Maciej Maryl, Institute of Literary Research of the Polish Academy of Sciences

Maciej Piasecki, Wrocław University of Technology

Ksenia Młynarczyk, Wrocław University of Technology

Contact: maciej.maryl@ibl.waw.pl

Submission type: Experience Report

Text Clustering Methods in Literary Analysis of Weblog Genres

The existing typologies of weblog genres are based on the blog topic (e.g. cooking blogs, travels, business) or its medium (e.g. vlogs, picture logs), Maryl, Niewiadomski and Kidawa conducted an interpretive study on the sample of 322 popular Polish blogs[1]. They adopted a new-rhetorical approach, basing on Carolyn Miller's concept of genre as a social action[2], concentrating mostly on: the blog's communicative purpose and functions.

Basing on the close reading the team created an empirical-conceptual typology which entailed following genres: diaries (subjective, self-referential discourse), reflection (subjective discourse on universal matters), criticism (subjective and expert discourse on general issues), information (objective facts), filter (gateway to the existing web content), advice (subjective and expert instructions on particular issues), modelling (serving as a role model for readers) and fictionality (description of fictional events). Weblogs in the sample were coded by three coders with 69% average pairwise percent agreement and Cohen's kappa of .622. Such a moderate agreement could be attributed to the fact that the genres represent ideal types, and most of the blogs share features of more than one genre.

The subsequent study aimed at linguistic verification of this typology with the use of linguistic tools following the distant reading perspective. The corpus of blogs was downloaded with the use of BlogReader, a tool for semi-automatic acquisition of weblogs.[3]

In order to find groups of blogs of similar in style, we have followed the typical paradigm of stylometry. Blogs were described by feature vectors that were next filtered, transformed, and automatically clustered. As we were looking for the original style of the blog authors

[1] Maciej Maryl, Krzysztof Niewiadomski, Maciej Kidawa, „Empirically Generated Typology of Weblog Genres" (Paper currently under review).
[2] Miller, Carolyn R. „Genre as Social Action" *Genre and the New Rhetoric*, eds. Aviva Freedman and Peter Medway, London: Taylor & Francis, 1994.
[3] Marcin Oleksy, Jan Kocoń, Maciej Maryl, Maciej Piasecki, "Linguistic analysis of weblog genres", a paper presented during the conference *Practical Applications of Linguistic Corpora*, University of Łódź, November 2014.

comments were omitted and posts of the one blog merged together before processing. We tried to avoid features sensitive to the semantics of the blog content. However, in the same time we wanted to explore a wider range of features as an extension of the traditional most frequent words of the given language. For the experiments we selected for the description: most frequent lemmas from the Polish National Corpus, punctuation signs, selected Proper Name classes, grammatical classes and bigrams of the grammatical classes. Blogs were preprocessed by language tools of CLARIN-PL. Features were extracted with the help of the Fextor systems and Cluto clustering tool was used to find groups. We will present several different experiments, their automated evaluation against the blog types defined manually and the influence of different features on the results.

This study is a part of an ongoing research project under the auspices of DARIAH-PL Digital Philology Working Group.