COST Action Distant Reading for European Literary History

Corpus design and text contribution for ELTeC

Distant [目] Reading

Lou Burnard, Carolin Odebrecht, Christian Reul and Martina Scholger

Introduction

- metadata for novels in ELTeC
- introduction to metadata schema
- metadata collection with world cat
- encoding metadata in TEI schema

Outline

Outline

Working Group 1: Scholarly Resources

- Creating an open source multi-lingual benchmark corpus for European literature of the 19th century (novels): European Literary Text Collection (ELTeC)[1]
- 34 Members from 22 countries
- Main tasks are
  - defining corpus design
  - developing basic encoding schemas
  - developing workflows

---

[1]https://www.distant-reading.net/wg-1/

Corpus design

Corpus design defines two things (cf. Hunston 2008; Lüdeling et al. 2016):

- candidates → sampling
  - Which text(s) can be included in the corpus? Which can't?
- proportion → balancing
  - How many texts with which characteristics should the corpus contain?

Corpus design – approach of WG1

- Sampling and balancing criteria[2] will
    - not define what a novel is (cf. WG 3)
    - follow a non-normative but metadata-based approach (not canon-based)[3]
    - aim to represent the variety of a population[4]
    - allow for a comparability of texts and individual sub-collections according to different metadata set(s)

---

[2] https://github.com/distantreading/WG1/blob/master/sampling_proposal.xml

[3] Each canon is a result of rating texts from different perspectives: intellectual, economical, or/and reader rating (a.o. Herrmann 2011; Winko 1996).

[4] Cf. for discussion of representativeness Biber (1993) and canonicity and corpus design Algee-Hewitt and McGurl (2018) and Bode (2018).

Sampling criteria

- **language**: European languages, no translations
- **prose**: 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

Balancing criteria

100 texts per language (language collection)

- **period**: distribution over time
  - group T1: 1840-1859
  - group T2: 1860-1879
  - group T3: 1880-1899
  - group T4: 1900-1920
- **gender**: min. 10% and max. 50% written by female authors
- **authorship**: 9 - 11 authors with exactly three novels (otherwise, only one text for each author)
- **length**: min. 20% are short novels (10-50k word tokens), min. 20% are long novels (>100k word tokens).
- **reprint**: min. 30% are highly canonized novels, min. 30% should be non-canonized novels, based reprint counts within the period 1970-2009 (work in progress)
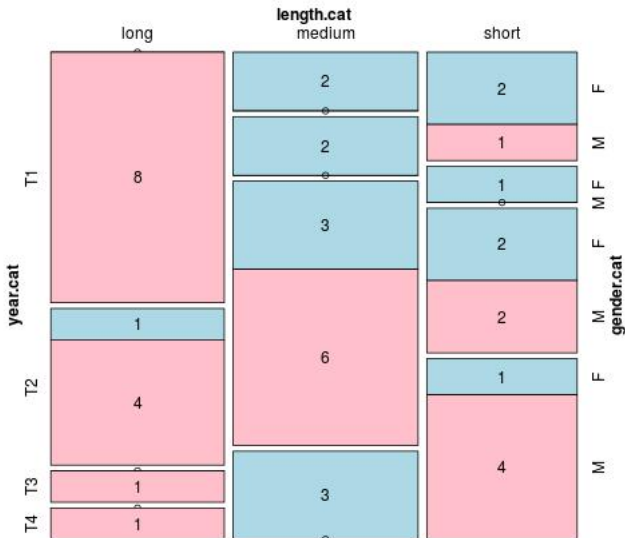
(ideal) Composition

Current composition of the language collection

Amount of texts with balancing categories (year, length, gender)

Example: German Text Collection

Title counts for each balance criterion

Outline

TEI encoding for ELTeC

Sources:

- ▶ different texts levels / editions (cf. e.g. IFLA 2009)
- ▶ different levels of digitization/annotation
- ▶ various languages, encodings, fonts

Corpus:

- ▶ minimal, uniform encoding (applicable for different sources)
- ▶ text representation focus on consistency and simplicity (not full complexity/richness needed for digital editions)
- ▶ basic metadata and references (not replicating work of libraries)
- ▶ serves as master file container for automatically generating transformations, annotations and derivations

TEI encoding for ELTeC

Sources:

- ▶ different texts levels / editions (cf. e.g. IFLA 2009)
- ▶ different levels of digitization/annotation
- ▶ various languages, encodings, fonts

Corpus:

- ▶ minimal, uniform encoding (applicable for different sources)
- ▶ text representation focus on consistency and simplicity (not full complexity/richness needed for digital editions)
- ▶ basic metadata and references (not replicating work of libraries)
- ▶ serves as master file container for automatically generating transformations, annotations and derivations

Encoding levels for ELTeC

TEI ODD-chain[5] provides different levels of encoding:

- ▶ **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
- → consistent metadata for each level of encoding

---

[5]Cf. TEI ODD Burnard and Rahtz (2004), for documentation of ELTeC ODD https://github.com/distantreading/WG1

Metadata schema

tba

TEI document, encoding level 1

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="de" xml:id="deu032">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Sibilla Dalmar: ELTeC Edition</title>
        <author ref="http://d-nb.info/gnd/pnd:11852643X">Dohm,
          Hedwig (1831-1919)</author>
        <respStmt>
          <resp>ELTeC conversion</resp>
          <name>Leonard Konle</name>
        </respStmt>
      </titleStmt>
      <extent>
        <measure unit="words">97252</measure><measure unit="pages"
          >374</measure>
      </extent><publicationStmt><p>Veröffentlicht als ein Teil von ELTeC <date>2

      <sourceDesc>
        <!-- converted from TextGrid (textgrid:mh3s.0) via Kallimachos: +71+ --
        <bibl type="copyText">
          <title>Sibilla Dalmar</title>
          <author>Dohm, Hedwig</author>
          <date>1896</date>
        </bibl>
      </sourceDesc>
    </fileDesc>
    <encodingDesc n="eltec-1">
```

6

Outline

Research for metadata

References I

Algee-Hewitt, Mark and 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), pp. 243–257.

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

Burnard, Lou and Sebastian Rahtz (2004). *RelaxNG with Son of ODD*. Extreme Markup Languages. URL: http://www.tei-c.org/cms/Talks/extreme2004/paper.html (visited on 01/08/2017).

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

Hunston, Susan (2008). "Collection strategies and design decisions". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 154–168.

IFLA (2009). *Functional Requirements for Bibliographic Records*. URL: http://www.ifla.org/publications/functional-requirements-for-bibliographic-records (visited on 12/23/2016).

Lüdeling, Anke, Julia Ritz, Manfred Stede, and Amir Zeldes (2016). "Corpus Linguistics". In: *OUP Handbook of Information Structure*. Ed. by Caroline Fery and Shinishiro Ishihara. Oxford: Oxford University Press, pp. 599–617.

Winko, Simone (1996). "Literarische Wertung und Kanonbildung". In: *Grundzüge der Literaturwissenschaft*. Ed. by Heinz Ludwig Arnold and Heinrich Detering. München: Deutscher Taschenbuch Verlag, pp. 585–600.