Introduction to Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice Tršić, December 4–8, 2023

Milica Rabrenović
milica.rabrenovic@isj.sanu.ac.rs
ORCID: 0000-0002-4441-4884
Institute for Serbian Language SASA
Belgrade, Serbia

PAPER SUBMITTED: 17 April 2024
PAPER ACCEPTED: 29 May 2024

1 Introduction

The University Library "Svetozar Marković" and the Society for Language Resources and Technologies – JeRTeh organized the first seminar, "Introduction to Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice", and it was held in Tršić from December 4 to December 8, 2023. The seminar was conducted through collaboration between its authors and the Ministry of Science, Technological Development and Innovation of the Republic of Serbia. The seminar participants were undergraduate, master's and doctoral students. The goal of the seminar was to introduce students of philological disciplines and humanities to digital humanities methods, as well as to prepare and process texts for the application of distant reading techniques. The seminar included a lecture component and a practical component, i.e., workshops.

On the first day of the seminar, two lectures were held, followed by the application of the theoretical segment in the afternoon. The first lecture, "Introduction to Digital Humanities", was delivered by PhD Vasilije Milnović. At the beginning of his presentation, PhD Milnović reflected on the results of their previous projects, introduced the participants to the COST Action CA16204: *Distant Reading for European Literary History* (2018–2022), and highlighted the importance of digital humanities. He discussed the significance of developing language technologies, with a particular focus on the so-called "small languages" (e.g., Icelandic, Estonian, or Hebrew, where substantial investments are made in the field) and their potential applications

in lexicography, the teaching of Serbian language and literature, translation, and other areas.

The lecture by PhD professor Cvetana Krstev was dedicated to the *SrpELTeC* corpus. PhD professor Krstev introduced the participants to the process of creating this corpus and the criteria that had to be met for Serbian novels from the given period to become a part of the corpus (equal representation of male and female authors; balanced coverage of the selected period (1840–1920); variety in terms of length; inclusion of both well-known and lesser-known works; diversity of authors). She discussed the challenges that occurred during the formation of the Serbian subcollection and the process of digitizing the first editions (or the oldest available ones). The digitization process includes the following steps: scanning, optical character recognition (OCR), correction, basic annotation, metadata entry and automatic advanced annotation.

The seminar participants were introduced to the process of how text correction and annotation are performed, and the way metadata is entered. On the first day of the workshop, they were given a text for correction following the annotation guidelines. In the following days, they entered metadata and worked on annotating linguistic entities in the text.

On the second day, PhD Aleksandra Trtovac, in her lecture "The University Library 'Svetozar Marković' and Distant Reading", discussed the connection between librarianship and digital humanities, as well as their role in the COST Action. PhD Trtovac dedicated her second lecture to the *Transkribus project* and digitization using mobile phones. During the workshop that day, participants created accounts on Transkribus[1] and were introduced to the *DocScan* app, which enables the uploading of handwritten works to generate text transcripts. A particularly useful segment was the introduction to advanced search in the COBISS[2] system. In addition to the selective searching, the focus was also placed on the command searching. One of the valuable pieces of information shared was that the *UNPAYWALL* section contains open-access scientific articles. Furthermore, all names and surnames of a single author (useful when dealing with pseudonyms or when a female author's surname has changed due to marriage) can be found in the *CONOR.SR* section.

On the third day, PhD professor Krstev introduced participants to the concept of named entities and the methods for annotating them in a text. PhD professor Ranka Stanković's lecture was built upon the previous one,

1. readcoop, visited 28. 6. 2024.
2. COBISS.SR, visited 28. 6. 2024.

focusing on the processing of digital texts. Workshop participants worked on the automatic annotation of text segments with named entities (places, people, organizations, professions, etc.). Notes on more advanced types of annotation were also provided, followed by the review and evaluation of automatically annotated data.

On the same day, during the workshop, participants were introduced to Wikidata entries about Serbian novels and entered basic information about the text they received on the first day using the tools *OpenRefine* and *QuickStatements*. Additionally, participants were introduced to data querying using the *SPARQL* language.

The fourth day began with a lecture by PhD professor Duško Vitas, who discussed the process of creating a culinary corpus. PhD professor Ranka Stanković continued the topic by presenting the corpora available on the NOSKE platform of the JeRTeh society.[3] Participants had the opportunity to learn the basics of the *CQL* language and how to query the *srpELTeC* corpus. The workshop focused on command searches, starting with simpler queries and progressing to more complex grammatical patterns and queries involving named entity tags.

On the final day of the seminar, participants attended the lecture "Initiative for the Development of Open NLP/NLU Resources and Tools for the Serbian Language" by Slobodan Marković. The lecture highlighted the challenges of using artificial intelligence (AI) and the possibilities and issues related to processing the Serbian language with AI. The problem lies in the fact that models have not been fundamentally adapted, their application is not practical for business solutions, and their expressive capabilities are limited. The solution is to have as much training text as possible. Ideally, this data should be publicly available, under a permissive license, and cover both Serbian dialects. The goal of the initiative for open NLP resources is to have better search results and improved text comprehension.

This seminar provided comprehensive knowledge about the methods of digital humanities and the significance of this scientific field. It has allowed for better insights into future research, especially when it comes to annotating resources and working with textual corpora, which is where the greatest application of what was learned during the seminar lies. What should be emphasized as the most important outcome of these lectures and workshops is the opportunity to connect with colleagues and develop potential collaborations on future projects and research. This is precisely the main reason why, as a participant, I would highly recommend the seminar "Introduction to

3. noske.jerteh.rs, visited 28. 6. 2024.

Digital Humanities: Workshops on the Implementation of Distant Reading in Research Practice".

Acknowledgment

This work was financed by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Contract No. 451-03-136/2025-03/200174, concluded with the Institute for Serbian Language, SASA.