Workshop "Methods and Tools of Distant Reading Adapted to Multiple European Languages" at the Galway Training School

Mihailo Škorić
mihailo.skoric@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia

PAPER SUBMITTED: 11 October 2021
PAPER ACCEPTED: 10 November 2021

Galway training school[1] was organized within the COST[2] action CA16204, project *Distant Reading for European Literary History*.[3] One of the main goals of this action was to create a network of researchers which would develop resources and methods within the Distant Reading paradigm – the usage of computational methods in the analysis of large amounts of literary texts. In order to achieve this, one of the action's outcomes is the preparation of a large literary corpus in several European languages, while another is the establishment of good, innovative methods and techniques for computational analysis of these specific texts.

In accordance with that, a training school was organized, and within it two workshops: a workshop on methods and tools of distant reading and a workshop on theoretical concepts and their confrontation with computational methods, with a goal of familiarizing participants with this topic. This training school, the second one organized by this action, was held at the premises of the National University of Ireland in Galway in December, 2018. The aim of the first workshop on methods and tools of distant reading was to present to participants a set of tools that can be applied for solving certain tasks within the distant reading paradigm.

The first tool presented was TXM,[4] which is used for corpus management and analysis using methods of textometry. The tool was presented by Serge Heiden from the *École normale supérieure de Lyon*, one of the authors of the software. From the corpus management point of view, TXM can create

1. Galway Training School 2018 General Information
2. COST Actions
3. D-reading
4. textometrie

corpora from text files, supporting both plain and annotated texts in different formats, such as XML, including TEI. Corpus-related metadata can be added from a separate XML file during the text import. Once created, the corpus can be browsed, searched and analyzed using various statistical and textometry methods, where values such as absolute and relative frequencies for word, word type and lemma can be calculated and exported for simple or complex queries. Software can also perform search for factorial correspondences, hierarchical classification, collocation analysis, etc. Participants had the practical task to create a corpus and then analyze the mentioned metrics using the TXM software.

The second theme of the workshop was topic modeling, and how it can be done using the *TopicsExplorer* package.[5] The lecturer was Steffen Pielström from the Würzburg University. Participants were introduced with topic modeling – grouping texts or documents by topic and were given the task to try it out using some texts and *TopicsExplorer* software. Participants were to experiment, changing the hyperparameters (number of topics searched, lists of stop words, etc.) and analyze the different results obtained in this way. After the experiments, participants discussed what the analyzed texts were about, what their key words were, and which texts are thematically similar and why.

The lecturer of the third course related to network analysis was Meliha Handžić from the *International University BURCH* in Sarajevo. Within this course, two software packages for visualizing text networks were presented. *Palladio*,[6] a software developed by a team from Stanford University that provides spacetime labeling and visualization of textual sources on an interactive map, and *Gephi*,[7] which is used for two-dimensional visualization of distances between the entities, in this case texts or authors. Distance values can represent any parameter and are imported in the form of a matrix table showing the mutual distance of all nodes. In addition to the predefined output, *Gephi* also offers the application of various graph transformation algorithms, to procure truer or better-looking images.

The last course, held on the last day of the training school, was devoted to the *stylo* package[8] for the programming language R, and the lecturer was Joanna Byszuk from the Polish Academy of Sciences – Polish language institute in Krakow, one of the institutions that developed the package. Par-

5. TopicsExplorer
6. Palladio
7. Gephi
8. stylo

ticipants were introduced with some possibilities that the package provides, such as stylometric analysis of texts and authors, finding of stylometric similarities and differences between documents, document clustering and dealing with the problem of authorship attribution.[9] Participants were required to create a set of documents for analysis, run *stylo* on their computer and analyze the texts. The corpus of short English novels was provided for participants that did not have access to a corpus of texts in their mother tongue. The last task for participants in this course was to experiment, using the imported corpus, with training and testing a computer model in order to determine document authorship.

Upon completion of this training school, participants were able to prepare a corpus, analyze it, present results, and were thus ready to do research and to study literary works through the distant reading paradigm.

9. Stylo in Galway