ELTeC: A comparable corpus of novels in many European literatures

Digital methods allow for new, additional and complementary ways of analysing texts. Tapping into recent advances in Digital Humanities, we present our collaborative work on the design and creation of a European corpus of novels published between 1840 and 1920 for evaluating distant reading methods in our network COST Action Distant Reading (CA16204).

Since the end of the 1990s there has been a massive and increasing digitization of various genres of texts establishing huge collections of digital texts. Well known examples are the HATHI Trust[1] and Google Books[2]. National libraries across Europe have initialized similar initiatives, e.g. BNF's project Gallica[3] with 2.4 million texts and the more modest Bookshelf project of the Norwegian National Library[4] with at least 500,000 digitized books. There are many such digitization projects, although their size and quality vary from country to country.

Besides these efforts, it has not been possible to obtain 'ready made' texts for our project. In many cases it has been necessary to digitize texts found only in printed editions to comply with the selection criteria.

The aforementioned digitization projects tend to collect texts without previously defined sampling criteria, using an 'opportunistic' corpus design. Despite some problems, this approach has made first available a huge variety of unknown and forgotten texts which have not been studied since they are not a part of the canon. We apply Distant Reading (Moretti, 2013), and approach emancipating non-canonical literature for large scale analysis in a straightforward manner in our project, applying clear criteria for corpus construction. In our view, Distant Reading is necessarily complemented by Close Reading - they are situated at the ends of a methodological continuum that can be applied to digital text corpora for different research contexts. These methods allow for purpose-tailored access to corpus data through analysis and visualisation methods in literary studies.

With our approach, we combine a rigorous approach to corpus design with the application of a diversity of methods. We are a multidisciplinary group of European scholars who work together in a COST Action (www.cost.eu), called Distant reading for European Literary history (Distant-Reading). The COST Action (CA16204) was initiated in 2017 and will last for four years. Our COST Action Distant Reading has three main objectives[5]

1. build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages primarily from the period 1850 to 1920, permitting to test methods and compare results across national traditions;

2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe's multilingual literary traditions;

3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

In the following section of this paper, we will focus on the first objective and discuss the collaborative effort of building an open access multilingual corpus of European novels (the European Literary Text Collection - ELTeC). We present the work done so far within Working Group 1 'Scholarly Resources' of the COST Action Distant Reading. Specifically, we address the link between the practical and technical aspects of corpus design on the one hand and the theoretical discussion on computational modeling of literature across languages and cultures on the other. This

1 https://www.hathitrust.org/

2 https://books.google.de/

3 https://gallica.bnf.fr

4 https://www.nb.no/search

5 Cf. https://www.distant-reading.net/ and MoU https://e-services.cost.eu/files/domain_files/CA/Action_CA16204/mou/CA16204-e.pdf

means paying attention to differences and similarities across different literary theoretical paradigms when setting up the corpus as a resource for Distant Reading, addressing the active role of corpus design for periodization and canonization in European literary history.

Working Group 1 is responsible for the development of the corpus design, the encoding schema and the workflow for data creation, maintenance and publication. The working group consists of European researchers from 23 countries and reflects an extraordinary field of different research disciplines such as corpus linguistics, computer linguistics, literary studies, social sciences, library science and philological studies. This European international scientific team is able to build a corpus that allows a European perspective on the novel.

The multilingual European Literary Text Collection (ELTeC, Odebrecht, Burnard, Navarro Colorado, Eder & Schöch, 2019) is an open access (CC-By 4.0 or CC0) of European novels from the period from 1840 to 1920. This is a period with large cultural and language changes and a formative phase for many European languages. For example, in this period the Norwegian written languages underwent very large changes: from pure Danish to an adapted Norwegian written standard at one hand and an introduction of a parallel dialect based Norwegian written standard on the other. The period is interesting to study also from a lexical/lexicographic point of view and raise large challenges for the design of language analyzing tools like lematizers and morphological analyzers. ELTeC will be used as a benchmark corpus to evaluate distant reading methods, and to discuss and even challenge established literary systems and publication history of the novel.[6] We organise the corpus in languages collections covering Romance, Slavic, Germanic and Finno-Ugric language families. Currently, we include Czech, English, French, German, Greek, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovenian and Spanish novels.

In order to allow comparability among the different European novels and foster the interoperability of the corpus, we use the TEI to encode the digitized texts. The aim is not to make a collection of scholarly text editions nor of plain texts: “we aim to facilitate a richer and better-informed distant reading than a transcription of lexical content alone would permit” (Burnard, Schöch & Odebrecht, 2019). We are therefore not aiming to represent various text structural or graphical features of the text but to enable a uniform and consistent encoding across the language collections. We are using ODD chaining to handle three encoding levels (cf. Rahtz &and Burnard, 2013): 1) a basic level covering the body of the texts including e.g. paragraphs, headers, and highlighted items, 2) a richer level including gaps, notes and quotes and 3) a third level with lexical information.[7] In addition, we provide basic text displays for each novel.[8]

Building ELTeC as a unique resource requires to reassess and even discard some established ways of defining literary systems and publication histories (see MoU, https://e-services.cost.eu/files/domain_files/CA/Action_CA16204/mou/CA16204-e.pdf . In the corpus design, we thus maintain a metadata-based approach that allows for representing the diversity of novels published between 1840 and 1920 across the multilingual, transnational, and pluri-cultural topologies of Europe. At the most general level, this approach addresses common textual and contextual features instead of solely relying on canonical definitions of novels in literary history. We defined sampling and balancing criteria that use metadata such as publication date and place, text length, reprint counts and authors’ gender.

We especially include novels not previously incorporated in the literary canons of the European countries. In our approach, ‘canonization’ cuts across ‘popularization,’ operationalized in terms of reprints. We assume that the different types of canons - defined nationally or by language, incarnated in the form of educational syllabus policies and reading lists at schools and universities - correlate with a relative high number of book reprints, documented in library records.

---

6 https://www.distant-reading.net/

7 See https://tei-c.org/guidelines/customization/getting-started-with-p5-odds/

8 See https://distantreading.github.io/ELTeC/index.html for the current state of ELTeC and for a Portuguese example of a text display https://distantreading.github.io/ELTeC/por/POR0045.html.

Using the TEI for encoding data fosters interoperability. By using the ODD mechanism, we also set a focus on clear schema definitions and documentation. The data and metadata of ELTeC are created collaboratively via GitHub.[9] Our working group provides an open access extensive documentation for (meta-)data schema, decisions and workflows.[10] We archive versions of ELTeC via Zenodo.[11] Thus, our (meta)data are re-usable, interoperable, accessible and findable (cf. FAIR Guiding Principles Wilkinson et al., 2016)

In the creation of ELTeC we do not aim at deductively defining what a ‘novel’ is, but to allow for different approaches in literary theory and history to be inductively explored and tested. This is likely to entail a re-evaluation and redefinition of key concepts for literary history, including genre, style or authorship as well as a debate about the advantages as well as limitations of Distant Reading methodologies and approaches to the study of European literary history.

Acknowledgements

The research described in this paper was conducted in the context of the COST Action "Distant Reading for European Literary History" (CA16204 - "Distant-Reading"). Find out more at: http://www.distant-reading.net. COST is funded by the Horizon 2020 Framework Programme of the EU.

References

Burnard, Lou; Schöch, Christof; Odebrecht, Carolin (2019) In search of comity: TEI for distant reading. Book of Abstracts TEI Conference Graz 2019.

Moretti, Franco (2013) *Distant Reading.* Verso, London, ISBN: 978-1-78168-112-1

Odebrecht, Carolin; Burnard, Lou; Navarro Colorado, Borja; Eder, Maciej, & Christof Schöch. (2019). The European Literary Text Collection ELTeC. Zenodo. http://doi.org/10.5281/zenodo.3546326

Rahtz, Sebastian, and Lou Burnard (2013) “Reviewing the TEI ODD System”. In *Proceedings of the 2013 ACM Symposium on Document Engineering. DocEng’13*. ACM, 2013. http://doi.acm.org/10.1145/2494266.2494321

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific Data* **3**, 160018 EP -. DOI: 10.1038/sdata.2016.18.

TEI Consortium, eds. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.6.0. [16th July 2019]. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (2019-09-25).

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

10 https://github.com/distantreading/WG1

11 https://zenodo.org/communities/eltec