian, Hungarian, Italian, Slovenian, Romanian, Czech, Swedish, Latvian, Russian, Polnish, Croatian

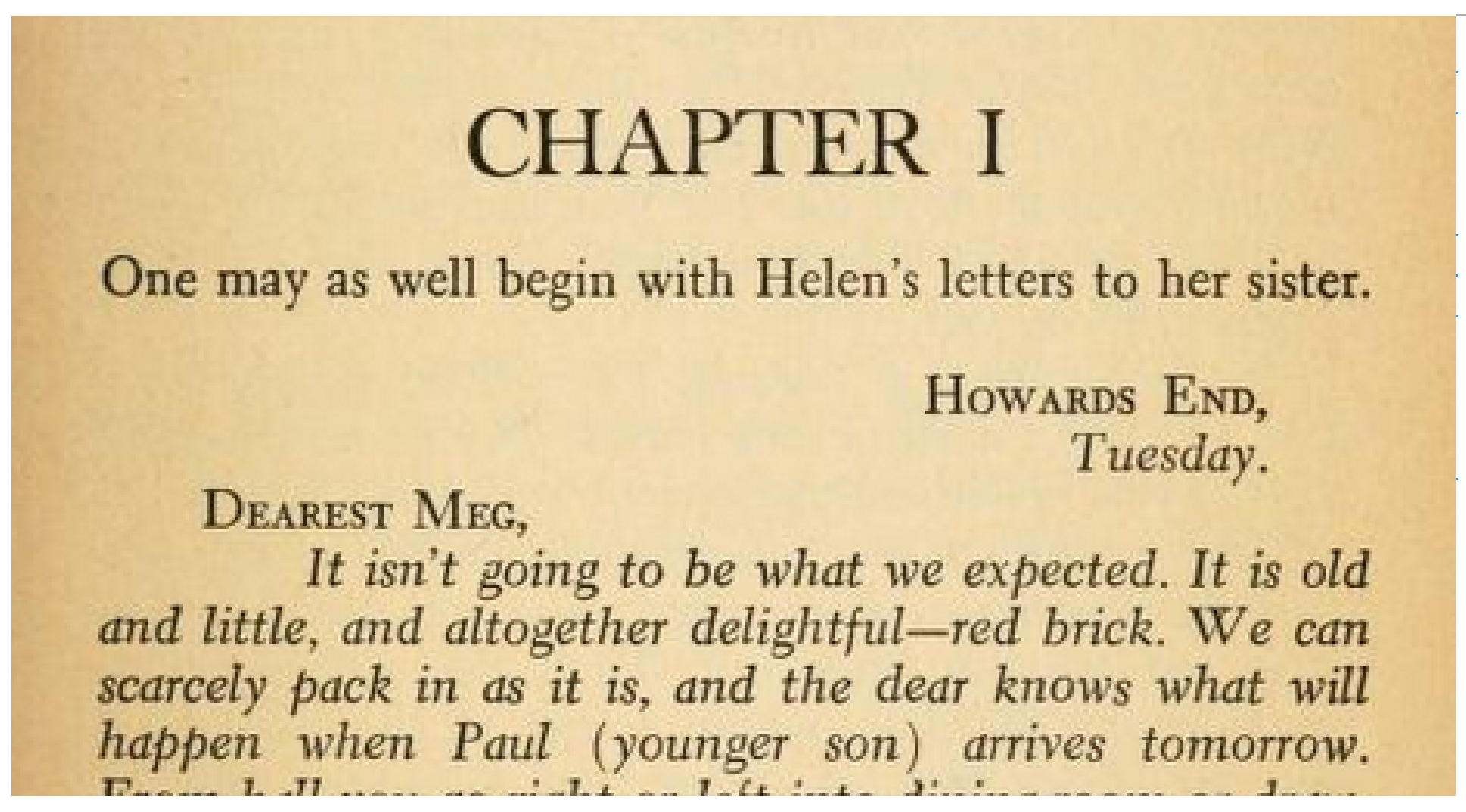
Encoding principles:

- minimal, uniform encoding (applicable for different sources)
- text representation focus on consistency and simplicity and basic metadata and references

TEI Encoding levels (4):

- **level0**: basic encoding with e.g. paragraph, heading, page break, text division,
- **level1**: richer encoding, adding e.g. font change, graphic, quotation, correction
- **level2**: token-based encoding with automatic lemmatization and part-of-speech annotation (work in progress)

Corpus data – Example of the English Language Collection



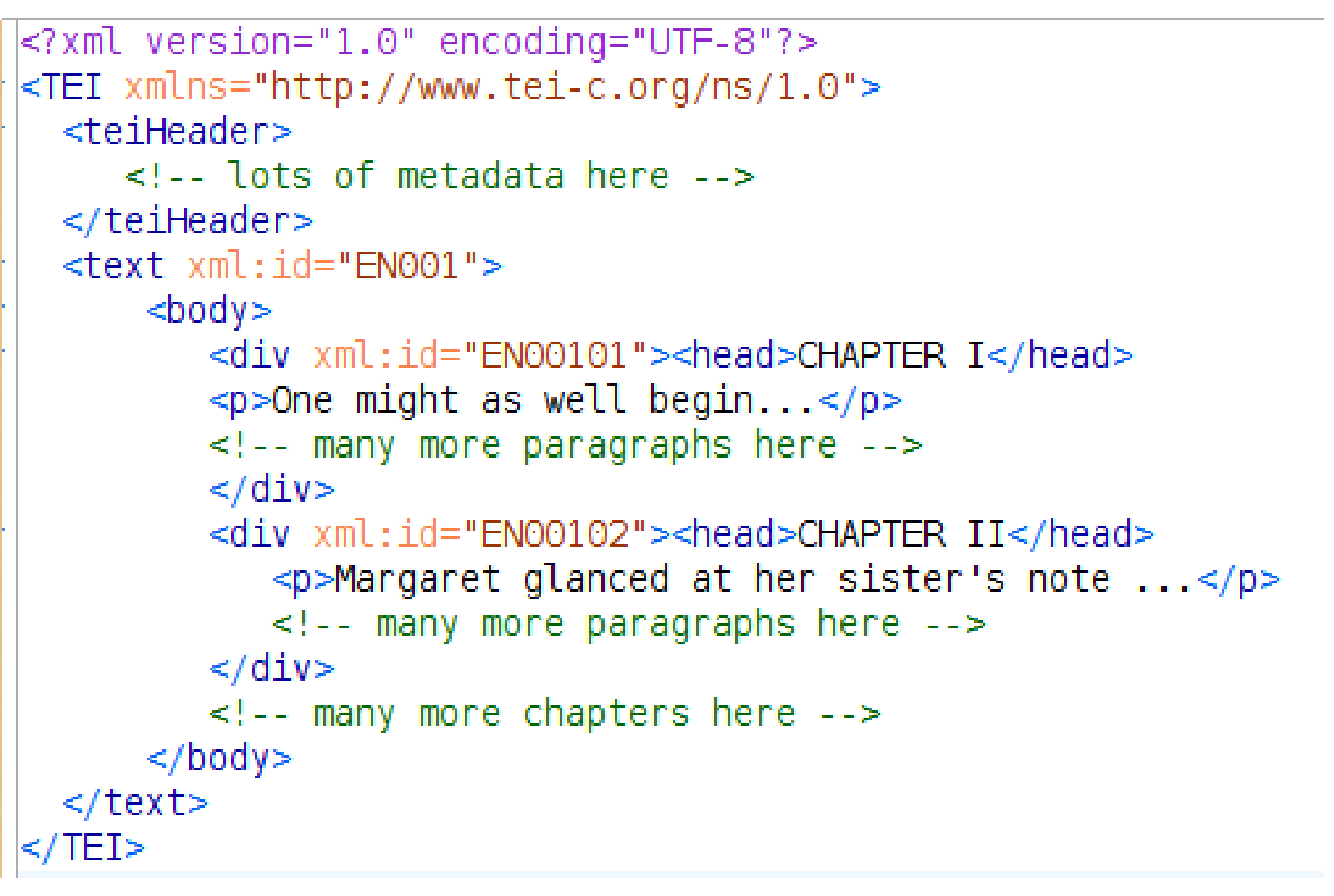
CHAPTER I

One may as well begin with Helen's letters to her sister.

Howards End,
Tuesday.

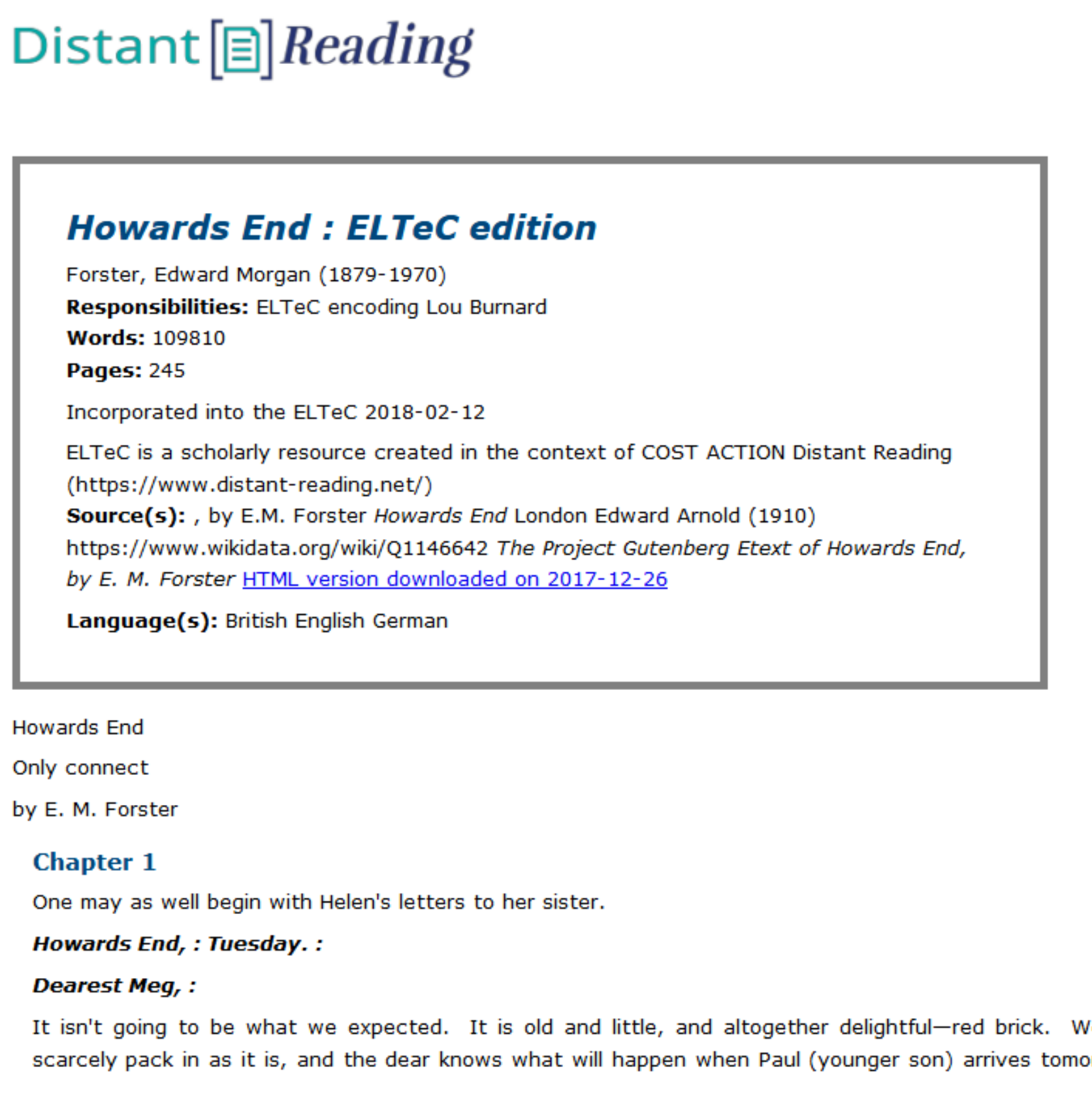
Dearest Meg,

It isn't going to be what we expected. It is old and little, and altogether delightful—red brick. We can scarcely pack in as it is, and the dear knows what will happen when Paul (younger son) arrives tomorrow.



```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
     <!-- lots of metadata here -->
  </teiHeader>
  <text xml:id="EN001">
      <body>
         <div xml:id="EN00101"><head>CHAPTER I</head>
         <p>One might as well begin...</p>
         <!-- many more paragraphs here -->
         </div>
         <div xml:id="EN00102"><head>CHAPTER II</head>
            <p>Margaret glanced at her sister's note ...</p>
            <!-- many more paragraphs here -->
         </div>
         <!-- many more chapters here -->
      </body>
  </text>
</TEI>
```

Workflow: Starting with digitized texts, manual or computed-aided trancription (OCR), encoding basic text features in TEI XML (TEI Consortium 2019).



Distant Reading

Howards End : ELTeC edition

Forster, Edward Morgan (1879-1970)
Responsibilities: ELTeC encoding Lou Burnard
Words: 109810
Pages: 245

Incorporated into the ELTeC 2018-02-12

ELTeC is a scholarly resource created in the context of COST ACTION Distant Reading (https://www.distant-reading.net/)
Source(s): , by E.M. Forster *Howards End* London Edward Arnold (1910) https://www.wikidata.org/wiki/Q1146642 *The Project Gutenberg Etext of Howards End, by E. M. Forster* HTML version downloaded on 2017-12-26

Language(s): British English German

Howards End

Only connect

by E. M. Forster

Chapter 1

One may as well begin with Helen's letters to her sister.

Howards End, : Tuesday. :

Dearest Meg, :

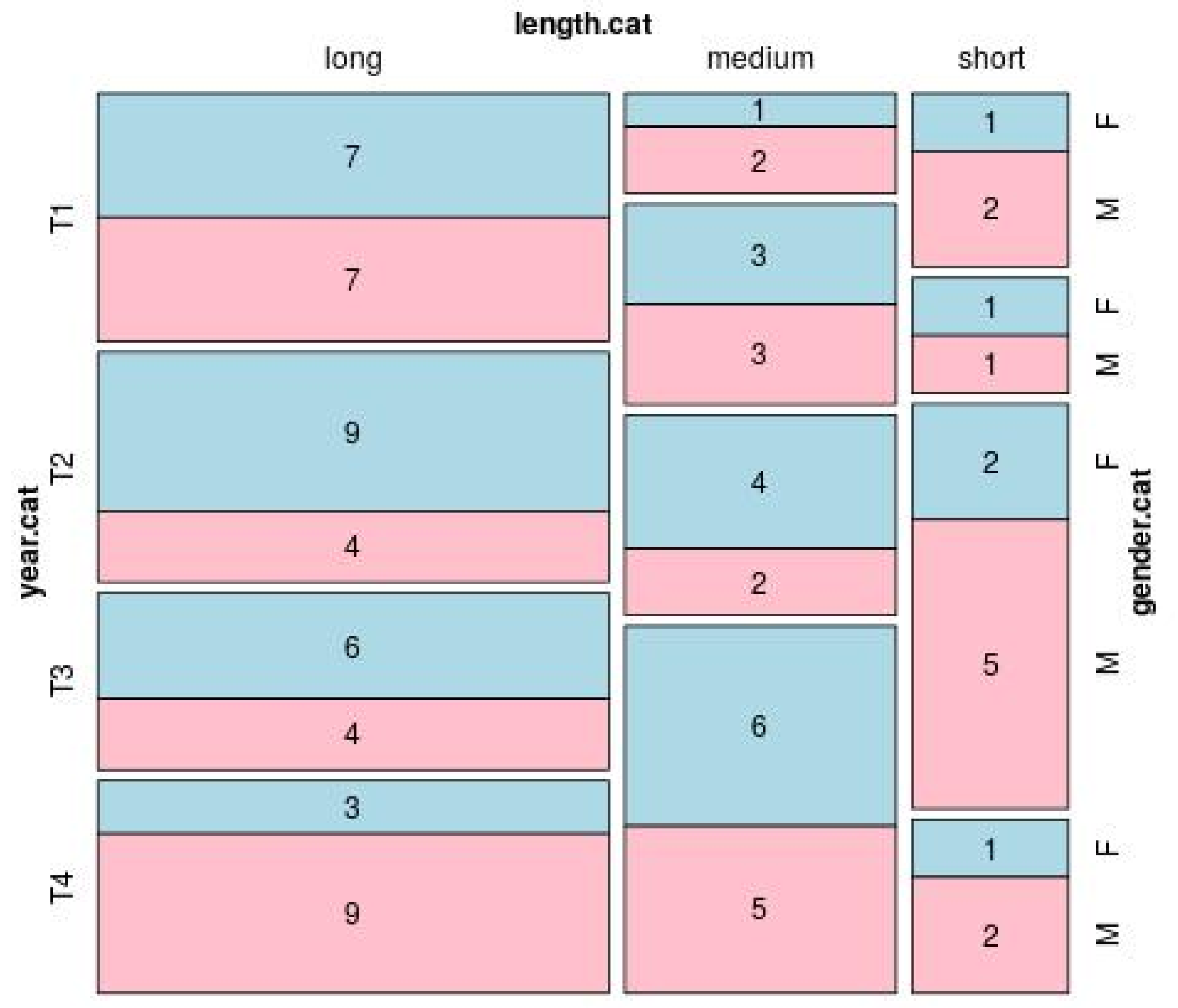
It isn't going to be what we expected. It is old and little, and altogether delightful—red brick. We can scarcely pack in as it is, and the dear knows what will happen when Paul (younger son) arrives tomorrow.

Display: Using TEI markup for css-based transformation processes to provide text display.

Future work

Corpus development: aim to get 100 texts per language according to sampling guidelines

Title counts for each balance criterion



Monitoring: The figure shows the current composition of the English language collection.

Next steps:

- Training school with three parallel tracks on corpus data, methods and theory, co-located with the DH Budapest 2019 conference
- adding more texts to language collections
- release of revised encoding schemas, release of encoding schema level2
- a first publication of ELTeC on Zenodo in 2019

References and Notes

Algee-Hewitt, Mark und Mark McGurl (2018). *Between canon and corpus: six perspectives on 20th-century novels*. URL: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Biber, Douglas (1993). „Representativeness in Corpus Design". In: *Literary and Linguistic Computing* (8), S. 243–257.

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

Herrmann, Leonhard (2011). „System? Kanon? Epoche?" In: *Kanon, Wertung und Vermittlung. Literatur in der Wissensgesellschaft*. Hrsg. von Claudia Stockinger Matthias Beilein und Simone Winko. Berlin: De Gruyter, S. 59–75.

TEI Consortium (2019). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. URL: http://www.tei-c.org/Guidelines/P5/ (besucht am 21. 06. 2019).

Winko, Simone (1996). „Literarische Wertung und Kanonbildung". In: *Grundzüge der Literaturwissenschaft*. Hrsg. von Heinz Ludwig Arnold und Heinrich Detering. München: Deutscher Taschenbuch Verlag, S. 585–600.

(1) Each canon is a result of rating texts from different perspectives: intellectual, economical, or/and reader rating (a.o. Herrmann 2011; Winko 1996).
(2) For discussion of representativeness Biber (1993) and canonicity and corpus design Algee-Hewitt und McGurl (2018) und Bode (2018).
(3) Design criteria https://distantreading.github.io/sampling_proposal.html.
(4) Schemas and documentation https://github.com/distantreading/WG1/wiki/textFeatures.

Acknowledgments and Funding

This poster describes the COST Action Distant Reading for European Literary History (CA16204 - Distant-Reading). More information: http://www.distant-reading.net. COST is funded by the Horizon 2020 Framework Programme of the EU.