Big Data, Small Literatures (An Introduction)

Lucija Mandić, Ivana Zajc

The digital turn, which is affecting virtually every aspect of our daily lives, has introduced new research approaches into literary studies as well. In a time where advanced technologies, including artificial intelligence, dominate global discourse, it is increasingly important for literary studies to explore the methodological potential of the latest technological advances. The use of computational tools is able to create an interdisciplinary space at the intersection of the humanities and information technology, fueling the growth of digital humanities. In literary studies, computer-assisted analyses of literature are usually collected under the umbrella term of *distant reading*, with *macroanalysis* and *algorithmic criticism* serving as near synonyms. While the paradigmatic turn to distant reading aims to decentralize the literary world-system, researchers predominantly focus on major world languages due to the easy accessibility of digital archives in those languages. This trend has led to the development of methodologies of digital humanities, which are rarely adapted to the study of literatures in less dominant languages, such as Slovenian. Distant reading involves computational analysis of literature using language models and machine learning algorithms, whereby computers extract data from corpora of literary texts in order to store and process them using quantitative methods. Such analyses can encompass extensive collections comprising thousands of literary works, zoom in on individual texts or excerpts, or focus on metadata related to the literary system. Computer-assisted reading based on statistics and related mathematical branches extracts specific formal or semantic data and identifies patterns that would remain imperceptible to close reading. It is important to note, though, that close and distant reading, far from being mutually exclusive, are complementary approaches, as evidenced by the articles clustered in this issue of *Primerjalna književnost* under the title "Big Data, Small Literatures."

The cluster brings together various digital approaches to literature in smaller languages, such as Czech, Romanian, Basque, and Slovenian. The articles address a gap produced by computational literary studies as they predominantly work with digitized texts in globally spoken

languages, while research on less dominant languages faces the challenge of digitizing literary texts and establishing necessary digital infrastructure before it can analyze and interpret texts. Originally proposed as a method of studying world literature, distant reading is effectively faced with the obstacle of an increasingly widening digital language divide which has made more than 3,700 languages virtually non-existent in the digital environment.

As a cluster, "Big Data, Small Literatures" explores digital literary studies in small literatures from two perspectives: the first group of articles concentrates on database construction, annotation, and metadata analysis, while the second group presents individual case studies in which machine learning and natural language processing have been applied to already existing corpora. In the first article, **Vlad Pojoga** presents an analysis of the Romanian literary journal *Convorbiri literare* as part of a broader initiative to construct a corpus of nineteenth-century Romanian poetry. The examination of hand-extracted metadata sheds light on the canonization process of late nineteenth-century Romanian authors based on their integration into the networks of world literature. Similarly, **Katja Mihurko**, **Ivana Zajc**, **Darko Ilin**, and **Mila Marinković** delve into networks of influence, but shift the focus to personal correspondences to address issues of gender pertaining to individual writers. Their semantic analyses and exploration of metadata show that non-literary corpora are an indispensable part of digital literary scholarship, as they offer new insights into the production and reception of literary texts. **Ranka Stanković**, **Cvetana Krstev**, and **Duško Vitas** take us back to exclusively literary corpora as they examine the ELTeC-srp corpus and its recent upgrade with enriched metadata from the Wikidata database. Their article outlines the challenges encountered by the designers of the corpus, from text digitization to automatic annotation, and proposes innovative solutions facilitated by interdisciplinary collaboration. This is followed by the article by **Silvie Cinková**, **Petr Plecháč**, and **Martin Popel**, which provides a kind of transition between the first and second parts of the cluster. The article describes the process of evaluating the automatic annotator UDPipe using nineteenth-century Czech poetry as a case study—a crucial step in the development of a linguistically annotated literary corpus. Given the distinctive stylistic word order prevalent in Czech poetry, the linguistic model trained on prose texts proved partially insufficient for the analysis of poetry.

Transitioning from corpus construction to textual analysis, **Dominika Werońska** applies a stylometric tool Stylo, which measures

the frequencies of the most frequent words, to a corpus of Basque novels. Her findings reveal that stylistic fluctuations attributable to translation or dialectical influences rival the authorial stylistic signal. Stylo is used by **Ivana Zajc** as well, as she adapts the tool to analyze a corpus of Slovenian mid-nineteenth- to early-twentieth-century fiction, highlighting the dominance of the authorial style over genre indicators. Her analysis of works by Ivan Cankar reveals the development of the authorial style, with Cankar's earlier period standing out in particular with its stylistic homogeneity. In the next article, **Andrejka Žejn**, **Marko Pranjić**, and **Senja Pollak** examine the authorial style in Slovenian prose of the same period by utilizing contextual word embeddings. Their computational semantic analysis unveils semantic shifts in the works of Josip Jurčič and Ivan Cankar, demonstrating that they are particularly evident in broader semantic domains. In the final aticle, **Lucija Mandić** broadens the analytical scope to encompass the entire Slovenian long narrative prose of the so-called long nineteenth century. Semantic analysis using word embeddings reveals a distinct convergence of semantic fields related to culture and politics across both canonized and non-canonized texts, both confirming and expanding the findings of traditional Slovenian literary studies pertaining to the so-called Prešernian structure of Slovenian canonical literature.

The editorial work on this thematic section of *Primerjalna književnost* has been funded by the Slovenian Research and Innovation Agency in the framework of the research program "Studies in Literary History, Literary Theory and Methodology" (P6-0024) at the Research Centre of the Slovenian Academy of Sciences and Arts and the research project "Transformations of Intimacy in the Literary Discourse of Slovene 'moderna'" (J6-3134) at the University of Nova Gorica.