COST Action Distant Reading for European Literary History

WG1 Meeting

Distant Reading

18/02/2020 Málaga

Outline

Schedule Tuesday

9:00-10:30	Session 1A: WG1 meeting ELTeC
10:30-11:00	Coffee break
11:00-12:30	Session 2A: WG1 and WG2: Annotation Schema Level 2
12:30-14:00	Lunch break
14:00-15:00	Session 3A: WG1 and WG4: Dissemination - organisation and workload
15:00-15:30	Coffee break
15:30-16:30	Session 4A: WG 1 and WG3: Literary Research with ELTeC
16:30-18:00	Joint final session: Publishing ELTeC Corpora using TextGrid and/or DraCor, Susan Schreibman on dariahTeach

See here for schedule of entire meetings

Schedule Wednesday

9:00-10:30	Tutorial "Walk through ELTeC"
10:30-11:00	Coffee break
11:00-12:30	Workshop session 2: Joint session on how to encode direct speech information in ELTeC
12:30-13:00	Farewell

Documention

Please use
https://semestriel.framapad.org/p/9f29-wg1malaga2020
for notes during the meeting.
On https://github.com/distantreading/WG1/tree/master/MalagaMeeting2020 we store these notes and the slides of our meeting.

Session WG1 and ELTeC

- State of the play
- “Missing” collections
- Goals for GP 4

Session WG2 and WG1 Annotation level 2

- Discussion and decision on WG2 proposal
- https://github.com/mikekestemont/speechies/tree/master/new/samples_correct_tei

Session WG4 and WG1 Dissemination

- Publication and dissemination strategies
- Responsibilities and access in WG1/WG4
- WG1 Meeting in June (Berlin)

Session WG3 and WG1 Literary Research

- Sampling criteria used in ELTeC and studying literary history and theory

Walk through ELTeC

The "Walk through ELTeC" tutorial introduces the design, structure and access points of ELTeC to all members of the Action. We start with the facsimile and present the workflow for data preparation, creating collections and present existing documentation and presentations. Finally, we present a small exploratory study of the novel titles in ELTeC to provide a first impulse for own studies and research.

COST Action Distant Reading

 [1]

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

WG1 Scholarly Resources

- Creating an open source multi-lingual benchmark corpus for European literature: European Literary Text Collection (ELTeC)[2]
- Lou Burnard and Borja Navarro-Colorado WG-Vice leads
- (currently) 35 Members of 23 countries
- Main tasks are
  - defining corpus design,
  - developing basic encoding schemas,
  - developing workflows.

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

ELTeC

- Digitized and annotated European novels of the 19th century
- Uniform sampling and balancing criteria
- Uniform and consistent encoding schemas in TEI XML
  - Basic encoding to facilitate distant reading
  - Applicable for different languages
  - Currently working on English, German, French, Spanish, Italian, Romanian, Slovenian, Polish, Hungarian, Portuguese, Serbian, Greek, Norwegian, Czech

MoU

WG1: Project Month 27

- initial goal: 2,500 full-text novels in at least 10 different languages: Dutch, English, French, German, Modern Greek, Italian, Polish, Portuguese, Russian and Spanish
- Additional languages as follow-up iterations
- Benchmark corpus for distant reading methods (!)

MoU

ID	month	description
D-1-1	4	Expert Meeting (EM) to agree on a basic common framework (data and metadata) for the creation of the ELTeC (in support of D-1-2, Guidelines)
D-1-2	7	Guidelines (version 1) on a common framework (data and metadata formats) for the creation of the first iteration of subcollections of the ELTeC.
D-1-3	9	Training School (TS) on corpus building, including data and metadata standards and requirements for linguistic and literary research.
M-1-1	12	The ELTeC, first iteration: subcollections for 6 languages.
D-1-4	18	EM to agree on common framework for subcollection annotation (in support of D-1-5, Guidelines; also involving WG 2)
D-1-5	18	TS on building corpora generally and contributing to the ELTeC specifically, including using the linguistic annotation framework.
D-1-6	21	Guidelines (version 2) elaborating on the common framework for subcollection creation, including shared framework for linguistic annotation.
M-1-2	24	The ELTeC, 2nd iteration: subcollections in at least 4 additional languages.
M-1-3	32	The ELTeC, 3rd iteration: at least 6 expansion subcollections.
D-1-7	37	EM on data and metadata quality requirements for linguistic & literary research (with WG 2 and library experts; for D-1-9)
D-1-8	37	TS on cross-language linguistic annotation and strategies for compatibility.
D-1-9	42	White paper on full-text data and metadata requirements directed at resource providers such as digital libraries (with WG 2)

Figure: Timeline for WG1. MoU online

First releases on Zenodo

- ELTeC Version 0.5.0, 11/2019
- nine languages with at least 20 texts
- German, English, French, Italian, Norwegian (Bokmål and Nynorsk), Portuguese, Romanian, Serbian, and Slovenian. The collections can be downloaded here: https://zenodo.org/communities/eltec/
- Spanish collection got it's first release recently!

Goals

- Add novels on existing collections and start new ones
- Curate the collections
  - Validation
  - Proportion (and text selection)
- Start working with extension
- Next release planned Nov 2020

Note on organisaton in GitHub

Lou has introduced a “bounced” folder to hold files which would otherwise break the build process because they are invalid XML.

Dissemination

- Please write WG4 if you got any publications on our WG topics or COST Action related topics which should be listed here: https://www.distant-reading.net/publications-and-presentations/

Research data management for ELTeC

- Data creation and update on GitHub[3]
- Encoding schema developed and documented with TEI ODD[4]
- Data and workflow documentation on GitHub[5]
- Persistent referencing and archiving on Zenodo[6]
- Free licence to foster re-usability: CC-BY 4.0[7]
- Further dissemination strategies are currently evaluated

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

[4] ODD https://github.com/distantreading/WG1/ and schema https://github.com/COST-ELTeC/Schemas

[5] https://github.com/distantreading/WG1/wiki

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

[7] https://creativecommons.org/licenses/by/4.0/

ELTeC – selection 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

ELTeC – proportion criteria

- 100 texts per language (language collection)
- period: distribution over time
  - T1: 1840-1859
  - T2: 1860-1879
  - T3: 1880-1899
  - T4: 1900-1920
- gender: min. 10% and max. 50% have been written by female authors for the language subcollection
- authorship: 9 - 11 authors with exact three novels
- 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

Language collection reports

- Could you report briefly on the current state of your language collections?

ELTeC – current state

Overview on ELTeC Language Collections:
https://distantreading.github.io/ELTeC/index.html

Metadata composition plot

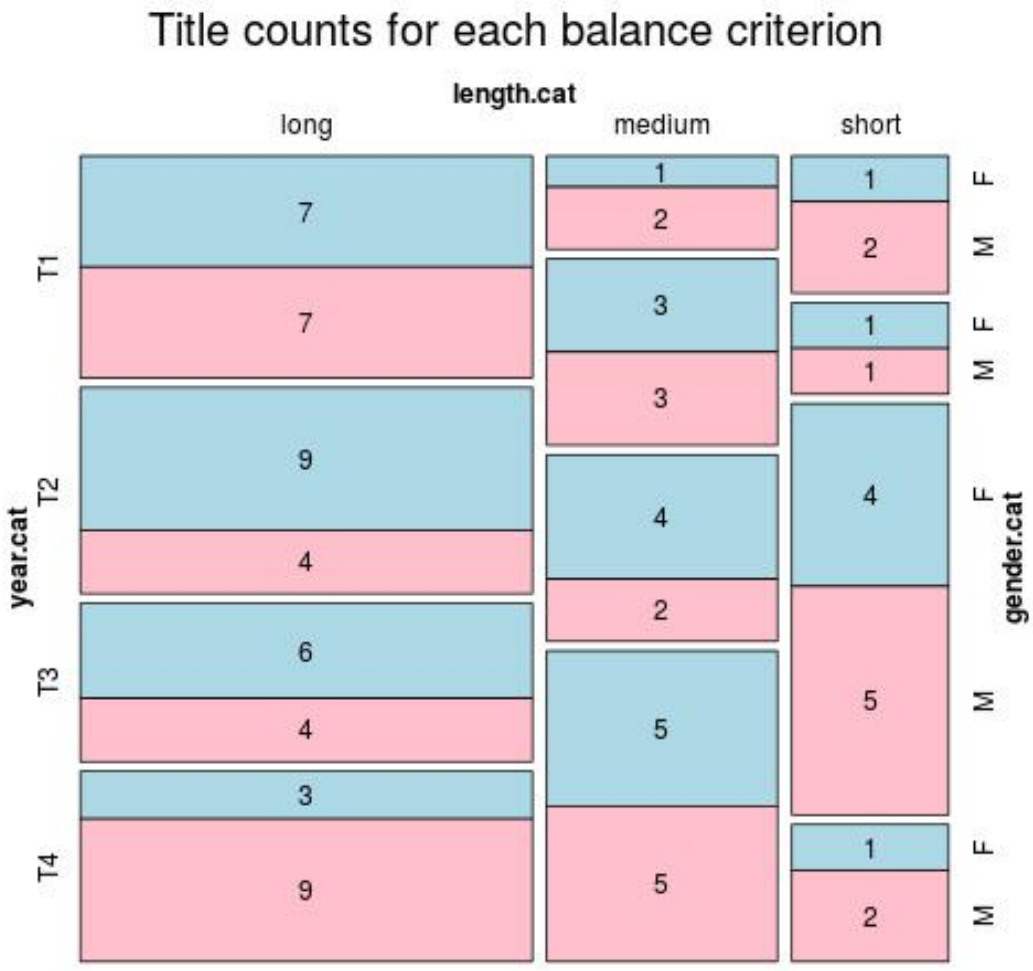

Figure: ELTeC-eng: Metadata in `teiheader` are parsed for each encoded file. Data is aggregated and visualized for corpus monitoring. Produced with ELTeC metadata and R package vcd by David Meyer [aut, cre], Achim Zeileis [aut], Kurt Hornik [aut], Florian Gerber [ctb], Michael Friendly [ctb][9].

[8] https://cran.r-project.org/web/packages/vcd/index.html

Proportion

Some general (not specific) issues

- hun: to many authors with more than three texts (e.g. Jókai Mór)
- cze: long novels are missing
- gre: all novels below 10.000 words
- lit: long and medium novels are missing
- nor: long novels are missing, gender is in balanced (short only male, medium only women?)
- rom: long novels are missing
- ...

Selection

- “Missing collections”: Dutch, Modern Greek, Polish, Swedish

Corpus data

- Different starting points for data creation, e.g.:
  - Exemplar of the book
  - Digitized book
  - Plain text
  - Previously encoded data set
- Metadata describe the digital or/and analogue source(s) of the data set
  - Library catalogues
  - Online databases for texts, ebooks, corpora

Validation

- Are you familiar with validating XML against a schema?

Organisation

- Meeting in March 2020?
- Ideas for STSM, topics and people?!

- Discussion and descision on WG2 proposal
- https://github.com/mikekestemont/speechies/tree/master/new/samples_correct_tei

Organisation

Discussion about

- Requirements
- Workload
- Responsibilies
- etc.

of platforms, cf. Christof's slides

WG1+WG4 meeting

- Preparing for meeting in June (Berlin)
- Dissemination strategies – working and testing phase

Question of WG3

- I only received the questions shortly before our meeting. Therefore the following answers could not be coordinated with WG1.
- I hope that the answers can contribute to the understanding with the work of/with ELTeC and can also be a starting point for common considerations.

- Question: All the required eligibility and composition criteria cannot be met by a small literatures corpus (e.g., historically late appearance of the first novel, low quantity of literary producers and readers, long-lasting patriarchy, preference for medium-length novels, undeveloped book market), hence achieving balance according to all required criteria doesn't allow for an insight into the development of a particular literature (diachronic overview of the development of the form as well as identifying the moment when women authors appear, for example)
- Answer: Do follow the criteria (selection and composition). Uniformity is necessary for the application scenarios in the Action. Distant reading methods need for example long enough novels and novels from different length.

- Question: It has been mentioned that the novels that dont fit the sampling criteria will be moved to an 'expanded collection'. What is its role in the overall corpus if it doesn't meet the criteria and cannot be used?
- Answer:ELTeC core is definitely priorized – if you got by chance more than the required novels you can put it into the extension. It is not intended to focus on the extension. Cf. example

- Question: How can we analyse, through a genuinely comparative perspective and distant reading, the European novel in all its linguistic and historical variety (which has to include peripheral literatures, not only the core ones)?
- Answer: Cf. “Walk through ELTeC”
- Answer: Test your concepts and hypothesis on unseen data (?).
- Answer: Be creative. It is hard to create research questions for somebody else.

- Question: How to encode extended metadata on the authors sex/first language/place of birth, topics, narrators, PoV (for the latter: which theoretical approach is to be used to get consistent results Genette's? Mieke Bal's? How to indicate switching between multiple narrators?)
- Answer: Interesting idea. Do you already have a fixed metadata schema? First idea: We could add a this information to each novel as a documentation note. Who could actually assign the metadata to each novel?

- Question: Can we afford intensive and in-depth encoding of the core corpus (who could do this and with what funding)? Encoding should be made economically, only with regard to the planned DH analyses.
- Answer: We have designed a basic encoding which does not require much in level 0.
- Answer: Is this related to WG3? Or to WG1?
- Answer: Complex annotation will be created automatically by WG2, if any.

- As definitions of the novels are not identical across languages (and especially in comparison to the Anglophone literary tradition) and sometimes contain multiple, partly equivalent or synonymous terms (as in Slovenian povest vs. roman), which texts should be included as 'novels' in the corpus? All, or only 'novels proper' i.e. those which in that particular literary tradition meet the criteria set by WG1's white paper: those texts that meet at least 1 of the following 3 criteria:
  - a) textual: length (>10.000 words NB: the lower limit seems too low), prose, fiction, narrative structure
  - b) peritextual (the term novel (or its equivalent) in the title or subtitle of the text) and
  - c) contextual: the text is bibliographically listed with the UDC: 82-31 Novels. Full-length stories.
- Answer: First: a minimum length is required (we fix this in the criteria). Thanks for your comments!
- Answer: We do not define exactly what a novel is. We collect texts that are good candidates to definitions of novels across times, place and language. (Resource to enable rewriting literary history, cf. MoU).

- Question: Is canonicity now defined only by reprint number? Among best measures of canonicity we suggest the inclusion of a particular novelist in the secondary-school curriculum. Reprint number is also important and we agree with the revised scale (low $< 1/$ medium $= 1$ / high > 1 number of reprints). Need to confirm the period in which reprints qualify a novel as canonical (from which period onwards 1980? Explain why that period for all literatures? Is this flexible?
- Answer: No! We see it as one indicator (among many others) that can be operationalised.
- Answer: We are running for over two years. No changes can be made. Any period would be difficult.

- What about applying literary concepts on novels/canons to texts in ELTeC that were not objects of research before?
- Compare concepts and definitions from one scholar history to another own which is covered ELTeC?
- Cross–close-read novels that are familiar and unfamiliar in the research discourse?
- ...