Kvantitativna analiza razmerij med semantičnimi polji v slovenski pripovedni prozi dolgega 19. stoletja

Lucija Mandić

Inštitut za slovensko literaturo in literarne vede ZRC SAZU, Novi trg 2, 1000 Ljubljana, Slovenija
https://orcid.org/0009-0000-7858-6513
lucija.mandic@zrc-sazu.si

V prispevku je uporabljena metoda vektorskih vložitev besed za analizo odnosov med semantičnimi polji v slovenski pripovedni prozi t. i. dolgega 19. stoletja. Z uporabo tehnologije Word2Vec v programskem jeziku Python je bila opravljena analiza Korpusa daljše slovenske pripovedne proze (KDSP 1.0). Za potrebe analize so bila konstruirana semantična polja za štiri družbene institucije: ekonomijo, politiko, kulturo in gospodinjstvo. Nabor besed za posamezno semantično polje je bil pridobljen z identifikacijo 50 besed z največjo kosinusno bližino vektorski predstavitvi vsake institucije. Nabor tako pridobljenih vektorjev je ponudil kvantitativno osnovo za raziskovanje odnosov med temi družbenimi institucijami, upodobljenimi v literarnih delih, zajetih v korpusu. Ugotovitve kažejo na pomembno prekrivanje med diskurzivnima poljema politike in kulture, s tem pa omogočajo kvantitativni pristop k pojavu, ki ga tradicionalna, kvalitativna literarna veda obravnava s konceptoma prešernovske strukture in slovenskega kulturnega sindroma.

Ključne besede: digitalna literarna veda / slovenska pripovedna proza / prešernovska struktura / slovenski kulturni sindrom / semantična analiza / vektorske vložitve besed

Računalniški pristopi k semantični analizi literarnih besedil

Pospešen tehnološki razvoj na področju tehnik strojnega učenja je sprožil metodološke premike tudi na področju humanistike.[1] Nadzorovano in nenadzorovano strojno učenje sta se izkazala za učinkoviti metodi za analize velikih količin besedilnih podatkov; predvsem nenadzorovane tehnike strojnega učenja so uporabne za luščenje informacij iz nestrukturiranega besedila. Vektorske vložitve besed (ang. *word embeddings*)

[1] Članek je nastal v okviru raziskovalnega programa »Literarnozgodovinske, literarnoteoretične in metodološke raziskave« (P6-0024), ki ga financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

so tehnika procesiranja naravnega jezika, ki uporablja strojno učenje za določanje konteksta posameznih besed v besedilu. V digitalni literarni vedi spada ta metoda med najbolj priljubljene pristope k semantičnim analizam (Hatzel idr. 203). Uporablja se za lociranje literarnih postopkov (Schneider idr.), za prepoznavanje avtorskega sloga (Eder in Šeļa), za prostorske analize (Herrmann, Byszuk in Grisot) in analize sentimenta (Brottrager idr.). V slovenski digitalni literarni vedi z izjemo semantične analize del Zofke Kveder (Pollak, Martinc in Mihurko) možnosti, ki jih ponujajo vektorske vložitve besed, še niso bile podrobneje raziskane.

Pri vektorskih vložitvah besed je semantična bližina besed izražena z numeričnimi koordinatami. Pri učenju jezikovnega modela so glede na kontekst vsaki besedi dodeljene koordinate večdimenzionalnih vektorjev, ki glede na določitve parametrov obsegajo med $n = 100$ in $n = 500$ dimenzij. Te določijo težo vektorja in njegovo numerično reprezentacijo, ki je za vsak vektor edinstvena. Pomensko sorodne besede, ki se pojavljajo v podobnih kontekstih, imajo posledično podobne vektorje in so si na ta način blizu v vektorskem prostoru. Razdalja med vektorji je običajno izračunana s pomočjo manhattanske, evklidske ali kosinusne razdalje.

Tehnologija Word2Vec (Mikolov idr.) v Pythonovi knjižnici Gensim, ki je uporabljena v analizi, predstavljeni v članku, je dvoslojno nevronsko omrežje, ki pretvori besedilo v nabor edinstvenih vektorjev, razporejenih v visokodimenzionalnem prostoru. Analize semantičnih relacij, ki jih zazna Word2Vec, so pokazale, da vektorske vložitve predvidijo položaje besed, ki presegajo preprosto semantično bližino; s seštevanjem in odštevanjem vektorskih reprezentacij besed model zazna tudi analogne odnose med pomeni besed. Avtorji orodja so z naslednjim izračunom vektorjev: »kralj« – »moški« + »ženska« prišli do vektorja, ki je najbližje vektorski reprezentaciji besede »kraljica«. Po podobnem načelu lahko jezikovni model predvidi na primer relacije med državami in njihovimi glavnimi mesti. Če vemo, kaj je Pariz v relaciji do Francije, model pravilno predvidi, da gre v primeru Berlina in Nemčije za primerljiv odnos. S tega vidika gre pri vektorskih vložitvah besed za kompleksnejšo reprezentacijo pomenskih odnosov med besedami kakor pri preprostih analizah kolokacij.

Word2Vec je nenadzorovana tehnika strojnega učenja, ki pa pri učenju jezikovnega modela omogoča nastavitve naslednjih parametrov: velikosti okna (ang. *window size*), s katero določimo obseg konteksta besede, ki bo uporabljen za učenje: na primeru povedi »Prvi slovenski roman je napisal Josip Jurčič.« bi bili v primeru velikosti okna *window*

= *2* za učenje uporabljeni dve besedi na vsaki strani ciljne besede – če je ciljna beseda »roman«, bo levi del konteksta »prvi« in »slovenski«, desni pa »je« in »napisal«; velikosti vektorja (ang. *vector size*), s katero določimo obseg skritega nivoja nevronske mreže, od te pa je odvisna že omenjena dimenzija vektorskega prostora; najnižje frekvence pojavitev besed v korpusu: če na primer uporabimo nastavitev *min_count = 100*, bodo pri učenju jezikovnega modela upoštevane le besede, ki se v korpusu pojavijo najmanj stokrat.

Word2Vec omogoča aplikacijo dveh različnih arhitektur prediktivnega modela za učenje: zvezne vreče besed (ang. *Continuous Bag-of-Words* oziroma CBOW) in preskočnega n-grama (ang. *Skip-gram* ali SG). CBOW poskuša prek konteksta priti do ciljne besede (za učenje besede »roman« bi kot vhod v nevronsko mrežo uporabili besede »prvi«, »slovenski«, »je« in »napisal« ter pri izhodu pričakovali besedo »roman«). Kontekst je na ta način predstavljen kot »vreča besed«, zajetih v oknu fiksne velikosti okoli ciljne besede. Arhitektura SG deluje v obratni smeri: kot vhod v nevronsko mrežo je uporabljena ciljna beseda (v tem primeru »roman«), na podlagi katere se model nauči njenega konteksta (pri izhodu iz nevronske mreže torej pričakujemo besede »prvi«, »slovenski«, »je« in »napisal«).

Izbira arhitekture je v veliki meri odvisna od velikosti korpusa. Medtem ko je arhitektura CBOW bistveno hitrejša, arhitektura SG natančneje predstavi redke besede in je zato primernejša za majhne nabore podatkov (Mikolov idr.), med katere z 11.454.627 besedami sodi tudi Korpus daljše slovenske pripovedne proze.

Korpus in predpriprava besedila

Za analizo je bil uporabljen Korpus daljše pripovedne proze KDSP 1.0 (Mandić in Erjavec). Korpus vsebuje 262 literarnih besedil, daljših od 20.000 besed, ki so izšla med letoma 1836 in 1918. Besedila so bila zajeta iz treh virov: iz Digitalne knjižnice Slovenije (dLib), projekta Slovenska leposlovna klasika na spletu, v okviru katerega so besedila v odprtem dostopu objavljena na portalu Wikivir, in korpusa ELTeC-slv (Erjavec idr.). Optično prepoznavanje znakov (OCR) vseh besedil je bilo ročno pregledano in popravljeno, korpus pa je bil nato stavčno segmentiran, lematiziran in tokeniziran. V korpusu so označene tudi imenske entitete ter morfosintaktične oznake MULTEXT-East in Universal Dependencies. Korpus je bil avtomatsko anotiran s programom CLASSLA (Ljubešić in Dobrovoljc).

Ta verzija korpusa je dostopna v formatu XML-TEI na repozitoriju CLARIN.SI.

Korpus je opremljen tudi z bibliografskimi in biografskimi metapodatki, kjer so bili ti na voljo. Preostali metapodatki vsebujejo tudi informacije o desetletju izida, literarni vrsti, literarni podvrsti, tematiki in stopnji kanoničnosti, ki je bila posebej relevantna za to analizo. Stopnja kanoničnosti je bila določena na podlagi vključenosti v slovenska šolska berila po letu 1980 oziroma v knjižno zbirko Zbrana dela slovenskih pesnikov in pisateljev. Dela, ki so vključena v omenjena izbora, so označena z visoko stopnjo kanoničnosti (teh je v celotnem korpusu 80), preostala pa z nizko stopnjo kanoničnosti (182).

Konstrukcija semantičnih polj

Analiza se v veliki meri naslanja na članek Laure Nelson, ki v svoji raziskavi konstruiranja družbenih institucij na ameriškem jugu t. i. dolgega 19. stoletja izhaja iz Marxove, Webrove ter feministične in kritične teorije družbe (Nelson 4–5). Osredotoča se na politiko, ekonomijo, kulturo in gospodinjstvo kot na osrednje družbene instance v obdobju industrijske revolucije v ZDA, ko politika, ekonomija in kultura izpodrivajo gospodinjstvo v zasebno sfero, podrejeno javni (4–5). Ker obravnava pol-periferno okolje v obdobju industrijske revolucije, so lahko družbene institucije, na katere se osredotoča, tudi izhodišče za semantične analize korpusa KDSP, ki zajema literarna dela, objavljena od druge četrtine 19. stoletja do konca 1. svetovne vojne. Po zgledu študije Laure Nelson bo kot ključen del analize reprezentacija teh družbenih institucij v korpusu prikazana s pomočjo konstrukcije semantičnih polj, ki se nanašajo nanje. Za razliko od omenjene študije, ki z vektorskimi vložitvami besed analizira korpus dnevniških, pisemskih, spominskih in drugih avtobiografskih spisov navadnih ljudi, bo v tem primeru analiziran korpus literarnih besedil s poudarkom na semantičnih poljih politike in kulture, namesto odnosa med intersekcionalnimi družbenimi identitetami in omenjenimi družbenih institucijami, pa bodo v ospredju odnosi med samimi institucijami.

Za te štiri institucije so bila skonstruirana semantična polja s pomočjo že omenjenega orodja Word2Vec v knjižnici Gensim v programskem jeziku Python. Jezikovni model je bil naučen na celotnem korpusu KDSP 1.0, ki pa je bil za potrebe učenja primerno razčlenjen in filtriran. Pri učenju so bile upoštevane samo leme glagolov in samostalnikov, izločene pa so bile imenske entitete. Na ta način so bile

ohranjene le polnopomenske besede z izjemo pridevnikov. Parametri so bili nastavljeni na *sg = 1* (uporabljena je bila arhitektura SG), *window size = 3*, *vector_size = 100* in *min_count = 10*. Z nastavitvijo spodnje meje frekvence besed, ki so bile uporabljene za učenje, je bil v veliki meri izključen šum (npr. napačno označene besede), ki se običajno pojavi ob avtomatskem označevanju starejših besedil. V prvem delu analize je bil za učenje modela Word2Vec porabljen celoten korpus besedil, v drugem delu analize pa sta uporabljena še podkorpus z nizko stopnjo kanoničnosti in podkorpus z visoko stopnjo kanoničnosti. Obsegi vseh treh korpusov po filtriranju nepolnopomenskih besed, pridevnikov in imenskih entitet so razvidni v Tabeli 1.

	Celoten korpus	**Visoka stopnja kanoničnosti**	**Nizka stopnja kanoničnosti**
Število dokumentov	262	80	182
Število pojavnic (ang. *token*)	3.790.873	1.197.035	2.593.838
Število različnic (ang. *type*)	62.392	36.690	48.506

Tabela 1: Velikosti korpusa in podkorpusov, ki so bili uporabljeni za učenje jezikovnih modelov Word2Vec.

Word2Vec vsebuje funkcijo *most_similar*, ki najde vektorje, ki so najbližje vektorski reprezentaciji besede, ki jo iščemo. Funkcija omogoča tudi, da s seštevanjem vektorjev iščemo semantično sorodne besede dveh ali več besed hkrati oziroma da semantično polje reduciramo z njihovim odštevanjem. To je uporabno predvsem pri večpomenskih besedah, ki jim na ta način lahko natančneje določimo kontekst. Semantično polje za sfero kulture je bilo pridobljeno z vektorji, ki so najbližje vektorjema »umetnost« + »literatura«, za sfero politike sta bila uporabljena vektorja »narod« + »država«, za ekonomijo »denar« + »kapital«, za gospodinjstvo pa vektor »družina« + »otrok«. Za vsako semantično polje je bilo izluščenih 50 vektorjev z najbližjo kosinusno razdaljo[2] (Tabela 2).

[2] Kosinusna razdalja meri kosinus kota med dvema vektorjema in se giblje od -1 do 1, pri čemer sta pri vrednosti 1 vektorja identična, pri -1 nasprotna, pri 0 pa ležita pravokotno drug na drugega. Odštevanje kosinusne podobnosti od 1 daje mero razdalje, kjer višje vrednosti implicirajo večjo podobnost med vektorji.

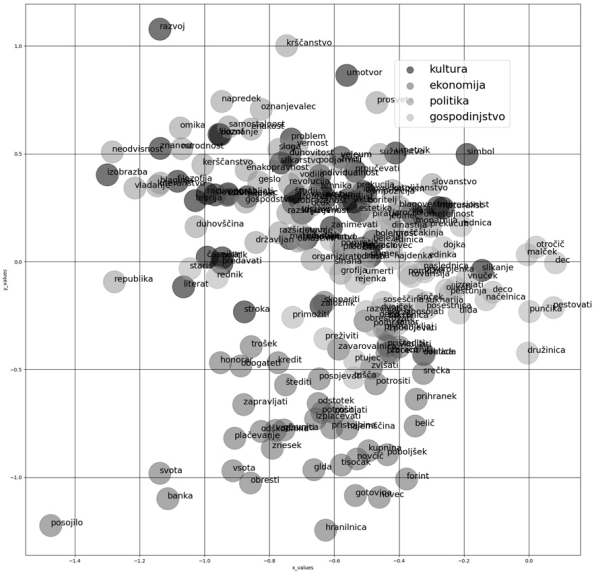
Izhodiščni vektorji	Semantično polje
umetnost + literatura	slikarstvo, ukus, slovstvo, kompozicija, proizvod, klasik, umotvor, stroka, realist, znanost, učenjak, matematika, časnikar, politik, duhovitost, impresionizem, naturalist, literat, filozofija, tehnika, predavati, proučevati, teorija, proučiti, izobraženost, problem, književnost, filozof, estetika, impresionist, izobraženec, razvoj, beležiti, študij, umeteljnost, veleum, razumnost, izobraženje, primes, zmisel, založnik, simbol, zanimevati, individualnost, jezikoslovec, umetnik, slikanje, izobrazba, poznanje, uporabljati
narod + država	ustava, boritelj, cesarstvo, revolucija, kerščanstvo, samostalnost, sloga, blaginja, enakopravnost, duhovenstvo, prekucija, geslo, gospodstvo, narodnost, vladanje, grofija, izobraženec, razširjanje, neodvisnost, enakost, živelj, napredek, blagovestnik, podjarmiti, državnik, duhovščina, oblastnik, oznanjevalec, prosveta, razširjevanje, organizirati, piratje, vernost, omika, suženjstvo, veleum, prekucuh, književnost, izročilo, katoličanstvo, individualnost, krščanstvo, slovanstvo, dinastija, republika, luteranstvo, državljan, vodilo, monarhija, očak
denar + kapital	svota, kredit, glavnica, gotovina, hranilnica, banka, izposojevati, šteditи, posojevati, potrošiti, glda, odstotek, novec, prihranek, prišteditи, novčič, belič, obresti, kupnina, varčevati, izplačevati, najemščina, posojati, znesek, trošek, procent, obrestovati, potrositi, vsota, primanjkljaj, tisočak, forint, svotica, honorar, zaračuniti, skopariti, plačevanje, zvišati, borza, odškodnina, srečka, poboljšek, vknjižiti, obogateti, izposojati, doklada, posojilo, zavarovalnica, pristojbina, zapravljati
družina + otrok	deca, dec, dvojček, otročič, sinček, preživiti, graščakinja, pestunja, punčika, pestovati, oskrbnica, edinica, prvorojenka, bolehnost, olikati, porod, posestnica, izrejati, najdenka, družinica, odrasti, tilda, polbrat, tašča, stariš, primožiti, edinka, dojka, suknarija, pomožiti, rednica, umerti, sinaha, vnuček, ranar, razvajati, očim, tovarišija, jedinec, rejenka, deco, načelnica, malček, pomreti, rednik, sneha, ptujec, soseščina, pomočnica, naslednica

Tabela 2: Semantična polja, pridobljena z iskanjem najbližjih vektorskih reprezentacij besed v celotnem korpusu.

Vizualizacija odnosov med diskurzivnimi polji

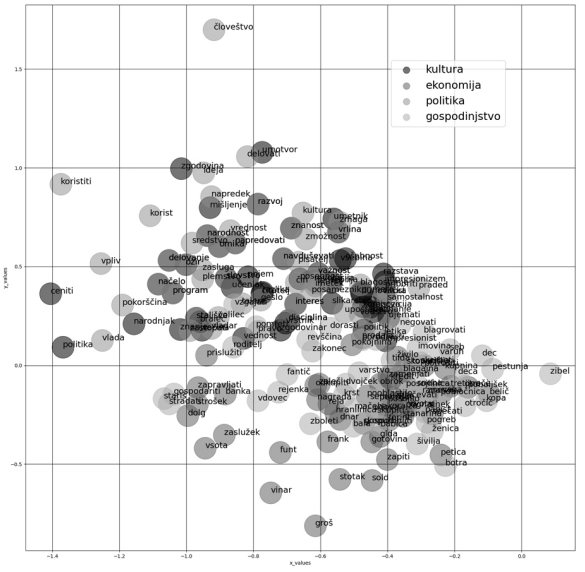
Vektorske predstavitve besed, ki so bile pridobljene z učenjem modela Word2Vec, obstajajo v stodimenzijskem vektorskem prostoru. Da bi bližino med vektorji lažje razumeli, moramo vektorski prostor zreducirati na tri ali celo dve dimenziji. Za jasnejšo predstavitev prekrivanja semantičnih polj je bila v tem primeru uporabljena dvodimenzionalna vizualizacija. Za redukcijo vektorskih dimenzij je bila uporabljena metoda glavnih komponent (ang. *principal component analysis* ali PCA), katere cilj je identifikacija ključnih komponent v naboru podatkov, ki v čim večji meri povzamejo variabilnost podatkov. S pomočjo PCA je tako mogoče zmanjšati dimenzionalnost podatkov in hkrati predstaviti informacije s čim manjšo izgubo natančnosti (Nelson 6).

Z redukcijo vseh 200 pridobljenih vektorskih predstavitev besed na dve dimenziji je bil ustvarjen diagram raztrosa, na katerem so razvidni odnosi med štirimi obravnavanimi semantičnimi polji (Slika 2). Ker so na diagramu predstavljeni le približki vektorskih koordinat, iz vizualizacije ne moremo interpretirati razdalj med posameznimi oznakami, še vedno pa lahko sklepamo, da so si gruče vektorjev, ki se prekrivajo, bolj podobne kakor tiste, ki so jasno ločljive. Z diagrama je razvidno, da so si vsa semantična polja blizu, pri čemer je gruča vektorjev, ki označujejo semantično polje ekonomije, najbližje gruči, ki predstavlja semantično polje gospodinjstva. To polje se dotika vektorskih gruč, ki predstavljata semantični polji kulture in politike in ki se prekrivata. Predvsem prekrivanje semantičnih polj politike in kulture nakazuje, da se v literarnih delih, ki so zajeta v korpus, temi kulture in politike pojavljata v skupnih kontekstih.

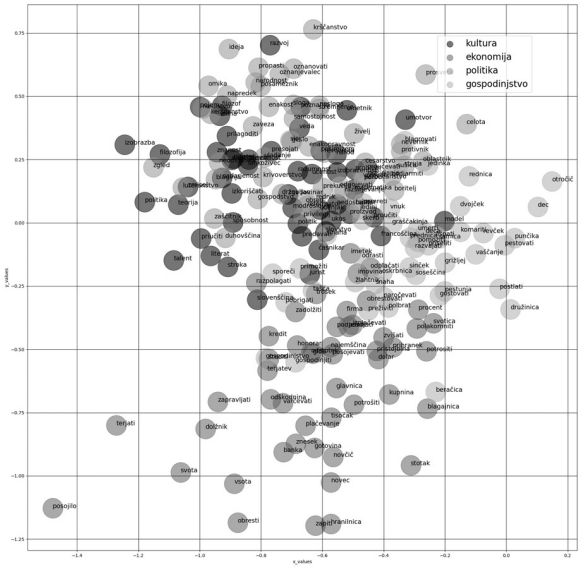


Graf 1: Vektorske vložitve besed na primeru jezikovnega modela, naučenega na celotnem korpusu.

Ker je bil namen analize preveriti, v kolikšni meri so odnosi med temi semantičnimi polji pogojeni s kanoničnostjo literarnih del, je bila analiza ponovljena še na podkorpusih del z visoko (Graf 2) in nizko (Graf 3) stopnjo kanoničnosti. Na grafih raztrosa je najočitnejša razlika v spremembi razmerja med gospodinjstvom in ekonomijo, pri čemer sta ti polji pri kanoničnih besedilih bolj prekrivni kakor pri nekanoničnih. Semantični polji kulture in nacionalne politike ostajata prekrivni v obeh podkorpusih, pri čemer je prekrivnost pri kanoničnih besedilih nekoliko očitnejša, saj je več besed v obeh semantičnih poljih identičnih (npr. »načelo«, »program«, »vrlina«, »znanost«, »navduševati«).



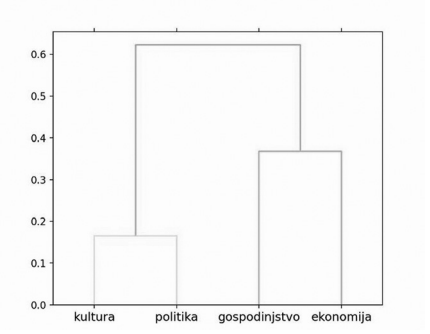
Graf 2: Vektorske vložitve besed na primeru jezikovnega modela, naučenega na podkorpusu besedil z visoko stopnjo kanoničnosti.



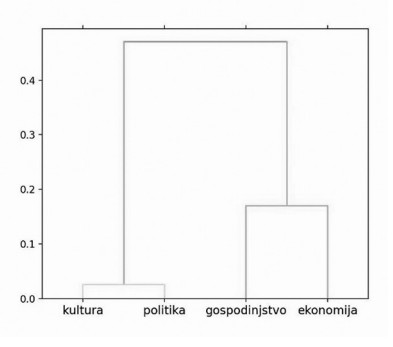
Graf 3: Vektorske vložitve besed na primeru jezikovnega modela, naučenega na podkorpusu besedil z nizko stopnjo kanoničnosti.

Evalvacija s pomočjo hierarhičnega gručenja

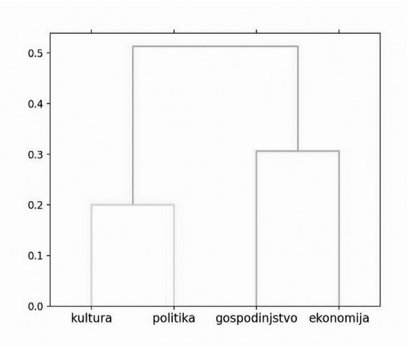
V kakšnih odnosih so posamezna semantična polja, lahko preverimo tudi s postopkom hierarhičnega gručenja (prim. Kljun, Teršek in Žitnik 8–9). Ker nas zanima razdalja med semantičnimi polji in ne posameznimi vektorji, so najprej izračunane povprečne vrednosti vseh vektorjev v posameznem semantičnem polju. Na ta način so pridobljeni štirje povprečni vektorji, ki imajo kakor izhodiščni vektorji 100 dimenzij. Razdalja med povprečnimi vektorji je izmerjena kot 1 simcos, kjer je simcos kosinusna podobnost med dvema povprečnima vektorjema. Za določanje razdalje med povprečnimi vektorji je bila izbrana Wardova metoda, ki je namenjena minimiziranju variance znotraj gruč. Rezultat so trije dendrogrami, na katerih je tako kakor na diagramih raztrosa razvidno, da sta si semantični polji kulture in politike bližji v primerjavi s semantičnima poljema ekonomije in gospodinjstva (Grafi 4–6). Večja ko je podobnost med gručami, nižje na osi y se stikata veji, ki povezujeta dve gruči.



Graf 4: Hierarhično gručenje vektorskih reprezentacij besed na primeru celotnega korpusa.



Graf 5: Hierarhično gručenje vektorskih reprezentacij besed na primeru podkorpusov besedil z visoko stopnjo kanoničnosti.



Graf 6: Hierarhično gručenje vektorskih reprezentacij besed na primeru podkorpusov besedil z nizko stopnjo kanoničnosti.

Kot gaže Graf 5, je za podkorpus besedil z visoko stopnjo kanoničnosti značilna bližina med semantičnima poljema kulture in politike, pa tudi med poljema ekonomije in gospodinjstva. Bližina semantičnih polj je (zlasti v primeru gospodinjstva in ekonomije) manj izrazita v korpusu del z nižjo stopnjo kanoničnosti.

Prešernovska struktura

Ker so rezultati analize pokazali najočitnejše prekrivanje semantičnih polj politike in kulture, se bo razlaga rezultatov podrobneje posvetila literarni tematizaciji teh dveh institucij v slovenskem prostoru v t. i. dolgem 19. stoletju. V zvezi s prekrivanjem semantičnih polj kulture in nacionalne politike se zdita relevantna dva koncepta slovenske preddigitalne literarne vede: Pirjevčev koncept prešernovske strukture iz leta 1969 in Ruplov koncept slovenskega kulturnega sindroma iz leta 1976. Oba koncepta temeljita na natančnem branju slovenskega literarnega kanona 19. stoletja. Pirjevec je s prešernovsko strukturo konceptualiziral delo in recepcijo Franceta Prešerna; koncept je (tam še pod imenom »Prešernova struktura«) vpeljal takole:

> [P]esnik je »voljan biti« pesnik in s tem tudi žrtev le, dokler mu je »zagotovljeno«, da je edini izvoljeni organ nebeške poezije in rajske lepote. Vse to mora biti »zagotovljeno«, sicer ostane narod ne le brez resnice, marveč tudi brez potrdil in brez mitologije. In vse dokler je narod blokirano gibanje, so takšna zagotovila tudi zares dana, iz česar hkrati sledi, da je Prešernova struktura zares plodna lahko le v določeni splošno zgodovinski strukturi. (Pirjevec 78)

Ruplov koncept slovenskega nacionalnega sindroma povzema to osnovno strukturo in jo s Prešerna razširi še na kanonične prozaiste 19. stoletja: Levstika, Stritarja, Jurčiča, Kersnika, Tavčarja in nazadnje Cankarja. Ruplova izhodiščna predpostavka je, da literatura »ni le preprost sektor družbene produkcije ali npr. družbene vrhnje stavbe, ampak skuša nadomeščati vse (oz. večino) funkcij, ki jih v razvitih družbah vršijo še (oz. predvsem) drugi sektorji vrhnje stavbe (pravno-političnі, izobraževalni, znanstveni … sektor)« (Rupel 424), njegova sklepna ugotovitev pa, da slovenska književnost 19. stoletja »nima tiste kvalitete, ki jo ima umetnost v razvitih družbah, kjer pretežno izraža oz. odraža družbene probleme in recimo osebne občutke oz. čustva«, saj po Ruplu »[l]iteratura pri Slovencih ni le odraz in izraz, temveč – ker nadomešča celo vrsto subsistemov – vtis, vzorec, naročilo, ukaz, ki ima določeno akcijsko, spodbujevalno funkcijo« (426–427).

Pirjevčevo fenomenološko študijo, ki je nastala kot del polemike o neoavantgardni poeziji poznih šestdesetih let, in Ruplovo kulturnosociološko analizo, ki si prizadeva za strožji družboslovni metodološki aparat, druži stališče, da slovenska literatura estetsko stagnira zaradi svoje vloge pri konstruiranju nacionalne identitete, in sicer naj bi bila literatura nadomestek za politično, deloma pa tudi gospodarsko uveljavljanje nedržavotvornega naroda, podrejenega tuji oblasti (Juvan 298).

Oba koncepta sta bila deležna številnih kritik, v zadnjih letih predvsem na račun dejstva, da slovenski književnosti že v 19. stoletju v resnici ni pripadal monopol pri konstruiranju nacionalne zavesti, da se estetske stagnacije ne da pripisati le narodotvorni vlogi literature, da tovrstna vloga literature ni izključno slovenski fenomen ter da se velik del trivialne literature s svojim pomenom za narodni preporod ni ukvarjal (Juvan 315–317; Dović 288–290).

Za razlago rezultatov zgornje analize je relevantno predvsem dejstvo, da omenjeni teoretiki večinoma izpeljujejo koncepta prešernovske strukture in slovenskega kulturnega sindroma predvsem iz Prešernove avtotematske poezije, kjer naj bi bila literatura kot sredstvo legitimacije nacionalnega gibanja tudi najjasneje tematizirana.[3]

Rupel sicer razširi nabor besedil tudi na daljšo prozo in publicistiko 19. stoletja, a je uspešnejši pri branju neliterarnih spisov, saj išče zgolj artikulacijo političnih nazorov, ne pa specifične literarne tematizacije nacionalne vloge literature, ki jo pri Prešernu iščejo ostali trije teoretiki.[4] To prekrivanje literarne in politične tematike, ki ga Pirjevec, Juvan in Dović odkrivajo predvsem pri Prešernu, kvantitativna semantična analiza odkriva v daljši prozi, v kakršni je Rupel iskal politično tematiko brez ozira na literarno. Rezultati analize napeljujejo na tezo, da gre pri tej semantični prekrivnosti za reprodukcijo širše uveljavljenega diskurza, ki polji približuje drugo k drugemu in ki se ne pojavlja le v besedilih, ki eksplicitno tematizirajo sočasno politično (ali literarno) dogajanje (kakor na primer Kersnikova romana *Ciklamen* in *Agitator*, ki jima Rupel namenja veliko pozornosti), temveč posredno učinkuje v celotnem korpusu.

Tako Pirjevec, Juvan in Dović kakor Rupel prešernovsko strukturo oziroma slovenski kulturni sindrom obravnavajo izključno v kontekstu

[3] Po Pirjevcu in Ruplu z natančnim branjem Prešerna tudi Marko Juvan pride do zaključka, da je »nastavke za oblikovanje podobe in vloge 'nacionalnega pesnika', ki vsekakor sodi v območje prešernovske strukture in slovenskega kulturnega sindroma, […] oblikoval že Prešeren sam v svoji avtotematski poeziji« (Juvan 313).

[4] Eksplicitno izražanje političnega stališča pri analizi kanoničnih besedil obravnavanih avtorjev najde pri Prešernu, Tavčarju, Kersniku in Cankarju, medtem ko pri Levstiku, Jurčiču in Stritarju odkriva le implicitno političnost (Rupel 431).

slovenskega literarnega kanona. Pirjevec ugotavlja, da je literarni prostor, v katerem je nastajala (kanonična) književnost, »reprodukcija temeljnih razsežnost naroda« in pripada tistemu delu ljudstva, ki je bilo aktivno vpleteno v oblikovanje slovenske nacionalne politike, torej narodnozavednemu meščanstvu (Pirjevec 77). Obstajal pa naj bi tudi del ljudstva, na katerega nacionalna politika ni vplivala, v skladu s tem pa tudi vzporeden sistem »literature za ljudstvo«. Drugi pol slovenskega literarnega prostora naj bi bila po Pirjevcu torej literatura, ki se s svojo nacionalno vlogo ni obremenjevala in je ostala tudi zunaj nacionalnega kanona. Poznejše kritike prešernovske strukture in slovenskega kulturnega sindroma te Pirjevčeve teze niso problematizirale; prej jim je njegova redukcija prešernovske strukture na problem, ki zadeva le literarni kanon, služila kot argument za njeno nedoslednost (Juvan 316). Kvantitativni pristop nam omogoča, da poleg kanoničnih avtorjev ti konceptualizaciji slovenskega literarnega kanona apliciramo tudi na ostalo slovensko literaturo 19. stoletja. V nasprotju z uveljavljenim prepričanjem je primerjava med kanoničnimi in nekanoničnimi avtorji namreč pokazala, da semantična bližina literature in nacionalne politike za nekanoničen del korpusa ni nič manj značilna kot za kanoničnega.

LITERATURA

Brottrager Judith, idr. »Modeling and Predicting Literary Reception. A Data-Rich Approach to Literary Historical Reception«. *Journal of Computational Literary Studies*, let. 1, št. 1, 2022, https://jcls.io/article/id/95/. Dostop 12. 4. 2024.

Dović, Marijan. *Prešeren po Prešernu: kanonizacija nacionalnega pesnika in kulturnega svetnika*. LUD Literatura, 2017.

Eder, Maciej, in Artjoms Šeļa. »One Word to Rule Them All: Understanding Word Embeddings for Authorship Attribution«. *Digital Humanities 2022 Combined Abstracts*, Univerza v Tokiu, 2022, str. 199–202, https://dh2022.adho.org/. Dostop 12. 4. 2024.

Erjavec, Tomaž idr. *Slovenian Novel Corpus (ELTeC-slv): April 2021 release (v2.0.0)*. Zenodo, 2021, https://doi.org/10.5281/zenodo.4662600. Dostop 12. 4. 2024.

Hatzel, Hans Ole, idr. »Machine Learning in Computational Literary Studies«. *it – Information Technology*, let. 65, št. 4–5, 2023, str. 200–217.

Herrmann, J. Berenike, Joanna Byszuk in Giulia Grisot. »Using Word Embeddings for Validation and Enhancement of Spatial Entity Lists«. *Digital Humanities 2022 Combined Abstracts*, Univerza v Tokiu, 2022, str. 239–241.

Juvan, Marko. *Prešernovska struktura in slovenski kulturni sindrom*. LUD Literatura, 2012.

Kljun, Maša, Matija Teršek, in Slavko Žitnik. »Pomenska analiza kategorij sovražnega govora v obstoječih označenih korpusih«. *Uporabna informatika*, let. 30, št. 1, 2021, str. 3–18.

Ljubešić, Nikola, in Kaja Dobrovoljc. »What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian«. *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, ur. Tomaž Erjavec idr., Association for Computational Linguistics, 2019, str. 29–34, https://aclanthology.org/W19-3704. Dostop 12. 4. 2024.

Mandić, Lucija, in Tomaž Erjavec. *Corpus of Longer Narrative Slovenian Prose KDSP 1.0*. ZRC SAZU, 2023, http://hdl.handle.net/11356/1823. Dostop 12. 4. 2024.

Mikolov, Tomáš, idr. »*Efficient Estimation of Word Representations in Vector Space«. International Conference on Learning Representations*, 2013, https://arxiv.org/abs/1301.3781. Dostop 12. 4. 2024.

Nelson, Laura K. »Leveraging the Alignment between Machine Learning and Intersectionality: Using Word Embeddings to Measure Intersectional Experiences of the Nineteenth Century U.S. South«. *Poetics,* št. 88, https://doi.org/10.1016/j.poetic.2021.101539. Dostop 12. 4. 2024.

Pirjevec, Dušan. *Vprašanje o poeziji. Vprašanje naroda*. Obzorja, 1978.

Pollak, Senja, Matej Martinc, in Katja Mihurko. »Natural Language Processing for Literary Text Analysis: Word-Embeddings-Based Analysis of Zofka Kveder's Work«. *Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020)*, ur. Maria José Finatto idr., CEUR-WS, Aachen, 2020, http://ceur-ws.org/Vol-2607/paper4.pdf. Dostop 12. 4. 2024.

Rupel, Dimitrij. *Svobodne besede od Prešerna do Cankarja*. Lipa, 1976.

Schneider, Felix idr. »Data-Driven Detection of General Chiasmi Using Lexical and Semantic Features«. *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, ur. Stefania Degaetano-Ortlieb idr., Association for Computational Linguistics, 2021, str. 96–100, doi.org/10.18653/v1/2021.latechclfl-1.11. Dostop 12. 4. 2024.

A Quantitative Analysis of Relations between Semantic Fields in the Slovenian Narrative Prose of the Long Nineteenth Century

Keywords: digital literary studies / Slovenian narrative prose / Prešernian structure / Slovenian cultural syndrome / semantical analysis / word embeddings

The article analyzes the relationships between semantic fields in Slovenian narrative prose of the long nineteenth century using the method of word embeddings. The corpus of longer Slovenian narrative prose (KDSP 1.0) was analyzed using the Word2Vec technology in the Python programming language. For the purposes of the analysis, semantic fields were constructed for four social institutions: economy, politics, culture, and the household. A set of words for each semantic field was obtained by identifying the 50 words with the greatest cosine proximity to the vector representation of each institution. The set of vectors obtained in this way became the quantitative basis of an investigation into the relations between these social institutions as they are narrated by the literary texts included in the corpus. The findings reveal a significant overlap between the semantic fields of politics and culture, thus offering a quantitative approach to a phenomenon that traditional literary scholarship tends to conceptualize as the Prešernian structure or the Slovenian cultural syndrome.

1.01 Izvirni znanstveni članek / Original scientific article
UDK 821.163.6.09:004
DOI: https://doi.org/10.3986/pkn.v47.i2.08