RAČUNALNIŠKA FOLKLORISTIKA

Semantična analiza in vizualizacija tematske porazdelitve pesemskih tipov

Izvirni znanstveni članek | 1.01

Izvleček: Članek predstavlja raziskavo latentne semantične strukture slovenskih ljudskih pesmi z metodami strojne (računalniške) analize naravnega jezika. Namen raziskave je ugotoviti primernost strojnih metod za odkrivanje splošnih vzorcev in razmerij na ravni pesemskih tipov in zvrsti ter podati osnovno specifikacijo postopkov, primernih za strojno analizo folkloristič nih vsebin. Rezultati analize kažejo, da lahko z izbrano strojno metodo generiramo večdimenzionalen semantični prostor, ki na podlagi tematske porazdelitve in mer podobnosti omogoča globljo tipološko analizo folkloristič nih vsebin.

Ključne besede: strojne metode, analiza naravnega jezika, LDA (Latentna Dirichletova razporeditev), računalniška semantika, ljudska pesem, folkloristika, tipologija

Abstract: The article presents research into the latent semantic structure of Slovenian folk songs using natural language processing (NLP) methods. The aim is to determine the appropriateness of NLP for discovering general patterns and relationships on the level of song types and genres, and to specify the basic procedure for the computational analysis of folkloristic materials. The results of the analysis show that the appropriate computational method can generate multidimensional semantic space on the basis of the distribution of topics and similarity measures, and therefore enable a more nuanced typological analysis of folkloristic materials.

Keywords: computational semantics, natural language processing (NLP), LDA (Latent Dirichlet Allocation), folk song, folkloristics, typology

Uvod

Razpoložljivost in eksponentna rast digitalnih virov je tudi v humanistiki odprla možnosti novih interdisciplinarnih raziskav, ki temeljijo na uporabi strojnih (računalniških) metod in jih poznamo pod skupnim imenom *digitalna humanistika* (Busa 1992; Berry 2011). Razvoj metod strojnega učenja (angl. *machine learning*) sega v 50. leta prejšnjega stoletja, ko so se z razvojem umetne inteligence začeli porajati izzivi uporabnosti računalniških sistemov v vsakdanjem življenju ter vprašanja, v kolikšni meri lahko umetna inteligenca pripomore k tehnološkemu razvoju in dopolni ali celo nadomesti vlogo človeka. Z vidika umetne inteligence je tako eno izmed temeljnih vprašanj, kako strojno avtomatizirati neki segment človeške kognicije in ga v nekaterih primerih celo preseči. Čeprav so bili ti sistemi sprva precej okorni in je koncept osebnega računalnika zaživel šele v poznih 70. letih prejšnjega stoletja, je današnje življenje prežeto z računalniško tehnologijo in digitalnimi viri. Ker ti eksponentno rastejo, se zanašamo na strojne metode za obdelavo in analizo podatkov in njihovo sposobnost, da informacije predstavijo v človeku razumljivi obliki. Slednje ni trivialen problem. Dihotomija med človekovim razumevanjem pomena in računalniško sposobnostjo njegove ekstrakcije iz binarnega zapisa namreč ustvarja semantično vrzel. Ekstrakcija latentnih semantičnih struktur (tj. skritih pomenskih struktur) je zato za strojne metode velik izziv in glavno vprašanje je, kako to vrzel premostiti? Na uspešnost metod strojnega učenja močno vpliva zasnova algoritma in reprezentacijskih struktur, ki povzemajo strukturo določenega področja, v kar pa je nujno potrebno vključiti človeški vidik kakor tudi posebnosti predmeta analize. Posledično strojne metode niso same sebi namen, marveč orodje v rokah raziskovalca.

Ena izmed prednosti metod strojnega učenja je, da raziskovalcu omogočajo inovativne pristope k analizi, predstavitvi in vizualizaciji latentnih semantičnih struktur, ne le v besedilih, marveč tudi v glasbi in vizualnih zapisih, in to v obsegu, ki ga klasična ročna analiza ne omogoča (npr. Shiffrin in Bömer 2004; Michel 2011; Greenfield 2013). V nadaljevanju so predstavljene strojne metode analize naravnega jezika oz. angl. *Natural Language Processing* (v nadaljnjem besedilu NLP). Te med drugim omogočajo strojno prevajanje, optično prepoznavanje znakov (OCR), avtomatsko segmentacijo (npr. slovnično analizo), prepoznavanje govora in naposled tudi semantično analizo (npr. analizo pomena besed, odkrivanje tematske strukture korpusa besedil, emotivnih elementov itn.), ki jo tu predstavljamo.

Prispevek nadaljuje raziskave semantične analize folkloristič nih vsebin (Strle in Marolt 2014). Tehnični vidiki priprave korpusa, primerjava metod in postopki strojne analize so izčrpno opisani v omenjenem prispevku, zato v nadaljevanju sledi le kratek oris. Glavni namen prispevka je ugotoviti primernost izbrane strojne metode za tematsko analizo in klasifikacijo latentnih semantičnih struktur slovenske ljudske pesmi na ravni pesemskih tipov in zvrsti. Pri tem izhajamo iz metodoloških vprašanj računalniške semantike in strojnega učenja, in ne folkloristike. To odločitev sta deloma narekovala sam obseg in narava korpusa, saj na številčno in tematsko močno omejenem obsegu pesemskih primerov težko pridemo do konkretnih sklepov. Po drugi strani so računalniške metode v folkloristiki razmeroma nove in in tudi zaradi pomanjkanja primerljivih raziskav (Abello idr. 2012) je za izhodišče primerna splošnejša predstavitev.

Strojna analiza naravnega jezika

Tradicionalno NLP predstavljajo statistične metode, kakršna je npr. latentna semantična analiza ali LSA (gl. Landauer in Dumais 1997; Landauer idr. 2007). Te temeljijo na vektorskih izračunih v visokodimenzionalnem semantičnem prostoru. Osnovna zamisel je, da semantične prostore pridobimo iz besedne statistike, pri čemer besedilni korpus obravnavamo kot »vrečo

* Dr. Gregor Strle, univ. dipl. filozof, asistent z doktoratom, Glasbenonarodopisni inštitut ZRC SAZU. 1000 Ljubljana, Novi trg 2, gregor.strle@zrc-sazu.si; doc. dr. Matija Marolt, univ. dipl. inž. rač. in inf., docent, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani. 1000 Ljubljana, Tržaška cesta 25, matija.marolt@fri.uni-lj.si.

besed«. Z izračunom razdalje med vektorji besed nato izluščimo »pomene« posamičnih besed, ki so opredeljeni z relativno bližino drugih besed v semantičnem prostoru. Vendar LSA ne uporablja samo informacije o tem, kako pogosto se beseda1 in beseda2 pojavita skupaj, marveč tudi, kako pogosto se pojavita z vsemi drugimi besedami korpusa. Na podlagi analize celotnega vzorca sopojavljanja besed nato LSA generira skupni visokodimenzionalni semantični prostor. Sledi dekompozicija prostora na manjše število dimenzij z matematičnim postopkom (Martin in Berry 2007), ki iz originalne matrike besed in dokumentov izlušči najpomembnejše dimenzije. Ta preprost princip semantične analize na večjih korpusih omogoča zavidljivo učinkovitost (gl. npr. Landauer in Dumais 1997).

Zdi se, da za neki zelo splošen uvid v semantično strukturo zadošča statističen izračun asociacijskih povezav med besedami, ki zanemari vse dodatne informacije (npr. zaporedje besed ali tematsko struktura korpusa), in kompleksnejša zasnova algoritma ni potrebna. Vendar imamo pogosto opraviti z manj obsežnimi korpusi, ki so pogosto dodatno omejeni s posebnostmi nekega področja. V takšnih primerih se hitro pokažejo pomanjkljivosti preprostega asociacijskega pristopa. Te so še posebej očitne pri analizi folklorističnega gradiva, ki je poleg lokalnih in regionalnih značilnosti pogosto zastopano le z nekaj deset variantami posamičnega pesemskega tipa, kar se je pokazalo tudi v konkretnem primeru.

Nedavna primerjalna raziskava metod NLP (Strle in Marolt 2014) je pokazala na nekaj bistvenih pomanjkljivosti tradicionalnega statističnega pristopa pri analizi folklorističnih vsebin. Raziskava je bila izvedena na korpusu slovenskih ljudskih pripovednih pesmi (podvrsti ljubezenskih in družinskih pripovednih pesmi), ki je poleg narečnosti in drugih jezikovnih posebnosti tudi močno medbesedilno obarvan. Glavno raziskovalno vprašanje je bilo, v kolikšni meri lahko izbrana strojna metoda omogoči uvid v semantično strukturo korpusa in zajame njegove splošne značilnosti. Ker semantični prostor, generiran z LSA, temelji zgolj na izračunu asociacij med besedami, posledično ne zazna latentnih tem, ki se prepletajo v posameznih pesemskih tipih, samo analizo pa dodatno oteži močna medbesedilnost. To se pokaže v semantični reprezentaciji razmeroma majhnega števila pesemskih tipov in tem, ki prevladujejo v celotni porazdelitvi. V takem semantičnem prostoru prevladujejo pesemski tipi, ki so v korpusu najštevilčnejši, ne glede na tematsko raznovrstnost variant manj zastopanih tipov. Ob nezmožnosti zaznavanja, generaliziranja in porazdelitve prek latentnih tem LSA posledično spregleda pomemben del semantičnih informacij in tematske strukture korpusa. Ugotovitve omenjene raziskave ustrezajo rezultatom predhodnih raziskav (Blei idr. 2003; Steyvers in Griffiths 2007), ki so za premostitev omenjenih težav namesto asociacijskega modela (LSA) predložile verjetnostni pristop k semantični analizi besedil na podlagi generativnega verjetnostnega tematskega modela. Ta model je bil uporabljen tudi v raziskavi tematske porazdelitve pesemskih tipov in zvrsti, ki jo predstavljamo v nadaljevanju.

Semantična analiza in vizualizacija tematske porazdelitve pesemskih tipov v pesemskih zvrsteh

Osnovno izhodišče predstavljene raziskave je ugotoviti primernost strojnega pristopa za vsebinsko oz. tematsko analizo slovenske ljudske pesmi. Na korpusu slovenskih ljudskih pripovednih pesmi smo želeli oceniti, ali semantičen prostor in porazdelitev tem pridobljenih z verjetnostnim tematskim modelom na kakršenkoli način ustrezata trenutni razvrstitvi variant posameznih pesemskih tipov v eno izmed družin pripovednih pesmi.[1] Čeprav so teme obeh pesemskih družin zelo sorodne in se nenehno prepletajo, smo s strojno analizo skušali zaznati specifične pomenske razsežnosti, na podlagi katerih strojna metoda razločuje med pesmimi o ljubezenskih usodah in sporih in tistih o družinskih usodah in sporih. V ta namen smo uporabili metodo hierarhičnega gručenja, ki pesemske tipe hierarhično strukturira po skupinah. Dodatni uvid v analizo nam omogoča semantični prostor, ki z verjetnostno porazdelitvijo tem na podlagi mer podobnosti odkriva kompleksnost razmerij in prehajanj semantičnih elementov med pesemskimi tipi.

Metoda: Latenta Dirichletova razporeditev (LDA)

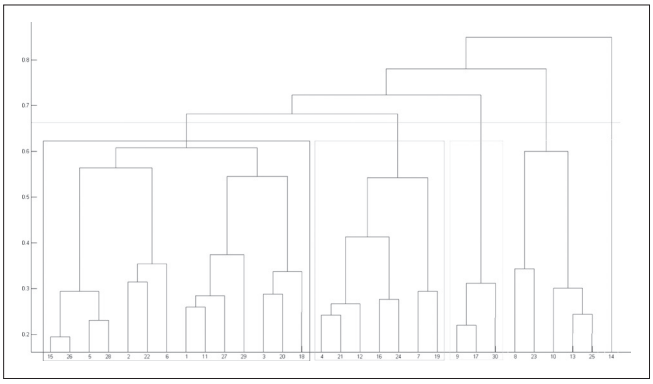
Latentna Dirichletova razporeditev (angl. *Latent Dirichlet allocation* oz. LDA; Blei idr. 2003) je generativni verjetnostni tematski model, ki za razloček od tradicionalnih statističnih metod temelji na predpostavki, da *dokumenti* poleg besed vsebujejo številne *teme*.

V analizi LDA (za izčrpno matematično predstavitev modela gl. prav tam) so podatki del generativnega procesa, ki določa *skupno verjetnostno porazdelitev* nad opazovanimi in skritimi naključnimi spremenljivkami, pri čemer so znane spremenljivke besede in vnaprej določeno število tem, skrita pa je tematska struktura dokumentov oz. korpusa. S tem LDA posnema vsebinsko strukturo korpusa, v katerem verjetnostna tematska porazdelitev dokumentov in besed tvori neko osnovno pomensko hierarhijo. V primerjavi s klasičnimi statističnimi metodami verjetnostni tematski modeli na splošno generirajo bolj interpretativne in posledično bolj uravnotežene pomenske prostore, in to prav zaradi njihove sposobnosti zaznave tematske porazdelitve (gl. Blei idr. 2003; Griffiths idr. 2007: Steyvers and Griffiths 2007; Strle in Marolt 2014).

Korpus

Pri raziskavi so bili uporabljeni isti korpus in postopki priprave besedil kakor v prvotni raziskavi, z izjemo dveh variantnih tipov hrvaških pripovednih pesmi o družinskih usodah in sporih. Korpus besedil zbirke *Slovenske ljudske pesmi* (SLP 4; SLP 5) tako obsega 1965 variant ljudskih pripovednih pesmi, od tega 36 pesemskih tipov pripovednih pesmi o ljubezenskih usodah in sporih s 1073 variantami in 52 (prvotno 54) pesemskih tipov pripovednih pesmi o družinskih usodah in sporih z 857 variantami. Gre za tematsko zelo sorodno gradivo, s prevladujočimi temami *smrt*, *uboj*, *umor*, *samomor*, *nezvestoba*, *kazen*, *nesreča* ipd., z močno izraženimi sorodstvenimi in družinskimi vezmi (npr. hči-mati-mačeha), opazni pa sta tudi močna medbesedilnost in za ljudsko pesem značilno potovanje verzov, kitic in tematskih vzorcev oz. motivov iz pesmi v pesem (gl. npr. Kumer 1996; Golež Kaučič 2003).

1 Da bi problematiko predstavili širšemu bralskemu krogu, smo tipologijo GNI ZRC SAZU za opis posamičnih ravni terminološko poenostavili. Tako npr. za opis strukture ‘pripovedne/družinske/224. Nezvestoba iz pohlepa/2. varianta’ namesto izrazov vrsta/podvrsta/tip/varianta uporabljamo zvrst/družina/pesemski tip/varianta.

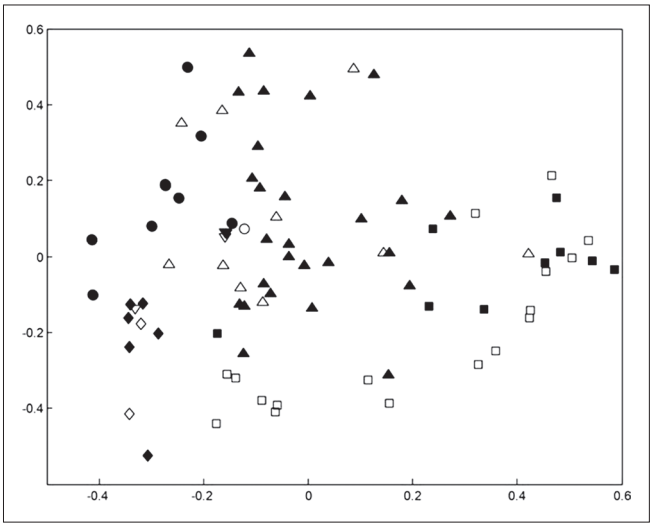


Slika 1: Dendrogram gručenja pesemskih tipov ljubezenskih in družinskih pripovednih pesmi.
Vir: lastni prikaz.

Zaradi morfološke raznovrstnosti, močne naročnosti in sintetične narave slovenskega jezika je bilo treba predhodno izvesti lematizacijo korpusa: z orodjem Obeliks (Grčar idr. 2012) smo naročne glasove zamenjali s slovničnimi ekvivalenti in besede pretvorili v osnovno obliko. S tem je bil korpus pripravljen za strojno analizo.

Analiza in vizualizacija pesemskih tipov ljubezenskih in družinskih pripovednih pesmi

Analiza in vizualizacija pesemskih tipov je bila opravljena z orodji Stanford TMT in Matlab TMT knjižnico za LDA. Vhodni podatki za analizo so vektorji, ki opisujejo pomembnost besed v pesmih, pomembnost pa je predstavljena z mero tf-idf (*term frequency-inverse document frequency*). To je standardna mera za uteževanje besed pri analizi tekstovnih zbirk, kjer je pomembnost besede predstavljena kot logaritem pogostnosti pojavljanja besede v pesmi, popravljena s količnikom pomembnosti besede. Slednji v obratnem razmerju upošteva število pesmi, ki vsebujejo besedo – več pesmi vsebuje besedo, manjši je količnik. Mera tf-idf torej pripiše večjo pomembnost besedam, ki se velikokrat pojavijo v posamični pesmi, hkrati pa malo v drugih pesmih.
Postopek tematske analize (izračun LDA) je optimizacija, ki je inicializirana naključno in ji kot parameter podamo število tem, ki jih želimo modelirati. Posledično večkratni izračuni prinesejo nekoliko drugačne rezultate, vendar relativna razmerja med temami in notranja struktura (npr. asociacije med besedami v temi) posameznih področij v semantičnem prostoru ostajajo večinoma enake. Mere podobnosti so sorazmerne s tematsko sorodnostjo vsebin korpusa, izbor števila tem pa vpliva na podrobnost njihove reprezentacije: manjši zbir tem pokaže splošnejšo tematsko porazdelitev, v kateri izstopajo predvsem tipične značilnosti korpusa, z večjim številom tem pa dobimo večjo segmentacijo in podrobnejši uvid v tematsko porazdelitev in razmerja med temami.
V nadaljevanju sta predstavljeni analiza in vizualizacija rezultatov. Za potrebe hierarhičnega gručenja in prikaza porazdelitve pesemskih tipov čez celoten korpus smo analizo omejili na pet gruč (Sliki 1 in 2, Tabela 1), za vizualizacijo posamezne podvrste pa na štiri gruče (Tabela 2, Slika 3). Potrebna sta bila dva različna izračuna (eden za analizo hierarhičnega gručenja in celostne porazdelitve, drugi za analizo tematske porazdelitve v posameznih podvrsti), zato rezultati obeh izračunov niso neposredno primerljivi.

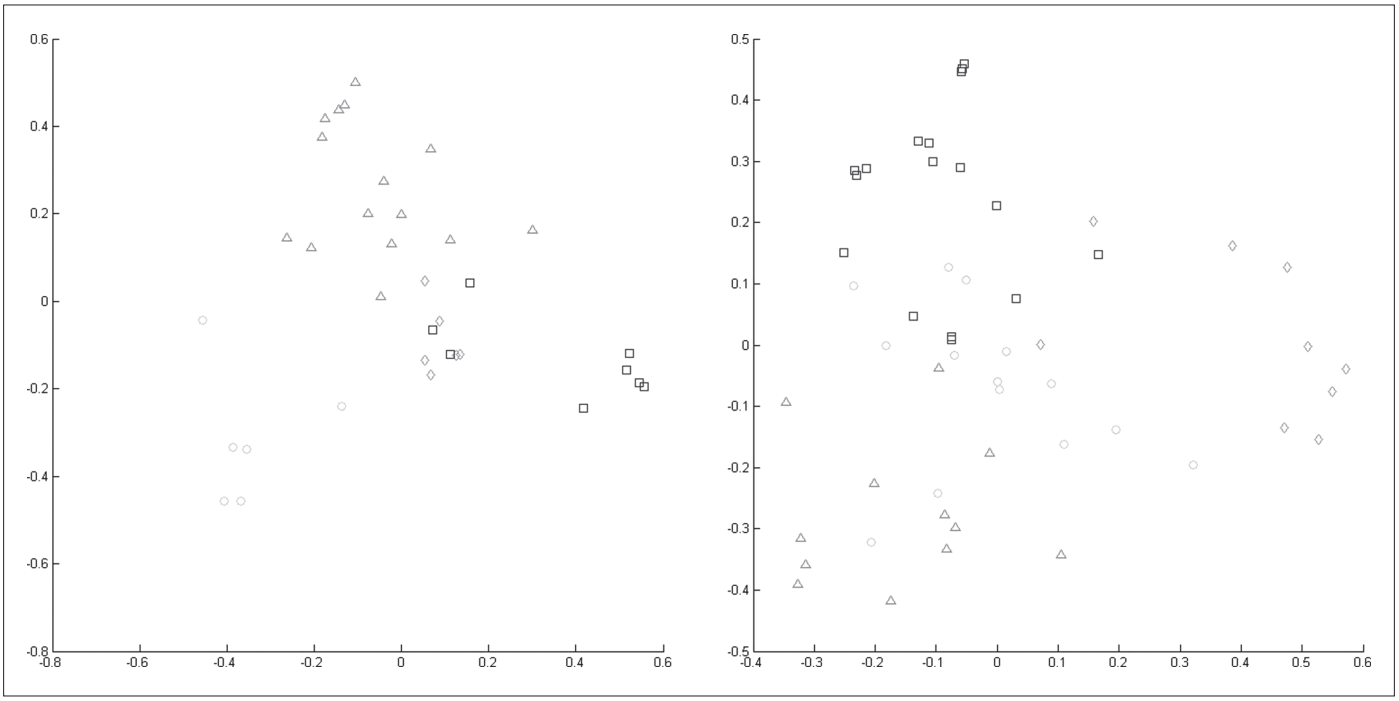


Slika 2: Vizualizacija porazdelitve rezultatov hierarhičnega gručenja. Skupine so ponazorjene z različnimi liki, ljubezenski tipi so označeni s *praznimi* liki, družinski z *zapolnjenimi.*
Vir: lastni prikaz.

Hierarhično gručenje pesemskih tipov – ljubezenske ali družinske?

Da bi zaznali specifične semantične dimenzije, na podlagi katerih metoda LDA razločuje med pesmimi o ljubezenskih usodah in sporih in tistih o družinskih usodah in sporih, smo sprva izračunali povprečen tematski vektor za vsak pesemski tip obeh družin s povprečenjem tematskih vektorjev variant, ki sodijo k tipu. Namen tovrstnega povprečenja je predvsem zmanjšanje nesorazmerja med številom variant pesemskih tipov, saj se to giblje med več kot 50 variant za nekatere pesemske tipe ali le do dveh ali treh variant za druge. S povprečenjem je vsak pesemski tip predstavljen z enim (povprečnim) tematskim vektorjem. Skupaj smo dobili 88 tematskih vektorjev za vseh 88 pesemskih tipov v analizirani zbirki.
Z metodo hierarhičnega gručenja smo nato zbir pesemskih tipov razdelili v gruče podobnih tipov, pri čemer smo podobnost pesemskih tipov izračunali kot kosinusno podobnost med tematskimi vektorji tipov. Kosinusna podobnost je mera, ki se pogosto uporablja v analizi besedilnih zbirk. Hierarhično delitev na gruče prikazuje Slika 1, razmerje ljubezenskih in družinskih tipov za posamično gručo je predstavljeno v Tabeli 1, njihova porazdelitev po meri podobnosti pa na Sliki 2.
Dendrogram (Slika 1) prikazuje hierarhično ureditev pesemskih tipov v gručah kakor tudi razmerja med gručami.[2] Veje dendrograma za vseh pet gruč sestavlja 30 podskupin, v katere je urejenih vseh 88 pesemskih tipov, od tega 36 ljubezenskih in 52 družinskih tipov pripovednih pesmi. Dolžina veje predstavlja sorodnost – v konkretnem primeru je npr. gruča 5 s podskupino 14 najbolj oddaljena od drugih gruč.

2 Barve posameznih gruč povezujejo prikaz gručenja na Sliki 1 z razporeditvijo tipov v Tabeli 1 in porazdelitvijo na Sliki 2.



Slika 3: Porazdelitev gruč tipov ljubezenskih (SLP 4; levo) in družinskih (SLP 5; desno) pripovednih pesmi.
Vir: lastni prikaz.

Številčna razmerja med ljubezenskimi in družinskimi tipi posameznih gruč dendrograma so prikazana v Tabeli 1 (tipi družinskih pesmi so sivo obarvani) in kažejo na približno 60 % prevlado tipov družinskih pripovednih pesmi (gruče 1, 3 in 4), kar je pričakovano glede na primerljivo številčno prevlado teh pesemskih tipov v samem korpusu. Na drugi strani prevlada ljubezenskih pesmi v gručah 2 in 5 kaže na to, da semantične dimenzije, generirane z LDA, uspešno zajamejo splošne značilnosti pesemskih družin, kar nadalje omogoča njihovo razločitev v gruče s prevladajočim številom ljubezenskih ali družinskih pripovednih pesmi (gl. Tabelo 1).

Slika 2 prikazuje gručenje povprečnih tematskih vektorjev variant vseh 88 pesemskih tipov v dvodimenzionalnem prostoru in ustreza zbiru tipov iz Tabele 1. Razmerja med tipi (ljubezenski tipi so označeni s *praznimi* liki, družinski z *zapolnjenimi*) temeljijo na merah podobnosti – tematsko sorodni tipi so si v prostoru blizu, nesorodni pa oddaljeni. Iz tega izhajajo tudi razmerja med gručami in razpršenost posamezne gruče. Obe največji gruči sta precej razpršeni, kar kaže na široko tematsko zastopanost predstavnikov obeh gruč v večjem delu semantičnega prostora kakor tudi na večje število tematskih podskupin, od katerih so nekatere po sorodnosti bližje podskupinam ali tipom drugih gruč. Kljub temu lahko gručo s prevladujočo družinsko tematiko (trikotniki) razmejimo od gruče, kjer prevladujejo ljubezenske pesmi (kvadrati). Navidezno prekrivanje gruč je deloma tudi posledica redukcije na dve dimenziji.

Porazdelitev pesemskih tipov v posamezni družini

Z gručenjem tipov za posamezno pesemsko družino smo želeli preučiti zgradbo pomenske predstavitve prostora ljubezenskih in družinskih pesmi in analizirati najznačilnejše predstavnike. Za gručenje smo uporabili metodo kmeans (Hartigan 1975), s katero smo prostor tematskih vektorjev razdelili na štiri skupine. Kmeans je standardna metoda za gručenje, ki razdeli tematske vektorje na skupine, vsak vektor pripada skupini z najbližjim povprečnim vektorjem. V Tabeli 2 so najvplivnejše besede štirih gruč ljubezenskih in družinskih pripovednih pesmi[3] in ustrezajo razdelitvi pesemskih tipov/gruč v Tabeli 3, Slika 3 pa prikazuje njihovo porazdelitev v semantičnem prostoru posamezne družine. Iz Tabele 2 je v prvi vrsti razvidno izrazito sopojavljanje besed, tako v posamezni družini (glej A) kakor tudi v celotnem korpusu (B). To kaže na sorodno tematiko in močno medbsedilnost ljubezenskih in družinskih pripovednih pesmi, na kar smo opozorili že v uvodu. Tako imajo variante, ki npr. tematizirajo neki motiv, z drugimi variantami podoben besedni sestav, ne glede na to, ali gre za ljubezenske ali družinske pripovedne pesmi (Slika 2). Posledično se v sami tematski analizi LDA posamezni pesemski tipi pojavljajo v družini, ki ji po tipologiji SLP ne pripadajo (npr. nekateri tipi ljubezenskih pesmi se razporejajo v družinske pesmi in nasprotno). Najnazorneje je to prikazano pri hierarhičnem gručenju pesemskih tipov v prejšnjem poglavju (primerjaj Sliko 1, Tabelo 1 in Sliko 2) in dodatno potrjeno s prikazom sopojavljanja besed v Tabeli 2. Kljub temu je iz gruč vseeno mogoče razločiti prevladujoče teme, občasno sklepati na konkretne pesemske tipe ali zaznati posamične motive.

Zgled: na podlagi besed ‘lovec’, ‘puška’, ‘gozd’, ‘listje’ (gl. Tabela 2/SLP 4/gruča 1) lahko sklepamo na pesemski tip ‘236. LOVEC USTRELI LJUBICO IN SEBE’, medtem ko ‘mačeha’, ‘mati’, ‘dete’ (gl. Tabela 2/SLP 5/gruča 2 in Tabelo 3) kažejo

3 Besede smo izbrali iz zbira prvih 40 najvplivnejših besed za posamezno gručo (tj. besed, ki so najbolj vplivale na tematsko porazdelitev LDA) in odstranili tiste, npr. veznike, zaimke, številnike, členke in večino glagolov (npr. hoteti, govoriti, stati, imeti), ki bistveno ne prispevajo k pomenu.

G1: 15, 26, 5, 28, 2, 22, 6, 1, 11, 27, 29, 3, 20, 18 (11:26)		**G3: 9, 17, 30 (2:6)**
15: 241. SMRT NEVESTE PRED POROKO/C	235. SMRT OB SNIDENJU	9: 210. DEKLE ZASTRUPI LJUBEGA
26: 264. PREŠUŠTNIK IN LJUBICA KAZNOVANA	247. SMRT DALEČ OMOŽENE	253. TREH HČERA NAGLA SMRT
5: 206. ZAPELJANI PUŠČAVNIK	254. ŽALOSTNA USODA TREH SINOV	266. NEZVESTA GOSPA S TREMI STRAŽARJI/A
257. MAČEHA IN SIROTA B (SIROTA JERICA)	272. SESTRA ZASTRUPI SESTRO	267. NEZVESTA GOSPA S TREMI STRAŽARJI/B
258. MAČEHA IN SIROTA C (SVETA KRISTINA)	11: 238. ODKLONITEV POROKE	276. SIN NA TASTOVO POBUDO UMORI MATER
282. ŽENA NOČE Z MOŽEM NA POT	280. BRAT IN SESTRA UBEŽITA IZ UJETNIŠTVA	17: 242. TAŠČI SE IZJALOVI UMOR SNAHE
28: 281. ŽENA REŠI MOŽA IZ UJETNIŠTVA	27: 273. BRAT UMORI SESTRO	30: 230. BOLNA LJUBICA UMRE
2: 203. PRIDOBITEV LJUBEGA Z NAVIDEZNO SMRTJO/B	287. OBSOJENA DETOMORILKA	268. NEZVESTA ŽENA POMAGA LJUBIMCU ...
212. LAHKOŽIVČEVE SANJE	29: 285. SINOVI ZAVRŽEJO MATER	
220. BRAT ALI LJUBI	3: 234. UTOPLJENI LJUBI	**G4: 8, 23, 10, 13, 25 (1:10)**
231. BOLNI LJUBI UMRJE	239. SMRT NEVESTE PRED POROKO/A	8: 236. LOVEC USTRELI LJUBICO IN SEBE
245. UGRABLJENA ŽENA NE SME DOMOV	243. SMRT ŽENINA PRED POROKO	251. POROD V GROBU
274. PASTOREK UMORI OČIMA	263. MOŽ KAZNUJE NEZVESTO ŽENO	252. VDOVEC NA ŽENINEM GROBU
22: 255. SMRT REJENKE	270. ŽENA UMORI OTROKA MOŽEVE LJUBICE	23: 260. MATI IZ GROBA TOLAŽI SIROTO
259. MRTVA MATI ZAGROZI MAČEHI	283. MOŽ SE VRNE NA ŽENINO SVATBO	10: 237. OČE DOLOČA USODO HČERE
6: 207. LJUBEZEN ZVABI MENIHA IZ SAMOSTANA	20: 248. Z ROPARJEM OMOŽENA	13: 240. SMRT NEVESTE PRED POROKO/B
1: 202. PRIDOBITEV LJUBEGA Z NAVIDEZNO SMRTJO	271. ŽENA ZASTRUPI MOŽEVO LJUBICO	256. MAČEHA IN SIROTA/A
204. ZVIJAČNA UGRABITEV NEVESTE	18: 244. ZVIJAČNA UGRABITEV MLADE MATERE	261. ZAPUŠČENE SIROTE/B
222. ZAVRNITEV VASOVALCA		277. OČE UMORI SINOVA
		286. NEVESTA DETOMORILKA
G2: 4, 21, 12, 16, 24, 7, 19 (18:9)		25: 261. ZAPUŠČENE SIROTE/A
4: 205. S SMRTJO REŠENA VSILJENE MOŽITVE	16: 217. UBOJ V FANTOVSKEM PRETEPU	
211. UMOR ZARADI ZAROKE Z DRUGIM	249. SMRT MATERE NA PORODU/A	**G5: 14 (4:1)**
275. PASTOREK UMORI MAČEHO	24: 225. NEZVESTI ŠTUDENT-NOVOMAŠNIK	14: 215. SAMOMOR NUNE ZARADI LJUBEZNI
21: 250. SMRT MATERE NA PORODU/B	232. SMRT VOZNIKA OB MRTVI LJUBICI	218. UMOR IZ LJUBOSUMJA
12: 213. ŽUPANOVA HČI IN GRAJSKI GOSPOD	233. SMRT ZAVRNJENEGA SNUBCA	219. UBOJ NA VASOVANJU
216. LJUBOSUMNI UBIJALEC OBSOJEN NA GALEJO	7: 208.[N] ČEZ MORJE V VAS	229. SPOKORJENA LJUBICA UMRJE
223. KAZNOVANI MLINAR ZAPELJIVEC	209. SMRT ČEVLJARJEVE LJUBICE	279. BRAT IN SESTRA SE NAJDETA/B
224. NEZVESTOBA IZ POHLEPA	214. OBUP ZAPUŠČENE LJUBICE	
226.MRTVEC OBIŠČE NEZVESTO LJUBICO	221. ZVESTOBA LJUBICE POPLAČANA	
227/A KAZNOVANJE DEKLETOVE ZAOBLJUBE	227. KAZNOVANJE NEZVESTOBE	
262. NEZVESTA ŽENA POBEGNE MOŽU	228.PREVARA PRI KAZNOVANJU NEZVESTOBE	
265. SMRT PREŠUŠTNIKA	278. BRAT IN SESTRA SE NAJDETA/A	
269. ŽENA DA UMORITI MOŽEVO LJUBICO	19: 246. PRISILNO DALEČ OMOŽENA	
284. KAZNJENEC ODKLONI VRNITEV K DRUŽINI		

Tabela 1: Razmerje tipov ljubezenskih in družinskih (obarvane sivo) pripovednih pesmi iz analize hierarhičnega gručenja.
Vir: lastni prikaz.

na tematiko ‘mačeha in sirota’, ki jo najdemo v več pesemskih tipih. Hkrati je treba poudariti, da so posamični pesemski tipi semantično globlje vpeti, kakor to prikazuje izbor besed v Tabeli 2, zato LDA zazna tudi specifičnejše primere. Tako se npr. beseda ‘gruntati’ pojavlja le v osmih variantah tipa ‘243. SMRT ŽENINA PRED POROKO’, a jo v porazdelitev (v SLP 5/gruča 4 je to knjižna oblika ‘meriti’!) veže semantična sorodnost z nekaterimi drugimi motivi te skupine: ‘gruntati’ v kontekstu omenjenega pesemskega tipa namreč pomeni »(gruntat sivo morje) ... meriti globino vode, do tal se potopiti« (SLP 5: 72). Varianto v to skupino dodatno postavljajo besede ‘morje’, ‘sin’, ‘črn’, ‘umreti’, ki jih najdemo tudi v temah nekaterih drugih pesemskih tipov (Tabela 3, SLP 5/gruča 4).

Slika 3 se navezuje na Tabelo 3 in prikazuje gručenje pesemskih tipov v štiri gruče za vsako od družin (SLP 4 in SLP 5). Porazdelitev v semantičnem prostoru kaže na tematsko sorodnost med tipi in gručami družine. Tako npr. porazdelitev ljubezenskih tipov tvori jasno ločene gruče in posledično tematsko jasneje razmejen semantični prostor (SLP 4, levo), medtem ko so tipi gruč družinskih pesmi (SLP 5, desno) bolj razpršeni, iz česar lah-

A: Sopojavljanje besed v posamezni družini								B: Sopojavljanje besed med družinama							
SLP 4								SLP 4							
šopek	zvesto	fant	sin	umreti	močen	sin	morje	šopek	zvesto	fant	sin	umreti	močen	sin	morje
puška	odsekati	mrtev	dete	zlat	polje	pristava	hlapec	puška	odsekati	mrtev	dete	zlat	polje	pristava	hlapec
brat	veselje	ljubica	sinoči	prstan	dom	neža	turški	brat	veselje	ljubica	sinoči	prstan	dom	neža	turški
ljubica	glavica	mati	umreti	kralj	bolan	konj	plavati	ljubica	glavica	mati	umreti	kralj	bolan	konj	plavati
gozd	tresoč	bog	marija	johana	sejem	žlahten	golob	gozd	tresoč	bog	marija	johana	sejem	žlahten	golob
ljubiti	karol	zelen	mož	grob	ljubica	kamra	voda	ljubiti	karol	zelen	mož	grob	ljubica	kamra	voda
dekle	fant	lipa	bel	bel	mati	bel	svetel	dekle	fant	lipa	bel	bel	mati	bel	svetel
lovec	laž	žlahten	grob	grad	jama	ljubljana	ljubica	lovec	laž	žlahten	grob	grad	jama	ljubljana	ljubica
ljubi	počiti	svet	ljubi	johan	matjaž	grad	kri	ljubi	počiti	svet	ljubi	johan	matjaž	grad	kri
vojska	listje	kamra	dekle	sestrica	hud	mati	turek	vojska	listje	kamra	dekle	sestrica	hud	mati	turek
meč	obraz			dekle	kačji	molitev	micka	meč	obraz			dekle	kačji	molitev	micka
				pokopati	dekla	spletična	prileteti					pokopati	dekla	spletična	prileteti
				puščavniški		oskrbnica						puščavniški		oskrbnica	
SLP 5								SLP 5							
bel	mrtev	mačeha	kri	voda	marija	morje	žena	bel	mrtev	mačeha	kri	voda	marija	morje	žena
grad	prstan	mati	ženin	mati	kraljič	sin	črn	grad	prstan	mati	ženin	mati	kraljič	sin	črn
ljubi	fant	kruh	črn	grad	bel	mati	hud	ljubi	fant	kruh	črn	grad	bel	mati	hud
mož	zlat	dete	česati	dar	zlat	bel	pristava	mož	zlat	dete	česati	dar	zlat	bel	pristava
umreti	marija	jokati	zemlja	venec	majhno	konj	neža	umreti	marija	jokati	zemlja	venec	majhno	konj	neža
žlahten	dom	star	vol	dete	prositi	voda	oče	žlahten	dom	star	vol	dete	prositi	voda	oče
dekle	barka	otrok	žena	pasti	maša	svetel	meriti	dekle	barka	otrok	žena	pasti	maša	svetel	meriti
svet	dom	grob	bog	človek	dom	grad	umreti	svet	dom	grob	bog	človek	dom	grad	umreti
ljubica	oče	jerca	ida	klada	krona	kri	gora	ljubica	oče	jerca	ida	klada	krona	kri	gora
		nevesta	oče	bog	obleka	hči	turek			nevesta	oče	bog	obleka	hči	turek
		hud	svet	micka	bukov	hlapec	bolan			hud	svet	micka	bukov	hlapec	bolan
		postelja		zelen						postelja		zelen			

Tabela 2: Prikaz najvplivnejših besed v posamezni družini. Sopojavljanje besed v družini (A) in v celotnem korpusu (B) je barvno poudarjeno.
Vir: lastni prikaz.

ko sklepamo na slabšo tematsko razmejenost gruč kakor tudi na prekrivanje tem pesemskih tipov iz različnih gruč.

Sklep

Z raziskavo smo želeli ugotoviti, v kolikšni meri lahko računalniške metode generirajo semantični prostor, ki bi bil dinamičen in hkrati dovolj interpretativen za tematsko analizo folkloristič-nih vsebin. Konkretno nas je zanimalo, kako uspešno metoda LDA razločuje med sorodnima družinama iste zvrsti, tj. med pesmimi o ljubezenskih usodah in sporih in tistih o družinskih usodah in sporih. Kot izhodišče smo vzeli formalno razvrstitev tipov ljubezenskih in družinskih pripovednih pesmi zbirk SLP 4 in SLP 5 ter jo primerjali z LDA rezultati hierarhičnega gručenja pesemskih tipov omenjenih zbirk. Rezultati kažejo na to, da lah-ko že na razmeroma majhnem in tematsko izrazito prepletenem korpusu ljudskih pesmi dobimo primerljivo razmejitev. Analiza je pokazala tematsko prevlado družinskih tipov nad ljubezenski-mi v razmerju 60/40 %, kar ustreza številčni prevladi družin-skih tipov v samem korpusu. Nadalje smo z analizo LDA lah-ko sklepali na nekatere skupne značilnosti pesemskih tipov in dobili pogled v splošno semantično strukturo korpusa. Po drugi strani pa porazdelitev gruč tipov kaže na to, da je za globljo se-mantično analizo in natančnejše razmejevanje pesemskih tipov med družinami in zvrstmi potreben obsežnejši korpus, z žanrsko (zvrstno) in tipološko raznovrstnejšim gradivom.

Na splošno je prednost računalniških metod predvsem v večdi-menzionalni semantični predstavitvi latentnih tem in razmerij, ki temeljijo na verjetnostni porazdelitvi in merah podobnosti ter posledično omogočajo kompleksnejša tipološka razločevanja. S tem lahko dobimo dinamično tipologizacijo vsebin, ki presega omejitve poznanih pristopov, saj pesmi ureja na podlagi poljub-no izbranih parametrov in konteksta ter hkrati omogoča uvid v semantično strukturo celotnega korpusa. Na podlagi mer podob-nosti lahko npr. po izbranih merilih izluščimo značilne variante ali tipe, ki povzemajo splošne značilnosti nekega pesemskega tipa ali zvrsti, kakor tudi mejne primere, ki niso izključno vezani na en pesemski tip ali zvrst, marveč imajo podobno tematsko porazdelitev kakor drugi tipi ali zvrsti. Generativna zasnova verjetnostnega tematskega modela LDA nam poleg omenjenega

SLP 4	SLP 5	
1: 219.UBOJ NA VASOVANJU	1: 240.SMRT NEVESTE PRED POROKO/B	3: 237.OČE DOLOČA USODO HČERE
220.BRAT ALI LJUBI	244.ZVIJAČNA UGRABITEV MLADE MATERE	238.ODKLONITEV POROKE
221.ZVESTOBA LJUBICE POPLAČANA	251.POROD V GROBU	245.UGRABLJENA ŽENA NE SME DOMOV
227.KAZNOVANJE NEZVESTOBE	252.VDOVEC NA ŽENINEM GROBU	247.SMRT DALEČ OMOŽENE
228.PREVARA PRI KAZNOVANJU NEZVESTOBE	253.TREH HČERA NAGLA SMRT	263.MOŽ KAZNUJE NEZVESTO ŽENO
234.UTOPLJENI LJUBI	254.ŽALOSTNA USODA TREH SINOV	269.ŽENA DA UMORITI MOŽEVO LJUBICO
235.SMRT OB SNIDENJU	255.SMRT REJENKE	278.BRAT IN SESTRA SE NAJDETA/A
236.LOVEC USTRELI LJUBICO IN SEBE	259.MRTVA MATI ZAGROZI MAČEHI	280.BRAT IN SESTRA UBEŽITA IZ UJETNIŠTVA
	260.MATI IZ GROBA TOLAŽI SIROTO	286.NEVESTA DETOMORILKA
2: 202.PRIDOBITEV LJUBEGA Z NAVIDEZNO SMRTJO	264.PREŠUŠTNIK IN LJUBICA KAZNOVANA	
203.PRIDOBITEV LJUBEGA Z NAVIDEZNO SMRTJO/B	266.NEZVESTA GOSPA S TREMI STRAŽARJI/A	4: 239.SMRT NEVESTE PRED POROKO/A
204.ZVIJAČNA UGRABITEV NEVESTE	267.NEZVESTA GOSPA S TREMI STRAŽARJI/B	243.SMRT ŽENINA PRED POROKO
209.SMRT ČEVLJARJEVE LJUBICE	268.NEZVESTA ŽENA POMAGA LJUBIMCU POBEGNITI	249.SMRT MATERE NA PORODU/A
213.ŽUPANOVA HČI IN GRAJSKI GOSPOD	279.BRAT IN SESTRA SE NAJDETA/B	250.SMRT MATERE NA PORODU/B
214.OBUP ZAPUŠČENE LJUBICE	281.ŽENA REŠI MOŽA IZ UJETNIŠTVA	265.SMRT PREŠUŠTNIKA
217.UBOJ V FANTOVSKEM PRETEPU	282.ŽENA NOČE Z MOŽEM NA POT	270.ŽENA UMORI OTROKA MOŽEVE LJUBICE
218.UMOR IZ LJUBOSUMJA	284.KAZNJENEC ODKLONI VRNITEV K DRUŽINI	271.ŽENA ZASTRUPI MOŽEVO LJUBICO
222.ZAVRNITEV VASOVALCA		272.SESTRA ZASTRUPI SESTRO
224.NEZVESTOBA IZ POHLEPA	2: 241.SMRT NEVESTE PRED POROKO/C	274.PASTOREK UMORI OČIMA
225.NEZVESTI ŠTUDENT-NOVOMAŠNIK	242.TAŠČI SE IZJALOVI UMOR SNAHE	275.PASTOREK UMORI MAČEHO
226.MRTVEC OBIŠČE NEZVESTO LJUBICO	246.PRISILNO DALEČ OMOŽENA	276.SIN NA TASTOVO POBUDO UMORI MATER
227/A KAZNOVANJE DEKLETOVE ZAOBLJUBE	248.Z ROPARJEM OMOŽENA (KATA KATALENA)	277.OČE UMORI SINOVA
231.BOLNI LJUBI UMRJE	256.MAČEHA IN SIROTA/A	285.SINOVI ZAVRŽEJO MATER
233.SMRT ZAVRNJENEGA SNUBCA	257.MAČEHA IN SIROTA B (SIROTA JERICA)	287.OBSOJENA DETOMORILKA
	258.MAČEHA IN SIROTA C (SVETA KRISTINA)	
3: 206.ZAPELJANI PUŠČAVNIK	261.ZAPUŠČENE SIROTE/A	
207.LJUBEZEN ZVABI MENIHA IZ SAMOSTANA	261.ZAPUŠČENE SIROTE/B	
210.DEKLE ZASTRUPI LJUBEGA	262.NEZVESTA ŽENA POBEGNE MOŽU	
229.SPOKORJENA LJUBICA UMRJE	273.BRAT UMORI SESTRO	
230.BOLNA LJUBICA UMRE	283.MOŽ SE VRNE NA ŽENINO SVATBO	
232.SMRT VOZNIKA OB MRTVI LJUBICI		
4: 205.S SMRTJO REŠENA VSILJENE MOŽITVE		
208.[N] ČEZ MORJE V VAS		
211. UMOR ZARADI ZAROKE Z DRUGIM		
212. LAHKOŽIVČEVE SANJE		
215. SAMOMOR NUNE ZARADI LJUBEZNI		
216. LJUBOSUMNI UBIJALEC OBSOJEN NA GALEJO		
223. KAZNOVANI MLINAR ZAPELJIVEC		

Tabela 3: Gručenje pesemskih tipov v SLP 4 in SLP 5.
Vir: lastni prikaz.

prinaša še en, bistven vidik: zmožnost učenja in posploševanja z novimi podatki. V praksi to pomeni, da opravljeno analizo širimo z dodajanjem novih pesemskih variant in da se posledično spreminja tematska porazdelitev semantičnega prostora – s tem lahko, npr., ujamemo pojav tematske transformacije, ki je inherentna slovenski ljudski pesmi. To je še posebej priročno v primerih, ko sčasoma pride do močne spremembe določene zvrsti, kar verjetnostna tematska analiza lahko zazna in omogoči, in tej transformaciji v semantičnem prostoru sledimo tako kronološko kot vsebinsko. Z vidika folklorista so predstavljene strojne me-

tode lahko orodje, ki mu, pri jasno zastavljeni raziskavi, ponudi dodaten uvid v semantično strukturo predmeta analize (Laudun in Goodwin 2013; Tangherlini 2013).

V prihodnosti želimo strojne metode vpeti v širše folkloristične in etnomuzikološke raziskave slovenske ljudske pesmi in glasbe ter izvesti semantično analizo na količinsko in tematsko obsežnejšem gradivu. Prednost strojnega pristopa je tudi, da se presežejo trenutna merila razvrščanja ljudskih pesmi (Klobčar 2010), saj lahko v analizo in razvrščanje hkrati vključimo več vidikov in se po drugi strani izognemo omejitvam ročne analize. Ker dokumentiranje gradiva v GNI ZRC SAZU že več let poteka v digitalni obliki (zvočni posnetki, vizualni zapisi, terenski zapisniki, transkripcije itn.), bodo strojne metode v prihodnosti zelo uporabne.

Literatura

ABELLO, James idr.: Computational folkloristics. *Communications of the ACM* 55/7, 2012, 60–70.

BERRY, David M: The Computational Turn: Thinking About the Digital Humanities. *Culture Machine* 12, 2011, 1–22.

BLEI, David M.: Probabilistic Topic Models. *Communications of the ACM* 55/4, 2012, 77–84.

BLEI, David M. idr.: Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3/4–5, 2003, 993–1022.

BUSA, Roberto: Half a Century of Literary Computing: Towards a »New« Philology. *Historical Social Research / Historische Sozialforschung* 17/2, 1992, 124–133.

GOLEŽ KAUČIČ, Marjetka: *Ljudsko in umetno. Dva obraza ustvarjalnosti*. Ljubljana: Založba ZRC, ZRC SAZU (Folkloristika), 2003.

GRČAR, Miha idr.: Obeliks. Statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. V: Tomaž Erjavec in Jerneja Žganec Gros (ur.), *Zbornik Osme konference Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan, 2012, 89–94.

GRIFFITHS, Thomas L. idr.: Topics in Semantic Representation. *Psychological Review* 114/2, 2007, 211–244.

GREENFIELD, Patricia M.: The Changing Psychology of Culture from 1800 through 2000. *Psychological Science* 24/9, 2013, 1722–1731.

HARTIGAN, John A.: *Clustering algorithms*. New York: John Wiley & Sons, Inc, 1975.

KLOBČAR, Marija: Zvrstnost slovenskih ljudskih pesmi. Refleksija pesemskega razvoja ali pogledov nanj. *Traditiones* 39/2, 2010, 125–147, DOI: 10.3986/Traditio2010390208.

KUMER, Zmaga: *Vloga, zgradba, slog slovenske ljudske pesmi.* Ljubljana: Založba ZRC, 1996.

LANDAUER, Thomas idr. (ur.): *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007.

LANDAUER, Thomas K. in Susan T. Dumais: A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104, 1997, 211–240.

LAUDUN, John in Jonathan Goodwin: Computing Folklore Studies: Mapping over a Century of Scholarly Production through Topics. *Journal of American Folklore* 126/502, 2013, 455–475.

MARTIN, Dian I. in Michael W. Berry: Mathematical Foundations behind Latent Semantic Analysis. V: Thomas K. Landauer idr. (ur.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007, 35–55.

MICHEL, Jean-Baptiste, idr.: Quantitative Analysis of Culture using Millions of Digitized Books. *Science* 331/6014, 2011, 176–182.

SHIFFRIN, Richard M. in Katy Börner: Mapping Knowledge Domains. *Proceedings of the National Academy of Sciences of the United States of America* 1, 2004, 5183–5185.

SLP 4. Marjetka Golež Kaučič idr. (ur.), *Slovenske ljudske pesmi. Knjiga 4. Pripovedne pesmi*. Ljubljana: Slovenska matica, 1998.

SLP 5. Marjetka Golež Kaučič idr. (ur.), *Slovenske ljudske pesmi. Knjiga 5. Pripovedne pesmi*. Ljubljana: Založba ZRC, ZRC SAZU, 2007.

STEYVERS, Mark in Thomas Griffiths: Probabilistic Topic Models. V: Thomas K. Landauer idr. (ur.), *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007, 424–440.

STRLE, Gregor in Matija Marolt: Novi pristopi. Odkrivanje semantičnih struktur znotraj etnoloških vsebin. *Glasnik SED* 54/1–2, 2014, 17–22.

TANGHERLINI, Timothy R.: The Folklore Macroscope. *Western Folklore* 72/1, 2013, 7–27.

Spletna vira

Spletni vir 1: Matlab Topic Modeling Toolbox 1.4. *UCI. Cognitive Scinece Experiments*; http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm, 18. 2. 2014.

Spletni vir 2: Stanford Topic Modeling Toolbox – Version 0.0.4. *The Stanford Natural Language Processing Group;* http://nlp.stanford.edu/software/tmt/tmt-0.4/, 12. 2. 2014.

Computational Folkloristics: A Semantic Analysis and Visualization of Topic Distribution of Song Types

This article presents an analysis of the latent semantic structure of Slovenian folk songs by using the natural language processing (NLP) techniques. The principal issues in this study that needed to be resolved were: 1. How effective are the NLP methods for the thematic analysis of folkloristic materials? 2. Can this computational approach uncover relevant typological boundaries within individual genres? In the experiments presented in the study, the topic model *Latent Dirichlet allocation* (LDA) was used on the corpus of love and family ballads from the collection *Slovenske ljudske pesmi* (*Slovenian folk songs*). While the themes of both ballad types are very similar and exhibit a strong intertextuality, we were nevertheless interested whether the LDA can detect semantic dimensions specific to individual song types, and thus disambiguate between love and family ballads. Hierarchical clustering has been applied to disambiguate individual variant types into relevant thematic groups. Additional insight into the corpus structure is given by representation of similarity measures in semantic space, based on probability distributions of semantic elements present in individual variant types. Results indicate that meaningful semantic delimitation is possible even when a relatively small and thematically intertwined corpus is concerned. The analysis shows thematic domination of family over love ballads by 60/40%, which reflects the formal classification of family and love variant types in the collection.