Izgradnja povijesnog i suvremenog putopisnog korpusa i računalna usporedna analiza jezika

Živičnjak, Klara

Undergraduate thesis / Završni rad

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Zagreb, Faculty of Humanities and Social Sciences / Sveučilište u Zagrebu, Filozofski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:131:801433

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-07-27**

Sveučilište u Zagrebu
Filozofski fakultet
University of Zagreb
Faculty of Humanities
and Social Sciences

Repository / Repozitorij:

ODRAZ - open repository of the University of Zagreb
Faculty of Humanities and Social Sciences

zir.nsk.hr

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2020./2021.

Klara Živičnjak

Izgradnja povijesnog i suvremenog putopisnog korpusa i računalna usporedna analiza jezika

Završni rad

Mentor: dr. sc. Petra Bago, doc.

Zagreb, rujan 2021.

Izjava o akademskoj čestitosti

Izjavljujem 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.

SADRŽAJ

1. UVOD

1.1 Kontekst rada

Tehnologija već godinama zahvaća gotovo sva znanstvena područja u sve većoj mjeri pa je tako danas ključni dio i društvenih i humanističkih znanosti. Kroz ovaj će se rad pokazati kako je tehnologija doprla i do jezika, odnosno do jezične analize. Danas postoje mnogi alati koji pomažu lingvistima, filolozima i ostalim znanstvenicima koji se bave jezikom da na što brži, jednostavniji, ali i precizniji način izvrše svoja istraživanja. Za ovaj je rad izabran alat Sketch Engine kao pomoć pri analizi dvaju putopisnih korpusa. Za ovakvo sam se istraživanje odlučila zbog toga što sam studentica informacijskih znanosti i kroatistike te mi je cilj bio ujediniti znanja iz tih dvaju područja.

U uvodna dva potpoglavlja govorit će se o razvoju digitalne humanistike te o Sketch Enginu kao leksikografskom alatu, alatu za analizu i podučavanje jezika. Nakon toga će se u poglavlju Prikupljeni podaci opisati način sastavljanja korpusa te će se sastavljeni korpusi i opisati. U poglavlju Metodologija navedene su hipoteze te metodologija kojom će se postavljene hipoteze provjeravati. Rezultati istraživanja govore o gotovim rezultatima do kojih se došlo provjeravanjem hipoteza, a u poglavlju Diskusija pisalo se o značenjima rezultata te zašto je došao do takvih rezultata.

Za provedbu istraživanja odabrana su dva putopisna teksta koja su u Sketch Enginu pretvorena u korpuse, odabrani su tekstovi pisani u različitim vremenskim razdobljima te drugačijim stilom. Cilj je rada provjeriti hipoteze kroz koje se postavlja pitanje o tome koji korpus ima razvedeniji jezik, koliko je lematizator Sketch Engina precizan te kako izgleda tipična terminologija dvaju navedenih putopisnih korpusa.

1.2 Digitalna humanistika

Digitalna humanistika znanstveno je područje u kojem se tehnologija preklapa s humanističkim znanostima, odnosno humanističke se znanosti preklapaju s informacijskim znanostima. Nastaje sredinom 20. stoljeća, nakon pojave prvih računala. Nastala je s

namjerom olakšanog proučavanja humanističkih tekstova pomoću digitalnih alata. Zapravo se svako korištenje računala u svrhu istraživanja u području humanistike smatra radom u digitalnoj humanistici. Ona donosi različite alate i metode koji se koriste pri proučavanju raznih humanističkih znanosti. Dolazi do toga da tiskana riječ više nema glavnu ulogu u distribuciji znanja, tehnologija preuzima glavnu ulogu u takvoj distribuciji (Bosančić, 2011). Proizvodnjom i korištenjem novih tehnika dolazi do nove vrste istraživanja i proučavanja, a to se prvenstveno odnosi na korištenje digitalnih jezičnih alata za analizu, proučavanje i poučavanje jezika. No povećano korištenje tih metoda utječe na kulturno nasljeđe i digitalnu kulturu. Više se bavi digitalizacijom naslijeđa, nego pojavom novih kulturnih oblika te dolazi do toga da interaktivnost u pojedinim slučajevima postaje važnija od sadržaja. Prepoznatljiva je značajka digitalne humanistike ta da ona uglavnom istodobno koristi tehnologiju u istraživanju humanističkih znanosti te podvrgava tehnologiju humanističkom ispitivanju.

Digitalna humanistika kao znanost obuhvaća raznolika područja. Uključuje materijale koji originalno nisu bili digitalni nego su digitalizirani, a isto tako i originalno digitalne materijale. Ova disciplina kombinira metodologije iz tradicionalnih humanističkih disciplina kao što su retorika, povijest, filozofija, lingvistika, književnost, umjetnost, arheologija, glazba te metodologije iz društvenih znanosti s alatima koji se koriste u računarstvu, kao što su hipertaksa, hipermediji, vizualizacija podataka, pronalaženje podataka, rudarenje podacima, statistika, rudarenje tekstom, digitalno mapiranje i ostali tome slični alati. Ti se alati i tehnike primjenjuju na arhive i zbirke koje su količinski znatno prevelike za pojedinog znanstvenika ili za skupinu znanstvenika. Metode koje se koriste omogućuju stvaranje velikih ambicioznih projekata s velikim interdisciplinarnim timovima koji se okupljaju za rad na opsežnim i složenim projektima (Berry, 2019).

Povijesno, digitalna humanistička znanost potječe iz područja humanističkog računarstva čiji počeci sežu u četrdesete i pedesete godine prošlog stoljeća. Značajan je bio pionirski rad jezuitskog učenjaka Roberta Buse koji je svoj rad započeo 1946., i profesorice engleskog jezika Josephine Miles koja je sa svojim radom u području digitalne humanistike započela ranih 1950-ih. Proučavala je kvalitativne i računalne metode u humanističkim znanostima te je provodila mnoge istraživačke projekte. U suradnji s IBM-om, Busa i njegov tim stvorili su računalno generiranu podudarnost ili konkordancije sa spisima Tome Akvinskog, a taj je rad poznat pod nazivom Index Thomisticus. Drugi su znanstvenici počeli koristiti računala za automatizaciju zadataka poput pretraživanja riječi, razvrstavanja i brojanja što je bilo mnogo brže od obrade podataka iz ručno napisanih tekstova ili od podataka otkucanih s indeksnih

kartica. U desetljećima koja su slijedila, arheolozi, povjesničari, književni znanstvenici i širok spektar istraživača humanističkih znanosti u drugim disciplinama, primjenjivali su nove računalne metode za transformiranje humanističke znanosti.

Računala su se vrlo rano u svojoj povijesti počela koristit u radu u humanističkim područjima. Prvenstveno je olakšanje i prednost u radu humanističkih znanosti dostupnost digitalnih datoteka s gotovo bilo kojeg mjesta na svijetu. Takav je pristup informacijama imao velik utjecaj na olakšavanje provođenja istraživanja u humanističkim znanostima. Fokus je ove discipline u novije vrijeme usmjeren na područja koja se povezuju sa strojnim učenjem i umjetnom inteligencijom. Digitalni humanisti donose novu ideju o tome što može biti humanistički istraživački projekt, daju nove načine viđenja prošlih i sadašnjih kultura. (Berry, 2019).

Nove zbirke povijesnih, književnih ili drugih umjetničkih artefakata često su u posljednje vrijeme dostupne na internetu ili u digitalnim bazama podataka, a materijal koji sadrže dostupniji je nego što je bio slučaj s tiskom. Takvi izvori povećavaju sposobnost humanista da kombiniraju skupove podataka, društvene medije, zvukove, internetske i slikovne arhive te da se s lakoćom njima koriste. Ključno je također bilo i stvaranje softvera za analizu, razumijevanje i transformaciju takvih digitalnih materijala. No sa sve većom rasprostranjenošću digitalnih alata u našim životima dolazi se i do pokojeg negativnog zaključka. Pojavile su se nove zabrinutosti oko sposobnosti koje te tehnologije moraju imati. Velike tvrtke poput *Facebooka, Googlea, Amazona i Netflixa* redovno prikupljaju i koriste osobne podatke potrošača na vrlo nametljive načine što je javno zabrinjavajuće. Ali i u ovom slučaju nam je korisna digitalna humanistika koja svojom stručnošću u mnogim područjima znanja može pomoći u razumijevanju ovih problema te donekle može pružiti uvide u takvu politiku. (Berry, 2019).

Digitalne humanističke znanosti bile su vrlo dobre u prenošenju digitalnih tehnika i metoda u humanističke znanosti i time su postavile temelje za zlatno doba humanističkih istraživanja u 21. stoljeću. U digitalnom dobu humanističke znanosti moraju više nego ikad komunicirati humanističke vrijednosti i vlastiti doprinos javnoj kulturi. Humanističke znanosti i dalje postavljaju važno pitanje „Što je život vrijedan življenja?“, a digitalna je humanistika dio te tradicije, ona nam pomaže da proširujemo naše razumijevanje ljudske kulture u digitalnom svijetu.

1.3 Sketch Engine

Sketch Engine je korpusni alat i alat za analizu teksta koji se vrlo često koristi u leksikografiji. Razvio ga je Lexical Computing Limited, a s radom kreće 2004. godine. Svrha Sketch Engina je omogućiti ljudima koji proučavaju ponašanje jezika da pretražuju i analiziraju velike količine teksta uz pomoć mnogih opcija koje alat nudi. Alat se također koristi i za podučavanje jezika. Trenutno sadrži korpuse na više od 90 jezika, a hrvatski je jezik uvršten u alat 2013. godine ulaskom Hrvatske u Europsku uniju, čime je i hrvatski postao službenim jezikom u Uniji (Baisa, Michelfeit, Medved, Jakubiček, 2016).

Sketch Engine ime je dobio prema riječi *sketch* (eng.), odnosno skica. Skica se naziva jedna stranica gramatičkog sažetka te kolokacije odrađene riječi. To je jedna od osnovnih značajki alata, uz brojne druge koje sadrži, a ona se naziva *word sketch* (eng.). Tu su rezultati organizirani u kategorije koje se nazivaju gramatički odnosi. Ako se koncentriramo na glagol, to mogu biti riječi koje služe kao objekt glagola, kao subjekt glagola, riječi koje modificiraju druge riječi i tako dalje. Te riječi uključene u analizu su definirane prema pravilima koja su napisana u gramatici skice (eng. *sketch).*

Kao još jedna značajka ističu se kolokacije. Kolokacija je slijed ili kombinacija riječi koje se zajedno javljaju češće nego što bi se očekivalo. Postoje slabe i jake kolokacije. Na primjer, *dobar fakultet.* Riječi *dobar* i *fakultet* mogu se kombinirati i s mnogo drugih riječi te je ovdje riječ o slaboj kolokaciji. S druge strane primjer jake kolokacije bio bi *Filozofski fakultet* jer je ta sintagma rjeđa i mnogo specifičnija.

Iduća je značajka *word sketch difference* (eng.) u kojoj se riječi uspoređuju putem njihovih kolokacija. Koristi se za uspoređivanje kontrastnih kolokacija, a za to su dostupne tri opcije. Prva je lema (eng. *lemma*), lema je riječ koja opisuje rječničku natuknicu, a u ovom se slučaju njome uspoređuje upotreba dviju lema putem njihovih kolokata. Druga se opcija naziva oblici riječi (eng. *word forms*). Ovdje se uspoređuje upotreba dvaju različitih oblika riječi iste leme putem njihovih kolokata. Treća je opcija podkorpus (eng. *subcorpora*). Ona uspoređuje upotrebu iste leme u dva različita podkorpusa istog korpusa putem njihovih kolokata.

Distribucijski tezaurus (eng. *distributional thesaurus)* je automatski proizvedeni tezaurus koji identificira riječi koje se javljaju u sličnom okruženju kao ciljana riječ. Automatski izrađen tezaurus dostupan je za svaku riječ u korpusu.

U opciji pretraživanja konkordanci (eng. *concordance search*) nailazimo na popis svih primjera pretraživanih riječi ili fraza koje se nalaze u korpusu koji pretražujemo. Obično se nalaze u KWIC (*key word in context)* formatu podudarnosti s riječju koja se pretražuje, a ona je istaknuta u središtu zaslona ili u pojedinim kontekstima s desna i s lijeva.

Alat za popis riječi (eng. *Wordlist tool)* generira popise učestalosti različitih skupina u koje se riječi svrstavaju, u prvu takvu skupinu ubrajamo imenice, glagole, pridjeve i druge vrste riječi, druga su skupina riječi koje počinju, završavaju ili sadrže određene znakove, a treće su oblici riječi, oznake, leme te drugi atributi. Također se u pretraživanju može tražiti i kombinacija triju gore navedenih mogućnosti. Na popisu riječi mogu se prikazati tri različite mjere frekvencije, a to su frekvencija, frekvencija na milijun i ARF (prosječna smanjena frekvencija, eng. *average reduced frequency)*.

Slijed niza stavki u Sketch Enginu naziva se n-gram (n zamjenjuje bilo koji broj, a najčešći su bigrami i trigrami). Ta stavka može se odnositi na više stvari, uključujući slovo, znamenku, slog, riječ ili nešto drugo. U kontekstu korpusa i korpusne lingvistike, n-grami se obično odnose na lekseme (ili riječi). Stvaranje popisa najčešćih n-grama pomaže nam u lingvističkim pojavama koje bi mogle ostati nezapažene pri korištenju drugih alata. N-grami mogu identificirati oznake diskursa ili dijelove jezika koje bi trebalo podučavati i učiti kao fiksne fraze u podučavanju nekog stranog jezika.

Ključne riječi i izdvajanje pojmova (eng. *keywords and term extraction*) podrazumijevaju prepoznavanje tipičnih riječi za korpus koji se analizira. Sketch Enginu podržava izdvajanje jednojezičnih ili dvojezičnih termina. Takvo izdvajanje ključnih riječi i pojmova koristi se za izdvajanje terminologije, za njezinu upotrebu u prijevodima te za tumačenje, za izdvajanje jedinica riječi i pojmova od više riječi koji su tipični za korpus koji se pretražuje. Također se koristi i za usporedbu dvaju korpusa, preko toga se ustanovljuje što je u jednom korpusu tipičnije u odnosu na drugi. Dobiveni rezultati su podijeljeni na ključne riječi i pojmove. Ključne riječi su stavke s jednom riječju, a pojmovi su stavke od više riječi.

Još jedan alat odnosi se na neologizme i dijakronijsku analizu upotrebe riječi (eng. *neologisms and diachronic analysis od word usage*). Pod njih spadaju trendovi, a oni su značajka za otkrivanje riječi koje se s vremenom mijenjaju prema učestalosti korištenja (dijakronijska analiza). Trendovi prepoznaju riječi čija se upotreba s vremenom povećava ili smanjuje. Pomoću trendova leksikolozi mogu identificirati nove riječi, neologizme, a mogu ih koristiti i povjesničari kako bi identificirali trenutak u kojem se riječ počela upotrebljavati ili kada se

prestala upotrebljavati te mogu vidjeti u kojem se trenutku riječ počela neobično često ili rijetko koristiti. Trendovi se mogu koristiti samo u korpusima koji sadrže vremensku oznaku. Ova značajka nije omogućena u jednojezičnim korpusima na hrvatskom jeziku.

Jedna od glavnih značajki Sketch Engina je ta da se u njemu mogu izrađivati vlastiti korpusi koji se potom mogu analizirati. Jedinstveni alat za izgradnju korpusa koristi tehnologiju WbBootCaT za automatsko stvaranje tekstualnog korpusa. Podaci za korpus mogu biti preuzeti s interneta ili se može umetnuti bilo koji drugi tekstualni dokument s računala. Podaci preuzeti s interneta čiste se, određeni dijelovi se uklanjaju automatski kako bi se dobio jezično vrijedan tekstualni materijal.

Sketch Engine može koristiti i dvojezične i višejezične tekstove da bi pretražio riječ ili frazu te pokazao primjere prijevoda u kontekstu, takvi tekstovi se nazivaju paralelnim korpusima, takav postupak je paralelno pretraživanje, a rezultat toga pretraživanja je paralelna podudarnost konkordancija (eng. *parallel concordance*). Sketch Engine sadrži gotove paralelne korpuse na mnogim jezicima, a također se i iz vlastitih tekstova mogu izgraditi paralelni korpusi.

1.4 Vizualizacija podataka

Vizualizacija podataka je grafički prikaz informacija i podataka. Korištenjem vizualnih elemenata poput grafikona i karata, alati za vizualizaciju pružaju pristupačan način za sagledavanje i razumijevanje izdvojenih vrijednost i obrazaca u podacima. U svijetu velikih podataka, alati za vizualizaciju podataka neophodni su za analizu velikih količina informacija i za donošenje odluka na temelju podataka. Naše oči privlače boje i uzorci te ih lako raspoznajemo i razdjeljujemo. Vizualizacija podataka još je jedan oblik vizualne umjetnosti koji plijeni naš interes i drži pogled na poruci. Kada vidimo grafikon, brzo vidimo trendove i izdvojenosti. Vizualizacija je sve ključniji alat za osmišljavanje bilijuna redaka podataka koji se generiraju svaki dan, ona pomaže ispričati priče smanjenjem podataka u oblik lakši za razumijevanje ističući trendove i izdvojenosti. Dobra vizualizacija priča priču, uklanja nepotrebne podatke, odnosno šum, i ističe korisne informacije. No teško je postići učinkovitu vizualizaciju podataka. Ako se, na primjer, grafikon previše pojednostavi, postaje previše dosadan da bi prenio bilo kakvu obavijest. S druge strane, previše kompliciran i uređena vizualizacija može biti potpuno disfunkcionalna u prenošenju prave poruke. U koristi je mnogih područja da podaci koje predstavljaju budu što razumljiviji i pristupačniji. Neka od

područja u kojima je vrlo bitno razumijevanje podataka su financije, marketing, povijest, uslužne djelatnosti, obrazovanje i tako dalje (Sadiku, Shadare, Musa, Akujuobi, 2016).

Što se tiče vizualizacije samog teksta, različite ga tvrtke najčešće koriste za sažimanje velikih količina teksta, automatsko označavanje ključnih pojmova i kategoriziranje tekstova prema temi. Vizualizacija teksta učinkovit je način pojednostavljivanja složenih podataka i prenošenja ideja putem oblaka riječi, grafova, grafikona, karata, mreža i vremenskih traka. Vizualizacija teksta jedan je od najvažnijih alata za rudarenje tekstom zbog svoje čitljivosti i razumljivosti i za ljude i za strojeve.

Za proučavanje teksta se znaju koristiti strojevi koji se služe novim istraživačkim poljem koje se naziva rudarenje tekstom. Rudarenje tekstom pokušava pronaći statističke obrasce analizom korpusa koji sadrži veliku količinu tekstova koji su čitljivi strojevima. Procesom rudarenja tekstom mogu se otkriti neki novi obrasci velikog podatkovnog sustava.

Oblaci riječi izvrsno su polazište za vizualizaciju kvalitativnih podataka. Oni pružaju osnovne i brze uvide te mogu biti korisni za istraživačku analizu. Oblaci riječi mogu biti najkreativniji i najsnažniji alat za vizualizaciju teksta. Ističući riječi veličinom fonta ili bojom prema učestalosti upotrebe boja, oni imaju svoj značaj i u analizi teksta i u digitalnoj humanistici. Rezultat analize teksta u obliku oblaka riječi može očigledno ukazivati na temu teksta ako se kao pretpostavka uzme da se važnije riječi češće pojavljuju. Jedan od alata za stvaranje oblaka riječi je Tagwedo. U tom alatu korisnik može stvarati oblik oblaka riječi, riječi koje će biti naglašene i kontrast riječi.

Što se tiče korištenja grafova za vizualizaciju teksta, najuspješniji je primjer Google Ngram Viewer, grafički alat za faznu uporabu koji se temelji na učestalosti faze upotrebe u materijalima objavljenim kroz različite godine, odnosno vremenski period. Ovaj alat podržava baza podataka koja sadrži više od pet milijuna digitaliziranih knjiga. Kada korisnik upiše nekoliko riječi ili izraza, alat će pretražiti bazu podataka kako bi pronašao podudaranja i generirao graf koji prokazuje odnos između učestalosti i objavljene godine. S ovakvom vrstom grafa istraživačima je lako proučiti razvoj fraza, a također i odnos među izrazima. Razvoj faza može u nekim aspektima pokazati i razvoj kulture i društva što ponovno povezuje vizualizaciju teksta s humanistikom.

Još jedna vrlo česta metoda kod vizualizacije teksta je upotreba grafikona. Oni se koriste u izvješćima, znanstvenim radovima, blogovima i tako dalje. Jednostavno ih je napravili, ali i koristiti. Postoji nekoliko različitih vrsta grafikona kao što su tortni grafikon, trakasti

grafikon, mjehurićasti grafikon. Sustav iz izradu grafikona može, na primjer, prikupljati podatke o radovima, a zatim generirati grafikon koji će prikazivati kako se citati jednog rada distribuiraju kroz vrijeme. Na taj način znanstvenici mogu lako pratiti razvoj istraživačkog polja. Iako ovdje vizualizirani elementi nisu riječi ili izrazi nego radovi, ovakav način je ipak vizualizacija teksta jer koristi citate svih tekstova u jednom radu isto kako drugi alati za vizualizaciju teksta koriste učestalost korištenja riječi.

Još jedan alat za vizualizaciju teksta su karte. Karte su uvijek bile važno oruđe za geoznanosti, ali u digitalnoj humanistici su one nova ideja. S razvojem rudarenja teksta, znanstvenici počinju koristiti alate drugih disciplina kako bi pronašli stvari koje se prije bile zanemarene i kako bi se što više razvili u što više smjerova. U vizualizaciji teksta postoje dvije različite vrste karata, zemljopisna i apstraktna karta. Za geografske je karte osnovna ideja staviti literaturu u odgovarajuće zemljopisno okruženje kako bi se pronašli neki odnosi ili interakcija između teksta i okoliša. Za apstraktne karte ideja je prikazati prostornu raspodjelu teksta, odnosno koji dijelovi teksta se koriste češće, a koji su manje značajni.

Mreža kao alat se koristi za prikaz odnosa između različitih jedinica koje čine cijelu mrežu. Vizualizacija teksta temeljena na mreži uglavnom kao cilj ima stvaranje mreže različitih dijelova određene teme i pokušaj pronalaženja odnosa između različitih dijelova i struktura te teme. Jedan važan oblik mreže je stablasta struktura koja više pažnje posvećuje odnosima između vodećih dijelova. Prateći strukture stabla, onaj koji ga proučava može lako saznati kako se jedan dio odnosi na drugi i bez dubljeg poznavanja teme koja se obrađuje. S ovom vrstom vizualizacije pojavljuju se neki skriveni aspekti u strukturi tekstova koji mogu pomoći da se tekstovi bolje razumiju.

Vremenska traka je vrlo često korišten alat za vizualizaciju podataka. Za razliku od karte koja se koristi za prikazivanje prostornih značajki teksta, vremenska traka se koristi za prikaz vremenskog aspekta teksta. Kada se vremenska traka koristi u vizualizaciji teksta, rezultati nalikuju elektroničkoj literaturi (Zhang, 2013). Na vremenskoj traci mogu biti postavljene ključne točke koje korisnik može odabrati te koje će mu dati određenu informacijo o tome što se na toj vremenskoj traci u tome trenutku dogodilo.

2. PRIKUPLJENI PODACI

2.1. Sastavljanje korpusa

Za provedbu ovog istraživanja bila su potrebna dva korpusa koja će se uspoređivati. Korpusi su napravljeni u korpusnom alatu Sketch Engine. Pristup Sketch Enginu studentima je osiguran preko AAI identiteta. Prvi korpus napravljen je iz putopisne knjige Antuna Nemčića koja se naziva *Putositnice,* a tekst za drugi korpus preuzet je s internetske stranice na kojoj se objavljuju putopisni blogovi, a ona se naziva *Putopisi net.*

Djelo *Putositnice* preuzet je s internetske stranice *E-lektire* (https://lektire.skole.hr/) na koju se prijaviljuje također putem AAI identiteta. Nakon preuzimanja PDF dokumenta, djelo je prebačeno u Word kako bi se mogao urediti i pripremiti za izradu korpusa u Sketch Enginu. Odlučila sam se za Word kako bi mi djelo ostalo u svojem originalnoj obliku, kako bi se jasno vidjela poglavlja i odlomci. Iz preuzetog su maknuti nepotrebni dijelovi koji ne doprinose u obradi i istraživanju. Iz cjelovitog teksta izbačen je sadržaj. Također su obrisani česti stihovi i rečenice na latinskom jeziku, takav tekst je većinom bio smješten ispod naslova poglavlja. Usred teksta su se također mogli naći i stihovi na drugim jezicima kao što su talijanski, francuski i njemački. Oni su također uklonjeni. Nepotreban je također bio i tekst koji se nalazio u fusnotama. I na samom kraju maknut je dio rječnika u kojem su se nalazile nepoznate strane, ali i hrvatske fraze i riječi. Nakon toga tekst je bio spreman za izradu korpusa u Sketch Enginu. Takvim je čišćenjem pripremljen tekst koji će se naknadno analizirati, a za analizu je važan samo tekst na hrvatskom jeziku te se zbog toga sav tekst na drugim jezicima briše. Tekst za stvaranje korpusa u program se umeće na način da se klikne na prozor za pretraživanje koji se otvori te se u njemu odabere opcija za stvaranje korpusa. Tamo je potrebno napisati ime korpusa, odabrati hoće li korpus biti jednojezičan ili višejezičan, u ovom slučaju je to jednojezični korpus, te odabrati jezik korpusa. Također postoji i opcija u kojoj se može napisati opis korpusa. Kada se taj dio ispuni, dolazi se do dijela na kojem se umeće tekst za željeni korpus. Postoje dvije opcije, pronalazak teksta na internetskim stranicama ili učitavanje već postojećeg teksta s računala. Za ovaj korpus sam odabrala opciju u kojoj učitavam vlastiti tekst s računala te sam učitala svoj dokument s uređenim tekstom. Takav tekst se učitava nekoliko sekundi te na kraju preostaje samo pritisnuti opciju sastavljanja i korpus je učitan u alatu. Sastavljeni korpus preuzet je tako što je

odabrana opcija upravljanja korpusom te preuzimanje (eng. *download*). Preuzeti dokument je u .txt formatu.

Što se tiče sastavljanja korpusa putopisnog bloga, proces je bio sličan, ali i ne potpuno isti. Cilj je bio pronaći što opširniji putopisni blog jednog autora. Naposlijetku je nađena internetska stranica *Putopisi net* (https://putopisi.net/) koja je između ostaloga podijeljena prema autorima i njihovim putopisima. Izabran je autor koji je od ponuđenih napisao najviše. O autoru se zna jedino da se zove Anton i dolazi iz Trogira. U izabranim putopisnim blogovima obiđeno je mnogo država svijeta. Pri sastavljanju ovoga korpusa izabran je nešto drugačiji redoslijed. Direktno u Sketck Engine, u koraku kada se dodaje tekst od kojeg će se raditi korpus, odabire se opcija find text on web. U ponuđeni se prostor kopiraju poveznice svih putopisnih tekstova autora Antona. Naposlijetku se izrađeni korpus preuzme u .txt formatu te se tada krene završno čistiti za ponovno vraćanje u Sketch Engine bez nepotrebnih dijelova. Iz tako preuzetog dokumenta maknuti su zaostali linkovi, nepotpuni opisi fotografija i rečenice koje su u potpunosti na engleskom jeziku. Takav uređeni dokument ponovno je učitan na Sketch Engine na isti način kao što je bio učitan i tekst za korpus *Putositnica.*

2.2. Opis korpusa

Prvi sastavljeni korpus, *Putositnice* Antuna Nemčića, izabran je kao stariji oblik putopisa. Djelo je napisano 1845. godine. Antun Nemčić je putopis originalno pisao pod pseudonimom A. N. Gostovinski. Djelo je prvi puta tiskano u Zagrebu u Narodnoj tiskarnici Ljudevita Gaja, a izdanje koje je korišteno u ovom istraživanju preuzeto je s portala eLektire. Autor opisuje svoj put po Hrvatskoj prema Italiji, a potom slijedi i opis puta natrag preko Italije kroz Sloveniju te putovanje završava u Austriji. Djelo je pisano vrlo životopisno i maštovito uz mnoge zanimljive i duhovite detalje. Uvode poglavlja autor je upotpunjavao stihovima na stranim jezicima. Te je stihove preuzimao od Horacija, Juvenala, Schillera i drugih stranih autora. Ti se stihovi u korpusu ne nalaze zbog toga što nisu pisani hrvatskim jezikom te nisu bitni za istraživanje. Ponekad zna preuzeti i stihove na hrvatskom od svojih suvremenika kao što su Petar Preradović i Dimitrije Demeter. Jezik kojim se koristi vrlo je bogat, a za neke riječi se gotovo sa sigurnošću može reći i da su upotrijebljene samo u njegovom djelu te da su izmišljene za potreban kontekst. Na primjer, govori o svojoj majci te kako bi opisao odnos s njom koristi riječ *premamiti,* a kasnije ju naziva *sofizmati.* U sintaksi je vidljivo da je dobro

znao latinsku gramatiku te da se njome vodio jer vrlo često stavlja glagol na kraju rečenice, a isto vrijedi i za njemački jezik. Također se može naići na mnoge izraze kojima su se tada pisci i intelektualci služili, a danas nam nisu toliko poznati te otežavaju razumijevanje djela. Ono je također prepuno latinizmima, germanizmima i romanizmima. Korpus se sastoji od 79033 pojavnica, a sa svim rečeničnim znakovima sadrži 97049 pojavnica. Različnica ima 21442. U Sketch Enginu pojavnica je svaka riječ koja se u analiziranom tekstu nalazi, pojavnice mogu biti tretirane samo kao riječi ili kao riječi i interpunkcijski znakovi. Različnica je svaka različita riječ koja se nalazi u tekstu te se ona računa kao jedna riječ iako se može pojavljivati više puta kroz tekst u različitim oblicima.

Drugi korpus nešto je jednostavniji od prvoga. Ovaj je korpus izabran kako predstavnik modernog oblika putopisa. Onog koji se ne piše s previše razmišljanja o stilu i jezičnoj preciznosti. U ovom se korpusu nalazi 34 različitih članaka koje je napisao autor Anton 2015. i 2016. godine. Piše o tome kako je putovao po Grčkoj, Tunisu, Americi, Albaniji, Srbiji te okolici navedenih lokacija. Svi članci su spojeni u jedinstveni korpus te su tretirani kao opreka korpusu *Putositnica*. Jezik kojim se u korpusu koristi pripada klasičnom razgovornom stilu. Blogovi su pisani na način da ih svatko razumije, jezik je vrlo jednostavan, a opet može biti zanimljiv svim uzrastima, tekstovi djeluju kao da se putovanja prepričava prijatelju, možemo naići i na mnoge poštapalice, umetnute riječi i usklike. Također, kao i u prvom korpusu, autor ponekad koristi izmišljene riječi. Ponekad piše na dijalektu što se prvenstveno može uočiti kod upotrebe jata, neke riječi su pisane ikavicom (čovik). Blog je čak pisan tako da se u ponekim slučajevima ne upotrebljavaju dijakritički znakovi (citao sam). Korpus se sastoji od 32522 pojavnica uključujući rečenične znakove, a bez njih je pojavnica 26472, a različnica 8671. Korpus je gotovo upola manji od prvoga korpusa.

3. METODOLOGIJA

Prva hipoteza koja je bila postavljena bila je da će vokabular kod Nemčićevih *Putositnica* biti razvijeniji nego u putopisnom blogu. Kako bi konačni rezultat za usporedbu bio što točniji, svela sam oba korpusa na približno jednak broj pojavnica. Iz korpusa *Putositnica* izdvojila sam početnih 25000 riječi. Korpusi su se svodili na sličan broj pojavnica kako bi u konačnoj usporedbi korpusi imali što sličniji broj pojavnica i kako bi dobiveni rezultati u usporedbi bili što sličniji. Dva navedena korpusa zapravo se tretiraju kao različite vrste tekstova te i o tome ovise dobiveni rezultati (Torruella, Capsada, 2013). Korpus putopisnog bloga ima 26472 pojavnica, a *Putositnice* 79033. Nakon skraćivanja dobiven je novi korpus koji ima 25431 pojavnicu. Do zaključka se došlo usporedbom omjera različnica i pojavnica u dvama korpusima. Iz toga se omjera zaključuje koliko se pojavnica pojavljuje u tekstu na jednu različnicu.

Druga provjeravana stavka bila je točnost rada lematizatora Sketch Engina u dvama korpusima. Lematizator obilježava vrste riječi te ih svodi na njihove leme. Pretpostavka je bila da će lematizator biti točniji u korpusu putopisnog bloga nego u korpusu *Putositnica* zbog arhaičnijeg jezika koji se u *Putositnicama* koristi. Takav jezik nije poznat Sketch Enginu te ga ne može obraditi i analizirati na ispravan način. Hipotezu je provjerena na način da su oba korpusa preuzeta u dokumentu u kojem su riječi, svaka zasebno, složene vertikalno. Uz svaku se riječ horizontalno nalaze njezini podaci. U ovom slučaju bitni su podaci vrsta riječi te postavljanje riječi u njezin osnovni oblik ili lemu. U oba korpusa postupak je bio jednak. Izdvojen je dio od otprilike 500 riječi na kojemu se provjeravala postavljena hipoteza. Ona se provjeravala putem Excela. Preuzimala sam ih iz dokumenta u kojemu su riječi postavljene vertikalno.

Treća je provjeravana stavka bila putopisna terminologija. Do nje se dolazilo na način da je u oba korpusa odabrana opcija ključne riječi. Kao referentni korpus s kojim su se korpusi uspoređuju kako bi se izvukle tipične riječi baš za ove tekstove, odabran je hrWaC. Prvo je odabrana opcija u kojoj se pronalaze ključne riječi (eng. *single-words*) gdje je izdvojeno 50 termina od jedne riječi, pod opcijom koja omogućava promjenu pregleda uključena je opcija koja pokazuje koliko se puta pojedine riječi u dokumentu pojavljuju te je prema broju pojavljivanja riječi u dokumentu odabrano 50 najučestalijih. Nakon toga je odabrana opcija koja prikazuje pojmove od više riječi te se ponovio postupak u kojem je izdvojeno 50 najučestalijih pojmova koji su sastavljeni od dvije ili više riječi.

4. REZULTATI ISTRAŽIVANJA

Prvom se hipotezom provjeravao omjer različnica i pojavnica u dvama korpusima. Pretpostavka je bila da će razvijeniji jezik u omjeru imati korpus *Putositnica* zbog toga što je navedeno djelo Antuna Nemčića književni, umjetnički tekst dok je korpus putopisnih blogova jezikom nešto jednostavniji. Takva je pretpostavka postavljena zbog članka u kojemu vrste tekstova, kakav je proučavani blog, imaju manje razvijeni vokabular od tekstova koji su po vrsti sličniji *Putositnicama.* Blog je gledan kao kronike, a *Putositnice* kao epistularno djelo (Torruella, Capsada, 2013). Rezultati ipak pokazuju suprotnu situaciju. U omjeru pojavnica i različnica u korpusu *Putositnica* dolazi se do toga da se na jednu različnicu u korpusu nalazi 2,8 pojavnica. Kod korpusa putopisnih blogova na svaku se različnicu pojavljuje 3,1 pojavnica. Zaključak je da je korpus putopisnih blogova ipak ima malo razvijeniji vokabular od korpusa *Putositnica.*

Nakon što je, prema opisanoj metodologiji, ispitana druga hipoteza, dobiveni su sljedeći rezultati: u Nemčićevu uzorku korpusa koji se sastojao od 505 riječi, odnosno pojavnica, lematizator Sketch Engina i ja za 422 riječi smo odredili jednaku vrstu. Prvo su ručno određene vrste riječi kako rezultati lematizatora ne bi utjecali na rješenja određivanja, a nakon toga su preuzeti rezultati lematizatora. Kod 83 pojavnica dolazi do neslaganja oko vrste riječi. Kako je jezik u ovome korpusu arhaičniji, ponekad je alatu bilo teže odrediti točnu vrstu riječi, mijenja pridjeve za imenice, i obrnuto, a u nekim slučajevima čak kao česticu odredi najobičniju imenicu (kao što je na primjer riječ proljeće). Svaku pojavu veznika *i* također određuje kao česticu umjesto kao veznik. Točnost (eng. *accuracy)* definira se kao bliskost slaganja određene vrijednosti i stvarne vrijednosti. Točnost lematizatora u ovom je slučaju 83,56 %.

Što se tiče drugog korpusa, korpusa putopisnog bloga, postupak je bio jednak. Nakon što sam uz svaku pojavnicu odredila vrste riječi, iz lematizatora sam prepisala ono što je on uz riječ odredio. Ovaj uzorak iz korpusa se sastoji od 508 pojavnica od kojih je kod 436 pojavnica vrsta riječi određena jednako, a vrsta riječi se razlikuje kod 69 pojavnica. Lematizator je vrlo često priloge mijenjao s imenicama, i obrnuto, imenice s prilozima. Isto je tako znalo biti problema s prepoznavanjem pridjeva i priloga. Lematizator je u nekoliko navrata teško mogao raspoznati priloge kao nepromjenjive riječi te ih je zato znao svrstavati pod imenice ili pridjeve. Točnost u određivanja vrste riječi kod ovoga je koprusa 86,42 %.

Točnost rada lematizatora, odnosno svođenje pojavnica na leme bila je druga stavka koju sam ispitivala. Slično kao ispitivanje prethodne hipoteze, radila sam na uzorku od oko 500 riječi. U korpusu *Putositnica* nakon ispisanih 505 pojavnica slijedi stupac u Excelu u kojem sam navedene riječi svela na njihove leme, odnosno na njihove osnovne oblike. Nakon toga nalaze se preuzete leme iz lematizatora Sketch Engina. Potom se događa usporedba mojega svođenja na leme te svođenja na leme pomoću alata. Pojedine su se pojavnice u svakom svom ponavljanju svode na jednaku krivu lemu. Na primjer, često se ponavlja riječ *nu* koju lematizator uvijek svede samo na *n*, *našo* iz teksta ponovno u lemi svodi na *našo* umjeto na *naći,* prijedlog *s,* u tekstu *š,* u lemi svodi na *ti*. Pojavnicu *se* uvijek svodi na lemu *sebe* iako to nije uvijek tako jer *se* ponekad može bit zamjenica *sebe*, a ponekad samo čestica *se*. Krajnji rezultat točnosti lematizatora kod ovoga korpusa je 74,85 %.

Kod drugog korpusa, korpusa putopisnog bloga, jednak je postupak dolaženja do rezultata. Ponovno možemo pronaći neke pojavnice koje lematizator ne prepoznaje dovoljno dobro te ih konstantno svodi na pogrešnu lemu. U ovom korpusu to se, na primjer, događa s pojavnicom *put*. U tekstu se ona nalazi u obliku *puta,* a lematizator nam govori da bi isto tako trebala glasiti i lema. Zanimljiva je i pojavnica *više.* Nju lematizator svaki puta svede na lemu *vio.* Konačni je postotak točnosti 82,67 %.

Treći dio rezultata do kojih sam došla vezani su uz terminologiju u dvama korpusima. Do tipičnih ključnih riječi za pojedini korpus sam došla preko alata Sketch Engina, odabrala opciju ključne riječi te je alat usporedio prvo odabrani korpus *Putopsitnica* s hrWaC-om. Putem te usporedbe, došli smo do riječi i fraza koje nisu toliko tipične u nekom općem korpusu, a više se puta ponavljaju u korpusu koji smo htjeli analizirati. U rezultatima prvoga korpusa zapravo je vidljivo da dobivene riječi nisu toliko povezane s putopisom i putovanjima općenito kao što je bilo pretpostavljeno. Zapravo su dane riječi izvučene kao tipične zbog toga što je korpus prepun arhaičnih, netipičnih riječi za današnje pojmove. Oblici riječi su drukčiji, a neke riječi su i potpuno nepoznate, neke čak i izmišljene samo za ovaj kontekst. Od odabranih pedeset pojedinačnih riječi i pedeset fraza, gotovo svi pojedinačni termini dolaze na vrh ljestvice kao one riječi koje su se ponavljale više puta, dok za višesložne fraze ta ponavljanja ipak nisu toliko česta. Riječ *ko* se ponavlja najviše puta, čak 276, a fraza *konj od mjeda* 3 puta, što je najmanji broj ponavljanja, uz još neke druge riječi i fraze, koji je uzet kao primjer. Kod korpusa putopisnog bloga moglo bi se reći da je situacija poprilično slična. Kod izdvojenih riječi ponovno se isprepliću termini od jedne riječi i one fraze koje imaju dvije ili više riječi. Također se može zaključiti da istaknute riječi nemaju u potpunosti veze s

putopisom, no u ovome slučaju ipak više nego u prethodnom korpusu zbog toga što su poneke istaknute riječi geografski pojmovi koje autor ovih blogova posjećuje, spominje iz dovoljno puta te postaju istaknutije od drugih riječi. Kao što se događa i kod Antuna Nemčića, i ovdje autor Anton ponekad upotrebljava izmišljene riječi, ali i mnogo anglizama koje piše fonološki, odnosno onako kako se na engleskom jeziku čitaju, takve su riječi također istaknute kao tipične za ovaj korpus. Najčešće korištena riječ u ovome korpusu je *Tunis* koji se u mnogo navrata spominje zbog putovanja u nj, a iz izabranih fraza jedna od manje spominjanih, ali i dalje tipičnih za ovaj korpus, a ne i za opći, je *afrički pijesak*.

5. DISKUSIJA

Prva se provjeravala hipoteza u kojoj je rečeno da je korpus Antuna Nemčića razvijeniji od korpusa putopisnih blogova. Istraživanje je dokazalo da ta pretpostavka nije točna. Oba su korpusa svedena na otprilike 25000 pojavnica te je uspoređen broj različnica i pojavnica. Na svaku različnicu kod korpusa putopisnog bloga se nalazi veći broj pojavnica nego kod korpusa *Putositnica.* Prema tome, razvijeniji je koprus putopisnog bloga.

Primjena lematizatora Sketch Engina na dvama putopisnim korpusima pokazala je da taj lematizator uspješno odrađuje svoj zadatak. Prosjek je točnosti lematizatora, koji je bio provjeravan u dvije stavke, 81,88 %. Navedeni je rezultat prilično dobar ako uzmemo u obzir da je jedan korišteni korpus zapravo povijesni tekst koji je nastao sredinom 19. stoljeća te se korišteni jezik razlikuje od suvremenog hrvatskog jezika, a drugi je tekst primarno pisan razgovornim stilom koji se ne slaže u potpunosti s hrvatskim standardnim jezikom. Može se zaključiti da niti jedan korpus nije potpuno sukladan s današnjim hrvatskim standardnim jezikom te je lematizator dolazio do pogrešnih rješenja u slučajevima kada jezik nije bio u skladu sa standardom.

Prva se provjera točnosti lematizatora odnosi na određivanje vrste riječi u korpusima. Kod *Putositnica* lematizator je točan u 83,56 % slučajeva. Rad lematizatora bio je otežan kada je analizirao riječi koje nisu sukladne s današnjim standardnim jezikom. Na primjer, riječ *desivši* koja je preuzeta iz teksta je određena kao prilog, a zapravo je glagol koji na standardu znači *dogoditi se.* Iako se riječ *desiti* i danas često upotrebljava u žargonu, lematizator je ne prepoznaje. Još jedan tipičan primjer pri kojemu je vrsta riječi krivo određena je slučaj u kojemu riječ ima drukčiji morfološki oblik od suvremenog. Na primjer, u korpusu se nalazi riječ *moro* koja bi u današnjem standardu u ovom licu i broju glasila *morao.* Lematizator Sketch Engina *moro* određuje kao imenicu umjesto kao glagol. Također se nailazi na pogreške pri određivanju vrste riječi vrlo često u slučajevima kada analizirana riječ ima isti korijen riječi koji može imati i neka druga vrsta riječi. Jedan je takav primjer riječ *žive* koja je u kontekstu teksta glagol, a lematizator ju je odredio kao imenicu.

Druga stavka čija se točnost provjerava u lematizatoru Sketch Engina je svođenje riječi na lemu. U korpusu *Putositnica* konačna je točnost te stavke 74,85 %. Do netočno svedenih riječi na lemu dolazi iz istog razloga zbog kojeg riječima nije određena točna vrsta. Zbog zastarjelog jezika koji nije uvijek jednak današnjem standardu. Često kao lemu riječi ponovno

uzima oblik koji je napisan u tekstu, na primjer riječi *zakonskimi* određuje jednaku lemu, *zakonskimi.* Ako je riječi prethodno krivo označena vrsta, to za sobom „povlači“ i određivanje leme. Riječ *aleopat* je određena kao glagol pa će prema tome i lema biti u infinitivu, a u ovom slučaju glasi *aleopatiti.*

U usporedbi tih dviju stavki za korpus *Putositnica* dolazi se do zaključka da je lematizator Sketch Engina kod povijesnog teksta točniji u određivanju vrsta riječi nego u svođenju riječi na leme.

Točnost lematizatora kod drugog korpusa, odnosno kod korpusa putopisnih blogova je nešto veća. Što se tiče određivanja vrste riječi u korpusu, točnost je lematizatora 86,42 %. Za analizu ovog korpusa ključan je podatak da je pisan razgovornim funkcionalnim stilom. Pisan je na taj način kako bi se što više približio čitateljima putopisa. No izbor riječi svejedno nije previše udaljen od standarda te greške do kojih dolazi i ne moraju ovisiti isključivo o tome. Češće dolazi do nevezanih pogrešaka. Na primjer, riječ *i* nije određena kao veznik, već je određena kao čestica, riječ *sada* nije određena kao prilog, nego kao zamjenica, dok ima i obrnutih primjera, ovdje je riječ *ništa* određena kao prilog umjesto kao zamjenica. Ne može se točno definirati zašto dolazi do takvih grešaka, za razliku od grešaka do kojih dolazi u povijesnom tekstu.

Svođenje na lemu u korpusu putopisnih blogova je također točnije. Ovdje dolazimo do nešto značajnije razlike od gotovo 8 %. Točnost svođenja na lemu je 82,68 %. Događa se razlika također zbog toga što je jezik puno bliži standardu nego u prethodnom korpusu. No svejedno dolazi do grešaka. Kao i kod *Putopsitnica* lematizator ponekad, kada nema spreman siguran odgovor, na lemu svodi na način da ju doslovno prepisuje, odnosno lema je identična riječi koju treba svesti na lemu, na primjer riječ *di* koja znači *gdje* lematizator svodi na *di*, isto tako riječ puta koju bi trebao svesti da put ponovno bilježi kao puta. Nailazi se i na slučajeve u kojima određivanje vrste riječi i svođenje na lemu nisu dosljedni. Riječ *vezan* se u tekstu pojavljuje kao pridjev te je tako i određena, ali u svođenju na lemu svedena je na glagol *vezati*.

U ovome je slučaju zaključak jednak kao i kod prethodnog korpusa, kod lematizatora Sketch Engina određivanje vrsta riječi je točnije od određivanja lema riječi.

Gledajući zajedno oba koprusa, dolazi se do tvrdnje da je lematizator Sketch Engina točniji kod korpusa putopisnog bloga, odnosno kod korpusa u kojemu je jezik jednostavniji i bliži današnjem standardnom jeziku. Time se potvrđuje druga postavljena hipoteza.

Proučavanjem terminologije u korpusima može se zaključiti da traženje ključnih riječi u korpusima postavljenim na Sketch Engin prvenstveno funkcionira tako da se izdvoje pojmovi koji se najviše razlikuju od usporednog korpusa, u ovome slučaju hrWaC-a, ali to nužno ne znači da se u putopisnim korpusima izdvaja isključivo putopisna terminologija. To vrijedi za oba istraživana korpusa. No u svakom se korpusu ipak ističu određene skupine riječi. U *Putositnicama* se izdvajaju arhaične riječi, one koje se više ne mogu naći u standardu ili se koriste vrlo rijetko. Također su izdvojene tuđice, posebno talijanizmi. Naravno, može se naći i pokoji pojam koji opisuje putovanja Antuna Nemčića, odnosno znamenitosti koje je vidio na svojem putu. Kod korpusa putopisnih blogova ipak je više istaknuta putopisna terminologija. Ona je vrlo jednostavna i tipična za svakodnevni govor. Izdvojena su i imena lokacija koje je autor posjetio jer se one ponavljaju više puta u tekstovima, a nisu toliko tipične da bi se isticale u općem korpusu kao što je hrWaC. Kao i u prethodnom korpusu, izdvojene su i posuđenice, ovoga se puta radi o anglizmima. U oba se korpusa više puta ponavljaju termini od jedne riječi nego termini od više riječi.

6. ZAKLJUČAK

Digitalni alati za analizu jezika danas se smatraju jednim od ključnih pomagala, uz čovjeka, pri jezičnoj analizi. Nalazimo se u vremenu u kojemu je tehnologija sveprisutna, sve češće znanstvena istraživanja teže interdisciplinarnosti. Odnosno, ono što je nekada u humanističkim znanostima bio samo razgovor, papir i olovka, danas je digitalizirano, ubrzano i olakšano. Promijenile su se metode rada, a u ovom slučaju govorimo o metodama analiziranja tekstova, a takve metode se uključuju u znanstveno polje digitalne humanistike.

Takav pristup rada na istraživanjima ima svoje pozitivne, ali i negativne strane. U provedenom istraživanju pomoću korpusnog alata Sketch Engine mogu se pronaći oba slučaja. Lematizator navedenog alata može momentalno analizirati sve riječi koje se nalaze u korpusu, jedino što je potrebno je preuzeti dokument u kojemu se nalaze rješenja, određena vrsta riječi te riječi svedene na leme. Prednost je u tome što je taj proces iznimno brz i nije potrebno zastati na svakoj riječi te ju posebno analizirati. Negativna je strana to što alat nije točan u 100 % slučajeva. Na dokazivanju toga se ovaj rad i temeljio. Točnost alata uvelike ovisi o stilu kojim je pisan tekst koji će biti analiziran. Na posljetku se zapravo treba vratiti tradicionalnom načinu analiziranja jezika kako bi se provjerila rješenja alata. U tom smislu ponovno treba posvetiti više vremena analizi. Točnost bi bila mnogo veća kada bi analizirani tekst bio pisan u potpunosti po pravilima standarda.

Iako se ovaj alat razvija već godinama, analiza hrvatskog jezika nije dostupna toliko dugo te s obzirom na vrijeme u kojem je hrvatski jezik razvijan u alatu, on funkcionira odlično. Naravno uvijek postoji mjesta za napredak. Razvoj alata temeljen je na radu na korpusima pisanim standardnim jezikom. Da se vokabular proširi i točnost poveća, potrebno je raditi na što raznovrsnijim tekstovima. Kako bi se jezik i tehnologija što više slagali i imali međusobnu korist, bitna je suradnja stručnjaka iz različitih područja.

LITERATURA

1. Baisa, V., Michelfeit, J., Medved, M., Jakubičel, M. (2016). *European Union language resources in Sketch Engine*. Masaryk University, Brno, Chech Republic. Lexical Computing Ltd, United Kingdom.
2. Berry, M. (2019). *What are the digital humanities.* The British Academy. https://www.thebritishacademy.ac.uk/blog/what-are-digital-humanities/ [pristupljeno 26. 8. 2021].
3. Bosančić, B. (2011). *Uloga opisnih označiteljskih jezika u razvoju digitalne humanistike.* Sveučilište J. J. Strossmayera u Osijeku, Filozofski fakultet, Odsjek za informacijske znanosti.
4. Kilgarriff, A., Baisa, V., Bušta, J., Jakubiček, M., Kovar, V., Michelfeit, J., Rychly, P., Suchomel, V. (2014). *The Sketch Engine: ten years on.* Masaryk University, Brno, Chech Republic. Lexical Computing Ltd, United Kingdom.
5. *Sketch Engine user guide. Corpus tools.* Sketch Engine. https://www.sketchengine.eu/guide/ [pristupljeno 11. 6. 2021].
6. Torruella, J., Capsada, R. (2013). *Lexical statistics and tipological structures: A measure of lexical richness.* Universitat autonoma de Barcelona, Spain.
7. *What is dana visualization? Definition, examples and learning resources.* Tableau. https://www.tableau.com/learn/articles/data-visualization [pristupljeno 26. 8. 2021].
8. Zhang, X. (2013). *Text vizualization: see more than text.* Augmenting Realities. https://sites.duke.edu/lit80s_02_f2013_augrealities/text-visualization-see-more-than-texts/ [pristupljeno 26. 8. 2021].

Izgradnja povijesnog i suvremenog putopisnog korpusa i računalna usporedna analiza jezika

Sažetak

U sklopu ovog rada izgradit će se dva korpusa iste tematike, ali iz različitih razdoblja. Cilj je rada računalna usporedna analiza jezika dvaju korpusa putopisne tematike objavljenih u razmaku od dvjestotinjak godina. Prvi će biti temeljen na tekstu *Putositnice* Antuna Nemčića (prvi puta objavljen 1845.), a drugi korpus na suvremenim putopisnim blogovima jednog autora (objavljivani 2015. i 2016. godine). Glavni alat koji će se u radu koristiti za izgradnju i usporednu analizu bit će Sketch Engine. Također će se provjeravati točnost lematizatora za hrvatski jezik koji dolazi u sklopu alata. U usporedbi korpusa stavit će se naglasak na omjere različnica i pojavnica zbog usporedbe bogatstva vokabulara različitih autora te na tipičnu putopisnu terminologiju autora različitih razdoblja.

Ključne riječi: digitalna humanistika, Sketch Engine, korpus, lematizator

Construction of historical and contemporary travel corpus and comparative computer analysis of language

Summary

As a part of this work, two corpora of the same topic but from different time periods will be built. The goal of this paper is comparative computer analysis of the languages in the two travel corpora which are published at an interval of two hundred years. The first one will be based on the text from Antun Nemčić's Putositnice (first published in 1845), and the second corpus will be based on contemporary travel blogs by an author (published in 2015 and 2016). The main tool for construction and comparative analysis that will be used in the paper will be Sketch Engine. The accuracy of the lematizer for the Croatian language that comes with the tool will also be checked. In the comparison of the corpus, emphasis will be placed on the ratio of diversity and occurrence due to the comparison of the vocabulary richness of different authors and on the typical travel terminology of authors from different periods.

Key words: digital humanities, Sketch Engine, corpus, lematizer