Short Term Scientific Mission to Krakow: Comparative Stylistic and Morphosyntactic Analysis of ELTeC Texts Using Stylo R Package

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

PAPER SUBMITTED: 12 October 2021
PAPER ACCEPTED: 10 November 2021

The COST action CA16204 - Distant Reading for European Literary History issued several STSM (short term scientific mission) calls enabling action participants to visit organizations in other action member countries, with the goal to learn more about a topic in which a host organization is specialized, or to solve together a specific problem. The mission to procure comparative stylistic and morphosyntactic analysis of ELTeC texts using *stylo R* package[1] was organized within the third call (the first quarter of 2020), and in March 2020 I visited one of the institutions that participated in the creation of the package, the Polish language institute in Krakow.

The purpose of the short-term scientific mission was to perform experiments that would apply different levels of morphosyntactic annotations to a text from the ELTeC collection, in order to obtain accurate numeric comparisons between these different text incarnations. The main motivation was to test the usability of these morphosyntactic annotations in stylometric analysis and find the optimal combination of annotations that can provide better classification results in various scopes (gender, time period and authorship) related to ELTeC texts. If such approaches were proven to increase classification accuracy, at least to some extent, they could be applied for further analysis of texts written in languages with developed morphosyntactic annotation tools.

Work was done through several working sessions in which researchers involved in creation and maintenance of the *stylo* package participated. These included methodology overview, a workshop on classification using the *stylo*

[1] Stylometry with *stylo*

R package, discussion meetings, and one session devoted to interaction with everyday users. All tests were performed using the Serbian ELTeC texts. Balanced subsets – incarnations of the documents - were prepared for the experiments by cross sectioning two groups of variants: one was using the metadata (author, authors gender and time-period) and the other was using different levels of morphosyntactic annotation (part-of-speech – POS, grained part-of-speech, lemma, and word forms, that is, words with no-annotation) totalling in twelve different incarnations. These were all tested on classification reliability to establish their efficiency and create a comparative analysis. The idea was to answer the following and similar questions:

- What defines an author's style? Is it the use of word forms or the use of lemmas?
- Do different authors of different gender use different word combinations or word forms with different lemmas?
- Is the use of a certain part of speech specific for a time period in literature?

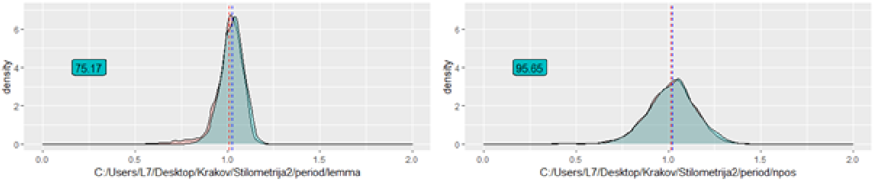

Figure 1. Delta distances distributions (same-class is blue and different-class is red) for the time period using lemmas (left) and for the gender using grained POS (right) outputted through R studio via R script prepared during the STSM.

The classification reliability was investigated by overlapping same-class and different-class cosine delta distance distribution. Overlapping areas of such distributions were procured as indicators of classification reliability for certain document incarnations. If same-class and different-class cosine delta distance distributions overlap, then accurate classification is improbable, and if, on the other hand, they do not overlap it indicates a higher probability of accurate classification using incarnations of that text, with the smaller area on the output graphs. In cases where distributions intersect multiple times, overlap or have close-by means, the classification reliability cannot be accurately measured; in that case, it is by default very low or none.

For testing purposes an R script that uses the *stylo* package was compiled, which can be reused for testing the results for different classes, annotations and even languages, provided with the required text incarnations.

Classifications by gender and period were both found to be unobtainable for Serbian ELTeC texts. This was observed through multiple intersections of the series, close-by means and overall multitude of overlapped areas for these cases of simple binary classification. Two out of eight graphs for these classifications are shown as an example of the low reliability classification in Figure 1.

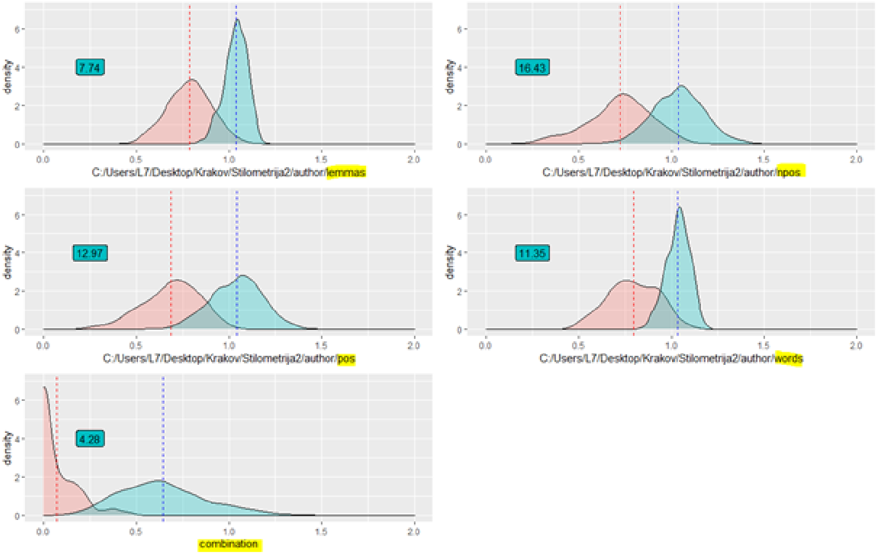

Figure 2. The comparison of delta distance distribution for different text incarnations regarding authorship attribution. Top left is for a text incarnation using only lemmas, top right is using only grained POS, middle left is using only POS, middle right is using an original text and bottom is using the combination of delta distances for all incarnations.

These results were explained by the narrow time window of the text origin for Serbian novels of ELTeC corpus, in regard to the period, and the lack of obvious differentiation of female and male writing styles in early 20 century Serbian novels (combined with the low representation of female authors in that period in general) regarding gender differentiation.

On the other hand, classification by authors (authorship attribution) yielded interesting results. The text collection comprised 80 novels written by 17 writers. Findings of previous stylometric research in Slavic languages were confirmed by this experiment. Classification using lemmatized text was found to be the most efficient, with a text incarnated into a series of part-of-speech tags (and using trigrams) falling behind the original text. The Grained POS (also using trigrams) was found to have the worst result in this scenario. The text incarnations with POS tags showed to have higher standard deviation. These results are presented in Figure 2.

However, the combination method (using distance matrices of different incarnations) developed on the last day of the STSM proved to be the most reliable, with up to two times better result than a lemmatized text, over a 17-class classification by authors, with possible error margin down to only 4.3% (see the bottom row of Figure 2). The obtained results prove that the STSM to Krakow was successful.

Acknowledgment

This STSM would not be possible without the hosts from the Polish language institute in Krakow – Maciej Eder and Joanna Byszuk, nor without the help of their keen colleagues from the Institute's Department of methodology.