Računalna obrada književnih tekstova na primjeru analize korpusa ruskih romantičara i realista

Kasunić, Lorena

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2018./2019.

Lorena Kasunić

Računalna obrada književnih tekstova na primjeru analize korpusa ruskih romantičara i realista

Završni rad

Mentor: dr. sc. Petra Bago, doc.

Zagreb, svibanj 2019.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

Uvod

Računala su postala neophodan instrument u svim aspektima ljudskog društva, kako u privatnom životu, tako i u profesionalnom. Tehnologija je omogućila automatiziranje radnji koje su prije znale oduzimati mnogo vremena, s obzirom na to da ih je čovjek morao ručno provoditi. Neizostavan je doprinos digitalnih tehnologija znanstvenom svijetu. Eksploziju razvoja znanosti uvelike treba zahvaliti primjeni tehnoloških postignuća, u novije vrijeme ponajprije računala. U tom kozmosu znanosti i s njom povezane tehnologije javile su se informacijske znanosti, relativno mlado znanstveno područje.

Informacijske znanosti interdisciplinarno su područje s velikim opsegom interesa u različitim granama ljudskog djelovanja. Iako definirane kao društvena znanost, vrlo su povezane s ostalim znanstvenim područjima, posebno tehničkim i humanističkim znanostima. U ovom završnom radu pokušat će se naglasiti ta povezanost kroz primjer računalne obrade književnih tekstova, polja koje kombinira tehničke alate na humanističkim predmetima interesa. Takva vrsta obrade dio je digitalne humanistike, područja koje potiče zajednički rad humanista, lingvista, informacijskih stručnjaka, informatičara, sociologa i još mnogih drugih stručnjaka.

Cilj rada je iznijeti temeljne i trenutno najraširenije metode koje se koriste prilikom analiziranja književnih tekstova uz pomoć računala, tj. prilikom kvantitativnog pristupa. U prvom poglavlju iznijet će se temeljni koncepti digitalne humanistike, koje su bile faze njezinog razvoja te kako je ona povezana sa znanosti o književnosti. Zatim će se izložiti fenomen računalne obrade s posebnim naglaskom na analizi književnih tekstova. Drugi dio rada bavit će se aktualnim metodama i primjenama tih metoda u kvantitativnom pristupu književnim djelima. Treći dio prikazat će istraživanje koje se provelo u svrhu izrade ovog završnog rada - koje metode su se koristile, kako je teklo istraživanje, do kojih rezultata je došlo te koji zaključci se iz njih mogu izvesti. Na samom kraju pokušat će se prikazati glavni zaključci cjelokupnog rada.

Digitalna humanistika

Digitalna humanistika mlado je i interdisciplinarno područje koje povezuje humanističke znanosti i digitalne tehnologije. Oksfordski rječnik engleskog jezika definira digitalnu humanistiku na sljedeći način: „Akademsko područje koje se bavi primjenom računalnih alata i metoda na tradicionalne humanističke discipline kao što su književnost, povijest i filozofija." (Oxford Dictionaries, 2019). *The Digital Humanities Manifesto 2.0* poima digitalnu humanistiku kao neujedinjeno područje koje istražuje znanje koje nije samo ono u tradicionalnom tiskanom izdanju, tj. proučava znanje u tiskanom obliku koje se sve više povezuje i stapa s multimedijskim aspektima. Digitalni alati nisu samo instrumenti koji služe humanistici, nego aktivno utječu na produkciju i diseminaciju znanja iz područja humanistike.

Gdje se može uočiti poveznica između humanistike informacijskih znanosti, postoji li neki temelj za koji se može reći da je povezao ova dva područja? Marijana Tomić (2015) u svome izlaganju navodi jednu bitnu karakteristiku humanistike zbog koje se može utvrditi poveznica koja je dovela do stvaranja digitalne humanistike. Naime, ono što je humanistici predmet istraživanja (književna djela, povijesni dokumenti, slike itd.) i čini njezin istraživački korpus, informacijski su objekti, koji se danas čuvaju u informacijsko-dokumentacijskim, odnosno baštinskim institucijama. Poveznica između informacijskih znanosti i digitalne humanistike njihova je interdisciplinarnost i bavljenje informacijama, točnije njihovim pretraživanjem, interpretacijom i rudarenjem (Koltay, 2015). Cecire (2011), kako je navedeno u radu Koltaya (2015), ističe još neke dodirne točke, kao što je interes za tekst te otvorenost prema drugim, suvremenim medijima na kojima tekstovi mogu biti zapisani. Koltay (2016) ih čak naziva znanostima o informacijama, dakle disciplinama koje zanima proučavanje zabilježenih informacija. Robinson et al. (2015) kako je navedeno u radu Koltaya (2016) primjećuju da obje discipline imaju problema s tenzijom između njihovog akademskog statusa i tretiranja kao podupirućih disciplina unutar istraživanja u drugim znanostima. Uz to, digitalna humanistika dio je polja interdisciplinarne humanističke znanosti, ali i društvenih znanosti, s obzirom na to da se njome bave i informacijske znanosti. U segmentu interdisciplinarnosti Fry (2006), kako je navedeno u radu Koltaya (2016), uspoređuje digitalnu humanistiku s korpusnom lingvistikom jer se nalazi na granici između humanistike, društvenih znanosti i primijenjenih znanosti.

Digitalna humanistika izvorno je nosila naziv *computing in the humanities* ili *humanities computing* (Nikolić, 2016). U hrvatskom jeziku često se koristi i naziv „društveno-

humanistička informatika", kako se naziva jedan od kolegija na diplomskoj razini studija na Odsjeku za informacijske i komunikacijske znanosti na Filozofskom fakultetu u Zagrebu. Iz prethodno navedenih termina može se vidjeti da se digitalna humanistika u svojim počecima promatrala kao „sluga" humanistici, popratni alat koji bi se mogao iskoristiti. Prelazak na naziv „digitalna humanistika" nije označavao samo jezičnu promjenu nego i promjenu u načinu poimanja područja. Time je ovo polje postalo intelektualno, s vlastitim praksama, pravilima, standardima, teorijskim postulatima (Hayles, 2012 navedeno u Nikolić, 2016). Digitalna humanistika omogućuje pregledavanje, pretraživanje, spremanje i dodavanje bilješki digitalnim objektima, nakon što su izvorno analogni objekti digitalizirani (Konfic, 2017). Takvi objekti mogu se analizirati, kritički propitivati, praviti njihove raznovrsne vizualizacije (pomoću, primjerice, grafikona i dijagrama).

Valovi razvoja digitalne humanistike

U drugome manifestu o digitalnoj humanistici Schnapp i Presner govore o valovima razvoja: prvi - kvantitativni, i drugi - kvalitativni val. Kvantitativni val razvio se u 1990-im i 2000-im godinama, naglasak je bio na digitalizaciji i tehnološkim postavkama. Kvalitativni val (ili DH 2.0, kako ga naslovljavaju autori *Manifesta*) orijentira se na interakciju sa znanjem koje nastaje digitalno (eng. *digital born*), ne ograničava se samo na tekstove koji su nastali tiskanjem, za razliku od prvog vala. Može se reći da se prvi val razvoja bavi prvenstveno analizom teksta u užem smislu, a drugi val pokušava povezati što više polja i razviti hibridne metodologije (Nikolić, 2016). Berry (2012) kako je navedeno u radu Nikolića (2016) predlaže da se umjesto pojma valova koristi pojam slojeva jer se tako bolje oslikava međudjelovanje i utjecaj dviju etapa jednu na drugu. Na tom tragu Berry uvodi i treći val (ili sloj) digitalne humanistike koji predstavlja zaokret prema računalnim metodama (Nikolić, 2016). Time se podcrtava prva riječ u sintagmi, a to je digitalnost koja se prije više činila kao „služavka" humanistici, popratna pojava. Treći val digitalne humanistike propituje metode i opažanja koja su duboko utemeljena u humanističkim znanostima - pomno čitanje, definiranje kanona, periodizacija književnosti i sl. Međutim, ne smije se zapasti u „zamku" računalnih tehnologija i valja uvijek imati na umu da se računalne metode trebaju proučavati humanistički (Nikolić, 2016). Ovdje treba naglasiti da se digitalna humanistika ne bavi samo literarnim tekstovima, kako bi se moglo zaključiti iz gore iznesenog, nego i ostalim proizvodima humanističkih znanosti (npr. slikama, muzejskim predmetima, dokumentima, glazbenim djelima), ali i

upotrebom proširene stvarnosti (eng. *augmented reality*) ili zračne fotografije u području digitalne arheologije.

Digitalna humanistika i znanost o književnosti

Zašto se javila potreba digitaliziranja književnih tekstova? Prvotni cilj bio je da se omogući veća dostupnost široj publici, a posebno u akademskim krugovima – studentima, znanstvenicima, profesorima (Piotrowski, 2012). Digitalna humanistika kombinira kvalitativne metode iz proučavanja književnosti s kvantitativnim metodama koje vrše računala. Iako računalno proučavanje povijesnih tekstova započinje još 1949. godine kada je isusovac Roberto Busa pokušao napraviti indeks riječi iz knjiga svetog Tome Akvinskog i sličnih autora, odnos između ova dva područja nije razvijen koliko bi se očekivalo da će biti (Schreibman et al., 2004 navedeno u Piotrowski, 2012). Jedni od oštrih protivnika računalnog pristupa analiziranju književnih djela su književni teoretičari Katie Trumpener i Stanley Fish. Postoji neko generalno mišljenje da tehnologija želi tekstove svesti na puki zbir podataka i činjenica te ih na taj način lišiti njihove „humanosti" (Kerr, 2017). Kako i navodi Stephen Marche (2012) u svom članku o digitalnoj humanistici: „Literature is not data. Literature is opposite of data." Kvantitativni i kvalitativni dokazi sve se lakše kombiniraju, zamagljuju se granice između znanstvenih područja. Underwood (2016) uspoređuje strah od računalnih metoda s upotrebom relativno novog pojma velike količine podataka (eng. *big data*) – njegova novost, nejasna definiranost izaziva preuveličanu zabrinutost. U svojoj knjizi *Reading Machines: Toward an Algorithmic Criticism* Stephen Ramsay navodi drugačiji razlog zašto se humanisti „opiru" utjecaju računalnih znanosti – računalnim znanostima nedostaju odvažne teze, previše su vođene hipotezama, pozitivistički su obilježene te počivaju na definiranosti i opredijeljenosti (Hoover, 2016). Već spomenuti Stanley Fish priznaje, kako ističe Hoover (2016), da računalni alati mogu otkriti uzorke koje humanist ne može uočiti vlastitim čitanjem. I baš zato se ne može znati što će se računalnom analizom dobiti. Zbog toga proučavatelji nastavljaju istraživanje u nasumičnom i hirovitom kontekstu.

Zanimljiva razmišljanja o problemu književnog kanona daje Matthew Wilkens u svome članku *Canons, Close Reading, and the Evolution of Method* gdje uočava da se većina slaže da je hermetičnost kanona loša stvar, ali nitko ne čini ništa konkretno da bi se problem riješio. Čak i digitalni humanisti velik broj svojih studija posvećuju kanonskim djelima o kojima je ionako već mnogo toga napisano i istraženo. Wilkensa (2012) ne zabrinjava mogućnost da će

udaljeno čitanje naštetiti ili potisnuti pomno čitanje: „My sense is that we'll come out all right and that it's a trade—a few more numbers in return for a bit less text—well worth making."

Međutim, nužno se primjećuje određeni paradoks. Naime, kad su se u znanosti o književnosti pojavile raznolike književne teorije koje su tumačile tekstove na sebi svojstven način, prigovaralo se da pravo na interpretaciju imaju samo „praktičari", tj. pisci koji stvaraju književna djela, te da naglasak mora biti na samom čitanju, a ne ostalim naknadnim procesima (Kerr, 2017). Danas ima nebrojeno mnogo teorija koje su sastavni i legitimni dio znanosti o književnosti. Zašto ne očekivati da će se ista stvar dogoditi i s digitalnom humanistikom, da će i ona postati važan dio i pomoć u proučavanju književnosti?

Računalna obrada

Računalna obrada kao jedan od postupaka u digitalnoj humanistici primjenjiva je na sve one digitalne i analogne objekte koji se mogu proučavati, što znači da nije ograničena samo na tekstualne objekte. Međutim, s obzirom na to da je naglasak u ovom radu na analiziranju teksta i jezika, postavke računalne obrade iznijet će se u tome kontekstu. Računalna obrada (eng. *computational analysis*) postupak je koji je moguć jedino ako se posjeduje digitalizirani tekst. U slučaju da istraživač ima samo analogni tekst, on se mora prvo podvrgnuti procesu digitalizacije (Text Analysis Resources, 2016). Tekstovi koji se analiziraju pomoću računala, a predmet su promatranja humanističkih i društvenih znanosti, mogu se podijeliti u dvije velike skupine: fikcionalni i nefikcionalni tekstovi. Fikcionalni tekstovi drugi je naziv za književne tekstove koji po vrsti mogu biti pjesme, romani, drame, novele, pripovijetke itd. Nefikcionalni tekstovi u području humanistike odnose se na različite povijesne dokumente, rječnike, rasprave, teorijske eseje i sl., a ako se uzmu u obzir društvene znanosti, tada popis uključuje i novinske članke, udžbenike, zakone, znanstvene članke itd. Tekstovi ove dvije skupine imaju vrlo različite strukture, vokabular, stil pisanja, a razlikuju se i po vremenu nastanka, kako unutar skupine fikcionalnih tekstova, tako i unutar skupine nefikcionalnih tekstova (primjerice, rječnik Fausta Vrančića koji datira iz 1595. godine i Anićev *Veliki rječnik hrvatskog jezika* iz 2003. godine). Upravo zbog toga računalna obrada uzima u obzir o kojoj vrsti teksta se radi i koje se sve metode mogu nad njime koristiti (Kerr, 2017).

Što općenito računalna obrada može pružiti u odnosu na tradicionalno proučavanje tekstova bez pomoći računala? Prvo, omogućuje čitanje velikog broja tekstova u kratkom vremenskom roku. Ovdje se termin „čitanje" odnosi na mogućnost prolaženja kroz tekst, pritom se orijentirajući na neke specifične parametre (npr. način uporabe vlastitih imena u romanima Jane Austen). Drugo, tekstovi se mogu automatski klasificirati, ili čak žanrovski odrediti. Računala se treniraju kako bi mogla prepoznati je li riječ o rječniku, grčkoj tragediji, povijesnom epu ili povelji. Također, moguće je odrediti autorstvo određenog teksta na temelju analize korpusa autora za kojeg se pretpostavlja da je napisao spomenuti tekst nepoznatog autorstva. Treće, olakšava se uočavanje poveznica između vremenski udaljenih tekstova (Computational Textual Analysis, 2018). Četvrto, računalnim programima može se napraviti vizualizacija podataka koji su prikupljeni (pomoću grafikona, tablica, obilježavanja tekstova itd.) Peto, ovakvim analizama istraživači mogu empirijski i statistički potvrditi valjanosti svojih hipoteza, mogu dobiti dokaze o svojim teorijama „crno na bijelo" (Kerr, 2017).

Da bi se računalna obrada provela u djelo nužna je suradnja stručnjaka iz različitih područja: lingvistike, humanistike, sociologije, statistike, računalnih znanosti, informacijskih znanosti (Text Analysis Resources, 2016). Način na koji će se računalna obrada izvesti uvelike ovisi o programerskim vještinama pojedinca, hoće li raditi svoje istraživanje u programskim jezicima kao što su *Python*[1] i *R*[2], ili u *out-of-the-box* alatima kao što su *MALLET*[3] ili *Paper Machines*[4].

Računalna obrada književnih tekstova

Književni tekstovi dijele se u dvije kategorije koje je bitno istaknuti kada se radi o njihovoj računalnoj obradi. Naime, ako je djelo nastalo u nekome od prethodnih književnih razdoblja koja ne pripadaju suvremenosti ili bližoj prošlosti, govorimo zapravo o povijesnim tekstovima. Treba napomenuti da se u ovom kontekstu pridjev „povijesni" odnosi na vremensku kategoriju, a ne nužno na povijesnu važnost određenog teksta. S druge strane, ako su djela napisana modernim jezikom, uvelike je olakšana njihova analiza, što se tiče samog jezika. Pojednostavljeno rečeno, povijesni tekstovi koriste određene riječi, fraze i pravila koja više ne postoje u suvremenom jeziku, ili se veoma rijetko koriste. Zbog toga je izazov analizirati takve tekstove, posebno zato što su alati koji se koriste većinom trenirani na korpusima suvremenih tekstova, kao što su primjerice novinski članci. Ipak, kako i Piotrowski (2012) naglašava, sve što se može vršiti nad modernim tekstovima, može se i nad povijesnima: pretraživanje cijelog teksta, konkordancija, lematizacija, morfološka i sintaktička analiza, rudarenje teksta, strojno prevođenje itd. Književni korpus koji će se analizirati u ovome radu sastoji se od povijesnih tekstova ruskih romantičara i realista jer su prijevodi nastali u 19., odnosno 20. stoljeću, na hrvatskom jeziku: *Evgenij Onjegi*n iz 1881. godine (prijevod Ivana Trnskoga), *Junak našeg doba* iz 1918. godine (prijevod Milana Bogdanovića), *Ana Karenjina* u prijevodu Martina Lovrenčevića (godina prijevoda nije navedena) i *Zločin i kazna* u prijevodu Ise Velikanovića (godina prijevoda nije navedena).

Iako se danas računalna obrada shvaća kao postupak koji se poglavito služi kvantitativnim pristupom, odnosno metodama koje se u njemu koriste, kvalitativni pristup također ima svoj udio u računalnoj obradi i ne smije ga se isključiti. Ovaj rad prvenstveno se orijentira na

1 Dostupno na: https://www.python.org/
2 Dostupno na: https://www.r-project.org/
3 Dostupno na: http://mallet.cs.umass.edu/
4 Dostupno na: http://papermachines.org/

kvantitativni pristup, međutim iz narednih konstatacija bit će vidljivo da kvantitativno i kvalitativno nužno idu ruku pod ruku, međusobno se nadopunjujući.

Kvantitativni pristup

Hoover (2008) daje svojevrsnu definiciju ovog pristupa pa tako kaže da je kvantitativni pristup književnim tekstovima onaj koji numerički predstavlja elemente ili obilježja književnih tekstova, pri tome primjenjuje snažne, precizne i široko prihvaćene metode iz matematike na mjerenje, klasifikaciju i analizu. Porast broja digitalno dostupnih tekstova povećao je zanimanje za ovaj pristup i naglasio njegovu inovativnost u načinu obrade književnih tekstova. Petrović i Vranešević (2015) definiraju ga kao inovativan način čitanja književnih tekstova. Ono čime se kvantitativni pristup u proučavanju književnih djela najviše bavi jest pitanje autorstva i stila, ali zanima se i za neka specifičnija i kompleksnija pitanja kao što su: pitanje žanra, tematike, tona teksta, periodizacije (Hammond, 2016). Ne smije se zanemariti ni lingvistička razina i bavljenje gramatičkim, pravopisnim, sintaktičkim i morfološkim istraživanjima teksta. Sposobnosti računala u procesu analiziranja dolaze na vidjelo kada treba izvući zaključke iz podataka koje čak ni profesionalni čitatelj (sveučilišni profesor, književni kritičar ili teoretičar književnosti) ne može „detektirati" pomnim čitanjem (eng. *close reading*), nego samo pomoću udaljenog čitanja (eng. *distant reading*). Podatak o broju prijelaznih glagola, posvojnih pridjeva ili ukupnom broju riječi u Dickensonovim *Velikim očekivanjima* može se dobiti samo kvantitativnim računalnim pristupom. Uz ova dva oprečna pojma, udaljeno čitanje i pomno čitanje, Berenike Herrmann (2017) spominje pojam čitanja na srednjoj udaljenosti (eng. *middle-distance reading*) kao termin koji predstavlja spoj prije spomenutih oprečnih načina čitanja. Mnogi stručnjaci koji u svome proučavanju književnosti upotrebljavaju računalne alate, tj. kvantitativni pristup sve više zagovaraju stajalište da bi se računalna obrada trebala „udružiti" s tradicionalnim proučavanjima tako da se međusobno nadopunjuju (Hammond, 2016). Kvantitativni pristup mora se uskladiti s postojećim idealima i praksama u humanistici. Iznošenje pukih činjenica o tome koliko imenica ima u nekom romanu nema nikakvu svrhu ako se ne razmotri i precizira svrha takvog podatka. Jedno pojavljivanje određenog jezičnog svojstva može biti puno zanimljivije i važnije od deset pojavljivanja drugog jezičnog svojstva (Petrović i Vranešević, 2015). Kvantitativna analiza odlična je za detektiranje onoga što se u tekstovima javlja rijetko ili na neuobičajen način. A to se jedino može otkriti brojanjem i uspoređivanjem koje omogućava računalna obrada (Hoover, 2008).

Metode unutar kvantitativnog pristupa

Svaka pojava unutar teksta koja se može pouzdano identificirati (primjerice riječi, značajke, stil) može se i izbrojiti (Hoover, 2008). Odluku o tome što će se brojati donosi onaj tko provodi analizu. Ta odluka može biti jednostavna ili kompliciranа. Loše odluke koje nisu prethodno detaljno promišljene mogu lako dovesti do nepotrebnog gubljenja vremena na istraživanja koja neće donijeti nikakvog rezultata. Ono što se najviše i najčešće prebrojava jesu riječi, najmanje jedinice koje nose neko značenje. Uz riječi, objekt interesa su i sintaktičke kategorije kao što su imenice, glagoli, pridjevi itd. Ovisno o vrsti teksta koja se proučava, interpunkcija također može biti podvrgnuta analizi (u lirskim pjesmama i dramskim tekstovima interpunkcija ima važnu ulogu). S obzirom na to da je riječ o kvantitativnom pristupu, statističke metode vrlo su popularne i korisne. Takvim metodama pokušava se doći do zaključaka koje nije moguće izvesti kvalitativnim pristupom, zbog praktičnih razloga (ne može se prikupiti tolika količina statističkih podataka).

Osim metoda preuzetih iz statistike, vrlo su važne i metode iz područja obrade prirodnog jezika. Književni tekstovi su lingvističke strukture, jezik je njihovo glavno sredstvo prenošenja ideja, poruka, estetskog užitka. Zato svaki književni korpus mora proći kroz neke temeljne metode obrade prirodnog jezika kao što je, primjerice, tokenizacija ili lematizacija. Na taj se način književna djela „ogoljuju" na svoje elementarne dijelove – riječi, rečenice, interpunkcije, poglavlja – koji se mogu pobrojati i kojima se može mnogo jednostavnije manipulirati nego tekstom u cjelini. Lopina (2012) potvrđuje prethodnu tezu govoreći da svako istraživanje korpusa započinje primjenom kvantitativnih metoda. Još jednom valja istaknuti da kvantitativne metode neće donijeti nikakva revolucionarna otkrića ako ne postoje književni stručnjaci koji će znati protumačiti dobivene rezultate.

Metode preuzete iz statistike

U skupini statističkih metoda ističe se uspoređivanje frekvencija ili prosjeka, standardna devijacija (unutar koje je moguće računanje z-rezultata) te omjer razlikovnosti (eng. *distinctiveness ratio*). Općenito govoreći, statistika je: „grana matematike koja obuhvaća sakupljanje, analizu, interpretaciju i prezentaciju podataka te izradu predviđanja koja se temelje na tim podacima" (Osnove statistike, n.d.). Dijeli se na dvije vrste: deskriptivnu i inferencijalnu. Deskriptivna statistika je ona koja pokriva područje statističkih metoda koje se koriste u računalnoj obradi književnih tekstova. Naime, takva vrsta statistike upotrebljava se

za opisivanje osnovnih značajki podataka; pruža jednostavne sažetke o uzorku i mjerama te čini temelj svake kvantitativne analize podataka, pa tako i onih koji se dobivaju analiziranjem književnih djela, stoga se deskriptivna statistika koristila u istraživanju odabranih realističkih i romantičarskih djela (The meaning of statistics and digital humanities, 2012).

Od spomenutih metoda najjednostavnije je uspoređivanje frekvencija gdje se uspoređuju prosjeci pojedinih grupa tekstova, najčešće pomoću statističkih testova važnosti, kao što su t-test (eng. *Student's T-test*) ili hi-kvadrat (eng. *Chi-square*), da bi se vidjelo radi li se o razlici koja je nastala slučajno ili je posrijedi primjena određene zakonitosti koja funkcionira unutar korpusa. T-test je statistički postupak kojim se testira značajnost razlike između dva uzorka. Uspoređuju se njihove aritmetičke sredine (Vježbe iz statistike, 2017). Potrebno je imati takozvanu nul-hipotezu (postoji li, primjerice, statistički značajna razlika u primjeni opisnih pridjeva kod ruskih realista i romantičara). Ako t-test pokaže da razlika između aritmetičkih sredina nije značajna, potvrdila se početna nul-hipoteza. Hi-kvadrat također gleda razlike između uzoraka, tj. kolika je mogućnost da je distribucija podataka podložna slučajnosti. Test funkcionira tako da analizira samo unaprijed kategorizirane podatke, dakle podatke koji su prebrojani i podijeljeni u skupine prije same analize. Zato se ne može provoditi na brojčanim ili kontinuiranim podacima (Tutorial: Pearson's Chi-square Test for Independence, 2008).

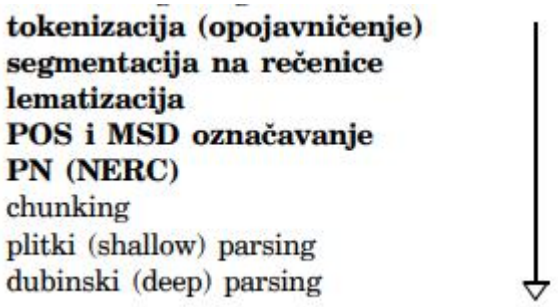
Standardnu devijaciju najsažetije moglo bi se definirati kao mjeru disperzije. Njome se mjeri koliko su skupovi podataka udaljeni od srednje vrijednosti. Što je veća udaljenost između srednje vrijednosti i pojedinih podataka, to je veća devijacija unutar skupa podataka. Korisnost ove metode ogleda se u mogućnosti predviđanja trendova, mjerenja odstupanja te ukazivanja na količinu raspršenosti podataka (Hargrave, 2019). U kontekstu književnosti to bi značilo da pomoću dobivenih podataka stručnjak može vidjeti na koji način se razlikuje zastupljenost neke riječi u korpusima koji se uspoređuju (Hoover (2008) kao primjer daje razlike u pojavnosti riječi *upon* i *on* u *Silasu Marneru* i *Sajmu taštine*). Dijeljenje frekvencije pojave u tekstu s njezinom frekvencijom u drugom tekstu još je jedna od jednostavnih statističkih metoda koja bi se mogla definirati kao mjera razlikovnosti između tekstova (Hoover, 2008). Ako je dobiven omjer ispod 0.67 ili iznad 1.5, vrijedi ga istražiti.

Statističke metode pokazale su se korisnima u analiziranju sadržaja, posebno kod korpusa koji sadrže velik broj tekstova i gdje čovjek bez asistencije računala ne može provesti istraživanje. Pripisivanje autorstva i stilometrija, kao jedni od najpopularnijih pristupa u kvantitativnoj analizi književnih tekstova, temelje se na primjeni statističkih metoda (Hoover, 2008). Očito

je da statistika igra bitnu ulogu u procesu analiziranja književnosti jer svojim objektivnim i empirijskim parametrima daje čvrsto polazište za daljnje istraživanje.

Metode preuzete iz obrade prirodnog jezika

Velik broj ovih statističkih metoda koristi se u obradi prirodnog jezika, polju koje je usko povezano s računalnom obradom književnih tekstova. Metode koje će se koristiti u ovom radu (kao što je POS označavanje) dolaze zapravo iz prakse prirodne obrade jezika. One koje se najviše primjenjuju u analizi književnih tekstova su morfosintaktičko označavanje, POS označavanje (eng. *Part-of-speech tagging*), lematizacija i morfološka analiza te sintaktičko parsiranje (Piotrowski, 2012). Sve navedene metode moraju prvo proći kroz proces tokenizacije, odnosno opojavničenja, što se i vidi na slici 1:



Slika 1. Redoslijed metoda u obradi prirodnog jezika (Bekavac, 2002)

Morfosintaktički označivači pridružuju pojavnicama u korpusu oznaku za vrstu riječi kojoj pripadaju (Hrvatski mrežni rječnik - Mrežnik). Smatraju se jednom od temeljnih metoda u obradi prirodnog jezika. Većina modernih MSD označivača statističke su prirode te su trenirani na označenom korpusu (Piotrowski, 2012). Prilikom označavanja povijesnih, odnosno književnih tekstova, postoji nekoliko načina kako se može napraviti POS označivač ili MSD označivač. Piotrowski (2012) iznosi njih šest:

1. Stvaranje označivača „od nule” (ručno ili poluautomatski se označi korpus na kojem se označivač trenira)
2. Proširivanje postojećeg modernog označivača dodavanjem riječi iz povijesnih tekstova u leksikon modernog označivača

3. Korištenje modernog označivača na povijesnom tekstu, ručno ispravljati pogreške te ga ponovno istrenirati na temelju ispravljene verzije
4. Moderniziranje pravopisa povijesnih tekstova te nakon toga primijeniti moderni označivač
5. „Postarivanje“ modernog korpusa i zatim ponovno istrenirati označivač za korištenje na povijesnim tekstovima
6. Uspoređivanje modernih i povijesnih tekstova i zatim se oznake dobivene analizom modernih tekstova koriste za analizu povijesnih.

Iz izloženog se vidi da morfosintaktičko obilježavanje književnih tekstova nije toliko jednostavno kao ono modernih tekstova (novinskih članaka, vijesti i sl.) iz razloga što velik broj književnih tekstova nije pisan suvremenim jezikom, struktura rečenica i upotreba riječi odstupa od svakodnevnog načina izražavanja koji je prisutan u neknjiževnim korpusima. Uz to, dostupnost MSD i POS označivača za velike svjetske jezike, kao što su engleski i njemački, nije upitna, ali manji jezici često uopće nemaju razvijene spomenute označivače (srećom, za hrvatski jezik su razvijeni).

Lematizacija je: „postupak svođenja riječi na osnovni oblik” (Hrvatski jezični portal). Termin je preuzet iz lingvistike. Osnovni oblik riječi naziva se lema. Da bi lematizacija bila uspješna, nužno je znati o kojoj se vrsti riječi radi, a u tome pomaže POS označavanje (Piotrowski, 2012). Istopisnice su problem kad se želi postići što veća točnost automatske lematizacije (Bekavac, 2002). Bitno je reći da lematizacija uzima u obzir kontekst u kojem se riječ nalazi kako bi mogla pravilno odrediti osnovni oblik riječi (Lemmatization Approaches with Examples in Python). Ovo je iznimno važno kod hrvatskog jezika s obzirom na to da hrvatski ima složenu tvorbu riječi. Izrada alata za lematizaciju zahtijeva dobro poznavanje strukture i zakonitosti jezika za koji će se taj alat koristiti.

Sintaktičko parsiranje proces je raščlanjivanja rečenice na sintaktičke strukture u računalnoj obradi jezika (Hrvatski mrežni rječnik - Mrežnik). Kako je već istaknuto da je morfosintaktičko označavanje puno zahtjevnije za povijesne nego moderne tekstove, tako je i sa sintaktičkim parsiranjem. Štoviše, Piotrowski (2012) ističe da je vrlo malo radova i konkretnih pomaka napravljeno u ovom segmentu obrade prirodnog jezika. Kad treba provesti kvalitetno sintaktičko obilježavanje, istraživači se često okreću ručnom označavanju. Ponekad sintaktičko parsiranje može čak biti jednostavnije za tekstove napisane na starijem jeziku. Primjer za to je istraživanje Huanga et. al. (2002) koji su istaknuli kako im je manje problema zadavalo parsiranje klasičnog kineskog jezika nego modernog kineskog, zbog toga što se

većina riječi sastoji od samo jednog znaka, dok moderni kineski koristi više znakova za jednu riječ. Međutim, ipak su imali poteškoća zbog dvosmislenosti leksičkih kategorija riječi u klasičnom kineskom. Metode preuzete iz obrade prirodnog jezika ne mogu zadovoljiti sve zahtjeve koje pred njih stavljaju književni korpusi, s obzirom na to da su oni lingvistički vrlo heterogeni. Kao glavni problem s kojim se ove metode moraju nositi jest oskudnost podataka (Piotrowski, 2012).

Primjena metoda

Navedenim metodama ne služi se samo računalna obrada književnih tekstova, nego i drugi pristupi koji se bave nefikcionalnim tekstovima. Struktura i pravila kojima podliježu fikcionalni i nefikcionalni tekstovi često su vrlo različiti. To se treba uzeti u obzir i kad se računalno analiziraju. Metode se mogu poklapati, ali kako će one biti usmjerene, na što će se fokusirati, na koja pitanja pokušati odgovoriti, to ovisi o vrsti teksta koja je pred njih stavljena. Iz tog razloga korisno je objasniti neke od trenutno najpopularnijih načina kako se spomenute metode koriste u analizi književnih tekstova. Naravno, gotovo sve primjene koje će se iznijeti mogu se koristiti i za nefikcionalne tekstove, ali ovdje će se one sagledati s isključivo književnog aspekta te oprimjeriti konkretnim situacijama iz književnih djela.

Kvantitativna tematska analiza

Kvantitativna tematska analiza često se koristi u računalnoj obradi književnih tekstova. Njome se može pratiti rast, pad i razvoj vokabulara unutar određene tematske cjeline, ili na koji način autori upotrebljavaju odabrane izraze kako bi iznijeli temu kojom se bave (npr. kako Gogolj i Dickens prikazuju temu malograđanštine) (Hoover, 2008). Analiza se provodi tako da se uzme neka smislena cjelina unutar teksta i zatim se gleda, na temelju riječi koje se koriste, koje sve teme su u tom bloku teksta obuhvaćene (Roberts 2000). Ono što Roberts (2000) vidi kao moguću zamku ovog pristupa je činjenica da se ponekad ne može odrediti je li pojava neke sintagme uzrok ili posljedica pojave druge sintagme. Kao što se može zaključiti iz dosad rečenog, i ova metoda počiva na statističkim temeljima kako bi došla do podataka koje će obraditi. Kvantitativna tematska analiza mora svoje podatke (tj. riječi) imati pohranjene u obliku dvodimenzionalne matrice da bi se provela statistička analiza – jedan red za svaki uzorak teksta i jedna kolona za svaku temu ili koncept koji se može pojaviti u uzorcima teksta (Roberts, 2000). Shema tablice vidljiva je na slici 2.

Ova vrsta analize većinom se usmjerava na vokabular i riječi. Pritom treba biti oprezan kod donošenja zaključaka o svrsi uporabe pojedinih riječi. Naime, pisci nerijetko svakodnevne i

ID-number	Theme 1	Theme 2	Theme 3
1	2	0	0
2	0	0	1
3	1	3	1
4	0	2	1
5	0	0	0
.	.	.	.
.	.	.	.
.	.	.	.

Slika 2. Shema matrice za kvantitativnu tematsku analizu (Roberts, 2000)

dobro poznate riječi transformiraju u značenju koje se može u potpunosti razlikovati od onog koje je zapisano u rječnicima i leksikonima. Nikad se ne smije smetnuti s uma da je književnost umjetnost i da je pisac umjetnik koji ima potpunu slobodu pridavanja značenja svojim konstruktima.

Pripisivanje autorstva

Glavna pretpostavka u pripisivanju autorstva (eng. *authorship attribution*) uz pomoć računala je da je mjerenjem nekih značajki teksta i uspoređivanjem s drugim tekstovima moguće uočiti razliku u autorstvu (Stamatatos, 2008). Hoover (2008) upozorava da postoje dvije vrste pripisivanja autorstva – forenzička i literarna. Obje koriste iste metode kako bi došle do rješenja, iako proučavaju različite vrste tekstova (forenzička analizira tekstove kao što su prijeteća pisma, poruke otmičara, ali i SMS poruke, komentare na društvenim mrežama itd., a literarna se najviše bavi književnim tekstovima). Razlika između ova dva tipa je u njihovom cilju. Forenzičko pripisivanje zanima tko je poslao poruku, a ne nužno tko je napisao poruku, dok se literarno pripisivanje fokusira na estetsku i kulturnu vrijednost teksta i njegovog autora (Hoover, 2008). Informacijsko pretraživanje, obrada prirodnog jezika i strojno prevođenje imali su veliki utjecaj na razvoj pripisivanja autorstva, kao i porast broja digitalno dostupnih tekstova (Stamatatos, 2008).

Proces započinje odabiranjem nekog teksta čije autorstvo je nepoznato, odabiru se potencijalni autori i analiziraju njihova djela za koja se može sa stopostotnom sigurnošću tvrditi da su njihova. Zatim se, na temelju određenih karakteristika pisanja svakog autora, određuje tko bi mogao biti autor anonimnog teksta (Stamatatos, 2008). Pripisivanje autorstva tradicionalno se služilo varijablama kao što su duljina riječi, rečenica ili bogatstvo vokabulara. Međutim, u zadnje vrijeme sve se više primjenjuju složenije metode kao što je analiza glavnih komponenti (eng. *principal components analysis*) te analiza klastera (eng. *cluster analysis*), kako tvrdi Hoover (2008). U svojem članku iz 2013. godine pod naslovom *Textual Analysis* Hoover navodi da postoje i primjeri gdje su slijed slova i sintaktičke oznake bile od veće koristi nego same riječi. Grieve (2007) navodi još neke frekvencije koje pripisivanje autorstva upotrebljava: frekvencija grafema, riječi, interpunkcijskih znakova, kolokacija, n-gram frekvencija na razini znaka. Za uspoređivanje vrijednosti većinom se služi hi-kvadrat testom. Neki od glavnih zadataka pripisivanja autorstva su: verifikacija autorstva, otkrivanje plagijatorstva, profiliranje, tj. izdvajanje osobina autora (dob, obrazovanje, spol autora), detektiranje stilskih nedosljednosti (Stamatatos, 2008). Bez obzira na to što pripisivanje autorstva postoji već više desetljeća, Rudman (1958; vlastiti prijevod) kako je navedeno u radu Statamatos (2008) je još u 50-im godinama pesimistički zaključio da se znanstvenici nisu „pomaknuli s mrtve točke": „Činjenica je da je u ovom polju ostalo još mnogo redundantnosti i metodoloških nepravilnosti, dijelom zbog interdisciplinarne naravi polja." Ono što je i dalje otvoreno pitanje jest koliko bi dugačak trebao biti tekst da bi se mogle detektirati njegove stilističke značajke, kao i pitanje važnosti nekih drugih faktora u utvrđivanju autorstva (Stamatatos, 2008). Pripisivanje autorstva usko je povezano sa stilometrijom, bitnim ogrankom unutar kvantitativnog pristupa književnim tekstovima.

Stilometrija

Stilometrija je kvantitativni pristup u proučavanju književnih stilova uz pomoć metode udaljenog čitanja (eng. *distant reading*) (Laramée, 2018). Počiva na pretpostavci da autori pišu na prepoznatljiv, jedinstven način koji se može strojno detektirati. Gomez-Adorno (2018) ističe da se stilometrijom posebno često služi pripisivanje autorstva budući da je stil jedan od temeljnih signala da neko djelo pripada baš tom određenom autoru. To nije jedina moguća uporaba stilometrije. Njome se identificiraju stilovi vezani uz pojedini žanr, nacionalnu književnost, književno razdoblje i sl. (Stewart, primjerice, proučava stil naracije u dva romana Charles Brockdena Browna). Hoover (2008) gleda na stilometriju (ili statističku stilistiku,

kako se ponekad naziva) kao najšire područje unutar kvantitativne analize koje sumira sve prije spomenute metode i primjene. Problemi koje stilometrija proučava najbliži su onima kojima se bavi znanost o književnosti, posebno se zanima za uzorke i ponavljanja koja su povezana s pitanjima interpretacije, značenja i estetike. Moglo bi se reći da stilometrija najdublje ulazi u samu srž pokušaja shvaćanja značenja književnog djela na temelju statističkih podataka. Da bi to bilo ostvarivo, ključno je koncentrirati se na relevantne aspekte teksta. Podatke o autorstvu, žarnu, vremenu nastanka Daelemans (2013) naziva metapodacima. Sam proces stilometrijske analize sastoji se od nekoliko složenih, višefaktorskih faza predobrade, izdvajanja značajki, statističke analize i prezentacije rezultata, najčešće vizualnim putem (Eder et al., 2016). Iako se neprestano poboljšavaju metode unutar stilometrije, i dalje postoje problemi na koje se zasad ne može ponuditi konkretno rješenje. Daelemans (2013) kaže da je to pitanje promjenjivosti stila autora tijekom života, tj. da se individualni stil mijenja kroz vrijeme, te pitanje podrijetla stila - može li se on imitirati ili je to ipak nesvjesna aktivnost. Stańczyk i Cyran (2007) zamjećuju da analiziranje sintaktičkih oznaka, koje se dobivaju morfosintaktičkim označavanjem, može puno objektivnije reći nešto o stilu autora s obzirom na to da ih pisci koriste podsvjesno te su stoga manje izloženi mogućnosti imitacije od strane drugih autora. Česta situacija jest da se analiziranjem vokabulara zapravo odredi tematika, a ne stil djela, zato što se proučavaju sadržajne riječi. Daelemans (2013) je mišljenja da kad bi se izostavile sadržajne riječi iz analize, mogao bi se odrediti stil. Nadalje, navodi i međužanrovsku generalizaciju kao situaciju s kojom se treba uhvatiti u koštac. Međužanrovska generalizacija je postupak izvođenja općenitih zaključaka o stilu pojedinog autora na temelju proučavanja samo jednog ili nekoliko žanrova u kojima se autor okušao. Mogu li se zaključci dobiveni analiziranjem stilskih obilježja Kafkinih pripovijedaka primijeniti na utvrđivanje autorstva anonimnih pisama za koje se pretpostavlja da pripadaju Kafkinom opusu?

Analiza ključnih riječi

Iako stilometrija i pripisivanje autorstva prednjače u popularnosti, analiza ključnih riječi također se pokazala kao obećavajuć i koristan postupak. Kao što i sam naziv govori, naglasak je na proučavanju ključnih riječi u određenom korpusu. Prvo treba definirati što se podrazumijeva pod terminom „ključna riječ". Ključna riječ je: "riječ ili skupina riječi uzeta iz naslova ili nekog teksta koja označuje njegov sadržaj i omogućuje nalaženje u spremljenom materijalu" (Hrvatski jezični portal). Fischer-Starcke (2009) ističe da ključne riječi mogu

svoju važnost imati u tome što su značajne za sadržaj teksta ili za njegovu strukturu. Smatra da se ključne riječi ne mogu odrediti intuitivno, već korištenjem kvantitativnog usporednog pristupa analizi. Dok je istraživala ključne riječi u romanu *Ponos i Predrasude* Jane Austen, Bettina Fischer-Starcke poslužila se dvama referentnim korpusima iz kojih je izvukla liste ključnih riječi i uspoređivala ih s riječima u samom romanu. Time je legitimnost dominantnih riječi i tema dvostruko provjerena te je moguće identificirati razlike između dva referentna korpusa i kako te razlike utječu na rezultate analize. Nigdje se ne precizira koliko bi opsežan trebao biti referentni korpus, ali Rayson (2008) kaže da mora biti opsežniji od korpusa koji se planira analizirati (Rayson koristi naziv *corpus of interest*). Ono zašto je analiza ključnih riječi bitna jest to da se pomoću tih riječi može objektivnije prosuđivati o stilu i preokupacijama nekog autora. Ne postoje puka nagađanja kojim se temama Jane Austen dominantno bavila, već se pogledaju rezultati analize i donesu empirijski utemeljeni zaključci. Da bi se poboljšala ključnost (eng. *keyness*) ključnih riječi, skovan je pojam *key-keywords*. To su riječi koje se pojavljuju kao ključne u više tekstova koji čine analizirani korpus, nisu ključne samo za jedan tekst unutar korpusa (Conway, 2009). Tako se ključne riječi mogu rangirati s obzirom na razinu svoje „ključnosti". Ključne riječi u listama mogu se podijeliti u dvije skupine, kako navodi Rayson (2008): pozitivne (one koje su neobično česte u istraživanom korpusu u odnosu na referentni korpus) i negativne (one koje su neobično rijetke u istraživanom korpusu u odnosu na referentni korpus). Scott (2000) kako je navedeno u radu Raysona (2008) zaključuje da ne postoji sigurnost da bi riječi koje je čovjek ručno istaknuo kao ključne bile iste onima koje su računalno tako kategorizirane. Nadalje, neki autori ključne riječi promatraju u kontekstu njihove važnosti u bilježenju političkih i društvenih činjenica o zajednici. To ovisi o tome na što se autor orijentira u svojoj studiji. Najčešće statističke metode koje se koriste prilikom uspoređivanja korpusa su: Yuleov koeficijent razlike, hi-kvadrat test, logaritamska funkcija vjerojatnosti (eng *log-likelihood function*), normaliziran omjer, Fischerov egzaktni test (Rayson, 2008). Malhberg (2007) kako je navedeno u radu Raysona (2008) smatra da tehnika ključnosti, primijenjena na sljedove riječi koje se ponavljaju, može pomoći u istraživanjima iz književne stilistike. Međutim, postoji jedan problem u korištenju ove tehnike, a to je da je dobiveni popis ključnih riječi preopsežan da bi ga se analiziralo. Zato Berber Sardhina (1999) kako je navedeno u radu Raysona (2008) predlaže da se reducira broj riječi - uzimanjem jednostavne većine ili odabiranjem značajnog podskupa pomoću hi-kvadrat testa.

Istraživanje korpusa ruskih romantičara i realista

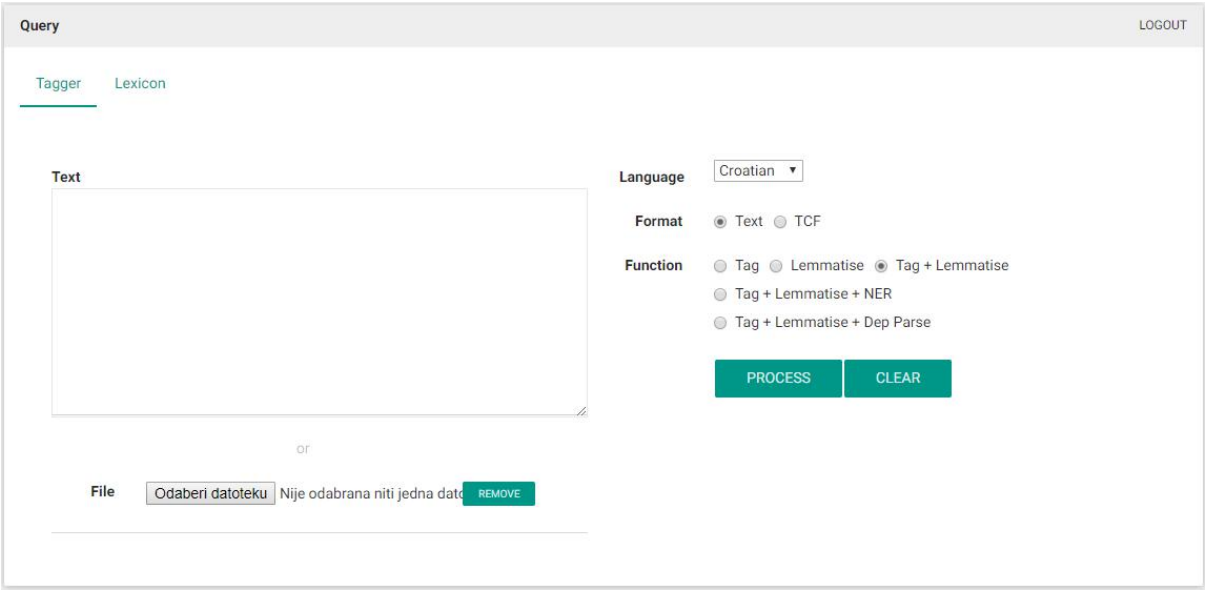
U sklopu ovog rada provedeno je istraživanje uporabe pridjeva u korpusu koji se sastojao od djela ruskih romantičara i realista prevedenih na hrvatski jezik. Korpus su činila zapravo dva potkorpusa - realistički i romantičarski. Realistički potkorpus činili su romani *Ana Karenjina* Lava Nikolajeviča Tolstoja (u prijevodu Martina Lovrenčevića) i *Zločin i kazna* Fjodora Mihajloviča Dostojevskog (u prijevodu Ise Velikanovića). Romantičarski potkorpus sastojao se od romana u stihovima, *Evgenija Onjegina* Aleksandra Sergejeviča Puškina (u prijevodu Ivana Trnskoga), i romana *Junak našeg doba* Mihaila Jurjeviča Ljermontova (u prijevodu Milana Bogdanovića). Izabrana su upravo ova djela jer predstavljaju vrhunac ruske literature i reprezentativna su za razdoblja u kojima su nastala. Namjerno su sva djela, izuzev *Evgenija Onjegina*, proznog oblika zbog toga što pripadnost istoj vrsti nalaže primjenu sličnih načela strukturiranja djela. Lirske pjesme i dramske vrste imaju specifični ustroj i stoga nisu uzete u obzir, a i činjenica je da je roman kao književna vrsta prevladavao u realizmu, dok je ruski romantizam ostao poznat upravo po ovim prije spomenutim djelima, iako je Puškin bio vrstan liričar. Glavna hipoteza bila je da će se uporaba pridjeva razlikovati u analiziranim potkorpusima budući da romantizam i realizam počivaju na, mogli bismo reći, posve oprečnim poetikama i načinima izražavanja, kao i tematskim preokupacijama. Za istraživanje se koristio programski jezik *Python* (verzija 3.6.0) te POS označivač za hrvatski jezik koji je dostupan na internetskoj stranici CLARIN.SI[5]. Svi tekstovi preuzeti su s internetske stranice *eLektire*[6], u internetskom formatu, kako bi se mogli spremiti u tekstualne datoteke koje su korištene za daljnji rad.

CLARIN.SI je nastao kao projekt unutar okvira europskog konzorcija CLARIN ERIC (*European Research Infrastructures Consortium*), čiji je cilj pružiti znanstvenicima s područja društvenih i humanističkih znanosti tehnologije, izvore i alate kako bi mogli provoditi svoja istraživanja (CLARIN Slovenia). Ovaj projekt prvenstveno je namijenjen obradi slovenskog i sličnih jezika, kao što su hrvatski i srpski jezik. Na internetskoj stranici dostupna je *online* usluga obrade tekstualnih podataka kojom se tekstovi mogu lematizirati, tokenizirati, sintaksno analizirati (parsirati), morfosintaktički označavati. Na slici 3 može se vidjeti prikaz sučelja usluge. Djela nisu jednakog broja pojavnica, stoga su u nastavku izneseni podaci o veličinama korpusa, potkorpusa i pojedinačnih književnih djela, s uključenom interpunkcijom:

5 Dostupno na: http://www.clarin.si/info/about/

6 Dostupno na: https://lektire.skole.hr/

- Ukupan broj pojavnica u korpusu: 686260
- Broj pojavnica u romantičarskom potkorpusu: 86650
- Broj pojavnica u realističkom potkorpusu: 599610
- Broj pojavnica u *Evgeniju Onjeginu*: 31516
- Broj pojavnica u *Junaku našeg doba*: 55134
- Broj pojavnica u *Zločinu i kazni*: 229328
- Broj pojavnica u *Ani Karenjinoj*: 370282



Slika 3. Snimka sučelja internetske usluge označavanja CLARIN.SI-a

Prethodno navedeni alati kojima je moguće analizirati tekstove razvijeni su zahvaljujući ReLDI[7]-ju (*Regional Linguistic Data Initiative*), mreži znanstvenika i istraživača koji se bave proučavanjem jezika. Inicijativa je nastala kao rezultat dvogodišnje institucionalne suradnje između švicarskih, hrvatskih i srpskih istraživača. Uglavnom se orijentiraju na analiziranje južnoslavenskih jezika, među kojima je i hrvatski jezik. Osim istraživačke djelatnosti, bave se i držanjem seminara te izradom tečajeva i potrebnih materijala za učenje na daljinu.

7 Dostupno na: https://reldi.spur.uzh.ch/hr-sr/

Metode

Istraživanje se temeljilo na primjeni statističkih metoda i metoda preuzetih iz obrade prirodnog jezika. Prvi korak bila je normalizacija teksta – izbacivanje izvantekstualnih podataka kao što su fusnote, zabilješke, nazivi djela, imena autora, rječnici manje poznatih riječi. Dakle, bilo je potrebno dobiti „čisti" književni tekst, bez bilo kakvih metapodataka jer bi oni utjecali na rezultate istraživanja. Zatim je uslijedila obrada tekstova pomoću CLARIN.SI internetske usluge tako da se tekst morfosintaktički označio i da se provela lematizacija. Svaka daljnja analiza vršila se nad tako obrađenim tekstom. Prije rada s dobivenim podacima trebalo je provjeriti točnost klasifikatora. To je učinjeno tako da se uzeo jedan segment teksta i ručno ga se označavalo. Budući da je istraživanje bilo fokusirano na pridjeve kao vrste riječi, ručno označavanje nije išlo u dubinsku morfosintaktičku analizu, nego se samo provjeravalo jesu li vrste riječi dobro označene (nije se gledalo je li dobro označen, primjerice rod, broj ili padež). Pomoću programskog jezika *Python* napravljena je frekvencijska distribucija riječi i interpunkcija za svaki tekst, s obzirom na oznake koje se koriste prilikom označavanja tekstova na hrvatskom jeziku:

V - glagol

A - pridjev

N - imenica

P - zamjenica

R - prilog

C - veznik

S - prijedlog

M - broj

Q - čestica

I - uzvik

Y - kratica

X - ostali znakovi

Z – interpunkcija (MULTEXT-East Morphosyntactic Specifications, Version 5 (draft)[8]).

8Dostupno na: http://nl.ijs.si/ME/V5/msd/html/msd-hr.html

Nakon toga izvukli su se samo pridjevi kako bi se vidjelo koji se najčešće pojavljuju u pojedinom tekstu. Primjer programskog koda napisanog u *Python*-u može se vidjeti u prilogu 3.

Rezultati

Prvi rezultati koji su bili dobiveni odnosili su se na provjeravanje točnosti samog klasifikatora. Na internetskoj stranici CLARIN.SI-a piše podatak da točnost za hrvatski jezik iznosi 92.53%. Kad se ručno označavalo prvih 500 pojavnica (uključujući interpunkciju) *Evgenija Onjegina*, došlo se do točnosti od 87.2%. Prilikom proučavanja podataka vidjelo se da klasifikator nije imao nijednu pogrješku prilikom označavanja interpunkcijskih znakova, stoga se izračunala i točnost kad bi se zanemarila interpunkcija. Točnost je tada iznosila 84.2%, dakle bila je nešto manja od one dobivene uz interpunkcijske znakove. Od izbrojene 61 pogreške njih 10 se odnosilo na krivo označavanje pridjeva, koji su izneseni u tablici 1:

Tablica 1. Usporedba ručnog i računalnog označavanja pridjeva

pojavnica	**ručna oznaka**	**ručna lema**	**oznaka klasifikatora**	**lema klasifikatora**
žive	A	živ	V	živjeti
polusmiješna	A	polusmiješan	N	polusmiješna
narodnjega	A	narodan	N	narodnjega
drugi	A	drugi	M	drugi
suho	A	suh	R	suho
stežne	A	stežan	N	stežna
bijeli	A	bijel	V	bijeliti
franceskoj	A	franceski	P	franceskoj
svijem	A	sav	V	sviti
lahka	A	lahki	P	lahka

Ako gledamo u obrnutom smjeru, koliko pridjeva je klasifikator detektirao a da se ne radi o pridjevima nego o nekoj drugoj vrsti riječi, dolazi se do brojke od 12 pogrešaka. Od 61 ukupne pogreške njih 22 je u nekoj poveznici s označavanjem pridjeva. Frekvencijskom distribucijom prikazuje se koliko se puta pojavila određena vrijednost, u ovom slučaju koliko se puta pojavila pojedina vrsta riječi. Izražava se cjelobrojno, a ne postotkom. U tablici 2 mogu se vidjeti frekvencijske distribucije riječi i interpunkcije za svaki tekst.

Tablica 2. Frekvencijske distribucije vrsta riječi i interpunkcije

	Evgenij Onjegin	*Junak našeg doba*	*Zločin i kazna*	*Ana Karenjina*
glagoli	4422	11838	45541	77475
imenice	7546	8945	33123	62494
prijedlozi	1730	3435	12690	23449
prilozi	1364	3660	17096	24184
pridjevi	3131	3642	13898	25218
zamjenice	3215	7284	28471	49032
brojevi	470	554	1886	2635
veznici	1260	4617	21388	32904
uzvici	52	98	480	447
čestice	666	1079	6217	7722
kratice	40	9	14	64
ostali znakovi	25	27	28	187
interpunkcija	7595	9946	48496	64471

Morfosintaktički označivač trenirao se na korpusu tekstova koji su dio hrvatskog jezičnog korpusa, te hrWaC-a, hrvatskog internetskog korpusa, i hrvatskog nacionalnog korpusa. Na internetu su dostupni podaci o prva dva korpusa te se pomoću njih izračunala relativna frekvencijska distribucija vrsta riječi, kako bi se rezultati usporedili s korpusom tekstova koji

su korišteni prilikom ovog istraživanja. U *Evgeniju Onjeginu* najveća razlika u postotcima vidljiva je u raspodjeli interpunkcijskih znakova: u djelu je njihova relativna frekvencijska distribucija 24.1%, a u hrWaC je 11.89%, odnosno 14.29% u hrvatskom jezičnom korpusu. U *Junaku našeg doba,* osim kod interpunkcije, zamjetne su razlike u zastupljenosti glagola (21.47% u djelu, 16.05% u hrWaC-u i 14.82% u jezičnoj riznici), imenica (16.22% u djelu, 26% u hrWaC-u, 27.91% u jezičnoj riznici) i zamjenica (13.21% u djelu, 8.2% u hrWaC-u i 6.88% u jezičnoj riznici). U romanu *Zločin i kazna* najveće su razlike ponovno u interpunkciji (mogli bismo reći da je to lajtmotiv svih analiziranih tekstova), zatim u imenicama (14.44% je relativna frekvencija u romanu, postoci za hrWaC i jezičnu riznicu identični su onima maloprije navedenima) i zamjenicama (12.41%). Isti slučaj je i kod *Ane Karenjine*: relativna frekvencijska distribucija imenica iznosi 16.88%, zamjenica 13.24%, interpunkcijskih znakova 17.41%, a glagola 20.92%. Što se tiče pridjeva, najveća relativna frekvencijska distribucija iznosi 9.93% i odnosi se na *Evgenija Onjegina*, dok ostala djela imaju vrlo sličnu relativnu frekvencijsku distribuciju za pridjeve: *Junak našeg doba* - 6.61%, *Zločin i kazna* - 6.06%, *Ana Karenjina* - 6.81%. Detaljnije informacije o frekvencijama pojedinih korpusa mogu se pronaći u prilogu 1 ovog rada.

Tablica 3. Broj pojedinih vrsta pridjeva u svakom djelu

	Evgenij Onjegin	***Junak našeg doba***	***Zločin i kazna***	***Ana Karenjina***
opisni pridjevi	2961	3389	12589	23160
posvojni pridjevi	53	32	379	587
glagolski pridjevi trpni	117	221	930	1471
ukupan broj pridjeva	3131	3642	13898	25218

Unutar pridjeva vršila se dodatna frekvencijska distribucija pojedinih vrsta pridjeva (opisni, posvojni i glagolski pridjevi trpni) što se može vidjeti u tablici 3. Klasifikator prepoznaje ove tri vrste pridjeva te su zato prikazani, iako u gramatici hrvatskog jezika postoji osnovna

podjela pridjeva na opisne, gradivne i posvojne. S obzirom na to da je analiza pretežito orijentirana na prisutnost pridjeva, iz svakog promatranog književnog djela izvučeni su podaci o 20 najučestalijih pridjeva koji se nalaze u pojedinačnom tekstu. Iz rezultata se vidi da se pridjevi "sam" i "sav" pojavljuju u sva četiri djela. Ovo su pridjevi koji su najučestaliji za pojedine tekstove (frekvencijske distribucije za svaki pridjev dostupne su u prilogu 2):

Evgenij Onjegin	***Junak našeg doba***	***Zločin i kazna***	***Ana Karenjina***
sav	sav	sam	sav
mlad	sam	sav	sam
sam	velik	isti	i
star	čitav	cijel	dobar
lijep	hladan	velik	isti
mio	moj	posljednji	nov
krasan	dobar	dobar	lijep
velik	mlad	vaš	njezin
nov	crn	nov	velik
ruski	čudan	neobičan	star
l	isti	Raskoljnjikov	njegov
tanak	njen	Svidrigajlov	mlad
sladak	prav	čudan	veseo
čudan	bijel	mali	moguć
živ	štaban	pijan	čitav
drag	uvjeren	osobit	potreban
prost	posljednji	jasan	Vronski
bijel	pun	neki	posljednji
hladan	lijep	prav	sretan
tih	smiješan	mlad	mali

Diskusija

Primjena klasifikatora na analiziranom korpusu književnih tekstova pokazala je da klasifikator vrši uspješno morfosintaktičko označavanje. Odstupanje u točnosti iznosi 5.33%, u odnosu na podatak o točnosti klasifikatora koji je dostupan na njegovoj internetskoj stranici (92.53%). Ne smije se zaboraviti da se ovdje radi o povijesnim tekstovima koji su nastali tijekom 19. i 20. stoljeća te se zato njihov jezik razlikuje od tipičnog suvremenog hrvatskog jezika. Za provjeravanje točnosti uzet je isječak teksta iz *Evgenija Onjegina* zato što to djelo po svojoj strukturi i leksiku najviše odstupa od svakodnevnih tekstova, te je stoga veća vjerojatnost da će se klasifikatoru "potkrasti" koja greška prilikom označavanja. Zbog 1nedostupnosti podataka o godinama prijevoda *Ane Karenjine* i *Zločina i kazne* ne može se kao jedan od razloga pogrešnog označavanja navesti činjenica da je prijevod *Evgenija Onjegina* najstariji od analiziranih tekstova, iako se ni ta mogućnost ne smije odbaciti.

Iz rezultata relativne frekvencijske distribucije može se zaključiti da se književni tekstovi, koji čine istraživani korpus, imaju različit udio imenica, glagola, zamjenica i interpunkcijskih znakova u odnosu na hrWaC i Hrvatsku jezičnu riznicu. *Ana Karenjina, Junak našeg doba i Zločin i kazna* više koriste glagole i zamjenice, a manje imenice, u odnosu na gore navedene korpuse. Kod *Evgenija Onjegina* ima drugačiju situaciju. Tamo ima puno manje glagola, zamjenica, pa i veznika, nego u ostalim romanima. S druge strane, imenice i pridjevi su više zastupljeni nego u ostalim tekstovima. Izgleda da je tome tako jer se radi o romanu u stihovima koji baštini dio lirskog utjecaja. Iako postoje događaji, oni nisu toliko u prvome planu i zato roman ima manje glagola. Veća količina pridjeva i imenica može se objasniti prisutnošću lirskog izraza koji teži opisivanju, izricanju osjećaja, nabrajanju. Nizak postotak veznika može se povezati s visokim postotkom interpunkcijskih znakova, koji su bitan element stihovanog načina pisanja. Ako se uspoređuju dva potkorpusa, realistički i romantičarski, onda se s obzirom na relativne frekvencijske distribucije *Junak našeg doba* može prije svrstati u realistički, nego romantičarski potkorpus. Mora se istaknuti da se do tog zaključka došlo bez ulaženja u sam sadržaj djela, dakle govori se isključivo na temelju statističkih podataka. Interpunkcija općenito u svim djelima ima puno veći udio nego u hrWaCu i Hrvatskoj jezičnoj riznici što je razumljivo jer se radi o književnim djelima koja se često služe složenim rečenicama, nizanjima, umetanjima, a to sve povećava uporabu zareza, dvotočja, trotočja, zagrada i sl.

Ako se promatraju frekvencijske distribucije pojedine vrste pridjeva (opisni, posvojni, glagolski pridjev trpni), uočava se ujednačenost u omjerima unutar svih tekstova. Najviše ima

opisnih pridjeva, zatim glagolskih pridjeva trpnih i na kraju posvojnih pridjeva. Iz tog se može iščitati zajedničko obilježje realizma i romantizma - težnja za opisivanjem, bilo fikcionalnog svijeta u kojem se radnja odvija, likova koji ga nastanjuju, ili osjećaja i emocionalnih stanja. Ono što je u najvećem fokusu interesa jest koji se zapravo pridjevi pojavljuju u pojedinom djelu. U prethodnom poglavlju napravljene su liste 20 najfrekventnijih glagola za svaki tekst. Primjećuje se da su pridjevi na vrhu liste vrlo slični, naime "sav" i "sam" nalaze se na prvom, drugom ili trećem mjestu kod svakog teksta.

U romantičarskom potkorpusu vidi se dosta sličnosti i ponavljanja pridjeva - i u *Evgeniju Onjeginu* i u *Junaku našeg doba* prisutni su sljedeći pridjevi: „sav", „sam", „lijep", „bijel", „velik", „hladan", „čudan", „mlad". Oba teksta imaju po jedan par međusobno oprečnih pridjeva: mlad i star u *Evgeniju Onjeginu* i bijel i crn u *Junaku našeg doba*. Na listama su se našle neke riječi koje nisu pridjevi (pa čak i slova): „l", „moj", „njen", „štaban" (vjerojatno se odnosi na riječ „štabni" ili se radi o zastarjeloj verziji riječi).

U realističkom potkorpusu ponovno se vide sličnosti. Pridjevi koji se javljaju i u *Zločinu i kazni* i u *Ani Karenjinoj* su: „sam", „sav", „isti", „nov", „velik", „posljednji", „dobar", „mali", „mlad". Ono što je osobito za ove tekstove jest da se u oba imena likova (Raskoljnikov, Svidrigajlov i Vronski) označavaju kao pridjevi. Razlog tome je u nastavcima -ov i -ski koji su tipični za pridjeve u hrvatskom jeziku, a ne za vlastita imena. Također se javljaju parovi pridjeva: „velik" - „mali" (*Zločin i kazna*) i „star" - „mlad" (*Ana Karenjina*). Zanimljivo je da se pridjev čudan" javlja u svim djelima, osim u *Ani Karenjinoj*. Zapravo velik broj pridjeva zajednički je svim četirima tekstovima iako pripadaju drugim književnim razdobljima. To bi se moglo shvatiti kao jedan od dokaza teorije koju često ističu teoretičari književnosti a to je da se između književnih razdoblja ne mogu povući jasne granice, utjecaji su uvijek prisutni i za nijedno djelo velike književne vrijednosti ne može se reći dar je, primjerice, tipično realističko ili romantičarsko djelo.

Koji su bili problemi prilikom provođenja istraživanja? Prvi problem javio se kod samog korištenja klasifikatora. Naime, realistički tekstovi puno su opsežniji od romantičarskih i kad su se trebali označavati, jednostavno su bili preveliki i klasifikator je zablokirao. Stoga su se ta djela morala podijeliti u više manjih tekstualnih datoteka koje su se zatim označile i nakon tog procesa opet spojile u jednu tekstualnu datoteku koja se koristila u daljnjoj analizi. Prilikom pokretanja *Python* programskog koda za svaki tekst uočilo se da je upravo zbog veličine datoteka *Ani Karenjinoj* i *Zločinu i kazni* trebalo više vremena dok bi sučelje Pythona izbacilo podatke. To je posebno bilo vidljivo u dijelu kad se izvlači 20 najfrekventnijih

pridjeva. Kod ručnog označavanja bilo je problema s definiranjem vrsta nekih riječi jer se radilo o zastarjelicama (npr. „priljem”, „ponevju”, „stežne”, „zgolje”).

Rezultati su pokazali da postoje neke razlike u uporabi pridjeva između dva potkorpusa, ali ne onako drastične kako se očekivalo. Mogući odgovor zašto je tome tako mogao bi ležati u činjenici da ono što književna djela diferencira jedno od drugog jesu jezične pojave koje nisu toliko brojčano česte, tj. one koje su specifične baš za to djelo. U ovom konkretnom slučaju to bi značilo da bi valjalo proučiti pridjeve koji su po rangu frekventnosti negdje u sredini - nisu toliko rijetki da bi ih se moglo smatrati iznimkama, a ni toliko učestali da bi se njihova pojava mogla pripisati općenitom načinu strukturiranja i upotrebe jezika u književnoj umjetnosti.

Zaključak

Suvremeno se društvo kreće prema sve većoj interdisciplinarnosti, suradnji različitih, naizgled nespojivih znanstvenih područja. Digitalna humanistika jedno je od takvih polja. Rad se bavio računalnom obradom književnih tekstova pozivajući se na podatke i trendove koje su razvili digitalni humanisti. Spomenuta interdisciplinarnost najbolje se ogleda u metodama koje se koriste prilikom tekstualnih analiza - statistički alati, metode iz obrade prirodnog jezika itd.

Kvantitativni pristup ima svoje pobornike, ali i protivnike. Važno je saslušati što obje strane imaju za reći i kako mogu pridonijeti poboljšanju i daljnjem razvoju praksi unutar digitalne humanistike. Primjeri konkretnih primjena metoda u računalnoj obradi književnih tekstova, kao što je kvantitativna tematska analiza, pripisivanje autorstva, stilometrija i analiza ključnih pojmova, pokazuju da empirijski podaci mogu biti od koristi u pokušajima tumačenja književnih djela.

Ono što se u prvom dijelu rada teoretski iznijelo pokušalo se oprimjeriti na analiziranju korpusa ruskih romantičara i realista. Istraživanje se orijentiralo na upotrebu isključivo kvantitativnih alata, kao što je frekvencijska distribucija. Služilo se i metodama iz obrade prirodnog jezika (POS označavanje, lematizacija). Proces provođenja istraživanja potvrdio je ono što je rečeno u teoretskom dijelu rada - kvantitativni pristup može biti vrlo koristan u tumačenju književnosti, ali jedino ako postoji ljudska strana koja će dobivene podatke znati protumačiti.

Iako počeci računalnog proučavanja povijesnih tekstova (stoga i književnih) potječu iz 50-ih godina prošlog stoljeća, prostora za napredak još uvijek ima. Mora se raditi na razvoju alata i metoda, posebno za tekstove koji nisu napisani na velikim svjetskim jezicima kao što je engleski. Istraživanja se uglavnom provode na velikim kanonskim djelima, a ona manje „razvikana" se zanemaruju. Nužna je suradnja stručnjaka iz različitih područja. Međutim, bez obzira na sve prepreke s kojima se susreću digitalni humanisti, računalna analiza postaje sve više korištena kao nova vrsta pristupa književnim djelima koja može pružiti drugačiji pogled i kreirati dosad nepostavljena ili previđena pitanja.

Literatura

1. Bekavac, B., (2002.), Strojno obilježavanje hrvatskih tekstova – stanje i perspektive, *Suvremena lingvistika*, [Online], 53-54 (1-2), str. 173-182, <raspoloživo na: https://suvlin.ffzg.hr/index.php/hr/arhiva-casopisa/14-vol-53-54-no-1-2-2002/443-strojno-obiljezavanje-hrvatskih-tekstova-stanje-i-perspektive>, [pristupljeno 5.5.2019.].
2. CLARIN Slovenia, Common Language Resources and Technology Infrastructure, <http://www.clarin.si/info/about/>, [pristupljeno 2.5.2019.].
3. Computational Textual Analysis (2018.), <https://guides.temple.edu/corpusanalysis>, [pristupljeno 15.3.2019.].
4. Conway, M., (2009.), Mining a corpus of biographical texts using keywords, *Literary and Linguistic Computing*, [Online] 25 (1), <raspoloživo na: https://academic.oup.com/dsh/article/25/1/23/926040>, [pristupljeno 18.4.2019.].
5. Ćavar, D., Brozović Rončević, D., (2012.), Riznica: the Croatian language corpus, *Prace filologiczne*, 63, str. 51-65.
6. Daelemans, W., (2013.), Explanation in Computational Stylometry, [Internet], <raspoloživo na: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1322&rep=rep1&type=pdf >, [pristupljeno 16.4.2019.].
7. Digital Humanities Manifesto 2.0. [Internet], <raspoloživo na: http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf>, [pristupljeno: 14.4.2019.].
8. Eder, M., Rybicki, J., Kestemont, M., (2016.), Stylometry with R: A Package for Computational Text Analysis, *The R Journal*, [Online], 8/1. <raspoloživo na: https://journal.r-project.org/archive/2016/RJ-2016-007/RJ-2016-007.pdf>, [pristupljeno 17.4.2019.].
9. eLektire, https://lektire.skole.hr/, [pristupljeno 10.2.2019.].
10. Fischer-Starcke, B., (2009.), Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*: A corpus-stylistic analysis. *International Journal of Corpus Linguistics*, [Online], 14 (4). <raspoloživo na: https://www.ingentaconnect.com/content/jbp/ijcl/2009/00000014/00000004/art00003>, [pristupljeno 16.4.2019.].

11. Gomez Adorno, H., (2018.), Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts, *Computation y Sistemas*, [Online], 22 (1), <raspoloživo na: https://www.researchgate.net/publication/324201958_Stylometry-based_Approach_for_Detecting_Writing_Style_Changes_in_Literary_Texts>, [pristupljeno 16.4.2019.].
12. Grieve, J., (2007.), Quantitative Authorship Attribution: An Evaluation of Techniques, *Literary and Linguistic Computing*, [Online], 22 (3), <raspoloživo na: https://academic.oup.com/dsh/article/22/3/251/951481>, [pristupljeno 15.4.2019.].
13. Hammond, A., (2016), Quantitative Approaches to the Literary, U: *Literature in the Digital Age: An Introduction*. [Internet], str. 82-130, Cambridge: Cambridge University Press, <raspoloživo na: https://books.google.hr/books?id=IE6SCwAAQBAJ&printsec=frontcover&dq=literature+in+the+digital+age&hl=hr&sa=X#v=onepage&q&f=false>, [pristupljeno 10.4.2019.].
14. Hargrave, M., (2019.), Standard Deviation Definition. <https://www.investopedia.com/terms/s/standarddeviation.asp>, [pristupljeno 9.4.2019.].
15. Hoover, D. L, (2013.), Textual Analysis, U: Price, K. M., Siemens, R., ed., *Literary Studies in the Digital Age: An Evolving Anthology*. [Internet], <raspoloživo na: https://dlsanthology.mla.hcommons.org/textual-analysis/>, [pristupljeno 18.4.2019.].
16. Hoover, D. L, (2016.), Arguments, Evidence, and the Limits of Digital Literary Studies, U: *Debates in the Digital Humanities 2016*. [Internet], University of Minnesota Press, <http://dhdebates.gc.cuny.edu/debates?data=toc-open>, [pristupljeno 16.4.2019.].
17. Hoover, D. L. (2008.), Quantitative Analysis and Literary Studies, U: Schreibman, S., Siemens, R., ed., *A Companion to Digital Literary Studies*. [Internet], Oxford: Blackwell, <raspoloživo na: http://digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.id=0&brand=9781405148641_brand>, [pristupljeno 28.2.2019.].
18. Hrvatski jezični portal, <http://hjp.znanje.hr/>, [pristupljeno 11.4.2019.].
19. Hrvatski mrežni rječnik – MREŽNIK, <http://ihjj.hr/mreznik/>, [pristupljeno 11.4.2019.].

20. Huang, Liang, et. al., (2002), PCFG parsing for restricted Classical Chinese texts, U: Proceedings of the SIGHAN workshop on Chinese language processing (SIGHAN '02). [Internet], str. 1-6, Stroudsburg, PA, USA: Association for Computational Linguistics, <https://www.aclweb.org/anthology/W02-1806>, [pristupljeno 22.4.2019.].
21. Kerr, Sarah J., (2017.), When Computer Science Met Jane Austen and Edgeworth, *NPPSH Reflections*, [Online], 1, raspoloživo na: <http://mural.maynoothuniversity.ie/8298/1/NPPSH%202016%20Reflections%20-%20Kerr.pdf>, [pristupljeno 28.3.2019.].
22. Koltay, T., (2016.), Library and information science and the digital humanities: Perceived and real strengths and weaknesses, *Journal of Documentation*, [Online], 72 (4), str. 781-792, raspoloživo na: https://www.emeraldinsight.com/doi/abs/10.1108/JDOC-01-2016-0008>, [pristupljeno: 17.5.2019.].
23. Koltay, T., (2015.), The Digital Humanities and Information Science: Remarks on the Epistemologies, *KIIT Journal of Library and Information Management*, [Online], 2 (2), str. 110-120, raspoloživo na: <https://www.researchgate.net/publication/277749624_The_Digital_Humanities_and_Information_Science_Remarks_on_the_Epistemologies_KIIT_Journal_of_Library_and_Information_Management_Vol_2_No_2_110-120>, [pristupljeno: 17.5.2019.].
24. Konfic, L., (2017), Predstavljanje osnovnih pojmova digitalne humanistike u vezi s glazbom i muzikologijom, ili: Što je digitalna muzikologija?, *Arti musices*, [Online], 48(2), <raspoloživo na: https://hrcak.srce.hr/192634>, [pristupljeno 29.3.2019.].
25. Laramée, F. D., (2018.), Introduction to stylometry with Python. [Internet], <raspoloživo na: https://programminghistorian.org/en/lessons/introduction-to-stylometry-with-python>, [pristupljeno 2.5.2019.].
26. Lemmatization Approaches with Examples in Python. [Internet], <raspoloživo na: https://www.machinelearningplus.com/nlp/lemmatization-examples-python/>, [pristupljeno 17.4.2019.].
27. Lopina, V., (2012.), *Računalna obrada jezika i stila u pjesništvu A. B. Šimića*, Doktorska disertacija, Zagreb: V. Lopina.
28. Ljubešić, N., Klubička, F., (2014.), {bs, hr, sr} wac-web corpora of Bosnian, Croatian and Serbian, *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*.
29. Marche, S., (2012.), Literature Is not Data: Against Digital Humanities. [Internet], <raspoloživo na: https://lareviewofbooks.org/article/literature-is-not-data-against-digital-humanities/#!>, [pristupljeno 1.5.2019.].
30. MULTEXT-East Morphosyntactic Specifications, Version 5 (draft), <http://nl.ijs.si/ME/V5/msd/html/msd-hr.html>, [pristupljeno 1.4.2019.].

31. Nikolić, D., (2016.), Digitalna humanistika i nacionalna filologija: o mogućim implikacijama računalnog obrata, *Croatica*, [Online], 50(60), str. 75-87. <raspoloživo na: https://hrcak.srce.hr/174512>, [pristupljeno 30.3.2019.].
32. Osnove statistike. [Internet], < https://www.pmf.unizg.hr/_download/repository/PREDAVANJE7.pdf>, [pristupljeno 15.4.2019.].
33. Oxford Dictionaries, <https://en.oxforddictionaries.com/definition/digital_humanities>, [pristupljeno 17.4.2019.].
34. Petrović, B., Vranešević, D., (2015.), Kvantitativna raščlamba *Čudnovatih zgoda šegrta Hlapića* Ivane Brlić-Mažuranić, U: Majhut, B., Narančić Kovač, S., Lovrić, S., ed., Šegrt Hlapić – *od čudnovatog do čudesnog*, Zagreb: Slavonski Brod: Hrvatska udruga istraživača dječje književnosti; Matica hrvatska, Ogranak, str. 251-267.
35. Piotrowski, M., (2012.), *Natural language processing for historical texts*, University of Toronto: Graeme Hirst.
36. Rayson, P., (2008.), For key words to key semantic domains, *International Journal of Corpus Linguistics*, [Online], 13 (4), <raspoloživo na: https://www.jbe-platform.com/content/journals/10.1075/ijcl.13.4.06ray>, [pristupljeno 16.4.2019.].
37. ReLDI, <http://clarin.si/services/web/>, [pristupljeno 29.4.2019.].
38. Roberts, Carl W., (2000.), A Conceptual Framework for Quantitative Text Analysis: On Joining Probabilities and Substantive Inferences about Texts, *Quality & Quantity*, [Online], 34 (3), <raspoloživo na: https://link.springer.com/content/pdf/10.1023%2FA%3A1004780007748.pdf>, [pristupljeno 17.4.2019.].
39. Stamatatos, E., (2008.), A Survey of Modern Authorship Attribution Methods. [Internet], <raspoloživo na: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1634&rep=rep1&type=pdf >, [pristupljeno 15.4.2019.g.
40. Stańczyk, U., Cyran, K. A.,(2007.), On employing elements of Rough Set Theory to stylometry analysis of literary texts, *International Journal of Applied Mathematics and Informatics*, [Online], 1 (4), <raspoloživo na: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.2815&rep=rep1&type=pdf >, [pristupljeno 16.4.2019.].
41. Text Analysis Resources (2016.). [Internet], <raspoloživo na: https://digitalhumanities.berkeley.edu/resources/text-analysis-resources>, [pristupljeno 15.3.2019.].

42. The meaning of statistics and digital humanities (2012.). [Internet], <raspoloživo na: http://lab.softwarestudies.com/2012/11/the-meaning-of-statistics-and-digital.html>, [pristupljeno 2.5.2019.].
43. The meaning of statistics and digital humanities. [Internet], <http://lab.softwarestudies.com/2012/11/the-meaning-of-statistics-and-digital.html>, [pristupljeno 10.4.2019.].
44. Tomić, M., (2015.), Digitalna humanistika kao izazov: promjena paradigme istraživanja u humanistici i praksa digitalizacije, prezentacija s *Petog festivala hrvatskih digitalizacijskih projekata*, [Online], <raspoloživo na: http://dfest.nsk.hr/2015/wp-content/themes/boilerplate/2015/prezentacije/Tomic_Marijana.pdf>, [pristupljeno 30.3.2019.].
45. Tutorial: Pearson's Chi-square Test for Independence. [Internet], <https://www.ling.upenn.edu/~clight/chisquared.htm>, [pristupljeno 9.4.2019.].
46. Underwood, T., (2016.), Distant Reading and Recent Intellectual History U: *Debates in the Digital Humanities 2016*. [Internet], University of Minnesota Press. <http://dhdebates.gc.cuny.edu/debates?data=toc-open>, [pristupljeno 16.4.2019.].
47. Vježbe iz statistike (2017.), Zdravstveno veleučilište u Zagrebu, [Internet], <raspoloživo na: https://ldap.zvu.hr/~oliverap/VjezbeIzStatistike/7_T-test%20vje%C5%BEbe.pdf>, [pristupljeno 2.5.2019.].
48. Wilkens, M., (2012.), Canons, Close Reading, and the Evolution of Method U: *Debates in the Digital Humanities 2012*. [Internet], University of Minnesota Press, <http://dhdebates.gc.cuny.edu/debates?data=toc-open>, [pristupljeno 16.4.2019.].

Popis slika

Popis tablica

Prilozi

Prilog 1 - Relativna frekvencijska distribucija riječi unutar korpusa

	Evgenij Onjegin	*Junak našeg doba*	*Zločin i kazna*	*Ana Karenjina*	*hrWaC*	*Hrvatska jezična riznica*
glagoli	14.03%	21.47%	19.86%	20.92%	16.05%	14.82%
imenice	23.94%	16.22%	14.44%	16.88%	26.00%	27.91%
pridjevi	9.93%	6.61%	6.06%	6.81%	9.41%	10.16%
zamjenice	10.20%	13.21%	12.41%	13.24%	8.20%	6.88%
brojevi	1.49%	1.00%	0.82%	0.71%	2.52%	2.97%
veznici	4.00%	8.37%	9.33%	8.89%	7.75%	6.80%
uzvici	0.16%	0.18%	0.21%	0.12%	0.11%	0.04%
čestice	2.11%	1.96%	2.71%	2.09%	1.68%	1.42%
kratice	0.13%	0.02%	0.006%	0.02%	0.41%	0.35%

ostali znakovi	0.08%	0.05%	0.01%	0.05%	1.78%	0.61%
interpunkcija	24.1%	18.04%	21.15%	17.41%	11.89%	14.29%
prijedlozi	5.49%	6.23%	5.53%	6.33%	8.47%	8.92%
prilozi	4.33%	6.64%	7.45%	6.53%	5.63%	6.88%

Prilog 2 – Frekvencijska distribucija najučestalijih pridjeva

Evgenij Onjegin	***Junak našeg doba***	***Zločin i kazna***	***Ana Karenjina***
sav - 153	sav – 253	sam – 1714	sav – 2058
mlad - 63	sam – 70	sav – 573	sam – 672
sam - 54	velik – 49	isti – 195	i – 493
star - 35	čitav – 42	cijel – 169	dobar – 387
lijep - 32	hladan – 35	velik – 128	isti – 341
mio - 28	moj – 32	posljednji – 118	nov – 323
krasan - 27	dobar – 30	dobar – 109	lijep – 277
velik - 26	mlad – 28	vaš – 105	njezin – 262
nov - 25	crn – 27	nov – 103	velik – 247
ruski - 25	čudan – 27	neobičan – 91	star – 229
1 - 25	isti – 26	Raskoljnjikov – 90	njegov – 221
tanak -24	njen – 24	Svidrigajlov – 90	mlad – 195
sladak -24	prav – 24	čudan – 87	veseo – 186
čudan -23	bijel – 24	mali – 86	moguć – 186
živ -23	štaban - 23	pijan – 73	čitav – 161
drag -22	uvjeren- 22	osobit – 70	potreban – 156

prost - 21	posljednji – 21	jasan – 68	Vronski – 147
bijel -21	pun – 21	neki – 67	posljednji – 136
hladan -20	lijep – 21	prav – 67	sretan – 128
tih -20	smiješan - 20	mlad - 66	mali - 126

Prilog 3 – Primjer programskog koda za frekvencijsku distribuciju i 20 najučestalijih pridjeva u *Evgeniju Onjeginu*

```
evgenij= open('evgenij_clarin_punkt.txt','r', encoding='utf8')
red= evgenij.readlines()
pridjevi=0
imenice=0
glagoli=0
zamjenice=0
prilozi=0
veznici=0
prijedlozi=0
brojevi=0
cestice=0
uzvici=0
kratice=0
residual=0
interpunkcija=0
general=0
possessive=0
participle=0

lista_oznaka=[]
leme_i_oznake=[]
```

```
for element in red:
    lista= element.split('\t')
    lista_oznaka.append(lista[2])

    leme_i_oznake.append(lista[3])
    leme_i_oznake.append(lista[2])

for oznaka in lista_oznaka:
    if oznaka.startswith('Ag'):
        pridjevi+=1
        general+=1
    elif oznaka.startswith('As'):
        pridjevi+=1
        possessive+=1
    elif oznaka.startswith('Ap'):
        pridjevi+=1
        participle+=1
    elif oznaka.startswith('N'):
        imenice+=1
    elif oznaka.startswith('V'):
        glagoli+=1
    elif oznaka.startswith('P'):
        zamjenice+=1
    elif oznaka.startswith('R'):
        prilozi+=1
    elif oznaka.startswith('C'):
        veznici+=1
    elif oznaka.startswith('S'):
        prijedlozi+=1
    elif oznaka.startswith('M'):
        brojevi+=1
```

```
    elif oznaka.startswith('Q'):
        cestice+=1
    elif oznaka.startswith('I'):
        uzvici+=1
    elif oznaka.startswith('Y'):
        kratice+=1
    elif oznaka.startswith('X'):
        residual+=1
    else:
        interpunkcija+=1

parovi = list()
while(leme_i_oznake):
    a = leme_i_oznake.pop(0); b = leme_i_oznake.pop(0)
    parovi.append((a,b))

pridjevi=[]
for par in parovi:
    if par[1].startswith('A'):
        pridjevi.append(par[0])

brojac = {}
for pridjev in pridjevi:
    brojac[pridjev] = brojac.get(pridjev, 0) + 1

from funkcije import *
sortirano= sortiraj_distr(brojac)
print(sortirano[:20])
```

Računalna obrada književnih tekstova na primjeru analize korpusa ruskih romantičara i realista

Sažetak

U radu će se objasniti što znači pojam računalne obrade književnih tekstova te u kojim područjima se upotrebljava. Iznijet će se glavne prednosti i nedostaci analiziranja književnih ostvarenja pomoću računalnih programa. Prikazat će se najpoznatije metode koje su svojstvene kvantitativnom pristupu književnim djelima (kvantitativna tematska analiza, pripisivanje autorstva, stilometrija, analiza ključnih riječi). Poseban naglasak će se staviti na analiziranje vokabulara pisaca, u ovom slučaju kako se koriste pridjevi u djelima ruskih romantičara Puškina i Ljermontova, te ruskih realista Dostojevskog i Tolstoja. U istraživanju će se upotrebljavati hrvatski prijevodi spomenutih djela, dostupni na internetskoj stranici *eLektire*. Ispitivanje će biti fokusirano na to kolike su razlike između pojedinih djela i potkorpusa (ruskog i romantičarskog). Pri tome će se primijeniti alat za označavanje vrsta riječi (tzv. *POS tagger*) kako bi se dobili označeni pridjevi. Dobiveni rezultati će se analizirati te će se pomoću njih pokušati izvesti zaključak o korisnosti kvantitativne analize književnih tekstova.

Ključne riječi: digitalna humanistika, računalna obrada, kvantitativne metode, književni tekst, POS označavanje

Computational analysis of literary texts using the example of the analysis of Russian literary corpus

Summary

The paper will analyze the concept of computational analysis of literary texts and investigate which scientific fields use this type of analysis. We will present the main advantages and disadvantages of analyzing literary works with computer programs. The most popular methods of quantitative analysis of literature will be shown (quantitative thematic analysis, authorship attribution, stylometry, keyword analysis). Special emphasis will be placed on analyzing the vocabulary of writers, in this case the usage of adjectives in the works of Russian romantics Pushkin and Lermontov, and Russian realists Dostoevsky and Tolstoy. In particular, we will investigate whether there are any differences or similarities in the way adjectives are used in subcorpora (romantics vs. realists). A word tagging tool (so-called "POS tagger") will be applied in order to obtain the tagged adjectives. The corpus will be made up of Croatian translations available on the *eLektire* platform. The obtained results will be analyzed and will be used to test the usability of quantitative analysis of literary texts.

Key words: digital humanities, computational analysis, quantitative methods, literary text, POS tagging