Sveučilište u Rijeci – Odjel za informatiku

Jednopredmetna informatika

Matea Smoković

Razrješavanje višeznačnosti riječi pomoću konkordanci

Završni rad

Mentor: dr. sc. Lucia Načinović Prskalo

Rijeka, rujan 2018.

ZADATAK ZAVRŠNOG RADA:

ODJEL ZA INFORMATIKU
SVEUČILIŠTE U RIJECI

Rijeka, 15.3.2018.

Zadatak za završni rad

Pristupnik: Matea Smoković

Naziv završnog rada: Razrješavanje višeznačnosti riječi pomoću konkordanci

Naziv završnog rada na eng. jeziku: Resolving ambiguities with concordances

Sadržaj zadatka: Zadatak završnog rada je identificirati parove višeznačnih riječi iz *Hrvatskog naglasnog rječnika* koje imaju imaju isti pisani oblik i isti naglasak, a različito značenje kao primjerice u riječima kòsa (na glavi) i kòsa (oruđe). Uz pomoć konteksta (pronalaženjem konkordanci koji se pojavljuju uz datu riječ) razriješit će se problem višeznačnosti kod takvih odabira parova.

Mentor

Dr. sc. Lucia Načinović Prskalo

Voditelj za završne radove

Dr. sc. Miran Pobar

Zadatak preuzet: 22.3.2018.

(potpis pristupnika)

SADRŽAJ:

SAŽETAK

Glavni cilj ovog završnog rada je opisati postupak razrješavanje višeznačnosti riječi s posebnim naglaskom na hrvatski jezik. Navedeni su i detaljno objašnjeni pristupi i metode koje se koriste pri razrješavanju navedenog problema. Ukratko je opisan hrvatski naglasni sustav te alati korišteni za pomoć pri izradi ovog rada. Na posljetku napravljena je detaljna analiza riječi u hrvatskom jeziku s istim grafemskim i prozodijskim slijedom, a različitim značenjem. Prikazani su postupci dobivanja rezultata korak po korak, odnosno preuzimanja datoteka s konkordancama, izbacivanja interpunkcijskih znakova i brojeva, razdvajanja sadržaja izvorne datoteke u više odvojenih datoteka ovisno o značenju, izbacivanja stopwords-a te dobivanja datoteka s najčešćim konkordancama i pripadajućim frekvencijama.

Ključne riječi: WSD, problem višeznačnosti, razrješavanje višeznačnosti riječi, konkordance, rječnik, korpus, homonimi, homografi, homofoni.

1. UVOD

Hrvatski jezik veoma je kompleksan i morfološki bogat jezik (G. Hržica, A. Ordulj, 2013.). U njemu postoje riječi istog grafemskog i prozodemskog slijeda koje mogu imati više različitih značenja. Kod takvog se slučaja pojavljuje pitanje kako se može odrediti različito značenje riječi u kontekstu. Na prvi pogled nama ljudima se to ne čini kao veliki problem. Naprotiv, ljudski mozak lako može za riješiti taj problem. No u informatičkoj znanosti i informatičkoj tehnologiji veliki je izazov razviti sposobnost računala da radi obradu prirodnog jezika, odnosno, jako je teško razviti algoritam koji bi replicirao ljudsku sposobnost rješavanja tog problema. Do danas su isprobane brojne metode, kao što su nadzirana metoda, potpuno nenadzirana metoda, metoda temeljena na rječnicima koja koristi znanje kodirano na leksičkim resursima i slično.

Načinović Prskalo (Načinović Prskalo, 2016.) je u svom doktorskom radu izradila hrvatski naglasni rječnik u kojem se svaka natuknica sastoji od naglašenog oblika, morfosintaktičke oznake (poput broja, padeža i sl.) i nenaglašenog oblika. Tim se rječnikom mogu rješiti slučajevi kada imenice imaju različitu morfosintaktičku oznaku (u daljnjem tekstu: MSD oznaka) i/ili različit naglasak i slučajevi imenica s istim pisanim oblikom, a različitim naglaskom. Međutim ostaje problem s riječima koje imaju istu MSD oznaku, isti naglasak, a dva ili više različitih značenja. Tada problem višeznačnosti ne možemo riješiti pomoću naglasnog rječnika, već isključivo preko konteksta. Jedan od načina rješavanja tog problema je pomoću konkordanci i kolokacija.

Konkordance služe za prikaz pojavljivanja određene riječi u kontekstu te njezin kontekst. Od kolokacija se razlikuju po tome što kolokacije prikazuju samo parove riječe koje se gotovo uvijek zajedno ponavljaju u tekstu (primjerice crno vino), a kod konkordanci se prikazuje sve riječi iz konteksta (R. B. Guru, 2004.).

U prvom poglavlju iznesen je problem leksičke višeznačnosti. Ukratko je opisan hrvatski naglasni sustav i način na koji je strukturiran, te u kratkim crtama opisana povijest nastanka i razvoja WSD-a. U drugom poglavlju detaljno su opisani pristupi i metoda razrješavanja višeznačnosti riječi. U trećem poglavlju navedene su teškoće u postupcima razrješenja problema višeznačnosti, dok je u zadnjem poglavlju napravljena detaljna analiza riječi u hrvatskom jeziku s istim grafemskim i prozodijskim slijedom, a različitim značenjem

te su prikazani postupci dobivanja rezultata korak po korak, odnosno preuzimanja datoteka s konkordancama, izbacivanja interpunkcijskih znakova i brojeva, razdvajanja sadržaja izvorne datoteke u više odvojenih datoteka ovisno o značenju, izbacivanja stopwords-a te dobivanja datoteka s najčešćim konkordancama i pripadajućim frekvencijama.

2.PROBLEM LEKSIČKE VIŠEZNAČNOSTI

U mnogim jezicima diljem svijeta postoje riječi kojima značenje ovisi o kontekstu u kojem se nalaze. Razrješenje problema višeznačnosti, odnosno Word Sense Disambiguation (WSD) obuhvaća pronalaženje rješenja problema, odnosno pronalaženje točnog značenja dvosmislene riječi u određenom kontekstu (Ide N., Véronis J., 1998).

Glavno područje primjene WSD-a je strojno prevođenje, ali se koristi u gotovo svim vrstama lingvističkih istraživanja (Information Retrieval, Information Extraction) (Sanderson, M.1994.; J. Y. Chai, A. W. Biermann).

WSD je najprije bio oblikovan kao poseban računalni zadatak tijekom ranih dana strojnog prevođenja u 1940-ima, što ga čini jednim od najstarijih problema računalne lingvistike. Warren Weaver, u svom poznatom memorandumu o prijevodu (Weaver, 1949.) iz 1949. godine, prvi je put predstavio problem u računalnom kontekstu.

1980. godine došlo je do izuzetnog razvoja na području WSD-a. Leksički izvori poput Oxford Advanced Learner Dictionary (OALD) (J. Turnbull, J. Bradbery, M. Deuter, 1948.) postali su dostupniji istraživačima. Ručno kodiranje zamijenjeno je znanjima koja su automatski izvučena iz ovih izvora.

1991. godine Guthrie upotrebljava kodove kako bi identificirao točno značenje riječi koristeći Longman Dictionary of Contemporary English (LDOCE) (P. Longman, 1978.).

Tijekom devedesetih godina statistička revolucija počela se primjenjivati i u računalnoj lingvistici, a WSD je postao problemom paradigme na kojemu se primjenjuju nadzirane tehnike strojnog učenja. Postaje dostupan online rječnik WordNet (Miller, 1990.) koji donosi revoluciju u ovom području istraživanja koji je, za zastupljenije jezike poput engleskog, vrlo važan i danas, no za manje zastupljenije jezike poput hrvatskog jezika još uvijek nema dovoljno jezičnih resursa.

2.1 Hrvatski naglasni sustav i problem višeznačnosti

Naglasak predstavlja isticanje sloga u riječi jačinom i visinom glasa te trajanjem (D. Dujmović Markusi, T. Pavić Pezer, 2014.)

U hrvatskom jeziku postoje četiri vrste naglaska: kratkosilazni, kratkouzlazni, dugosilazni, dugouzlazni. U pravilu su u riječima naglašeni samo samoglasnici (a, e, i, o, u) pošto su temelj sloga, no moguće je da se naglasak nađe i na suglasniku 'r' ukoliko se radi o slogotvornom 'r' (vrt, smrt,...).

Obzirom da se u hrvatskom jeziku ne pišu naglasci, u pisanim riječima postoje riječi i oblici koji imaju isti slijed grafema. Takve se riječi nazivaju homogrami (*mol* i *mol*). Ukoliko je kod takvih riječi prisutan i jednak slijed prozodema, tj., jednako se i izgovaraju, onda se nazivaju homofoni (*mol* i *mol*), dok se homografi (istopisnice) podudaraju potpuno i po grafemskom i po prozodemskom slijedu, što je vidljivo iz slovopisa, npr. *môl* i *môl*. Homonimi su riječi istog fonetskog sastava, istog naglaska, iste vrste riječi te istog grafemskog sastava (*bôr* i *bôr*). Svi homonimi su istovremeno i homografi, ali nisu svi homografi homonimi (Tarfa B.).

Morfološko bogatim jezicima uz vrstu riječi dodaju se i dodatne morfološke (MSD) oznake i takve se slučajeve primjerice za hrvatski jezik može riješiti pomoću hrvatskog naglasnog rječnika (*róda (N jd. im. róda)* i *ròda (G jd. im. rôd)*) (L. Načinović Prskalo, 2016.). Ali što s riječima koje imaju istu MSD oznaku, isti naglasak, a različito značenje (*fàks – fakultet* i *fàks - telefaks*)? Tada problem dvosmislenosti ne možemo riješiti pomoću naglasnog rječnika, već isključivo preko konteksta. Takve nas riječi zanimaju u ovom radu.
Slijedi primjer riječi 'bôr' koja može imati dva značenja:

1. Vrsta stabla
2. Kemijski element

Primjer rečenica u kojima se imenica 'bôr' pojavljuje u dva konteksta:

1. Današnji neverin srušio je bor u obližnjoj šumi.
2. Za svoj pokus koristio je nekoliko kemijskih elemenata među kojima bor i cink.

Čovjeku je iz konteksta prilično jasno da je u prvoj rečenici značenje 'bôr' vrsta stabla, a u drugoj rečenici kemijski element, međutim kod automatskih postupaka razrješavanja višeznačnosti, nije lako odrediti pravo značenje.

3. PRISTUPI I METODE RAZRJEŠAVANJA VRSTE RIJEČI

Postoje dva glavna pristupa WSD-u - dubinski pristupi i plitki pristupi.

Dubinski pristupi podrazumijevaju dostupnost sveobuhvatnom tijelu svjetskog znanja. Znanje poput „ulovit ću kukca koji skače, ali ne mogu uloviti nož koji skače“ i „nožem ću rezati sir, a ne kukcem“ se koristi kako bi se odredilo u kojem se smislu riječ „skakavac“ koristi. Ovi pristupi nisu vrlo uspješni u praksi, uglavnom zato što takvo znanje ne postoji u računalnom obliku, osim vrlo ograničenih domena. Međutim, kad bi takvo znanje postojalo, onda bi dubinski pristupi bili mnogo precizniji i davali bolje rezultate od plitkih pristupa. Također postoji dugogodišnja tradicija u računalnoj lingvistici, odnosno iskušavanje takvih pristupa u smislu kodiranog znanja (X. Zhou, H. Han, 2005.).

Plitki pristupi ne pokušavaju razumjeti tekst. Oni samo razmatraju okolne riječi, koristeći informacije poput „ako se uz riječ bor u blizini pojavljuju riječi stablo, šuma tada je riječ o vrstu stabla, a ako se pojavljuje riječ element, kemija, cink, tada je riječ o kemijskom elementu“. Ova pravila mogu biti automatski izvedena od strane računala, koristeći korpus riječi s oznakom značenja (X. Zhou, H. Han, 2005.).

Postoje četiri konvencionalna pristupa WSD-u (Cucerzan, R.S., C. Schafer, D. Yarowsky, 2002.):

- Pristupi utemeljeni na rječniku i znanju: ovakvi se pristupi prvenstveno oslanjaju na rječnike, tezauruse i leksičke baze znanja, bez korištenja korpusa.
- Polu-nadzirani ili minimalno nadzirani pristupi: upotrebljavaju sekundarni izvor znanja kao što su mali označeni korpus kao sjeme podataka u procesu pokretanja ili primjerice uparene dvojezične korpuse.
- Nadzirani pristupi: koriste korpuse s oznakama značenja na kojima temelje postupak učenja modela.
- Nenadzirani pristupi: koriste netaknute korpuse te rade direktno s njima, bez sekundarnog izvora znanja.

3.1 Pristupi temeljeni na rječniku i znanju

Ovi su pristupi bazirani na različitim izvorima znanja kao što su rječnici, tesaurusi i slično. Sadrže nekoliko metoda među kojima je Leskov algoritam, heuristička metoda, metoda semantičke sličnosti i metoda prednosti odabira.

3.1.a Leskov algoritam

Temelji se na hipotezi da su riječi korištene zajedno u tekstu međusobno povezane i da se odnos može promatrati u definicijama riječi i njihovih značenja (Lesk, 1986.; Banerjee S., Pedersen T., 2002). Dvije (ili više) riječi razriješene su pronalaženjem para značenja u rječniku s najvećim preklapanjem riječi. Sličan pristup traži najkraći put između dvije riječi: drugu riječ iterativno pretražuje među definicijama svake semantičke varijante prve riječi, zatim među definicijama svake semantičke varijante svake riječi u prethodnim definicijama i tako dalje. Konačno, prva riječ je razriješena odabirom semantičke varijante koja minimizira udaljenost od prve do druge riječi.

3.1.b Metoda semantičke sličnosti

Metoda semantičke sličnosti (Mittal K. and Jain A., 2015) temelji se na pretpostavci da se riječi koje su povezane zajedničkim kontekstom i određenim značenjem nalaze na najmanjoj semantičkoj udaljenosti. Koriste se različite mjere sličnosti koje određuju koliko jako su dvije riječi semantički povezane.

3.1.c Metoda prednosti odabira

Ova metoda (Diana M.C., Carroll J; Patrick Y. and Timothy B., 2006.) pronalazi informacije o vjerojatnim odnosima vrsta riječi te označavaju značenja riječi koristeći izvore znanja. U navedenoj metodi neoznačene riječi su izostavljene i odabrane su samo one riječi koje su prethodno bile označene. Glavna ideja je brojanje koliko puta se neki par sintaktički povezanih riječi pojavljuje u korpusu.

3.1.d Heuristička metoda

Heuristika se procjenjuje iz različitih jezičnih svojstava kako bi se otkrilo značenje riječi (E. Agirre, P. Edmonds, 2007.). Kao temelj za procjenu WSD koriste se 3 osnovne vrste heuristike:

- najčešće značenje: pronalazi sva moguća značenja koja riječ može imati te odabire ono koje se pojavljuje češće od ostalih.
- jedno značenje po diskursu: riječ će zadržati svoje značenje bez obzira na njegova ostala pojavljivanja u tekstu.
- jedno značenje po kolokaciji: slično kao i jedno značenje po diskursu, samo što riječi koje se nalaze bliže jedna drugoj pružaju jaču vezu koja utječe na značenje riječi.

3.2 Nadzirani pristup

Nadzirani pristup zasniva se na pretpostavci da kontekst može dati dovoljno informacija za razrješavanje problema višeznačnosti. Ovaj pristup koristi tehniku strojnog učenja iz ručno označenih značenja te daje bolje rezultate od ostalih pristupa. Nadzirani se pristup sastoji od niza metoda koje su detaljno opisane ispod.

3.2.a Stabla odlučivanja

Slabla odlučivanja (Singh R. L., Ghosh K., Nongmeikapam, K., Bandyopadhyay, S., 2014) koriste se za označavanje pravila klasifikacije u strukturi stabla koja rekurzivno dijeli skup podataka o učenju. Unutarnji čvor stabla odluke označava test koji će biti primijenjen na trenutnu vrijednost, a svaka grana označava izlaz testa.

3.2.b Naive Bayes-ova metoda

Naive Bayes-ov klasifikator je probabilistički klasifikator utemeljen na Bayesovom teoremu (Le C. and Shimazu A., 2004; Aung N.T.T., Soe K.M., Thein N.L., 2011.). Ovaj pristup klasificira tekstualne dokumente koristeći dva parametra: uvjet vjerojatnosti svakog

značenja (Si) riječi (w) i značajke (fj) u kontekstu. Maksimalna vrijednost procijenjena iz formule predstavlja najprikladnije značenje u kontekstu.

$$\hat{S} = \underset{S_i \in Sense_D(w)}{\operatorname{argmax}} P(S_i \mid f_1, \ldots, f_m) = \underset{S_i \in Senses_D(w)}{\operatorname{argmax}} \frac{P(f_1, \ldots, f_m | S_i) P(S_i)}{P(f_1, \ldots, f_m)}$$

$$= \underset{S_i \in Senses_D(w)}{argmax} P(S_i) \prod_{j=1}^{m} P(f_j | S_i)$$

Slika 1 – Naive Bayes-ova formula

3.2.c Neuronske mreže

U računalnom modelu baziranom na neuronskim mrežama, za obradu podataka koriste se umjetni neuroni (Azzini, C. da Costa Pereira, Dragoni, Tettamanzi, 2008.). Neuronske mreže mogu se koristiti pri predstavljanju riječi kao čvorova. Ulazi se šire od ulaznog sloja preko međuslojeva pa sve do izlaznog sloja. Ulazi putuju kroz mreže te je teško izračunati jasan izlaz jer se veze u mreži neprestano šire u svim smjerovima i stvaraju petlje. Neuronske mreže su bolji izbor za probleme koji su neovisni o vremenu i predviđaju raznolik raspon primjena.

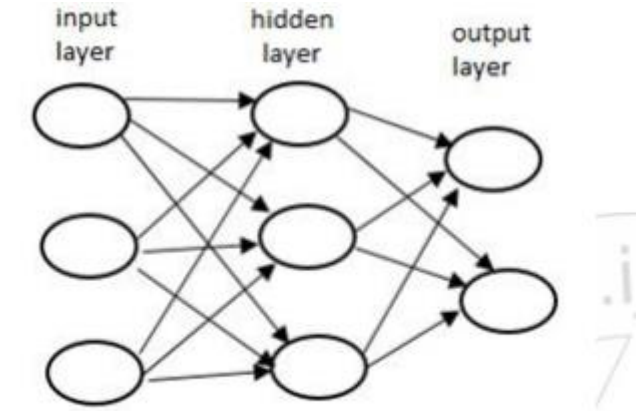


Figure 2: A feed forward neural network WSD with 3 features and 2 Responses

Slika 2 – Neuronske mreže (Azzini, C. da Costa Pereira, Dragoni, Tettamanzi, 2008.)

3.2.d Učenje temeljeno na primjerima ili instancama

Ovaj algoritam gradi klasifikacijski model iz primjera koji će kasnije biti dodani u model (H. T. Ng, 1997.). Najprije se prikuplja određeni broj primjera. Nakon toga izračuna se Hammingova udaljenost primjera koristeći k-NN algoritam. Hammingova udaljenost računa najbliži ulaz u odnosu na pohranjene primjere.

3.2.e Metoda potpornih vektora (Support vectore machine)

Cilj ove metode (Buscaldi D., Rosso P., Pla F., Segarra E., Arnal E. S., 2006.) je razdvajanje pozitivnih primjera od negativnih primjera sa maksimalnom granicom, a granica je udaljenost of hiperravnine do najbližeg pozitivnog i negativnog primjera. Ti primjeri koji su najbliži hiperravnini nazivaju se potporni vektori. Algoritmi bazirani na SVM (Vapnik, 1995.) koriste se za klasifikaciju nekoliko primjera u dvije različite klase. Ovaj algoritam nalazi hiperravninu između te dvije klase tako da odvajanje granica između te dvije klase postaje maksimalno. Klasifikacija testnog primjera ovisi o strani hiperravnine, odnosno strani gdje se testni primjer nalazi.

3.2.f AdaBoost

AdaBoost je metoda (Escudero, G., M`arquez, L. and Rigau, G., 2000.) za stvaranje jakih klasifikatora uz pomoć lineranih kombinacija nekoliko slabijih klasifikatora. Ova metoda pronalazi pogrešno klasificirane instance iz prethodnog klasifikatora tako da se može koristiti za daljnje klasifikatore. U svakom koraku obavlja se određena iteracija za svaki klasifikator. U svakoj iteraciji broj pogrešnih težina smanjuje se tako da se nadolazeći klasifikatori mogu usredotočiti na te pogrešne primjere.

3.3 Polu-nadzirani pristup

Zbog nedostatka podataka o učenju, mnogi algoritmi za razrješenje višeznačnosti riječi koriste polu-nadzirano učenje. Ono koristi označene i neoznačene podatke za učenje (Ankita Sati, 2013.), ali i podatke koji su potrebni za nadzirani i nenadzirani pristup te zbog toga nosi takav naziv. Kao i kod ostalih pristupa, polu-nadzirani pristup ima nekoliko različitih metoda među kojima su Yarowsky Bootstrapping metoda i dvojezična bootstrapping metoda (Bilingual bootstrapping method).

3.3.a Yarowsky Bootstrapping metoda

Yarowsky-jeva metoda (Yarowsky, D., 1995.) koristi jedan od najjednostavnijih iterativnih algoritama koji ne zahtijeva puno učenja i ovisi o jako malom broju instanci. Kako

polu-nadzirani pristup koristi označene instance, te instance se kasnije koriste za učenje klasifikatora. Zatim se klasifikatori koriste zbog razdvajanja većih podataka za učenje iz neoznačenog korpusa.

Glavna značajka ovog pristupa je mogućnost postupnog učenja iz sve šireg skupa podataka iz prvotno male skupine podataka za učenje. Također ova metoda postiže veliku preciznost.

3.3.b Dvojezična bootstrapping metoda (Bilingual bootstrapping method)

Ovo je jedna od novijih metoda razrješenja problema višeznačnosti. Metoda (Li H. & Li C., 2004.) koristi malu količinu klasificiranih podataka, a veliku količinu neklasificiranih podataka na dva različita jezika. Podaci u oba jezika moraju biti iste domene, ali ne trebaju biti paralelni. Na taj način stvaraju se klasifikatori u oba jezika ponavljajući dva koraka:

1. stvoriti klasifikatore za svaki jezik posebno na temelju klasificiranih podataka na oba jezika
2. iskoristiti napravljeni klasifikator za svaki jezik kako bi klasificirali neklasificirane podatke koji se zatim dodaju klasificiranim podacima

Možemo koristiti klasificirane podatke za oba jezika u prvom koraku, jer riječi u jednom jeziku imaju prijevod u drugom jeziku, a mi možemo primijeniti podatke iz jednog jezika u drugi. To povećava učinkovitost klasifikatora klasificiranjem neklasificiranih podataka u oba jezika i razmjenom informacija o kvalificiranim podacima oba jezika.

3.4 Nenadzirani pristup

Nenadzirano učenje je najveći izazov za WSD istraživače. Temeljna pretpostavka je da se slična značenja pojavljuju u sličnim kontekstima, a time i značenja mogu biti inducirana iz teksta grupiranjem pojava riječi pomoću neke mjere sličnosti konteksta. Taj zadatak naziva se indukcija riječi ili diskriminacija. Zatim se nove pojave riječi mogu svrstati u najbliže inducirane klastere/značenja. Izvedba ove metode daje slabije rezultate od gore navedenih

metoda, a i usporedbe su otežane jer značenja moraju biti mapirana u poznatom rječniku značenja.

Nenadzirani pristup (Martín-Wanton, T. , Berlanga-Llavori, R., 2012.) ne ovisi o vanjskim izvorima znanja, strojno čitljivim rječnicima ili podacima čija su značenja označena. Ovaj pristup ima dvije vrste distribucijskih pristupa: jednojezični korpus i ekvivalentni prijevod na temelju paralelnih korpusa. Te su tehnike dalje podijeljene u dvije vrste: pristup temeljen na tipu i pristup temeljen na tokenima. Pristup temeljen na tipu razdvaja se grupiranjem instanci ciljane riječi, dok se pristup temeljen na tokenu razdvaja grupiranjem konteksta ciljane riječi.

3.4.a Klasteriranje konteksta (Context clustering)

Metoda klasteriranja konteksta (Niu C., Li W., Srihari R. K., Li H., Crist L., 2004.) temelji se na grupiranju u kojem su stvoreni vektori koji će biti grupirani u klastere kako bi se utvrdilo značenje riječi. Ova metoda koristi vektorski prostor kao prostor riječi i njezine su dimenzije samo riječi. Riječ koja je u korpusu bit će označena kao vektor i brojati će se njezino pojavljivanje u kontekstu. Nakon toga stvara se matrica njezinog pojavljivanja i primjenjuju se mjere sličnosti. Tada se diskriminacija izvodi pomoću bilo koje tehnike klasteriranja.

3.4.b Klasteriranje riječi (Word clustering)

Ova je tehnika (Agirre E., López O, 2003.) slična klasteriranju konteksta u smislu pronalaženja značenja, ali klasterira one riječi koje su semantički jednake. Ovaj pristup za klasteriranje koristi Linovu metodu. Ona provjerava identične riječi koje su slične ciljanoj riječi. Riječi koje su slične dijele istu vrstu ovisnosti u korpusu. Nakon toga, algoritam klastera primjenjuje se na diskriminaciju među značenjima. Ako se uzme popis riječi, najprije se pronađe sličnost među njima te se tada riječi poredaju prema toj sličnosti i nastaje stablo sličnosti. U početnoj fazi postoji samo jedan čvor i za svaku dostupnu riječ u popisu napravi se iteracija koja dodaje najsličniju riječ početnom čvoru u stablu. Slablo se „obreže" i nastaju podstabla. Podstablo čiji je korijen početna riječ koju smo uzeli da bi našli značenje, daje značenje toj riječi.

3.4.c Graf zajedničkih pojavljivanja

Ova metoda (Cho J. K., Shin K. C., 2014.) stvara graf sa vrhom V i rubom E, gdje V predstavlja riječ u tekstu, a E se dodaje ako se riječi zajedno pojavljuju u istom poglavlju ili tekstu. Za određenu ciljnu riječ stvara se graf i matrica susjedstva. Nakon toga primjenjuje se Markova metoda klasteriranja kako bi se pronašlo značenje riječi. Svaki rub grafa ima dodijeljenu težinu koja je zajednička frekvencija tih riječi. Težina za rub {m, n} dana je formulom:

$$\boldsymbol{w_{mn} = 1 - \max\{P(w_m|w_n), P(w_n|w_m)\}}$$

gdje $\boldsymbol{P(w_m|w_n)}$ označava $freq_{mn}/freq_n$ a $freq_{mn}$ je zajedničko pojavljivanje frekvencija riječi $\boldsymbol{w_m}$ i $\boldsymbol{w_n}$, dok je $freq_n$ pojavljivanje frekvencije w_n. Riječi s visokom frekvencijom dodjeljuje se težina 0, a riječima koje se rijeđe pojavljuju dodjeljene su težine 1.

Na graf se primijenjuje iterativni algoritam i čvor koji ima najveći relativni stupanj odabire se kao središte. Cijelo središte označeno je kao značenje danoj ciljnoj riječi.

3.4.d Pristup utemeljen na rasprostranjenom stablu

Indukcija značenja riječi je zadatak identificiranja skupa značenja dvosmislenih riječi. Te metode pronalaze značenja riječi iz teksta s pretpostavkom da zadana riječ ima određeno značenje u određenom kontekstu kada se zajedno pojavljuje sa istim susjednim riječima (Agirre E., Martinez D., de Lacalle O.L., Soroa A., 2006).

Najprije se stvara graf zajedničkih pojavljivanja (Gq). Nakon toga izvodi se niz koraka kako bi se pronašlo točno značenje dvosmislenih riječi:

- svi čvorovi sa stupnjem 1 eliminiraju se iz grafa Gq,
- napravi se maksimalno rasprostranjeno stablo TGq,
- rubovi s minimalnom težinom uklanjaju se iz grafa TGq jedan po jedan, sve dok ne nastanu n-spojene komponente ili svi rubovi budu uklonjeni.

4. TEŠKOĆE U POSTUPCIMA RAZRJEŠAVANJA VIŠEZNAČNOSTI RIJEČI

U postupku razrješavanja višeznačnosti riječi može doći do različitih poteškoća zbog korištenja različitih sekundarnih izvora znanja poput rječnika ili označenih korpusa.

4.1 Razlike među rječnicima

Glavni cilj kod postupka razrješavanja višeznačnosti riječi je odlučivanje o tome koje je njihovo značenje. U slučajevima poput riječi „bor“, neka značenja vidno su različita. U drugim slučajevima, različita značenja mogu biti tijesno povezana (jedno znači metaforično ili metonimijsko proširenje drugog), a u takvim slučajevima podjela riječi po značenju postaje mnogo teža (npr. „griz“ kao hrana i kao zalogaj). Razni rječnici pružit će različite podjele riječi po značenjima. Jedno rješenje koje su istraživači koristili jest odabrati određeni rječnik i upotrijebiti svoj niz značenja. Općenito, rezultati istraživanja koji koriste riječi čija su značenja vidno različita bili su mnogo bolji nego od onih koji koriste riječi čija su značenja usko povezana.

4.2 Dodjeljivanje oznaka riječi (Part of speech tagging)

Part of speech tagging (u daljnjem tekstu: dodjeljivanje vrsta oznaka riječima) i razrješavanje višeznačnosti u nekim su slučajevima vrlo usko povezani. Pitanje je trebaju li ti zadaci biti provedeni zajedno ili odvojeno, ali znanstvenici nastoje zasebno testirati ove stvari (Wilks and Stevenson, 1996).

Problem razrješenja značenja riječi može se usporediti s problemom dodjeljivanjem vrsta oznaka riječima. Oboje uključuju razrješavanje ili označavanje riječi, bilo sa značenjem ili oznakom vrste riječi. Međutim, algoritmi koji se koriste za jedan problem nužno ne rade za drugi, zato što je postupak dodjeljivanja vrsta oznaka riječima prvenstveno određen jednom do tri neposredno susjednih riječi, dok se značenje riječi može odrediti širim okvirom riječi, odnosno kontekstom. Morfološko bogatim jezicima uz vrstu riječi dodaje se i dodatne morfološke (MSD) oznake i takve se slučajeve primjerice za hrvatski jezik može riješiti pomoću hrvatskog naglasnog rječnika (L. Načinović Prskalo, 2016.; L. Načinović Prskalo, M. Brkić Bakarić, 2018.; L. Načinović Prskalo, M. Brkić Bakarić, 2017.).

5. ANALIZA VIŠEZNAČNIH RIJEČI I NJIHOVA KONKEKSTA

U okviru ovog završnog rada odabrane su pojedine višeznačne riječi u hrvatskom jeziku koje imaju isti oblik i iste su vrste te imaju isti naglasak, ali dva ili više različitih značenja. Te su riječi pronađene u hrvatskim korpusima odakle je izvučen njihov kontekst te se je njihovo značenje ručno klasificiralo na temelju konteksta. U nastavku je opisan rječnik korišten u postupku analize i alat za pronalaženje odabranih riječi i njihova konteksta u hrvatskom korpusu.

5.1 Rječnik

Za ovaj rad korišten je Veliki rječnik hrvatskog jezika (Anić V., 2009.) kao jedan od najvećih suvremenih hrvatskih rječnika te jedini hrvatski rječnik s označenim naglascima na osnovnom obliku riječi dostupan u računalno pretraživom obliku. Rječnik sadrži 70 576 osnovnih riječi i 125 000 izvedenica na 1873 stranice, opširne definicije, gramatičke oblike, sinonime, frazeologiju, regionalizme, žargonizme i slično.

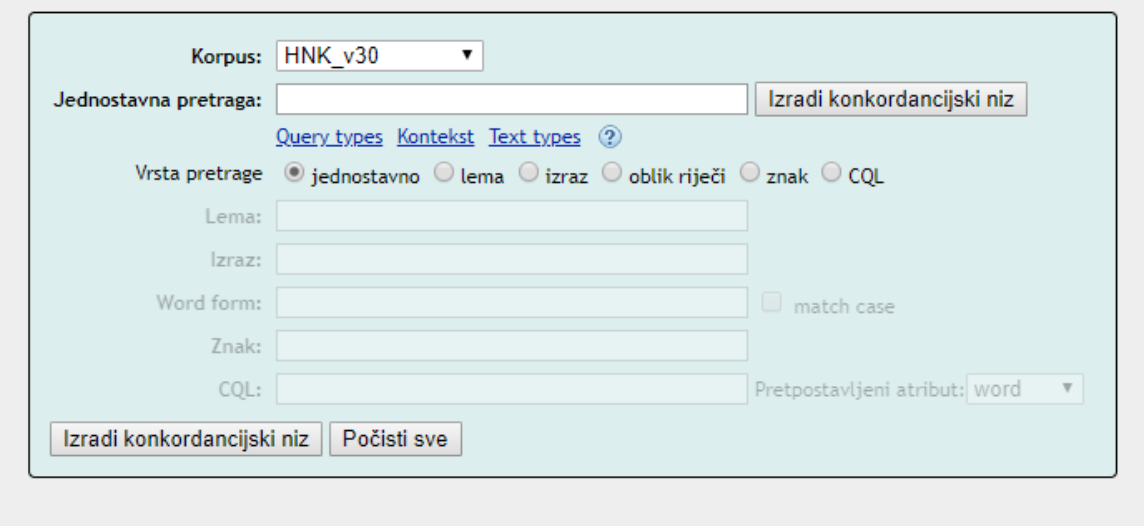
Iz njega su izvučene riječi koje imaju 2 ili više značenja i isti naglasak, primjerice bôr1 koji označava vrstu stabla i bôr2 koji označava kemijski element.

U rječniku su bile navedene i riječi koje imaju dva ili više značenja, ali su različite vrste, primjerice bòrdō prid. - tamnocrven i bòrdō m - francusko crno vino. Takve su riječi iz rječnika izbačene pomoću skripte u Pythonu. U konačnici je identificirano 547 riječi s istim oblikom i naglaskom, ali s dva ili više značenja. U ovom su radu za 43 takve riječi izvučene konkordance iz hrvatskih korpusa.

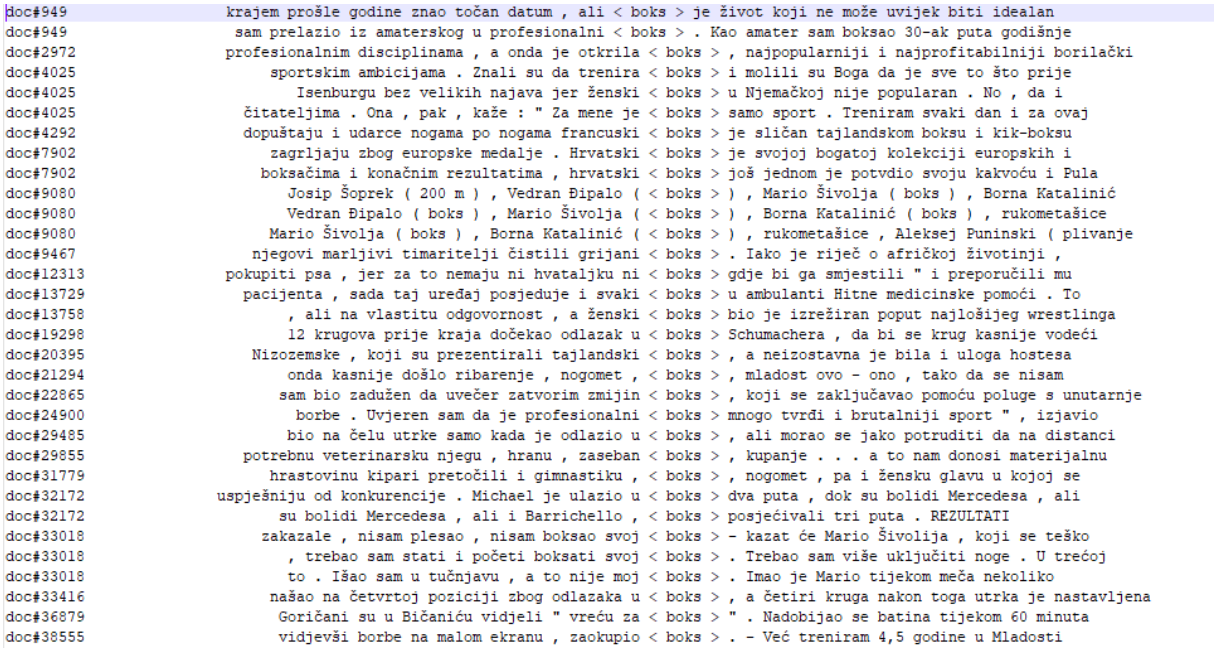
5.2 Sketch Engine

Sketch Engine (Kilgarriff A., Baisa V., Bušta J, Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V., 2014.) je alat za stvaranje i pretraživanje korpusa čiji su autori Adam Kilgarriff i Pavel Rychlý. Oni su 2003. sa svojom tvrtkom Lexical Computing Limited napravili Sketch Engine kako bi omogućili leksikografima, prevoditeljma ili ljudima željnih učenja jezika da s lakoćom pretražuju velike rječnike te nad njima izvode razne upite. Alat trenutno sadrži oko 500 otvorenih korpusa na više od 90 jezika.

S obzirom da se Sketch Engine plaća, pri izradi rada korištena je njegova besplatna inačica NoSketch Engine (Rychlý P., 2007.) koja je softverska podloga za Hrvatski nacionalni korpus (HNK) (Tadić M.,2000.) te hrWac - hrvatski web korpus (Ljubešić, Erjavec, 2011.) koji je ujedno i najveći računalni korpus hrvatskoga jezika. Alat pretražuje konkordance, te se datoteka s konkordancama lako preuzima. U ovom radu korištene su tri verzije korpusa: HNK 20, HNK 30 te hrWac korpus. Koji korpus je korišten za koju riječ naznačeno je u posebnoj datoteci zajedno s oznakama.



Slika 3 – početni prikaz NoSketch Engine-a (Rychlý P., 2007.)



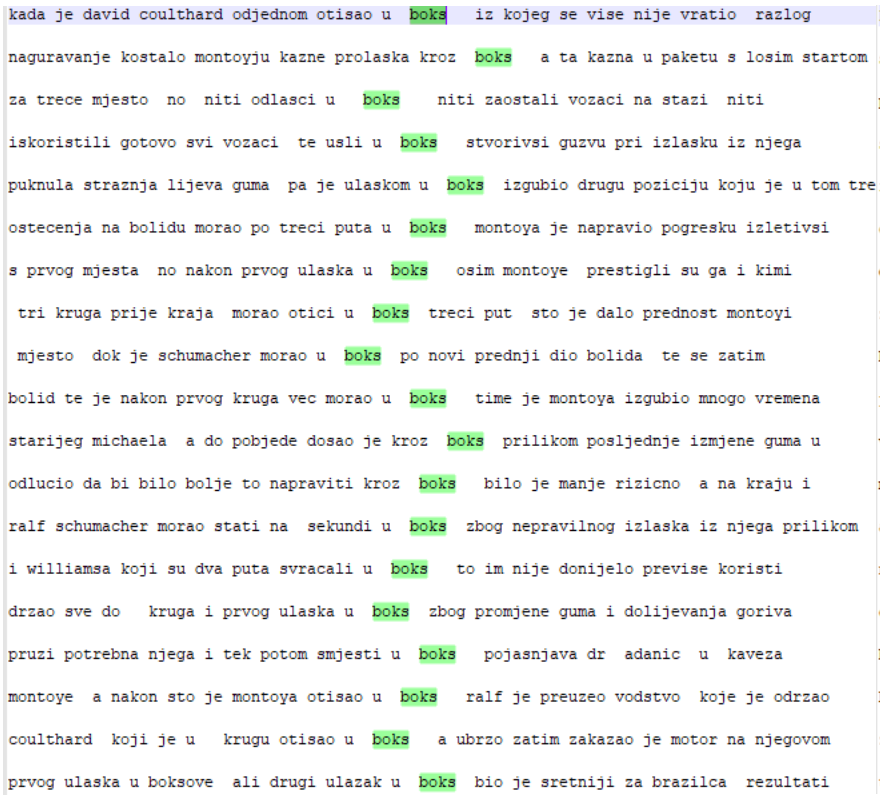
Slika 4 – izgled datoteke s preuzetim konkordancama

5.3 Rezultati analize

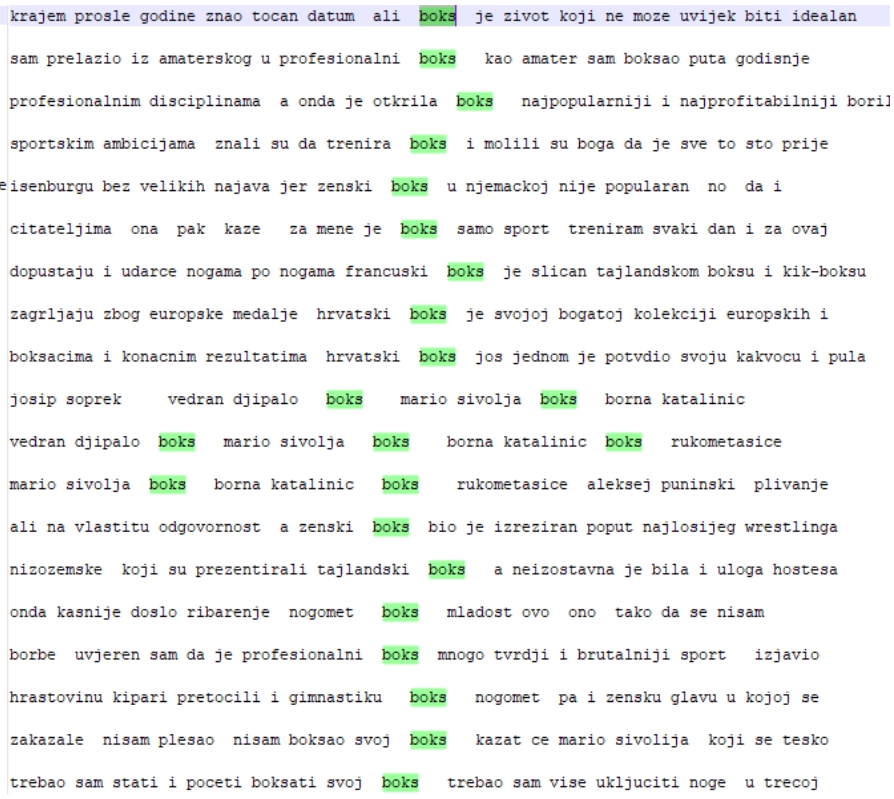
Nakon preuzimanja datoteka s konkordancama, bilo je potrebno uz pomoć Pytonove skripte iz datoteke izbaciti interpunkcijske znakove i brojeve jer su takvi znakovi nepotrebni u našim rezultatima. Također, svaki redak u preuzetoj datoteci počinje oznakom *doc#*, *text#* ili *http:* pa su i ti dijelovi izbačeni kako ne bi dobili pogrešne rezultate.

Nakon izbacivanja interpunkcijskih zakova i brojeva bilo je potrebno ručno razdvojiti sadržaj datoteke u dvije ili više posebnih datoteka, ovisno o značenju ciljne riječi. Ukoliko riječ primjerice ima 3 moguća značenja, onda su za tu riječ stvorene tri datoteke, a u svakoj datoteci spremljene su pripadajuće riječi i njihove konkordance ovisno o dodijeljenom značenju.

Sljedeće slike prikazuju primjer. Na prvoj slici vidimo datoteku s konkordancama u slučaju kada riječ *boks* označava neki zagrađeni prostor, dok druga slika prikazuje datoteku u slučaju kada riječ *boks* označava borilački sport.



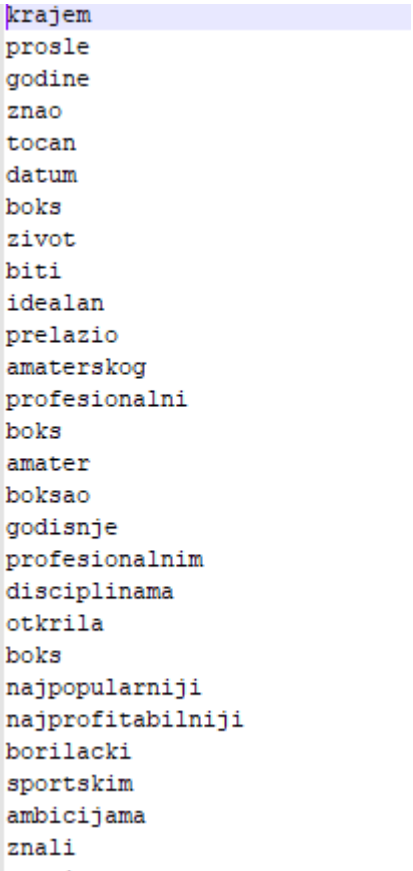
Slika 6 – boks označava prostor



Slika 5 – boks označava borilački sport

Nakon razdvajanja u odvojene datoteke iz istih je potrebno izbaciti riječi koje se nazivaju stopwords. Stopwords (zaustavne riječi) su riječi koje se u nekom jeziku najčešće pojavljuju, a nemaju veliko leksičko značenje. To su u hrvatskom jeziku riječi poput *i, ili, jer, kada, kakva, zato, to* itd. Popis stopwords-a koji sam koristila preuzet je s interneta u obliku datoteke. Ukoliko se takve riječi ne izbace, za oba značenja dobili bi jako slične najčešće riječi koje se pojavljuju u njihovom kontekstu.

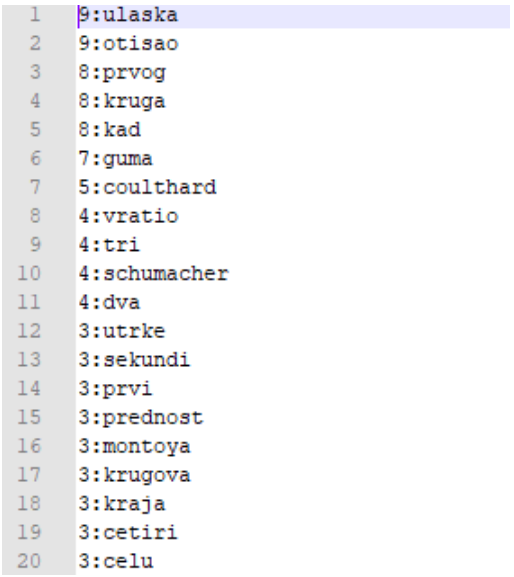
Nakon prolaska datoteka s konkordancama kroz Pythonovu skriptu dobivene su nove datoteke u kojima je napravljen popis konkordanci bez stopwords-a.



Slika 7 - popis konkordanci bez stopwords-a za boks: prostor

Slika 8 – popis konkordanci bez stopwords-a za boks: borilački sport

Zadnji je korak provlačenje obje datoteke s popisom konkordanci u Pythonovu skriptu koja će zbrojiti pojavljivanje svake riječi u datoteci te ih ispisati od najčešće pojavljivane do najmanje pojavljivane. Uz riječ biti će ispisan i broj pojavljivanja u datoteci. Veoma je bitno napomenuti da Pythonova skripta radi na način da najčešće pojavljivanu riječ briše iz konačnog popisa konkordanci iz razloga što je najčešće pojavljivana riječ upravo ona za koju tražimo konkordance (u prethodnim primjerima to je riječ *boks*).



Slika 10 – konačna lista konkordanci za boks: prostor

Slika 9 – konačna lista konkordanci za boks: borilački sport

U sljedećoj su tablici prikazani konačni rezultati svih riječi za koje su izvučene konkordance. S obzirom da su korištena tri različita korpusa, razdvojila sam riječi u zasebne tablice. U prvom se stupcu nalazi riječ za koju su tražene konkordance. U drugom stupcu nalazi se jedno značenje riječi, a u trećem drugo značenje, (i u četvrtom ako ima treće značenje) te ispod toga 5 najčešće pojavljivanih konkordanci uz broj njihovih pojavljivanja. Primjerice, prvo značenje riječi *balun* je *lopta*, a drugo značenje je *ples*.

Tabela 1 - Tablica konkordanci izvučenih iz hrWaC korpusa

Balun	Lopta	Ples
	– kad: 173 – šta: 142 – igra: 104 – san: 80 – ima: 74	– istarski: 9 – plesati: 5 – ples: 5 – sati: 4 – polka: 4
Biber	Papar	Crijep
	– sol: 305 – luk: 148 – dodati: 104 – meso: 100 – minuta: 74	– crijep: 43 – crijepa: 50 – crijepom: 37 – oblika: 21 – krov: 20
Bukva	Stablo	Riba
	– hrast: 211 – jasen: 80	– gira: 8 – srdela: 7

	– grab: 80 – vrste: 66 – drveta: 65	– koristi: 6 – svježa: 5 – riba: 5
Čičak	Biljka	Kopča
	– maslačak: 19 – kopriva: 17 – veliki: 15 – korijen: 15 – trava: 13	– trakom: 61 – traka: 51 – džepa: 45 – ima: 42 – cm: 37
Griz	Hrana	Zalogaj
	– mlijeko: 91 – kad: 73 – mlijeku: 70 – brašno: 65 – dodajte: 52	– jedan: 54 – kad: 40 – daj: 30 – sad: 21 – prvi: 17
Faks	Fakultet	Telefaks
	– kad: 120 – posao: 88 – sad: 75 – ima: 72 – znam: 56	– email: 42 – tel: 30 – telefon: 25 – poruka: 15 – broj: 15
Friz	Frizura	Greda
	– kad: 50 – ima: 47 – kosu: 45 – sad: 37 – ko: 36	– nalazi: 13 – teče: 9 – biti: 9 – vijenac: 8 – profilacija: 8
Gem	Gemišt	Pojam u tenisu
	– zapravo: 7 – gemišt: 5 – stari: 4 – kad: 4 – zeđ: 3	– servis: 245 – jedan: 207 – set: 169 – break: 166 – prvi: 150
Kamenica	Školjka	Posuda
	– uzgoj: 57 – kamenice: 56 – dagnji: 49 – školjaka: 36 – godine: 35	– nalazi: 9 – kamena: 9 – kamenih: 7 – vodu: 6 – vode: 6
Kapica	Pokrivalo	Školjka
	– ima: 72 – kad: 60 – biti: 37 – kapice: 36 – glavi: 34	– jakobovih: 33 – jakobova: 16 – jakovljeva: 15 – školjka: 14 – tri: 12
Klen	Biljka	Riba

	– jasen: 18 – grab: 17 – javor: 15 – hrast: 15 – bukva: 11	– štuka: 20 – mrena: 20 – šaran: 18 – riba: 18 – pastrva: 18
Leut	Brod	Lutnja
	– gajeta: 54 – brod: 35 – kad: 20 – mreže: 18 – plivarica: 17	– klapa: 46 – publike: 35 – znatni: 30 – svira: 29 – pusti: 27
List	Papir/novine	Biljka
	– večernji: 159 – jutarnji: 139 – novi: 131 – godine: 74 – piše: 69	– lista: 17 – bolesti: 11 – peršina: 9 – koristi: 9 – korijen: 9
Matirati	U sportu	Oduzeti sjaj
	– minuti: 100 – uspio: 93 – vratara: 58 – metara: 38 – loptu: 35	– puder: 26 – prahu: 14 – kožu: 14 – lice: 13 – kamenu: 11
Mišica	Ženska miša	Mišić u tijelu
	– umišljena: 15 – miš: 14 – kad: 11 – dana: 8 – predstavu: 7	– kad: 17 – srčanog: 13 – snagom: 10 – snagu: 8 – ruke: 8
Mrena	Bolest oka	Riba
	– siva: 96 – oka: 84 – leće: 79 – očna: 64 – katarakta: 62	– šaran: 19 – klen: 19 – štuka: 16 – ribe: 16 – riba: 15
Naložiti	Narediti	Zapaliti
	– sud: 138 – članak: 58 – odnosno: 55 – roku: 49 – članka: 47	– vatru: 97 – peć: 33 – drva: 28 – pećі: 14 – ima: 14
Skakavac	Kukac	Nož
	– kad: 25 – jedan: 12 – biti: 8 – mislim: 7	– nož: 43 – vozila: 6 – ulici: 6 – izvadio: 6

	– zeleni: 6	– ubo: 5	
Sunčanica	Gljiva	Bolest	
	– gljiva: 33 – vrganja: 25 – gljive: 24 – ima: 18 – macrolepiota: 14	– udar: 20 – suncu: 15 – biti: 15 – toplinski: 12 – dana: 9	
Šlep	Haljina	Tegljač, šleper	
	– haljine: 14 – veo: 13 – metara: 10 – haljina: 10 – dugački: 10	– auto: 62 – službu: 40 – služba: 37 – službe: 26 – kad: 23	
Tartan	Pod	Roba	
	– stazu: 109 – stazi: 104 – staze: 45 – minuti: 45 – staza: 41	– uzorak: 23 – uzorka: 14 – nositi: 14 – uzorkom: 10 – boje: 9	
Vlasulja	Biljka	Perika	
	– festuca: 21 – vrlo: 12 – morskih: 12 – crvena: 11 – meduza: 10	– kose: 13 – kosu: 7 – kad: 6 – vlasulje: 5 – kosa: 5	
Zaštopati	Začepiti	Onemogućiti nešto	
	– kad: 5 – glava: 5 – znam: 4 – wc: 4 – problem: 4	– loptu: 8 – igrača: 6 – uspio: 5 – vremena: 4 – kad: 4	
Mol	Kemijski element	Ljestvica u glazbi	Pristanište uz more
	– ml: 85 – otopine: 68 – vode: 43 – kiseline: 40 – otopina: 37	– dur: 110 – ton: 17 – ima: 15 – akord: 15 – dura: 13	– more: 30 – brod: 21 – ima: 20 – mola: 18 – godina: 18
Stolica	Za sjedenje	Izmet	Sjedište pape
	– stolova: 32 – stol: 30 – kad: 29 – stolice: 28 – biti: 27	– stolice: 31 – dana: 30 – tvrda: 18 – zelena: 17 – biti: 16	– sveta: 28 – prazna: 12 – petrova: 11 – papa: 9 – biskup: 7
Tuš	Za tuširanje	Šminka	Doživjeti šok (preneseno

			značenje)
	– wc: 101 – ima: 75 – kabine: 70 – sobe: 64 – kabina: 60	– oči: 60 – crni: 23 – tuša: 18 – sjenila: 18 – olovka: 17	– hladan: 116 – minuti: 15 – hladni: 14 – kad: 13 – uslijedio: 11
Vodenjak	Opna u maternici	Horoskopski znak	Životinja
	– pukao: 178 – kad: 141 – trudovi: 125 – porod: 89 – puknuo: 87	– hvala: 39 – biti: 37 – reno: 30 – ima: 28 – podznak: 27	– veliki: 9 – planinski: 8 – živi: 6 – staništa: 6 – cm: 6
Zvončić	Biljka	Granata	Zvonce
	– campanula: 30 – istarski: 27 – raste: 16 – vrsta: 15 – učke: 11	– mina: 12 – jedan: 8 – jednu: 6 – minu: 5 – komada: 5	– jedan: 9 – zvoni: 8 – mali: 8 – kraju: 8 – kad: 8

Tabela 2 - Tablica konkordanci izvučenih iz HNK20 korpusa

Bob	Sport	Mahunarka
	– hrvatski: 14 – četverosjed: 9 – šola: 7 – ivan: 6 – prvi: 5	– grašak: 12 – grah: 10 – kuhan: 7 – kasnije: 6 – suhi: 4
Boks	Prostor	Sport
	– ulaska: 9 – otišao: 9 – prvog: 8 – kruga: 8 – kad: 8	– tajlandski: 7 – profesionalni: 7 – hrvatski: 6 – sport: 5 – mario: 5
Gore	Suprotno od dolje	Jako loše
	– palac: 10 – dole: 9 – tri: 8 – navedene: 8 – jedan: 7	– stvari: 37 – bolje: 27 – stanje: 23 – moglo: 19 – puno: 12
Konac	Za šivanje	Kraj
	– kirurške: 20 – igle: 11 – škare: 9 – iglodržač: 9 – skalpele: 7	– godine: 15 – početak: 11 – biti: 8 – kad: 5 – života: 4

Pijavica	Životinja	Vrtlog
	– puževa: 7 – vrsta: 6 – sakupljanje: 6 – žaba: 5 – endemska: 4	– odnijela: 5 – morska: 4 – krova: 4 – pola: 3 – nevrijeme: 3 –
Polica	Osiguranje	Polica za knjige
	– osiguranja: 163 – kuna: 24 – životnog: 22 – dopunskog: 18 – cijene: 17	– biti: 16 – knjige: 14 – odnosno: 11 – smještaj: 10 – ormara: 9
Pop	Glazba	Svećenik
	– rock: 93 – glazbe: 56 – zvijezda: 26 – album: 25 – jazz: 19	– kad: 12 – biti: 9 – zna: 7 – đurić: 5 – župnik: 4
Predavati	Predavati nekome nešto	Predati nekome nešto
	– školi: 12 – počeo: 11 – hrvatski: 10 – fakultetu: 9 – profesori: 9	– oružje: 9 – prijevoz: 5 – mogli: 5 – zahtjeve: 4 – unaprijed: 4
Prevesti	Prevesti iz jednog jezika u drugi	Prevoziti nešto
	– hrvatski: 27 – jezik: 24 – jeziku: 8 – tekst: 7 – jezike: 7	– putnika: 53 – prijevoznik: 24 – tona: 16 – milijuna: 16 – teret: 14
Reket	Reketarenje	Tenis
	– hdz: 10 – medijski: 8 – plaćao: 6 – stranački: 4 – pravilo: 4	– prvi: 80 – hrvatski: 60 – drugi: 39 – ljubičić: 21 – ivan: 21
Studijski	Studij	Stidio
	– programi: 12 – program: 8 – novi: 7 – centar: 7 – boravak: 7	– album: 24 – projekt: 18 – novi: 14 – sljedeći: 5 – posljednji: 4

Žal	Obala	Žalost	
	– točke: 4 – pješčani: 4 – oči: 3 – zjenica: 2 – vremena: 2	– ostaje: 20 – propuštenom: 7 – vremenima: 6 – tri: 5 – prilikom: 5	
Kosa	Alat	Kosina	Vlas
	– kosi: 5 – puška: 3 – usitnjivač: 2 – trimer: 2 – strižna: 2	– metara: 11 – crta: 10 – strane: 8 – loptu: 6 – desne: 5	– glavi: 102 – diže: 67 – crna: 42 – lice: 33 – plava: 29
Marka	Novac	Brend	Poštanska markica
	– njemačka: 51 – konvertibilna: 24 – jedna: 19 – kuna: 14 – euro: 12	– vozila: 14 – broj: 13 – tip: 10 – robna: 10 – poznata:9	– doplatna: 34 – tiska: 21 – dnevnog: 21 – časopisa: 21 – poštanska: 19

Tabela 3 - Tablica konkordanci izvučenih iz HNK20 korpusa

Bor	Stablo	Kemijski element
	– pinus: 214 – crni: 84 – šumarija: 70 – upisnim: 68 – brojem: 67	– napomena: 45 – sadrže: 39 – sub: 26 – smjese: 25 – vrijednosti: 20

6. ZAKLJUČAK

U hrvatskom jeziku postoje riječi s istim grafemskim i prozodemskim slijedom, ali s različitim značenjem. Kod takvih je riječi teško razriješiti višeznačnost u automatskim postupcima. U ovom se radu pristupilo analizi takvih riječi i njihovih konteksta u hrvatskim korpusima. Iz konačnih rezultata možemo vidjeti koje se konkordance najčešće pojavljuju oko dvosmislenih riječi te na temelju njih možemo pretpostaviti koje je značenje ciljne riječi.

Za budući rad predviđeno je pronalaženje konkordanci za preostale nađene višeznačne riječi s istim naglaskom i oblikom, a različitim značenjem. Također, predviđena je primjena pronađenih najčešćih riječi za klasifikaciju značenja višeznačnih riječi u nekom neviđenom tekstu, analiza dobivenih rezultata te izračun postotka uspješnosti. Također, u ovom radu tražile su se konkordance u cijeloj okolini ciljne riječi, a u budućnosti planira se grupiranje konkordanci zasebno na konkordance ispred i zasebno na konkordance iza riječi kako bi se dobili još bolji rezultati.

7.LITERATURA

1. Ide, N., Véronis, J., (1998) “Word Sense Disambiguation: The State of the Art”, Computational Linguistics, Vol. 24, No. 1
2. Weaver, Warren (1949). "Translation" (PDF). In Locke, W.N.; Booth, A.D. Machine Translation of Languages: Fourteen Essays. Cambridge, MA: MIT Press.
3. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J., (1990) “WordNet An on-line Lexical Database”, International Journal of Lexicography
4. Banerjee, S., Pedersen, T.,(2002) "An adapted Lesk algorithm for word sense disambiguation using WordNet", In Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February.
5. Lesk, M.,(1986) "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone", Proceedings of SIGDOC.
6. Mittal, K. and Jain, A.,(2015)“Word sense disambiguation method using semantic similarity measures and owa operator”, ictact journal on soft computing: special issue on soft –computing theory, application and implications in engineering and technology, January, 2015.
7. Diana, M.C., Carroll, J., “Disambiguating Nouns, Verbs, and Adjectives Using Automatically Acquired Selectional Preferences”, Computational Linguistics, Volume 29, Number 4
8. Patrick, Y. and Timothy, B.,(2006) “Verb Sense Disambiguation Using Selectional Preferences Extracted with a State-of-the-art Semantic Role Labeler”, Proceedings of the 2006 Australasian Language Technology Workshop
9. Singh, R. L., Ghosh, K. , Nongmeikapam, K. and Bandyopadhyay, S.,(2014) “a decision tree based word sense disambiguation system in manipuri language”, Advanced Computing: An International Journal (ACIJ), Vol.5, No.4, July 2014.
10. Le, C. and Shimazu, A.,(2004)“High WSD accuracy using Naive Bayesian classifier with rich features”, PACLIC 18, December 8th-10th, 2004, Waseda University, Tokyo

11. Aung, N. T. T., Soe, K. M., Thein, N. L.,(2011)“A Word Sense Disambiguation System Using Naïve Bayesian Algorithm for Myanmar Language”, International Journal of Scientific & Engineering Research Volume 2, Issue 9, September-2011.
12. Buscaldi, D., Rosso, P., Pla, F., Segarra, E. and Arnal, E. S.,(2006)“Verb Sense Disambiguation Using Support Vector Machines: Impact of WordNet-Extracted Features”, A. Gelbukh (Ed.): CICLing 2006, LNCS 3878
13. H. T. Ng, ‘Exemplar-Base Word Sense Disambiguation: Some Recent Improvements’, in Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing, EMNLP, (1997).
14. Martín-Wanton, T. , Berlanga-Llavori, R.,(2012)“A clustering-based Approach for Unsupervised Word Sense Disambiguation”, Procesamiento del Lenguaje Natural, Revista no 49 septiembre de 2012.
15. Niu, C., Li, W., Srihari, R. K., Li, H., Crist, L.,(2004) “Context Clustering for Word Sense Disambiguation Based on Modeling Pairwise Context Similarities”, SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.
16. Ankita Sati, “Review: Semi-Supervised Learning Methods for Word Sense Disambiguation”, volume 12, issue 4. IOSR-JCE, 2013.
17. Li, H. & Li, C., “Word Translation Disambiguation Using Bilingual bootstrapping”, Computational Linguistics, 30(1), 1-22, 2004.
18. Escudero, G., M`arquez, L. and Rigau, G. Boosting Applied to Word Sense Disambiguation. Technical Report LSI-00-3-R, LSI Department, UPC, 2000.
19. Agirre, E., Martinez, D., de Lacalle, O.L., Soroa, A.: Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm. In: Proc. of TextGraphs 2006, New York, USA, pp. 89–96 (2006)
20. Yarowsky, D. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods". Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, 1995.
21. Agirre, E., López, O.: Clustering wordnet word senses. In: Proceedings of the Conference on Recent Advances on Natural Language Processing, Bulgary, 2003.
22. Cho J. K., Shin K. C.: A Graph-based Word Sense Disambiguation using Measures of Graph Connectivity. In: KIIT, Vol. 12, No. 6, (2014)
23. Načinović Prskalo, Lucia: Automatsko predviđanje i modeliranje hrvatskih prozodijskih obilježja na temelju teksta, Zagreb 2016.

24. Vladimir Anić: Veliki rječnik hrvatskog jezika, 2009.
25. [MREŽNO] https://en.wikipedia.org/wiki/Sketch_Engine#cite_note-4, (1.9.2018)
26. [MREŽNO] https://bib.irb.hr/datoteka/544723.Tafra_Istopisnice_i_istoslovnice.pdf (1.9.2018)
27. Eneko Agirre, Philip Edmonds: Word Sense Disambiguation Algorithms and Applications, 2007.
28. Dragica Dujmović-Markusi, Terezija Pavić - Pezer, FON FON 1, Profil 2014.
29. Nacinovic Prskalo, Lucia; Brkic Bakaric, Marija: The Role of Homograms in Machine Translation // International journal of machine learning and computing (IJMLC), 8 (2018), 2; 90-97 doi:10.18178/ijmlc.2018.8.2.669 (međunarodna recenzija, članak, znanstveni)
30. Nacinovic Prskalo, Lucia; Brkic Bakaric, Marija: Disambiguation of Homograms in a Pitch Accent Language // Proceedings of 2017 International Conference on Computer Science and Artificial Intelligence CSAI 2017 New York: ACM, 2017. str. 32-37 (predavanje, međunarodna recenzija, cjeloviti rad (in extenso), znanstveni)
31. Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít (10 July 2014). "The Sketch Engine: ten years on". Lexicography. Springer Berlin Heidelberg.
32. Rychlý, Pavel. Manatee/Bonito - A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno : Masaryk University, 2007. p. 65-70. ISBN 978-80-210-4471-5.
33. Ljubešić, N., Erjavec, T. (2011) hrWaC and slWac: Compiling Web Corpora for Croatian and Slovene. In Proceedings of the 14th International Conference Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 6836, Springer, Heidelberg, pp. 395-402.
34. Tadić, M. (2009) New version of the Croatian National Corpus. U: Hlaváčková, Dana ; Horák, Aleš ; Osolsobě, Klara ; Rychlý, Pavel (ur.) After Half a Century of Slavonic Natural Language Processing. Masaryk University, Brno, str. 199-205.
35. [MREŽNO] http://filip.ffzg.hr/cgi-bin/run.cgi/first_form (11.9.2018.)
36. [MREŽNO] https://www.clarin.si/noske/run.cgi/first_form?corpname=hrwac;align= (11.9.2018.)
37. Gordana Hržica, Antonia Ordulj: Dvočlane glagolske konstrukcije u usvajanju hrvatskoga jezika, znanstveni rad, 2013.

38. Ramakrishnan B. Guru: A Similarity Based Concordance Approach to Word Sense Disambiguation, 2004.
39. Sanderson, M.,(1994) "Word Sense Disambiguation and Information Retrieval", Proceedings of the 17th Annual International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR'94, July 03-06, Dublin, Ireland, Springer, New York, pp 142-151.
40. Joyce Yue Chai, Alan W. Biermann: The Use of Word Sense Disambiguation in an Information Extraction System
41. Joanna Turnbull, Jennifer Bradbery, Margaret Deuter: Oxford Advanced Learner's Dictionary, 1948.
42. Pearson Longman, Longman Dictionary of Contemporary English, 1978.
43. Xiaohua Zhou, Hyoil Han ,"Survey of Word Sense Disambiguation Approaches", College of InformationScience & Technology, Drexel University 3401Chestnut Street, Philadelphia, PA 19104, Appeared inThe 18th FLAIRS Conference, Clearwater Beach,Florida, May 15-17, 2005.
44. Cucerzan, R.S., C. Schafer, and D. Yarowsky, (2002) "Combining classifiers for word sense disambiguation", Natural Language Engineering, Vol. 8, No. 4, Cambridge University Press, Pp. 327- 341.
45. Azzini, C. da Costa Pereira, Dragoni, Tettamanzi, "Evolving Neural Networks for Word Sense Disambiguation", WSPC – Proceedings, 2008.
46. Wilks, Y. and M. Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? Technical Report CS-96- 05, University of Sheffield
47. [MREŽNO] https://bitbucket.org/trebor74hr/texthr/src/tip/text_hr/std_words.txt?fileviewer=file-view-default (12.9.2018.)

8. PRILOZI

Svi prilozi su predani na CD-u uz ovaj završni rad.

1. Python skripta za izvlačenje riječi s istim naglaskom i oblikom, a različitim značenjem iz rječnika.
2. Python skripta koja iz preuzetih datoteka s konkordancama miče oznake *doc#, text#* ili *http:*
3. Python skripta koja iz datoteka izbacuje stopwords
4. Python skripta koja iz datoteke ispiše najčešće konkordance s pripadajućim frekvencijama
5. Sve oblike preuzetih datoteka s konkordancama, odnosno izvorne datoteke, datoteke bez interpunkcijskih znakova i brojeva te oznaka *doc#, text#* ili *http:*, datoteke s izbačenim stopwords, datoteke s popisom najčešćih konkordanci i pripadajućim frekvencijama.
6. Datoteka s popisom naziva svih datoteka s popisom najčešćih konkordanci uz komentar koje je značenje ciljne riječi
7. Datoteka u kojoj su upisane sve riječi korištene za analizu grupirane po korištenim korpusima
8. Datoteka sa stopwords riječima
9. Datoteka u kojoj su izvučene riječi s istim naglaskom i oblikom, a različitim značenjem
10. Datoteka u kojoj se nalaze sve riječi iz Anićevog Velikog rječnika hrvatskog jezika