Distant Reading Training School – *Exploring ELTeC: Use-Cases for Information Extraction and Analysis*

Aleksandra Marković
aleksandra.markovic@isj.sanu.ac.rs
The Institute for Serbian Language SASA
Belgrade, Serbia

PAPER SUBMITTED: 28 May 2022
PAPER ACCEPTED: 8 June 2022

Training school *Exploring ELTeC: Use-cases for Information Extraction and Analysis* took place in Belgrade, from 22th till 24th of March 2022. This workshop, dedicated to ELTeC (abb. for European Literary Text Collection) was the final one within the COST Action 16204, under the title *Distant Reading for European Literary History*.

The training school was organized at the Faculty of Mining, University of Belgrade, and organizers were Prof. Dr Ranka Stanković (from the same faculty), Prof. Dr Cvetana Krstev (Faculty of Philology, University of Belgrade),[1] and Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences). The workshop was organized in a hybrid form, which means that attendees could choose whether they will take part in person or online. Among the attendees who participated in person were ones from Serbia, Slovenia, Romania, Belgium and Lithuania, and participants who were present remotely were from Austria, Britain, Portugal, Hungary. There were no formal knowledge requirements for the participation, but it was advised for the attendees to have at least basic computer skills. The target audience for this training school were researchers from the countries which took part in the Distant Reading project, interested in Digital Literary Studies, Corpus and Computational Linguistics, Literary Theory and their methodological uses across national traditions.

The training school offered hands-on approaches to information extraction and analysis of textual data, especially ELTeC corpora, developed within

1. Professors Krstev and Stanković were in charge for the production of the Serbian part of ELTeC. Those who are interested in the Serbian collection digitized within Distant Reading may find information about the work in the journal Infotheca, Vol 21, No 2, which is wholly dedicated to the Serbian novels collection. ELTeC can be reached at the Github of the Action

the aforementioned COST action. Among covered topics were different aspect of work with named and geographical entities: their recognition (NER), extraction (NEE), as well as their analysis; work with Wiki-ELTeC data, linking (historical) data with Nodegoat, the platform made for research in humanities),[2] semantic analysis with word embeddings and language models, and comparing corpora with stylometry.

Among the trainees there were many participants from the COST Action Distant Reading for European Literary History. Let's mention some of them: Christof Schöch (Professor of Digital Humanities at the University of Trier, Germany; the chair of this COST Action); Maciej Eder (the Director of the Institute of Polish Language, Polish Academy of Sciences); Diana Santos (Professor of Portuguese language, and Statistics for Humanities at the University of Oslo), Fotis Yannidis (Professor of Digital Humanities at the University of Würzburg in Germany); Dr Cvetana Krstev (Professor of Information Sciences at the Library and Information Science Department, Faculty of Philology, University of Belgrade, retired); Dr Ranka Stanković (Professor of Mathematics and Information sciences at the Chair of Applied Mathematics and Informatics, The Faculty of Mining and Geology, University of Belgrade).

The training school was organized in nine modules, and the attendees were expected to participate in all sessions.[3]

1. Christof Schöch held (online) an introductory lecture about the project and ELTeC (*What is ELTeC all about?*). This lecture was an introduction to the objectives of the COST Action *Distant Reading for European Literary History*, with a particular focus on the structure of the core deliverable of the project – the multilingual European Literary Text Collection (ELTeC).
2. Maciej Eder, Joanna Byszuk, and Artjoms Šeļa held a lecture (online) under the topic: *Exploring and comparing ELTeC corpora with stylometry*. The lecture was followed by the Stylo intro hands-on.
3. Diana Santos held an online session: *NER exploitation and analysis*; the lecture was followed by hands-on in R.
4. Benedikt Perak was talking about ELTeC Data Analysis, representation of the Geo-Entities and interlinking with knowledge bases; his lecture (held on site) was also followed by hands-on.

2. Nodegoat
3. The program of the training school.

5. Fotis Jannidis and Leonnard Konle held an online lecture: *Semantic analysis using word embeddings and language models*, followed by a practical part.
6. Ranka Stanković and Milica Ikonić Nešić held (on site) Wiki-ELTeC data session: *Wikidata introduction; Wiki-ELTeC schema* (all metadata from header plus main characters, their relations, places); *pipeline for Wiki-ELTeC data population*; *predefined SPARQL query exploration.* The talk was followed by hands-on: *population of Wikidata for other languages.*
7. Denis Maurel, Eric Laporte, and Cvetana Krstev held a lecture (online): *Unitex for processing of literary text: the case of NER automata. Enriching ELTEC texts by Named Entity Recognition using CasSys to parse texts with Unitex graph cascade of finite state transducers in different languages* (hands-on).
8. Jessie Labov, Pim van Bree, and Geert Kessels held an online lecture: *ELTec in Nodegoat: Introduction to the Nodegoat interface and how it works with this kind of data.*
9. Pim van Bree and Geert Kessels talked about *Using Nodegoat for working with the ELTeC data* (specifically the NER), demonstrating how to enrich it by linking it to open data sources (hands-on).

The workshop was intensive, informative and dynamic. All materials needed for work and hands-on were available at the GitHub.[4] The organization and lead through topics were great, topics were interesting and relevant, and practical work was conceived very well (even for those who lack higher level of computer skills, as in the case of the author of this review).

4. Workshop material