PODIPLOMSKA ŠOLA ZRC SAZU

Lucija Mandić

RAČUNALNIŠKO BRANJE SLOVENSKE PRIPOVEDNE PROZE DOLGEGA 19. STOLETJA V KONTEKSTU LITERARNE KANONIZACIJE

Doktorska disertacija

Mentorica: doc. dr. Andrejka Žejn

Somentorica: doc. dr. Monika Deželak Trojar

Ljubljana, 2024

Povzetek

Doktorska disertacija *Računalniško branje slovenske pripovedne proze dolgega 19. stoletja v kontekstu literarne kanonizacije* razvije novo branje slovenske daljše pripovedne proze z uporabo računalniško podprtih metod za raziskovanje literature, ki so se v zadnjih desetletjih razvile v okviru digitalne humanistike. Osrednje raziskovalno vprašanje disertacije je, ali in kako lahko uporaba računalniških metod, s katerimi lahko razširimo predmet raziskovanja s posameznih literarnih del, avtorjev in smeri na celotno produkcijo obravnavanega obdobja (v tem primeru t. i. dolgega 19. stoletja), prispeva k razumevanju procesa literarne kanonizacije na Slovenskem. Kanonizacijo, ki jo je tradicionalna literarna veda obravnavala predvsem z gledišča družbenega konteksta literarnih tekstov, bomo z uporabo metod računalniške literarne vede obravnavali tudi na tekstualni ravni. Tema disertacije zahteva interdisciplinarni pristop, katerega osnovno vodilo je prepletanje kvantitativnih in kvalitativnih analiz. Kvantitativna komponenta temelji na metodah računalniške literarne vede, tj. na računalniško podprtih tehnikah za analizo literarnih del, medtem ko se kvalitativni del opira na ugotovitve literarnega zgodovinopisja pa tudi občega zgodovinopisja in interpretativnih družbenih ved.

Predmet disertacije je izvirna slovenska daljša pripovedna proza, ki je nastala v nekaj manj kot stoletnem obdobju od njenih začetkov do konca prve svetovne vojne in razpada Avstro-Ogrske. Izbiro obdobja je narekoval razvojni lok slovenske pripovedne proze od prve slovenske izvirne prozne pripovedi (1836), prek začetka in razvoja slovenskega romana in vzporednega pojava mohorjanskih povesti v drugi polovici 19. stoletja, do estetskega vrhunca v času moderne. Poleg teh zgodovinskih premen na področju literature pa izbrano obdobje zaznamujejo tudi pomembne družbene spremembe, kot sta vstop žensk v javno sfero ter razvoj slovenskega nacionalnega gibanja od njegovih začetkov do razpada Avstro-Ogrske v prvih desetletjih 20. stoletja.

Za disertacijo je bil sestavljen elektronski Korpus daljše slovenske pripovedne proze (KDSP), ki vsebuje 261 izvirnih pripovednih proznih besedil v slovenščini, daljših od 20.000 besed, ki so izšla med letoma 1836 in 1918. Korpus je avtomatsko tokeniziran, lematiziran, stavčno segmentiran in opremljen z oblikoskladenjskimi oznakami in oznakami imenskih entitet. Besedila v korpusu so opremljena z bibliografskimi in biografskimi (avtorje in avtorice zadevajočimi) metapodatki ter so dodatno kategorizirana glede na desetletje izida, dolžino, zvrst, tematiko in kanoničnost, pri čemer je 209 besedil nekanoničnih, preostalih 52 pa kanoničnih. Za merilo kanoničnosti je bil izbran sodobni slovenski srednješolski kanon, in sicer so bila kot kanonična opredeljena tista dela, ki so citirana ali vsaj omenjena v slovenskih srednješolskih berilih, izdanih od leta 1980 naprej.

Disertacija je razdeljena na tri vsebinske sklope. Prvi sklop začenja niz metapodatkovnih analiz o zvrstni in žanrski strukturi korpusa KDSP v odnosu do ravni kanoničnosti, ki pokaže, da se v literarni kanon niso uvrstila tista literarna dela, ki ne ustrezajo sodobnim pojmovanjem literarnih zvrsti, nekanonična dela v korpusu pa so zvrstno in žanrsko bolj raznolika kakor kanonična. Uporaba metode modeliranja tem z orodjema Gensim in Mallet pokaže, da je kanon tudi na ravni tem homogenejši kakor nekanon. Tako v kanoničnih kakor v nekanoničnih delih so najbolj zastopane meščanske in kmečke teme, a so te v kanoničnih delih navzoče v bistveno večjem deležu kakor v nekanoničnih. Nekanonična dela so tematsko bolj raznolika. Večja tematska homogenost kanona je opazna tudi na ravni notranje strukture literarnih tekstov. Modeliranje tem izpostavi Ivana Cankarja, najbolj kanoničnega avtorja celotnega korpusa kot avtorja, čigar dela so tematsko najbolj homogena.

V drugem sklopu so s pomočjo stilometričnega orodja Stylo, uporabe konkordančnika Sketch Engine in analize oblikoskladenjskih oznak kanonična in nekanonična dela primerjana z vidika rabe besednih vrst. Z orodjem Stylo so s kontrastivno analizo pridobljene besede, ki so značilne za kanonična dela v primerjavi z nekanoničnimi na eni strani in za nekanonična dela v primerjavi s kanoničnimi na drugi, s konkordančnikom Sketch Engine pa sta podkorpusa primerjana na podlagi ekstrakcije ključnih besed. Primerjava rezultatov obeh metod pokaže, da v kanoničnih delih statistično bolj izstopajo predmetnopomenske, v nekanoničnih delih pa funkcijske besede. Te rezultate dodatno osvetli analiza razmerja med pojavnicami in različnicami, ki razkrije, da imajo kanonična dela na ravni predmetnopomenskih besed manj raznoliko besedišče kakor nekanonična. V nadaljevanju sklopa je izvedena semantična analiza besedišča s pomočjo vektorskih vložitev besed. Jezikovni model pokaže prekrivnost semantičnih polj politike in kulture predvsem v kanonični, a tudi v nekanonični literaturi, kar empirično potrdi in razširi ugotovitve tradicionalne literarne vede o t. i. prešernovski strukturi slovenske književnosti obdobja, ki ga zajema raziskava.

Tretji sklop zastavlja vprašanje spola v kontekstu kanonizacije. Metapodatkovne analize potrdijo marginalnost avtoric v slovenskem srednješolskem kanonu. Analiza imenskih entitet podobno pristopi k ženskim literarnim likom, ki so tako v kanoničnem kakor v nekanoničnem delu korpusa zastopane v manjši meri kakor moški literarni liki. Dodatna primerjava zastopanosti ženskih in moških literarnih likov v delih avtoric in avtorjev pokaže, da tako avtorji kakor avtorice omenjajo moške like pogosteje kakor ženske like, čeprav je razmerje med omembami ženskih likov in omembami moških likov pri avtoricah bolj uravnoteženo kakor pri avtorjih. Analiza s pomočjo vektorskih vložitev besed dopolni analizo frekvenc imenskih entitet in osvetli še hierarhijo med spoloma, ko gre za način pripovedovanja o ženskah na eni strani in o moških na drugi. Več jezikovnih modelov, naučenih na različnih naborih besednih vrst, pokaže, da avtorji pri opisovanju moških dosledno izbirajo drugačne nabore besednih vrst kakor pri opisovanju žensk, pri avtoricah pa so razlike v opisovanju moških in žensk bistveno manjše. Primerjava med kanoničnimi in nekanoničnimi deli ter primerjava med deli avtorjev in deli avtoric skupaj razkrijeta, da avtorice v nasprotju z avtorji v nekaterih primerih opisujejo moške in ženske na način, ki je značilen za kanonična dela, čeprav njihova dela ostajajo zunaj kanona.

V vseh treh sklopih se kvantitativne metode, podprte z računalniškimi orodji, izkažejo za uspešen pristop k razumevanju procesa kanonizacije slovenske pripovedne proze dolgega 19. stoletja. Rezultati izvedenih analiz ponekod empirično potrjujejo ugotovitve tradicionalne literarne vede, drugod pa ponujajo sklepe, ki bi tradicionalnim metodam natančnega branja bržkone ostali nedostopni. To velja zlasti za ugotovitve o relativni homogenosti kanona, ki se bolj približa monotematskosti, t. i. prešernovski strukturi in patriarhalnosti. Disertacija tako omogoča sklep, da digitalnohumanistične metode kažejo na tekstualno razsežnost kanonizacije, ki jo je slovenska literarna veda doslej raziskovala predvsem kontekstualno.

Ključne besede: oddaljeno branje, računalniška literarna veda, literarna kanonizacija, 19. stoletje, slovenska pripovedna proza

Abstract

The doctoral dissertation *A Computational Reading of Slovenian Narrative Prose of the Long 19th Century in the Context of Literary Canonization* offers a new interpretation of Slovenian long narrative prose through the application of computational methods of literary research, which have evolved over recent decades within the framework of digital humanities. At the core of the dissertation lies the research question whether the use of these methods, which allow the analysis to extend beyond individual works, authors, and trends to encompass the entire literary production of the period (in this case the so-called long nineteenth century), can enhance our understanding of literary canonization in Slovenian literature. Canonization, traditionally examined through the lens of social context and using methods of close reading, will be approached at the textual level as well, using computational techniques. The topic of the dissertation requires an interdisciplinary approach, the basic guideline of which is the interweaving of quantitative and qualitative analyses. The quantitative aspect relies on methods of computational literary studies comprising computer-assisted technics of analyzing literary texts on a larger scale, while the qualitative component draws upon the state of the art in literary history as well as general history and interpretative social studies.

The subject of the dissertation is original Slovenian long narrative prose produced over a period of less than a century, from its beginnings up until the end of the First World War. The selection of this period is based on the evolution of Slovenian prose, starting with the publication of the first original Slovenian prose narrative in 1836, continuing through the emergence and growth of the Slovenian novel and the parallel rise of stories published by the Hermagoras Society (Mohorjeva družba) in the latter half of the nineteenth century, and culminating aesthetically during the *moderna* period of the first two decades of the twentieth century. Alongside these literary shifts, the period is marked by significant social transformations, such as the entry of women into the public sphere and the development of the Slovenian national movement from its emergence and up to its role in in the collapse of Austria-Hungary in the opening decades of the twentieth century.

For the purposes of the dissertation, the electronic Corpus of Longer Slovenian Narrative Prose (KDSP) was compiled, collecting 261 original narrative prose works that exceed 20,000 words, are written in the Slovenian language and were published between 1836 and 1918. Each text is automatically annotated with bibliographic and biographical metadata and further categorized by decade of publication, length, text form, theme, and canonicity, with 52 of the texts treated as canonical and the remaining 209 as non-canonical. The corpus is fully processed—tokenized, sentence-segmented, lemmatized, and tagged with morphosyntactic features and named entities—and publicly accessible. As the representative criterion of canonicity, the contemporary high school canon was chosen: canonical texts were defined as those that are cited or are at least referred to in Slovenian high school textbooks published from 1980 onward.

The dissertation is divided into three main sections. The first section begins with metadata analyses aimed at examining form and genre in relation to canonicity. The analysis of the KDSP corpus reveals that works excluded from the literary canon often do not align with contemporary formal classifications and are more diverse in terms of genre and form than their canonical counterparts. Topic modeling using Gensim and Mallet tools further indicates that canonical works tend to be more homogeneous in terms of themes in comparison with non-canonical. Both canonical and non-canonical works prominently feature themes related with the bourgeoisie and peasantry, which, however, appear in canonical literature with a significantly higher proportion than in non-canonical texts, which display a greater thematic diversity. This

thematic homogeneity of the canon is also evident in the internal structure of literary texts, with topic modeling singling out the extraordinary thematic homogeneity of the works by Ivan Cankar, the most canonical author in the corpus.

The second section employs the stylometric tool Stylo, the Sketch Engine concordancer, and morphosyntactic analysis to compare canonical and non-canonical works based on their use of word types. Both contrastive analysis using Stylo to identify characteristic words of each group and keyword extraction using Sketch Engine reveal that content words are more statistically prominent in canonical works, whereas function words dominate in non-canonical ones. This finding is supported by an analysis of lexical diversity which includes calculating the type/token ratio for each word type to show that canonical works employ a less varied vocabulary in terms of content words. Further semantic analysis using word embeddings reveals overlaps between the semantic fields of politics and culture in both canonical and non-canonical texts, with canonical literature showing a stronger connection. These results empirically confirm and extend the findings of traditional literary studies about the so-called Prešernian structure of Slovenian literature of the period covered by the KDSP corpus.

The third section explores the issue of gender in the context of literary canonization. Metadata analysis highlights the marginalization of female authors in the Slovenian secondary school canon. An analysis of named entities reveals that female protagonists are underrepresented compared to their male counterparts in both canonical and non-canonical works. However, a closer comparison between male and female authors shows that, while both tend to feature more male characters than female ones, female authors portray a more balanced ratio between male and female characters. Word embeddings trained on different selections of word types shed further light on gender representation, demonstrating that male authors consistently choose different sets of word types when describing men than when describing women, while female authors show less disparity in their descriptions than male authors. Together, a comparison between canonical and non-canonical works and a comparison between male and female authors reveal that, despite adopting the stylistic features of canonical literature, female authors are nevertheless excluded from the canon.

In all three sections, quantitative methods supported by computational tools demonstrate their efficacy in enhancing the understanding of the canonization of Slovenian longer narrative prose of the long nineteenth century. The results of the analyses presented in the dissertation, while empirically corroborating certain findings of traditional literary scholarship, often provide new insights that traditional methods of so-called close reading could hardly reach, notably findings concerning the relative homogeneity of the canon, which is more monothematic, more receptive to the so-called Prešernian structure, and more patriarchal than the rest of the corpus. Consequently, the dissertation supports the conclusion that methods of digital humanities reveal a textual dimension of the process of canonization that has mostly been approached contextually in Slovenian literary studies.

Keywords: distant reading, computational literary studies, literary canonization, nineteenth century, Slovenian narrative prose