

Towards Diachronic Corpus of Polish Latin

Marszałek, Jagoda

jagoda.marszalek@ijp.pan.pl
Institute of Polish Language, Polish Academy of Sciences,
Poland

Nowak, Krzysztof

krzysztof.nowak@ijp.pan.pl
Institute of Polish Language, Polish Academy of Sciences,
Poland

Krawczyk, Iwona

iwona.krawczyk@ijp.pan.pl
Institute of Polish Language, Polish Academy of Sciences,
Poland

Latin was for centuries an essential element of the Polish culture. From the beginning of the 11th century, it was the language of religious and lay communication in the Kingdom of Poland and as such provides an access to information concerning both state-wide events and policy making, and the daily life of individuals. During the Middle Ages, Latin kept playing essential role in diplomacy, administration and trade, but for a long time it was also the exclusive language of scientific writing, theological speculation or historiography. With Polish starting to take over some of its functions, the role of Latin has gradually diminished, but it remained the official language of the Polish-Lithuanian Commonwealth until 1795, when the Third Partition of Poland took place.

For decades, scholars have been voicing the need for critical editions and tools to facilitate research on Polish Latin. The written production of the Middle Ages and the early Renaissance is now roughly covered in the eFontes corpus. There also exist a number of minor text collections that focus on single authors or domains (e.g. Corpus of Ioannes Dantiscus' Texts, IURA). However, they do not allow for an integrated access and as such are by no means amenable to distant reading methods or corpus study. Recent advances in automatic text recognition for historical texts (Weichselbaumer et al. 2020) and the availability of public recognition models trained on large and sufficiently varied datasets promise that this situation may improve in the years to come.

In this paper, we present the preliminary results of our proof of concept study in which we assessed the feasibility of small-scale automatic acquisition of Latin diachronic corpus for linguistic research. By building a custom pipeline from open source components, we expect to identify major challenges such a project may face. We selected a set of works by Polish authors composed in Latin from the 1550 to 1800. For the sake of convenience, we divided the period into 5 parts covering 50 years and consisting of ca. 500 printed pages each. The texts were pre-selected on a basis of their availability in digital libraries. Apart from the date of writing and printing, we did not, however, control for domain or genre, except for a rough distinction between prose and poetry, as it correlates with text layout.

The images, either in PDF or PNG etc. format, were automatically processed with scantailor. The region and line segmentation was handled using kraken. As for the text recognition, we used

the Calamari OCR (Wick et al. 2020) with the publicly available deep3_antiqua-hist model (Reul et al. 2021).

The model turned out to produce by far the best results in our tests carried out on a dozen of manually proofread pages. The text required standardization as the model preserves the typographic conventions typical of early modern editions, but which are not relevant to corpus or lexicographic study. The resulting XML files along with corresponding images are presented in a TEI Publisher instance.

Although not aimed at rigorous evaluation, our preliminary study showed that existing technologies can be easily combined to produce a corpus which is “good enough” for lexicography and general corpus queries. We also identified a number of challenges that need to be addressed the future. Majority of them have to do with the complex layout (e.g. decorative devices, combining various typesets on a single page etc.) and may require either manual intervention in the pre-processing phase or more precise text detection. In order to make informed decisions about choosing and refining OCR models, we will also need to proofread a subset of the texts we have retrieved and produce a ground truth we could use in subsequent evaluation. Finally, basing on our previous research (Wróbel et al. 2022), we will adapt lemmatization and part of speech models to cover early modern Latin.

Bibliography

Weichselbaumer, Nikolaus / Seuret, Mathias / Limbach, Saskia / Dong, Rui (2020): “New Approaches to OCR for Early Printed Books”, in: *Digitalia* 15, 2: 74–87.

Wick, Christoph / Reul, Christian / Puppe, Frank (2020): “Calamari - A High-Performance Tensorflow-based Deep Learning Package for Optical Character Recognition”, in: *Digital Humanities Quarterly* 14, 2.

Reul, Christian / Wick, Christoph / Nöth, Maximilian / Büttner, Andreas / Wehner, Maximilian / Springmann, Uwe (2021): “Mixed Model OCR Training on Historical Latin Script for Out-of-the-Box Recognition and Finetuning”, in: *The 6th International Workshop on Historical Document Imaging and Processing*. New York: 7–12.

Wróbel, Krzysztof / Nowak, Krzysztof (2022): “Transformer-based Part-of-Speech Tagging and Lemmatization for Latin”, in: *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*. Marseille: 193–197.