

TNA ACTIVITY REPORT

Can't you see what I'm going through? Stylometrical analysis of Marcus Tullius Cicero's works in their correlation to the author's life.

Author: Wiktor Dziemski

Current position: specialist at Institute of Source Edition, Faculty of Law and Administration

Affiliation: Jagiellonian University

Host institution: UNED

Mentor(s): Salvador Ros, Álvaro Pérez Pozo

Period of stay: 26.11.2023 – 8.12.2023

1. Introduction

The primary objective of this fellowship was to conduct a stylometrical analysis of Marcus Tullius Cicero's works to identify and verify changes in his literary style corresponding to significant events in his life. Cicero's extensive body of work, including speeches, philosophical treatises, and personal letters, provides a unique opportunity to explore the relationship between an author's life experiences and stylistic evolution. The research focused on applying stylometric techniques, particularly Most Frequent Words and n-grams, to discern patterns in Cicero's writing style over time.

2. Objectives

The fellowship aimed to achieve the following objectives: a) Utilize web scraping tools to collect Cicero's works from www.thelatinlibrary.com. b) Create a comprehensive literary corpus by aggregating the collected texts. c) Apply stylometric analysis techniques to explore linguistic patterns and writing styles within the corpus. d) Identify and analyze significant findings, potentially paving the way for future research in literary analysis.



3. Methodology

a) Web Scraping:

- Analyzed the HTML structure of www.thelatinlibrary.com.
- Implemented the Wget application for ethical and legal web scraping.

b) Corpus Creation:

- Cleaned and preprocessed texts by removing HTML tags, punctuation, and non-textual elements.
- Organized texts into a structured corpus with appropriate categorization and metadata.

c) Stylometric Analysis:

- Employed the R package stylo for stylometric analysis.
- Utilized statistical and computational methods to quantify and analyze stylistic patterns.
- Visualized stylometric developments across Cicero's works chronologically and by genre.

d) Data Interpretation:

- Analyzed stylometric results to identify trends, patterns, and characteristics within the literary corpus.
- Interpreted findings in the context of Cicero's life, offering insights into genre-specific traits and stylistic evolution.
- Summarized results in a comprehensive report, highlighting implications and potential avenues for further research.

4. Timeline

Due to unexpected personal circumstances, the duration and implementation of the planned tasks had to be rearranged. It was only possible to organize two weeks of onsite attendance at UNED. During this time, however, much has been done. Scraping the site www.thelatinlibrary.com was the first and easiest step. Also cleaning the data was unproblematic. The creation of corpus required much deliberation, especially putting the texts in chronological order. This was only possible after the dating of particular pieces was unified. The raw files from the scraped www site offered dating employed by the ancient Romans i.e. according to the founding of Rome with the use of the Roman calendar (Idus, Nonas, Kalendas). Furthermore, the letters of Cicero required much attention. Firstly, because of their dating, and secondly because of their sheer amount and uneven dispersion.

Most of the letters come from just a few years. Also, letters written to Cicero and not by him had to be disregarded as well as letters, whose dating was unknown.

5. Infrastructure of Laboratorio de Innovación en Humanidades Digitales (UNED)

The Laboratorio de Innovación en Humanidades Digitales (LINHD) at UNED provided essential support for both building the corpus and analyzing literary works. The infrastructure offered tools and expertise necessary for successful execution. Professor Salvador Ros was really helpful, especially at the preliminary period of the fellowship where much theoretical and ontological discussion was necessary. Working with Álvaro Pérez Pozo, a computational linguist from the Laboratorio was a great opportunity to learn how to apply Python coding to literary research. The decision to collaborate with LINHD was influenced by previous positive interactions with staff during the CLS Infra training school in Madrid (2023).

6. Conclusion

This fellowship contributes to literary analysis by creating a comprehensive corpus and conducting a stylometric analysis of Cicero's works. The support from LINHD has been instrumental in achieving the research goals. The acquired knowledge and skills will be applied in future research, contributing to the development of my academic career. Anticipating publication in a peer-reviewed journal, the project sheds new light on the correlation between an author's life and literary evolution.