SylLab – software for semi-automatic stylometric analysis for poetry

Rykowska, Aleksandra
aleksandra.rykowska@student.uj.edu.pl
Jagiellonian University in Kraków, Poland

By focusing on the quantitative ratios of the different types of speech sounds in a poem, it is not so much possible to establish the authorship of the text (which remains a potential topic for future research), but it allows us to determine the mood of a poem.

Considering the semantics of a poem, all attempts to determine the mood of a poem using stylometric methods operate on words using a 'bag-of-words method' (e.g., Kao and Jurafsky, 2015) and it is worth noticing that the forementioned study using word-based methods did not provide any successful results. The most intriguing work seems to confirm the 'semantic halo' of verse theory (Šeļa et al., 2022), which supports the conjecture that differen-ces at the phonetic level for the poems with the same versifica-tion system (and having the same semantic halo) will differ in the mood when differences at the phonetical level are observed in them. Knowing this might then lead to the research question – if the meter imposes a certain topic to the poem, then the differences regarding the mood of a poem might be visible on the phonetic level of the text.

The idea is based on Maria Dłuska's (manual and non-computational) method of dealing with Polish accentual-syllabic verse poetry (Dłuska, 2001). On the basis of the steps she initially proposed for the analysis, an algorithm was established, which was then used to prepare the SylLab computer program that automates most of the steps involved in a similar analysis of a poem. The program is prepared in the Eclipse environment, in the object-oriented programming language Java. The program interface is divided into 6 tabs, each of them responsible for a different stage of the verse analysis.

The program operates on texts written both orthographically and phonetically, but because of its focus on the sound layer of the text, the most accurate results are obtained when working on text already transcribed into the IPA alphabet. Although the program automatically transcribes the orthographically written text, due to the redundancy of certain features and phonetic similarities in the further steps of the analysis, the transcription generated by the program is subjected to manual analysis by a human operator to detect and eliminate all inconsistencies if necessary.

The next step that the program performs is to divide the words in each line into syllables. The program calculates the sonority trait of the speech sounds. They are then divided into classes, each annotated with a given sonority value - the syllable boundary runs where there is a sharp rise or fall in the sonority value for the speech sounds present in the piece (Śledziński, 2016). A tilde mark separates all syllables. Further on, the program marks stressed syllables and displays a diagram of the rhythmic structure of the analysed poem. The program considers word boundaries and marks the syllables accented according to the rules of Polish accentuation, where the accent is persistent and paroxytonic in most words (Wiśniewski, 2007). A list of proclitics and enclitics has also been created, which, when detected by the program, are 'attached' in the analysis to the following word due to them not having their own stress. The results of the versological analysis of the works are saved to text files (as a plain text and in JSON format), which is a prelude to the creation of the planned corpus of Polish poetry annotated in terms of versification.

The program further determines the characteristics of the individual speech sounds. The most important features here are if they are opened/closed and palatalised. According to Dłuska's original analysis, closed and palatal sounds and their accumulation in work have a significant impact on the mood of the poem - a predomi-nance of closed and palatal speech sounds is associated with the poem's positive valence. In contrast, a higher proportion of opened and not palatalised sounds is associated with its gloomy mood. The program then generates a CSV-encoded file with the results of the quantitative analysis of the speech sounds included in the poem, together with an indication of where each speech sound is located in the poem and the sums of different classes of sounds in order to compare the results with the reference corpus.

The program presented here is just a promising seed of what results it could present once its functionality is developed with more advanced information technologies, such as the use of machine learning and NLP techniques.

Bibliography

Dłuska, M. (2001): „Kazimierz Wierzyński «Gdzie nie posieją mnie»", in *Liryka polska. Interpretacje*, edited by Jan Prokop, Janusz Sławiński, 293–319, Gdańsk: słowo/obraz terytoria.

Kao T., and Jurafsky D. (2015): "A computational analysis of poetic style: Imagism and its influence on modern professional and amateur poetry", in *Linguistic Issues in Language Technology*, volume 12 - Literature Lifts up Computational Linguistics, CSLI Publications.

Šeļa A, Plecháč P., and Lassche A. (2022): "Semantics of European poetry is shaped by conservative forces: The relationship between poetic meter and meaning in accentual-syllabic verse", PLoS ONE 17(4): e0266556.

Śledziński, D. (2016): „Tworzenie reguł dla programu dzielącego tekst w języku polskim na sylaby", Biuletyn Polskiego Towarzystwa Językoznawczego, vol. 72: 151-161.

Wiśniewski, M. (2007): „Zarys fonetyki i fonologii współczesnego języka polskiego", Toruń: Wydawnictwo Uniwersytetu Mikołaja Kopernika.