Veliki podatki, male literature (predgovor)

Lucija Mandić, Ivana Zajc

Digitalni obrat, ki posega v tako rekoč vsa področja našega vsakdana, je tudi v literarno vedo prinesel nove raziskovalne pristope. V času, ko so se napredne tehnologije, vključno z umetno inteligenco, znašle v središču globalnih razprav, postaja tudi za literarno vedo vse pomembneje, da smiselno izkoristi možnosti, ki jih ponuja tehnološki razvoj na področju literarnovedne metodologije. Uporaba računalniških orodij namreč odpira prostor za interdisciplinarne raziskave na preseku med humanistiko in informacijsko tehnologijo, kjer se vse intenzivneje razvija digitalna humanistika. Na področju literarne vede za računalniško podprte analize literature pogosto uporabljamo termin *oddaljeno branje*, uveljavili pa sta se tudi poimenovanji *makroanaliza* in *algoritemska kritika*. Čeprav paradigmatski obrat k oddaljenemu branju temelji na želji po decentralizaciji sistema svetovne literature, raziskovalci in raziskovalke s tovrstnimi pristopi zaradi dostopnosti digitalnih arhivov najpogosteje posegajo po literaturah v svetovnih jezikih. To je botrovalo tudi specifičnemu razvoju samih metod digitalne humanistike, ki so le redko prilagojene raziskovanju literatur v manjših jezikih, tudi v slovenščini. Pri oddaljenem branju gre za t. i. računalniško branje, ki literaturo analizira z uporabo jezikovnih modelov in strojnega učenja, pri čemer računalniki iz korpusov literarnih besedil pridobivajo podatke ter jih shranjujejo in obdelujejo s kvantitativnimi metodami. Tovrstne analize lahko zajemajo obsežne zbirke, ki vsebujejo po več tisoč literarnih del, lahko se ukvarjajo s posameznimi besedili ali njihovimi deli, včasih pa se osredotočajo tudi na metapodatke, povezane z literarnim sistemom. Z računalniškim branjem, ki temelji na statistiki in sorodnih vejah matematike, je mogoče izluščiti specifične formalne ali semantične informacije in v njih prepoznati vzorce, ki so s pomočjo natančnega branja nezaznavni. Natančno branje kot klasična metoda literarne vede in oddaljeno branje pa se ne izključujeta, temveč se dopolnjujeta, kot pokažejo tudi raziskave v pričujočem tematskem sklopu *Primerjalne književnosti*.

Članki tematskega sklopa »Veliki podatki, male literature« prinašajo nabor digitalnih pristopov k literarnovednim raziskavam v manjših literaturah, kakršne so na primer češka, romunska, baskovska in nenazadnje

tudi slovenska. S tem raziskovalke in raziskovalci zapolnjujejo vrzel v računalniških raziskavah književnosti, ki večinoma temeljijo na digitaliziranih besedilih, ta pa še vedno večinoma pripadajo literaturam v svetovnih jezikih. Pri raziskovanju literatur v manj razširjenih jezikih smo tako postavljeni pred izziv, kako zagotoviti digitalizacijo literarnih besedil in vso drugo potrebno digitalno infrastrukturo. Oddaljeno branje, katerega predmet raziskovanja naj bi bila svetovna literatura, se v praksi sooča z oviro vse bolj očitnega digitalnega jezikovnega razkoraka (ang. *digital language divide*), zaradi katerega več kot 3.700 jezikov v digitalnem okolju tako rekoč ne obstaja.

Tematski sklop obravnava problematiko digitalne literarne vede na področju malih literatur z dveh vidikov: članki v prvem delu se posvečajo izgradnji podatkovnih zbirk, njihovi anotaciji in analizam metapodatkov, v drugem delu pa so zbrane študije primerov, v katerih so metode strojnega učenja in obdelave naravnega jezika aplicirane na že sestavljene korpuse. V prvem članku **Vlad Pojoga** predstavi analizo romunskega literarnega časopisa *Convorbiri literare* kot preliminarno študijo, ki je nastala v procesu izgradnje obsežnega korpusa romunske poezije 19. stoletja. Ročno izluščeni metapodatki o lokalni literarni produkciji, prevodni literaturi in medliterarnih vplivih osvetljujejo proces kanonizacije romunskih avtorjev poznega 19. stoletja z vidika vpetosti v svetovno književnost. **Katja Mihurko**, **Ivana Zajc**, **Darko Ilin** in **Mila Marinković** prav tako proučujejo omrežja vpliva med pisatelji in pisateljicami, a težišče prenesejo na osebne korespondence. Na primeru metapodatkovnih in semantičnih analiz korpusa Pisma pokažejo, da so tudi neliterarni korpusi nepogrešljiv del digitalne literarne vede, saj ponujajo nov vpogled v produkcijo pa tudi recepcijo literarnih besedil. Članek **Ranke Stanković**, **Cvetane Krstev** in **Duška Vitasa** se vrača k izključno literarnemu korpusu, namreč ELTeC-srp, ki je v novi različici obogaten z metapodatki iz podatkovne zbirke Wikidata. Članek predstavi izzive, s katerimi so se soočali oblikovalci korpusa vse od digitalizacije besedil do avtomatske anotacije, ter ponuja inovativne rešitve, ki jih omogoča interdisciplinarno sodelovanje. Vezni člen med prvim in drugim delom sklopa je članek **Silvie Cinkove**, **Petra Plecháča** in **Martina Popela**, ki opisuje proces evalvacije avtomatskega označevalnika UDPipe na primeru češke poezije 19. stoletja kot nujen del procesa sestavljanja jezikovno označenega literarnega korpusa. Ker je za češko poezijo značilen slogovno zaznamovani besedni red, se je izkazalo, da je jezikovni model, ki je bil naučen na proznih besedilih, za poezijo pomanjkljiv.

V drugem delu tematski sklop prehaja s sestavljanja korpusov in analiz metapodatkov k tekstualnim analizam s pomočjo računalniških

orodij. **Dominika Werońska** aplicira stilometrično orodje Stylo, ki meri frekvence najpogostejših besed, na korpus baskovskih romanov. Z analizo pokaže, da sta spremembi sloga, ki se izkažeta za posledico prevajanja iz tujih jezikov ali vpliva narečij, prav tako močni kakor avtorski signal. **Ivana Zajc** prilagodi uporabo orodja Stylo analizi korpusa slovenske proze iz obdobja realizma in moderne, pri tem pa pokaže, da je avtorski signal močnejši kakor na primer žanrski. Podrobneje se posveti Ivanu Cankarju, pri katerem stilometrična analiza razkrije razvoj avtorskega sloga, pri čemer kot slogovno bolj homogeno izstopa predvsem Cankarjevo zgodnejše obdobje. Tudi **Andrejka Žejn**, **Marko Pranjić** in **Senja Pollak** pod drobnogled vzamejo slovensko prozo realizma in moderne, a za analizo avtorskega sloga uporabijo metodo kontekstualnih vektorskih vložitev besed. S pomočjo računalniške semantične analize zaznajo pomenske premike v delih Josipa Jurčiča in Ivana Cankarja ter ugotovijo, da so ti najizrazitejši pri splošnejših pomenskih poljih. V zadnjem prispevku tematskega sklopa **Lucija Mandić** razširi predmet analize s kanoniziranih avtorjev na celotno slovensko daljšo pripovedno prozo t. i. dolgega 19. stoletja. Semantična analiza vektorskih vložitev besed razkrije izrazito prekrivanje semantičnih polj kulture in politike tako v kanonizirani kakor v nekanonizirani literaturi, kar naveže na preddigitalne ugotovitve slovenske literarne vede o t. i. prešernovski strukturi.

Tematski sklop je nastal na Inštitutu za slovensko literaturo in literarne vede ZRC SAZU in na Raziskovalnem centru za humanistiko Univerze v Novi Gorici v okviru raziskovalnega programa »Literarnozgodovinske, literarnoteoretične in metodološke raziskave« (P6-0024) in raziskovalnega projekta »Transformacije intimnosti v literarnem diskurzu slovenske moderne« (J6-3134), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.