Primerjava Josipa Jurčiča in Ivana Cankarja z računalniškimi metodami za zaznavanje semantičnih premikov

Andrejka Žejn, Marko Pranjić, Senja Pollak

Inštitut za slovensko literaturo in literarne vede ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenija
https://orcid.org/0000-0001-8028-0193
andrejka.zejn@zrc-sazu.si

Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana, Slovenija
https://orcid.org/0000-0002-8645-9714
marko.pranjic@ijs.si

Institut »Jožef Stefan«, Jamova cesta 39, 1000 Ljubljana, Slovenija
https://orcid.org/0000-0002-4380-0863
senja.pollak@ijs.si

V prispevku uvodoma predstavimo heterogenost in interdisciplinarnost digitalne humanistike kot dva osrednja in medsebojno povezana koncepta. V osrednjem delu vpeljemo uporabo metode zaznavanja semantičnih premikov z uporabo kontekstualnih besednih vložitev pri analizi literarnih del. Potencial metode prikažemo na primerjalni analizi pripovednih opusov Josipa Jurčiča in Ivana Cankarja, kanoničnih slovenskih avtorjev, ki sta ustvarjala v različnih obdobjih, in sicer z avtomatskim prepoznavanjem besed, katerih pomeni se med avtorjema najbolj razlikujejo. Nadaljnja interpretacija temelji na kvalitativni analizi avtomatsko pridobljenih rezultatov in na uvrščanju besed, prepoznanih kot relevantnih za razlike med Jurčičevim in Cankarjevim slogom, v pomenska polja. Pokazali smo, da je pristop, ki temelji na kontekstualnih besednih vložitvah, mogoče uporabiti za analize literature z zadovoljivimi rezultati. S tem z vidika literarne zgodovine ponujamo nov vpogled v Cankarjevo in Jurčičevo pripovedništvo, saj pokažemo, da razlika med Jurčičevim (romantičnim) realizmom in Cankarjevo moderno sloni tudi na semantiki besed, povezani z gibanjem ter družbenimi in psihološkimi dejanji in procesi.

Ključne besede: slovenska književnost / Jurčič, Josip / Cankar, Ivan / obdelava naravnega jezika / semantična analiza / digitalna literarna veda / digitalna humanistika

Uvod

Digitalna humanistika, ki se kot široko polje raziskovalnih praks razvija od šestdesetih let 20. stoletja naprej, ima zaradi svoje heterogenosti množico definicij, kar hkrati pomeni, da obče veljavne definicije področja ni mogoče podati (Alvaro 50; Kuhn in Callahan 291; Rodríguez Ortega 2; Terras idr.).[1] Polje digitalne humanistike je namreč nenehen proces (Vanhoutte) oziroma amorfno, fluidno in fleksibilno področje raziskav.[2] Zato tudi vprašanje, kaj je digitalna humanistika, ne anticipira odgovora, ampak stalno raziskovanje in redefiniranje (McCarty 1233; Earhart 1, 117–119). Kolektivno ime *digitalna humanistika* (Svensson, »Humanities« 42) in njena posplošena definicija ne pokrijeta niti različnih digitalnih pristopov niti tega, kako digitalne prakse spreminjajo posamezna polja humanistike, zato je po Amy Earhart smiselna njena segmentacija (Earhart 119). Eden od segmentov digitalne humanistike je digitalna literarna veda (1–10),[3] ki je obenem področje v okviru literarnih ved (Zajc in Purg) in njihova kontinuiteta (Ganascia 3). Ker je bilo vse od začetkov digitalne humanistike temeljni predmet preučevanja besedilo, je bila digitalna literarna veda dolgo časa eno njenih najplodnejših področij (Murray § 10; Ganascia 1). V digitalni literarni vedi je mogoče vsaj v grobem ločevati med praksami, posvečenimi ohranjanju, dokumentaciji, prezentaciji in diseminaciji literarnih besedil (najbolj široko lahko te prakse zajamemo s pojmom digitalne izdaje), in kvantitativnimi analizami

[1] Članek je nastal v okviru raziskovalnega programa »Tehnologije znanja« (P2-0103), raziskovalnega programa »Literarnozgodovinske, literarnoteoretične in metodološke raziskave« (P6-0024) in raziskovalnega projekta »Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami« (J6-2581), ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

[2] Nedefiniranost kot posledica obsežnosti digitalne humanistike je vključena tudi v diskurz o digitalni humanistiki kot »krovnem pojmu« (Svensson, »Beyond«; Presner 195–196; Alvarado) in njegovih ironičnih izpeljankah, kakršna je na primer ugotovitev, da »šotor ni dovolj velik« za pokritje celotnega spektra digitalne humanistike (Terras idr.). Ironiziranje metafore je zaznavno tudi pri Stephenu Ramsayju, ki ekstenzivno naštevanje, kaj vse je mogoče uvrstiti v digitalno humanistiko, sklene z besedno zvezo »prostorno platno« (»capacious canvas«; Ramsay 240).

[3] Angleški izraz za digitalno literarno vedo, *digital literary studies*, se je v literarni vedi »razplamtel« (Hoover idr.) v prvi polovici drugega desetletja 21. stoletja. Glede na Googlov n-gram za obdobje do leta 2019 se je ta besedna zveza na internetu – ki je ne nazadnje pogost ali celo najpogostejši medij digitalnohumanističnih objav – začela pojavljati sredi prvega desetletja 21. stoletja; njena pojavnost je izrazito naraščala do leta 2017, ko se je začel izrazit upad.

literature z najrazličnejšimi metodami t. i. oddaljenega branja ter interpretativne vizualizacije raznovrstnih podatkov iz literature, pridobljenih z računalniškimi metodami (sem sodita na primer analiza omrežij in literarna geografija). Področji se ne izključujeta med seboj in se medsebojno prepletata. Tako kot digitalno humanistiko tudi digitalno literarno vedo zaznamujeta odprtost in razvoj, povezan predvsem z razvojem računalništva in novih metodologij.

Poleg heterogenosti in z njo povezane predefiniranosti in/ali nedefiniranosti ima digitalna humanistika močan pečat kontroverznosti. Ta se zdi najbolj izrazit v digitalni literarni vedi, saj je k izzivalnosti v največji meri prispeval teoretik oddaljenega branja, ki je sicer metaforična in provokativna (Juvan idr. 57; Jannidis in Lauer 30) izraza *natančno branje* in *oddaljeno branje* predstavil kot opozicijo med branjem in nebranjem literarnih besedil (Moretti 11) in s tem sprožil ostre razprave med zagovorniki kvantitativnega oddaljenega branja in zagovorniki tradicionalnega natančnega branja, ki so se razplamtele v prvem desetletju 21. stoletja in so segle tudi zunaj ozkega akademskega prostora. Kljub mnenju, da so te debate danes postane in neplodne (Murray § 3), se digitalna literarna veda še vedno spopada z očitki, da so njene raziskave neuporabne (ne povedo ničesar, česar še ne vemo), trivialne (omejene so na štetje besed), celo neoliberalne (Eve 1). Ahistorični pogled na digitalno literarno vedo bi morda navedenim očitkom celo pritrdil, s historičnega vidika pa so te prve raziskave, kot ugotavlja Simon Mahony (Mahony 372), postavljale temelje discipline in sprožile nadaljnje preusmerjanje pozornosti od tehnologije kot nekakšne služabnice humanistike k interdisciplinarnosti kot nujni specifiki, na katero se opirajo opredelitve digitalne humanistike in njenih podpodročij (McCarthy 1225; Presner 195–196).

Jean-Gabriel Ganascia pri presojanju interdisciplinarnosti digitalne humanistike ugotavlja, da digitalna humanistika pripada tako t. i. kulturnim znanostim kakor t. i. naravnim znanostim, saj s stališča predmeta obravnave sodi med prve, po metodološki plati in po tem, da uporablja obsežne podatkovne zbirke, ki so avtomatsko procesirane, pa je bliže drugim (Ganascia 2–3).[4] V digitalni literarni vedi ni antagonizma med logiko znanosti kulture in razvijanjem orodij, ki pomagajo interpretirati ogromne količine podatkov z upoštevanjem obstoječih teorij.[5] Victoria Kuhn in Vicki Callahan pa v zvezi z

[4] Pri tem izhaja iz postavk nemških filozofov Heinricha Rickerta in njegovega učenca Ernsta Cassirerja, ki sta v prvi polovici 20. stoletja utemeljila razlikovanje med kulturnimi in naravnimi znanostmi.

[5] Za nasprotne poglede gl. Ganascia 4.

interdisciplinarnostjo digitalne humanistike ugotavljata, da doba digitalne kulture premešča meje med znanstvenimi disciplinami, ki so bile zarisane med vzponom kulture tiska (Kuhn in Callahan 292). V primerjavi s horizontalno interdisciplinarnostjo dveh disciplin, kakršni sta na primer zgodovinopisje in literarna veda, kjer dodajanje ali nadomeščanje elementov poteka brez bistvene spremembe v strukturi ali logiki udeleženih disciplin, je radikalnost digitalne humanistike v njenem potencialu, da vertikalno razširi interdisciplinarnost. Vertikalna razširitev postavlja najrazličnejše discipline pred nove izzive ter spreminja njihovo strukturo in logiko. Računalniške metode niso zgolj aplicirane na literaturo, ampak spremenijo sam način raziskovanja literature, s tem ko zastavljajo nova vprašanja. Digitalna humanistika ponuja predvsem miselno obzorje, ki omogoča nov pogled na analiziranje in interpretiranje kulturne produkcije (Bode 1, Piper 2). Inherentno prekrivanje računalništva in humanistike, v njenem okviru pa tudi literarne vede, in mrežne povezave med njima vzpostavljajo prostor združevanja vednosti, ki presega enostavno ločevanje in v katerem običajno sodelujejo strokovnjaki tako iz naravoslovja kot iz humanistike. Potreba po harmonizaciji vodi v razvijanje nove hermenevtike, v kateri potenciale kvantitativnih in matematičnih formulacij kombiniramo s humanističnimi (literarnovednimi) vidiki. V skladu z vsebinsko odprto naravo digitalne humanistike na novo vzpostavljeni prostor epistemološke produkcije še vedno potrebuje razvoj, uskladitev in utrditev (Rodríguez Ortega 4), kar je ne nazadnje imanentna lastnost vsake metode, o čemer priča nenehen razvoj metodoloških pristopov tako znotraj kot zunaj digitalne literarne vede, in sicer največkrat kot odziv na pomanjkljivosti že uveljavljenih metod. S tem povezan je tudi metodološki pluralizem, ki je še posebej izrazit v digitalni literarni vedi, kjer gre že v izhodišču za kombiniranje oddaljenega in natančnega branja oziroma kvantitativnega in kvalitativnega pristopa.

Čeprav glavni argumenti zagovornikov oddaljenega branja poudarjajo možnost sočasnega pogleda na obsežne količine literarnih del in vključitev literarnih del, ki jih tradicionalna literarna veda z metodami natančnega interpretativnega branja ni vključila v kanon in s tem tudi ne v svoje raziskave,[6] se digitalna literarna veda pogosto ukvarja s

[6] Čeprav ta argument močno odmeva, odkar ga je zapisal Moretti (Moretti 8), velja poudariti, da je že Burrows več kot desetletje prej zapisal, da literarna veda ne more več zanemarjati potenciala »elektronskega medija« in nadaljevati z literarno zgodovino, ki temelji na majhnem vzorcu in ne upošteva »tretjine, dveh petin, polovice« gradiva (Burrows 1).

kanoniziranimi avtorji,[7] ne pa nemara z obsežnimi korpusi, kakršnih posameznik ne bi uspel niti prebrati. Vzrok za osredotočenost na kanon bi lahko iskali v tem, da so ti avtorji in njihova dela najbolj raziskani, tako da je mogoče iz njih lažje izpeljati nova raziskovalna vprašanja, na katera je mogoče odgovoriti v okviru digitalne literarne vede. Ne glede na obseg korpusa je, kot ugotavlja Thomas Rommel, s kvantitativnimi metodami mogoče med drugim določati značilnosti stila, ki jih pri natančnem branju besedila ni mogoče prepoznati ali izluščiti, toda vseeno vplivajo na splošni vtis besedila (Rommel 90). Računalnik lahko učinkovito prepozna distribucijske vzorce, ki pomagajo razumeti učinek besedila. Za pomenljive rezultate je poleg literarnega koncepta, iz katerega izhaja računalniška kvantitativna analiza, in samega računalniškega orodja nujna ustrezna literarnovedna interpretacija.

Prepoznavanje, katere jezikovne značilnosti imajo v besedilu funkcijo markerjev stila in katere so stilistično nevtralne, temelji na njihovi pojavnosti v različnih, a povezanih kontekstih (Enkvist 34–35). V različna, a povezana konteksta spadata tudi pripovedna opusa kanoniziranih slovenskih literatov Josipa Jurčiča (1844–1881) in Ivana Cankarja (1876–1918), ki sta v literarnozgodovinskem središču prispevka. Jurčič in Cankar sta pisala žanrsko raznovrstno pripovedno prozo, Jurčič v obdobju realizma oziroma romantičnega realizma, Cankar v obdobju moderne, predvsem pa sta oba postavila mejnike v razvoju slovenskega romanopisja in pripovedne proze t. i. dolgega 19. stoletja: Jurčič kot avtor prvega slovenskega romana in s tem začetnik klasičnega slovenskega romana (Kos, »Cankar«), Cankar pa kot avtor, po zaslugi katerega je slovenski roman v obdobju moderne postal pravi, umetniški roman (Pirjevec 73). V nadaljevanju predstavljena računalniško podprta kvantitativna raziskava, ki je rezultat interdisciplinarnega sodelovanja med literarno vedo in informatiko, je bila usmerjena v razkrivanje razlik v diskurzu teh dveh avtorjev na podlagi pomenske primerjave z uporabo kontekstualnih besednih vložitev. Ta računalniška metoda za zaznavanje semantičnih premikov je usmerjena na pomen besede, ki izhaja iz njene tipične besedne okolice.

Za računalniško analizo besedil je ključnega pomena način predstavitve besed in dokumentov. Velik preskok na področju obdelave naravnega jezika so omogočile globoke nevronske mreže za učenje gostih vektorskih vložitev besed in dostopnost velikih prednaučenih jezikovnih modelov (BERT: Bidirectional Encoder Representations

[7] Značilno je, da je v stilometričnih in drugih računalniško podprtih kvantitativnih raziskavah med najpogosteje obravnavanimi literati William Shakespeare.

from Transformers; Devlin idr.). Medtem ko so bile prve goste vložitve statične oziroma globalne, tako da je bil eni besedi pripisan en vektor glede na vse kontekste, v kateri se dana beseda pojavlja, so novejše predstavitve, ki jih uporabljamo tudi v tem prispevku, kontekstualne. Pri teh ima vsaka raba besede (oziroma žetona, ki lahko predstavlja tudi del besede) svojo predstavitev, ki se spreminja glede na lokalni kontekst. Tovrstne vložitve bolje zajamejo polisemijo in zaznavajo razlike v pomenu določenega leksema v različnih kontekstih. Druga pomembna novost novejših metod pa je v dostopnosti prednaučenih modelov: modele, naučene na velikih jezikovnih korpusih, lahko donaučimo za specifični korpus ali nalogo; s tem obdržimo splošno jezikovno znanje in razumevanje odnosa med besedami, ki pa ga prilagodimo specifičnosti posamezne uporabe.

Metode za zaznavanje semantičnih premikov na podlagi vektorskih vložitev so bile uporabljene za vrsto nalog. Različne raziskave (Kutuzov idr., »Tracing«; Tang; Tahmasebi idr.) zaznavajo pomenske premike posameznih besed s pomočjo kontekstualnih vložitev. Sorodne metode pa so bile uspešno uporabljene tudi na primer za analizo stališča (Azarbonyad idr.; Martinc idr., »EMBEDDIA«), zaznavanja dogodkov (Kutuzov idr., »Diachronic«), premikov v diskurzu (Schlechtweg idr.) in diahrone razsežnosti novic (Martinc idr., »Leveraging«), vključno z novicami o covidu-19 (Montariol idr.).

Glavna inovacija v tem prispevku je uporaba metode zaznavanja semantičnih premikov na področju digitalne humanistike, in sicer pri primerjalni analizi literarnih del. Izkaže se, da je metoda, ki temelji na primerjavi distribucij gruč (*clusters*) kontekstualnih besednih vložitev (Montariol idr.), primerna za primerjavo avtorskih literarnih opusov in za nova spoznanja o literaturi. Z vidika literarne vede prispevek prinaša nov vpogled v diskurz dveh kanoničnih slovenskih avtorjev in razgrinja enega od sestavnih delov stila kot skupka številnih formalnih značilnosti, večplastnega celovitega sistema, ki ga lahko preučujemo kvalitativno ali kvantitativno (Herrmann idr. 44–45). Analiza je namreč pokazala, da razlika med Jurčičevim realizmom in Cankarjevo moderno med drugim sloni na semantiki besed, ki jih je mogoče uvrstiti v določena pomenska polja.

V nadaljevanju so predstavljeni izhodiščni literarni koncepti, metodologija, proces analize, interpretacija in sklepne ugotovitve, pri tem pa se izmenjujeta in medsebojno dopolnjujeta kvantitativni in kvalitativni pristop oziroma oddaljeno in natančno branje.

Slovenska pripovedna proza od Jurčiča do Cankarja

Ko se je v šestdesetih letih 19. stoletja začel vzpon slovenskega pripovedništva, so bile literarne težnje povezane z vznikom slovenskega meščanstva, v njih pa je oporo našlo tudi slovensko nacionalno gibanje (Kmecl, »Problematika« 164–165). Josip Jurčič, eden vodilnih predstavnikov tega obdobja, je svoje pripovedništvo začel razvijati leta 1861 z objavami krajših in daljših povesti, leta 1866 pa je postal zastavonoša slovenskega romanopisja z romanom *Deseti brat*. Modele za roman je prevzel po literarnem programu Frana Levstika in zgodovinskih romanih Walterja Scotta (Kos, *Pregled* 160). Še isto leto, ko je izšel njegov prvi roman, je nadaljeval s pisanjem povesti, romanov in krajših proznih del (Kmecl, *Josip*). Njegova začetna dela so izraziteje navezana na ljudsko izročilo, tudi z elementi pravljičnosti, kasneje pa je vedno bolj ubesedoval zgodovinsko in meščansko tematiko (Kmecl, *Josip* 18–19, 37). Jurčič je svojo ustvarjalnost pojmoval kot realizem, medtem ko ga je literarna veda uvrstila v romantični realizem, na prehod med romantiko in realizmom, ali celo bližje romantiki (Kmecl, *Josip* 137; Kos, *Pregled* 131).

Na začetku osemdesetih let 19. stoletja se v slovenski pripovedni prozi začne razmah realizma (Kos, *Pregled* 131), ki proti koncu stoletja preraste v naturalizem, od preloma stoletja do konca 1. svetovne vojne pa pripovedništvo zaznamuje slovenska moderna kot specifično križanje različnih literarnih smeri. Vodilni predstavnik tega obdobja je Ivan Cankar, čigar pripovedni in dramski opus še danes veljata za vrhunec slovenske literature. Žanrsko v njegovi pripovedni prozi že na začetku prevladuje kratka proza, ki proti koncu njegovega ustvarjanja postane celo edina zvrst (Čeh Steger 89). Od daljše proze je po letu 1900 objavil devet romanov in z njimi v slovensko književnost uvedel nove tipe romana. Njegov pripovedni slog velja za novost v primerjavi s prejšnjimi obdobji. Ko je na začetku devetdesetih let 19. stoletja objavil svoje prve spise, se je sicer še zgledoval po domači realistično-naturalistični tradiciji, ki jo je močno zaznamoval Jurčič, ob koncu stoletja pa so na njegovo kratko prozo pa tudi romane močno vplivali moderni evropski tokovi, zlasti dekadenca, impresionizem in simbolizem. Od tod subjektivistični pogled na svet, usmerjenost dogajanja v čutno-čustveni svet in uvajanje pesniških postopkov (Kos, *Pregled* 239; Zupan Sosič 232; Čeh Steger 89–90).

Korpus besedil

Jurčičeva in Cankarjeva pripovedna besedila so v elektronski obliki javno dostopna v spletni digitalni knjižnici *Wikivir: Slovenska leposlovna klasika*, ki je nastala in se dopolnjuje v okviru projekta digitalizacije slovenske leposlovne klasike v javni domeni. Tako je bilo mogoče razmeroma hitro pridobiti pripovedna besedila obeh avtorjev za računalniško analizo. Korpus, ki zajema dobra dva milijona žetonov (*tokens*), tj. jezikovnih enot, ki ustrezajo besedi ali delu besede, obsega 68 literarnih besedil; podrobnejši podatki o velikosti korpusa so prikazani v Tabeli 1.

	Josip Jurčič	Ivan Cankar
Število dokumentov	36	32
Število stavkov	31646	49217
Število žetonov	842604	1223164

Tabela 1: Pregled obsega analiziranega korpusa, ki ga sestavljajo ročno zbrana, javno dostopna dela Josipa Jurčiča in Ivana Cankarja; tokenizacija z modelom SloBERTa besedilo razdeli na žetone.

Metodologija

Za odkrivanje razlik v besedni rabi med avtorjema je bilo treba metodologijo za odkrivanje semantičnih premikov, ki so jo opisali Syrielle Montariol, Matej Martinc in Lidia Pivovarova (Montariol idr.), prilagoditi s prostodostopno kodo (dostopno na naslovu https://github.com/RSDO-DS3/SloSemanticShiftDetection). Medtem ko je bila izvirna metoda namenjena analizi besednih premikov skozi čas, je tako prilagojena metoda omogočila primerjavo avtorjev.

Kontekstualne besedne vektorske vložitve smo pridobili z modelom SloBERTa (Ulčar in Robnik Šikonja): prednaučeni model smo z maskiranjem besed prilagodili uporabljenemu literarnemu korpusu (učenje poteka pet epoh). Za gradnjo kontekstualnih besedilnih vložitev smo dokumente najprej razdelili na stavke, vsak stavek pa obrezali na 256 žetonov. Od konteksta odvisne vložitve za vsak žeton so bile ustvarjene s seštevanjem zadnjih plasti izhoda kodirnika modela. Zaradi tokenizacije, ki v modelih tipa BERT preslika eno vhodno besedo v več žetonov, so bile besede, sestavljene iz več žetonov, predstavljene kot povprečje vseh vektorjev žetonov, ki besedo sestavljajo. Za vsako besedo smo nato

kontekstualne vektorske vložitve združili v gruče podobnih rab, kjer smo uporabili algoritem k-means (in nastavitev za pet gruč po zgledu Montariol idr.).

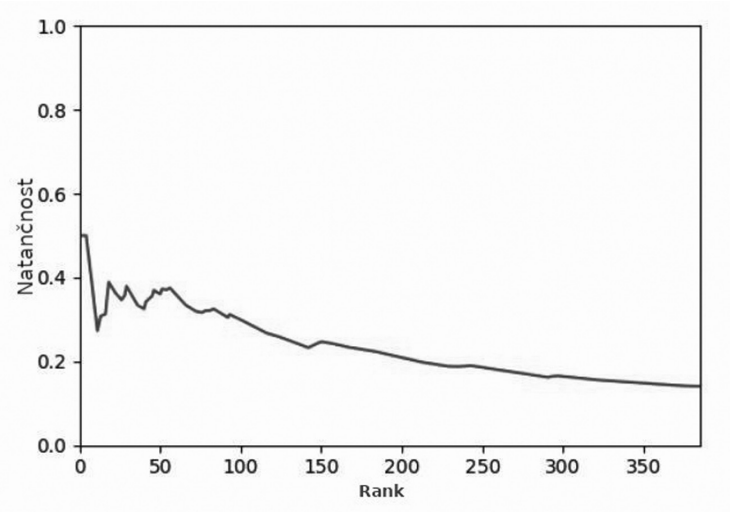
V naslednjem koraku smo metodo uporabili za primerjavo distribucij gruč. V nasprotju s pristopom Syrielle Montariol idr., kjer so avtorji primerjali rabe v različnih obdobjih, smo komponento za delitev korpusa uporabili za delitev po imenu avtorjev. Besede smo tako rangirali glede na razliko v distribuciji gruč besednih rab med avtorjema, pri čemer smo se naslonili na metriko Jensen-Shannon divergence oziroma JSD (Lin). Prav tako smo lahko za vsako besedo vizualizirali in interpretirali gruče rab, in sicer preko besed, ki gručo opisujejo (z uteževanjem tf-idf), in vizualno primerjali distribucijo rab med avtorjema.

Rezultati in izbor besed za analizo

Z uporabo opisane metodologije smo dobili rangiran seznam besed z največjo zaznano pomensko spremembo med dvema avtorjema. V nadaljevanju smo rezultate, natančneje prvih 400 besed z največjo zaznano pomensko spremembo, kvalitativno (ročno) preverili. Pri tem smo presojali, katere besede na seznamu so najbolj povedne kot diferencialni markerji stila med Jurčičem in Cankarjem, in sicer s pomočjo grafičnih prikazov razporeditve gruč besednih rab pri posameznem avtorju in kontekstualnih izpisov posameznih besed,[8] ki so bili generirani ob analizi. Z ročnim pregledom smo dobili kvalitativni vpogled v uspešnost zaznavanja pomenskih razlik z vidika uporabe metode pri literarni analizi razlik v rabi besed v besedilih različnih avtorjev. Izmed 400 besed z največjo razliko v rabi jih je bilo 57 izbranih kot ustreznih.

Natančnost, kot je opredeljena v kontekstu iskanja informacij, je del ustreznih rezultatov med vsemi vrnjenimi rezultati. Sistem za zaznavanje semantičnih premikov vrne vse besede iz korpusa, ki so bile najdene, ne glede na njihovo semantično spremembo. Zato ocenjujemo natančnost pri rangu oziroma natančnost@k, ki je opredeljena kot delež ustreznih rezultatov med najvišjimi vrnjenimi k rezultati. To metriko, ki pokaže ročno ocenjen odstotek relevantnih besed za analizo v odnosu do pozicije na rangiranem seznamu, predstavljamo na Sliki 1. Med zgornjimi 50 besedami je tako 38 % besed zanimivih za analizo, na vseh 400 besedah pa je ta odstotek 14,25.

[8] Ob kontekstualnem primeru je podan tudi podatek o avtorju in besedilu, iz katerega je primer.



Slika 1: Natančnost glede na rangiranje (natančnost@k), ki prikazuje odstotek besed, ki so zaznane z metodo za analizo semantičnih premikov in so zanimive tudi z vidika literarne vede.

Metodo za zaznavanje pomenskih razlik s kontekstualnimi vložitvami ocenjujemo kot dobro izhodišče za primerjavo avtorjev literarnih besedil. Je pa pri interpretaciji odstotka nerelevantnih besed za analizo potrebna previdnost. V primerih nerelevantnih besed ne gre nujno le za omejitev pristopa k odkrivanju pomenskih sprememb, ampak tudi za to, da so lahko nekatere besede, kjer pride do različne rabe, vsaj na prvi pogled težje prepoznavne kot relevantne, mogoče pa je tudi, da je del besed nepomemben z vidika ciljne raziskave na področju literarne analize.

Natančna primerjava avtorjev

V iskanju skupnih točk in razlik med besedami, ki so bile med rezultati kvantitativne analize izbrane za podrobnejšo presojo, smo izbrane besede v naslednji fazi[9] združevali v pomenska polja. Pri uvrščanju v pomenska polja načeloma združujemo besede, ki imajo skupno semantično komponento in ki jih uporabljamo, ko govorimo o istem pojavu ali isti topiki (Heuser in Le-Khac 4), v posameznem pomenskem polju pa so besede organizirane glede na medsebojna pomenska razmerja, in

[9] Poudariti velja, da pri presojanju, kateri pomenski premiki se zdijo najrelevantnejši, še nismo razmišljali o uvrščanju v pomenska polja, tako da je izbor nastal neodvisno od tu prikazanega grupiranja besed.

sicer je relevantna ne le sopomenskost, ampak tudi protipomenskost ter nad- in podpomenskost (Stubbs 36).

Ker splošna taksonomija pomenskih polj za slovenščino ni izdelana, smo se pri razvrščanju oziroma združevanju besed oprli na semantično taksonomijo USAS (UCREL[10] Semantic Analysis System). Ta vsebuje 21 preddefiniranih osnovnih semantičnih kategorij ali konceptov, ki so nadalje diferencirani z različnim številom podkategorij, tako da je skupno število podkategorij 232. Vsak leksem je mogoče uvrstiti v več pomenskih polj, a smo zaradi preglednosti izbrali vsakič le eno polje in zaradi razmeroma majhnega števila besed upoštevali le osnovno taksonomijo z 21 razredi. Razporeditev besed v pomenska polja[11] je prikazana v Tabeli 2, v kateri si pomenska polja sledijo od tistih z največ leksemi navzdol.

Pomensko polje	Uvrščene besede
»gibanje in lokacija«	kod, spustiti, prevzeti, seči, nazaj, odpeljati, naravnost, zapreti, stopiti, poslati, jug, pot, hoditi, popotnik, najti
»družbena dejanja in stanja«	skupaj, služabnik, tuj, neznan, nebesa, tujec, slovenski, poljub, stisniti
»psihološka dejanja in stanja«	slišati, vzbuditi, čuti, premišljevati, zagledati, spati, želeti, želja
»govorna dejanja«	zaklicati, zasmejati, molčati, potihoma, izgovorjen, glasen, praviti
»predmeti, snovi in lastnosti«	hlad, mraz, ogenj, kaplja, luč, mrzel
»telo in posameznik«	lice, bolečina, smrt
»čustvena dejanja, stanja in procesi«	veselje, sreča, težko
»čas«	čas, tema, mrak
»arhitektura, zgradbe, hiša/dom«	hiša
»uprava in javne domene«	vojska
»številke in mere«	tisoč

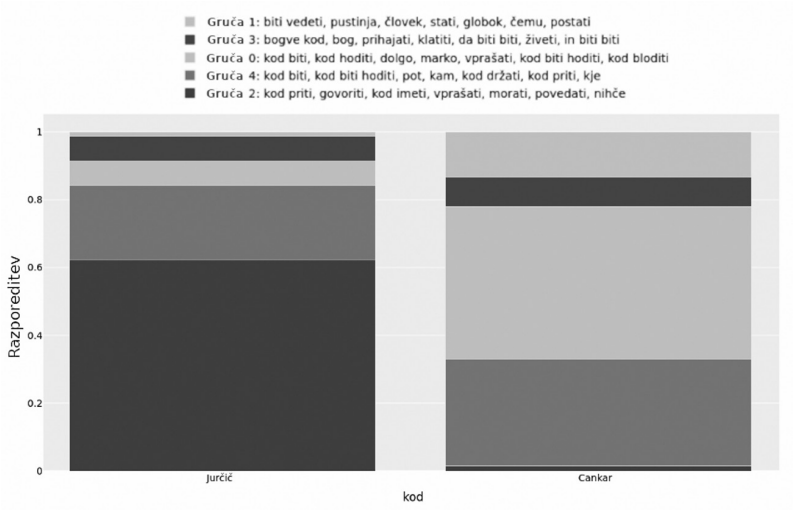
Tabela 2: Razvrstitev izbranih besed v pomenska polja.

[10] Sistem je bil razvit v Univerzitetnem središču za računalniške korpusne raziskave jezika (University Centre for Computer Corpus Research on Language) na Univerzi Lancaster v Veliki Britaniji. Taksonomija t. i. semantičnih domen je bila sicer zasnovana za avtomatsko analizo pomenskih polj oziroma semantičnih domen v korpusni stilistiki, vpeljana v številnih kvantitativnih korpusnih raziskavah literature in prevedena v več jezikov.

[11] Prepoznanih je bilo 11 semantičnih polj, torej polovica vseh, ki so v taksonomiji USAS. Nobena beseda ni bila uvrščena v sledeča pomenska polja: »splošno in abstraktno«; »umetnost in obrt«; »hrana in kmetijstvo«; »denar in trgovanje«; »zabava, šport in igre«; »življenje in živa bitja«; »izobraževanje«; »svet in okolje«; »znanost in tehnologija«; »imena in slovnične besede«.

Iz te razporeditve je mogoče razbrati, da večina besed, v katerih je bila zaznana pomembna pomenska sprememba glede na sobesedilni kontekst v opusih izbranih avtorjev, spada v semantično polje »gibanje in lokacija«. Razmeroma obsežni sta tudi pomenski polji »družbena dejanja in stanja« (to polje v najširšem smislu opisuje družbene in druge medčloveške odnose) in »psihološka dejanja in stanja« (besedišče iz tega pomenskega polja opisuje človekov notranji svet). Tem po obsegu tesno sledi pomensko polje »govorna dejanja«, temu pa polje »predmeti in snovi in lastnosti«, pri čemer – kot bomo na enem od primerov prikazali v nadaljevanju – gre pri pomenskem premiku predvsem za razmerje med osnovnim in simbolnim pomenom. Bistveno manj obsežna so pomenska polja »telo in posameznik«, »čustvena dejanja, stanja in procesi« ter »čas«, ki vsebujejo po tri lekseme. V zadnje tri kategorije (tj. »arhitektura, zgradbe, hiša/dom«, »uprava in javne domene« ter »številke in mere«) je bil uvrščen le po en leksem, zato je njihova vloga razmeroma zanemarljiva.

Iz kategorizacije v pomenska polja je mogoče izpeljati ugotovitev, da je pomembna diskurzivna razlika med avtorjema vgrajena v kronotop njunih pripovednih del; tega predstavlja »gibanje in lokacija«, najobsežnejše pomensko polje, v kombinaciji s pomenskim poljem »čas«, ki je sicer med manj obsežnimi. Za ilustracijo na Sliki 2 podrobneje predstavljamo grafični prikaz in sobesedilne primere za leksem *kod* iz najobsežneje zastopanega pomenskega polja.



Slika 2: Grafični prikaz pomenskih razlik v rabi leksema *kod*, zaznanih z metodo za zaznavanje semantičnih premikov; posamezna gruča predstavlja tipično besedno okolico izbranega leksema.

Leksem *kod* je četrti na seznamu vseh leksemov, ki izkazujejo visoko stopnjo pomenske spremembe (JSD 0,291). Primerjava pokaže, da v Jurčičevem opusu pri *kod* izrazito prevladuje 2. gruča – *kod priti*, *govoriti*, *kod imeti*, *vprašati*, *morati*, *povedati* in *nihče* – medtem ko pri Cankarju prevladuje gruča 0 – *kod biti*, *kod biti hoditi*, *dolgo*, *Marko*,[12] *vprašati*, *kod biti hoditi* in *kod bloditi*. Ta gruča je tudi pojmovno povezana s 1. gručo, ki je pri Cankarju prav tako pogosta, namreč *biti vedeti*, *pustinja*, *človek*, *stati*, *globok*, *čemu* in *postati*. Razliko med tipičnim sobesedilom lahko prepoznamo kot razmerje med Jurčičevim ciljnim gibanjem likov v konkretnem literarnem prostoru in času, medtem ko se pri Cankarju leksem *kod* navezuje na prostorsko nedoločljivost in brezciljno tavanje. Spodnji kontekstualni primeri ponazarjajo, kako je ta semantični premik tipično izražen v besedilih:

> Vedel in videl je, kod se pot vije in kje zopet v hosti izgine. (Josip Jurčič, *Sin kmečkega cesarja*)

> Leonu pak je še posebej pokazal, kod se pride na stranski kor, kjer se dobro k oltarju in na prižnico vidi in kamor navadno hodijo civilni ljudjé njegove fare. (Josip Jurčič, *Cvet in sad*)

> Kod blodim tri dni, ali morda že teden dni? (Ivan Cankar, *Troje povesti*)

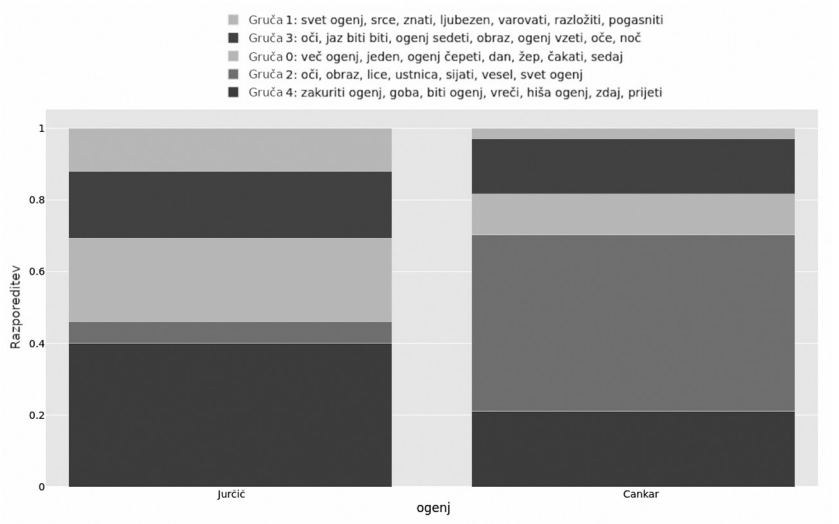
> Kod bega mati, da bi prinesla mesá in morda celó jabolk in orehov? (Ivan Cankar, *Hiša Marije Pomočnice*)

Drugi izbrani primer, ki ga podrobneje interpretiramo, je leksem *ogenj* iz pomenskega polja »predmeti, snovi in lastnosti« (gl. Sliko 3). Gruča 2 – *oči*,[13] *obraz*, *lice*, *ustnica*, *sijati*, *vesel*, *svet ogenj* –, ki pri Cankarju prevladuje, izraža povezave ognja z obrazom, kar nakazuje simbolnost, medtem ko je pri Jurčiču zaznana izrazita povezava s konkretnimi dejanji in samim obstojem ognja, na primer *zakuriti*, *vreči*, *prijeti*, *biti* (gl. gručo 4).

---

[12] Lastna imena so sicer posebna, za nas nepomembna kategorija. Če je lastno ime zaznano kot beseda, ki se spreminja, kaže na primer na isto ime dveh literarnih likov. Če se pojavlja pri opisu gruče, sicer kaže na sopojavljanje z besedo, vendar nam o sami rabi in analizi razlik ne pove veliko, tako da pri interpretaciji njihova posamezna pojavnost na celoto rezultatov tako rekoč ne vpliva.

[13] V tem kontekstu velja opozoriti na kvantitativno raziskavo besed in pomenskih polj v pripovedni prozi in dramatiki Ivana Cankarja, ki ugotavlja, da je najpogostejša polnopomenska beseda pri Cankarju *oči*. Raziskava izpostavlja tudi pogostost besed za označevanje delov telesa, zlasti podrobnosti obraza (Mikolič 23, 32).



Slika 3: Grafični prikaz pomenskih razlik v rabi leksema *ogenj*, zaznanih z metodo za zaznavanje semantičnih premikov.

Poleg tega primerjava kontekstualne rabe pokaže, da je pri Cankarju leksem *ogenj* v povezavi z deli obraza, v katerih se zrcali notranjost človeka, najobičajneje simbol strasti, življenjske energije:

> Nobenega ognja ni bilo v črnih, motnih očeh. (Ivan Cankar, *Vinjete*)

> Njen obraz je bil, kakor jih je ljubil: svež in vesel, ves poln ognja in mladosti, njene temnosive oči so gledale prešerno in odkrito[.] (Ivan Cankar, *Tujci*)

> Na potepuhovih kolenih je sedela ženska in ga je objemala z golo roko okoli vratu; nanjo je sijala vsa luč, kakor v ognju so bila lica, ampak iz vsega ognja so žarele velike, pregrešne ciganske oči. (Ivan Cankar, *Zgodbe iz doline Šentflorjanske*)

Drugače je pri Jurčiču, kjer leksem *ogenj* najpogosteje poimenuje dejansko substanco:

> Nazaj prišedši, je celo v kuhinjo stopil, iz velikega ognja, pri katerem je Franica kosilo kuhala, žareč ogel za pipo vzel ter molče zopet odšel[.] (Jurčič, *Sosedov sin*)

Kot je prav tako mogoče razbrati iz grafičnih prikazov, ima *ogenj* tudi pri Jurčiču lahko simbolni pomen, a v bistveno manjši meri kakor pri

Cankarju, večinoma pa tudi v drugačnem kontekstu. Navedimo značilen primer:

> Nikakor ne morem tajiti, da me srce ne vleče do hčere mojega gospodarja, da, to moram izpovedati se tebi, predragi moj, če je eno bitje na božjem daljnem svetu, za katero bi se jaz z vsem mladostnim ognjem vnel, za katero bi norel in gorel, kakor je kdaj kak človek mogel, bila bi to enaka deklica, kakor je ta, katere sem ti v zadnjem pismu omenil. (Josip Jurčič, *Deseti brat*)

Sklepne ugotovitve

Zgornja primera vsebinske interpretacije grafičnih prikazov rezultatov kvantitativne analize in delitev besed, za katere je bil prepoznan pomenljiv semantični premik, v pomenska polja kažejo, da metoda omogoča globlji primerjalni uvid v sam način, na katerega avtorja na semantični ravnini gradita svoj literarni stil, in v to, katera pomenska polja, ki so obenem tudi skupna pomenska polja pripovedne proze izbrane dvojice avtorjev, so nosilci prehoda od Jurčičevega (romantičnega) realizma k Cankarjevi simbolistični moderni. Analiza in interpretacija sta pokazali, da sprememba v pomenski plati ni toliko vezana na neposredno izražanje čutno-čustvene komponente, kot bi nemara pričakovali, kolikor je izražena v vsaj na videz splošnejših pomenskih poljih, kot so – poleg »gibanja in lokacije« – »družbena dejanja in stanja«, »psihološka dejanja in stanja« in »govorna dejanja«.

V tej raziskavi so bile besede oziroma pomenski premiki, ki naj bi bili najbolj pomenljivi z vidika razmerja slogov izbranih avtorjev, prepoznani z (ročno ali neavtomatsko) kvalitativno analizo, nato pa so bile izbrane besede razporejene v pomenska polja. Za primerjavo med ročno izbranimi besedami in prvimi 100 besedami na seznamu smo v pomenska polja naknadno razporedili vseh prvih 100 besed na avtomatsko rangiranem seznamu. Izkazalo se je, da to na obseg posameznih pomenskih polj ni imelo bistvenega vpliva,[14] pa tudi nova pomenska polja se niso oblikovala, iz česar je mogoče sklepati, da bi do enako relevantnih rezultatov kakor s kvalitativno analizo rezultatov avtomatske analize prišli tudi z razporeditvijo vseh prvih 100, 150[15] ali celo več besed v pomenska polja, kar bi zahtevalo bistveno manj ročnega dela.

[14] Nekoliko obsežnejše je bilo le polje »govorna dejanja«.

[15] Prim. Sliko 1, iz katere je razvidno, da natančnost, ki je izračunana iz primerjave med kvantitativno in kvalitativno (ročno) analizo, po približno 150 prvih besedah na rangiranem seznamu izraziteje upade.

Rezultati nakazujejo tudi možnosti nadaljnjih raziskav, saj bi bilo mogoče pomenski premik opazovati glede na obdobja ustvarjanja teh dveh avtorjev (zlasti ker za oba velja, da sta iz zgodnje faze, navezane na tradicijo, ter preko vrhunca ustvarjalnosti prešla v svoje zrelo obdobje) ali glede na žanrsko delitev njune pripovedne proze. Vsekakor je raziskavo mogoče razširiti tudi na druge avtorje in obdobja pripovedne proze dolgega 19. stoletja in širše. Poleg tega bi bila zelo zanimiva medjezična analiza literarnih del, ki bi temeljila na večjezičnih jezikovnih modelih, kot sta mBERT (Devlin idr.) in XLM-R (Conneau idr.), saj bi omogočila preučevanje evolucije literarnega sloga ne le med žanri, ampak tudi med različnimi jeziki in kulturami.

Z računalniškega vidika se zdi, da pristop za odkrivanje semantičnih sprememb, ki so ga predlagali S. Montariol, Martinc in L. Pivovarova, dobro ustreza literarni analizi in preučevanju razlik v rabi besed med avtorji. Možna pomanjkljivost pri uporabi metod, ki temeljijo na velikih prednaučenih jezikovnih modelih, je v koraku prilagoditve domene za uporabo v literarni vedi. Čeprav model v fazi dotreniranja prilagodimo novi porazdelitvi besed, je velikost korpusa v naši študiji majhna. Zato bi ga bilo smiselno najprej prilagoditi na velikem korpusu vseh dostopnih literarnih besedil in ga šele nato prilagoditi posebni študiji primera. Razlike med rabo besed v literarni umetnosti so lahko kompleksnejše od korpusov, na katerih je bil usposobljen model, ki poleg literature zajema tudi veliko neliterarnih besedil. Poleg tega je predpogoj za računalniško primerjavo besedne semantike med dvema korpusoma tudi število pojavitev iste besede, to pa je v nasprotju s človeško sposobnostjo sklepanja o pomembnih razlikah v semantiki z enim samim primerom.

V prispevku smo se dotaknili primerjalne analize v kontekstu literarne umetnosti z uporabo velikih jezikovnih modelov. Ta vrsta analize z uporabo zaznavanja semantičnih sprememb je sama po sebi omejena na semantične razlike besed. Toda umetniškega sloga seveda ne opredeljujeta le semantika in sposobnost avtorja, da izčrpno izrazi in celo preobremeni semantiko besede, ampak tudi druge razsežnosti pisanja.

LITERATURA

Alvarado, Rafael C. »The Digital Humanities Situation«. *Debates in the Digital Humanities*, ur. Matthew K. Gold, University of Minnesota Press, 2012, str. 50–55.

Azarbonyad, Hosein, idr. »Words are Malleable: Computing Semantic Shifts in Political and Media Discourse«. *CIKM '17: Proceedings of the 2017 ACM Conference on Information and Knowledge Management*, Association for Computing Machinery, 2017, str. 1509–1518.

Bode, Katherine. *Reading by Numbers: Recalibrating the Literary Field*. Anthem Press, 2012.

Burrows, John F. *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press, 1987.

Conneau, Alexis, idr. »Unsupervised Cross-lingual Representation Learning at Scale«. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, ur. Dan Jurafsky idr., The Association for Computational Linguistics, 2020, str. 8440–8451, https://aclanthology.org/2020.acl-main.747.pdf. Dostop 5. 2. 2024.

Čeh Steger, Jožica. »Kratka proza«. *Ivan Cankar: literarni revolucionar*, ur. Aljoša Harlamov, Cankarjeva založba, 2018, str. 88–141.

Devlin, Jacob, idr. »BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding«. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 1. zv., ur. Jill Burstein idr., The Association for Computational Linguistics, 2019, str. 4171–4186, https://aclanthology.org/N19-1423.pdf. Dostop 5. 2. 2024.

Earhart, Amy E. *Traces of the Old, Uses of the New: The Emergence of Digital Literary Studies*. University of Michigan Press, 2015.

Enkvist, Nils Erik. »On Defining Style«. *Linguistics and Style*, ur. Nils Erik Enkvist idr., Oxford University Press, 1964, str. 1–56.

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

Ganascia, Jean-Gabriel. »The Logic of the Big Data Turn in Digital Literary Studies«. *Frontiers in Digital Humanities*, št. 2, 2015, https://doi.org/10.3389/fdigh.2015.00007. Dostop 5. 2. 2024.

Herrmann, J. Berenike, idr. »Revisiting Style, a Key Concept in Literary Studies«. *Journal of Literary Theory*, let. 9, št. 1, 2015, str. 25–52.

Heuser, Ryan, in Lang 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 5. 2. 2024.

Hoover, David L., idr. *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. Routledge, 2014.

Jannidis, Fotis, in Gerhard Lauer. »Burrows's Delta and Its Use in German Literary History«. *Distant Readings: Topologies of German Culture in the Long Nineteenth Century*, ur. Matt Erlin in Lynne Tatlock, Camden House, 2014, str. 29–54.

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

Kmecl, Matjaž. *Josip Jurčič: pripovednik in dramatik*. Zavod Republike Slovenije za šolstvo, 2009.

Kmecl, Matjaž. »Problematika slovenske proze 19. stoletja«. *Zbornik predavanj / XV. seminar slovenskega jezika, literature in kulture, 2.–14. julija 1979*, ur. Breda Pogorelec

in Ljubica Črnivec, Filozofska fakulteta, Pedagoško-znanstvena enota za slovanske jezike in književnosti, 1979, str. 157–182.
Kos, Janko. »Cankar in problem slovenskega romana«. *Sodobnost*, let. 24, št. 5, 1976, str. 413–423.
Kos, Janko. *Pregled slovenskega slovstva*. DZS, 2010.
Kuhn, Virginia, in Vicki Callahan. »Nomadic Archives: Remix and the Drift to Praxis«. *Digital Humanities Pedagogy*, ur. Brett D. Hirsch, Open Book Publishers, 2012, str. 291–308, https://books.openbookpublishers.com/10.11647/obp.0024.pdf. Dostop 5. 2. 2024.
Kutuzov, Andrey, idr. »Diachronic Word Embeddings and Semantic Shifts: A Survey«. *Proceedings of the 27th International Conference on Computational Linguistics*, ur. Emily M. Bender idr., The Association for Computational Linguistics, 2018, str. 1384–1397, https://aclanthology.org/C18-1117. Dostop 5. 2. 2024.
Kutuzov, Andrey, idr. »Tracing Armed Conflicts with Diachronic Word Embedding Models«. *Proceedings of the Events and Stories in the News Workshop*, ur. Tommaso Caselli idr., The Association for Computational Linguistics, 2017, str. 31–36, https://aclanthology.org/W17-2705.pdf. Dostop 5. 2. 2024.
Lin, Jianhua. »Divergence Measures Based on the Shannon Entropy«. *IEEE Transactions on Information Theory*, let. 37, št. 1, 1991, str. 145–151.
Mahony, Simon. »Cultural Diversity and the Digital Humanities«. *Fudan Journal of the Humanities and Social Sciences*, št. 11, 2018, str. 371–388.
Martinc, Matej, idr. »EMBEDDIA Hackathon Report: Automatic Sentiment and Viewpoint Analysis of Slovenian News Corpus on the Topic of LGBTIQ+«. *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, ur. Hannu Toivonen in Michele Boggia, The Association for Computational Linguistics, 2021, str. 121–126, https://aclanthology.org/2021.hackashop-1.17. Dostop 5. 2. 2024.
Martinc, Matej, idr. »Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift«. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, ur. Nicoletta Calzolari idr., European Language Resources Association, 2020, str. 4811–4819, https://aclanthology.org/2020.lrec-1.592. Dostop 5. 2. 2024.
McCarty, Willard. »Humanities Computing«. *Encyclopedia of Library and Information Sciences*, ur. Miriam Drake, Marcel Dekker, 2003, str. 1224–1235.
Mikolič, Vesna. *Ali bereš Cankarja?*. Slovenska matica, 2021.
Montariol, Syrielle, idr. »Scalable and Interpretable Semantic Change Detection«. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ur. Kristina Toutanova idr., The Association for Computational Linguistics, 2021, str. 4642–4652, https://aclanthology.org/2021.naacl-main.369. Dostop 5. 2. 2024.
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.
Murray, Simone. »Varieties of Digital Literary Studies: Micro, Macro, Meso«. *DHQ: Digital Humanities Quarterly*, let. 16, št. 2, 2022, http://www.digitalhumanities.org/dhq/vol/16/2/000616/000616.html. Dostop 5. 2. 2024.
Piper, Andrew. »There Will Be Numbers«. *Journal of Cultural Analytics*, let. 1, št. 1, 2016, https://culturalanalytics.org/article/11062-there-will-be-numbers. Dostop 5. 2. 2024.
Pirjevec, Dušan. »Problem slovenskega romana«. *Literatura*, let. 9, št. 67–68, 1997, str. 63–75.

Presner, Todd. »Comparative Literature in the Age of Digital Humanities«. *A Companion to Comparative Literature*, ur. Ali Behdad in Dominic Thomas, Blackwell, 2011, str. 193–208.

Ramsay, Stephen. »Who's In and Who's Out«. *Defining Digital Humanities: A Reader*, ur. Melissa Terras idr., Ashgate, 2013, str. 239–241.

Rodríguez Ortega, Nuria. »Five Central Concepts to Think of Digital Humanities as a New Digital Humanism Project«. *Artnodes*, št. 22, 2018, https://doi.org/10.7238/a.v0i22.3263. Dostop 5. 2. 2024.

Rommel, Thomas. »Literary Studies«. *A Companion to Digital Humanities*, ur. Susan Schriebman idr., Blackwell, 2004, str. 88–96.

Schlechtweg, Dominik, idr. »SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection«. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, 2020, https://aclanthology.org/2020.semeval-1.1. Dostop 5. 2. 2024.

Stubbs, Michael. *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell, 2005.

Svensson, Patrik. »Beyond the Big Tent«. *Debates in the Digital Humanities*, ur. Matthew K. Gold, University of Minnesota Press, 2012, str. 36–72.

Svensson, Patrik. »Humanities Computing as Digital Humanities«. *Digital Humanities Quarterly*, let. 3, št. 3, 2009, http://www.digitalhumanities.org/dhq/vol/3/3/000065/000065.html. Dostop 5. 2. 2024.

Tahmasebi, Nina, idr. »Survey of Computational Approaches to Lexical Semantic Change Detection«. *Computational Approaches to Semantic Change*, ur. Nina Tahmasebi idr., Language Science Press, 2021, str. 1–91.

Tang, Xuri. »A State-of-the-art of Semantic Change Computation«. *Natural Language Engineering*, let. 24, št. 5, 2018, str. 649–676.

Terras, Melissa, idr. »Selected Definitions From the Day of Digital Humanities: 2009–2012«. *Defining Digital Humanities: A Reader*, ur. Melissa Terras idr., Ashgate, 2013, str. 279–287.

Ulčar, Matej, in Marko Robnik Šikonja. *Slovenian RoBERTa Contextual Embeddings Model: SloBERTa 2.0*. Fakulteta za računalništvo in informatiko, 2021, http://hdl.handle.net/11356/1397. Dostop 5. 2. 2024.

Vanhoutte, Edward. »The Gates of Hell: History and Definition of Digital | Humanities | Computing«. *Defining Digital Humanities: A Reader*, ur. Melissa Terras idr., Ashgate, 2013, str. 119–156.

Zajc, Ivana, in Peter Purg. *Digitalna humanistika in literatura*. Založba Univerze, 2023, https://www.ung.si/sl/zalozba/43/digitalna-humanistika-in-literatura. Dostop 5. 2. 2024.

Zupan Sosič, Alojzija. »Romani«. *Ivan Cankar: literarni revolucionar*, ur. Aljoša Harlamov, Cankarjeva založba, 2018, str. 200–233.

Comparing Josip Jurčič and Ivan Cankar Using Computational Semantic Change Detection Methods

Keywords: Slovenian literature / Jurčič, Josip / Cankar, Ivan / natural language processing / semantical analysis / digital literary studies / digital humanities

The article begins with a presentation of the heterogeneity and interdisciplinarity of the digital humanities as two central and interrelated concepts. In the main part of the article, the method of detecting semantic changes based on contextual word embeddings for the analysis of literary works is introduced. The method's potential is demonstrated through a comparative analysis of the narrative works of two canonical Slovenian authors belonging to two distinct literary periods, Josip Jurčič and Ivan Cankar, in particular through the automatic recognition of words whose meanings differ between the authors. The differences in literary style are further interpreted via a qualitative analysis of the automatically obtained results, followed by a manual categorization into semantic fields of the words that were qualitatively identified as informative of stylistic differences between Jurčič and Cankar. The article shows that the approach based on contextual word embeddings can be used for literary analysis with satisfactory results. This enables narratology to gain new insight into the oeuvres of Cankar and Jurčič, as the article shows that the difference between Jurčič's (romantic) realism and Cankar's kind of modernism (moderna) is also based on the semantics of discourses related to movement and social and psychological actions and processes.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.163.6.09:004
DOI: https://doi.org/10.3986/pkn.v47.i2.07