COST Action Distant Reading for European Literary History

European Literary Text Collection (ELTeC)

Distant Reading

Christof Schöch, Carolin Odebrecht, Lou Burnard, Borja Navarro-Colorado et al.

TS Budapest Track 1, Corpus design and text contribution for ELTeC

TS – Organisation, schedule and data

- Schedule
- Slides
- Training data
- Corpus data

TS – Track 1

- Introduction to ELTeC
- Metadata in the teiHeader
- Optical character recognition
- Hands-on experience in creating ELTeC TEI-XML versions of source texts
- Start from scanned page images or from a pre-existing HTML version
- TS output: contribute new TEI encoded texts to the ELTeC GitHub repository

Outline

1. COST Action Distant Reading and Working Group Scholarly Resources

2. ELTeC Corpus Design

3. WG Workflow

COST[2] Action Distant Reading

- Christof Schöch, University of Trier
- CA16204 will
  - „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“[1]
  - started autumn 2017
- Working groups
  - WG 1: Scholarly Resources
  - WG 2: Methods and Tools
  - WG 3: Literary Theory and History
  - WG 4: Dissemination

[1] www.distant-reading.net

[2] European Cooperation in Science and Technology = COST[3]

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)[4]
- 34 Members from 22 countries
- Main tasks are
  - defining corpus design
  - developing basic encoding schemas
  - developing workflows

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

Working Group 1: Scholarly Resources

Context of the WG:

- closely related to WG2 Tools and methods and WG3 Theory, cf. parallel tracks of the TS
- aims to support many different perspectives on distant reading
- many members with different scientific background and experience in data creation
- enabling distant collaborative work on creating ELTeC

Outline

1. COST Action DISTANT READING and Working Group SCHOLARLY RESOURCES

2. ELTEC Corpus Design

3. WG Workflow

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[5] will
  - not define what a novel is
  - follow a non-normative but metadata-based approach (not canon-based)[6]
  - aim to represent the variety of a population[7]
  - allow for a comparability of texts and individual sub-collections according to different metadata set(s)

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

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

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

Sampling criteria

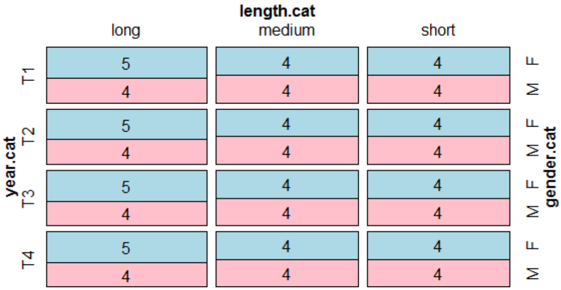
- **language**: European languages, no translations
- **prose**: narrative fictional prose
- **period**: 1840–1920
- **length**: min. 10.000 words
- **publication**: prefer books over novels published in serial publications
- **access**: only freely available digitizations

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 (rarely or never reprinted), based reprint counts within the period 1970-2009 ("unmarked", information not yet available)

(ideal) Composition



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

ELTeC summary `https://distantreading.github.io/ELTeC/`

Outline

1. COST Action Distant Reading and Working Group Scholarly Resources

2. ELTeC Corpus Design

3. WG Workflow

WG workflow

- WG1 on GitHub[8]
    - organization and documentation
- ELTeC on GitHub[9]
    - data and scripts
- Archiving on Zenodo[10]

[8] https://github.com/distantreading/WG1

[9] https://github.com/COST-ELTeC

[10] https://zenodo.org/communities/eltec/

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.

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.

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.