CLSINFRA COMPUTATIONAL LITERARY STUDIES INFRASTRUCTURE

TNA ACTIVITY REPORT

Polish poetry corpus: creation, integration with PoeTree, computational poetics

Author: Anna Mędrzecka-Stefańska

Current position: researcher-technical specialist

Affiliation: Institute of Literary Research of the Polish Academy of Sciences

Host institution: Charles University, Prague

Mentor(s): Silvie Cinkowa, Michal Křen

Period of stay: 15.12.2024 – 28.02.2025

Introduction and objectives

The project comprised three pivotal components. The first involved creating a comprehensive corpus of Polish poetry using publicly accessible data and preparing it for seamless integration into the multilingual poetry corpus, PoeTree. The second component focused on embedding this corpus within the PoeTree infrastructure, while the third centered on conducting initial research to analyze and interpret the gathered material.

Prior to commencing the scholarship, I had some corpus material at my disposal; however, it did not yet meet the stringent requirements of PoeTree. Consequently, the foremost and most critical task was to meticulously adapt the existing material and meticulously gather supplementary data to significantly expand and enrich the corpus.

Methodology

During my fellowship with CLARIN-PL, several poetry corpora were developed with my involvement, marking significant advancements in the field.

- **Corpus of Four Bards (K4W)**: Led by Professor Marek Troszyński's team, this corpus encompasses the complete works of eminent poets Adam Mickiewicz, Juliusz Słowacki, Zygmunt Krasiński, and Cyprian Norwid. The corpus includes 389 texts by Mickiewicz (94,020 tokens), 353 by Norwid (105,284 tokens), 283 by Słowacki (75,967

CLS INFRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004984

tokens), and 177 by Krasiński (90,231 tokens). While the entire K4W cannot be incorporated into other projects due to balancing considerations, it provides an excellent foundation for future expansion and refinement.

- **Corpus of 19th-Century Polish Women's Poetry (KPPK XIX)**: Created during my internship with CLARIN-PL, this corpus includes 120 poems (12,868 verses) by ten Polish female poets from the 19th century.
- **Corpus of Polish Poetry for Artistic Expression Analysis**: Developed as part of CLARIN-PL's ongoing work on artistic expression detection tools, this corpus is expanding and currently contains 59 texts (14,297 tokens) from various periods.

Throughout the fellowship, I focused on enriching these corpora with texts from freely accessible resources, aiming to establish a broader corpus of 19th-century Polish poetry. Notable sources include Wikisource, Polona, and Wolne Lektury. Utilizing Python scripts and institutional support, I targeted underrepresented periods and authors, particularly those from 1775-1822 and 1822-1918, as well as lesser-known Romantic poets.

Fellowship outcome

Upon arrival in Prague, I had at my disposal the aforementioned collection of texts in txt format. Consequently, I undertook two parallel tasks: firstly, adapting the format of the existing files to meet the integration requirements of PoeTree, and secondly, identifying available texts online that could supplement the corpus. This resulted in a list of over 220 titles that could be acquired from the Wolne Lektury service, as well as more than 200 volumes of poetry accessible through the Polona digital library.

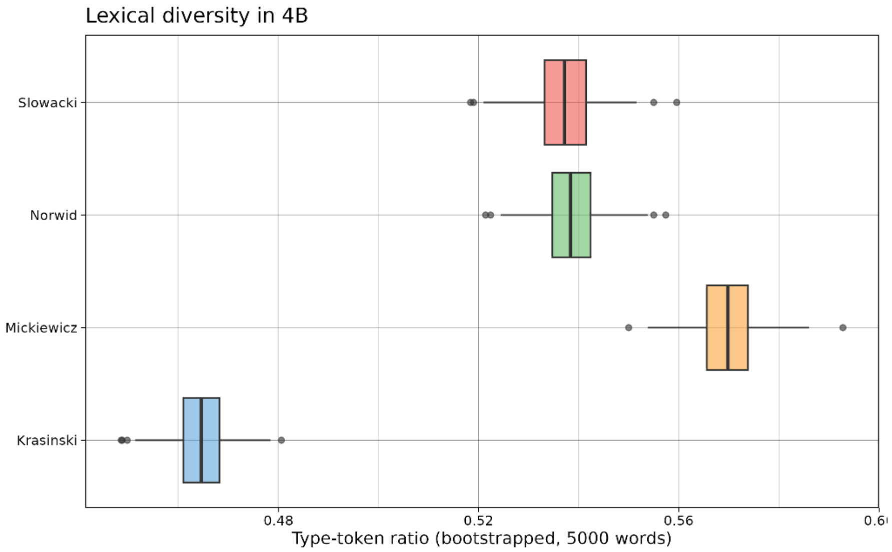
The texts available from the Wolne Lektury service could be downloaded in XML format, allowing for automated processing and relatively swift implementation. Unfortunately, the texts sourced from the Polona library primarily consisted of scans and OCR output of varying quality, necessitating extensive correction. This process proved to be significantly more labor-intensive and time-consuming than initially anticipated. However, a positive outcome is that the large quantity of acquired texts will ultimately enable the creation of a comprehensive corpus of Polish poetry.

At present, the corpus comprises 4,000 poems. Each file has been processed into a format compatible with PoeTree integration and annotated with appropriate metadata. The metadata pertaining to authors has been linked to corresponding WikiData entries. Corpus had been deduplicated and processed with UDPipe. Petr Plechac and Artjom Sela from PoeTree developement team were of great help during this process.

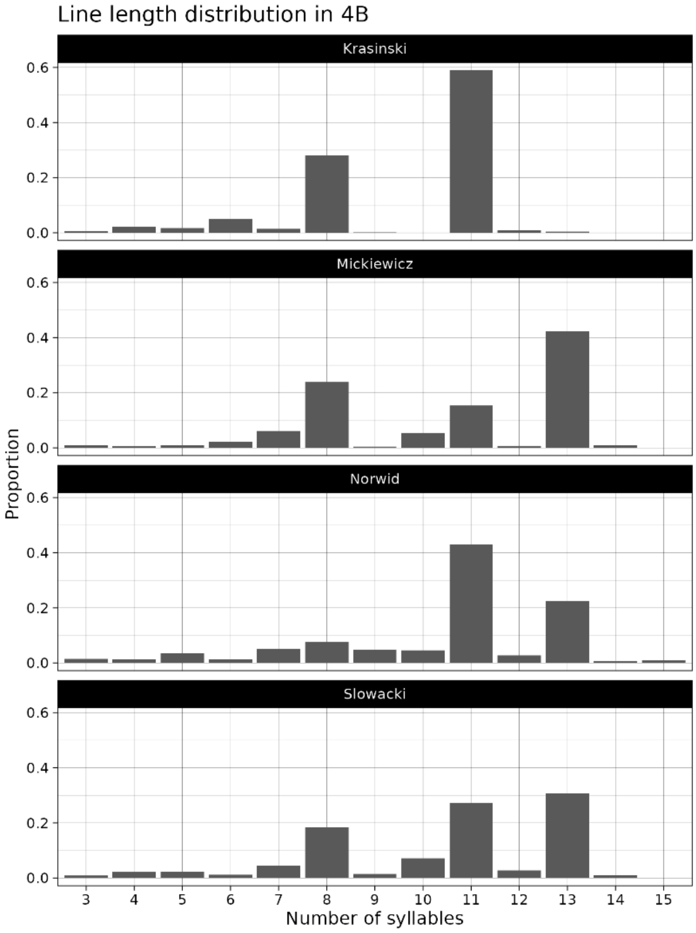
During my fellowship, I was also able to conduct preliminary stylometric analyses on materials from the Corpus of Four Bards using the Stylo tool. Below, I present some exemplary results. The study proved particularly intriguing, as it revealed that Zygmunt Krasiński significantly diverges from the other three poets. This finding is especially interesting given that Krasiński, Mickiewicz, and Słowacki belong to the same generation, known as the first generation of Polish Romantics. Their works are stylistically similar, and they moved in comparable social circles. Biographically and literarily, one would expect Norwid - a representative of a younger generation whom most scholars no longer classify as a Romantic - to be the outlier.

Among others, the following analyses were conducted:

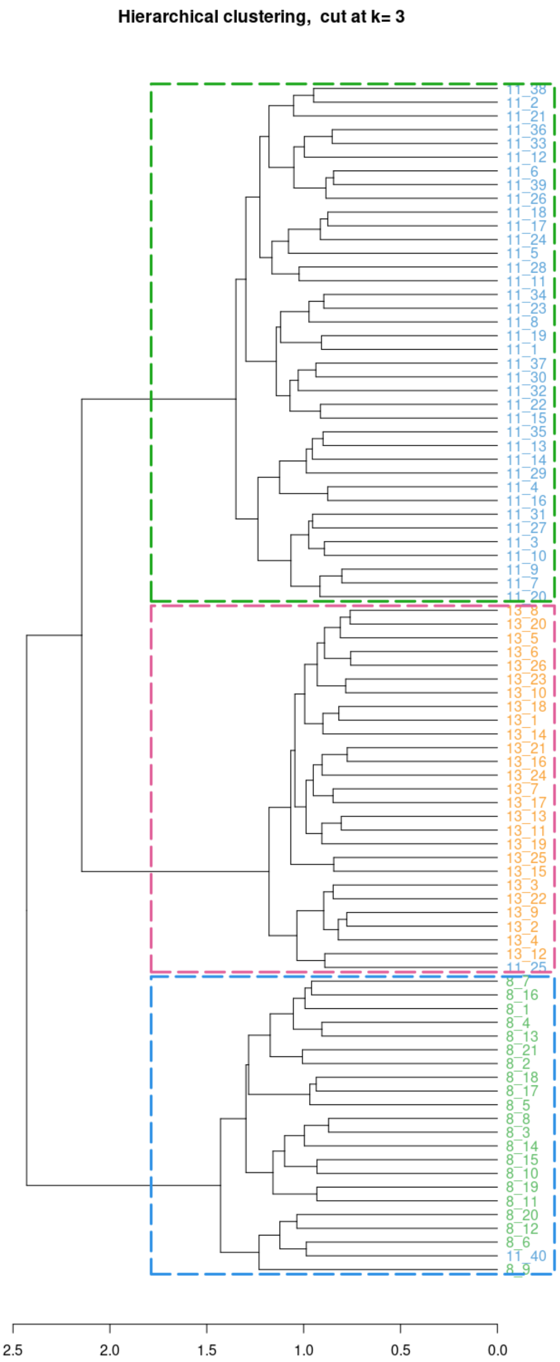
1. Type-token bootstrap plot - Krasiński's work stands out distinctly here.



2. Distribution of line lengths across authors - the result of this analysis corresponds well with the subsequent study, demonstrating a relationship between the poems' themes and their versification structure.



3. Clustering of samples of lines of the same lengths (showing genre x form association).



4. Words associated with the samples of different lengths (demonstrating a relationship between the poems' themes and their versification structure).

Cluster 1

my
takie
patrz
wy
sen
znów
drogi
piersi
dotąd
kwiaty
coraz
góry
imię
wszystkie
masz
idzie
nią
woła
pierś
jedno
czoło
oto
dwa

Cluster 2

chociaż
cóż
chce
których
zaś
człowiek
stąd
zawsze
znowu
nieraz
ów
tyle
człowieka
oczy
szczęścia
ją
była

Cluster 3

nawet
panie
wiem
pani
miłości
twój
bądź
mego
twej
wtedy
twych
mej
bym
tobą
pieśń
krew
moją
świecie
dzisiaj
niż
będą
nic
mną
wieki
twego
wszystkich
widziałem
tę
będziesz
mych
anioł
nigdy
duszy
ducha
me
mojej
śmierć
moich
będę

These preliminary findings open up intriguing avenues for further research into the stylistic peculiarities of these key figures in Polish Romantic literature. The unexpected divergence of Krasiński's work, despite his generational and social proximity to Mickiewicz and Słowacki, warrants deeper investigation. Additionally, the apparent correlation between thematic content and versification structure could provide valuable insights into the poets' compositional techniques and stylistic evolution.

Further studies could explore the reasons behind Krasiński's stylistic divergence and investigate how Norwid's work, despite his generational difference, aligns more closely with the older Romantics in certain stylometric aspects. In fact, together with Artjom Sela we plan to work on a paper on this subject.

Future work

The primary and most crucial task to be undertaken following the conclusion of my fellowship is the completion of the text correction process for the acquired materials and the expansion of the Polish poetry corpus. Due to technical considerations, implementing an unfinished corpus into the PoeTree infrastructure was deemed impractical. Therefore, the next step will be the publication of the corpus and its integration into this resource.

This integration will facilitate conducting further research on the corpus and continued expansion of its volume, contingent upon the acquisition of additional data sources.

The completion of this task is essential for several reasons:

- It will ensure the integrity and reliability of the corpus, which is crucial for any subsequent analyses.
- The publication will make this valuable resource available to the wider academic community, potentially spurring new research directions in Polish literary studies.
- Integration with PoeTree will enhance the corpus's accessibility and usability, allowing for more sophisticated computational analyses.

The expanded and refined corpus will serve as a robust foundation for various research endeavors, including but not limited to:

- Comprehensive stylometric analyses across different periods of Polish poetry
- Investigations into thematic and stylistic evolution over time
- Comparative studies with other language corpora within the PoeTree framework

Summary

In summary, the outcomes of my fellowship have been substantial and have significantly contributed to the advancement of my academic career. The planned deliverables and future research directions are as follows:

1. Publication of the corpus in open access format.
2. A previously mentioned article on stylometric analyses of the Corpus of Four Bards.
3. A publication accompanying the release of the Polish poetry corpus, which will likely detail its composition, methodology, and potential applications.

Future research directions enabled by this work include:

- Versification studies
- Analyses of language patterns and rhyme schemes
- Investigations based on the corpus metadata
- Comparative studies with poetry corpora in other languages

These research avenues represent the next steps in my academic career and will build upon the foundation established during this fellowship.

Despite the challenges encountered in acquiring and processing the source material, I consider the achieved results to be satisfactory. The goals set for the immediate future appear to be readily achievable, thanks to the groundwork laid during this fellowship period.

The fellowship has thus proven instrumental in advancing my research capabilities, expanding the available resources for Polish literary studies, and opening up new possibilities for future investigations. The skills and knowledge acquired during this period will undoubtedly continue to inform and enhance my academic work in the years to come.