Analiza 90 slovenskih romanov in opusa Ivana Cankarja z računalniško stilometrijo

Ivana Zajc

Univerza v Novi Gorici, Raziskovalni center za humanistiko, Vipavska 13, 5000 Nova Gorica, Slovenija
https://orcid.org/0000-0002-5502-6772
ivana.zajc@ung.si

Članek predstavi oddaljeno branje slovenske književnosti iz obdobij slovenskega literarnega realizma in moderne. Prvi del članka vsebuje prikaz rezultatov računalniške stilometrične analize korpusa 90 slovenskih romanov iz omenjenih obdobij, izvedene v programskem okolju R s paketom Stylometry with R. Analiza je pokazala, da je ključni signal, po katerem se vključeni romani medsebojno razlikujejo, avtorski, pri čemer najbolj izstopajo romani Pavline Pajk in Ivana Cankarja. Ker je izstopanje najbolj izrazito pri Pavlini Pajk, članek predstavi tudi stilometrično raziskavo besedišča, ki je za njene romane najznačilnejše. Na tej podlagi članek ponudi sklep, da k specifičnosti njenih romanov med drugim prispeva žanr sentimentalnega ljubezenskega romana, ki zaznamuje njen opus. V drugem delu članka je z isto metodo analiziran Cankarjev literarni opus, v korpus pa so poleg romanov vključena tudi njegova druga dela iz različnih ustvarjalnih obdobij. Iz rezultatov analize sta razvidna obdobni in žanrski razvoj Cankarjevega literarnega sloga.

Ključne besede: slovenski roman / Cankar, Ivan / oddaljeno branje / računalniška stilometrija / digitalna humanistika

Digitalne metode v literarni vedi

Kadar k literarnim delom pristopamo s pomočjo oddaljenega branja, jih raziskujemo s kvantitativnimi metodami, v sodobnem času z uporabo računalniške tehnologije.[1] Pristop oddaljenega branja je leta 2000 predlagal Franco Moretti, pozneje pa je o tovrstnem raziskovanju literature Matthew L. Jockers vplivno govoril kot o makroanalizi, Stephen

[1] Članek je nastal v okviru raziskovalnega projekta »Transformacije intimnosti v literarnem diskurzu slovenske moderne« (J6-3134), ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

Ramsay pa kot o algoritemski kritiki. Prvi poskusi oddaljenega branja so bili izvedeni preštevno, brez uporabe računalniške tehnologije,[2] danes pa kvantitativne raziskave praviloma analizirajo obsežnejše literarne zbirke ali druge podatke s pomočjo digitalnih orodij in statističnih pristopov. Pri sodobnem oddaljenemu branju ne gre le za analizo besedišča v književnih delih, ampak za razumevanje pisanja kot kompleksnega polja odnosov, ki ga je mogoče kvantitativno modelirati (Underwood, »Distant Reading«). Tovrstne raziskave se lahko nanašajo na različne elemente, na primer na korpus literarnih besedil v celoti, na dele literarnega besedila (npr. na dialoge ali karakterizacijo likov), na posamezno besedilo ali na metapodatke o književnosti in literarnem sistemu. Raziskave se lahko ukvarjajo tudi z ustaljenimi predpostavkami literarne zgodovine, ki jih lahko s pomočjo digitalnih tehnologij prevpraašujemo, na primer s časovnim potekom literarnozgodovinskih obdobij in smeri ter z njihovimi lastnostmi. S tem področje literarne vede obogatimo z novim tipom kvantitativnih podatkov (Eve 153), ki jih interpretiramo s kvalitativnimi metodami (Underwood, »Distant Reading«). Kvantitativni podatki so se v literarni vedi uporabljali že prej, vendar so se z vzponom digitalnih informacijskih tehnologij možnosti za raziskave razširile. Massimo Salgaro izpostavi naslednje razlike med kvantitativnim in kvalitativnim raziskovanjem v literarni vedi: drugačen odnos med teorijo in raziskovanjem: kvantitativne raziskave so deduktivne in strukturirane po stopnjah, medtem ko je kvalitativno raziskovanje bolj odprto; različne tipe konceptov: kvantitativno raziskovanje predvideva operativne koncepte, medtem ko so koncepti pri kvalitativnem raziskovanju odprti in se nenehno razvijajo; različne tipe psihološke interakcije med raziskovalcem in opazovanim fenomenom: kvantitativno raziskovanje je znanstveno in nevtralno, medtem ko kvalitativno raziskovanje predvideva empatično identifikacijo s predmetom raziskovanja oziroma prevzemanje njegove perspektive; razlike v reprezentativnosti podatkov: kvantitativno raziskovanje teži k statistično reprezentativnim vzorcem, medtem ko kvalitativno raziskovanje izhaja iz posamičnih primerov, ki niso statistično pomembni; razlike v odnosu do matematičnih in statističnih tehnik: v nasprotju s kvalitativnim raziskovanjem kvantitativno raziskovanje te tehnike intenzivno uporablja; različne implikacije rezultatov: pri kvantitativnem raziskovanju so rezultati posplošeni, pri kvalitativnem pa specifični (Salgaro 51).

[2] Za genealogijo oddaljenega branja od starejših analognih tehnik do sodobnih računalniških metod gl. Underwood, »A Genealogy«.

Bemma Adwetewa-Badu opozarja, da digitalna orodja ne morejo opravljati kritičnega in analitičnega dela, ki ga opravljajo literarni raziskovalci, ki pa po drugi strani prav tako ne morejo opravljati funkcij digitalnih orodij (Adwetewa-Badu). Digitalne metode v literarni vedi s tega vidika ne prelamljajo s tradicionalnim bližnjim branjem in kvalitativnimi literarnovednimi metodami, ampak ponujajo dodatne podatke, ki lahko te raziskave poglobijo. Med kritikami digitalnih raziskav književnosti so na primer opozorila, da vzorci, ki jih v obsežnih zbirkah literarnih del najdevajo statistični algoritmi, niso pomenljivi (Nan 605), poleg tega digitalni humanistiki t. i. analogna humanistika očita »teoretsko podhranjenost, opiranje na zastarela pojmovanja teksta, naivni realizem ter trivialnost, nezadostnost ali celo zgrešenost računalniških rezultatov, na prvi pogled neovrgljivih« (Juvan, Šorli in Žejn 55). S strani literarne vede se pojavlja očitek, da se digitalnohumanistične raziskave ne naslanjajo na literarnovedne ugotovitve, pač pa zgolj prikazujejo podatke (Nan 605), oziroma da lahko raziskovalec med številnimi metodami računalniške analize besedil izbere tisto, ki najbolj podpre njegovo tezo (Juola, »The Rowling Case« 102). Bližnje branje nam omogoča vpogled v omejeno število literarnih del, ki jih lahko posameznik prebere, velika količina besedil v literarni zgodovini pa je ob tem spregledana kot del tega, čemur Margaret Cohen in za njo Moretti pravita »veliko neprebrano« (Moretti 8). Literarne raziskave se tako osredotočajo na besedila, ki so izpostavljena v kanonu, vendar so po obsegu omejena, njihova izbira pa je pristranska (Eve 9). A niti oddaljeno branje samo po sebi ne opravi te težave, saj ne vključuje vseh obstoječih literarnih besedil, ampak samo tista, ki so digitalizirana, kar so običajno predvsem kanonizirana besedila. Poleg tega ni konsenza o merilih za določanje reprezentativnosti korpusa. Slovenska književnost je v primerjavi z literaturami v jezikih, ki jih govori veliko več govorcev, v tem pogledu izjema, saj je digitaliziran že velik delež starejše književnosti.

Računalniška stilometrija

Sodobna stilometrija omogoča računalniško analizo literarnega stila (Eder idr. 107). S to metodo pogosto določamo avtorstvo literarnih del; med znanimi primeri so prepoznavanje avtorice knjig o Harryju Potterju J. K. Rowling, ki je po letu 2012 začela pisati pod psevdonimom Robert Galbraith (prim. Juola, »The Rowling Case«; Juola, »Rowling«) in preverjanje avtorstva del Shakespearea in Marlowa (Fox

idr.). Stilometrična analiza upošteva in primerja pogostost pojavljanja posameznih besed v vseh delih, vključenih v korpus. Individualen avtorski slog sicer pomembno določajo funkcijske besede, ki pogosto nimajo leksikalnega pomena (Kestemont). V dosedanjih raziskavah slovenske književnosti je Andrejka Žejn (Žejn, »Računalniško podprta«) z računalniško stilometrijo primerjalno analizirala literarna dela Janeza Ciglerja in Christopha Schmida, obravnavala pa je tudi pripovedno prozo od sredine 17. do sredine 19. stoletja (Žejn, *Izhodišča*). Stilometrična raziskava je podprla tudi analizo dramatike avtorice Simone Semenič (glej Zajc), z metodo »rolling stylometry« pa je bil obravnavan roman *Rokovnjači* Josipa Jurčiča in Janka Kersnika (Mandić in Zajc).

Tako kot omenjene raziskave je bila tudi raziskava, predstavljena v tem članku, izvedena v paketu Stylo v programskem jeziku R, ki ga je razvila poljsko-belgijska Skupina za računalniško stilistiko (Eder idr.) za analizo literarnih besedil na podlagi statističnih izračunov pogostosti besed ali zlogov.[3] Z meritvijo uporabe najpogostejših besed v besedilih iz korpusa so besedila razporejena glede na slogovne podobnosti; v vizualnem prikazu so tako dela, ki so si glede na uporabo besed podobna, bližje drugo drugemu. Z uporabo klastrske analize so podatki o romanih razvrščeni glede na njihove podobnosti ali razlike. Analiza upošteva in primerja frekvenco pojavljanja besed enakovredno v vseh delih korpusa. Članek prikazuje klastrsko analizo dveh korpusov slovenskih literarnih del, kjer so bili podatki o romanih razvrščeni glede na njihove podobnosti ali razlike. Analiza je temeljila na relativni frekvenci od 100 do 1000 najpogostejših besed v posameznem korpusu. Uporabljena je bila statistična metoda za merjenje razdalje Delta (prim. Burrows; Evert idr.). Rezultati raziskav so predstavljeni v obliki dendrograma in v obliki dvodimenzionalne vizualizacije bližine oziroma oddaljenosti vključenih literarnih besedil. Stilometrična analiza samodejno razporedi besedila glede na t. i. signale, ki izstopajo s tem, da določajo razvrščanje besedil glede na njihovo oddaljenost oziroma bližino z vidika stila.

Korpus 90 slovenskih romanov

V prvem delu raziskave sem oblikovala korpus 90 slovenskih romanov iz 19. in začetka 20. stoletja: romane sem shranila v besedilne datoteke (txt), urejene na način, ki je dostopen paketu Stylo in omogoča tudi

[3] Ta del raziskave, predstavljene v članku, je rezultat sodelovanja v akciji COST Distant Reading for European Literary History (CA16204) pod mentorstvom dr. Joanne Byszuk z Inštituta za poljski jezik Poljske akademije znanosti v Krakovu.

vizualizacijo v programski opremi Gephi. Besedila sem za analizo pripravila tako, da sem korpus besedil pregledala in uredila, iz njih izključila paratekst ter jih shranila v format txt in specifično urejene mape, do katerih program lahko dostopa. Besedila v korpusu sem črpala iz različnih virov: iz slovenskega dela Evropske zbirke literarnih besedil (ELTeC-slv)[4] in iz literarnih besedil, ki so prosto dostopna na slovenskem Wikiviru.[5] Iz tega osnovnega korpusa sem ustvarila druge sklope besedilnih korpusov za izvajanje specifičnih analiz. Korpus 90 romanov vsebuje dela slovenskih avtoric in avtorjev iz obdobij realizma in moderne.

Signal predstavlja merljiv podatek, ki ga lahko razberemo tudi iz vizualizacij, na primer kot jasne razdalje med določenimi skupinami besedil, kar nakazuje na njihovo slogovno raznolikost (prim. Evans). Kadar se literarna dela razlikujejo predvsem glede na njihovo avtorstvo, gre za avtorski signal, kadar pa se razlikujejo na primer glede na žanr ali spol, gre za žanrski oziroma spolni signal (prim. Jockers). Poleg tega se lahko pokaže izrazit vzorec kronološkega razvoja literarnega sloga posameznega avtorja (Evans). Koncept po drugi strani določen signal pojasnjuje, na primer z ugotovitvijo, da besedila, med katerimi so očitne razlike, spadajo v različne literarne žanre. V digitalni humanistiki tvorimo argumente na osnovi konceptov, medtem ko lahko računalniki izmerijo le signale (Heuser in Le-Khac).

S stilometrično analizo 90 slovenskih romanov odgovarjam na naslednji raziskovalni vprašanji: ali najdemo signale, ki niso povezani zgolj z avtorstvom, na primer spolne signale? in ali kateri od avtorjev in avtoric iz korpusa odstopa glede na svoj slog pisanja?

Stilometrična analiza 90 slovenskih romanov: rezultati in razprava

Slika 1 prikazuje rezultate raziskave: posamezna točka označuje določeno literarno besedilo, ki vsebuje pripis avtorja in naslova, črte pa označujejo povezave med besedili glede na splošno slogovno podobnost. S temnejšim barvnim odtenkom so označena dela avtoric, s svetlejšim pa dela avtorjev.

[4] Korpus vsebuje 100 daljših proznih del v slovenskem jeziku, ki so izšla med letoma 1836 in 1921.

[5] Wikivir kot projekt Mirana Hladnika je »pomemben za razumevanje branja v digitalnem okolju, ker si zadaja cilj ustvarjanja edinstvene spletne knjižnice slovenske književnosti, uporabnicam in uporabnikom pa omogoča, da se srečajo z besedili, ki niso znana ali kanonizirana« (Ilin 195).

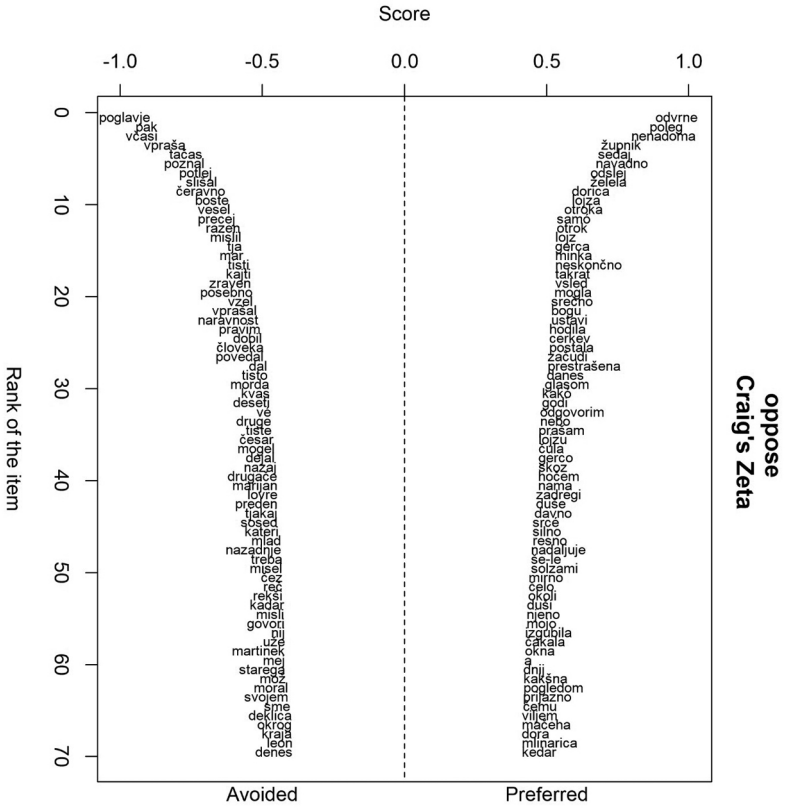
Slika 1: Stilometrična analiza 90 slovenskih romanov.

V zgornjem delu prikaza vidimo skupino avtorjev slovenskega realizma, od katerih sta si po stilu na primer blizu Josip Jurčič in Janko Kersnik. Njun roman *Rokovnjači*, ki ga je začel pisati Jurčič in dokončal Kersnik, je razvrščen med Kersnikova dela.[6] Deloma se kaže signal spola, ki pa ni izrazit, saj se skupaj razvrščajo le nekatere avtorice, med njimi pa so tudi avtorji, tako da se na primer delo Frana Levstika *Gadje gnezdo* razvrsti v bližini romanov Lee Fatur in Josipine Turnograjske. Med drugim so dela Zofke Kveder slogovno blizu delu Marice Nadlišek Bartol, dela Miroslava Malovrha so blizu delom Frana Zbašnika, dela Frana Saleškega Finžgarja pa romanu Ivana Preglja *Tlačani*. Kot je razvidno iz vizualizacije, spol ni ključen signal za razvrščanje teh del v skupine, temveč je najizrazitejši signal njihovo avtorstvo: na primer besedila Ivana Cankarja ali Lee Fatur so se razvrstila v skupini v spodnjem delu prikaza. Skrajno levo so se v ločeno skupino razvrstila dela Pavline Pajk, iz česar lahko razberemo, da so izrazito stilno drugačna od ostalih besedil v tem korpusu. Sklepamo, da je razlogov za to več: Pavlina Pajk se je slovenskega jezika, v katerem je pozneje pisala romane, naučila šele pri 16 letih, njen prvi jezik pa je bila italijanščina; otroštvo je preživela v Milanu in Trstu, po smrti staršev je živela pri stricu v Solkanu in nato

[6] Za natančnejšo stilometrično analizo Jurčič-Kersnikovega romana *Rokovnjači* gl. Mandić in Zajc 2020.

pri bratu, hodila je v uršulinsko šolo v Gorici; zaradi specifične jezikovne situacije v njenem otroštvu in mladosti sklepamo, da je bila tudi njena raba slovenščine specifična; poleg tega je pisala žanrsko drugačna besedila kot drugi avtorji in avtorice v korpusu, in sicer sentimentalne ljubezenske romane. Da bi pridobila še dodatne informacije o rabi jezika pri Pavlini Pajk, sem izvedla dodatno raziskavo, pri kateri sem uporabila funkcijo Oppose iz paketa Stylo, ki omogoča, da ugotavljamo razlike med najpogosteje uporabljenimi besedami v posamezni skupini literarnih besedil oziroma vidimo, katere besede se v posamezni skupini besedil pojavljajo bistveno pogosteje kot v drugi skupini. Ustvarila sem korpus besedil Pavline Pajk na eni strani (t. i. primarni sklop) in enako količino drugih literarnih besedil različnih njenih sodobnikov na drugi strani (t. i. sekundarni sklop), da bi izvedla primerjalno analizo besedišča del Pavline Pajk v primerjavi z drugimi reprezentativnimi romani tega obdobja slovenskega realizma.[7] Ugotavljala sem, katere besede so najznačilnejše za dela Pavline Pajk, tj. primarnega sklopa, kar pomeni, da so v besedilih sekundarnega sklopa redke, pa tudi, katere besede so najznačilnejše za sekundarni sklop in torej redke v besedilih primarnega sklopa. Naslednja slika prikazuje seznam najpogostejših besed, ki se jim primarni sklop izogiba (na levi strani), in seznam besed, ki so v primarnem sklopu pogostejše v primerjavi s sekundarnim sklopom (na desni strani).

[7] Janez Cigler, *Sreča v nesreči*; Fran Detela, *Malo življenje*; Fran Govekar, *V krvi*; Josip Jurčič, *Deseti brat*; Janko Kersnik, *Agitator*; Anton Koder, *Luteranci*; Fran Levstik, *Gadje gnezdo*; Josip Stritar, *Sodnikovi*.



Slika 2: Besedišče del Pavline Pajk v primerjavi z drugimi besedili njenega časa.

Kot lahko vidimo na Sliki 2, med najbolj priljubljenimi besedami v korpusu del Pavline Pajk najdemo besede, povezane z ženskimi liki – omembe ženskih literarnih oseb (»mačeha«, »Dora«, »njeno« ipd.) ter glagolske oblike in pridevnike, ki se nanašajo na ženske like (»prestrašena«, »čakala«, »hodila«, »postala«, »želela« ipd.) –, oziroma besede, ki so značilne za žanr sentimentalnega romana (»srce«, »duše«, »solzami« ipd.), omembe župnika, cerkve, otrok ipd. Avtoričina specifična raba jezika sicer zahteva obsežnejše raziskave z bližnjim branjem, ki presegajo namen te študije.

Raziskava korpusa besedil Ivana Cankarja

Čeprav se računalniške stilometrične raziskave najpogosteje nanašajo na dela z različnim (pogosto tudi neznanim) avtorstvom, se v drugem delu raziskave osredotočam na žanrsko raznolik opus posameznega avtorja, in sicer Ivana Cankarja. Tudi ta eksperiment sem izvedla v paketu Stylo v programskem jeziku R.

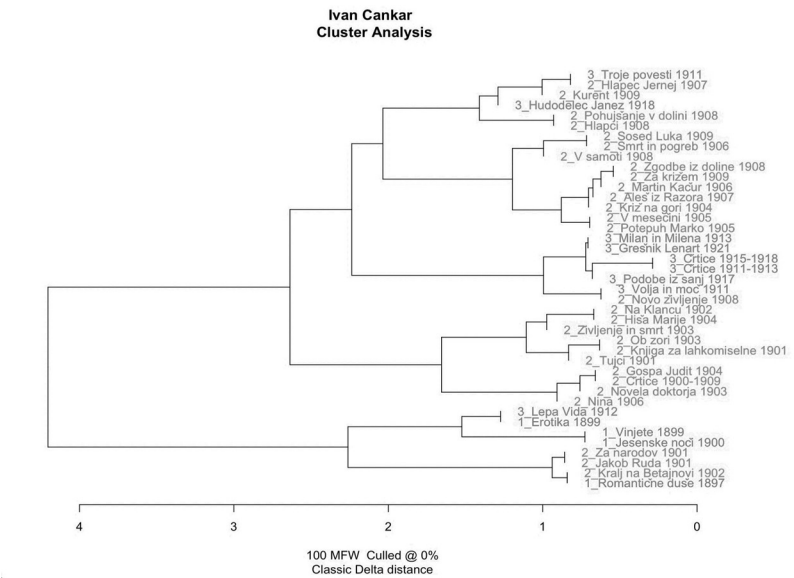
Cankar velja za enega glavnih predstavnikov slovenske literarne moderne. Njegov opus šteje približno 30 samostojnih knjižnih publikacij, številna dela pa je objavljal tudi v tedanji periodiki. Teme njegovih del so socialne, moralne in nacionalne, pisal je poezijo, dramatiko in prozo, uveljavil je žanr simbolistične skice oziroma črtice.

Dekadenca in simbolizem sta se v slovenski literaturi pojavila po letu 1897, ko je Cankar ustvaril prve pesmi in črtice s tovrstnimi vplivi. Do sredine devetdesetih let 19. stoletja je bilo v ospredju njegove ustvarjalnosti pesništvo: »/s/ pripovedno prozo se do leta 1896 ni intenzivneje ukvarjal in jo je pisal bolj ali manj iz neliterarnih nagibov ter po tradicionalnem realistično-naturalističnem vzorcu, vendar ga je že naslednje leto z bolj ali manj deepiziranimi vinjetami […] začel tudi razkrajati.« (Čeh 45) Te črtice vnašajo prelom v Cankarjevo ustvarjalnost in so zametek poznejše zbirke *Vinjete* (Kocijan 36). Cankarjeve romane po drugi strani zaznamujeta zvrstni sinkretizem, tj. neklasična epska struktura, in žanrski sinkretizem, tj. preplet različnih žanrov, ki so bili v slovenskem prostoru pogosto inovativni (glej Zupan Sosič). Avtorja je močno zaznamovalo bivanje na Dunaju, kjer je s presledki preživel deset let svojega življenja (1898–1909), ob tem pa je »sprejel periferijo Dunaja kot svoj dom, kot samotarski, kontemplativni avtor se ni poskusil uveljaviti v dunajski kulturi, čeprav je prav prek nje sprejemal različne filozofske in literarne tokove« (Jensterle-Doležal 68).

Kot je pokazal prvi del raziskave, se izbor iz Cankarjevega opusa (prozna besedila *Martin Kačur*, *Tujci*, *Hiša Marije Pomočnice* in *Na Klancu*) razlikuje od opusov njegovih sodobnikov, saj se v vizualizaciji stilometrične analize razporedi v svoj sklop (gl. skrajno desno spodaj na Sliki 1), ki je še najbližje Finžgarjevemu delu *Moja duša vasuje* in Pregljevemu delu *Tlačani*. Drugi del raziskave se posveča izključno Cankarjevemu opusu in preverja, ali je stilno homogen ali heterogen. Za namene te stilometrične analize sem oblikovala korpus 47 žanrsko raznolikih Cankarjevih del na podlagi besedil, zbranih v Wikiviru, ki vsebuje mnoga Cankarjeva besedila. Med vključenimi proznimi deli najdemo tako krajše črtice kot romane in druga daljša besedila. Da bi zadostila kriteriju primerljivosti vključenih besedil glede na dolžino,

sem v korpus poleg proznih povesti in romanov vključila zbirke črtic in ne posameznih črtic, enako pa velja tudi za poezijo. Prvo izhodišče raziskave je raznolikost Cankarjevega opusa, ki vsebuje vse od proznih del prek poezije do dram. Poleg tega raziskava upošteva dejstvo, da je literarna zgodovina Cankarjev opus delila na različna ustvarjalna obdobja, in sicer sem izbrala tisto periodizacijo njegovega pisanja, ki je del šolskih kurikulov v slovenskem prostoru, saj gre za uveljavljeno periodizacijo (prim. Bernik, *Obzorja slovenske*). Ta Cankarjevo delo deli na zgodnje obdobje med letoma 1897 in 1900, srednje obdobje med letoma 1901 in 1909 ter pozno obdobje, ki traja od leta 1910 do avtorjeve smrti.[8] Cilj drugega dela raziskave je odgovoriti na naslednji raziskovalni vprašanji: ali lahko Cankarjeva dela uvrstimo v različna ustvarjalna obdobja? In ali lahko Cankarjeva dela uvrstimo v različne literarne vrste oziroma zvrsti?

Korpus del Ivana Cankarja: rezultati in razprava



Slika 3: Klastrska analiza del Ivana Cankarja.

[8] Irena Avsenik Nabergoj za zaključek Cankarjevega prvega ustvarjalnega obdobja glede na kraj, kjer je deloval, določi leto prej, tj. 1899 (Avsenik Nabergoj 96).

S klastrsko analizo sem identificirala skupine literarnih besedil v korpusu, ki so si stilno podobna (Eder 51). Slika 3 prikazuje rezultate na podlagi analize 100 najpogostejših besed; rezultati tovrstne analize so se ujemali tudi, kadar sem uporabila izhodišče 500 najpogosteje uporabljenih besed v posameznih besedilih korpusa. Kot je razvidno iz vizualizacije, se Cankarjev opus deli na dva dela: dela, ki so na sliki skrajno spodaj, spadajo v prvo obdobje njegovega ustvarjanja in se najbolj razlikujejo od ostalih, višje razvrščenih Cankarjevih besedil. Tudi ta se sicer nadalje delijo v dve skupini, ki se kronološko ujemata. V prvo skupino, kjer najdemo predvsem Cankarjeva zgodnja dela, se ob pesniško zbirko *Erotika* (1899) izjemoma razvrsti tudi avtorjeva poznejša drama *Lepa Vida* (1912), kar si razlagam kot avtorjevo vrnitev k poetičnemu slogu, ta drama pa je sicer nastajala dlje časa. Bernik (*Ivan Cankar* 375) opozarja, da se je Cankar z *Lepo Vido* vrnil k poeziji. Poleg tega navaja, da se ta drama vsebinsko povezuje z avtorjevo prvo »dramatično sliko« *Romantične duše*, kar potrjuje tudi stilometrična analiza, ki obe deli razvrsti v bližino (Bernik, *Ivan Cankar* 375). Cankar je v začetnem obdobju poleg poezije ustvarjal tudi dramatiko. Njegovi zgornji drami *Jakob Ruda* (1901) in *Romantične duše* (1897) se skupaj s poezijo iz prvega obdobja razvrstita v bližini drame *Kralj na Betajnovi* (1902), ki očitno stilno spada še v zgodnji del opusa, kar pomeni, da nad žanrskim signalom prevladuje časovni oziroma da se dela stilno razlikujejo glede na čas nastanka. To potrjuje tudi dejstvo, da se tudi drami *Pohujšanje v dolini šentflorjanski* in *Hlapci* razporedita sicer skupaj, kar pomeni, da sta si stilno podobni, vendar se umestita med pripovedna besedila. Ko govorimo o dramatiki, v Cankarjevem opusu torej prevlada časovni signal, saj se dela razvrstijo predvsem glede na obdobje njihovega nastanka, avtorjev dramski opus pa je torej stilno heterogen. Na sredino in v skrajno zgornji del prikaza se razporedijo dela, ki so nastala v srednjem obdobju, kar pomeni, da je to obdobje stilno heterogeno. Izrazito vejo, ki se stilno razlikuje od drugih, predstavljajo črtice, ki so nastale v poznem obdobju, tj. zbirki črtic iz obdobij 1911–1913 in 1915–1918 ter *Podobe iz sanj* (1917), medtem ko se črtice iz obdobja 1900–1909 od drugih črtic razlikujejo in stilno razporedijo med ostala dela iz drugega obdobja Cankarjevega ustvarjanja.

Sklep

Raziskava, predstavljena v prispevku, je bila izvedena v paketu Stylo v programskem jeziku R in se deli na dva eksperimenta: na stilometrično analizo korpusa 90 slovenskih romanov in stilometrično analizo opusa del Ivana Cankarja. Analiza je pokazala, da je ključen signal, po katerem se ta literarna dela razlikujejo, njihovo avtorstvo. Posebej izstopata Pavlina Pajk in Ivan Cankar, saj so njuna besedila v vizualizaciji razvrščena izrazito ločeno od drugih del, vključenih v korpus. Poleg tega se deloma kaže signal spola, ki pa ni izrazit, saj se skupaj razvrščajo le nekatere avtorice, med njimi pa najdemo tudi avtorje: na primer Levstikovo delo *Gadje gnezdo* se razvrsti v bližini romanov Lee Fatur in Josipine Turnograjske. Stilometrična analiza Cankarjevega opusa je pokazala na razpoznavna časovna obdobja njegove ustvarjalnosti. Stilno homogeno in izrazito drugačno od drugih delov Cankarjevega opusa je njegovo prvo ustvarjalno obdobje, v katero se je sicer razvrstila tudi drama iz leta 1902 *Kralj na Betajnovi*. Drugo obdobje je stilno bolj heterogeno, v tretjem obdobju pa kot stilno specifičen del Cankarjevega opusa izstopajo njegove poznejše črtice.

LITERATURA

Adwetewa-Badu, Bemma. »Poetry from Afar: Distant Reading, Global Poetics, and the Digital Humanities«. *Modernism/Modernity Print*+, let. 5, št. 1, 2020, https://modernismmodernity.org/forums/posts/adwetewa-badu-poetry. Dostop 12. 4. 2024.

Avsenik Nabergoj, Irena. »Ivan Cankar med domovino in tujino«. *Dve domovini*, št. 20, 2004, str. 95–111.

Bernik, France. *Obzorja slovenske književnosti*. Slovenska matica, 1999.

Bernik, France. *Ivan Cankar: Monografija*. Litera, 2006.

Burrows, John. »'Delta' – A Measure of Stylistic Difference and a Guide to Likely Authorship«. *Literary and Linguistic Computing*, let. 17, št. 3, 2002, str. 267–287.

Da, Nan, Z. »The Computational Case against Computational Literary Studies«. *Critical Inquiry*, let. 45, št. 3, 2019, str. 601–639.

Čeh, Jožica. »Cankarjeva metafora v vinjetnem obdobju«. *Slavistična revija*, let. 45, št. 1–2, 1997, str. 45–58.

Eder, Maciej. »Visualization in Stylometry: Cluster Analysis Using Networks«. *Digital Scholarship in the Humanities*, let. 32, št. 1, 2017, str. 50–64.

Eder, Maciej, Jan Rybicki in Mike Kestemont. »Stylometry with R: A Package for Computational Text Analysis«. *The R Journal*, let. 8, št. 1, 2016, str. 107–121.

Evans, Mel. »Style and Chronology: A Stylometric Investigation of Aphra Behn's Dramatic Style and the Dating of *The Young King*«. *Language and Literature*, let. 27, št. 2, 2018, str. 1–49.

Eve, Martin Paul. *The Digital Humanities and Literary Studies*. Oxford University Press, 2022.

Evert, Stefan, idr. »Understanding and Explaining Delta Measures for Authorship Attribution«. *Digital Scholarship in the Humanities*, let. 32, št. 2, 2017, str. 4–16.

Fox, Neal P., Omran Ehmoda in Eugene Charniak. »Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate«. *Proceedings of the Georgetown University on Language and Linguistics*, Georgetown University, 2012, https://cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf. Dostop 12. 4. 2024.

Heuser, Ryan, in Long Le-Khac. »A Quantitative Literary History of 2,958 Nineteenth-century British Novels«. *Stanford Literary Lab*, 2012, https://litlab.stanford.edu/assets/pdf/LiteraryLabPamphlet4.pdf. Dostop 12. 4. 2024.

Ilin, Darko. »Wikisource platforma kao biblioteka slovenačke književnosti i kulture«. *Slovenika: časopis za kulturu, nauku i obrazovanje*, št. 9, 2023, str. 195–211.

Jensterle-Doležal, Alenka. »Fenomen mesta v opusu Ivana Cankarja in Zofke Kveder«. *Jezik in slovstvo*, let. 55, št. 5–6, 2010, str. 57–70.

Jockers, Patrick. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, 2013.

Juola, Patrick. »Rowling and 'Galbraith': An Authorial Analysis«. *Language Log*, 16. 7. 2013, http://languagelog.ldc.upenn.edu/nll/?p=5315. Dostop 12. 4. 2024.

Juola, Patrick. »The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions«. *Digital Scholarship in the Humanities*, let. 30, št. 1, 2015, str. 100–113.

Juvan, Marko, Mojca Šorli in Andrejka Žejn. »Interpretiranje literature v zmanjšanem merilu: 'oddaljeno branje' korpusa 'dolgega leta 1968'«. *Jezik in slovstvo*, let. 66, št. 4, 2021, str. 55–76.

Kestemont, Mike. »Function Words in Authorship Attribution: From Black Magic to Theory?«. *Proceedings of the 3rd Workshop on Computational Linguistics for Literature*, ur. Anna Feldman, Anna Kazantseva in Stan Szpakowicz, The Association for Computational Linguistics, 2014, str. 59–66.

Kocijan, Gregor. *Kratka pripovedna proza v obdobju moderne*. Znanstveni inštitut Filozofske fakultete, 1996.

Mandić, Lucija, in Ivana Zajc. »Stilometrična analiza avtorskega sloga Jurčič-Kersnikovega romana *Rokovnjači*«. *Slovansko jezikovno in literarno povezovanje ter zgodovinski kontekst*, ur. Ina Poteko, Študentska organizacija Filozofske fakultete, 2020, str. 7–13.

Moretti, Franco. »Domneve o svetovni literaturi«. *Grafi, zemljevidi, drevesa in drugi spisi o svetovni literaturi*, ur. in prev. Jernej Habjan, Studia humanitatis, 2011, str. 5–25.

Salgaro, Massimo. »The Digital Humanities as a Toolkit for Literary Theory: Three Case Studies of the Pperationalization of the Concepts of 'Late Style,' 'Authorship Attribution,' and 'Literary Movement'«. *Iperstoria*, let. 12, št. 2, 2018, str. 50–60.

Underwood, Ted. »Distant Reading and Recent Intellectual History«. *Debates in the Digital Humanities* 2016, ur. Matthew K. Gold in Lauren F. Klein, University of Minnesota Press, https://dhdebates.gc.cuny.edu/read/untitled/section/3b96956c-aab2-4037-9894-dc4ff9aa1ec5. Dostop: 12. 4. 2024.

Underwood, Ted. »A Genealogy of Distant Reading«. *Digital Humanities Quarterly*, let. 11, št. 2, 2017, http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html. Dostop 12. 4. 2024.

Zajc, Ivana. »Elementi monodrame in avtobiografskosti v besedilih Simone Semenič«. *Amfiteater*, let. 7, št. 2, 2019, str. 80–98.

Zupan Sosič, Alojzija. »Deset romanov Ivana Cankarja in sodobna definicija romana«. *Slavia Centralis*, let. 13, št. 1, 2020, str. 136–152.

Žejn, Andrejka. »Računalniško podprta stilometričcna analiza pripovedne literature Janeza Ciglerja in Christopha Schmida v slovenščini«. *Fluminensia*, let. 32, št. 2, 2020, str. 137–158.

Žejn, Andrejka. *Izhodišča slovenske pripovedne proze. Dvestoletna tradicija slovenske pripovedne proze: od sredine 17. do sredine 19. stoletja*. ZRC SAZU, Inštitut za slovensko literaturo in literarne vede, 2021, https://ispp.zrc-sazu.si/. Dostop 12. 4. 2024.

Analyzing 90 Slovenian Novels and the Oeuvre of Ivan Cankar Using Computational Stylometry

Keywords: Slovenian novel / Cankar, Ivan / distant reading / computational stylometry / digital humanities

The article presents a distant reading of Slovenian literature from the periods of realism and early modernism (*moderna* in Slovenian). The first part of the article displays the results of a computational stylometric analysis of a corpus of 90 Slovenian novels from the two periods, conducted in the programming environment R with the Stylometry with R package. The results show that the primary signal differentiating the included novels is the authorial one, with Pavlina Pajk and Ivan Cankar standing out the most. Since Pajk is the most distinct author in the corpus, the article also includes a stylometric investigation into the most typical vocabulary of her works. On this basis, the article concludes that her writing differs from others due to the genre of sentimental romance novels which characterizes her work. In the second part of the article, Cankar's literary oeuvre is analyzed using the same methodology, and the corpus includes not only his novels but also other works from different periods of his writing career. The results of the analysis reveal the development of Cankar's literary style in terms of genre and periodization.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.163.6.09-31Cankar I.:004
DOI: https://doi.org/10.3986/pkn.v47.i2.06