Towards Modeling the European Novel. Introducing ELTeC for Multilingual and Pluricultural Distant Reading

Herrmann, J. Berenike (1); Odebrecht, Carolin (2); Santos, Diana (3); Francois, Pieter (4)
1: University of Basel, Switzerland; 2: Humboldt-University Berlin, Germany; 3: University of Oslo, Norway; 4: University of Oxford, UK

Full Abstract

This contribution reports on the collaborative effort of building an open access multilingual corpus of European novels published 1840-1920 (the European Literary Text Collection - ELTeC) within the COST Action “Distant Reading”.¹

Working at the intersection of many languages and cultures, we address practical and technical aspects of corpus design based on a theoretical discussion of pluri-cultural computational modeling of literature. In the corpus design, we adopt a metadata-based approach that allows for representing the diversity of novels published 1840-1920 across Europe. Our sampling and balancing criteria use metadata including publication date, text length, reprint counts and authors’ gender,² and we deliberately focus on inclusion of non-canonical novels.

We have built a workflow for systematically sampling and encoding novels, as well as a consistent annotation model of data and metadata (cf. Burnard, Schöch, Odebrecht, 2019). Currently, ELTeC constitutes a dynamic intersection of fictional discourse in fourteen languages, including Czech, English, French, German, Greek, Hungarian, Italian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovenian, and Spanish (ca. 600 text candidates amounting to ca. 52 mio. words).³ Figure 1 depicts the current state of the Portuguese sample, across

¹ [https://www.distant-reading.net/](https://www.distant-reading.net/)
² See our white paper on [https://distantreading.github.io/sampling_proposal.pdf](https://distantreading.github.io/sampling_proposal.pdf)
³ Current state of ELTeC: [https://distantreading.github.io/ELTeC/index.html](https://distantreading.github.io/ELTeC/index.html)
categories “text length” (long, medium, short), “gender” (female, male), and “date slot” (T1-4; https://distantreading.github.io/ELTeC/por/index.html).

Figure 1. Screenshot of Portuguese sample

Our TEI-XML (TEI Consortium, 2019) encoding scheme is minimal, but aims at facilitating a rich and well-informed distant reading. ELTeC is rooted in the open data movement, with collaborative data creation and an open access extensive documentation for (meta-)data schema, decisions and workflows. Each version of ELTeC is archived via Zenodo. Thus, our (meta)data are re usable, interoperable, accessible and findable (cf. FAIR Guiding Principles; Wilkinson et al., 2016).

In view of the challenge of ‘representativeness’ (see Biber, 1993), ELTeC deliberately refrains from modeling the statistical distribution of populations of publication or reception (cf. Herrmann & Lauer, 2019). Rather, we address the inevitable bias included in the sampling (see Bode, 2018), as well as the explicit link to research questions (Underwood, 2019; Lüdeling, 2011) and the act of construction (Piper, 2019).

ELTeC caters to the computational modeling of literature at the intersection of cultures, nations, languages, genders, but also poetics, trends, and traditions, in a historical period of extreme aesthetic change and diversity. Giving one example, in collaboration with other working groups, using demonym and named entity recognition, it is possible to comparatively explore images of ‘the other’ (ethnic, national, regional; Leerssen, 2016). Generally, in the creation of ELTeC we aim at inductively defining what a ‘novel’ is, allowing for diverse approaches in literary theory and history to be explored and tested.

---

4 Corpus data: https://github.com/COST-ELTeC, documentation: https://github.com/distantreading/WG1
5 https://zenodo.org/communities/eltec
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


