Developing a Pipeline for Automatic Linguistic Analysis of Historical Manuscripts and Early Printings: The Pre-Modern Slavic Case

Rabus, Achim
achim.rabus@slavistik.uni-freiburg.de
University of Freiburg, Germany

Arnold, Eckhart
Arnold@badw-muenchen.de
Bavarian Academy of Sciences and Humanities, Germany

Jouravel, Anna
anna.jouravel@slavistik.uni-freiburg.de
University of Freiburg, Germany

Lendvai, Piroska
Piroska.Lendvai@badw-muenchen.de
Bavarian Academy of Sciences and Humanities, Germany

Meindl, Martin
martin.meindl@slavistik.uni-freiburg.de
University of Freiburg, Germany

Polomac, Vladimir
vladimir.polomac@gmail.com
University of Kragujevac, Serbia

Renje, Elena
elenarenje@outlook.de
University of Freiburg, Germany

Recent advances in Handwritten Text Recognition (HTR) revolutionize access to historical sources and make mass digitization (i.e., by means of AI-assisted automatic transcription) of large amounts of manu-scripts feasible for the first time in history. Using the arguably most widespread and advanced HTR platform Transkribus (Muehlberger et al. 2019), we trained and published HTR models for both pre-modern East and South Slavic (also called Church Slavonic) Cyrillic sources (Rabus 2019; Polomac 2022) that are capable of transcribing manuscripts and early printings with a low character error rate (CER) of below 4% and 2%, respectively.

In our paper, we report on experiments where we use these HTR models to automatically create large pre-modern Slavic text cor-pora (several millions of word tokens) and to use these corpora without manual post-correction for quantitative linguistic analysis, since manually correcting large amounts of linguistically annotated data is slow and expensive, thus not always feasible. Our approach is twofold: First, we use uncorrected raw text data for applying state-of-the art descriptive and inferential statistical methods such as mixed-effects logistic regression or random forest analysis. Furthermore, we experiment with stylometric approaches for the analysis of linguistic variation. Using the stylo package (Eder et al. 2016) in RStudio and particularly its Nearest Shrunken Centroids (NSC) algorithm, it was shown to be possible to pinpoint potential linguistic variables of interest in an inductive, bottom-up manner and to uncover subtle, previously undetected differences (e.g., with respect to the frequency and distribution of function words) between two or more subcorpora of a given language (Lahjouji-Seppälä et al. 2022). In the current paper, we apply this method to pre-modern Slavic data for the first time. Second, we conduct full-morphology tagging of the uncorrected raw texts using the standalone version of the stanza tagger (https://github.com/yvesscherrer/stanzatagger) and the pre-modern Slavic model presented in

Scherrer et al. (2018) as well as lemmatization using UDPipe (https://github.com/bnosac/udpipe). As has been shown (Besters-Dilger / Rabus 2021), the noise introduced by the uncorrected full-morphology tagging of pre-modern Slavic data does not prevent linguistic analysis from being conducted successfully. In our paper, we explore the challenges posed by working with lingu-istically annotated data in which incorrect annotation labels may origin from various analysis levels, first introduced by uncorrected HTR transcriptions and second by uncorrected full-morphology tagging and lemmatization. We demonstrate the possibilities and limits of quantitative linguistic analysis on such data, compare the results with results obtained in a traditional (analogue and qualitative) way and show quantitative and qualitative approaches to evaluate the actual amount of noise in the data. Finally, we reflect on how the specific features of pre-modern Slavic data (rich morphology, high amount of variation at different linguistic levels) affect our pipeline and give an outlook of how to apply the pipeline to other use cases (different historical epochs, script styles, and languages).

Our paper showcases both the opportunities of interdisciplinary and international collaboration and the important contribution by scholars from South-Eastern Europe. It demonstrates that data from (South-) Eastern Europe are an interesting test case for the language- and text-focused Digital Humanities community and helps overcome the monolingual and English-focused bias.

Bibliography

Besters-Dilger, Juliane / Rabus, Achim (2021): “Neural Morphological Tagging for Slavic: Strengths and Weaknesses”, in: Scripta & E - Scripta 21: 79–92.

Eder, Maciej / Rybicki, Jan / Kestemont, Mike (2016): “Stylometry with R: a package for computational text analysis”, in: R Journal 8, 1: 107–121 https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf [29.04.2023].

Lahjouji-Seppälä, M. Zaidan / Rabus, Achim / von Walden-fels, Ruprecht (2022): “Ukrainian standard variants in the 20th century: stylometry to the rescue”, in: Russian Linguistics 46: 217–232. DOI: 10.1007/s11185-022-09262-9.

Muehlberger, Guenter / Seaward, Louise / Terras, Melissa / Ares Oliveira, Sofia / Bosch, Vicente / Bryan, Maximilian / Colutto, Sebastian / Déjean, Hervé / Diem, Markus / Fiel, Stefan / Gatos, Basilis / Greinoecker, Albert / Grüning, Tobias / Hackl, Guenter / Haukkovaara, Vili / Heyer, Gerhard

/ **Hirvonen, Lauri** / **Hodel, Tobias** / **Jokinen, Matti** / **. . . Zagoris, Konstantinos** (2019): “Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study”, in: Journal of Documentation 75, 5: 954–976.

Polomac, Vladimir (2022): “Serbian Early Printed Books from Venice: Creating Models for Automatic Text Recognition Using Transkribus”, in: Scripta & E - Scripta 22: 11–29.

Rabus, Achim (2019): “Recognizing Handwritten Text in Slavic Manuscripts: a Neural - Network Approach Using Transkribus”, in: Scripta & E - Scripta 19: 9–32.

Scherrer, Yves / **Mocken, Susanne** / **Rabus, Achim** (2018): “New Developments in Tagging Pre - modern Orthodox Slavic Texts”, in: Scripta & E - Scripta 18: 9–33.