Laboratorij za analizu teksta i inženjerstvo znanja – TakeLab

Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 283

Računalni modeli distribucijske leksičke semantike hrvatskoga jezika

Vedrana Janković

Zagreb, lipanj 2011.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 21. veljače 2011.

DIPLOMSKI ZADATAK br. 283

Pristupnik: **Vedrana Janković**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Računalni modeli distribucijske leksičke semantike hrvatskoga jezika**

Opis zadatka:

Računalna semantika ima važnu ulogu u sustavima za obradu i razumijevanje prirodnoga jezika. Distribucijski semantički modeli značenje riječi prikazuju kontekstnim vektorima u višedimenzijskom vektorskom prostoru. Nadogradnju prestavljaju modeli distribucijske semantičke složivosti, kojima je moguće modelirati semantiku višerječnih izraza.
U radu je potrebno proučiti i opisati postojeće distribucijske semantičke modele i modele semantičke složivosti te postupke njihove izgradnje i vrednovanja, s naglaskom na model nasumičnog indeksiranja. Potrebno je oblikovati, programski ostvariti i vrednovati distribucijski semantički model za hrvatski jezik primijenjen na zadatak određivanja semantičke sličnosti riječi. Kao nadogradnju, potrebno je oblikovati, programski ostvariti i vrednovati semantički model složivosti za hrvatski jezik te razmotriti njegovu primjenu u detekciji idioma. Radu priložiti izvorni programski kod i ispitne uzorke.

Zadatak uručen pristupniku: 25. veljače 2011.
Rok za predaju rada: 10. lipnja 2011.

Mentor:

Prof.dr. sc. Bojana Dalbelo-Bašić

Djelovođa:

Doc.dr.sc. Domagoj Jakobović

Predsjednik odbora za diplomski rad profila:

Prof.dr.sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Semantika je lingvistička disciplina koja se bavi proučavanjem značenja i smisla riječi. Računalna semantika bavi se automatiziranim predstavljanjem i oblikovanjem značenja te zaključivanjem na temelju izraza prirodnoga jezika. Preduvjet automatizacije bilo kojeg aspekta semantike jest definicija koncepta značenja, koje je u distribucijskom pristupu semantici određeno isključivo kontekstom u kojemu se izraz pojavljuje. Navedeni pristup formalno je definiran distribucijskom hipotezom: stupanj semantičke sličnosti između dva jezična izraza jest funkcija sličnosti jezičnih konteksta u kojima se pojavljuju, što znači da se izrazi slična značenja pojavljuju u sličnim kontekstima.

Formalni model značenja teorijski utemeljen na distribucijskoj hipotezi jest distribucijski semantički model, jezični model u kojemu su izrazi predstavljeni visokodimenzionalnim vektorima na temelju statističke analize konteksta u kojima se pojavljuju. Vektori koji predstavljaju izraze povezane snažnijom semantičkom relacijom u vektorskom su prostoru međusobno bliži nego vektori značenjski nepovezanih izraza. Nadogradnju predstavljaju modeli distribucijske semantičke složivosti, kojima je moguće modelirati semantiku višerječnih izraza.

Važnost distribucijskih semantičkih modela dvojaka je. Jezično, distribucijski pristup značenju kao jednom od središnjih lingvističkih i filozofskih problema specifičan je jer ne zahtijeva njegovu eksplicitnu i konkretnu ontologijsku definiciju, a za izgradnju jezičnog modela dovoljan je samo dostatno opsežan korpus. Praktično, distribucijski semantički modeli primjenjuju se u zadacima pretrage dokumenata, ispitivanju sinonimije, razdvajanju značenja višeznačnih riječi, crpljenju informacija iz dvojezičnih resursa, konstrukciji taksonomije, automatiziranom grupiranju riječi i dr.

U ovom diplomskom radu proučeni su i opisani postojeći distribucijski semantički modeli i modeli semantičke složivosti, s naglaskom na modele nasumičnog indeksiranja. Opisani su postupci njihove izgradnje i vrednovanja. Oblikovano je i programski ostvareno 350 distribucijskih semantičkih modela za hrvatski jezik koji su primijenjeni na zadatak određivanja semantičke povezanosti riječi. Modeli su vrednovani uspored-

bom sa zlatnim standardom dobivenim od ljudskih ocjenjivača te je proučen utjecaj pojedinih parametara na izvedbu modela. Oblikovana su, programski ostvarena i vrednovana tri modela distribucijske semantičke složivosti za hrvatski jezik te je razmotrena njihova primjena u detekciji idioma.

U drugom poglavlju dan je pregled važnijih semantičkih teorija i pristupa definiciji značenja. U trećem poglavlju definiran je formalan način predstavljanja značenja. Razrada, argumentacija, definicija, vrste i primjena distribucijske hipoteze dani su u četvrtom poglavlju. U petom poglavlju formalno je definiran distribucijski semantički model i opisan način njegove izgradnje, uz detaljnu formalnu definiciju i opis ostvarenja modela nasumičnog indeksiranja. Pregled i primjena postojećih modela distribucijske semantike dani su u šestom poglavlju. U sedmom poglavlju definirani su i opisani postojeći modeli distribucijske semantičke složivosti. Ostvareni distribucijski semantički modeli, njihovo programsko ostvarenje, vrednovanje i dobiveni rezultati opisani su u osmom poglavlju, dok su u devetom poglavlju opisani ostvareni modeli distribucijske semantičke složivosti i njihovo vrednovanje. U desetom je poglavlju dan zaključak, kratak pregled rezultata i smjernice za budući rad.

2. Značenje u filozofiji jezika

Problematika jezika postala je u filozofskim okvirima jedinstvenom tematskom preokupacijom tek krajem 18. stoljeća, a u središte filozofskog zanimanja dospijeva u 20. stoljeću (Šoljić, 2010). Prvo veće zanimanje za filozofiju jezika simbolično označava naslov zbirke ogleda o filozofskoj metodi američkog filozofa Richarda Rortyja, Linguistic turn (Chicago University Press, 1967), gdje je jezična prekretnica zapravo shvaćanje jezika kao načina konceptualizacije, odnosno kao središnje točke u razumijevanju i objašnjavanju shvaćanja i percepcije svijeta (Wolf, 2009). Do zaokupljenosti jezikom došlo je također uslijed sve intenzivnijih kritika filozofije, koju se pokušalo svesti na filozofiju jezika, odnosno na njegovu kritiku. Povijesni pregled filozofije jezika može se ugrubo tematski i vremenski podijeliti u nekoliko cjelina: referencijalne teorije o jeziku, razdoblje ranog logičkog pozitivizma s idejama idealnog ili čak formaliziranog jezika, ranu fazu analitičke filozofije, prijelazno razdoblje kritike analitičkih teorija sredinom dvadesetog stoljeća i suvremenu filozofiju jezika.

Semantika je lingvistička disciplina koja se bavi izučavanjem značenja i smisla riječi. Automatiziranim predstavljanjem značenja i zaključivanjem na temelju izraza prirodnoga jezika bavi se računalna semantika. Osnovni preduvjet računalnoj automatizaciji semantike jest definiranje samoga koncepta značenja. Ta definicija predstavlja jedan od središnjih lingvističkih i filozofskih problema, koju se pristupa iz, između ostalih, konceptualne, objektne, semiotičke i pragmatičke perspektive.

Krajem devetnaestog stoljeća engleski filozof John Stuart Mill u okviru empirijske filozofije koju zastupa, razvija i empirijski pristup jeziku i značenju, shvaćajući značenje riječi jedino u kontekstu koncepata, odnosno objekata koje označavaju, a koje je pak moguće doživjeti jedino putem osjetila (Kalin, 2003). Taj pristup naziva se *referencijskom teorijom značenja*. Mill konotaciju podređuje denotaciji, definirajući time značenje samo kao oznaku, a ne i kao implikaciju šireg spektra atributa koje određeno značenje povlači za sobom. U referencijskoj teoriji značenja, rečenica *Mačka leži u košari* tumači se kao složeni uređeni skup oznaka u kojemu, primjerice, *mačka* označava životinju vrste Felis catus, a *košara* pleteni predmet koji može služiti za nošenje

stvari. Takvo shvaćanje značenja nailazi na dva ozbiljna problema, nemogućnost objašnjavanja pojmova koji se ne odnose ni na što, odnosno negativnih egzistencijalnih rečenica te informativnost rečenica koje izražavaju identitet (engl. *identity sentences*).

Prvi problem moguće je ilustrirati rečenicom *Atlantida ne postoji* (Wolf, 2009). Ako Atlantida ne postoji, izraz *Atlantida* ne označava ništa te stoga ni nema značenje. Definiranje Atlantide ne kao (postojećega) potonulog grada, već kao koncepta takvoga grada, rezultira paradoksom. Naime, rečenica *Atlantida ne postoji* tada postaje neistinitom s obzirom na to da tada izraz *Atlantida* označava koncept čije je postojanje nemoguće poreći. Problem informativnosti rečenica koje izražavaju identitet moguće je ilustrirati na sljedećim rečenicama: *Kilimandžaro je Kilimandžaro* i *Kilimandžaro je najviša planina u Africi* (Wolf, 2009). Prva rečenica izražava povratni identitet (engl. *self-identity*) te je zbog svoje forme logički istinita, ali ne sadrži nikakvu novu informaciju. Intuitivno bi se moglo zaključiti kako je, za razliku od prve, druga rečenica čitatelju mnogo informativnija. Međutim, pretpostavka je referencijske teorije značenja kako se *Kilimandžaro* i *najviša planina u Africi* odnose na istu stvar te su stoga prva i druga rečenica po značenju identične i jedna ne može nositi više informacija od druge. Takvo shvaćanje značenja negira spoznajnu vrijednost razumijevanja izraza. Naime, ako izrazi označavaju istu stvar, pojam ili koncept i ako je njihovo značenje sadržano samo u toj oznaci, ne postoji razlika u kognitivnom procesu razumijevanja prve i druge rečenice iz primjera.

Gottlib Frege predlaže rješenje navedena dva problema sagledavajući značenje kroz dva semantička aspekta koja naziva *smislom* i *značenjem*. Smisao izraza Frege definira kao način izražavanja putem kojega se na specifičan način prenosi informacija. Informacija prenesena izrazom određuje označenu stvar, odnosno definira značenje. Značenje je, dakle, određeno smislom, a njihov odnos Frege definira ovako: "Pravilna veza između znaka, njegova smisla i njegova značenja takva je da znaku odgovara određen smisao, a smislu opet određeno značenje, dok jednome značenju (jednome predmetu) ne pripada samo jedan znak" (Frege, 1995). Promišljajući nadalje o razlici između iskaza uobličenih kao $a = a$ i $a = b$, pod uvjetom da je posljednja jednakost istinita, Frege ističe da se u oba slučaja označava isti predmet, ali da je kod jednakosti $a = b$ potrebno uočiti i smisao (Šoljić, 2010). Važna posljedica proizašla iz razlikovanja smisla i značenja jest razlikovanje iskaza oblika $a = a$ od onih oblika $a = b$ ne samo konfiguracijski, već i različitim načinom danosti označenoga. Njihove su spoznajne vrijednosti različite, što Frege objašnjava tvrdnjom da za spoznajnu vrijednost smisao rečenice, to jest u njoj izražena misao, nije manje relevantna od njezina značenja, koje je njezina istinitosna vrijednost. Ako je $a = b$, značenje izraza b isto je kao i značenje

izraza a, zbog čega je i istinitosna vrijednost iskaza $a = b$ jednaka onoj iskaza $a = a$. Usprkos tome, smisao izraza a može biti različit od smisla izraza b, zbog čega misao izražena u $a = b$ može biti različita od misli izražene u $a = a$, čime su i spoznajne vrijednosti tih iskaza različite (Frege, 1995). Dva izraza jednakih značenja u većini se slučaja mogu unutar rečenice međusobno zamijeniti bez da se mijenja njezina istinitost. Primjerice, Thomas Alan Waits i Tom Waits označavaju istu osobu, stoga ako je rečenica *Thomas Alan Waits 1983. godine izdao je album Swordfishtrombones* istinita, onda je istinita i rečenica *Tom Waits 1983. godine izdao je album Swordfishtrombones*. Međutim, Frege ukazuje na postojanje konteksta u kojemu ovo ne vrijedi. Rečenica *Ana zna da je Tom Waits 1983. godine izdao album Swordfishtrombones* može biti istinita i onda kada je rečenica *Ana zna da je Thomas Alan Waits 1983. godine izdao album Swordfishtrombones* lažna, primjerice kada Ana ne zna da su Tom Waits i Thomas Alan Waits ista osoba, odnosno izrazi istoga značenja. Frege tvrdi kako u ovome slučaju značenje surečenica *Tom Waits 1983. godine izdao je album Swordfishtrombones* i *Thomas Alan Waits 1983. godine izdao je album Swordfishtrombones* nije njihova istinitosna vrijednost, već smisao same rečenice. Budući da je moguće razumjeti smisao samo jedne od njih, moguća su i različita značenja početnih rečenica iz primjera. Frege takve slučajeve naziva indirektnim kontekstima.

Ferdinand de Saussure u djelu *Cours de linguistique générale* (1916./1983.) postavlja temelje lingvističkoga strukturalizma na koji se u velikoj mjeri kasnije oslanja i distribucijska hipoteza. Strukturalističko poimanje jezika u središte svojega zanimanja stavlja strukturu jezika, dok manju važnost pridaje njegovoj individualnoj upotrebi. Apstraktna načela jezika kao sustava sastavni su dio svakog pojedinačnog iskaza. Tu ideju Saussure slikovito objašnjava analogijom s igrom šaha. Šah je definiran pravilima igre, figuricama i šahovskom pločom. Informacije o pojedinačnim potezima i odigranim partijama nisu prijeko potrebne za definiciju igre ili za njezino razumijevanje. Nadalje, svaku je figuricu moguće identificirati prema njezinim funkcionalnim razlikama u odnosu na sve druge. Primjerice, kralj se može pomicati za jedno polje u bilo kojem smjeru, a lovac dijagonalno za proizvoljan broj polja. Analogno, znakovi u jeziku definirani su svojim funkcionalnim razlikama. Saussure ulogu pojedinoga znaka u jeziku označava terminom *valeur*, koji definira ulogu, odnosno funkciju znaka unutar jezičnoga sustava. Razlikovna definicija koncepta *valeur* znači da znak ima ulogu u jeziku isključivo na temelju svoje različitosti od drugih znakova. Naglašava se važnost jezičnoga sustava kao cjeline jer razlike, a time i *valuers*, ne mogu postojati samostalno i odvojeno od sustava (Sahlgren, 2008). Jezični sustav određuje termin, iskaz, to jest znak, a znak određuje vrijednost, značenje, odnosno *valeur* (de Saussure,

1993). Funkcionalne razlike između znakova unutar jezika Saussure dijeli na sintagmatske i paradigmatske odnose. Sintagmatske relacije odnose se na pozicioniranje riječi i na riječi koje su supojavljuju unutar teksta, definirajući odnos *in praesentia*. Relacija je linearna i primjenjuje se na jezične entitete koji se mogu međusobno kombinirati. Paradigmatske relacije odnose se na zamjenu riječi i na jezične entitete koji se ne supojavljuju unutar teksta, ali se, iako ne istovremeno, pojavljuju u sličnim kontekstima, ostvarujući odnos *in absentia*. Paradigmatski je odnos nadomjesni odnos, a dvije se riječi u njemu nalaze kada odabir jedne isključuje odabir druge riječi (Sahlgren, 2008). Primjer sintagmatskih i paradigmatskih odnosa koje definira Saussure prikazan je u tablici 2.1.

Tablica 2.1: Sintagmatski i paradigmatski odnosi

	Paradigmatski odnosi oblika "α ili β ili ..."			
	dijete	*obožava*	*zelenu*	*boju*
Sintagmatski odnosi	*roditelj*	*voli*	*crvenu*	*farbu*
oblika "α i β i ..."	*sestra*	*preferira*	*plavu*	*nijansu*

Pojmovi *obožava*, *voli* i *preferira* u paradigmatskom su odnosu, a pojmovi *roditelj*, *voli*, *zelenu* i *boju* u sintagmatskom odnosu.

3. Formalno predstavljanje značenja

Tradicionalna formalna semantika značenje objašnjava preciznim logičkim formalizmima, pokušavajući strogoćom forme nadići nepotpunost, nepreciznost i višeznačnost prirodnoga jezika. Značenje riječi predstavljeno je formalnom simboličkom strukturom na temelju koje se izvode semantička svojstva riječi, ali i sintagmi i rečenica. U tablici 3.1 prikazan je primjer korištenja predikatne logike i lambda računa za predstavljanje značenja pojedinih riječi i rečenica te modeliranje formalnoga leksičkog zaključivanja (Stefan Evert, Alessandro Lenci, 2009).

Tablica 3.1: Predstavljanje značenja i zaključivanja formalnom logikom

Ivan	$\rightarrow$	**ivan**
loviti	$\rightarrow$	$\lambda x \lambda y.[\mathbf{loviti}(x, y)]$
jedan	$\rightarrow$	$\lambda P \lambda Q.\exists x[P(x) \wedge Q(x)]$
šišmiš	$\rightarrow$	$\lambda x.[\mathbf{šišmiš}(x)]$
Ivan lovi šišmiša	$\rightarrow$	$\exists x[\mathbf{šišmiš}(x) \wedge \mathbf{loviti}(\mathbf{ivan}, x)]$
Ivan lovi šišmiša	$\Rightarrow$	Ivan lovi životinju
ubiti $\rightarrow \lambda x \lambda y.[\mathbf{ubiti}(x, y)]$	$\Leftrightarrow$	$\lambda x \lambda y.[\mathbf{uzrok}(x, \mathbf{postati}(\mathbf{mrtav}(y)))]$

Opisano formalno predstavljanje značenja ne uspijeva riješiti problem višeznačnosti riječi, ne uzima u obzir način na koji kontekst utječe na njezino značenje te ni na koji način ne obuhvaća koncept učenja značenja.

4. Distribucijska hipoteza

Strukturno, prirodni se jezik može opisati kao slijed simbola, odnosno riječi, čije je značenje određeno odnosom s drugim riječima u nizu (sintagmi ili rečenici). Relacijski aspekt značenja moguće je demonstrirati promatranjem bilo koje višeznačne riječi, primjerice riječi *jagodica*, koja ima različito značenje u kontekstu *A* koji se odnosi na voće i u kontekstu *B* koji se odnosi na dijelove ljudskoga tijela. Kad bi značenje bilo određeno svojstvenom značenjskom karakteristikom, ta bi se karakteristika morala moći mijenjati s obzirom na to da je značenje u kontekstu *A* u ovom slučaju različito od značenja u kontekstu *B*. Ako je pak značenje definirano u okvirima odnosa s konteksnom okolinom, različito značenje iste riječi tumači se različitošću konteksta u kojima se riječ nalazi. Distribucijska svojstva jezičnih entiteta koriste se kao semantički građevni blokovi.

Na ovoj se pretpostavci temelji distribucijska hipoteza: stupanj semantičke sličnosti između dva jezična izraza α i β jest funkcija sličnosti jezičnih konteksta u kojima se α i β pojavljuju (Lenci, 2008). Distribucijska hipoteza definirana je i na sljedeće načine: riječi sličnih značenja pojavljuju se u sličnim kontekstima (H. Rubenstein, J. Goodenough, 1965); riječi sličnih značenja pojavit će se u sličnim susjedstvima ukoliko je korpus dovoljno velik (H. Shutze, J. Pedersen, 1995); prikaz načina korištenja riječi u njihovom prirodnom kontekstu obuhvatit će značajni dio koncepta značenja (Landauer, 1997) te riječi koje se pojavljuju u sličnim kontekstima sklone su i sličnostima u značenju (Pantel, 2005).

Koncept značenja prema distribucijskoj hipotezi djelomično se temelji na Wittgensteinovom objašnjenju značenja koje je određeno upotrebom riječi u jeziku. Upotreba riječi ostvaruje se njezinom distribucijom, zbog čega se distribucijski uzorci predstavljeni konktestima pojedine riječi mogu koristiti za određivanje (nekih aspekata) njezina značenja. Važno svojstvo takvog načina predstavljanja značenja jest činjenica da pritom nije potrebna njegova eksplicitna i konkretna ontologijska definicija (Jussi Karlgren, Magnus Sahlgren, 2001).

Promatraju li se distribucijska svojstva koja se uzimaju u obzir i njihove razlike,

distribucijsku hipotezu moguće je ostvariti kroz dva različita pristupa. Prvi je izgradnja distribucijskog profila pojedine riječi na temelju riječi koje ju okružuju ili su u nekoj drugoj relaciji s njom. Drugim pristupom distribucijski se modeli grade na temelju većih cjelina teksta (odlomaka ili dokumenata) u kojima se pojavljuju riječi koje se modeliraju (Sahlgren, 2008).

Razmatra li se aspekt značenja koji je pokriven jezičnim modelima temeljenima na distribucijskoj hipotezi, odnosno semantički status koji je dodijeljen kontekstnim prikazima značenja, distribucijska hipoteza može biti slaba ili jaka. Slaba distribucijska hipoteza temelj je kvantitativnih metoda semantičke analize. Na temelju distribucije riječi unutar konteksta pokušavaju se utvrditi uzorci semantičkih svojstava promatranog izraza, pod pretpostavkom da (eksplicitno nedefinirano) značenje riječi, odnosno izraza određuje način raspoređivanja unutar konteksta te da se semantička obilježja izraza ponašaju kao svojevrsna ograničenja koja određuju sintagmatsko ponašanje izraza. Posljedično, proučavanjem dovoljno velikoga broja konteksta trebalo bi biti moguće prepoznati one aspekte značenja koji su zajednički izrazima sličnih kontekstnih distribucija, a kojima bi se zatim mogla objasniti upravo ta distribucijska sličnost. Slaba distribucijska hipoteza pretpostavlja postojanje isključivo korelacijske veze između semantičkog sadržaja i kontekstnih distribucija, koju se koristi za bolje razumijevanje semantičkog ponašanja riječi. Slaba hipoteza ne podrazumijeva pretpostavku da su izravno u leksičkim distribucijama sadržana i semantička svojstva riječi i izraza (Lenci, 2008). Koristi se za jezično modeliranje, za izgradnju leksikona i tezaurusa, za razrješavanje višeznačnosti, ekstrakciju relacija i automatsko traženje odgovora na pitanja (engl. *question answering*).

Jaka distribucijska hipoteza leksičkoj distribuciji riječi daje sastavnu ulogu u tvorbi semantičke reprezentacije. Riječi koje se najčešće pojavljuju u kontekstu izraza formiraju njegovu kontekstnu reprezentaciju, odnosno apstrahirani prikaz najznačajnijih jedinica konteksta unutar kojih se izraz koristi (Lenci, 2008). Kontekstni prikazi riječi imaju uzročnu ulogu u formiranju njihovih semantičkih prikaza. Distribucijsko ponašanje riječi unutar konteksta nije shvaćeno samo kao način opisa semantičkoga ponašanja, već kao način objašnjavanja značenjskoga sadržaja na spoznajnoj razini. Koristi se u jezičnom modeliranju, modeliranju koncepata, za semantičko uvjetovanje (engl. *semantic priming*) i učenje riječi (engl. *word learning*).

Kritike distribucijskoga pristupa značenju očituju se u činjenici da pojam semantičke sličnosti definiran u okviru distribucijske hipoteze obuhvaća širok raspon različitih semantičkih relacija, poput sinonimije, antonimije i hiponimije. Ovako definirani koncept kritizira se kao preširok da bi mogao biti koristan, a nemogućnost razliko-

vanja između različitih kategorija semantičke sličnosti poput, primjerice, sinonima i antonima, shvaća se kao veliki nedostatak distribucijske hipoteze (Sahlgren, 2008). Iz normativne perspektive, gdje su navedeni odnosi unaprijed definirani unutar jezične ontologije, takve su kritike opravdane. Međutim, iz deskriptivnoga gledišta, relacije semantičke sličnosti nisu aksiomatske te je ideja širokog shvaćanja pojma sličnosti vjerodostojna i prihvatljiva.

5. Distribucijski semantički modeli

Distribucijski semantički model (DSM) jezični je model u kojemu su riječi predstavljene visokodimenzionalnim vektorima na temelju statističke analize konteksta u kojima se pojavljuju (Stefan Evert, Alessandro Lenci, 2009). Distribucijski semantički modeli teorijsku podlogu nalaze u ditribucijskoj hipotezi i predstavljaju alternativu tradicionalnim simboličkim shvaćanjima koncepta i strukture značenja. Distribucijski pristup semantici naziva se još i korpusnom semantikom (engl. *corpus-based semantics*), statističkom semantikom, geometrijskim modeliranjem značenja i vektorskom semantikom.

5.1. Formalna definicija

Distribucijski semantički model definiran je kao matrica supojavljivanja (engl. *co-occurrence matrix*), odnosno distribucijska matrica M u kojoj svaki redak predstavlja kontekstni vektor, to jest distribuciju ciljnog elementa (riječi, izraza ili dokumenta) u kontekstu (Evert, 2010). DSM se formalno može definirati kao n-torka:

$$DSM = (T, C, R, W, M, d, S),$$

pri čemu su T ciljni elementi, odnosno jezični entiteti čije se značenje modelira, C označava kontekst u kojemu se ciljni elementi promatraju, R je relacija između ciljnih elemenata i konteksta, a W predstavlja sustav težinskoga označavanja elemenata konteksta (engl. *context weighting scheme*). Oznaka M predstavlja distribucijsku matricu $|T| \times |C|$, funkcija $d : M \rightarrow M'$ je funkcija dimenzijske redukcije, dok je S mjera udaljenosti između vektora u M, odnosno M'.

5.2. Izgradnja distribucijskog semantičkog modela

Matrica supojavljivanja izgrađuje se tako da se pri svakom pojavljivanju ciljnog elementa $t_i \in T$ u korpusu zabilježi pojavljivanje svih elemenata c_j koji se nalaze u

kontekstu ciljnog elementa t_i, odnosno za koje vrijedi relacija $R(t_i, c_j)$. U tablici 5.1 prikazan je primjer jednostavne matrice supojavljivanja za morfološki normalizirani korpus.

Tablica 5.1: Jednostavni primjer matrice supojavljivanja

	uzica	*hodati*	*trčati*	*vlasnik*	*noga*	*lajati*
pas	3	5	1	5	4	2
mačka	0	3	3	1	5	0
lav	0	3	2	0	1	0
svjetlost	0	0	0	0	0	0
lajati	1	0	0	2	1	0

Ciljne riječi navedene su u prvom stupcu, jedinice konteksta su riječi, a relacija između ciljnih riječi i kontesta je, primjerice, supojavljivanje unutar iste rečenice.

Izgradnja DSM-a može se detaljnije prikazati u dvije faze, lingvističkoj i matematičkoj. Lingvistička faza obuhvaća pretprocesiranje korpusa, definiranje ciljnih riječi i konteksta te njihov odabir. Matematički koraci uključuju prebrojavanje supojavljivanja ciljnih riječi i konteksta, definiranje i određivanje kontekstnih težina, izgradnju distribucijske matrice, reduciranje dimenzija distribucijske matrice te računanje vektorskih udaljenosti temeljem (reducirane) matrice. Svaki od navedenih koraka određuje širok spektar parametera DSM-a, a odabrani parametri definiraju specifičan tip DSM-a. Odabir parametara značajno utječe na svojstva modela.

5.2.1. Lingvistički koraci izgradnje DSM-a

Pretprocesiranje korpusa nužno uključuje njegovo opojavničenje, a može uključivati i morfosintaktičko označavanje (engl. *POS tagging*), lematizaciju i parsiranje. Međutim, dublja lingvistička obrada rezultira manjom količinom jezičnih specifičnosti, potencijalno unosi dodatne pogreške svakim novim stupnjem obrade te uvodi dodatnu parametarsku slobodu, time povećavajući opsežnost njihova ugađanja. Strategija i razina pretprocesiranja utječe na definiranje ciljnih riječi i odabir vrste konteksta. Kontekst ciljne riječi definira se kao dokument ili riječ.

Odabir konteksta utječe i na definiciju relacije R kojom je kontekst povezan s ciljnim riječima. Ako je kontekst definiran kao dokument ili odlomak, R se definira kao pojavljivanje ciljne riječi u kontekstu C. Ako je kontekst definiran kao riječ, relacija R može biti sintagmatska relacija koja ciljnu riječ povezuje s kontekstom,

poput prozora riječi prije ili poslije ciljne riječi, dijela rečenice, cijele rečenice ili odlomka u kojoj se ciljna riječ pojavljuje, proizvoljne vrste morfosintaktičke veze ili nekog leksičko-sintaktičnog uzorka (Stefan Evert, Alessandro Lenci, 2009). Definira li se kontekst C kao prozor od n riječi prije i poslije ciljne riječi, potrebno je definirati veličinu prozora, oblik prozora, to jest sustav težinskoga označavanja elemenata konteksta (primjerice pravokutni prozor, u kojemu sve riječi imaju jednaku težinu ili trokutni prozor, gdje riječi koje se nalaze bliže ciljnoj riječi imaju veći značaj, odnosno težinu), zatim simetričnost prozora i granicu prozora, odnosno dodatnu suženost prozora s obzirom na eventualne završetke rečenica ili odlomaka. Ako je relacija R definirana kao sintaktična povezanost, potrebno je precizirati vrstu i način povezanosti (primjerice veze poput subjekt-objekt, subjekt-predikat ili predikat-objekt te vrstu veze s obzirom na izravnost).

Najveća razlika između odabranih vrsta konteksta očituje se u prirodi obuhvaćenih semantičkih odnosa, koji s obzirom na veličinu konteksta mogu biti sintagmatski i paradigmatski. Sintagmatski odnosi obuhvaćeni su kontekstima koji su dokumenti, a dvije su riječi tada distribucijski sličnije ako se pojavljuju u više istih dokumenata, dok su odlomcima, rečenicama i riječima obuhvaćeni paradimatski odnosi, gdje je distribucijska sličnost proporcionalna supojavljivanju s istim riječima (Sahlgren, 2008).

U računalnoj lingvistici popularnije je korištenje opsegom manjih i jezično obrađenijih konteksta, dok je u kognitivnoj znanosti zabilježena veća sklonost širim kontekstima kao što su dokumenti.

5.2.2. Matematički koraci izgradnje DSM-a

Nakon određivanja čiste se frekvencije pojavljivanja logaritmiziraju, čime se postiže zaglađivanje drastičnih razlika u broju pojavljivanja. Sama informacija o učestalosti riječi i nakon logaritmiziranja ima dva nedostatka. Prvi je činjenica da samo frekvencije nisu značajne u smislu određivanja jačine povezanosti dviju riječi. Ukoliko su dvije riječi dovoljno učestale, njihovo supojavljivanje može biti sasvim slučajno, bez ikakvog dubljeg semantičkog razloga. Potrebno je stoga statistički interpretirati podatke o frekvencijama kako bi se odredio stupanj statističke povezanosti između riječi (engl. *statistical association between the words*). Drugi nedostatak sakupljenih informacija jest njihova specifičnost za konkretni korpus iz kojega su ekstrahirane, no ne i za prirodni jezik u cijelosti, odnosno njegov dobro definirani podskup. Generalizaciju je moguće ostvariti metodama statističkoga zaključivanja. Najčešće korištene metode za definiranje i izgradnju takvoga statističkog distribucijskog modela u kojemu

je moguće predvidjeti u kolikoj su mjeri supojavljivanja riječi slučajna, a u kojoj su posljedica stvarnih jezičnih ili semantičkih veza te modela na temelju kojega je moguće s određenim stupnjem sigurnosti generaliziranjem protumačiti specifični model kao reprezentativni uzorak jezika ili njegova podskupa objedinjene su pod nazivom mjere povezanosti (Evert, 2005). Mjere povezanosti detaljno su kategorizirane i objašnjene u radu (Evert, 2004).

Računanje udaljenosti između vektora

Prostor u kojemu je definiran koncept udaljenosti naziva se metričkim prostorom. Metrika je mjera udaljenosti $d(u, v)$ između točaka u i v koja zadovoljava sljedeće aksiome (Christopher D. Manning, Hinrich Schutze, 2003):

1. $d(u, v) = d(v, u)$
2. $d(u, v) > 0 \forall u \neq v$
3. $d(u, u) = 0$
4. $d(u, w) \leq d(u, v) + d(v, w)$ (Ovaj se izraz naziva još i nejednakošću trokuta.)

U nastavku su navedene neke često korištene mjere udaljenosti.

Minkowski mjera udaljenosti između dvije točke $P = (x_1, x_2, \ldots, x_n)$ i $Q = (y_1, y_2, \ldots, y_n)$ iz skupa $\mathbb{R}^n$:

$$dist_{minkowski}(P, Q) = \left(\sum_{i=1}^{n} \mid x_i - y_i \mid^p \right)^{\frac{1}{p}} \tag{5.1}$$

Manhattan mjera udaljenosti je Minkowski mjera s parametrom p = 1:

$$dist_{manhattan}(\vec{l_1}, \vec{l_2}) = \sum_{i=1}^{N} \mid l_{1i} - l_{2i} \mid \tag{5.2}$$

Euklidska udaljenost je Minkowski mjera s parametrom p = 2:

$$dist_{euklidska}(\vec{l_1}, \vec{l_2}) = \sqrt{\sum_{i=1}^{N} (l_{1i} - l_{2i})^2} \tag{5.3}$$

Kosinusna udaljenost, kojom se računa sličnost dva n-dimenzionalna vektora na temelju kosinusa kuta koji vektori zatvaraju:

$$sim_{kosinus}(\vec{l_1}, \vec{l_2}) = \frac{\sum_{i=1}^{N} l_{1i} * l_{2i}}{\sqrt{\sum_{i=1}^{N} l_{1i}^2}\sqrt{\sum_{i=1}^{N} l_{2i}^2}} \tag{5.4}$$

Jaccardov indeks ili Jaccardov koeficijent sličnosti, koji predstavlja količnik zajedničkih značajki i ukupnog broja značajki, može se primijeniti za binarne vektore u obliku:

$$sim_{Jaccard\_bin}(\vec{l_1}, \vec{l_2}) = \frac{\mid l_1 \cap l_2 \mid}{\mid l_1 \cup l_2 \mid} \tag{5.5}$$

Ukoliko vektori nisu binarni, već predstavljaju težinske supojavne vektore, koristi se modificirana formula definirana u radu (Curran, 2008):

$$sim_{Jaccard}(\vec{l_1}, \vec{l_2}) = \frac{\sum_{i=1}^{N} min(l_{1i}, l_{2i})}{\sum_{i=1}^{N} max(l_{1i}, l_{2i})} \tag{5.6}$$

Dice mjera sličnosti za binarne vektore dana je formulom:

$$sim_{Dice\_bin}(\vec{l_1}, \vec{l_2}) = \frac{2* \mid l_1 \cap l_2 \mid}{\mid l_1 \mid + \mid l_2 \mid} \tag{5.7}$$

Ako vektori nisu binarni, Dice mjera sličnosti za težinske supojavne vektore dana je modificiranim izrazom definiranim u radu (Curran, 2008):

$$sim_{Dice}(\vec{l_1}, \vec{l_2}) = \frac{2 * \sum_{i=1}^{N} min(l_{1i}, l_{2i})}{\sum_{i=1}^{N} (l_{1i}, l_{2i})} \tag{5.8}$$

5.3. Skalabilnost i dimenzionalnost

Najveći se nedostatak distribucijskih semantičkih modela očituje u skalabilnosti. Dimenzionalnost kontekstnih vektora, odnosno matrice supojavljivanja, ovisna je o broju podataka na temelju kojih se izgrađuje jezični model. Primjerice, ako je kontekst definiran kao dokument, broj dimenzija jednak je broju dokumenata, dok je u slučaju pojedinačnih riječi kao kontekstnih elemenata dimenzionalnost približno jednaka broju jedinstvenih pojavnica. Posljedica prevelikog broja dimenzija jest računska neiskoristivost takvih matrica supojavljivanja.

Drugi problem matrice supojavljivanja jest rijetka popunjenost podacima koje sadrži. Većina je elemenata matrice jednaka nuli, neovisno o veličini korpusa. Rijetkost podataka moguće je objasniti Zipfovim zakonom: vrlo maleni broj riječi pojavljuje se u raznorodnim kontekstima i s velikom učestalošću, dok se većina riječi u jeziku pojavljuje u ograničenim i malobrojnim kontekstima te sa znatno manjom učestalošću

te je u prosjeku 99% elemenata matrice supojavljivanja jednako nuli (Christopher D. Manning, Hinrich Schutze, 2003).

Problem rijetke popunjenosti i prevelike dimenzionalnosti u nekim je postojećim distribucijskim semantičkim modelima, poput LSA modela (Landauer, 1997) i modela Infomap NLP (Widdows, 2004.), riješen dekompozicijom singularnih vrijednosti (engl. *singular value decomposition, SVD*), metodom faktorizacije kojom se matrica dekomponira te potom aproksimira gušćom matricom s mnogo manjim brojem stupaca, najčešće nekoliko stotina. Osim dekompozicije singularnih vrijednosti, dimenzije matrice moguće je smanjiti srodnim statističkim metodama poput analize svojstvenih komponenti (engl. *principal component analysis, PCA*) ili analize nezavisnih komponenti (engl. *independent component analysis, ICA*). Daljna analiza podataka provodi se nad reduciranom matricom.

5.3.1. Dekompozicija singularnih vrijednosti

Dekompozicija singularnih vrijednosti formalno se definira na sljedeći način (Golub i Reinsch, 1970):

Neka je A realna $m \times n$ matrica uz $m \geq n$. Vrijedi

$$A = U\Sigma V^T \tag{5.9}$$

gdje je

$$U^T U = V^T = VV^T = I_n \text{i} \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n).$$

Matrica U sastoji se od n ortogonalnih i normaliziranih svojstvenih vektora povezanih s n najvećih svojstvenih vrijednosti matrice AA^T. Matrica V sastoji se od ortogonalnih normaliziranih svojstvenih vektora matrice $A^T A$. Dijagonalni elementi Σ predstavljaju nenegativne druge korijene svojstvenih vrijednosti matrice $A^T A$ te ih se naziva singularnim vrijednostima. Pretpostavimo da je

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0.$$

Tada, ako vrijedi $rang(A) = r$, mora biti $\sigma_{r+1} = \sigma_{r+2} = \cdots = 0$. Dekompozicija 5.9 je dekompozicija singularnih vrijednosti.

5.4. Nasumično indeksiranje

Svim je tehnikama redukcije dimenzionalnosti navedenima u prethodnom odlomku (skalabilnost, više nije prethodni) (SVD, PCA, ICA) zajednička osnovna metodologija: prvo se svi podaci uzorkuju i na temelju njih se izgradi matrica supojavljivanja koja se tek potom reducira u manji i gušći oblik. Iako su matematički opravdane, statističke metode smanjenja dimenzionalnosti imaju nekoliko nezanemarivih nedostataka, od kojih je prvi zahtjevnost u smislu potrebne radne memorije te vremena izvođenja. Nadalje, redukcija dimenzionalnosti za pojedinu se matricu supojavljivanja provodi jednokratno, što znači da nije moguće naknadno u matricu dodati nove podatke, već se za promijenjeni skup podataka, odnosno nadopunjenu matricu, postupak redukcije mora u cijelosti ponovno provesti. Stvaranje početne prevelike matrice nije moguće zaobići te je nemoguće dobiti ikakve međurezultate prije cjelovitog postupka redukcije.

Alternativni pristup izgradnji reducirane matrice supojavljivanja jest metoda nasumičnog indeksiranja (engl. *random indexing*), odnosno nasumične projekcije (engl. *random projection*) (Kanerva, 1988; Sahlgren, 2005; Bingham, 2001). Tehnika nasumičnog indeksiranja koristi matricu nasumičnih vektora (engl. *random matrix*) M_R dimenzija $|C| \times k$ za projiciranje originalne matrice supojavljivanja M dimenzija $|T| \times |C|$ u reduciranu matricu M' dimenzija $|T| \times k$ ($k \ll |C|$). Retci matrice M_R nazivaju se indeksnim vektorima i dimenzije su k. Svaki indeksni vektor $\mathbf{r}_j$ predstavlja jednu jedinicu, odnosno element konteksta $c_j \in C$. Indeksni vektori rijetko su popunjeni vektori, s malim brojem η nasumično raspoređenih elemenata koji poprimaju vrijednosti $+1$ i -1, dok su ostali elementi jednaki nuli. Svakom je ciljnom elementu t_i pridružen kontekstni vektor $\mathbf{t}_i$, početno nul-vektor, koji se gradi uzimanjem u obzir svih kontekstnih elemenata c_j koji se nalaze u kontekstu ciljnog elementa t_i, odnosno kontekstnih elemenata za koje je relacija $t_i R c_j$ zadovoljena. Za svaki se c_j vektoru $\mathbf{t}_i$ pribraja odgovarajući indeksni vektor $\mathbf{r}_j$.

Matrica izgrađena nasumičnim indeksiranjem aproksimacija je standardne matrice u smislu da su odgovarajući retci u obje matrice međusobno u jednakoj mjeri slični, to jest, udaljenosti između točaka u nasumično razapetom prostoru dovoljno velike dimenzije k očuvane su u usporedbi s na tradicionalan način izgrađenom matricom supojavljivanja.

Budući da je nasumično indeksiranje inkrementalna metoda, jedna od njezinih prednosti jest činjenica da se kontekstni vektori mogu koristiti za računanje sličnosti i prije uzorkovanja cjelokupnoga korpusa. Nadalje, s obzirom na to da je dimenzion-

alnost indeksnih vektora parametar postavljen na početku postupka izgradnje modela, dimenzije matrice supojavljivanja fiksno su određene i ne povećavaju se prilikom prebrajanja konteksta neovisno o broju različitih jedinica konteksta. Budući da je ovime dimenzijska redukcija provedena implicitno, nasumično indeksiranje manje je vremenski i izračunski složeno. Također, nasumično indeksiranje upotrebljivo je neovisno o vrsti, odnosno definiciji konteksta. Skalabilnost nasumičnog indeksiranja dovedena je međutim u pitanje u radu (Gorman i Curran, 2006) u kojemu se na temelju ispitivanja izvedbe metode nasumičnog indeksiranja na većim korpusima zaključuje kako je metoda robustna samo za manje količine podataka, ali i kako se njezina izvedba na većim korpusima može značajno poboljšati upotrebom odgovarajućih funkcija težinskog označavanja konteksta.

6. Postojeći modeli distribucijske semantike

Distribucijski semantički modeli primjenjuju se, između ostaloga, i u zadacima pretrage dokumenata (engl. *document retreival*), ispitivanju sinonimije, razdvajanju značenja višeznačnih riječi (engl. *word sense disambiguation*), ekstrakciji informacija iz dvojezičnih resursa, ocjenjivanju eseja (engl. *essay grading*), vizualizaciji, konstrukciji taksonomije, automatiziranom grupiranju riječi i dr. U poglavlju je dan pregled metodologijom značajnijih distribucijskih semantičkih modela, postupaka njihove izgradnje te načina vrednovanja i primjene. Opisani su i neki aktualni alati za izgradnju distirbucijskih modela.

6.1. Hiperprostorna analogija jeziku

Hiperprostorna analogija jeziku (engl. *Hyperspace Analogue to Language, HAL*) ostvaruje osnovnu metodologiju distribucijske semantike i temeljnu praktičnu potvrdu postojanja informacija o značenju riječi u kontekstu u kojemu se riječ pojavljuje te mogućnosti nenadzirane ekstrakcije takvih semantičkih informacija (Lund, 1995; Lund i Burgess, 1996). Kontekst je definiran kao prozor riječi koje se pojavljuju prije i poslije ciljne riječi. Matrica supojavljivanja oblika je *riječ* $\times$ *riječ*, a kao ciljne riječi i kao riječi konteksta uzimaju se sve riječi koje se pojavljuju u razmatranom korpusu, što znači da se za svaku riječ iz vokabulara može iz matrice supojavljivanja dohvatiti i redak i stupac koji označavaju željenu riječ. Na ovaj način jedna matrica supojavljivanja nudi dvostruku, na neki način recipročnu kontekstnu informaciju: primjerice, ako je relacija ciljne riječi i konteksta prozor on n riječi nakon ciljne riječi, tada će retci matrice supojavljivanja sadržavati informaciju u supojavljivanju ciljne riječi i riječi koje dolaze nakon nje, dok će stupci matrice sadržavati informaciju o supojavljivanju ciljne riječi i riječi koje se u korpusu nalaze prije nje. Frekvencije pojavljivanja riječi nisu obrađene ni na koji način, razmatrane dimenzije su one s najvećom vari-

jancom, a kao mjera za računanje udaljenosti između dva kontekstna vektora uzeta je Euklidska udaljenosti i udaljenost Manhattan. Vrednovanje je pokazalo da je model bolji u prepoznavanju paradigmatskih nego u otkrivanju sintagmatskih odnosa. Ovaj je model značajan kao vrlo rana i prototipna implementacija distribucijske hipoteze.

6.2. Latentna semantička analiza

Latentna semantička analiza (engl. *Latent Semantic Analysis, LSA*) distribucijski semantički model temelji se na klasičnoj matrici supojavljivanja, u kojoj retci predstavljaju ciljne elemente, primjerice riječi ili dokumente, a stupci kontekste s kojima su ciljni elementi u definiranoj relaciji, na primjer, paragrafi ili dokumenti u kojima se ciljni elementi pojavljuju (Landauer, 1997). Elementi matrice predstavljaju broj jedinica konteksta s kojima su ciljni elementi u relaciji, odnosno frekvencije supojavljivanja konteksta i ciljnih elemenata. Takve sirove vrijednosti zatim su transformirane funkcijom

$$f(freq) = \frac{ln(1 + freq)}{\textit{entropija ciljne riječi kroz sve kontekste}} \tag{6.1}$$

Nad matricom supojavljivanja provedena je nadalje dekompozicija singularnih vrijednosti, na temelju koje je izdvojeno n (Landauer (1997) navodi 300) stupaca s obzirom na singularnu vrijednost dimenzije, uzimajući stupce s najvećim singularnim vrijednostima, odnosno one koji u originalnoj matrici imaju najveću varijancu.

Dimenzijska redukcija značajan je korak u izgradnji modela i ključan za njegove performanse. Reducirani ciljni kontekstni vektor računa se SVD-om kao linearna kombinacija svih elemenata matrice. Bolje rečeno, reducirani vektor nije određen samo distribucijom ciljnog elementa koji je vektorom predstavljen (kao što je to slučaj kod originalnog, nereduciranog kontekstnog vektora), već SVD koristi sve linearne relacije određene dimenzionalnosti za računanje kontekstnih vektora koji će najbolje modelirati uzorke u kojima se ciljni element pojavljuje. Promjena jedne vrijednosti u originalnoj matrici rezultirat će stoga promjenom svih koeficijenata u svim kontekstnim vektorima. Reducirana matrica sadrži informacije neizravno zaključene na temelju originalne matrice.

Nakon redukcije dimenzije matrice supojavljivanja dekompozicijom singularnih vrijednosti svaki je ciljni element predstavljen n-dimenzionalnim vektorom. Udaljenost između vektora moguće je računati za sve moguće parove redaka i stupaca matrice supojavljivanja, odnosno moguće je računati udaljenost između dva ciljna ele-

menta, između ciljnog elementa i konteksta ili između dva kontekstna elementa. Za računanje udaljenosti koristi se kosinusna mjera udaljenosti.

Model latentne semantičke analize analizom pojmova koji se pojavljuju unutar dokumenata konstruira skup koncepata koji opisuju apstrahirano značenje ciljnih elemenata (LSA, 2010). Model u potpunosti zanemaruje višeznačnost riječi, a koncepti kojima se ciljni elementi opisuju uzorci su riječi koje se supojavljuju unutar dokumenata. Poredak riječi potpuno se zanemaruje (Landauer, 1997).

Za vrednovanje modela iz zadataka u *Test of English as a Foreign Language* (TOEFL) uzeto je 80 njih koji provjeravaju poznavanje značenja riječi. Za pojedinu problemsku riječ ponuđene su četiri riječi od kojih je potrebno odabrati onu koja je značenjem najbliža ili ista problemskoj riječi. Modelu su ponuđene problemske riječi i četiri potencijalna sinonima za svaku. Za sva četiri para kontekstnih vektora izračunata je kosinusnom mjerom udaljenost između riječi. Na temelju izračunate udaljenosti za pojedinu problemsku riječ odabran je njezin sinonim, odnosno bliskoznačnica.

Model je sinonime odredio točno za 51.5 od 80 zadataka, odnosno 64.4% (52.2% uzimajući u obzir pogađanje). Za usporedbu, veliki uzorak TOEFL testova s istim zadacima koje su za prijavu na američka sveučilišta rješavali ljudi kojima engleski jezik nije materinji u prosjeku za 80 zadataka daje 51.6 točno riješenih, odnosno 64.5% (52.7% uzimajući u obzir pogađanje). U prosjeku, korelacija između krivo odabranih sinonima za LSA i ljudske testove iznosi 0.44.

Zanimljiva je paralela u radu (Landauer, 1997) povučena između LSA i ljudskog učenja značenja novih riječi. Što je više jezičnih resursa dostupnih modelu LSA, to je više elemenata (riječi, dokumenata) čije se značenje može (bolje) modelirati. Ovo se uspoređuje s čovjekom, koji čitanjem uči značenje novih riječi, odnosno, što više čita, to više novih riječi poznaje. Čitanjem, čovjek susreće nove riječi čije mu je značenje nepoznato, ali unutar konteksta koji može razumjeti te na temelju takovoga razumljivog konteksta izvodi značenja dotad nepoznatih riječi. Kao i u distribucijskom semantičkom modelu, nove riječi nije potrebno eksplicitno definirati.

U radu (Phil Katz, 2008) opisano je korištenje modela latentne semantičke analize kao dodatnog klasifikatora u sustavu za razdvajanje značenja višeznačnih riječi, međutim izvedba modela LSA bila je sustavno lošija od drugih klasifikatora.

Modeli LSA korišteni su i u zadacima pretrage dokumenata (engl. *document retrieval*), kako bi se proširili sustavi temeljeni isključivo na podudaranju ključnih riječi. Nadogradnja se očituje u korištenju LSA za dohvaćanje latentnih, neizravnih distribucijskih sličnosti između dokumenata koji se pretražuju i kriterija pretrage (S. Deerwester, 1990; D. Widows, 2008). Usprkos određenim poboljšanjima, nedostaci takve

nadogradnje očituju se u nemogućnosti pouzdanog i konzistentnog poboljšanja preciznosti i odziva, neovisno o veličini i broju testnih podataka.

6.3. Vektori ovisnosti

Radno okruženje za izgradnju distribucijskih semantičkih modela temeljenih na vektorima ovisnosti (engl. *dependency vectors*) specifično je po definiciji relacije u kojoj se nalaze ciljne riječi i riječi iz konteksta, a koja u obzir uzima i sintaktičke odnose između riječi, formirajući time distribucijski model na temelju dodatnih lingvističkih informacija (Pado i Lapata, 2007). Na ovaj se način isključivo distribucijski pristup semantici spaja s formalnim sintaksnim modelima, a izgrađeni kontekstni vektori lingvistički su vjerodostojniji budući da njima oblikovana apstrakcija značenja ne ovisi samo o leksičkom supojavljivanju, nego uključuje i sintaksnu ovisnost između ciljnog izraza i konteksta.

Distribucijski semantički modeli obogaćeni sintaktičkim informacijama ispitani su na zadacima razrješavanja višeznačnosti, detekcije sinonima i semantičkog uvjetovanja (engl. *semantic priming*) i njihova je izvedba u svim slučajevima nadmašila izvedbu modela koji zanemaruju sintaktičke informacije. Nadogradnja sustava ostvarena je u radu (Erk i Padó, 2008), uz naglasak važnosti sintaktičkih informacija za konstrukciju robustnih i kvalitetnih sustava za oblikovanje značenja rečenice kompozicijom kontekstnih vektora.

6.4. Infomap NLP i Semantički vektori

Infomap NLP[1] je programski paket otvorenog koda koji omogućuje konstrukciju distribucijskih semantičkih modela korištenjem latentne semantičke analize. Kontekst je definiran kao prozor od n riječi, a udaljenost između kontekstnih vektora računa se kosinusnom mjerom udaljenosti. Paket je ostvaren u programskom jeziku C i može obrađivati velike korpuse. Međutim, matrice supojavljivanja izgrađene paketom Infomap NLP prvotno su vrlo velikih dimenzija i s neobrađenim učestalostima, a nakon uzorkovanja cjelokupnog korpusa minimiziraju se dekompozicijom singularnih vrijednosti. Ovaj je algoritam računski i resursno zahtjevan, teško ga je paralelizirati i onemogućava inkrementalnu izgradnju distribucijskog semantičkog modela, što rezultira lošim svojstvima modela u smislu skalabilnosti. Dokumenti obrađeni kao korpusni

[1] `http://infomap-nlp.sourceforge.net/`

materijal indeksiraju se te se modele izgrađene paketom Infomap NLP može koristiti za pretraživanje informacija (engl. *information retrieval*) i računanje semantičke sličnosti između dvije riječi. Razvoj paketa obustavljen je, a iz njega je proizašao paket Semantičkih vektora (engl. *Semantic Vectors*), pisan programskim jezikom Java.

Semantički vektori[2] omogućuju konstrukciju vektorskog semantičkog prostora, odnosno kontekstnih (semantičkih) vektora za ciljne riječi i dokumente iz korpusa novinskih članaka (D. Widows, 2008). Izgradnja vektora ostvarena je algoritmom nasumičnog indeksiranja, što uvelike doprinosti skalabilnosti izgrađenih distribucijskih semantičkih modela, a paket ostvaruje i funkcionalnost latentne semantičke analize. Dodatno, paket ostvaruje funkcionalnost pretraživanja izgrađenih kontekstnih vektora.

6.5. Distribucijska memorija

Model distribucijske memorije (engl. *Distributional Memory*) specifičan je pristup korpusnoj semantici koji omogućuje rješavanje različitih semantičkih zadataka na temelju jednog repozitorija koji sadrži informacije o distribucijskim svojstvima obrađenoga korpusa. Model je detaljnije opisan u radovima (Baroni i Lenci, 2009; Baroni, 2010a). Informacije o distribucijskim svojstvima korpusa bilježe se u obliku težinski označenih n-torki strukture *riječ – veza – riječ*. Takve n-torke tvore tenzore trećeg reda iz kojih se generiraju matrice čiji retci i stupci oblikuju vektorske prostore koji obuhvaćaju različite semantičke informacije. Na ovaj su način iste distribucijske informacije dostupne zadacima poput određivanja semantičke sličnosti riječi, otkrivanja sinonima i kategorizacije koncepata. Izvedba modela distribucijske memorije ne zaostaje za izvedbom zadatkovno specifičnih modela.

6.6. Primjena u razdvajanju značenja višeznačnih riječi

Distribucijski semantički modeli uspješno se primjenjuju i u razdvajanju značenja višeznačnih riječi (Schutze, 1998; Yarowsky, 1995). Osnovni algoritam temelji se na značenjskoj usporedbi riječi predstavljenih kontekstnim vektorima, ali na temelju supojavljivanja drugog reda: dvije jedinice konteksta *A* i *B* višeznačne riječi pridjeljuju se istoj značenjskoj grupaciji ako se riječi s kojima se *A* i *B* supojavljuju u korpusu

[2] `http://code.google.com/p/semanticvectors/`

također pojavljuju sa sličnim kontekstima u korpusu za učenje. Algoritam je automatiziran i nenadziran u primjeni i učenju budući da se o smislovima zaključuje isključivo iz korpusa i bez dodatnih informacija.

6.7. Distribucijski semantički modeli u slavenskim jezicima

Većina ostvarenih distribucijskih semantičkih modela izgrađena je za engleski jezik. Međutim, slavenski se jezici morfološki i leksički nezanemarivo razlikuju od engleskoga, prvenstveno u bogatsvu fleksije, slobodnijem redoslijedu riječi u rečenici, brojnijim morfološkim oblicima i glasovnim promjenama. Distribucijski semantički modeli engleskoga jezika nisu stoga u potpunosti i bez prilagodbi primjenjivi na slavenske jezike. Osim toga, jezični resursi i alati za slavenske jezike nerijetko su malobrojni (Broda i Piasecki, 2008), što može onemogućiti ili ograničiti fazu pripremne obrade korpusa prije izgradnje modela. Slavenski jezici za koje su ostvareni distribucijski semantički modeli su bugarski, češki, poljski i ruski jezik.

Model latentne semantičke analize za književni korpus bugarskoga jezika opisan je u radu (Nakov, 2001a). LSA model koristi se za usporedbu bugarskih književnih tekstova, preciznije, za automatizirano otkrivanje tekstova istoga autora te za automatizirano razlikovanje tekstova koji pripadaju različitim književnim razdobljima. Model je izgrađen na temelju 3032 teksta grupirana u četiri književna razdoblja, a koje je napisalo 48 autora. Tekstovi istih autora formirali su jasno odvojene grupe u korelacijskoj matrici te je tekstove bilo moguće automatski pridijeliti odgovarajućem autoru. Očekivano, performanse su bolje za autore s više tekstova u modelu te za autore specifičnoga stila pisanja i vokabulara. Tekstove autora sličnog stila i rječnika nije bilo moguće automatski razlikovati. Zadatak automatiziranoga razlikovanja tekstova iz različitih književnih razdoblja nije ostvario zadovoljavajuće performanse, a poboljšanje izvedbe predloženo je proširenjem korpusa tekstovima iz vremenski udaljenijih razdoblja.

Distribucijski semantički model za češki jezik ostvarili su Smrž i Rychlỳ (2001). Središnji zadatak rada je automatizirano grupiranje značenjski povezanih riječi, uz što se istražuju performanse mjera semantičke povezanosti između riječi te načini obrade kolokacija. Na temelju hijerarhijskog grupiranja velikog broja semantički povezanih riječi dani su, u obliku dendrograma, primjeri mogućih načina podjele riječi u grupe. Ostvareno je i sučelje za lakšu anlizu dobivenih grupa riječi.

Distribucijska semantika za poljski jezik istražena je u radovima (Piasecki, 2009; Broda et al., 2008; Broda i Piasecki, 2008) i usko je vezana za izgradnju i optimizaciju WordNet-a za poljski jezik. Preciznije, osnovna je primjena ostvarenih distribucijskih semantičkih modela i alata temeljenih na njima automatizirana ekstrakcija semantičkih relacija koja olakšava pronalazak sinonima ili bliskoznačnica, antonima, hiponima i dr. potrebnih za definiciju i opis riječi u okviru WordNet-a. Skup programskih biblioteka i korisničkih alata za izgradnju, pohranjivanje i obradu matrica supojavljivanja distribucijskih semantičkih modela za poljski jezik, *SuperMatrix*, opisan je u radu (Broda i Piasecki, 2008). U okviru sustava ostvarena je biblioteka za pohranu matrica, biblioteka za računanje semantičke sličnosti između supojavnih vektora korištenjem različitih mjera semantičke sličnosti, odnosno vektorske udaljenosti, zatim alati za transformaciju, smanjivanje dimenzionalnosti i spajanje matrica te pregledavanje sadržaja pohranjenog u matrici, biblioteka za grupiranje, alati za vrednovanje i dr. Sustav se prvenstveno upotrebljava kao pomoćni alat za izgradnju WordNet-a. U radu (Piasecki, 2009) uspoređen je distribucijski pristup ekstrakciji semantičkih relacija s pristupom temeljenim na leksičkim uzorcima (primjerice, uzorak "**A***, odnosno* **B**" može označavati sinonimiju). Zbog izražene sklonidbe riječi u poljskom jeziku, naglašena je važnost lematizacije u pripremi korpusa. Također, učinjen je kritički osvrt na poteškoće pri vrednovanju modela, s naglaskom na nedostatke ljudske interpretacije rezultata opisane nužno pristranom i selektivnom. Razvijen je i alat koji na temelju distribucijskog pristupa semantici za određenu riječ iz korpusa izdvaja skup značenjski povezanih riječi, predstavljajući dopunsko sredstvo za konstrukciju WordNet-a za poljski jezik koje funkcionira kao automatizirani sustav preporuka leksikografima potencijalno zanimljivih pojmova (Broda et al., 2008). Sustav je učinkovit za glagole i pridjeve te nešto slabijih performansi za imenice.

Za ruski jezik ostvareni su modeli latentne semantičke analize koji se koriste za grupiranje riječi i analizu književnih tekstova. U radu (Mitrofanova, 2007) opisan je postupak automatskog grupiranja riječi iz tekstova kombiniranjem algoritama grupiranja (aglomerativni i *K-means*) i modela latentne semantičke analize, uz konstrukciju matrica supojavljivanja i analizu konteksta. Slično kao i za bugarski jezik, Nakov (2001b) latentnu semantičku analizu koristi za razlikovanje ruskih tekstova različitih autora te automatizirano razlikovanje proze i poezije. Uz identičnu metodologiju kao u radu (Nakov, 2001a), zadatak razlikovanja autora ostvaruje prihvatljive rezultate, uz iznimke za pojedine autore, dok je razlikovanje proze i poezije iznimno uspješno.

6.8. Distribucijski semantički modeli u hrvatskome jeziku

Jedini distribucijski semantički modeli dosad ostvareni za hrvatski jezik opisani su u radu (Nikola Ljubešić, Damir Boras, Nikola Bakarić, Jasmina Njavro, 2008), u kojemu su izneseni rezultati uspoređivanja različitih metoda određivanja semantičke sličnosti iz korpusa. Automatsko određivanje semantičke sličnosti temelji se isključivo na distribucijskoj hipotezi, a svaki je leksem predstavljen kao vektor u kojemu je sadržana informacija o frekvenciji supojavnih leksema (vektor supojavljivanja, odnosno kontekstni vektor). Semantička udaljenost računa se kao udaljenost između tako definiranih vektora u tri osnovne faze izgradnje distribucijskog semantičkog modela: (1) izgradnja supojavnih vektora, (2) računanje veze s kontekstom, (3) računanje sličnosti vektora.

Eksperiment je proveden na Vjesnikovu korpusu koji sadrži članke iz Vjesnika u razdoblju od 1999. do 2007. godine. Korpus je morfosintaktički označen korištenjem označivača opisanoga u radu (Ž. Agić, M. Tadić, 2006), sadrži 79 566 904 pojavnica, 3 730 729 rečenica, 1 300 785 odlomaka i 205 686 članaka. Prva faza eksperimenta, izgradnja supojavnih vektora, definirana je na dva skupa leksema, $V1$ i $V2$. Skup $V1$ predstavlja ciljne lekseme, odnosno one koji su predstavljeni supojavnim vektorima, dok se u skupu $V2$ nalaze svi leksemi iz korpusa koji su utjecali na izgradnju supojavnih vektora.

Matrica supojavljivanja izgrađena je za 1000 najčešćih imenica, izostavivši prvih stotinu, uz definiciju konteksta kao odlomka u kojemu se pojavljuje ciljni leksem. Od tisuću ciljnih leksema, nasumično je odabrano njih pet: *ustav*, *istup*, *suđenje*, *serija* i *prihod*. Primijenjene mjere povezanosti s kontekstom su frekvencija, procijenjena maksimalna izglednost, uzajamna obavjesnost i t-test. Korišteno je osam mjera za izračunavanje sličnosti vektora: udaljenost Manhattan, euklidska udaljenost, kosinusna ualjenost, binarna Jaccardova udaljenost, Jaccardova udaljenost, binarna Dice mjera udaljenosti, Dice mjera udaljenosti te Jensen-Shannonova divergencija. Budući da binarna Jaccardova udaljenost i binarna Dice mjera udaljenosti zahtijevaju binarne vektore, prethodno nije moguće primijeniti mjere povezanosti s kontekstom. Za svaki od pet odabranih leksema, odnosno odgovarajućih supojavnih vektora, provedeno je stoga $2 + 6 * 4 = 26$ različitih mjerenja.

Vrednovanje je ostvareno na temelju standarda koji je odredilo troje ocjenjivača. Za svaki ciljni leksem definiran je popis leksema potencijalno značenjski povezanih s njime kao unija prvih dvadeset leksema koje je svih dvadeset šest metoda rangiralo

kao semantički najsličnije ciljnom leksemu. Za svaki ciljni leksem određeno je u prosjeku 78 potencijalno sličnih leksema koje su ocjenjivači ocijenili ocjenama od 1 do 4. Podudaranje dvaju ocjenjivača računa se po formuli:

$$IAA(\vec{g_1}, \vec{g_2}) = 1 - \frac{\sum_{i=1}^{N} \mid g_{1i} - g_{2i} \mid}{\sum_{i=1}^{N} 3} \tag{6.2}$$

Standard za vrednovanje svake od 26 metoda određivanja sličnosti sastoji se od izabranih pet leksema i svakome od njih pridijeljenih semantički sličnih leksema te njihovih ocjena sličnosti. Ocjene su izračunate kao srednja vrijednost dobivenih pojedinačnih ljudskih ocjena. Pri vrednovanju metoda za određivanje sličnosti koristi se funkcija gubitka (engl. *loss function*) definirana formulom:

$$L(\vec{r}, \vec{g}) = \sum_{i=1}^{N} \frac{4}{i} - \sum_{i=1}^{N} \frac{g_i}{r_i} \tag{6.3}$$

Funkcija gubitka temelji se na inverzu ranga i koristi vektor rangova dobiven metodom koja se vrednuje i vektor ocjena dobiven na temelju izračunatoga standarda. Funkcija predstavlja razliku između maksimalne vrijednosti (slučaj u kojemu svi leksemi dobivaju maksimalnu ocjenu 4) i vrijednosti dobivene standardom. Budući da se svaka od ove dvije vrijednosti računa kao količnik ocjene i ranga, nagrađeni su slučajevi u kojima leksemi s visokim ocjenama imaju i visoke rangove.

Od mjera sličnosti vektora najboljima su se pokazale Jensen-Shannonova, Manhattan i euklidska udaljenost. Od ispitanih mjera povezanosti s kontekstom najbolji su rezultati postignuti s procijenjenom maksimalnom vjerojatnosti. Prema pretpostavci autora, sofisticiranije mjere poput t-testa ili uzajamne obavjesnosti pokazale su se lošijima zbog toga što su informacije o supojavljanju dobivene samo na temelju najčešćih imenica. Metode računanja sličnosti vrlo sličnima su ocijenile lekseme koji su od ljudskih ocjenjivača dobili većinom srednje ocjene sličnosti, najčešće 2 i 3. Autori takav rezultat tumače širinom prozora kojim je definiran kontekst (odlomkom) te predlažu ispitivanje s užim prozorom postavljajući teoriju da će u tom slučaju automatske metode računanja sličnosti pronaći više snažnijih bliskoznačnica i istoznačnica. Predlažu daljne eksperimentiranje uz promjenu širine kontekstnoga prozora te uz veći broj leksema koji se uzimaju u obzir. Također predlažu parsanje ili djelomično parsiranje (engl. *light parsing, chunking*) korpusa te, pri izgradnji supojavnih vektora, uzimanje u obzir i sintaktičkih relacija. U radu nije napravljen osvrt na specifičnosti hrvatskog jezika u odnosu na modele izgrađene za druge jezike.

7. Modeli distribucijske semantičke složivosti

Značenje složenih izraza predstavljeno formalnom logikom može se definirati kao funkcija značenja pojedinih elemenata izraza i relacija kojima su elementi povezani. Nedostatak takvog pristupa jest kvalitativno, a ne kvantitativno shvaćanje razlika u značenju te ograničena izražajnost stupnjeva sličnosti. Za razliku od formalne logike, distribucijski semantički modeli značenje modeliraju kvantitativno te omogućuju uočavanje stupnjeva jačine semantičke sličnosti. Međutim, oblikuju značenje samo pojedinačnih riječi. Nadogradnja distribucijskog semantičkog modela s idejom oblikovanja značenja sintagmi ili rečenica opisana je u radovima (Jeff Mitchell, Mirella Lapata, 2008; Widdows, 2008; Guevara, 2010, 2011; Baroni, 2010b). Budući da je značenje pojedinačne riječi određeno njezinim kontekstnim vektorom, značenje složenoga izraza oblikuje se kao funkcija kontekstnih vektora riječi od kojih se izraz sastoji. Dosad najčešće korišten način kompozicije konteksnih vektora jest računanje geometrijske sredine sastavnih vektora. Metoda nije osjetljiva na poredak riječi ili bilo kakve sintaksne strukture i odnose te različite složene izraze koji se sastoje od istih riječi prikazuje na jednak način. Zbog toga bi, primjerice, rečenice *Ana voli Marka* i *Marko voli Anu* (uz prethodnu morfološku normalizaciju) rezultirale istim prikazom. Mitchell i Lapata predlažu dva nova načina oblikovanja semantičke složivosti, aditivni i multiplikavitni model.

7.1. Aditivni i multiplikativni modeli složivosti

Semantička složivost definirana je kao funkcija dvaju vektora, **u** i **v**. Vektori predstavljaju kontekstne prikaze pojedinačnih riječi oblikovane jednom od prethodno opisanih metoda izgradnje distribucijskog semantičkog modela. Oznaka **p** predstavlja kompoziciju vektora **u** i **v** koji označavaju dvije sastavne riječi povezane proizvoljnom jezičnom relacijom R. Oznaka K predstavlja dodatno znanje o načinu slaganja višer-

ječnog izraza. Općenita definicija modela semantičke kompozicije u okviru DSM-a je (Jeff Mitchell, Mirella Lapata, 2008):

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}, R, K) \tag{7.1}$$

Navedeni izraz pretpostavlja da se **p** nalazi u istom prostoru kao i **u** i **v**, a obuhvaća modele koji koriste dodatne informacije K o kompoziciji te modele u kojima postoji ovisnost kompozicije o sintaksi, R. Parametar R fiksiran je, čime se model složivosti usredotočava samo na jednu dobro definiranu jezičnu strukturu, primjerice odnos predikat-subjekt. Ako se parametar K ignorira, model semantičke kompozicije svodi se na

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}) \tag{7.2}$$

Za cjelovitu definiciju modela potrebno je odrediti funkciju f. U radu (Jeff Mitchell, Mirella Lapata, 2008) definirani su aditivni i multiplikativni model. Aditivni model pretpostavlja da **p** leži u istom prostoru kao i **u** i **v** te definira f kao linearnu funkciju:

$$\mathbf{p} = \mathbf{Au} + \mathbf{Bv} \tag{7.3}$$

Oznake **A** i **B** predstavljaju matrice koje određuju intenzitet doprinosa svakog od vektora **u** i **v** kompoziciji **p**. Multiplikativni model definira f kao funkciju vektorskog produkta vektora **u** i **v**. Matrica **C** ranga je 3 i projicira vektorski produkt **u** i **v** u prostor kompozicijskog produkta **p**.

$$\mathbf{p} = \mathbf{Cuv} \tag{7.4}$$

Oba je modela moguće dodatno pojednostaviti smanjujući broj slobodnih parametara. Na temelju pretpostavke da na i-tu komponentu vektora **p** utječu samo i-te komponente vektora **u** i **v**, da elementi nisu ovisni o indeksu i te da je funkcija simentrična u smislu zamjene vektora **u** i **v**, dobivaju se jednostavniji izrazi adititvnog i multiplikativnog modela:

$$p_i = u_i + v_i \tag{7.5}$$

$$p_i = u_i \cdot v_i \tag{7.6}$$

U tablici 7.1 prikazan je hipotetski distribucijski semantički model za riječi *konj* i *trčati*.

Tablica 7.1: Hipotetski distribucijski semantički model za riječi *konj* i *trčati*

	životinja	*staja*	*selo*	*galop*	*džokej*
konj	0	6	2	10	4
trčati	1	8	4	4	0

Semantička kompozicija kontekstnih vektora tih dviju riječi korištenjem aditivnog modela opisanog formulom 7.5 daje rezultat **konj + trčati** $= [1, 14, 6, 14, 4]$, dok multiplikativni model sažet u 7.6 oblikuje vektor **konj**·**trčati** $= [0, 48, 8, 40, 0]$. Iako se u literaturi češće upotrebljava 7.5, u radu (Jeff Mitchell, Mirella Lapata, 2008) navodi se kako je iz lingvističke perspektive zanimljiviji model 7.6. Naime, zbrajanjem vektora njihovi se konteksti naprosto gomilaju, dok se ni na koji način na temelju sadržaja jednoga vektora ne može odabrati relevantan dio sadržaja drugog vektora. Model 7.6 postiže upravo skaliranje doprinosa i-te komponente vektora **u** ovisno o njezinoj važnosti za **v** i obrnuto. Budući da je pretpostavljena simetrija u smislu zamjene vektora **u** i **v**, oba modela neosjetljiva su na poredak riječi. Odbaci li se navedena pretpostavka, formula jednostavnog aditivnog modela 7.5 mijenja se u:

$$p_i = \alpha u_i + \beta v_i \tag{7.7}$$

Težine komponenti sada su različite, što povećava razinu osjetljivosti na sintaksu izraza budući da semantički važniji elementi mogu imati veću važnost u kompoziciji. Pretpostavi li se da i-ta komponenta vektora **u** i **v** ne utječe samo na i-tu komponentu vektora **p**, već i na druge komponente, kružnom konvolucijom multiplikativni se model mijenja u oblik:

$$p_i = \sum_j u_j \cdot v_{i-j} \tag{7.8}$$

Također, moguće je uvesti na početku zanemarenu ovisnost modela o dodatnim informacijama K. U aditivnom se modelu dopunske informacije dodaju zbrajanjem dodatnih kontekstnih vektora koji su u nekoj relaciji s riječima promatranog višerječnog izraza. Predlaže se uključivanje n distribucijski najsličnijih kontekstnih vektora:

$$p_i = \mathbf{u} + \mathbf{v} + \sum \mathbf{n} \tag{7.9}$$

Potencijalni nedostatak ovako definiranih multiplikativnih modela očituje se u elementima vektora jednakima nuli, s obzirom na to da se množenjem s nulom nužno gubi dio

informacija. Ovu je pojavu moguće ublažiti kombiniranjem multiplikativnog i aditivnog modela:

$$p_i = \alpha u_i + \beta v_i + \gamma u_i v_i \tag{7.10}$$

Oznake α, β i γ predstavljaju težinske koeficijente, ali u radu (Jeff Mitchell, Mirella Lapata, 2008) nije opisan način njihova određivanja.

7.2. Vrednovanje aditivnog i multiplikativnog modela složivosti

Osnovna ideja vrednovanja opisanih modela distribucijske semantičke složivosti oslanja se na mogućnost predočavanja modeliranih višerječnih izraza jednom riječi, koja je značenjski što bliža sintagmi. Primjerice, sintagma *Sveti Otac* istoga je značenja kao i riječ *papa*. Iz ovoga slijedi da će metoda kompozicije biti bolja ako je njome ostvareni složeni vektor bliži jednorječnom sinonimu modelirane sintagme. Opisanim načinom vrednovanja utvrđeno je kako su multiplikativni modeli uspješniji od aditivnih (Jeff Mitchell, Mirella Lapata, 2008).

7.3. Tenzorski produkt

U radu (Widdows, 2008) kao kompleksniji model složivosti od aditivnog i multiplikativnog predlaže se tenzorski produkt (engl. *tensor product*), definiran kao

$$\mathbf{p} = \mathbf{u} \otimes \mathbf{v} \tag{7.11}$$

Oznaka $\mathbf{p}$ predstavlja matricu čiji je element ij jednak umnošku $\mathbf{u}_i \times \mathbf{v}_j$. Vrednovanje metode ostvareno je jednostavnim i opsegom malenim eksperimentima ekstrakcije relacija između riječi te modeliranjem kompozicije glagola i objekta. Modelirana relacija semantički je odnos između grada i države u kojoj se on nalazi, dok se kompozicijom modelira sličnost između sintagmi (odnosno ostvaruje model u kojemu bi, primjerice, sintagma *jesti jabuku* trebala biti sličnija sintagmi *jesti rajčicu* nego sintagmi *baciti jabuku*). Slični eksperimenti s tenzorskim produktom ostvareni su i u radu (Giesbrecht, 2009), a oba rada potvrđuju kako performanse modela ostvarenoga na temelju tenzorskog produkta nadmašuju one multiplikativnih i aditivnih modela. Potrebno je napomenuti kako tenzorski produkt nije moguće izravno uspoređivati s

multiplikativnim i aditivnim, s obzirom na to da, za razliku od multiplikativnog i aditivnog koji kao rezultat kompozicije daju vektor smješten u istom vektorskom prostoru kao i izvorni vektori, tenzorski produkt rezultira matricom, odnosno većim brojem dimenzija u odnosu na polazne vektore.

Svim navedenim pristupima zajedničko je dobivanje informacija o složenom vektoru $\mathbf{p}$ isključivo na temelju izvornih vektora $\mathbf{u}$ i $\mathbf{v}$, odnosno riječi koje tvore sintagmu, dok se informacije na razini same sintagme ne uzimaju u obzir (ni u kojoj fazi ne koriste se informacije o distribuciji sintagme u korpusu). Osim toga, sve navedene metode kompozicije temelje se na jednokratnom provođenju jedne geometrijske operacije nad izvornim vektorima, što daje opravdanja dovođenju u pitanje toga koliko je vjerojatno da se samo jednom transformacijom mogu vjerno obuhvatiti sve semantičke promjene u sintagmi, a u odnosu na gradivne riječi, neovisno o sintaktičkim odnosima ili jezičnim specifičnostima.

7.4. Modeli složivosti temeljeni na strojnom učenju

Drugačiji pristup modeliranju distribucijske semantičke složivosti opisan je u radovima (Baroni, 2010b; Guevara, 2010) i može ga se shvatiti kao nadogradnju opisanih metoda koje u obzir ne uzimaju složeni vektor $\mathbf{p}$, odnosnu sintagmu koja se modelira. Osim za gradivne riječi, primjerice *lijepa* i *kuća*, Guevara (2010) gradi i kontekstni vektor za dobivenu sintagmu *lijepa kuća*. Koristeći ove podatke, model kompozicije sintagmi oblika *pridjev – imenica* ostvaruje se primjenom nadziranog strojnog učenja, konkretno multivarijantnom multiplom linearnom regresijom metodom najmanjih kvadrata (engl. *multivariate linear regression analysis by partial least squares*). Navedenom metodom moguće je naučiti funkciju kompozicije koja najbolje aproksimira $\mathbf{p}$ na temelju $\mathbf{u}$ i $\mathbf{v}$. Nešto drugačija metodologija korištena je u radu (Baroni, 2010b). Pretpostavljeno je da je svaki pridjev zapravo funkcija linearne transformacije, to jest ona funkcija koju je algoritmom potrebno naučiti, zbog čega se kompozicija imenice i pridjeva modelira aproksimacijom vektora $\mathbf{p}$ samo na temelju $\mathbf{u}$ (kontekstnog vektora imenice), ali uz provođenje dodatnih regresijskih analiza za svaki pridjev u modelu.

Opisani pristup iz rada (Guevara, 2010) zapravo je samo proširenje aditivnog modela iz rada (Jeff Mitchell, Mirella Lapata, 2008) pronalaskom optimalnih vrijednosti skalarnih težinskih faktora za $\mathbf{u}$ i $\mathbf{v}$ pomoću nadziranog učenja linearnom regresijom. Daljna nadogradnja ideje ostvarena je u radu (Guevara, 2011), gdje se sintagme oblika *pridjev – imenica* i *glagol – imenica* oblikuju također aditivnim modelom uz multivarijantnom multiplom linearnom regresijom naučene težinske faktore, ali s njihovim

proširenjem u težinske matrice.

Multivarijantna multipla linearna regresija metodom najmanjih kvadrata prilagođena je upotrebi u slučajevima problematičnima zbog dimenzionalnosti. Ova metoda regresije predviđa matricu Y na temelju informacija u ulaznoj matrici X, ali i samoj izlaznoj matrici Y. Kovarijanca između X i Y pokušava se objasniti na temelju skupa latentnih varijabli koje istovremeno dekomponiraju obje matrice. Koristeći se dekompozicijom matrice X, regresijom se predviđa Y, a konačni model predviđanja dobiva se ekstrakcijom latentnih varijabli iz kojih je moguće izvući najviše predikcijskih informacija. Tehnika je robustna i posebno učinkovita u slučajevima s velikim brojem prediktora (engl. *predictors*), ali malo opažanja (engl. *observations*) (Guevara, 2011).

7.5. Značenjski neprozirne sintagme

Frazemi su ustaljene sveze riječi koje se upotrebljavaju u gotovu obliku, a ne stvaraju se u tijeku govornoga procesa, i kod kojih je bar jedna sastavnica promijenila značenje, tako da značenje frazema ne odgovara zbroju značenja njegovih sastavnica (Menac, 2003). Važne su odrednice frazema ekspresivnost, slikovitost i konotativnost.

Frazemi se mogu kategorizirati u tri osnovna oblika. Najzastupljeniji su frazemi koji imaju oblik *sveze riječi*, pri čemu su bar dvije sastavnice samostalne i naglašene riječi; veza među njima može biti zavisna, kao na primjer u frazemima *zlatni rudnik*, *rame za plakanje*, *dirati stare rane*, ili nezavisna, na primjer u frazemima *vedriti i oblačiti*, *ni živ ni mrtav*, *milom ili silom*.[1] Manje zastupljeni u jeziku su frazemi koji imaju oblik *fonetske riječi*, tj. kod kojih je samo jedna sastavnica samostalna i naglašena, a druga ili druge dvije su nesamostalne nenaglašene riječi, na primjer *iza rešetaka*, *ne bez razloga*, *ni u ludilu*. Treći tip su frazemi koji imaju oblik rečenice, na primjer *ne cvjetaju ruže komu*, *vrag ne spava*, *makar sjekire padale s neba*, *trla baba lan da joj prođe dan*.

Frazemi, osobito oni u prva dva spomenuta osnovna oblika, najčešće se uvrštavaju u rečenicu kao njezin sastavni dio, i to u imeničkim funkcijama kao subjekt, objekt ili predikat, u glagolskoj predikatnoj funkciji i u priložnim funkcijama, ali mogu biti i izvan rečenice, kao replika u vezi s prethodnim tekstom (*za babino brašno* - replika na pitanje *zašto?*), kao uzvučni frazem (*glavu gore!*, *jezik za zube!*, *ni govora!*) i dr.

Frazemi mogu biti različita podrijetla. Mnogi su nastali frazeologizacijom slobodnih sveza pa se paralelno mogu rabiti slobodna sveza (u kojoj sastavnice zadržavaju

[1] Ovi primjeri, primjeri u nastavku te cjelokupna taksonomija preuzeti su iz Hrvatskog frazeološkog rječnika (Menac, 2003)

svoje leksičko značenje) i frazem s istim sastavom (ali s promjenom značenja bar jedne sastavnice), npr. *imati prazan džep (džepove)*, *zagristi u kiselu jabuku*. Izvor su nekim frazemima Biblija ili drugi crkveni tekstovi, npr. *glas vapijećeg u pustinji*, *dolina suza*, *nositi svoj križ*, *mana s neba*. Neki frazemi potječu iz književnih djela, npr. *biti ili ne biti*, *ružno pače*, *posljednji Mohikanac*, neki mogu imati u svom sastavu poneki podatak iz povijesti, npr. *kocka je bačena*, ili poneki zemljopisni naziv, npr. *Martin u Zagreb, Martin iz Zagreba*. Neki su frazemi nastali na osnovi izraza iz raznih područja ljudske djelatnosti koji su proširili svoje značenje, npr. na osnovi termina nekih znanosti (*svesti na zajednički nazivnik*, *izazivati lančanu reakciju*, *iks puta*), termina iz glazbe (*prva violina*), iz kazališta (*igrati glavnu ulogu*, *iza kulisa*) ili iz sporta (*nizak udarac*, *greška u koracima*).

U kontekstu distribucijskog pristupa značenju, iz perspektive distribucijske semantičke složivosti, frazemi su specifični upravo zbog toga što njihovo značenje ne odgovara zbroju značenja sastavnih riječi. Na temelju ovoga dovodi se u pitanje primjena modela složivosti koji uspješno modeliraju značenje semantički prozirnih sintagmi na značenjski neprozirne sintagme, odnosno frazeme. Ako postoji konzistentna razlika u izvedbi, nju je moguće iskoristiti za detekciju frazema. U okviru modela složivosti ostvarenih u ovome radu ispitana je i kvaliteta izvedbe modela složivosti na frazemskim sintagmama.

8. Ostvareni distribucijski semantički modeli

8.1. Korpus

Korpus na temelju kojega su izgrađeni distribucijski semantički modeli u okviru ovog diplomskog rada jest kolekcija novinskih članaka iz Vjesnika[1] od 1999. do 2009. godine. Korpus se sastoji od 276 231 dokumenata i 85 000 različnica. Prije izgradnje modela, korpus je obrađen: sva velika slova pretvorena su u mala (engl. *case folding*), izbačeni su svi interpunkcijski znakovi, drugi posebni (nealfanumerički) znakovi i znamenke te zaustavne riječi. Iz korpusa su zatim izbačene i sve riječi kraće od tri slova, nakon čega je korpus lematiziran (Šnajder). U slučaju više mogućih lema, uzima se prva. Nakon lematizacije iz korpusa su izbačene sve riječi koje se pojavljuju najviše dvaput. Ovako obrađeni korpus sastoji se od 49 000 različnica i 155 274 000 pojavnica.

8.2. Modeli

Za ciljne riječi odabrano je 185 najučestalijih riječi iz korpusa. Ukupno je ostvareno 350 distribucijskih semantičkih modela: 70 frekvencijskih modela i 280 modela nasumičnog indeksiranja. Modeli se međusobno razlikuju u odnosu na vrstu konteksta i mjeru sličnosti između kontekstnih vektora ciljnih riječi, a modeli nasumičnog indeksiranja dodatno se razlikuju i u dimenziji k indeksnoga vektora.

Kontekst je definiran kao rečenica, simetrični i asimetrični prozor. Simetrični i asimetrični prozori obuhvaćaju 5, 10 i 20 riječi s obje strane ciljne riječi u slučaju simetričnog prozora, odnosno 5, 10 i 20 riječi s lijeve ili desne strane kad je riječ o asimetričnom prozoru. Dodatno, rečenica i simetrični prozor mogu biti težinski oz-

[1] www.vjesnik.hr

načeni ili ne, dok se asimetrični prozor težinski ne označava. Težinsko označavanje jedinica konteksta ostvareno je dvostrukom L funkcijom:

$$ll(x,\alpha,\beta,\gamma,\delta) = \begin{cases} l(x,\alpha,\beta), & x \geq 0 \\ l(-x,\gamma,\delta), & x < 0 \end{cases}, \quad l(x,\alpha,\beta) = \begin{cases} 1 & x < \alpha \\ \frac{\beta - x}{\beta - \alpha} & \alpha \leq x \leq \beta \\ 0 & \beta < x \end{cases} \quad (8.1)$$

Na primjer, težinski faktor za elemente konteksta definiranoga kao simetrični težinski označen prozor od pet riječi slijeva i pet riječi zdesna dohvaća se funkcijom $ll(x, 1, 5, 1, 5)$. Dimenzija k indeksnih vektora kod modela nasumičnog indeksiranja poprima vrijednosti 100, 200, 500 i 1000, uz $\eta = 2$, 4, 6 i 8. Za svih 350 modela izračunata je sličnost za ukupno $\binom{185}{2} = 17020$ parova ciljnih riječi (svaki par predstavljen je, dakle, parom kontekstnih vektora). Sličnosti odnosno udaljenosti između kontekstnih vektora, dobivene su sljedećim mjerama udaljenosti: Manhattan mjera udaljenosti, euklidska udaljenost, kosinusna udaljenost te Diceov i Jaccardov koeficijent vektorske udaljenosti. Za posljednje dvije mjere korištene su modificirane formule 5.6, 5.8 definirane u radu (Curran, 2008). Sve ocjene sličnosti parova skalirane su zatim na interval $[1, 5]$, gdje 5 označava najveću sličnost.

8.3. Vrednovanje

Model je vrednovan za ručno odabranih 450 (od ukupno 17 020 mogućih) parova usporedbom s ocjenama sličnosti dobivenima na temelju ocjena ljudskih ocjenjivača. Šesnaest ocjenjivača ocijenilo je sličnost, odnosno jačinu semantičke povezanosti 450 parova riječi ocjenama od 1 do 5, gdje 5 označava najveću sličnost. Koncept semantičke povezanosti definiran je kao unija paradigmatskih semantičkih odnosa i sintagmatskih odnosa, ali i tvorbenih te širih asocijativnih veza među riječima. Paradigmatski semantički odnosi obuhvaćaju antonimiju, hiponimiju, kohiponimiju, meronimiju i sinonimiju. Sintagmatski odnosi obuhvaćaju leksikalizirane izraze poput idioma, složenih imenica i klišeja, odnosno, na općenitijoj razini, sve riječi koje se često koriste zajedno, to jest često zajedno tvore sintagme. To su najčešće parovi atributa (ili apozicije) i imenice te glagola i objekta. Šira asocijativna veza označava povezanost na značenjskoj, smisaonoj razini (primjerice riječi *čitanje* i *knjiga*) te je ujedno i najširа, najsubjektivnije shvaćena kategorija. Ako u paru koji se ocjenjuje postoji višeznačna riječ, ocjenjivači su upućeni odabrati ono značenje koje će potencirati jačinu veze s drugom riječi u paru. Na ovaj način pokušalo se unificirati interpretaciju višeznačnih

parova, odnosno doskočiti situaciji u kojoj će neki ocjenjivači višeznačne parove ocjenjivati značajno drugačije nego drugi. Potonje se može dogoditi u dva osnovna slučaja. Prvo, odaberu li ocjenjivači različite interpretacije višeznačnih riječi, jačinu povezanosti riječi u paru interpretirat će na nužno drugačiji način. Drugo, moguće je da ocjenjivač procijeni kako je riječ višeznačna i zbog toga snagu semantičke veze ocijeni slabijom ocjenom. U ovom slučaju problem predstavlja subjektivna procjena mjere u kojoj višeznačnost oslabljuje jačinu značenjske veze. Čak i uz ovako definirane upute, ocjenjivanje snage semantičkih veza iznimno je subjektivan zadatak koji uvelike ovisi o odabranom skupu ocjenjivača.

Ukupno slaganje svih ocjenjivača te slaganje između parova ocjenjivača izračunato je korištenjem Fleissove kappe (Davies i Fleiss, 1982):

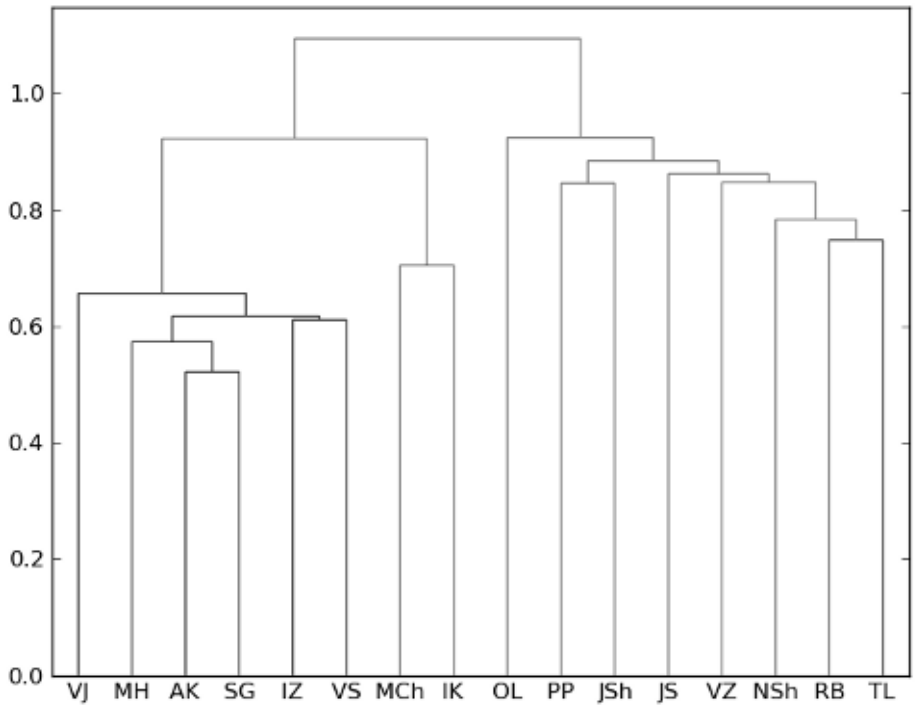
$$\kappa_f = \frac{p_a - p_r}{1 - p_r},$$
$$p_a = \frac{\sum_{i=1}^{j} \sum_{j=1}^{m} K_{ij}^2 - nK}{nK\,(K-1)}, \quad p_r = \sum_j p_j^2, \quad p_j = \frac{1}{nK} \sum_{n=1}^{i} K_i j \tag{8.2}$$

Oznaka p_a predstavlja uočeni stupanj slaganja, a p_r je procijenjeno slučajno slaganje. Broj ocjenjivača označen je s K, n označava broj zadataka, a m broj kategorija.

Na temelju izračunatog slaganja parova ocjenjivača dobivena je matrica slaganja (dimenzija 16×16) koja je zatim hijerarhijski aglomerativno grupirana (engl. *hierarchical agglomerative clustering*) uz povezivanje na temelju prosjeka (engl. *average linkage*) korištenjem programske biblioteke *hcluster: Hierarchical Clustering for SciPy* (Eads, 2008).

Dobiveni dendrogram prikazan je na slici 8.1 (oznake na x-osi predstavljaju inicijale ocjenjivača). Na temelju podataka dobivenih grupiranjem izdvojene su grupe od 12 i 6 ocjenjivača s jačim međusobnim slaganjem. Ukupno slaganje grupe od 12 ocjenjivača je $\kappa = 0.27$, interpretirano kao umjereno slaganje, odnosno $\kappa = 0.35$ za 6 ocjenjivača, također umjereno slaganje. Ocjene obiju grupa ocjenjivača uprosječene su, čime su dobivena dva vektora sličnosti koji su definirani kao zlatni standard i zatim korišteni za usporedbu sa strojno generiranim ocjenama sličnosti.

Od svih 450 parova, sljedeći parovi riječi ocijenjeni su najsnažnije povezanima (u grupi od 6 ocjenjivača ocijenjeni su ocjenom 5.00 uz devijaciju 0.0): *domaći – stran*, *država – državni*, *grad – građanin*, *grad – gradski*, *igra – igrač*, *igra – igrati*, *jak – snaga*, *kuna – novac*, *mali – velik*, *ministar – ministarstvo*, *mlad – star*, *momčad – tim*, *nov – star*, *početak – kraj*, *politika – politički*, *sud – sudac*, *sud – zakon*, *svijet – svjetski*, *zemlja – država*. Zanimljivo je primijetiti kako su riječi u čak devet od



Slika 8.1: Slaganje ocjenjivača mjereno Fleissovom kappom prikazano hijerarhijskim grupiranjem

navedenih devetnaest parova čvrsto tvorbeno povezane, dok su ostali parovi u odnosu sinonimije i antonimije. U tablici 8.1 prikazano je odabranih trideset parova riječi te prosječna ocjena i standardna devijacija procijenjene jakosti semantičke relacije para za obje grupe ocjenjivača. Parovi su poredani po prosječnoj ocjeni za grupu od 6 ocjenjivača. Osim tvorbenih veza, antonimije i sinonimije, u prosjeku visoke ocjene dobivali su parovi koji ostvaruju odnos hiponimije, odnosno kohiponimije. Sintagmatski su odnosi ocijenjeni neznatno slabijim ocjenama, najviša prosječna ocjena koji su dobili sintagmatski povezani parovi jest 4.5 (parovi *francuski – film*, *glavni – grad*, *pravi – trenutak*). Očekivano, standardne devijacije grupe od 6 ocjenjivača u pravilu su manje od devijacija grupe od 12 ocjenjivača.

8.4. Rezultati

Od ukupno 17 020 parova riječi, za svaki od 350 generiranih modela izdvojen je za 450 ciljnih parova vektor pripadajućih skaliranih ocjena sličnosti, odnosno jačine semantičke povezanosti. Izdvojenih 350 vektora sličnosti uspoređeno je s oba standardna vektora sličnosti dobivena na temelju ljudskih ocjena. Vektori su uspoređivani pomoću

formule za srednju kvadratnu pogrešku (engl. *mean square errror, MSE*):

$$MSE\left(\mathbf{v}_1, \mathbf{v}_2\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{v}_1\left[i\right] - \mathbf{v}_2\left[i\right]\right)^2 \tag{8.3}$$

Razlika između standardnih ocjena povezanosti riječi u parovima i ocjena generiranih od strane modela, odnosno iznos srednje kvadratne pogreške, bit će manji za modele koji oblikuju koncept distribucijske semantičke povezanosti sličniji ljudskom poimanju značenjske sličnosti.

Izabrani rezultati, poredani po rastućoj veličini srednje kvadratne pogreške, prikazani su u tablici 8.2. Vrsta modela definira model nasumičnog indeksiranja (RI-k), gdje k predstavlja dimenziju indeksnoga vektora, te frekvencijski model (Raw). Kontekst je definiran kao rečenica (R), težinski označena rečenica (Rt), prozor (P) ili težinski označen prozor (Pt). Lijeva i desna dimenzija prozora, x i y, u tablici su prikazane kao xL–yD. Prikazani rezultati uključuju najbolje performanse modela za svaku mjeru sličnosti te najbolji frekvencijski model.

Iz rezultata se zaključuje kako modeli nasumičnog indeksiranja performansama nadmašuju frekvencijske modele. Najbolji frekvencijski model rangiran je kao 31. pri usporedbi sa standardnim vektorom dobivenim od 6 ocjenjivača, odnosno 42. pri usporedbi s vektorom od 12 ocjenjivača. Za svaku dimenziju k indeksnoga vektora postoji model nasumičnog indeksiranja čije su performanse bolje od svih frekvencijskih modela. Ovakvi rezultati potvrđuju zaključke iz radova (Broda i Piasecki, 2008; Landauer, 1997) o različitoj količini informacija sadržanoj u kontekstima, odnosno njihovim učestalostima, te postojanju skrivenih, latentnih veza između kontekstnih vektora koje je moguće naglasiti tek postupcima matričnih transformacija (primjerice, dekompozicijom singularnih vrijednosti).

Iz rezultata se također zaključuje kako se performanse modela nasumičnog indeksiranja nužno ne poboljšavaju s povećanjem dimenzije k, odnosno da ne postoji jasna korelacija između dimenzije modela i njegove izvedbe. Sve četiri dimenzije $k = 100, 200, 500, 1000$ indeksnoga vektora pojavljuju se u deset najbolje rangiranih, ali i u deset najgore rangiranih modela.

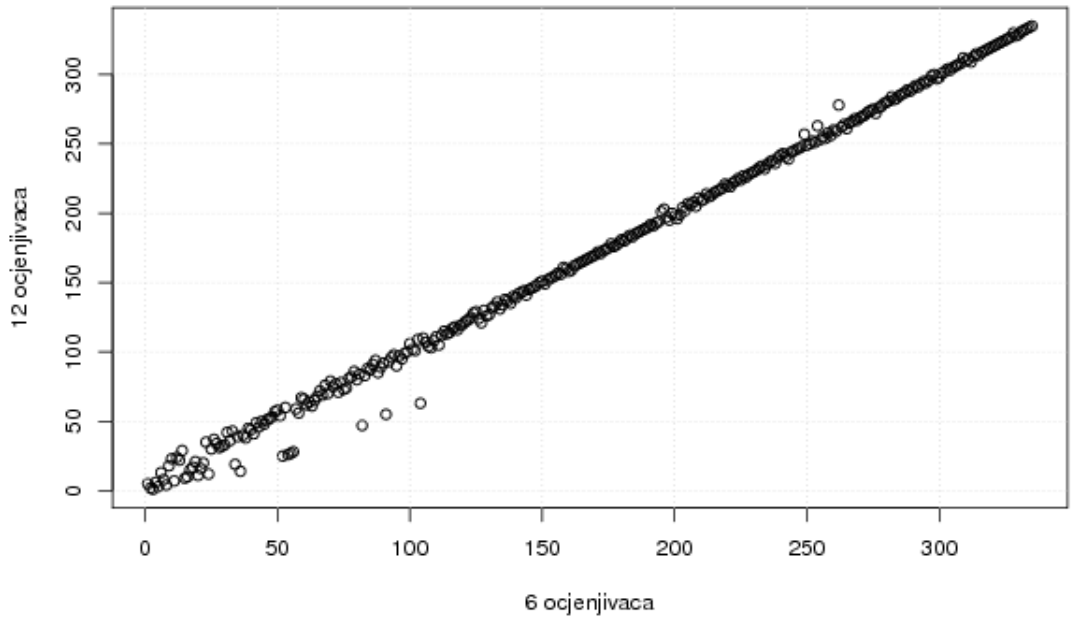
Odabir mjere udaljenosti između kontekstnih vektora pokazao se kao parametar sa značajnim utjecajem na performanse modela. Najbolji modeli kao mjeru udaljenosti koriste Diceov i Jaccardov koeficijent: najboljih 58 modela koristi isključivo Diceov i Jaccardov koeficijent, dok od najboljih 100 modela njih 93 koristi Diceov i Jaccardov koeficijent. Međutim, i najgorih deset modela također koristi isključivo Diceov koeficijent. Modeli koji koriste kosinusnu mjeru udaljenosti nalaze se u sredini rang

liste: najviši rang modela s kosinusnom udaljenošću je 58, najniži 185, a modeli rangirani između 147. i 185. mjesta koriste isključivo kosinusnu udaljenost. Udaljenost Manhattan i euklidska udaljenost u pravilu daju lošije rezultate, što se ne podudara s rezultatima iz rada (Nikola Ljubešić, Damir Boras, Nikola Bakarić, Jasmina Njavro, 2008). Najbolji model koji koristi euklidsku udaljenost rangiran je kao 186., a najbolji Manhattan model je na 190. mjestu.

Postojanje korelacije između vrste konteksta i performansi modela nije bilo moguće utvrditi. Naime, najbolji modeli značajno se razlikuju u vrsti konteksta, dok se istovremeno modeli slično definiranoga konteksta pak značajno razlikuju u svojoj izvedbi. Primjerice, u dvadeset najbolje rangiranih modela dvanaest je različitih definicija konteksta. S druge strane, kontekst najboljeg modela definiran je kao težinski neoznačeni prozor koji u obzir uzima samo pet riječi desno od ciljne riječi. Modeli s istim takvim kontekstom, ali s drugim kombinacijama ostalih parametara, rangirani su kao 13., 16., 43., 47., 65., 70., 71., 79., 87., 97., 99., 100., 103., 132., 199., 204., 209., 217., 219., 220., 224., 235., 244. i 256.

U tablici 8.3 prikazano je trideset parova riječi ocijenjenih najsnažnije značenjski povezanima u dva modela, modelu nasumičnog indeksiranja dimenzije 100 s kontekstom definiranim kao težinski neoznačeni prozor od pet riječi nakon ciljne riječi te Diceovim koeficijentom kao mjerom udaljenosti (RI–100, P–0L–5R, Dice) i modelu nasumičnog indeksiranja dimenzije 500 s kontekstom definiranim kao težinski neoznačen prozor od 20 riječi prije i poslije ciljne riječi i Diceovim koeficijentom kao mjerom udaljenosti (RI–500, P–20L–20R, Dice). U usporedbi sa zlatnim standardom od šest ocjenjivača, rezultati vrednovanja pokazali su kako je prvi model najbolji, a drugi osmi po redu. Međutim, uspoređivanjem najjače povezanih parova kod ljudi i kod modela, teško je bilo što zaključiti, a iako se određene pravilnosti mogu izdvojiti, upitna je njihova reprezentativnost i opravdanost na njima temeljenih poopćavanja (Piasecki, 2009). Riječi u parovima kojima su ocjenjivači dali najviše ocjene u znatnoj su mjeri u relaciji sinonimije i antonimije te derivacijski povezane. Među navedenim parovima riječi za prvi model u velikoj su mjeri obuhvaćeni parovi koji ostvaruju semantičku relaciju kohiponimije i hiponimije (dvanaest parova), sinonimi, odnosno bliskoznačnice slabije su zastupljeni (pet parova), dok je samo jedan par riječi antonim i jedan par u sintagmatskom odnosu. Od najjače povezanih 30 parova za drugi model, šest parova su bliskoznačnice, tri para su hiponimi, a četiri para nalaze se u sintagmatskom odnosu. Neke od ostalih parova s obje liste moguće je okarakterizirati kao asocijativno povezane, ali kod većine ostalih ne postoji jasno prepoznatljiva relacija semantičke povezanosti.

Izračunato je i Kendall Tau podudaranje između rangova modela uspoređenih sa zlatnim standardom dobivenim od 6, odnosno od 12 ocjenjivača i ono iznosi $\tau = 0.9784$, što je iznimno visoko podudaranje i iz čega se zaključuje kako ne postoji značajna razlika između dva korištena načina vrednovanja. Razlika između rangova modela prikazana je na slici 8.2, gdje se može uočiti kako je neslaganje najprimjetnije za bolje, odnosno više rangirane modele.



Slika 8.2: Kendall Tau podudaranje rangova modela vrednovanih usporedbom sa 6 i 12 ocjenjivača

Pri interpretaciji rezultata potrebno je u obzir uzeti dva faktora koji na njih utječu. Prvo, prilikom izgradnje kontekstnog vektora za višeznačnu riječ, distribucijski semantički model ni na koji način ne diferencira različite načine, odnosno značenja, na koje se ciljna riječ može interpretirati, nego sva moguće interpretacije (to jest, njihove kontekste) spaja u jedan kontekstni vektor. Nasuprot tome, ljudi odabiru jednu interpretaciju višeznačne riječi. Takva razlika u pristupu višeznačnim riječima nužno unosi pogrešku pri vrednovanju modela, budući da se modeli uspoređuju s ljudskim ocjenama. Nadalje, bitno je napomenuti kako distribucijski semantički modeli modeliraju značenje isključivo na temelju korpusa koji se koristi za njihovu izgradnju te da je korpus korišten u izgradnji modela u okviru ovog diplomskog rada domenski specifičan, novinski korpus koji nije reprezentativan prikaz hrvatskoga jezika u cjelini. Izgrađeni modeli, dakle, oblikuju domenski specifičan koncept značenja.

8.5. Programsko ostvarenje

Projekt je ostvaren u programskom jeziku Python 2.7.[2] Struktura programskoga ostvarenja može se prema fazama izgradnje modela ugrubo podijeliti u četiri cjeline: (1) obrađivanje i priprema korpusa, (2) izgradnja modela, (3) obrada ocjena dobivenih od ljudi te (4) vrednovanje modela. Shema projekta prikazana je na slici 8.3.

Prva faza izgradnje modela jest priprema korpusa. Dio projekta koji ostvaruje tu funkcionalnost sastoji se od skripti s funkcijama za pretvaranje cjelokupnoga korpusa u mala slova, izbacivanje interpunkcije i ostalih nealfanumeričkih znakova i znamenaka te funkcijama za filtriranje zaustavnih riječi na temelju popisa danog u [REF] i riječi kraćih od tri slova. U ovoj fazi izgradnje modela provodi se i lematizacija korpusa, pomoću programskog paketa *fer-lematizator.jar* (Šnajder), ostvarene u programskom jeziku Java. U slučaju višeznačne lematizacije, odnosno više ponuđenih lema, zbog nepostojanja drugih jezičnih informacija o korpusu uzima se prva lema. Pogreška unesena na ovaj način zanemaruje se. Nakon lematizacije, skriptom za dobivanje osnovnih statističkih podataka o korpusu dobiva se informacija o broju rečenica, pojavnica i različnica te popis svih različnica u korpusu zajedno s pripadajućim učestalostima pojavljivanja. Na temelju dobivenoga popisa iz korpusa se filtriraju riječi koje se pojavljuju manje od tri puta. Nakon obrade korpusa na temelju osvježenog popisa različnica i njihovih učestalosti, 185 najčešćih različnica odabrano je za ciljne riječi. U ovom dijelu projekta ostvarene su i pomoćne funkcije za baratanje korpusom.

Druga faza jest sama izgradnja modela za ciljne riječi izdvojene u prethodnome koraku. U ovom dijelu programskoga ostvarenja nalaze se funkcije i strukture podataka kojima se definira vrsta konteksta i funkcija za njegovo težinsko označavanje te ostvaruje dohvaćanje konteksta za pojedinu ciljnu riječ. Ostvarena je biblioteka za izgradnju kontekstnoga vektora za frekvencijski model i za model nasumičnog indeksiranja, uz pomoćne funkcije za generiranje indeksnih vektora za modele nasumičnog indeksiranja. Skripta koja ostvaruje funkcionalnost glavnog programa, odnosno izgradnje svih modela, ostvarena je u dva dijela, od kojih je prvi izgradnja kontekstnih vektora za sve željene vrste konteksta, a drugi računanje udaljenosti između kontekstnih vektora korištenjem pet mjera vektorske udaljenosti.

Treća cjelina projekta ostvaruje funkcionalnosti za sakupljanje i obradu ljudskih ocjena semantičke povezanosti parova ciljnih riječi. U programskom jeziku Java ostvareno je jednostavno grafičko sučelje za ljudsko ocjenjivanje parova. Sakupljene ocjene analizirane te su dobivene informacije poput prosječne ocjene koju je dao poje-

[2] `http://www.python.org/`

dini ocjenjivač te prosječne ocjene i standardne devijacije pojedinog para riječi. Ostvareno je nekoliko funkcija za računanje podudaranja između ocjenjivača (engl. *interannotators accordance, agreement*), Pearson, Fleiss, Kendall Tau, od kojih je na posljetku korištena samo funkcija koja računa podudaranje između n ocjenjivača formulom Fleissove kappe. Podudaranje je izračunato za sve parove ocjenjivača te je korištenjem biblioteke *hcluster* (Eads, 2008) ostvareno hijerarhijsko aglomerativno grupiranje uz povezivanje na temelju prosjeka. Na temelju informacija dobivenih grupiranjem odabrana su dva skupa vektora ljudskih ocjena koji sadrže ocjene 12 i 6 ocjenjivača. Dva zlatna standarda dobivena su uprosječivanjem odabranih skupova vektora.

Vrednovanje je ostvareno skriptom koja iterira kroz ostvarene modele, odnosno kroz vektore ocjena semantičke povezanosti u svakom modelu, te između njih i svakoga od dva zlatna standarda računa srednju kvadratnu pogrešku.

Tablica 8.1: Odabrane ljudske ocjene semantičke povezanosti parova riječi

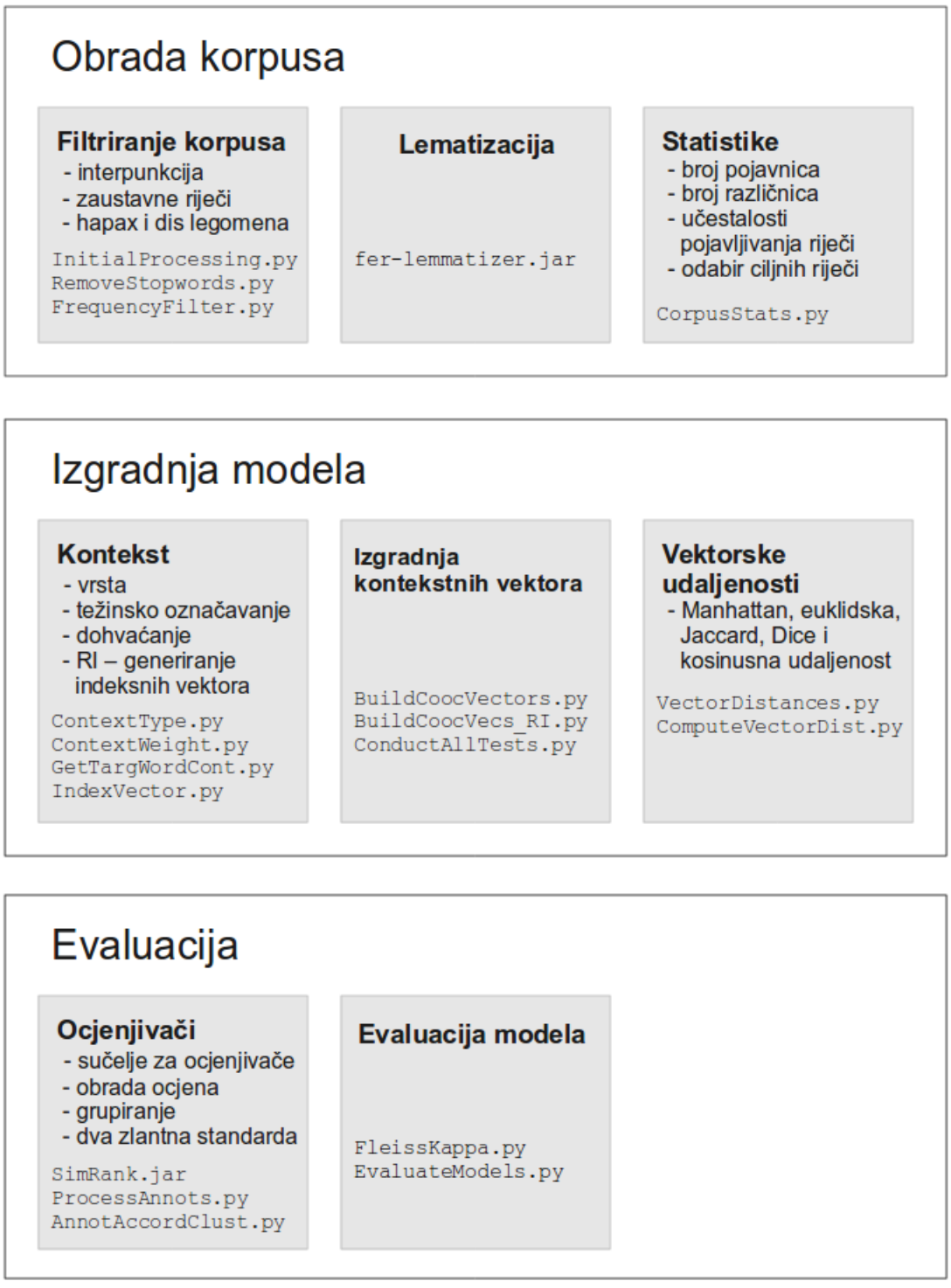
	Prosječna ocjena	
Par riječi	6 ocjenjivača	12 ocjenjivača
politika – politički	5.0 ± 0.0	4.9 ± 0.3
igra – igrač	5.0 ± 0.0	4.8 ± 0.4
kuna – novac	5.0 ± 0.0	4.8 ± 0.4
početak – kraj	5.0 ± 0.0	4.7 ± 0.5
zemlja – država	5.0 ± 0.0	4.6 ± 0.8
reći – govoriti	4.8 ± 0.4	4.8 ± 0.4
imati – nemati	4.8 ± 0.4	4.7 ± 0.5
banka – novac	4.8 ± 0.4	4.5 ± 0.5
utorak – tjedan	4.8 ± 0.4	4.2 ± 1.2
djeca – škola	4.7 ± 0.5	4.1 ± 0.8
dolar – novac	4.7 ± 0.5	4.7 ± 0.5
kuna – dolar	4.7 ± 0.5	4.3 ± 0.8
pravi – trenutak	4.5 ± 0.6	3.9 ± 0.8
sat – tjedan	4.5 ± 0.6	4.1 ± 0.7
mlad – mjesec	4.3 ± 0.5	3.7 ± 0.9
razgovor – izjaviti	4.3 ± 0.8	4.2 ± 0.7
smatrati – misliti	4.3 ± 1.2	4.3 ± 0.9
imati – obzir	4.2 ± 0.8	3.8 ± 0.8
dan – vrijeme	4.0 ± 0.6	4.0 ± 0.6
domaći – svjetski	3.8 ± 1.5	3.8 ± 1.0
vojni – snaga	3.8 ± 0.8	3.9 ± 0.8
europa – postupak	3.0 ± 0.6	2.2 ± 1.0
momčad – obitelj	2.8 ± 0.8	2.3 ± 0.9
pitanje – područje	2.7 ± 1.0	2.0 ± 1.1
trenutak – dionica	2.7 ± 1.2	2.2 ± 1.0
trebati – kultura	2.2 ± 0.9	1.8 ± 0.9
trenutak – sudac	2.0 ± 0.9	1.5 ± 0.8
sustav – izjaviti	1.7 ± 1.2	1.5 ± 0.9
trenutak – građanin	1.2 ± 0.4	1.2 ± 0.4

Tablica 8.2: Srednja kvadratna pogreška za odabrane distribucijske semantičke modele

Model			Srednja kvadratna pogreška (rang)			
DSM	Kontekst	Mjera sličnosti	6 ocjenjivača		12 ocjenjivača	
RI–100	**P–0L–5D**	**Dice**	1.96	**(1)**	1.68	(5)
RI–500	P–20L–0D	Dice	1.96	(2)	1.65	(2)
RI–100	P–0L–10D	Dice	1.98	(3)	1.64	**(1)**
RI–500	**R**	Dice	1.98	(4)	1.69	(6)
RI–200	**Pt–5L–5D**	Dice	2.00	(5)	1.65	(3)
RI–200	P–5L–0D	Dice	2.01	(6)	1.77	(13)
RI–100	P–5L–0D	Dice	2.05	(7)	1.7	(8)
RI–500	P–20L–20D	Dice	2.09	(8)	1.68	(4)
RI–1000	P–20L–0D	Dice	2.09	(9)	1.95	(18)
RI–1000	P–0L–20D	Dice	2.09	(10)	1.96	(23)
RI–200	P–10L–10D	Dice	2.10	(11)	1.70	(7)
RI–100	Pt–10L–10D	Dice	2.11	(12)	1.98	(24)
RI–1000	P–0L–5D	**Jaccard**	2.13	(13)	1.96	(22)
RI–1000	P–5L–0D	Jaccard	2.16	(14)	2.01	(29)
RI–500	P–10L–0D	Dice	2.17	(15)	1.71	(9)
RI–500	**Rt**	Jaccard	2.24	(25)	2.03	(30)
Raw	Pt–5L–5D	Jaccard	2.27	(31)	2.15	(42)
RI–100	R	Jaccard	2.35	(40)	2.22	(44)
Raw	P–0L–5D	Jaccard	2.47	(47)	2.4	(52)
RI–200	Rt	Dice	2.59	(55)	2.01	(27)
RI–200	Pt–5L–5D	**Cosine**	2.60	(59)	2.72	(67)
Raw	Pt–5L–5D	Cosine	2.61	(60)	2.71	(66)
Raw	R	Jaccard	3.03	(81)	3.07	(84)
RI–200	Pt–5L–5D	**Euclidean**	5.97	(186)	6.37	(186)
RI–200	Pt–5L–5D	**Manhattan**	6.23	(190)	6.66	(192)
RI–200	R	Euclidean	6.50	(222)	6.97	(222)
Raw	Rt	Manhattan	6.59	(237)	7.06	(238)
RI–1000	P–10L–0D	Euclidean	6.69	(250)	7.15	(249)
RI–100	P–5L–5D	Manhattan	6.75	(258)	7.24	(258)
RI–100	Rt	Dice	7.65	(327)	8.28	(327)
RI–1000	R	Dice	7.69	(328)	8.35	(330)
RI–100	Pt–20L–20D	Dice	7.70	(329)	8.33	(328)
RI–500	P–5L–5D	Dice	8.11	(332)	8.79	(332)
RI–100	P–0L–20D	Dice	8.46	(350)	9.16	(350)

Tablica 8.3: Parovi riječi poredani po procijenjenoj jačini semantičke povezanosti za dva distribucijska semantička modela

Model RI–100, P–0L–5R, Dice (1)		Model RI–500, P–20L–20R, Dice (8)	
Par riječi	Sličnost	Par riječi	Sličnost
unija – djeca	5.000	znati – tvrtka	5.000
kuća – kuna	4.592	europski – hrvatski	2.415
poznat – unija	4.473	hrvatski – francuski	2.297
kuna – euro	4.372	trebati – htjeti	2.254
unija – trenutak	3.635	unija – trenutak	2.205
unija – ukupan	2.638	glavni – grad	2.179
pravi – trenutak	2.590	tržište – reći	2.172
četvrtak – petak	2.331	momčad – tim	2.172
utorak – srijeda	2.331	trenutak – sudac	2.157
ponedjeljak – petak	2.331	važan – posljednji	2.154
utorak – četvrtak	2.331	vojni – snaga	2.151
ponedjeljak – utorak	2.330	reći – izjaviti	2.148
srijeda – petak	2.330	kultura – mjesec	2.147
srijeda – četvrtak	2.329	dan – obitelj	2.146
ponedjeljak – srijeda	2.329	grad – trenutak	2.145
ponedjeljak – četvrtak	2.328	djeca – dionica	2.141
kazati – morati	2.320	mlad – mjesec	2.134
početak – kraj	2.319	kazati – vidjeti	2.128
državni – međunarodni	2.318	znati – uvjet	2.128
važan – posljednji	2.314	vidjeti – dionica	2.126
misliti – vidjeti	2.314	govoriti – ponedjeljak	2.125
njemački – francuski	2.314	znati – zemlja	2.124
smatrati – misliti	2.312	posljednji – suradnja	2.121
razgovor – izjaviti	2.312	ministarstvo – kuća	2.120
svijet – svjetski	2.311	vrijeme – trenutak	2.119
poznat – mali	2.311	rad – posao	2.119
sustav – izjaviti	2.308	reći – kazati	2.118
govoriti – kazati	2.307	američki – hrvatski	2.117
očekivati – željeti	2.306	obitelj – rezultat	2.115
istaknuti – isticati	2.306	željeti – dogoditi	2.114



Slika 8.3: Struktura programskog ostvarenja distribucijskih semantičkih modela

9. Ostvareni modeli distribucijske semantičke složivosti

U radu su ostvarena tri osnovna modela distribucijske semantičke složivosti: aditivni, multiplikativni i konvolucijski model. Modeli složivosti ostvareni su za 500 ciljnih bigrama. U sljedećim poglavljima opisano je ostvarenje modela složivosti i njihovo vrednovanje te su interpretirani rezultati.

9.1. Odabir ciljnih sintagmi i konstrukcija složenih vektora

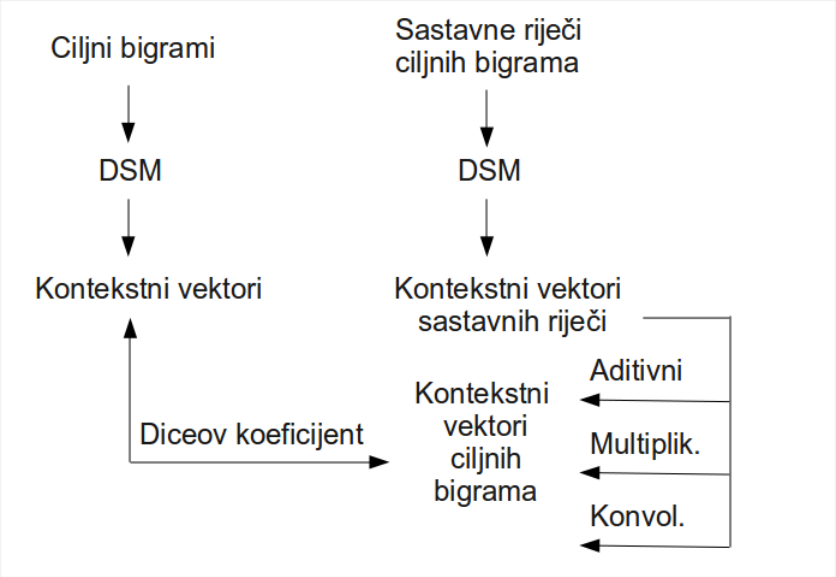
Za izgradnju modela distribucijske semantičke složivosti korišten je isti korpus kao u prvome dijelu projekta. Sintagme čije će se značenje modelirati kompozicijom supojavnih vektora njezinih sastavnica izdvojene su iz korpusa na sljedeći način. U korpusu su pronađeni svi bigrami koji se u njemu pojavljuju, odnosno iz svake su rečenice izdvojeni svi parovi riječi koje se nalaze jedna pored druge. Budući da se sintagma ne može protezati kroz više rečenica, zanemareni su bigrami koji se sastoje od riječi iz različitih (susjednih) rečenica. Na opisani način iz korpusa je izdvojeno 14 141 753 bigrama zajedno s učestalosti pojavljivanja svakoga od njih. U tablici 9.2 prikazano je 20 najčešćih bigrama. Za ciljne sintagme odabrano je 350 najčešćih bigrama, uz izostavljanje bigrama koji ne tvore sintagme i značenjski neprozirnih sintagmi. Uz sintagme atributa i imenice, odnosno apozicije i imenice, među odabranim se ciljnim bigramima pojavljuju i glagolske sintagme predikata i objekta te predikata i glagolske dopune u infinitivu. Dodatno, iz Hrvatskog frazeološkog rječnika (Menac, 2003) izdvojeno je 150 frazemskih bigrama.

Složeni vektori za ciljne bigrame konstruirani su slaganjem vektora riječi sastavnica. Shematski prikaz postupka konstrukcije i vrednovanja složenih vektora prikazan je na slici 9.1.

Tablica 9.1: Bigrami s najviše pojavljivanja u korpusu

Bigram	Učestalost	Bigram	Učestalost
prošao godina	27641	imati pravo	7622
europski unija	22365	sljedeći godina	7396
vanjski posao	13624	bosna hercegovina	7278
konferencija novinar	12466	kraj godina	7252
godina dan	12073	kaznen djel	6977
županijski sud	11815	iznositi kuna	6931
posljednji godina	10731	velik broj	6921
haaški sud	10361	europski komisija	6754
idući godina	9369	mjesec dan	6719
republika hrvatski	9282	velik britanija	6478
ljudski pravi	8628	hrvatski sabor	6409
new york	8603	predsjednik mesić	6389
ministar vanjski	8520	prošao tjedan	6353
međunarodan zajednica	8204	ivica račan	6333
moći reći	8139	stjepan mesić	6311
radni mjesto	7860	predsjednik republika	6210
hrvatski vlada	7772	kaznen prijava	6208
domovinski rata	7724	ustavan sud	6128
državan odvjetništvo	7699	zaštita okolišati	5970
ratni zločin	7698	predsjednik hrvatski	5943

Za riječi sastavnice ciljnih bigrama, ukupno njih 617, izgrađeni su kontekstni vektori distribucijskim semantičkim modelom koji je u prošloj fazi projekta vrednovan kao najbolji, model indeksiranja dimenzije 100 s kontekstom definiranim kao težinski neoznačeni prozor od pet riječi nakon ciljne riječi te Diceovim koeficijentom kao mjerom udaljenosti (RI–100, P–0L–5R, Dice). Kontekstni vektori riječi sastavnica zatim su složeni u vektor koji ostvaruje značenjsku reprezentaciju ciljne sintagme korištenjem formule 7.5 za aditivni model složivosti, formule 7.6 za multiplikativni model i formule 7.8 za konvolucijski model značenjske složivosti.



Slika 9.1: Shematski prikaz postupka konstrukcije i vrednovanja modela distribucijske semantičke kompozicije

9.2. Vrednovanje modela složivosti

Za sve ciljne bigrame izgrađeni su kontekstni vektori korištenjem istoga modela kao i za riječi sastavnice (RI–100, P–0L–5R, Dice), uz bitnu razliku definicije ciljnih elemenata kao ciljnih bigrama. Na ovaj način dobiveni kontekstni vektori ekvivalentni su shvaćanju sintagme, odnosno njezina značenja, kao cjeline određene kontekstom u kojemu se pojavljuje. Budući da su izgrađeni korištenjem vrednovanjem utvrđenog najboljeg distribucijskog semantičkog modela, ovako oblikovano značenje sintagme uzete je kao standardni, odnosno reprezentativni prikaz njezina značenja.

Dva skupa konteksnih vektora za 500 sintagmi zatim su uspoređena. Za svaku je sintagmu izračunat Dicoev koeficijent između kontekstnog vektora dobivenog slaganjem kontekstnih vektora riječi sastavnica te kontekstnog vektora dobivenog izgradnjom distribucijskog semantičkog modela za ciljnu sintagmu u cjelini. Model složivosti utoliko je bolji ukoliko je udaljenost između složenoga vektora izgrađenog modelom složivosti i standardnog vektora za istu sintagmu manja. Diceov je koeficijent korišten u vrednovanju modela složivosti jer je vrednovanjem distribucijskih semantičkih modela utvrđeno da značenjsku povezanost modelira najbliže ljudskom shvaćanju tog koncepta.

9.3. Rezultati

U tablicama 9.2 i 9.3 navedene su Diceove udaljenosti između složenih i standardnih vektora odabranih značenjski prozirnih, odnosno neprozirnih sintagmi. Sintagme su poredane po rastućoj vrijednosti Diceove udaljenosti u multiplikativnome modelu. Razlika između složenih i standardnih vektora značenjski prozirnih sintagmi najmanja je u multiplikativnom modelu, što je konzistentno s rezultatima opisanima u radu (Jeff Mitchell, Mirella Lapata, 2008). Do najvećeg neslaganja između složenih i standardnih vektora dolazi kod konvolucijskog modela. Nije moguće primijetiti konzistentnu razliku između udaljenosti za pojedine vrste sintagmi u smislu strukture, budući da su kao jednako slične vrednovane i imenske sintagme (npr. *grad zagreb*, *mjesec dan*[1], *ratni zločin*) i glagolske sintagme (npr. *moći govoriti*, *moći učiniti*), kao i vlastita imena (npr. *filozofski fakultet*, *los angeles*, *saddam hussein*, *janica kostelić*).

U sedmom je poglavlju u pitanje dovedena izvedba modela složivosti koji uspješno modeliraju značenje semantički prozirnih sintagmi na značenjski neprozirne sintagme, odnosno frazeme. Budući da se značenje frazema ne poklapa sa zbrojem značenja njegovih sastavnih riječi, za očekivati je kako primjena modela kompozicije koji je uspješan u slaganju značenjski prozirnih sintagmi neće biti jednako uspješna kod slaganja frazemskih sintagmi, što je zatim moguće iskoristiti za detekciju frazema. Razlika između složenih i standardnih vektora frazemskih sintagmi također je najmanja u multiplikativnome modelu. Nadalje, uspoređuju li se vrijednosti Diceove udaljenosti multiplikativnoga modela za prvih trideset frazema s onima značenjski prozirnih sintagmi, udaljenosti frazema u prosjeku su za red veličine veće od udaljenosti semantički prozirnih sintagmi, ali ako se usporedba proširi na sve sintagme i frazeme, prosječne su udaljenosti složenih i standardnih vektora za obje skupine bigrama vrlo slične, -1.0859 za značenjski prozirne i -0.9943 za frazemske sintagme. Usporedbom izvedbe aditivnog i konvolucijskog modela za prozirne sintagme u odnosu na frazeme nije moguće doći do konzistentnih razlika, odnosno zaključaka.

Budući da nije moguće uočiti dosljednu razliku u izvedbi modela semantičke složivosti za značenjski prozirne u odnosu na frazemske sintagme, ovako oblikovani modeli složivosti ne mogu se koristiti u detekciji frazema.

[1]Radi se o lematizacijom izmijenjenoj sintagmi *mjesec dana*.

Tablica 9.2: Diceova udaljenost između složenih i stadardnih vektora prozirnih sintagmi

	Model složivosti		
Sintagma	Aditivni	Multiplikativni	Konvolucijski
los angeles	−4.1104	0.0003	−46.6699
saddam hussein	−2.9430	0.0002	−17.5895
bin laden	2.8502	−0.0010	−18.9632
žut karton	6.4496	−0.0033	54.8112
igrač utakmica	5.5002	−0.0060	−3.8943
tjedan dan	−7.7047	−0.0061	−14.5864
liga prvak	5.9690	−0.0069	−14.5376
osmina final	14.6635	−0.0070	627.7151
mjesec dan	−5.9636	−0.0075	−11.0719
izmjena dopuna	5.7724	−0.0083	−16.6191
sunčan hvar	−2.1028	−0.0085	−22.6010
moći govoriti	−35.3106	−0.0086	−334.6210
grad zagreb	36.9652	−0.0094	−320.4286
hrvatski država	−98.4040	−0.0096	129.5874
filozofski fakultet	4.5625	−0.0105	−5.2837
moći učiniti	−40.8686	−0.0106	260.2194
posljednji mjesec	−4.5970	−0.0112	−11.5750
moći raditi	−26.0065	−0.0115	−43.8781
janica kostelić	4.1158	−0.0132	−4.0810
ljudski pravi	15.8977	−0.0140	−102.1838
moći doći	−31.9560	−0.0143	−105.7671
ratni zločin	−8.0714	−0.0146	−59.3974
george bush	11.7575	−0.0155	409.0517
ustavan zakon	21.8221	−0.0165	51.0821
moći reći	−116.1708	−0.0168	156.4518
vladajući koalicija	−11.3392	−0.0174	106.3679
katolički crkva	−20.1412	−0.0176	−158.9865
posljednji dan	−7.1965	−0.0180	−15.5348
zadnji godina	−12.1116	−0.0181	−21.9555
posljednji godina	−10.3976	−0.0191	−27.8961

Tablica 9.3: Diceova udaljenost između složenih i originalnih vektora frazemskih bigrama

	Model složivosti		
Sintagma	Aditivni	Multiplikativni	Konvolucijski
brat brat	6.5018	0.0000	−4.1439
dobar doći	−35.1668	−0.0162	−137.3287
živ primjer	15.4929	−0.0382	−51.2767
posljednji riječ	−9.2678	−0.0508	−42.9451
jasan glasan	−6.2521	−0.0539	−9.0862
meden mjesec	−4.2270	−0.0646	−6.7335
dobar proći	−34.5326	−0.0654	−82.3237
starati priča	18.3322	−0.0659	83.4423
težak riječ	21.3777	−0.0682	36.4300
puk slučaj	−123.9817	−0.0705	18,377.7364
zub vrijeme	21.4437	−0.0770	−22.8761
dan noć	−9.5263	−0.0820	15.4790
imati pravo	−23.4865	−0.0823	15.0589
kupovati vrijeme	20.1912	−0.0853	−27.6410
dobar duša	−62.8131	−0.0857	58.1069
nemati riječ	−28.8631	−0.0864	−147.4234
prst sudbina	64.5245	−0.0876	−1,114.7335
napraviti čovjek	25.4100	−0.0920	31.0877
prati ruka	16.4930	−0.1040	−109.5391
napuniti džep	19.7475	−0.1064	7.0993
soliti pamet	66.4719	−0.1069	343.2183
lud kuća	8.0563	−0.1095	−8.0314
slijep ulica	16.0454	−0.1101	−206.1272
dvosjekli mač	−8.2482	−0.1119	−12.3301
učiti pamet	−19.0264	−0.1127	−195.3803
miran duša	10.5928	−0.1194	−15.1426
težak srce	8.8918	−0.1228	−32.8416
loš strana	7.5648	−0.1256	29.1150
držati korak	19.2582	−0.1452	−178.8276
doći glava	17.8896	−0.1461	17.1735

10. Zaključak

Distribucijski pristup značenju temeljen je na distribucijskoj hipotezi i definira značenje izraza kroz kontekst u kojemu se izraz pojavljuje, izjednačavajući značenjsku sličnost dvaju izraza sa sličnošću konteksta u kojima se upotrebljavaju. Koncept značenja nije potrebno eksplicitno definirati, za izgradnju distribucijskog jezičnog modela dovoljan je dostatno opsežan korpus. Distribucijski semantički modeli značenje riječi prikazuju kontekstnim vektorima u višedimenzijskom vektorskom prostoru. Primjenjuju se u zadacima pretrage dokumenata, ispitivanju sinonimije, razdvajanju značenja višeznačnih riječi, crpljenju informacija iz dvojezičnih resursa, konstrukciji taksonomije, automatiziranom grupiranju riječi i dr. Nadogradnju predstavljaju modeli distribucijske semantičke složivosti, kojima je moguće modelirati semantiku višerječnih izraza.

U ovom su diplomskom radu proučeni i opisani postojeći distribucijski semantički modeli i modeli semantičke složivosti, s naglaskom na modele nasumičnog indeksiranja. Opisani su postupci njihove izgradnje i vrednovanja. Oblikovano je i programski ostvareno 350 distribucijskih semantičkih modela za hrvatski jezik koji su primijenjeni na zadatak određivanja semantičke povezanosti 450 parova riječi. Ostvareni modeli razlikuju se u načinu izgradnje supojavne matrice (frekvencijski modeli i modeli nasumičnog indeksiranja), definiciji konteksta (rečenica, simetrični i asimetrični prozor) i njegovu težinskom označavanju te mjeri vektorske udaljenosti između konstekstnih vektora (udaljenost Manhattan, euklidska udaljenost, kosinusna udaljenost, Jaccardov i Diceov koeficijent). Modeli nausmičnog indeksiranja dodatno se razlikuju u dimenziji indeksnog vektora. Ostvareni modeli vrednovani su usporedbom sa zlatnim standardom dobivenim od ljudskih ocjenjivača. Utvrđen je model s najboljom izvedbom te je proučen utjecaj pojedinih parametara na izvedbu modela. Uočena je izravna povezanost između odabira mjere vektorske udaljenosti i kvalitete izvedbe modela, dok je utjecaj ostalih parametara manje jasan. Očekivano, modeli nasumičnog indeksiranja izvedbom nadmašuju frekvencijske modele.

Oblikovana su, programski ostvarena i vrednovana tri modela distribucijske semantičke složivosti za hrvatski jezik, aditivni, multiplikativni i konvolucijski model

složivosti. Vrednovanjem je utvrđeno kako izvedba multiplikativnoga model nadmašuje aditivni i konvolucijski model neovisno o tome je li oblikovana sintagma značenjski prozirna ili frazemska. Budući da nije bilo moguće uočiti dosljednu razliku u izvedbi modela semantičke složivosti za značenjski prozirne u odnosu na frazemske sintagme, zaključuje se da se ovako oblikovani modeli složivosti ne mogu koristiti u detekciji frazema.

Kao dio budućeg rada predlaže se ostvarivanje distribucijskih semantičkih modela na temelju opsegom i sadržajem drugačijih korpusa uz uključivanje novih načina težinskog označavanja konteksta i novih relacija ciljnih elemenata s kontekstom koje uključuju i informacije o sintaksnim odnosima između riječi. Predlaže se izgradnja modela deriviranih iz frekvencijskog modela korištenjem metoda dimenzijske redukcije i drugih matričnih transformacija te usporedba njihove izvedbe s modelima nasumičnog indeksiranja. Nadalje, detaljnija analiza izvedbe modela u odnosu na specifične vrste semantičkih relacija mogla bi otkriti kako su pojedini distribucijski modeli osjetljiviji na, odnosno bolje prepoznaju određene semantičke relacije. Kao daljni rad na modelima distribucijske semantičke složivosti predlaže se korištenje metoda strojnog učenja i genetskog programiranja kao načina ugađanja parametara modela.

LITERATURA

Latent semantic analysis (lsa) tutorial. http://www.puffinwarellc.com/index.php/news-and-articles/articles/33.html?showall=1, 2010.

Lenci A. Baroni, M. Distributional memory: A general framework for corpus-based semantics. U *Computational Linguistics*, 2010a.

Marco Baroni i Alessandro Lenci. One distributional memory, many semantic spaces. U *GEMS '09: Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, stranice 1–8, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

Zamparelli R. Baroni, M. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. U *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-10)*, 2010b.

Mannila H. Bingham, E. Random projection in dimensionality reduction: applications to image and text data. U *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001. ISBN 1-58113-391-X.

Bartosz Broda i Maciej Piasecki. Supermatrix: a general tool for lexical semantic knowledge acquisition. U *Speech and Language Technology*, svezak 11, stranice 239–254. Polish Phonetics Assocation, 2008. The first version was published in the Proceedings of the International Multiconference on Computer Science and Information Technology — 3rd International Symposium Advances in Artificial Intelligence and Applications (AAIA'08).

Bartosz Broda, Magdalena Derwojedowa, Maciej Piasecki, i Stanisław Szpakowicz. Corpus-based semantic relatedness for the construction of polish wordnet. U *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*,

2008. URL http://www.lrec-conf.org/proceedings/lrec2008/pdf/459_paper.pdf.

Christopher D. Manning, Hinrich Schutze. *Foundations of Statistical Natural Language Processsing*. The MIT Press, 6 izdanju, 2003.

J.R. Curran. *From Distributional to Semantic Similarity*. Doktorska disertacija, University of Edinburgh, 2008.

K. Ferraro D. Widows. Semantic vectors: A scalable open source package and online technology management application. *Sixth International Conference on Language Resources and Evaluation*, 2008.

M. Davies i J. Fleiss. Measuring agreement for multinomial data. *Biometrics*, 38: 1047–1051, 1982.

Ferdinand de Saussure. *Third Course of Lectures in general linguistics (1910-1911)*. Pergamon Press, 1993.

Damian Eads, 2008. URL http://scipy-cluster.googlecode.com/. hcluster: Hierarchical Clustering for SciPy.

Katrin Erk i Sebastian Padó. A structured vector space model for word meaning in context. U *Proceedings of EMNLP*, 2008.

Lenci A. Evert, S. Foundations of distributional semantic models. http://wordspace.collocations.de/lib/exe/fetch.php/course:acl2010:naacl2010_part1.slides.pdf, 2010.

Stefan Evert. Computational Approaches to Collocations - Association measures. http://www.collocations.de/AM/index.html, 2004.

Stefan Evert. *The statistics of word cooccurrences : word pairs and collocations*. Doktorska disertacija, Institute of Computational Linguistics, Faculty of Humanities, Stuttgardt, 2005.

Gottlob Frege. *O smislu i značenju (Osnove aritmetike i drugi spisi)*. Kruzak, 1995.

Eugenie Giesbrecht. In search of semantic compositionality in vector spaces. U *Proceedings of the 17th ICCS*, stranice 173–184. Springer-Verlag, 2009.

G. Golub i C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14:403–420, 1970. ISSN 0029-599X. URL http://dx.doi.org/10.1007/BF02163027. 10.1007/BF02163027.

James Gorman i James R. Curran. Random indexing using statistical weight functions. U *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, stranice 457–464. Association for Computational Linguistics, 2006. ISBN 1-932432-73-6. URL http://portal.acm.org/citation.cfm?id=1610075.1610139.

Emiliano Guevara. A regression model of adjective-noun compositionality in distributional semantics. U *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics*, 2010.

Emiliano Guevara. Computing semantic compositionality in distributional semantics. U *Proceedings of the 9th International Conference on Computational Semantics*, 2011.

H. Rubenstein, J. Goodenough. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

H. Shutze, J. Pedersen. Information retreival based on word senses. *Proceedings of the 4th Annual Symposium od Document Analysis and Information Retreival*, stranice 161–175, 1995.

Jeff Mitchell, Mirella Lapata. Vector-based models of semantic composition. *Proceedings of ACL-08: HLT*, stranice 236–244, 2008.

Jussi Karlgren, Magnus Sahlgren. *Foundations of Real World Intelligence*, poglavlje From Words to Understanding. Stanford: CSLI Publications, 2001.

Boris Kalin. *Povijest filozofije*. Školska knjiga, 26 izdanju, 2003.

P. Kanerva. *Sparse Distributed Memory*. MIT Press, 1988.

Dumais S. Landauer, T. A solution to plato's problem: The latent semantic analysis theory of aquisition, induction and representation of knowledge. *Psychological Review*, 1997.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. A foreword. *Rivista di Linguistica*, 20(1):1–30, 2008.

Burgess C. Atchley R. A. Lund, K. Semantic and associative priming in high-dimensional semantic space. U *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, stranice 660–665, 1995.

Kevin Lund i Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, 28:203–208, 1996. ISSN 1554-351X. URL `http://dx.doi.org/10.3758/BF03204766`.

Fink-Arsovski Željka Venturin Radomir Menac, Antica. *Hrvatski frazeološki rječnik*. Naklada Ljevak, 2003.

Mukhin-A. Panicheva P. Savitsky V. Mitrofanova, O. Automatic word clustering in Russian texts. U *Text, Speech and Dialogue*, stranice 85–91. Springer, 2007.

P. Nakov. Latent semantic analysis for bulgarian literature. U *Proceedings of Spring Conference of Bulgarian Mathematicians Union. Borovetz*, 2001a.

P. I. Nakov. Latent semantic analysis for russian literature investigation. U *In Proceedings of the 120 years Bulgarian Naval Academy Conference*. Citeseer, 2001b.

Nikola Ljubešić, Damir Boras, Nikola Bakarić, Jasmina Njavro. Comparing measures of semantic similarity. *Proceedings of the ITI 2008* 30^{th} *International Conference of Information Technology Interfaces*, June 2008.

Sebastian Pado i Mirella Lapata. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199, 2007.

P. Pantel. Inducing ontological co-occurrence vectors. *Proceedings of the 43rd Conference of the Association for Computational Linguistics, ACL'05*, 2005.

Paul Goldsmith-Pinkham Phil Katz. Word sense disambiguation using latent semantic analysis. `http://www.sccs.swarthmore.edu/users/07/pkatz1/cs65f06-final.pdf`, 2008.

Maciej Piasecki. Automated extraction of lexical meanings from corpus: A case study of potentialities and limitations. U *Representing Semantics in Digital Lexicography. Innovative Solutions for Lexical Entry Content in Slavic Lexicography. MONDILEX Fourth Open Workshop. Warszawa, Poland, 29 June – 1 July, 2009. Proceedings*, stranice 32–43. Institute of Slavic Studies, Polish Academy of Sciences, 2009. URL `http://www.ii.pwr.wroc.pl/~piasecki/publications/Mondilex09-piasecki.pdf`.

G.W. Furnas T.K. Landauer R. Harshman S. Deerwester, S.T. Dumais. Indexing by latent semantic analysis. *J Amer Soc Inf Sci*, 1990.

Magnus Sahlgren. An introduction to random indexing. *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering*, 2005.

Magnus Sahlgren. The distributional hypothesis. *Rivista di Linguistica*, 20, 2008.

H. Schutze. Automatic word sense discrimination. *Comput Linguist*, stranice 97–123, 1998.

P. Smrž i P. Rychlỳ. Finding semantically related words in large corpora. U *Text, Speech and Dialogue*, stranice 108–115. Springer, 2001.

Stefan Evert, Alessandro Lenci. Foundations of Distributional Semantic Models. http://wordspace.collocations.de/lib/exe/fetch.php/course:esslli2009:esslli.dsm.1.pdf, July 2009.

Dalbelo Bašić B. Tadić Šnajder, J. Automatic acquisition of inflectional lexica for morphological normalisation.

Dominic Widdows. Semantic vector products: Some initial investigations. U *Proceedings of the Second Symposium on Quantum Interaction, Oxford, UK*, 2008.

Michael P. Wolf. Philosophy of language. http://www.iep.utm.edu/lang-phi/, svibanj 2009.

D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. U *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, stranice 189–196. ACL, 1995.

Ante Šoljić. Skripta iz kolegija Filozofija jezika. Filozofski fakultet u Mostaru, siječanj 2010.

Ž. Agić, M. Tadić. Evaluating morphosyntactic tagging of croatian texts. *LREC2006 Proceedings*, 2006.

Računalni modeli distribucijske leksičke semantike hrvatskoga jezika

Sažetak

Računalna semantika važna je u sustavima za obradu i razumijevanje prirodnog jezika. Distribucijski semantički modeli značenje riječi prikazuju kontekstnim vektorima u višedimenzijskom vektorskom prostoru. Nadogradnju predstavljaju modeli distribucijske semantičke složivosti, kojima je moguće modelirati semantiku višerječnih izraza.

U radu su proučeni i opisani postojeći distribucijski semantički modeli i modeli semantičke složivosti te postupci njihove izgradnje i vrednovanja, s naglaskom na model nasumičnog indeksiranja. Oblikovani su, programski ostvareni i vrednovani distribucijski semantički modeli za hrvatski jezik primijenjeni na zadatak određivanja semantičke sličnosti riječi. Dodatno, oblikovani su, programski ostvareni i vrednovani distribucijski modeli semantičke složivosti za hrvatski jezik te je razmotrena njihova primjena u detekciji idioma.

Ključne riječi: distribucijski semantički model, distribucijski model semantičke složivosti, računalna semantika, nasumično indeksiranje, hrvatski jezik

Computational Models of Distributional Lexical Semantics in Croatian Language

Abstract

Computational semantics is of high importance in systems for processing and understanding natural language. Distributional semantic models represent meanings of lexical expressions as multi-dimensional context vectors, and are upgraded to models of distributional semantic composition to model meaning of multiword expressions.

In this thesis existing distributional semantic models, their construction and evaluation, with an emphasis on random indexing models, have been studied and described. Distributional semantic models for Croatian language applied to the task of semantic similarity assessment have been modeled, implemented and evaluated. Furthermore, models of distributional semantic composition have been modeled, implemented and evaluated and their application in idiom detection has been considered.

Keywords: distributional semantic model, distributional semantic compositionality, computational semantics, random indexing, Croatian language