Towards modeling the European novel

Introducing ELTeC for Multilingual and Pluricultural Distant Reading

J. Berenike Herrmann, Carolin Odebrecht, Diana Santos & Pieter Francois

DH2020

Outline

1. COST Action Distant Reading and ELTeC
2. Collection design
3. Challenges
4. An empirical characterization of 'the novel'

Outline

Distant reading for European literary history

Distant [📄] *Reading*

- Chair: Christof Schöch (Trier)
- Vice-Chair: Maciej Eder (Kraków)
- `www.distant-reading.net`
- COST action 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"

Organization

- Working groups
  - WG 1: Scholarly Resources
  - WG 2: Methods and tools
  - WG 3: Literary Theory and History
  - WG 4: Dissemination

34 members from 22 countries

 [1]

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

ELTeC - European Literary Text Collection

- An open source multi-lingual benchmark corpus for European literature
- Aim: 2,500 full-text novels (covering at least 10 different languages)

Main tasks of WG1[2]

- Defining corpus design
- Developing basic encoding schemas (XML-TEI)
- Developing workflows

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

Outline

Collection design

- Definition of text characteristics criteria
  - "What *kinds* of texts do we choose?"
- Definition of text selection criteria
  - "Which *particular texts* do we choose?"
- Definition of proportion criteria
  - "*How many* texts with which characteristics do we choose?"
- All of these criteria are challenging
  - European perspective
  - Different traditions and contexts

Contents of ELTeC

- Goal for creating ELTeC: Faciliate the application of distant reading methods for data creation and analysis
- Digitized and annotated European novels of the 19th and early 20th century
- Uniform sampling and balancing criteria: length, publication date, author gender, & reprint count (1970-2010)
- Basic encoding to facilitate distant reading: Uniform and consistent encoding schemas in TEI XML
- Currently working on Czech, English, French, German, Greek, Hungarian, Italian, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovenian, Spanish and Swedish

ELTeC sampling criteria

- Presented/assumed as a 'novel' (or corresponding term)
- In original language (no translations)
- Narrative fictional prose (no memoirs, no poetry, no drama)
- Date of publication of first edition: 1840-1920
- Minimal length: 10,000 words
- Publication form: avoid novels published in serial form if possible
- Access: only works in the public domain, freely available

ELTeC proportion criteria

- Size: at least 20% should be over 100,000 words and 20% under 50,000 words
- Author gender: at least 10%, and not more than 50%, written by women
- Four temporal bins: 1840-1859; 1860-1879; 1880-1899; and 1900-1920
- Different reprint status: used reprint count in print form (no ePubs etc.) 1970-2010. At least 30% with one reprint, at least 30% with no reprint
- Authorship: per collection 9-11 authors with each three works

ELTeC data management

- Data creation and update on GitHub, `https://github.com/COST-ELTeC`
- Encoding schema developed and documented with TEI
  - ODD: `https://github.com/distantreading/WG1/`
  - schema: `https://github.com/COST-ELTeC/Schemas`
- TEI Documentation also on GitHub, `https://distantreading.github.io/ELTeC/index.html`
- Persistent referencing and archiving on Zenodo, `https://zenodo.org/communities/eltec/`
- Free licence to foster re-usability: CC-BY 4.0

Markup with TEI in ELTeC

- Our markup is *not* meant to ...
  - represent texts in all their original complexity
  - duplicate the work of scholarly editors
- Our markup *is* meant to...
  - facilitate a rich and well-informed distant reading
  - provide more information than a transcription of lexical content alone (plain text)
- Three encoding levels (via ODD chaining)

  level0: basic encoding

  level1: richer encoding

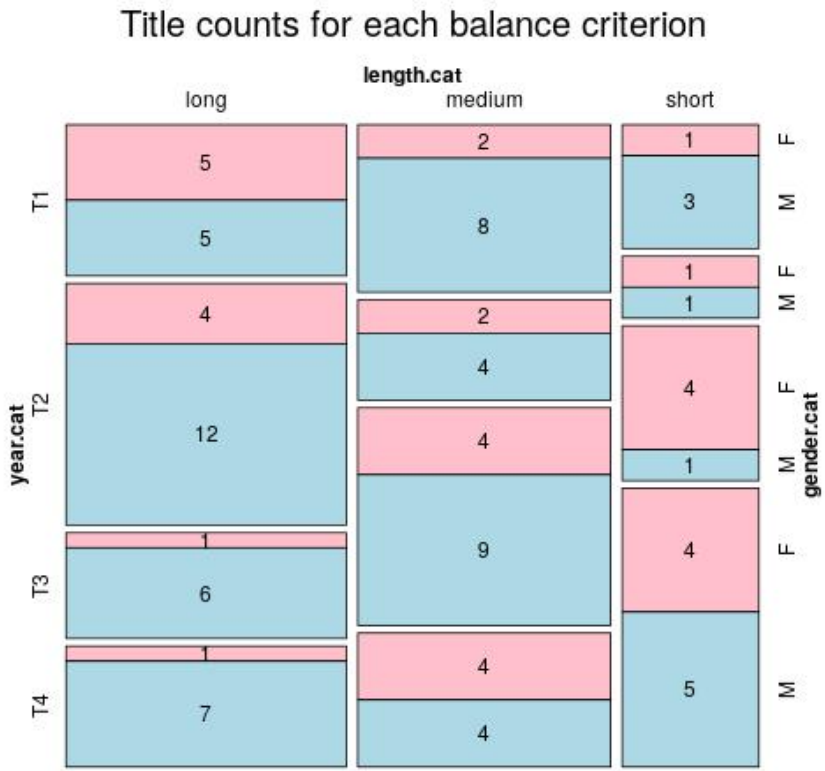
  level2: tokenization and linguistic annotation (work in progress)

  - See Burnard, Schöch and Odebrecht (2020) for more specific information[3]

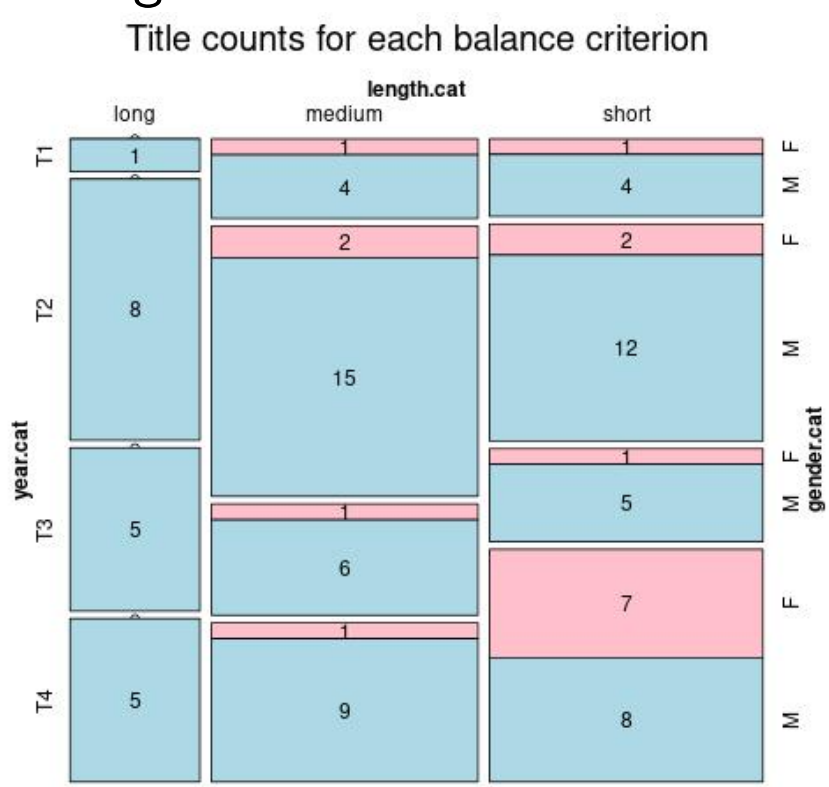
[3] `http://gams.uni-graz.at/context:tei2019`

Bird's eye view of some collections

German



Portuguese



Mosaic plots done by Carolin Odebrecht and Lou Burnard, available from `https://distantreading.github.io/ELTeC/`

Outline

Challenges when creating the collections

- Variation
  - Publication histories in Europe
  - Literary scholars, literary histories and traditions
  - Accessibility of information and resources
- Almost never "one language, one literature, one country"
  - Colonial/imperial languages: novel needs to be published in Europe in 10 years after the 1st edition
  - Some languages, like German, have several countries in Europe with native writers
  - Some countries, like Norway, have several languages or writing practices
  - Some countries were created, or literatures begun, after 1840
  - Political factors in the 20th century tend to influence reprinting

Practical challenges as well

- Different starting points for data creation, e.g.:
  - Printed book
  - Digitized book (image)
  - Plain text
  - Previously encoded data set (German Text Archive, Gutenberg...)

  –> may correspond to objects with highly different orthographies
- Metadata describing the digital or/and analogue source(s)
  - Library catalogs
  - Online databases for texts, ebooks, corpora

  –> may be missing, sketchy, or unreliable

Promises of ELTeC

The above challenges are offset by the promise of:

- Ability to contextualise one national literature in Europe-wide trends
- Assess how similar or unique issues of a balanced sample are across national cultures
- Share solutions cross-culturally
- Use the collections for multilingual and pluricultural Distant Reading

and ... ELTeC allowed us to understand the listed challenges much better!

The case of Swiss literature

Four official Swiss languages, and thus at least four different 'literatures.'

- ELTeC first of all 'language-'driven (not 'nationality-'driven)

Ernst Ludwig Kirchner (1880-1938): Rückkehr der Tiere, 1919

- Solution 1: small samples of Swiss novels incorporated each by German, French, and Italian ELTeC collections
- Solution 2: Building independent ELTeC-CH collections for Swiss novels in German, French, Italian (excluding dialect literatures)[a]

[a] See research project on the Swiss German novel, `https://mountain-sentiment.github.io/`

- Operational solution for determining nationality
  - Linked-Open Data - GND and VIAF (Virtual International Authority File)
  - Socialization ('went to school')

The case of literature in Portuguese

Portuguese as an imperial language spoken on the five continents (Brazil's independence relatively recent – in 1822)

Columbano Bordalo Pinheiro (1857-1929): O Grupo do Leão, 1885

- Novels written in Portuguese: not always European authors, even though published in Portugal
- Brazilian literature: extremely high price of paper in Brazil, thus common that Brazilian authors published in Lisbon (or Paris)

- Brazilian / Portuguese literature was 'common', including literary (and linguistic) discussions across the Atlantic
- Requirements of uniformity within the COST action: included only Portuguese citizens as authors

The case of Norwegian literature

Theodor Kittelsen (1857–1914): Kvitebjørn kong Valemon, 1912

- Norwegian in period 1840 - 1920 changed from Danish to a *Norwegianized* Danish (*Bokmål*) and Norwegian written language (*Nynorsk*).
  - *Bokmål* and *Nynorsk* have since continued to develop into the current standards.
  - Challenge 1: two languages, which are languages "in flux"
  - Challenge 2: our taggers and other tools are adapted for *modern* Norwegian: need for adaption for "*ELTeC* period"
- See Ore et al. (2020)

Outline

An empirical characterization of 'the novel'

Emerging questions about...

- form and structure
  - number and length of chapters
  - paratexts (titles and subtitles, epigraphs, whole first pages)
  - citations in other languages (measure of 'cosmopolitism')
  - etc.
- plot and fictional worlds
  - historical characters or events (e.g., Napoleon, First World War)
  - fictional characters (Othello, Penelope)
  - locations (non-fictional and fictional)
  - professions
  - nations/nationalities/demonyms
  - etc.

Preliminary data on named entities in ELTeC

- A manually annotated NER collection from ELTeC in 9 languages, with categories devised for literary questions (persons, roles (professions and titles), locations, demonyms, works, events and other) was created (Stanković et al 2019)
- Preliminary studies on evaluating off-the-shelf named entity recognizers on English, French, Portuguese and Serbian for person names and locations show that precision and recall are much lower than the usual NER results for modern newspaper text. (Frontini et al. 2020)

Next steps

- Exploratory study on ELTeC's titles (and sub titles) (Odebrecht et al., in prep.)
- Adding more and more novels to ELTeC
- Designing an encoding schema for tokenized and annotated representations of the novels

Conclusions

- Multilingual and pluricultural Distant Reading
  - "Who can read the entire ELTeC?" :-)
  - Tools for tokenization, part of speech, named entity recognition, sentiment analysis
  - Publication as epubs for close reading
- ELTeC is NO! Representative! Corpus!
  - Comparative and exploratory
  - Best practices (documentation, operationalization, TEI-ELTeC)
  - Fertilize research within particular 'national literary historiographies'
- Generate
  - New avenues for theory development (the novel, (trans-) national literatures, titles, etc.)
  - Towards standards across languages: Domain-adapt modern tools (literary & historical languages; NER, SA) (Frontini et al. 2020)
  - Possibly scale up ... 'WoLTeC' and/or representative collections

Acknowledgements

Thank you! Grazie! Obrigada! Merci! Takk! Danke schön! Gracias! Hvala! Mulțumiri! Ačiū! Dziękuję! Dank u wel!

References

- Burnard, Lou, Christof Schöch & Carolin Odebrecht. “In search of comity: TEI for distant reading”, *TEI 2019*, `https://gams.uni-graz.at/o:tei2019.106`
- Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos & Ranka Stanković. “Named Entity Recognition for Distant Reading in ELTeC”, *CLARIN Conference 2020*.
- Odebrecht, Carolin, et al. “Thresholds to the “great unread:” titling practices in eleven ELTeC collections”. In preparation.
- Ore, Christian-Emil, J. Berenike Herrmann, Carolin Odebrecht & Diana Santos. "ELTeC: a comparable corpus of novels in many European literatures". Accepted abstract to DHN2020, Riga, 2020.
- Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec & Carmen Brando. “Named entity recognition for Distant Reading in several European literatures". *DH Budapest 2019*.