TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 771

Model za određivanje semantičke kompozicionalnosti višerječnih izraza hrvatskoga jezika

Petra Almić

Zagreb, lipanj 2014.

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

Zagreb, 10. ožujka 2014.

DIPLOMSKI ZADATAK br. 771

Pristupnik: **Petra Almić**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Model za određivanje semantičke kompozicionalnosti višerječnih izraza hrvatskoga jezika**

Opis zadatka:

Višerječni izrazi, poput frazema, strukovnog nazivlja i leksičkih kolokacija, iziskuju posebnu pažnju u obradi prirodnog jezika zbog njihovih sintaktičkih i semantičkih osobitosti. Posebno su zanimljive višerječni izrazi koji su semantički neprozirni odnosno nekompozicionalni i koje zbog toga nije moguće modelirati raščlambom na sastavne riječi, poput izraza "morski pas", "žuta minuta" ili "ležeći policajac". Automatsko određivanje semantičke kompozicionalnosti višerječnih izraza važno je za mnoge primjene obrade prirodnog jezika, poput strojnog prevođenja i pretraživanja informacija.
U okviru diplomskoga rada potrebno je proučiti postupke za ekstrakciju višerječnih izraza iz korpusa i određivanje njihove semantičke kompozicionalnosti, s naglaskom na postupke temeljene na modelima distribucijske semantičke kompozicije. Razraditi model za određivanje semantičke kompozicionalnosti višerječnih izraza hrvatskoga jezika odabrane sintaktičke strukture, po uzoru na radove (Katz i Giesbrecht, 2006) i (Biemann i Giesbrecht, 2011). Izgraditi reprezentativnu ispitnu zbirku višerječnih izraza hrvatskoga jezika ručno označenu ocjenama semantičke kompozicionalnosti. Razviti programsku implementaciju postupka određivanja semantičke kompozicionalnosti višerječnih izraza te provesti detaljno eksperimentalno vrednovanje i analizu pogrešaka na ispitnome uzorku. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 14. ožujka 2014.
Rok za predaju rada: 30. lipnja 2014.

Mentor:

Doc. dr.sc. Jan Šnajder

Djelovođa:

Doc. dr.sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr.sc. Siniša Srbljić

SADRŽAJ

1. Uvod

Semantika je grana jezikoslovlja koja se bavi proučavanjem značenja. No jezikoslovci nisu jedini istraživači koji su zainteresirani za semantiku, to područje zanima i filozofe, psihologe i računarce. Kako ljudi uče jezik, kako shvaćaju i povezuju značenja riječi i rečenica? Mala djeca sposobna su brzo svladati bilo koji jezik, gramatiku i njegova produkcijska pravila samo kroz izloženost njegovoj upotrebi, i to prilično ograničenu. Mogu li računala nekako imitirati taj postupak?

Distribucijska semantika (engl. *distributional semantics*) nagovještava da mogu. Distribucijski ili statistički pristup semantici omogućava računalima da povezuju značenja iza riječi kroz njihov odnos u velikim zbirkama tekstova (korpusu). Hipoteza jest da će se riječi koje imaju slično značenje pojavljivati u sličnim kontekstima. Matematički, riječi su predstavljene vektorima u visoko dimenzionalnom prostoru. Vektori modeliraju značenje riječi na temelju frekvencija (su)pojavljivanja riječi u korpusu i mogu se međusobno uspoređivati: što su bliži, semantički su sličniji. Distribucijski semantički modeli pokazali su se prilično učinkovitima u primjeni na različitim zadacima: pretraživanju i crpljenju informacija, klasifikaciji i grupiranju dokumenata ili riječi, mjerenju sličnosti riječi i razrješavanju višeznačnosti riječi (Turney i Pantel, 2010).

Višerječni je izraz (engl. *multiword expression*, *MWE*) sveza dvije ili više riječi koja ima neka zanimljiva semantička, pragmatička, sintaktička ili statistička obilježja te se zbog toga treba tretirati kao cjelina, poput jedne riječi. Riječi u izrazu ne moraju biti slijedne i nisu ograničene na određene morfosintaktičke obrasce (Baldwin, 2006; Šnajder, 2010). U glavnom fokusu ovog rada su višerječni izrazi koji su semantički nekompozicionalni odnosno neprozirni, te ih zbog toga nije moguće modelirati raščlambom na sastavnice; značenje neprozirnog izraza ne odgovora zbroju (ili općenito kompoziciji) značenja njegovih sastavnica. Primjeri takvih izraza su: *ležeći policajac*, *žuta minuta*, *morski pas*. Automatsko određivanje kompozicionalnosti važno je u mnogim područjima obrade prirodnog jezika, primjerice u strojnom prevođenju i pretraživanju informacija.

Ideja ovog rada jest istražiti kako iskoristiti svojstvo nekomopozicionalnosti u distribucijskim semantičkim modelima za određivanje semantičke kompozicionalnosti višerječnih izraza hrvatskog jezika. Primjerice, izraz *prodavati maglu* semantički je neproziran jer se njegovo značenje ne može zaključiti iz poznavanja značenja riječi *prodavati* i *magla*. Zbog toga, za očekivati je da će u distribucijskom semantičkom modelu vektor koji predstavlja izraz $\overrightarrow{\boldsymbol{prodavati\ maglu}}$ biti značajno udaljen od kompozicije vektora $\overrightarrow{\boldsymbol{prodavati}}$ i $\overrightarrow{\boldsymbol{magla}}$. Cilj ovog rada jest provjeriti jesu li ta očekivanja uistinu točna i opravdana, što su na kraju rezultati i potvrdili.

Postoje neki višerječni izrazi koji se mogu koristiti u prozirnom i neprozirnom značenju poput izraza *desna ruka* koji osim svog doslovnog značenja može nositi značenje pouzdanog oslonca ili osobe od glavne pomoći. Potrebno je napomenuti da ovaj rad ni na koji način ne razmatra koliki je potencijal izraza da bude proziran (ili neproziran), već se bavi stupnjem neprozirnosti.

Sadržaj rada organiziran je po poglavljima ukratko opisanim u nastavku. U drugom poglavlju opisana je teorijska podloga potrebna za razumijevanje narednih poglavlja tj. opisani su distribucijski semantički modeli i modeli distribucijske semantičke kompozicije. Zatim je u trećem poglavlju opisan postupak izgradnje reprezentativne zbirke višerječnih izraza hrvatskog jezika. U četvrtom poglavlju opisan je postupak izgradnje modela i prezentirani su rezultati evaluacije modela.

2. Distribucijska semantika

Distribucijska semantika (engl. *distributional semantics*), poznata još kao i statistička semantika (engl. *statistical semantics*) ili semantika vektora (engl. *vector semantics*), svoje začetke nalazi u radovima (Harris, 1954; Firth, 1957) koji postavljaju teoriju o distribucijskoj hipotezi (engl. *distributional hypothesis*): *riječi koje se pojavljuju u sličnim kontekstima imaju slično značenje* (Harris, 1954) ili prema Firthu: *znat ćeš riječ po društvu u kojem se nalazi* (Firth, 1957).

U distribucijskoj je semantici riječ predstavljena distribucijom svih tekstovnih cjelina (konteksta) u kojim se pojavljuje. Za određenu riječ distribucija konteksta može se dobiti iz frekvencija supojavljivanja konteksta i te riječi. Prema tome, dvije riječi smatraju se sličnima ako imaju sličnu distribuciju konteksta. Drugi način na koji se može razmišljati o reprezentaciji riječi u distribucijskim semantičkim modelima jest kao prosjek značenja svih konteksta u kojima se pojavljuje (Landauer i Dumais, 1997). Kontekst se može definirati kao cijeli dokument, paragraf, rečenica, simetrični (ili asimetrični) prozor riječi fiksne veličine, određeni sintaktički uzorak ili neka kombinacija navedenog.

Na primjer ako se riječi *kuća* i *stan* često pojavljuju uz riječi *iznajmljivanje*, *prodaja*, *soba*, *zgrada*, *kat* itd., distribucijski modeli mogu zaključiti da su riječi *kuća* i *stan* slične. Slično kao i za riječ, višerječni izraz može se predstaviti distribucijom konteksta u kojim se pojavljuje. S druge strane, modeli distribucijske semantičke kompozicije pokušavaju predviditi distribuciju konteksta cijelog izraza na temelju distribucija pojedinih sastavnica izraza. Razlika između predviđene i stvarne distribucije konteksta može upućivati na neka zanimljiva svojstva izraza, poput nekompozicionalnosti i to je upravo ono što ovaj rad želi istražiti. Glavna je prednost distribucijskih modela mogućnost kvantifikacije značenja, a pri tome im je potreban samo jedan jezični reusurs: korpus. U nastavku slijedi pregled distribucijskih semantičkih modela.

2.1. Distribucijski semantički modeli

Distribucijski semantički modeli (engl. *distributional semantic models*, *DSM*) predstavljaju konkretnu realizaciju distribucijske hipoteze. U DSM-u svaka riječ predstavljena je kao matematički vektor u visoko dimenzionalnom prostoru. Cijeli model predstavljen je matricom čiji retci odgovaraju ciljnim riječima, a stupci kontekstu u kojem se ciljne riječi pojavljuju. Osnovno svojstvo DSM-a je da su vrijednosti njegovih elemenata izvedene iz frekvencija pojavljivanja, npr. broj puta koliko se jedna riječ pojavljuje pored druge (Turney i Pantel, 2010).

Formalno se DSM može predstaviti kao sedmorka (Lenci, 2008):

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

gdje je T skup ciljnih riječi (engl. *target elements*), C označava kontekst (dokumenti iz kolekcije, paragrafi, riječi, etc.), R je relacija između ciljnih riječi i konteksta, W je mjera dodjeljivanja težina elementima relacije, M je matrica frekvencija reda $|T| \times |C|$, d je funkcija koja radi smanjivanje dimenzionalnosti $d : M \rightarrow M'$, a S je mjera udaljenosti između vektora u matrici M'.

Kontekst, mjera dodjeljivanja težina i funkcija smanjivanja dimenzionalnosti poznati su još kao i parametri modela. Obično se pojedini tip DSM-a povezuje s određenim skupom parametara, no o tome nešto detaljnije u pregledu tipičnih distribucijskih semantičkih modela u poglavlju 2.1.3.

Postupak izgradnje DSM-a može se podijeliti u dvije glavne faze navedene u nastavku.

1. Lingvistička obrada:
   - tokeniziranje (opojavničenje)
   - normalizacija
   - označavanje.
2. Matematička obrada:
   - izgradnja frekvencijske matrice M
   - dodjeljivanje težina
   - smanjivanje dimenzionalnosti
   - uspoređivanje vektora.

Svaka od ovih faza uključuje po nekoliko koraka koji su detaljnije objašnjeni u potpoglavljima 2.1.1 i 2.1.2.

2.1.1. Lingvistička obrada

Prije izgradnje same matrice, korisno je prvo napraviti lingvističku obradu korpusa. Tokenziranje ili opojavničenje je postupak dovođenja korpusa u stanje u kojem su sve riječi (pojavnice) identificirane i označene (Bekavac, 2002). Naizgled jednostavan zadatak, budući da je riječi lako identificirati jer su odvojene bjelinama, no u složenijim slučajevima tokenizacija može biti mnogo zahtjevnija jer se pojavnicama mogu smatrati i višerječni izrazi, što znači da bi tokenizacija između ostalog uključivala i prepoznavanje imenovanih entiteta.

Drugi je korak normalizacija. Motivacija iza normalizacije jest pojava postojanja sličnih, ali različitih nizova znakova koji nose isto značenje. Najučestaliji oblici normalizacije su lematizacija i pretvaranje velikih slova u mala. Lematizacija je svođenje različitih pojavnica (članova iste paradigme) na zajedničku lemu te je izuzetno važna kod jezika koji imaju bogatu morfologiju, poput hrvatskog (Bekavac, 2002).

Označavanje je suprotno od normalizacije, jer kao što različiti nizovi znakova mogu nositi isto značenje, tako i jednaki nizovi znakova mogu nositi različito značenje, ovisno o kontekstu (istopisnice). Označavanje vrsta riječi[1] (engl. *part-of-speech tagging, POS tagging*) jest pridruživanje gramatičke kategorije svakoj pojavnici u tekstu (Bekavac, 2002). Druga je vrsta označavanja sintaktičko označavanje koje se odnosi na sintakičke relacije u rečenici.

2.1.2. Matematička obrada

Nakon što je tekst tokeniziran, normaliziran i označen, slijedi matematička obrada. Matematički koraci u izgradnji DSM-a opisani su u nastavku.

Izgradnja frekvencijske matrice

Prvi je korak generirati matricu frekvencija. Element u matrici frekvencija odgovara sljedećem događaju: određena stavka (pojam, riječ, par riječi) pojavila se u određenoj situaciji (dokument, kontekst, uzorak, prozor) određeni broj puta (frekvencija). Dakle, izgradnja frekvencijske matrice svodi se na brojanje događaja (Turney i Pantel, 2010). Ilustrativni primjer frekvencijske matrice može se vidjeti u tablici 2.1. Retci predstavljaju ciljne riječi i u tom primjeru to su: *pas, mačka, glazba, umjetnost*. Stupci predstavljalju kontekst, u primjeru to su riječi: *hraniti, popularan, opasan, kultura*. Element u matrici označava da se odgovarajuća ciljna riječ pojavila u blizini (prozor)

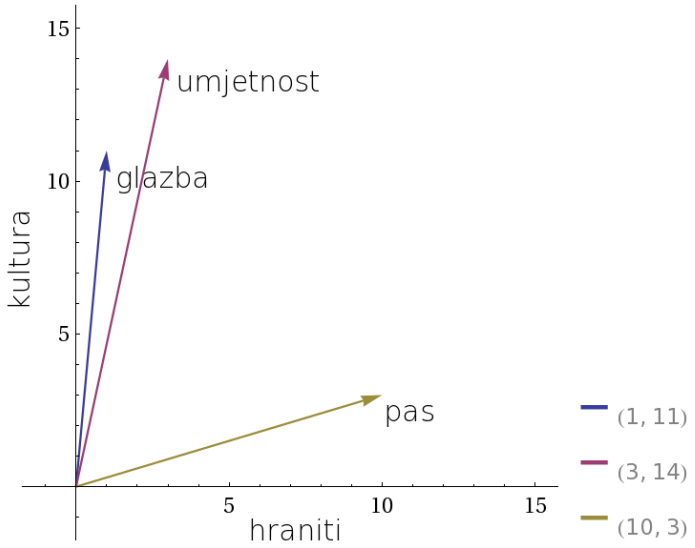
[1] Ponekad se naziva gramatičko označavanje ili morfosintaktičko obilježavanje (Bekavac, 2002).

odgovarajuće kontekst riječi određeni broj puta. Npr. riječ *pas* pojavila se u blizini riječi *hraniti* 10 puta.

Tablica 2.1: Frekvencijska matrica

	hraniti	popularan	opasan	kultura
pas	10	1	8	3
glazba	1	12	2	11
umjetnost	3	10	1	14
mačka	13	2	1	0

Frekvencijska je matrica sama srž distribucijskog semantičkog modela. Retci su matrice vektori koji nose značenje riječi. Na slici 2.1 prikazana je geometrija značenja. Za primjer su uzeti vektori iz tablice 2.1 koji su dodatno pojednostavljeni da bi se mogli prikazati u dvije dimenzije. Prvu dimenziju čini kontekstna riječ *hraniti*, a drugu dimenziju čini kontekstna riječ *kultura*. Dakle, iz tablice 2.1 u ovom primjeru razmatraju se samo prvi i zadnji stupac. Slika prikazuje kako su u tom malenom dvodimenzijskom svijetu riječi *glazba* i *umjetnost* prilično bliske, dakle međusobno slične, a opet znatno udaljene od riječi *pas* koja nije povezana s njima. Primjer je pretjerano pojednostavljen u svrhu ilustracije, ali isti koncept vrijedi i u N dimenzija, s puno više riječi i s većim vektorima.



Slika 2.1: Geometrija značenja u DSM-u

Izgradnja frekvencijske matrice osnovni je korak, ostali su koraci kod matematičke

obrade (dodjeljivanje težina frekvencijama, smanjivanje dimenzionalnosti) opcionalni, preporuča ih se napraviti zbog poboljšanja učinkovitosti, ali nisu nužni.

Težinske mjere

Sljedeći je korak izmijeniti elemente matrice tako da se rječim događajima dodijeli veće značenje. Npr. ako se uz riječ *kuća* pojave riječi *trijem* i *htjeti*, ideja je dati veće značenje pojavi riječ *trijem* jer riječ *htjeti* je prilično učestala i zbog toga ne nosi puno korisne informacije. U pretraživanju informacija (engl. *information retrieval*) neke od uobičajenih težinskih mjera koje se koriste su: mjera TF-IDF (engl. *term frequency-inverse document frequency*), mjere uzajamne informacije PMI (engl. *pointwise mutual information*) i LMI (engl. *local mutual information*) te logaritam entropije (engl. *log entropy*). Njihove formule dane su u nastavku.

Neka je f_{ij} frekvencija pojavljivanja ciljne riječi i u kontekstu j, a f_i broj pojavljivanja riječi i po svim kontekstima (zbroj svih stupaca u retku i) i f_j broj svih riječi u kontekstu j (zbroj svih redaka u stupcu j), n_c je broj stupaca, n_r je broj redaka, a n_{ci} je broj stupaca u retku i različitih od 0. Mjera tf-idf definirana je onda kao:

$$tfidf(i,j) = \frac{f_{ij}}{f_j} \times \log\left(\frac{n_c}{n_{ci}}\right) \tag{2.1}$$

Dodatno, neka su:

$$p_{i*} = \frac{\sum_{j=1}^{n_c} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \tag{2.2}$$

$$p_{*j} = \frac{\sum_{i=1}^{n_r} f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \tag{2.3}$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n_r}\sum_{j=1}^{n_c} f_{ij}} \tag{2.4}$$

Mjera PMI definirana je onda kao:

$$pmi(i,j) = \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right) \tag{2.5}$$

Mjera LMI definirana je kao:

$$lmi(i,j) = f_{ij} \times \log\left(\frac{p_{ij}}{p_{i*}p_{*j}}\right) \tag{2.6}$$

Mjera logaritam entropije definirana je kao:

$$le(i,j) = g_i \times \log\left(f_{ij} + 1\right) \tag{2.7}$$

gdje je

$$g_i = 1 - \sum_{j=1}^{n_c} \frac{p_{ij} \log p_{ij}}{\log n_c} \tag{2.8}$$

Za dodatne informacije o težinskim mjerama pogledati (Landauer, 2007; Lenci, 2008; Turney i Pantel, 2010; Evert, 2008; Nakov et al., 2001; Jones, 1972).

Smanjivanje dimenzionalnosti

Matrica koja predstavlja distribucijski semantički model obično je velika, ali rijetko popunjena, odnosno većina elemenata jest jednaka nuli. Ideja je smanjivanja dimenzionalnosti dobiti manju, ali gusto ispunjenu matricu. Na ovaj način olakšava se uspoređivanje vektora, uklanja se šum i zadržava se samo visoka razina informacije. Neke od matematičkih metoda kojima se postiže smanjivanje dimenzionalnosti su: dekompozicija singularnih vrijednosti (engl. *singular value decomposition*, *SVD*) (Deerwester et al., 1990), nasumično indeksiranje (engl. *random indexing*, *RI*) (Sahlgren, 2005), analiza svojstvenih komponenti (engl. *principal component analysis*, *PCA*) (Smith, 2002) ili analize nezavisnih komponenti (engl. *independent component analysis*, *ICA*) (Hyvärinen et al., 2004).

Uspoređivanje vektora

Na kraju, za uspoređivanje sličnosti vektora odnosno značenja riječi može se koristiti bilo koja matematička mjera udaljenosti ili sličnosti, no obično se najčešće upotrebljava kosinus kuta. Kosinus kuta između dva vektora $\vec{x}$ i $\vec{y}$ koji imaju N elemenata računa se kao:

$$\cos\left(\vec{x}, \vec{y}\right) = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{N} {x_i}^2 \cdot \sum_{i=1}^{N} {y_i}^2}} \tag{2.9}$$

Kosinus može biti između -1 i 1, kad su vektori okomiti (kut je 90 stupnjeva), onda je 0, a ako su paralelni, onda je -1 ili 1, ovisno o tome jesu li istog smjera. Neke druge mjere koje se koriste su Diceov indeks (Dice, 1945):

$$dice\left(\vec{x}, \vec{y}\right) = \frac{\sum_{i=1}^{N} \min\left(x_i, y_i\right)}{\sum_{i=1}^{N} \max\left(x_i, y_i\right)} \tag{2.10}$$

i Jaccardov indeks (Jaccard, 1901):

$$jaccard\left(\vec{x}, \vec{y}\right) = \frac{2 \cdot \sum_{i=1}^{N} \min\left(x_i, y_i\right)}{\sum_{i=1}^{N} \left(x_i + y_i\right)} \tag{2.11}$$

Za više informacija o mjerama sličnosti pogledati (Bullinaria i Levy, 2007; Manning et al., 2008; Curran, 2004).

2.1.3. Vrste distribucijskih semantičkih modela

Ranije je već spomenuto da su različite vrste distribucijskih semantičkih modela određene različitim parametrima (kontekst, težinska mjera, funkcija smanjivanja dimenzionalnosti). Pregled tipičnih distribucijskih semantičkih modela slijedi u nastavku.

Vektorski prostor

Vektorski prostor (engl. *vector space model*, *VSM*) najjednostavnija je verzija distribucijskog semantičkog modela. Originalno je razvijen za sustav pretraživanja informacija *SMART* (Salton, 1971). Ideja je bila da svaki dokument iz kolekcije predstavlja točku u prostoru (vektor), a korisnički upit čini *pseudo-dokument* koji se onda uspoređuje s vektorima ostalih dokumenata. Kontekst čine cijeli dokumenti, tj. u matrici frekvencija jedan stupac predstavlja jedan dokument. Frekvencije se ne modificiraju niti se provodi smanjivanje dimenzionalnosti. U širem smislu, pojam vektorski prostor može se odnositi na bilo koji distribucijski semantički model s proizvoljno odabranim parametrima (Turney i Pantel, 2010).

Hiperprostorna analogija jeziku

Hiperprostorna analogija jeziku (engl. *hyperspace analogue to language*, *HAL*) verzija je distribucijskog semantičkog modela opisanog u radu (Lund i Burgess, 1996). Kontekst čini klizni prozor od n riječi. Matrica je frekvencija dimenzija $|V|\times|V|$, gdje V predstavlja vokabular. Izvorni autori ne spominju reduciranje dimenzija niti dodjeljivanje težina. Kao parametre modela navode samo veličinu prozora i mjeru sličnosti vektora.

Latentna semantička analiza

Latentna semantička analiza (engl. *latent semantic analysis*, *LSA*) poznata i kao latentno semantičko indeksiranje u pretraživanju informacija (engl. *latent semantic indexing*, *LSI*), predstavlja nadogradnju vektorskog prostora s moćnim matematičkim alatom za smanjivanje dimenzionalnosti – dekompozicija singularnih vrijednosti (engl. *singular value decomposition*, *SVD*). Model je opisan u (Landauer i Dumais, 1997). Kontekst je proizvoljan, kao i odabir mjere za dodjeljivanje težine, no autori predlažu transformiranje frekvencija pojavljivanja funkcijom logaritma entropije (formula 2.7). Ključan korak latente semantičke analize jest singularna dekompozicija matrice

(Deerwester et al., 1990). Frekvencijska matrica A dekompozicijom singularnih vrijednosti rastavlja se na produkt tri nove matrice:

$$A = UDV^T \tag{2.12}$$

Konceptualno, matrica U predstavlja značenje riječi kao presjek konteksta u kojima se pojavljuje, a matrica V^T predstavlja značenje konteksta kao presjek riječi koji se pojavljuju u njemu. Matrica D jest dijagonalna matrica koja sadrži singularne vrijednosti; one predstavljaju skalirajuće faktore takve da kad se tri matrice pomnože, rekonstruira se originalna matrica A.

Nasumično indeksiranje

Model nasumičnog indeksiranja (engl. *random indexing*, *RI*) sličan je latentnoj semantičkoj analizi, opisan u (Sahlgren, 2005). Budući da je dekompozicija singularnih vrijednosti računalno zahtjevan postupak koji se mora svaki put prilikom promjene frekvencijske matrice ponovno napraviti, Sahlgren predlaže drugačiji pristup izgradnji konteksnih vektora. Model prvo za svaki kontekst (npr. riječ ili dokument) generira jedinstven nasumični vektor dimenzije d (indeksni vektor), uglavnom popunjen nulama. Zatim prolazi kroz korpus i svaki put kad se neka riječ pojavi u kontekstu, distribucijskom vektoru te riječi pribroji se indeksni vektor tog konteksta. Obično je d puno manji od broja konteksta, pa se zbog toga događa smanjenje dimenzionalnosti u postupku.

2.2. Distribucijski modeli semantičke kompozicije

Glavna ideja distribucijskih modela semantičke kompozicije temelji se na prinicipu semantičke kompozicionalnosti (engl. *principle of semantic compositionality*, *PSC*), koji kaže da je značenje kompleksnog izraza određeno značenjima njegovih dijelova i pravilima po kojima se ta značenja kombiniraju (Pelletier, 1994; Partee, 1995). Taj prinicip poznat je još i kao Fregeovo načelo jer se Gottlob Frege, njemački matematičar i filozof, smatra zaslužnim za njegovu prvu modernu formulaciju.

U okviru distribucijskih semantičkih modela princip kompozicionalnosti može se formalno definirati na sljedeći način (Mitchell i Lapata, 2008, 2010). Neka je $\vec{p}$ vektor koji se dobije kompozicijom vektora sastavnica $\vec{u}$ i $\vec{v}$. Distibucijski model semantičke kompozicije onda je predstavljen izrazom:

$$\vec{p} = f(\vec{u}, \vec{v}) \tag{2.13}$$

Jedna od najjednostavnijih strategija za kompoziciju vektora je njihovo zbrajanje (Landauer i Dumais, 1997), međutim, nedostatak ovog pristupa je neosjetljivost na poredak riječi. U slučaju zbrajanja isti kompozicijski vektor bi imali izrazi *Iva voli Antu* i *Ante voli Ivu* iako imaju različito značenje. Kako bi razriješio problem osjetljivosti na poredak riječi, Kintsch (2001) predlaže varijaciju aditivnog modela u kojem zbroju vektora sastavnica pridodaje zbroj vektora najbližih susjeda glavne sastavnice (predikat u njegovim primjerima).

Mitchell i Lapata (2008, 2010) na zadatku mjerenja semantičke sličnosti fraza (npr. *napraviti efekt* vs. *postići rezultat*) isprobali su niz različitih modela kompozicije vektora:

1. jednostavni aditivni
$$\vec{p} = \vec{u} + \vec{v} \tag{2.14}$$
2. multiplikativni
$$\vec{p} = \vec{u} \odot \vec{v} \tag{2.15}$$
3. težinski aditivni
$$\vec{p} = \alpha\vec{u} + \beta\vec{v} \tag{2.16}$$
4. tenzorski produkt
$$P = \vec{u} \otimes \vec{v} \tag{2.17}$$
5. dilatacija
$$\vec{p} = (1 - \lambda)(\vec{u} \cdot \vec{v})\vec{u} + (\vec{u} \cdot \vec{u})\vec{v} \tag{2.18}$$

U tablici 2.2 na ilustrativnom primjeru iz poglavlja 2.1 prikazan je dodatno multiplikativni i aditivni vektor, prema formulama 2.14 i 2.15. Kod aditivnog se modela elementi vektora samo zbroje, dok se kod multiplitkativnog modela pomnože. U te-

Tablica 2.2: Aditivni i multiplikativni model

	hraniti	popularan	opasan	kultura
pas	10	1	8	3
mačka	13	2	1	0
pas + mačka	23	3	9	3
pas $\odot$ mačka	130	2	8	0

žinskom aditivnom modelu pojedina sastavnica pridonosi ukupnom zbroju u određenoj mjeri i taj model nije osjetljiv na poredak riječi u izrazu. Na taj način veći naglasak

može se dati glavnoj sastavnici (npr. imenicama u imeničkim sintagmama). Skalare α i β odredili su eksperimentalno i ovise o vrsti izraza. Težinski model po rezultatima je superiorniji u odnosu na jednostavni aditivni. Općenito, najboljim se pokazao multiplikativni model iako i on pati od neosjetljivosti na poredak riječi.

Widdows i Ferraro (2008) problem poretka riječi rješavaju tenzorskim produktom, no on otvara drugi problem – neskalabilnost, rezultantni "vektor" nalazi se u drugom prostoru.

Baroni i Zamparelli (2010) i Guevara (2010) predlažu korištenje linearnog modela (regresija) za predviđanje vektora izraza koji se sastoji od imenice i pridjeva koji ju modificira. Osnovna je ideja osim vektora sastavnica ($\vec{u}, \vec{v}$), izgraditi i vektor cijelog izraza ($\vec{p}$) te na temelju tih podataka linearnom regresijom naučiti nepoznatu kompozicijsku funkciju f iz izraza 2.13. Postupak se pokazao učinkovitim, ali ostaje nejasno kako ga generalizirati na duže ili drugačije izraze (Turney, 2013). Nešto slično pokušali su Socher et al. (2012), samo što su oni nepoznatu funkciju f pokušali naučiti neuronskom mrežom.

2.3. Određivanje kompozicionalnosti izraza

Ovaj je rad motiviran idejom da značenje neprozirnog višerječnog izraza ne odgovora zbroju značenja njegovih sastavnica i da upravo po tome onda možemo razlikovati neprozirne višerječne izraze od prozirnih. Prepoznavanje neprozirnih izraza važno je u mnogim zadacima obrade prirodnog jezika poput pretraživanja infromacija (Acosta et al., 2011), strojnog prevođenja (Carpuat i Diab, 2010), razrješavanja višeznačnosti (Finlayson i Kulkarni, 2011) itd., i u zadnjim godinama sve se više pozornosti pridaje tom problemu.

Lin (1999) tvrdi da uzajamna informacija (engl. *mutual information*) može diskriminirati kompozicionalne izraze od nekompozicionalnih. Metoda koju Lin predlaže jest usporediti uzajamnu informaciju sastavnica neprozirnog višerječnog izraza s uzajmnom infromacijom sličnog izraza koji se dobije tako da se jedna od sastavnica promatranog izraza zamijeni sa sličnom riječi iz rječnika. Na primjer, engleski idiom za birokraciju *red tape* (*crvena vrpca* u doslovnom prijevodu na hrvatski) usporedio je s izrazima *žuta vrpca* i *narančasta vrpca*. U ovom slučaju, uzajamna informacija *crvene vrpce* je 5,87, dok za *žutu vrpcu* i *narančastu vrpcu* iznosi 3,75 odnosno i 2,64. Linov algoritam postiže preciznost od 13,7 % i odziv od 15,7 % u usporedbi sa zlatnim standardom generiranim na temelju rječnika.

Baldwin et al. (2003) uspoređuje distribuciju glavne sastavnice izraza s distribucijom cijelog izraza. Njihova je hipoteza da kompozicionalni izrazi imaju sličnu distribuciju kao i njihove sastavnice. Za mjerenje distribucijske sličnosti koriste model latentne semantičke analize (LSA). Iako nisu pokušali klasificirati izraze kao kompozicionalne ili nekompozicionalne, pokazali su da postoji korelacija između kompozicionalnosti i izmjerene distribucijske sličnosti.

Katz i Giesbrecht (2006) nastavljaju istraživanje na Linovu i Baldwinovu ideju. Koriste latentnu semantičku analizu za modeliranje značenja riječi sastavnica i cijelog izraza te pokušavaju saznati može li razlika između konteksta u kojem se cijeli izraz pojavljuje i konteksta u kojem se sastavnice pojavljuju same identificirati nekompozicionalne izraze. Pretpostavka je da kod nekomponzicionlnih izraza vektor koji predstavlja zbroj njihovih sastavnica u DSM-u neće odgovarati pravom vektoru izraza. Za uspoređivanje aproksimiranog vektora izraza i pravog vektora izraza koristili su mjeru kosinusa kuta (formula 2.9). Na temelju zlatnog standarda dobivenog iz ručnih oznaka odredili su optimalnu udaljenost koja graniči kompozicionalne od nekompozicionalnih izraza (0,2; F1-mjera: 48 %).

Biemann i Giesbrecht (2011) organiziraju radionicu na temu distribucijske semantike i kompozicionalnosti *DisCO* (engl. *Distributional Semantics and COmpositionality*) te postavljaju zadatak izgradnje sustava koji će automatski predviđati kompozicionalnost višerječnog izraza isključivo na temelju korpusa, bez upotrebe rječnika i baze znanja. Zadatak se pokazao prilično teškim: od 7 prijavljenih timova i 19 različitih sustava nije bilo jasnog pobjednika, no sustavi temeljeni na distribucijskim semantičkim modelima u pravilu pokazali su se boljim od sustava temeljenih na leksičkim asocijacijskim mjerama.

Krcmár et al. (2013) eksperimentira s različitim vrstama distribucijskih semantičkih modela (VSM, LSA, HAL, RI, COALS) i s različitim mjerama kompozicionalnosti:

- mjere zasnovane na supstituciji: iskorištavaju svojstvo nezamjenjivosti (okamenjenosti) kod višerječnih izraza, sastavnice u izrazu ne mogu se zamijeniti sličnim riječima (*vodeni pas* vs. *morski pas*)
- mjere zasnovane na komponentama: iskorištavaju svojstvo nepromjenjivosti kod višerječnih izraza, značenje izraza ne može se aproksimirati njegovom glavnom sastavnicom (*pas* vs. *morski pas*)
- mjere zasnovane na kompozicionalnosti: iskorištavaju svojstvo nekompozicionalnosti, značenje izraza ne odgovora kompoziciji značenje dijelova

– mjere zasnovane na zajedničkim susjedima: uspoređuju zajedničke susjede izraza i njegovih sastavnica i traže preklapanje (*pas* → *{mačka, lajanje}* vs. *morski pas* → *{more, riba}*).

Rezultati pokazaju da su najučinkovitiji model LSA i model COALS (engl. *correlated occurrence analogue to lexical semantics*) (Rohde et al., 2006), dok mjere imaju varirajući uspjeh, ovisno o različitim tipovima izraza i frekvenciji pojavljivanja u korpusu.

U uvodu je već napomenuto da se ovaj rad ne bavi problemom određivanja kompozicionalnosti višerječnih izraza koji se mogu upotrebljaviti u svom prozirnom, ali i neprozirnom značenju (npr. *desna ruka, okrugli stol, čupati kosu, otvoriti vrata*). Taj je problem zapravo sličan problemu razrješavanja višeznačnosti riječi; po kontekstu se pokušava odrediti koristi li se izraz u prozirnom ili neprozirnom značenju. Zbog potpunosti ovog pregleda, u nastavku su navedeni radovi koji se bave tom tematikom iako to ovaj rad ne razmatra.

Katz i Giesbrecht (2006) napravili su eksperiment nad njemačkim višerječnim izrazom *ins Wasser fallen* koji može imati značenje:

– *(u)pasti u vodu* (prozirno)

– *izjaloviti se, ne realizirati se* (neprozirno).

U korpusu su za 67 pojava izraza ručno označili radi li se o prozirnom ili neprozirnom značenju te su u distribucijskom semantičkom modelu (LSA) izgradili vektore za prozirnu i neprozirnu upotrebu izraza. Vektori su ispali gotovo ortogonalni (kosinus kuta 0,02) čime su dokazali da postoji razlika u kontekstu između upotrebe izraza u prozirnom i neprozirnom značenju. Kako bi potvrdili da se ta razlika može koristiti u određivanju kompozicionalnosti izraza, izgradili su klasifikator temeljen na metodi najbližih susjeda te su na ispitnom skupu postigli točnost od 72 %.

Cook et al. (2007) nastavljaju na ideju iz rada (Katz i Giesbrecht, 2006), no kako bi izgradili oba vektora za izraz, umjesto ručnog označavanja koriste tri različite metode nenadziranog učenja koje se oslanjaju na pretpostavku da su neprozirne upotrebe izraza često u kanonskom obliku izraza (oblik koji se navodi u rječniku), a da su prozirne upotrebe obično u nekanonskom obliku izraza. Također za klasificiranje koriste metodu najbližih susjeda i postižu točnost od 72,4 %.

Sporleder i Li (2009) prozirnu upotrebu izraza od neprozirne razlikuju po tome što kod prozirne upotrebe izraza postoje poveznice između okolnog teksta i sastavnica izraza. Predlažu dvije metode koje iskorištavaju tu pretpostavku, jedna temeljena na leksičkim lancima, a druga na kohezijskim grafovima. Oba izgrađena klasifikatora postižu F1 mjeru od 60-ak %.

3. Zbirka višerječnih izraza

Modeliranje semantičke kompozicionalnosti zahtijeva reprezentativan skup višerječnih izraza s (ručnim) oznakama semantičke kompozicionalnosti. Budući da takav skup podataka ne postoji za hrvatski jezik, pripremljen je u okviru ovog rada. Zbirka višerječnih izraza izgrađena je po uzoru na rad (Biemann i Giesbrecht, 2011). Za izgradnju zbirke i modela odabran je korpus *fhrwac* (Šnajder et al., 2013), filtrirana verzija hrvatskog web korpusa *hrwac* (Ljubešić i Erjavec, 2011). Korpus je već razdvojen u rečenice, tokeniziran te gramatički i sintaktički označen. Sadrži 50 940 598 rečenica i 1 232 632 208 pojavnica. Prosječna je duljina rečenice 24 pojavnice.

U potpoglavlju 3.1 dana je definicija višerječnog izraza koju slijedi ovaj rad, a u potpoglavlju 3.2 objašnjen je postupak ekstrakcije i odabira izraza iz korpusa. Nakon što su pripremljeni izrazi za zbirku, među neprozirnim izrazima uočene su neke kategorije neprozirnosti koje su opisane u potpoglavlju 3.3. Potom je u potpoglavlju 3.4 opisan način dobivanja zlatnog standarda, a u zadnjem potpoglavlju (3.5) dan je pregled zbirke.

3.1. Definicija višerječnog izraza

Ne postoji univerzalna i jedinstvena definicija višerječnog izraza – različiti autori navode različite definicije, ovisno o problemu koji rješavaju ili podatcima koje trebaju. U literaturi se pojam višerječnog izraza blisko vezuje uz pojam kolokacije (engl. *collocation*). Opet postoji neslaganje oko toga jesu ta dva pojma potpuno istoznačna ili se samo u većoj mjeri preklapaju. Evert (2008) navodi da je kolokacija empirijski pojam, a višerječni izraz teorijski pojam. Kolokacijske definicije obično naglašavaju konvencionalnost, *uobičajen način da se nešto izrazi* (Manning i Schütze, 1999), *niz riječi koji se supojavljuje češće nego što je očekivano* (Firth, 1957; Sinclair, 1991), *bilo kakvo statistički značajno supojavljivanje* (Sag et al., 2002). Definicije za višerječni izraz više naglašavaju sintaktička, semantička i statistička svojstva sveze riječi. U sklopu ovog rada koristit će se definicija iz (Baldwin, 2006) prenesena iz (Šnajder,

2010): *kombinacija od dvije ili više (ne nužno slijednih) riječi čija semantička, sintaktička ili statistička obilježja nisu u potpunosti predvidiva, stoga takav izraz treba biti naveden u leksikonu.*

Svojstva koja višerječni izrazi imaju su (Manning i Schütze, 1999):

1. nekompozicionalnost (semantička neprozirnost): značenje izraza ne može se izravno zaključiti iz značenja dijelova

2. nezamjenjivost (okamenjost): dijelovi izraza ne mogu se zamijeniti sličnom riječi, npr. *zlatno vino* umjesto *bijelo vino*

3. nepromjenjivost: izraz se ne može modificirati ni na koji način, npr. *bacati lijepo biserje pred svinje* umjesto *bacati biserje pred svinje*.

Dodatno, višerječni izrazi mogu imati i neka druga zanimljiva svojstva (Mihaljević, 1991). Jedna riječ u jednom jeziku može odgovarati višerječnom izrazu u drugom jeziku (*iskoristiti* $\rightarrow$ *take advantage*, *programska podrška* $\rightarrow$ *software*). Nekad se višerječni izraz može zamijeniti jednom riječju (*otegnuti papke* $\rightarrow$ *umrijeti*, *stara cura* $\rightarrow$ *usidjelica*). Idiomatski izrazi mogu imati potpuno drugačije prijevode u drugim jezicima (*otegnuti papke* $\rightarrow$ *kick the bucket*).

Obično se višerječni izrazi dijele u različite tipove radi lakšeg prepoznavanja i boljeg razumijevanja. Jedna od podjela preuzeta iz rada (Delač, 2009; Šnajder, 2010) navedena je u nastavku:

1. frazemi: *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 et al., 2003) (*prodavati maglu, crna ovca*)

2. vlastita imena: imena osoba, institucija i sl. (*Jutarnji list, Cvjetni trg*)

3. stručno nazivlje: terminološki izrazi iz različitih struka (*operacijski sustav, koronarna arterija*)

4. leksičke kolokacije: standardan način da se nešto izrazi (*morski pas, javna tajna*)

5. ustaljene fraze i klišeji: izrazi koji se često upotrebljavaju u jeziku (*plan i program, dobar dan*).

Ovaj rad posebno razmatra frazeme, stručno nazivlje i leksičke kolokacije. Vlastita imena nisu uključena u izradi zbirke.

3.2. Ekstrakcija kandidata

Višerječni izrazi dolaze u različitim oblicima i jezičnim konstrukcijama i mogu se sastojati od proizvoljnog broja leksičkih jedinica. Kako bi olakšali zadatak izbora i evaluacije kandidata, u ovom radu razmatraju se samo dvorječni izrazi koji su u jednom od sljedećih gramatičko-sintaktičkih odnosa:

a) pridjev-imenica (AN): imenica i pridjev koji ju modificira (atribut) (npr. *plavo nebo, žuti karton*)

b) glagol-subjekt (VS): imenica u službi subjekta i glagol (npr. *stručnjak tvrdi, podatak govori*)

c) glagol-objekt (VO): imenica u službi objekta i glagol (npr. *bilježiti rast, raskinuti ugovor*).

Odabir prikladnih kandidata napravljen je u tri koraka:

1. ekstrakcija kandidata iz korpusa pomoću sintaktičkih i gramatičkih oznaka te sortiranje kandidata po frekvenciji pojavljivanja

2. filtriranje mogućih kompozicionalnih i nekompozicionalnih kandidata

3. ujednačavanje finalnog skupa (200 izraza) tako da uključuje jednak broj kompozicionalnih i nekompozicionalnih kandidata.

U prvom se koraku generiraju svi izrazi iz korpusa koji zadovoljavalju jedan od prethodno navedenih gramatičkih-sintaktičkih obrazaca. Međutim, velik broj takvih izraza nema određeno svojstveno značenje, tj. ne uklapa se u definiciju višerječnog izraza iz potpoglavlja 3.1, stoga se u drugom koraku takvi izrazi uklanjaju. Na kraju, nasumično iz skupa potencijalnih kandidata odabiru se finalni izrazi koji ulaze u zbirku, ali na takav način da su u jednakom broju zastupljeni kompozicionalni i nekompozicionalni izrazi, unatoč tome da su kompozicionalni izrazi puno učestaliji u korpusu nego nekompozicionalni. Taj korak opravdan je činjenicom da je u glavnom fokusu ovog rada svojstvo nekompozicionalnosti. U slučaju da su izrazi odabrani potpuno nasumično, velik broj izraza u zbirci bio bi kompozicionalan te bi sustav koji predviđa visoku razinu kompozicionalnosti postizao visoku točnost. Budući da je određivanje kompozicionalnosti subjektivan zadatak, odabir izraza u tom koraku pristran je s obzirom na osobu koja radi predselekciju, no prave ocjene kompozicionalnosti dobivaju se usrednjavanjem ocjena dobivenih od više označivača.

U tablici 3.1 prikazana je distribucija izraza po tipu. Najviše je izraza oblika imenica-pridjev. Izrazi glagol-subjekt su prilično rijetki i njih ima najmanje.

Tablica 3.1: Zastupljenost višerječnih izraza po vrstama

Vrsta izraza	Broj izraza
pridjev-imenica (AN)	125
glagol-subjekt (VS)	10
glagol-objekt (VO)	65

3.3. Vrste neprozirnosti

Tijekom postupka filtriranja prozirnih i neprozirnih izraza uočeno je da kod neprozirnih izraza postoji više različitih stupnjeva neprozirnosti. Primjerice, izraz *žuti karton* uistinu jest žuti karton, ali ima preneseno značenje: upozorenje. Izraz *siva ekonomija* jest vrsta ekonomije, ali nije doslovno *siva*. Kod izraza *trgovački lanac* riječ *lanac* ne predstavlja baš lanac. Sve su to neprozirnosti, ali različitog stupnja i inteziteta. Stoga je u okviru ovog rada razvijena podjela na tri vrste neprozirnosti navedene u nastavku (*NP* označava *neprozirno*).

NP1: Tip 1 predstavlja izraze koji su neprozirni u potpunosti, tj. značenje obje sastavnice jest neprozirno. Primjeri takvih izraza su: *žuti karton, preliti čašu, trljati ruke, mrtva točka*. S lingvističkog stajališta ova kategorija bi se dodatno mogla podijeliti na dva podskupa:

(a) Izrazi kod kojih uopće ne postoji poveznica između značenja izraza i sastavnica. Iz gore navedenih primjera to su izrazi *trljati ruke* i *mrtva točka*. Zašto *trljanje ruku* predstavlja zadovoljstvo? Zašto je točka *mrtva*?

(b) Izrazi kod kojih postoji poveznica između značenja izraza i sastavnica. U gornjem primjeru to su izrazi *žuti karton* (jer se u sportskim natjecanjima žuti karton dodjeljuje kao znak upozorenja) i *preliti čašu* (prijelomni trenutak u pretjerivanju).

NP2: Tip 2 predstavlja izraze koji su djelomično neprozirni; jedna od riječi u izrazu je neprozirna, a druga je prozirna. Primjeri takvih izraza su: *siva ekonomija* (vrsta ekonomije), *bilježiti rast* (ne misli se na doslovno bilježenje).

NP3: Tip 3 predstavlja izraze koji su djelomično neprozirni ako se gleda dominantno značenje riječi. Primjeri takvih izraza su: *trgovački lanac*, *modna kuća*, *skrenuti pažnju*. Tip 3 jest dosta sličan tipu 2, a posebice kod glagolskih izraza jer nekad nije jasno koja sve značenja riječ službeno nosi. Stoga je razlika između tipa 2 i tipa 3 definirana rječnikom tj. Hrvatskim jezičnim portalom.[1] Ako je tamo pod natuknicom kao jedno od značenja riječi navedeno njeno neprozirno značenje, to je onda tip 3, inače se radi o tipu 2.

Opažanje iz predselekcije jest da je najučestalija vrsta prozirnosti tip 3 (52 %), zatim tip 1 (27 %) te na kraju tip 2 (21 %). Od izraza koji spadaju u tip 1, samo ih je 6 koji spadaju u tip 1a. Tip je 3 dosta zanimljiv jer je negdje na granici između prozirnih i neprozirnih izraza. Neki od tih izraza su toliko uobičajeni da ih vjerojatno na prvi dojam ne bi doživjeli kao neprozirne, čak su i u rječniku definirane, npr. jedno od značenja riječi *lanac* jest *niz poduzeća, prodavaonica i sl. istoga vlasnika*, pa se zapravo nameće pitanje o porijeklu tog neprozirnog značenja.

3.4. Označavanje

Nakon što je odabrano dvjesto višerječnih izraza koji tvore zbirku, sljedeći je korak bio odrediti njihovu ocjenu kompozicionalnosti. Zlatni standard dobiven je usrednjavanjem ocjena ljudskih označivača. U svrhu smanjivanja vremenske zahtjevnosti označavanja, dvjesto je izraza nasumično podijeljeno u 4 grupe (A, B, C, D) po 50 izraza te je dodatno u svaku grupu dodano po 5 izraza iz preostale tri grupe. Preklapanje je napravljeno radi određivanja razine slaganja među svim označivačima. Dakle, zadatak svakog označivača bio je za 65 različitih izraza odrediti na ljestvici 1–5 koliko su doslovni u određenom kontekstu. Kontekstna rečenica odabrana je nasumično iz korpusa, s napomenom da je za svaki nekompozicionalni (nedoslovni) izraz provjereno da je odabran kontekst u kojem se izraz zaista koristi u svom nedoslovnom značenju.[2]

Upute za označivače s uzorcima zadataka mogu se naći u dodatku B. Eksperimentu označavanja pristupilo je 24 označivača volontera. Svakom je označivaču nasumično dodijeljena jedna od četiri grupe. Za ispunjavanje obrasca (65 izraza) bilo je potrebno oko 15–20 minuta. Osim 20 izraza koje su označili svi označivači, preostalih 180 izraza označilo je po 6 označivača. Za finalnu ocjenu izraza odabran je medijan njego-

[1] http://hjp.novi-liber.hr/

[2] Ovo je važno zbog izraza koji se mogu koristiti u doslovnom i nedoslovnom značenju, poput izraza *desna ruka* ili *okrugli stol*.

vih ocjena. Medijan nije osjetljiv na ekstremne vrijednosti i bolji je indikator srednje vrijednosti na malim uzorcima.

Slaganje označivača izračunato je upotrebom Krippendorffove alfe (Krippendorff, 2004; Hayes i Krippendorff, 2007). Krippendorffova je alfa koeficijent slaganja koji mjeri razinu složnosti među više označivača, pogodna je za različite vrste mjernih varijabli te je otporna na nedostajuće vrijednosti. Za razliku od nekih sličnih statistika za mjerenje slaganja među označivačima (Shrout i Fleiss, 1979; Cohen, 1968; Scott, 1955) koji računaju omjer između dobivenog i očekivanog slaganja, Krippendorffova alfa računa omjer između dobivenog i očekivanog neslaganja. Općenita forma za računanje koeficijenta prikazana je u formuli 3.1.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{3.1}$$

Oznaka D_o predstavlja uočeno neslaganje, a oznaka D_e predstavlja očekivano slučajno neslaganje. Ako koeficijent α iznosi 0, slaganje uopće ne postoji, a ako koeficijent α poprimi iznos 1, znači da je složnost između označivača savršena i dobiveni su podatci potpuno pouzdani. Za detaljnije informacije o računanju α pogledati (Hayes i Krippendorff, 2007).

U tablici 3.2 prikazano je slaganje za svaku grupu (6 označivača) i slaganje na preklapanju (24 označivača). Rezultati pokazuju umjereno slaganje što znači da je zadatak bio prilično težak i subjektivan, no unatoč tomu označivači su uspjeli u razumnoj mjeri dodijeliti konzistentne oznake.

Tablica 3.2: Slaganje označivača

Uzorak	Krippendorffova α
Grupa A	$0,587$
Grupa B	$0,506$
Grupa C	$0,490$
Grupa D	$0,586$
Svi (20 izraza)	$0,456$

Standarna devijacija odgovora je 1.342. Detaljnijim pregledom dobivenih ocjena može se zaključiti da su označivači lakše postizali složnost na izrazito kompozicionalnim izrazima i na izrazito nekompozicionalnim izrazima. U tablici 3.3 naveden je primjer višerječnih izraza koji su imali najveće slaganje (standardna devijacija odgovora je 0) i onih koji su imale najmanje slaganje (standardna devijacija odgovora u intervalu $[1,2;\ 1,5]$). Najbolje slaganje imaju izrazi koji su nedvojbeno kompozicionalni (*igrati*

nogomet, *financijska pomoć*, *pjevati pjesmu*) ili izrazi koji su vidljivo nekompozicionalni (*punom parom*, *trljati ruke*, *mrtva točka*). Kod izraza koji su imali najmanje slaganje najviše je izraza tipa NP3 (nekompozicionalni u odnosu na dominantno značenje riječi) što je zapravo i očekivano jer ljudi različito tumače što je dominantno značenje neke riječi (*mjere* u *poduzeti mjere*, *korak* u *prvi korak*). Zanimljivo je također primjetiti da su se u listi riječi s najmanjim slaganjem našli neprozirni izrazi, ali s prenesenim značenjem (*žuti karton*, *crveni karton*, *platiti cijenu*). npr. izraz *žuti karton* dobio je ocjene $5, 3, 5, 3, 1, 2$ što znači da su jedni označivači njegovo značenje doživljavali potpuno doslovno, kao komadić žutog papira, dok su drugi njegovo značenje doživljavali kao znak upozorenja, a treći kao nešto između.

Tablica 3.3: Višerječni izrazi koji su postigli najveću razinu slaganja i neslaganja

Višerječni izrazi s najboljim slaganjem	Višerječni izrazi s najgorim slaganjem
igrati nogomet	*zabilježiti rast*
služiti kaznu	*žuti karton*
financijska pomoć	*prvi korak*
pjevati pjesmu	*telefonska linija*
nemati sumnje	*crveni karton*
kardiovaskularna bolest	*zaštitna mjera*
punom parom	*koncentracijski logor*
plastična vrećica	*nemati pojma*
toplinska izolacija	*imati prostor*
mrtva točka	*platiti cijenu*
počiniti samoubojstvo	*poduzeti mjere*
trljati ruke	*ostvariti cilj*
poslati pismo	*cijena pasti*
riješiti problem	*nemoguća misija*
preliti čašu	*životno djelo*

Zanimljivo je još pogledati kako su se označivači slagali u odnosu na vrste izraza (pridjev-imenica, glagol-objekt i glagol-subjekt). U tablici 3.4 prikazano je slaganje označivača (Krippendorffova α) prema vrstama izraza. Glagolski izrazi grupirani su u jednu kategoriju jer njih općenito ima manje od pridjevskih, a pogotovo izraza glagol-subjekt, njih ima tek desetak u cijeloj zbirci. Osim u grupi D, generalno opažanje jest da je slaganje nešto više u pridjevskim izrazima (AN) nego u glagolskim (VS-VO). Ova činjenica ne iznenađuje jer je i tijekom predselekcije uočeno da su glagoli

problematični jer često znaju biti višeznačni (npr. *voditi, podnijeti, pasti*).

Tablica 3.4: Slaganje označivača po vrstama izraza

Uzorak	AN	VS-VO
Grupa A	$0,620$	$0,535$
Grupa B	$0,510$	$0,478$
Grupa C	$0,544$	$0,337$
Grupa D	$0,505$	$0,648$
Svi (20 izraza)	$0,452$	$0,439$

3.5. Opis zbirke

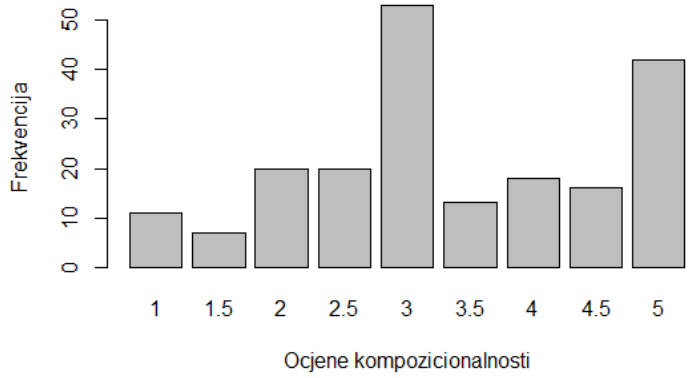
Dvjesto izraza iz zbirke nasumično je podijeljeno u dva podskupa – skup za učenje i skup za ispitivanje. Statistika po vrsti izraza prikazana je u tablici 3.5.

Tablica 3.5: Broj izraza po vrstama u skupovima za učenje i ispitivanje

	AN	VO	VS	ukupno
skup za učenje	60	35	5	100
skup za ispitivanje	65	30	5	100

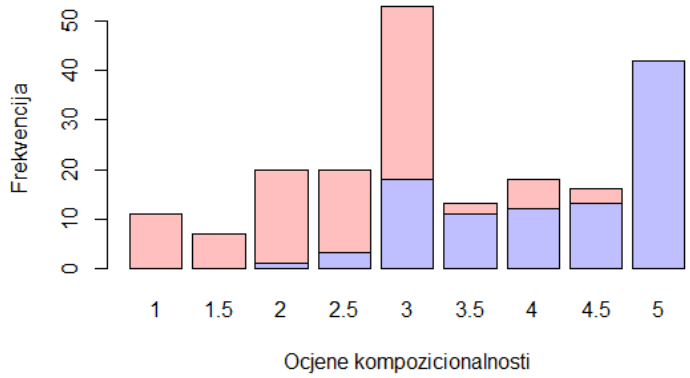
Dodatno, svakom izrazu dodijeljena je binarna oznaka kompozicionalnosti (*proziran* ili *neproziran*). Te oznake motivirane su činjenicom da je nekad dovoljno kategorizirati je li određeni izraz proziran ili neproziran, bez potrebe za detaljnijom informacijom o stupnju prozirnosti. Oznaka kompozicionalnosti dodijeljena je na temelju dobivenih ocjena. Na slici 3.1 prikazana je distribucija ocjena za sve izraze iz zbirke. Distribucija je ocjena bimodalna, što ukazuje na postojanje dvije grupe izraza – one s nižim ocjenama (neprozirna) i one s višim ocjenama (prozirna). Zbog toga se oznaka *proziran* dodijelila izrazima koji imaju ocjenu u intervalu $\langle 3, 5]$, a oznaka *neproziran* dodijelila se izrazima koji imaju ocjenu u intervalu $[1, 3]$. Prema ovom kriteriju, u zbirci se onda nalazi 111 neprozirnih (nekompozicionalni) izraza i 89 prozirnih (kompozicionalnih) izraza. Popis izraza s ocjenom i oznakom kompozicionalnosti može se pronaći u dodatku A.

U poglavlju 3.2 navedeno je da je napravljena predselekcija kandidata za zbirku takva da je ujednačen broj kompozicionalnih i nekompozicionalnih izraza. Budući da je taj korak pristran u odnosu na osobu koja radi predselekciju, zanimljivo je pogledati



Slika 3.1: Distribucija ocjena

u kakvom su odnosu ocjene označivača i grupe iz predselekcije. Upravo to jest prikazano na slici 3.2. Crveni stupci predstavljaju izraze koji su u predselekciji označeni kao nekompozicionalni, a plavi stupci predstavljaju izraze koji su u predselekciji označeni kao kompozicionalni. Grafikon pokazuje da su ocjene označivača uglavnom u skladu s grupama iz predselekcije. Najveće je preklapanje kod srednje ocjene (3), ali to je očekivano jer izrazi u sredini skale su neodređeni ("negdje između"), lako mogu prevagnuti na jednu ili drugu stranu, ovisno o perspektivi.



Slika 3.2: Distribucija ocjena u odnosu na grupe iz predselekcije

Nakon binarizacije ocjena kompozicionalnosti, može se ponovno razmotriti koliko su se označavači slagali ako se njihove ocjene pretvore u oznake *proziran* i *neproziran*. U tablici 3.6 prikazano je to slaganje. Slaganje je nešto slabije u odnosu na slaganje po ocjenama. Razlog tome može biti to što je sada razlika između ocjena 3 i 4 puno "teža"

nego što je bila na skali 1–5. Pretpostavka je da bi slaganje bilo više da su označivači sami dodjeljivali binarne oznake.

Tablica 3.6: Slaganje označivača nakon binarizacije

Uzorak	Krippendorffova α
Grupa A	$0,467$
Grupa B	$0,411$
Grupa C	$0,473$
Grupa D	$0,445$
Svi (20 izraza)	$0,401$

4. Model za određivanje semantičke kompozicije

Nakon danog pregleda područja i opisa zbirke, u ovom poglavlju opisano je kako je izgrađen model (potpoglavlje 4.1), dani su rezultati (potpoglavlje 4.2) i napravljena je analiza rezultata i pogreška (potpoglavlje 4.3).

4.1. Izgradnja modela

U poglavlju 2.1 dana je teorijska osnova za izgradnju modela i opisane su različite vrste distribucijskih semantičkih modela. Za evaluaciju zadatka ovog rada odabran je model latentne semantičke analize (LSA), po uzoru na rad (Katz i Giesbrecht, 2006) i rad (Karan et al., 2012) u kojem se model LSA pokazao najboljim na zadatku detektiranja sinonima hrvatskog jezika. Kontekst je definiran kao simetrični prozor od pet riječi. Budući da višerječni izraz nije nužno sastavljen od slijednih riječi, za kontekst cijelog izraza gleda se simetrični prozor od pet riječi za svaku sastavnicu, s tim da ako dođe do preklapanja prozora, riječi unutar tog preklapanja broje se samo jednom. Vektori za pojedinu sastavnicu grade se na temelju konteksta u kojima se sastavnica pojavljuje isključivo sama, a ne kao dio nekog izraza iz zbirke. Ideja iza toga postupka jest dodatno naglasiti nezavisnost sastavnice u izgradnji kompozicionalnog modela (Katz i Giesbrecht, 2006; Reddy et al., 2011)

Za kontekst riječi (stupci matrice) odabrano je 10 000 najčešćih riječi u korpusu (zaustavne riječi nisu uključene). U retcima matrice nalazi se 5000 najčešćih riječi u korpusu te svi izrazi iz zbirke i njihove sastavnice zasebno. Od težinskih funkcija eksperimentirano je s mjerom LMI (formula 2.6) i mjerom logaritam entropije (formula 2.7). Pomoću dekompozicije singularnih vrijednosti (formula 2.12) matrica je reducirana s 10 000 na k dimenzija. Optimalan iznos parametra k određen je na temelju skupa za učenje i iznosi 100.

Od distribucijskih modela semantičke kompozicije (poglavlje 2.2) isprobani su

multiplikativni, jednostavni i težinski aditivni model (formule 2.15, 2.14, 2.16) po uzoru na rad (Mitchell i Lapata, 2010). Sličnost vektora uspoređuje se mjerom kosinus kuta. Eksperimentirano je s dvije vrste težinskog aditivnog modela:

1. Parametri α i β odrede se na temelju skupa za učenje (Mitchell i Lapata, 2010), no s tom razlikom da su parametri neovisni o vrsti izraza. Ovo je opravdano činjenicom da je broj izraz po kategorijama malen.

2. Parametri α i β odrede se za svaki izraz posebno kao

$$\alpha = \frac{\cos(\overrightarrow{xy}, \overrightarrow{x})}{\cos(\overrightarrow{xy}, \overrightarrow{x}) + \cos(\overrightarrow{xy}, \overrightarrow{y})} \tag{4.1}$$

$$\beta = 1 - \alpha \tag{4.2}$$

   gdje je $\overrightarrow{xy}$ vektor višerječnog izraza, $\overrightarrow{x}$ vektor prve sastavnice, $\overrightarrow{y}$ vektor druge sastavnice, a $\cos$ je kosinus kuta (formula 2.9). Ideja je preuzeta iz rada (Reddy et al., 2011). Intuicija iza ovog dinamičkog računanja težina jest dati veći naglasak u težinskom zbrajanju onoj sastavnici koja je semantički sličnija (bliža) vektoru izraza.

Osim toga, dodatno su uspoređeni vektori samih sastavnica s vektorom izraza. To je zapravo posebna vrsta težinskog zbrajanja u kojem su α tj. β nula.

Sličnost između vektora pravog izraza i vektora izraza aproksimiranog jednom od funkcija kompozicije uspoređuje se mjerom kosinusa kuta (formula 2.9). Za predviđanje oznaka (*proziran*, *neproziran*) korištena je linearna kombinacija aditivnog modela, multiplikativnog modela i sastavnica po uzoru na jedan od najboljih sustava s radionice *DisCo* (Reddy et al., 2011):

$$\begin{aligned} MLR = a_0 + a_1 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x + y}) + a_2 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x * y}) \\ + a_3 \cdot \cos(\overrightarrow{xy}, \overrightarrow{x}) + a_4 \cdot \cos(\overrightarrow{xy}, \overrightarrow{y}) \end{aligned} \tag{4.3}$$

Radi se zapravo o višestrukoj linearnoj regresiji (engl. *multiple linear regression, MLR*) temeljenoj na metodi najmanjih kvadrata. Prag koji graniči prozirne od neprozirnih izraza određen je pretragom u intervalu $[0, 5]$ s korakom $0,01$ tako da maksimizira F1-mjeru na skupu za učenje.

4.2. Rezultati

Za vrednovanje modela odabran je Spearmanov koeficijent korelacije (engl. *Spearman's rank correlation coefficient*). Mjeri se povezanost između ocjena označivača

i ocjena koje daje model. Što je koeficijent veći (po apsolutnom iznosu), to je povezanost veća. Koeficijent može poprimiti iznos iz intervala $[-1, 1]$ gdje -1 ili 1 označavaju savršenu pozitivnu odnosno negativnu korelaciju, a 0 označava da nema korelacije. Formula za računanje Spearmanovog koeficijenta je:

$$\rho = 1 - 6 \sum_{i}^{n} \frac{d_i}{n(n^2 - 1)} \tag{4.4}$$

gdje je d_i razlika u rangovima para varijabli (x_i, y_i), a n je broj parova.

U poglavlju 4.2.1 dani su rezultati korelacije numeričkih ocjena modela s ocjenama označivača, a u poglavlju 4.2.2 dani su rezultati binarne klasifikacije oznaka kompozicionalnosti.

4.2.1. Predviđanje ocjene kompozicionalnosti

U tablicama 4.1 i 4.2 dani su rezultati za dva distribucijska semantička modela koji se razlikuju u težinskoj mjeri primijenjenoj na frekvencije supojavljivanja riječi, a to su logaritam entropije odnosno mjera LMI. Rezultati koji su podebljani statistički su značajni (p-vrijednost < $0,05$) u odnosu na broj izraza. P-vrijednost je vjerojatnost da promatrani uzorak podataka generira dobivenu vrijednost Spearmanove korelacije (ili veću) ako uopće ne postoji korelacija među podatcima. Za računanje p-vrijednosti korišten je dvostrani t-test. Pri vrednovanju rezultata po vrstama izraza, glagolski izrazi (glagol-subjekt i glagol-objekt) spojeni su u jednu kategoriju jer izraza vrste glagol-subjekt ima premalo za samostalnu evaluaciju.

Tablica 4.1: Rezultati korelacije za DSM 1; mjera logaritam entropije

model	ρ-AN-VO-VS	ρ-AN	ρ-VO-VS
multiplikativni	$-0,19$	$-0,20$	$-0,18$
jednostavni aditivni	$\mathbf{0,45}$	$\mathbf{0,54}$	$\mathbf{0,35}$
težinski aditivni (1)	$\mathbf{0,46}$	$\mathbf{0,56}$	$0,28$
težinski aditivni (2)	$\mathbf{0,46}$	$\mathbf{0,57}$	$0,26$
prva sastavnica	$\mathbf{0,41}$	$\mathbf{0,50}$	$0,19$
druga sastavnica	$\mathbf{0,28}$	$\mathbf{0,31}$	$0,31$
MLR (4.3)	$\mathbf{0,48}$	$\mathbf{0,56}$	$\mathbf{0,34}$

Rezultati pokazuju da je za ovaj zadatak primjerenija mjera logaritma entropije. Aditivni modeli kod oba DSM-a nadmašuju multiplikativni model. Rezultati su nešto

Tablica 4.2: Rezultati korelacije za DSM 2; mjera LMI

model	ρ-AN-VO-VS	ρ-AN	ρ-VO-VS
multiplikativni	$\mathbf{0,22}$	$\mathbf{0,24}$	$0,18$
jednostavni aditivni	$\mathbf{0,27}$	$0,21$	$\mathbf{0,40}$
težinski aditivni (1)	$\mathbf{0,29}$	$\mathbf{0,26}$	$0,31$
težinski aditivni (2)	$\mathbf{0,27}$	$0,21$	$\mathbf{0,42}$
prva sastavnica	$\mathbf{0,28}$	$\mathbf{0,27}$	$0,28$
druga sastavnica	$0,20$	$0,16$	$\mathbf{0,35}$
MLR (4.3)	$\mathbf{0,24}$	$0,21$	$0,27$

bolji u odnosu na rezultate sustava iz (Biemann i Giesbrecht, 2011). Kod njih je najbolji sustav imao ukupnu korelaciju od $0,35$, a ostali su imali između $0,27$ i $0,35$. No treba uzeti u obzir da su oni raspolagali s većim brojem izraza i imali su malo drugačiji sistem označavanja izraza. U svakom slučaju, može se zaključiti da DSM 1 daje relativno dobre rezultate. Korelacija je viša kod pridjevskih izraza, što je u skladu s radovima iz (Biemann i Giesbrecht, 2011) i (Krcmár et al., 2013). Vrijedi se podsjetiti da je i slaganje između označivača bilo veće kod pridjevskih izraza. Zanimljivo je primjetiti kako je prva sastavnica informativnija u odnosu na drugu kod pridjevskih izraza, dok je kod glagolskih izraza obrnuto.

4.2.2. Binarna klasifikacija kompozicionalnosti

Drugi pogled na modeliranje semantičke kompozicionalnosti jest binarna klasifikacija odnosno dodjeljivanje oznake *kompozicionalan* (*proziran*) ili *nekompozicionalan* (*neproziran*) višerječnom izrazu. Ovaj je zadatak u pravilu nešto lakši od predviđanja numeričke ocjene kompozicionalnosti, no opet mogu postojati neke situacije u kojima nije baš jasna granica između kompozicionalnosti i nekompozicionalnosti. Za binarnu klasifikaciju odabran je model MLR (formula 4.3), linearna kombinacija aditivnog i multiplikativnog modela te prve i druge sastavnice. Taj je model ujedno imao najviše ocjene korelacije u pobjedničkom DSM-u u prethodnom zadatku. Prag koji odvaja kompozicionalne od nekompozicionalnih izraza optimiran je na skupu za učenje na temelju F1-mjere. U tablici 4.3 i 4.4 prikazani su rezultati klasifikacije za dva DSM-a.

Ponovno DSM 1 daje nešto bolje rezultate, no ne i statistički značajnije. Može se primjetiti da su rezultati malo slabiji kod glagolskih izraza. U usporedbi s radom (Katz i Giesbrecht, 2006) koji je imao sličan zadatak, rezultati su malo bolji. Oni su klasificirali izraze kao kompozicionalne ili nekompozicionalne (no samo na temelju

Tablica 4.3: Rezultati klasifikacije za DSM 1; mjera logaritam entropije

	AN-VO-VS	AN	VO-VS
preciznost	$0,58$	$0,74$	$0,43$
odziv	$0,73$	$0,65$	$0,77$
točnost	$0,65$	$0,72$	$0,54$
F1-mjera	$\mathbf{0,65}$	$\mathbf{0,69}$	$\mathbf{0,56}$

aditivnog modela) te su postigli F1-mjeru od $0,48$.

Tablica 4.4: Rezultati klasifikacije za DSM 2; mjera LMI

	AN-VO-VS	AN	VO-VS
preciznost	$0,55$	$0,6$	$0,41$
odziv	$0,59$	$0,58$	$0,54$
točnost	$0,61$	$0,59$	$0,54$
F1-mjera	$\mathbf{0,57}$	$\mathbf{0,61}$	$\mathbf{0,47}$

Iako je automatsko određivanje semantičke kompozicionalnosti težak zadatak, rezultati upućuju na to da distribucijski modeli semantičke kompozicije mogu ponuditi relativno pristojno rješenje. Istraživanja su na tu tematiku to već potvrdila za višerječne izraze engleskog jezika, a rezultati ovog rada daju naznaku da slično vrijedi i za višerječne izraze hrvatskog jezika.

4.3. Analiza pogrešaka

Cilj je ovog poglavlja napraviti analizu pogrešaka, otkriti gdje model najviše griješi, odnosno koji izrazi imaju najslabiju korelaciju s ocjenama označivača, te zašto se to događa. Za analiziranje rezultata uzet je najbolji model iz prethodnog poglavlja, a to je model MLR (formula 4.3) iz DSM-a 1 koji postiže korelaciju od $0,48$ na skupu za ispitivanje.

Iako su ocjene tog modela i ocjena označivača na istoj skali, možda nemaju istu distribuciju, stoga, kako bi usporedili te dvije populacije, treba ih svesti na standardnu normalnu distribuciju (engl. *standard normal distribution*). Za obje populacije ocjena izračunate su njihove Z-vrijednosti (engl. *Z-score*) prema formuli 4.5.

$$Z = \frac{x - \mu}{\sigma} \tag{4.5}$$

gdje je x vrijednost elementa iz populacije, μ je artimetička sredina populacije, a σ je standardna devijacija. Kako bi otkrili koji izrazi imaju najveće odstupanje između ocjena označivača i ocjena modela, izračunata je apsolutna razlika njihovih Z-vrijednosti. U tablici 4.5 prikazano je 10 pridjevskih i 10 glagolskih izraza koji imaju najveće odstupanje, odnosno na kojima model najviše griješi. U tablici do izraza u zagradama naveden je tip neprozirnosti iz poglavlja 3.3 (NP1, NP2, NP3) za neprozirne izraze ili P za prozirne izraze.

Tablica 4.5: Višerječni izrazi na kojima model najviše griješi

AN	VO-VS
organizacijski odbor (P)	*nemati sumnje* (P)
izvršna vlast (P)	*dati život* (NP3)
financijska pomoć (P)	*optužnica teretiti* (P)
novi val (NP1)	*spasiti život* (P)
misno slavlje (NP3)	*uroditi plodom* (NP1)
internetski portal (P)	*zabiti gol* (P)
oglasna ploča (P)	*otvoriti vrata* (NP1)
životno djelo (P)	*poduzeti korak* (NP3)
rodni grad (P)	*uložiti žalbu* (P)
morski pas (NP2)	*smanjiti rizik* (P)

Lista nagovještava da model najviše griješi kod prozirnih izraza, i to uglavnom kod onih za koje su se označivači više-manje bez problema složili da se radi o visokom (5) stupnju prozirnosti (*organizacijski odbor*, *izvršna vlast*, *financijska pomoć*, *nemati sumnje*, *rodni grad*, *smanjiti rizik*). Pitanje je zašto se to događa? Teško je reći, npr. riječ *organizacijski* i riječ *odbor* tek su ponešto slične izrazu *organizacijski odbor* (kosinus kuta za obje sastavnice jest negdje oko $0,2$); pa onda ni njihova kompozicija ne može biti slična pravom vektoru izraza. Treba ponovno napomenuti da su vektori sastavnica izgrađeni samo iz onih konteksta u kojima se sastavnice pojavljuju same. Možda se sastavnice i izraz jednostavno ne pojavljuju u istim kontekstima, iako su sastavnice potpuno prozirne i ne nose idiomatsko značenje.

Zanimljivo je da se u 20 izraza s najvećim odstupanjem nalazi otprilike jednak broj glagolskih i pridjevskih izraza iako pridjevskih izraza ima dvostruko više. Ta činjenica nije iznenađujuća jer su glagolski izrazi imali i manje slaganje kod označivača i manji iznos korelacije kod gotovo svih modela.

U tablici 4.6 prikazana je dvodimenzionalna tablica primjera koja daje pogled na

odnos između korelacije ocjena modela s ocjenama označivača i razine slaganja među označivačima. Već je ranije zaključeno da su označivači lakše postizali slaganje na izrazima koji su očigledno prozirni ili neprozirni. U vezi s korelacijom unutar te podgrupe izraza ne može se puno zaključiti; od 26 prozirnih izraza iz te podgrupe, 15 ih ima nižu korelaciju, a neprozirni izrazi tipa 3 (NP3) postižu veću korelaciju u odnosu na neprozirne izraze tipa 1 (NP1). Unutar podgrupe s nižom razinom slaganja izrazi vrste NP3 i NP1 postižu višu razinu korelacije u odnosu na prozirne izraze. U tablici 4.7 prikazan je odnos između razine slaganja označivača i korelacije modela, ali po grupama prozirnosti. Generalno je opažanje da ocjene neprozirnih izraza imaju nešto višu razinu korelaciju u odnosu na ocjene prozirnih izraza.

Tablica 4.6: Odnos razine slaganja s korelacijom po izrazima

Visoka razina slaganja označivača	
visoka korelacija	***niska korelacija***
mrtva točka (NP1)	*nemati sumnje* (P)
pjevati pjesmu (P)	*morski pas* (NP2)
počiniti samoubojstvo (P)	*novi val* (NP1)
plastična vrećica (P)	*financijska pomoć* (P)
širok krug (NP3)	*plaćati porez* (P)
medicinska sestra (NP3)	*izvršna vlast* (P)
plastična vrećica (P)	*otvoriti vrata* (NP1)
dnevni boravak (NP3)	*organizacijski odbor* (P)
vatreno oružje (NP3)	*misan slavlje* (NP3)
istraživanje pokazati (NP3)	*motorno vozilo* (P)
Niska razina slaganja označivača	
visoka korelacija	***niska korelacija***
ostaviti dojam (NP3)	*kućni ljubimac* (NP3)
zatvorena vrata (NP1)	*uroditi plodom* (NP1)
policijske snage (NP3)	*zabiti gol* (P)
podnijeti ostavku (NP3)	*dati život* (NP3)
crna kutija (NP1)	*internetski portal* (P)
nemati veze (NP3)	*obiteljsko gospodarstvo* (P)
voditi ljubav (NP1)	*modna kuća* (NP3)
igrati ulogu (NP1)	*nemati pojma* (NP3)
zaštitna mjera (P)	*oglasna ploča* (P)
zapaliti svijeću (P)	*mobilna telefonija* (P)

Tablica 4.7: Odnos razine slaganja s korelacijom po vrstama neprozirnosti izraza

Visoka razina slaganja označivača		*Niska razina slaganja označivača*	
visoka korelacija	*niska korelacija*	*visoka korelacija*	*niska korelacija*
P: 11	P: 15	P: 8	P: 14
NP1: 1	NP1: 7	NP1: 5	NP1: 1
NP2: 2	NP2: 3	NP2: 2	NP2: 0
NP3: 8	NP3: 5	NP3: 13	NP3: 5

5. Zaključak

Višerječni izrazi zbog svojih sintaktičkih i semantičkih obilježja iziskuju posebnu pažnju u obradi prirodnog jezika. Obilježje koje je bilo u fokusu ovog rada jest semantička nekompozicionalnost ili neprozirnost. Izraze koji su semantički neprozirni ili nekompozicionalni nije moguće modelirati raščlambom na sastavne riječi, poput izraza *morski pas* ili *ležeći policajac*. Njihovo značenje ne odgovora kompoziciji značenja sastavnica. Cilj ovog rada bio je iskoristiti upravo to svojstvo nekompozicionalnosti u distribucijskim semantičkim modelima kako bi diskriminirali kompozicionalne od nekompozicionalnih izraza.

U okviru ovog rada izgrađena je zbirka višerječnih izraza koja se sastoji od 200 izraza koji su ručno označeni ocjenama semantičke kompozicionalnosti. Također je izgrađen distribucijski model latentne semantičke analize za određivanje semantičke kompozicionalnosti na temelju statističkog pojavljivanja riječi i izraza u korpusu. Eksperimentirano je s dvije vrste težinskih mjera, logaritam entropije i mjera LMI, od kojih se logaritam entropije pokazao boljom mjerom. Od distribucijskih modela semantičke kompozicije, aditivni modeli pokazali su se superiornijim u odnosu na multiplikativni. Dobiveni su rezultati statistički značajni i u rangu onih (Biemann i Giesbrecht, 2011) i (Katz i Giesbrecht, 2006) te nagovještaju da distribucijska semantika može ponuditi odgovor na automatsku identifikaciju semantičke kompozicionalnosti višerječnih izraza. No, problem je daleko od riješenog, potrebno je još istraživanja i eksperimentiranja kako bi se rezultati unaprijedili.

Kao dio budućeg rada predlaže se kao prvo proširenje zbirke višerječnih izraza jer veći broj izraza omogućava pouzdano analiziranje i zaključivanje. Kao drugo, predlaže se detaljnije eksperimentiranje s parametrima distribucijskih semantičkih modela i modela distribucijske semantičke kompozicije, moguće čak i uz kombiniranje metoda strojnog učenja ili genetskog programiranja. Na kraju, predlaže se detaljnije analiziranje kompozicionalnosti višerječnih izraza s lingvističke strane, u svrhu boljeg razumijevanja problema, a time i izgradnje boljih modela.

LITERATURA

Otavio Costa Acosta, Aline Villavicencio, i Viviane P. Moreira. Identification and treatment of multiword expressions applied to information retrieval. U *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, stranice 101–109, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 978-1-932432-97-8. URL `http://dl.acm.org/citation.cfm?id=2021121.2021141`.

Timothy Baldwin. Compositionality and multiword expressions: Six of one, half a dozen of the other. U *Invited talk given at the COLING/ACL'06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July*. Citeseer, 2006.

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, i Dominic Widdows. An empirical model of multiword expression decomposability. U *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, stranice 89–96. Association for Computational Linguistics, 2003.

Marco Baroni i Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. U *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, stranice 1183–1193. Association for Computational Linguistics, 2010.

Božo Bekavac. Strojno obilježavanje hrvatskih tekstova-stanje i perspektive. *Suvremena lingvistika*, (53-54):173–182, 2002.

Chris Biemann i Eugenie Giesbrecht. Distributional semantics and compositionality 2011: Shared task description and results. U *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, stranice 21–28, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284022. URL `http://dl.acm.org/citation.cfm?id=2043121.2043125`.

John A Bullinaria i Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3): 510–526, 2007.

Marine Carpuat i Mona Diab. Task-based evaluation of multiword expressions: A pilot study in statistical machine translation. U *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, stranice 242–245, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858028`.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

Paul Cook, Afsaneh Fazly, i Suzanne Stevenson. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. U *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, stranice 41–48, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1613704.1613710`.

James Richard Curran. From distributional to semantic similarity. 2004.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, i Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.

Davor Delač. Postupci ekstrakcije kolokacija iz zbirki tekstova. 2009.

Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, July 1945. URL `http://www.jstor.org/pss/1932409`.

Stefan Evert. Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2:223–233, 2008.

Mark Alan Finlayson i Nidhi Kulkarni. Detecting multi-word expressions improves word sense disambiguation. U *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, MWE '11, stranice 20–24, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

ISBN 978-1-932432-97-8. URL http://dl.acm.org/citation.cfm?id=2021121.2021128.

J. R. Firth. A synopsis of linguistic theory 1930-55. 1952-59:1–32, 1957.

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*, stranice 33–37. Association for Computational Linguistics, 2010.

Zellig Harris. Distributional structure. *Word*, 10(23):146–162, 1954.

Andrew F Hayes i Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89, 2007.

Aapo Hyvärinen, Juha Karhunen, i Erkki Oja. *Independent component analysis*, svezak 46. John Wiley & Sons, 2004.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

Mladen Karan, Jan Šnajder, i Bojana Dalbelo Bašić. Distributional semantics approach to detecting synonyms in croatian language. *Information Society*, stranice 111–116, 2012.

Graham Katz i Eugenie Giesbrecht. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. U *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, MWE '06, stranice 12–19, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. ISBN 1-932432-84-1. URL http://dl.acm.org/citation.cfm?id=1613692.1613696.

Walter Kintsch. Predication. *Cognitive Science*, 25(2):173–202, 2001.

Lubomır Krcmár, Karel Jezek, i Pavel Pecina. Determining compositionality of word expressions using various word space models and measures. *ACL 2013*, stranica 64, 2013.

Klaus Krippendorff. Reliability in content analysis. *Human Communication Research*, 30(3):411–433, 2004.

Landauer. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007. ISBN 0805854185. URL `http://www.lob.de/cgi-bin/work/suche2?titnr=248463920&flag=`.

Thomas K Landauer i Susan T. Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, stranice 211–240, 1997.

Alessandro Lenci. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31, 2008.

Dekang Lin. Automatic identification of non-compositional phrases. U *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, stranice 317–324. Association for Computational Linguistics, 1999.

Nikola Ljubešić i Tomaž Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. U *Text, Speech and Dialogue*, stranice 395–402. Springer, 2011.

Kevin Lund i Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.

Christopher D Manning i Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

Christopher D Manning, Prabhakar Raghavan, i Hinrich Schütze. *Introduction to information retrieval*, svezak 1. Cambridge university press Cambridge, 2008.

A. Menac, Ž. Fink-Arsovski, i R. Venturin. *Hrvatski frazeološki rječnik*. Naklada Ljevak, 2003. ISBN 9789531785877. URL `http://books.google.hr/books?id=m2tSNQAACAAJ`.

Milica Mihaljević. Višerječne natuknice i podnatuknice u jednojezičnom općem rječniku hrvatskoga jezika. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, 17(1): 133–144, 1991.

Jeff Mitchell i Mirella Lapata. Vector-based models of semantic composition. U *ACL*, stranice 236–244, 2008.

Jeff Mitchell i Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

Preslav Nakov, Antonia Popova, i Plamen Mateev. Weight functions impact on lsa performance. *EuroConference RANLP*, stranice 187–193, 2001.

Barbara Partee. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360, 1995.

Francis Jeffry Pelletier. The principle of semantic compositionality. *Topoi*, 13(1): 11–24, 1994.

Siva Reddy, Diana McCarthy, Suresh Manandhar, i Spandana Gella. Exemplar-based word-space model for compositionality detection: Shared task system description. U *Proceedings of the Workshop on Distributional Semantics and Compositionality*, DiSCo '11, stranice 54–60, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284022. URL `http://dl.acm.org/citation.cfm?id=2043121.2043131`.

Douglas L. T. Rohde, Laura M. Gonnerman, i David C. Plaut. An improved model of semantic similarity based on lexical co-occurence. *COMMUNICATIONS OF THE ACM*, 8:627–633, 2006.

Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, i Dan Flickinger. Multiword expressions: A pain in the neck for nlp. U *Computational Linguistics and Intelligent Text Processing*, stranice 1–15. Springer, 2002.

Magnus Sahlgren. An introduction to random indexing. U *Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE*, svezak 5, 2005.

G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

William A Scott. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, 1955.

Patrick E Shrout i Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

John Sinclair. *Corpus, concordance, collocation*, svezak 1. Oxford University Press Oxford, 1991.

Lindsay I Smith. A tutorial on principal components analysis. *Cornell University, USA*, 51:52, 2002.

Jan Šnajder. Genetičko programiranje susreće lingvistiku: računalni postupci ekstrakcije kolokacija iz korpusa. 2010.

Jan Šnajder, Sebastian Padó, i Željko Agić. Building and evaluating a distributional memory for croatian. U *51st Annual Meeting of the Association for Computational Linguistics*, stranice 784–789, 2013.

Richard Socher, Brody Huval, Christopher D Manning, i Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. U *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, stranice 1201–1211. Association for Computational Linguistics, 2012.

Caroline Sporleder i Linlin Li. Unsupervised recognition of literal and non-literal use of idiomatic expressions. U *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, stranice 754–762, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1609067.1609151`.

Peter D Turney. Domain and function: A dual-space model of semantic relations and compositions. *arXiv preprint arXiv:1309.4035*, 2013.

Peter D. Turney i Patrick Pantel. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Siječanj 2010. ISSN 1076-9757. URL `http://dl.acm.org/citation.cfm?id=1861751.1861756`.

Dominic Widdows i Kathleen Ferraro. Semantic vectors: a scalable open source package and online technology management application. U *LREC*, 2008.

Dodatak A
Zbirka višerječnih izraza

U tablici A.1 navedena je lista višerječnih izraza koji su ušli u zbirku izraza. Također, uz svaki izraz navedena je njegova ocjena semantičke kompozicionalnosti na ljestvici 1–5 gdje 1 predstavlja potpuno nekompozicionalan (neproziran) izraz, a 5 predstavlja potpuno kompozicionalan (proziran) višerječni izraz. Ocjena kompozicionalnosti jest medijan svih ocjena dobivenih od više označivača. Oznaka je kompozicionalnosti binarna oznaka dobivena zaokruživanjem ocjene. Oznaku *neproziran* (nekompozicionalan) dobili su izrazi koji imaju ocjenu u intervalu $[1, 3]$, dok su oznaku *proziran* (kompozicionalan) dobili izrazi koji imaju ocjenu u intervalu $\langle 3, 5]$.

Tablica A.1: Zbirka višerječnih izraza

Zbirka višerječnih izraza		
Višerječni izraz	Ocjena kompozicionalnosti	Oznaka kompozicionalnosti
hladni rat	1.0	neproziran
desni centar	1.0	neproziran
punom parom	1.0	neproziran
zeleno svjetlo	1.0	neproziran
crna rupa	1.0	neproziran
novi val	1.0	neproziran
mrtva točka	1.0	neproziran
preliti čašu	1.0	neproziran
dići ruke	1.0	neproziran
trljati ruke	1.0	neproziran
voditi račun	1.0	neproziran
pun pogodak	1.5	neproziran
crna kutija	1.5	neproziran

slobodni radikal	1.5	neproziran
zaštitno lice	1.5	neproziran
mali čovjek	1.5	neproziran
voditi ljubav	1.5	neproziran
uroditi plodom	1.5	neproziran
imati težinu	2.0	neproziran
pronaći put	2.0	neproziran
odigrati ulogu	2.0	neproziran
lijevi centar	2.0	neproziran
životni put	2.0	neproziran
napraviti korak	2.0	neproziran
vrijeme pokazati	2.0	neproziran
imati riječ	2.0	neproziran
otvoriti vrata	2.0	neproziran
odmjeriti snage	2.0	neproziran
morski pas	2.0	neproziran
carski rez	2.0	neproziran
riješiti pitanje	2.0	neproziran
siva ekonomija	2.0	neproziran
časna sestra	2.0	neproziran
zajednički jezik	2.0	neproziran
okrugli stol	2.0	neproziran
širok krug	2.0	neproziran
radna snaga	2.0	neproziran
igrati ulogu	2.0	neproziran
odnijeti pobjedu	2.5	neproziran
činjenica govoriti	2.5	neproziran
izgubiti život	2.5	neproziran
dnevni boravak	2.5	neproziran
zaštitna mjera	2.5	neproziran
podatak govoriti	2.5	neproziran
poduzeti korak	2.5	neproziran
državni vrh	2.5	neproziran
zdrav razum	2.5	neproziran
lagana vatra	2.5	neproziran
crveni tepih	2.5	neproziran

modna kuća	2.5	neproziran
mlada nada	2.5	neproziran
optužnica teretiti	2.5	neproziran
zatvorena vrata	2.5	neproziran
ostaviti dojam	2.5	neproziran
teško vrijeme	2.5	neproziran
imati prostor	2.5	neproziran
državno tijelo	2.5	neproziran
široka potrošnja	2.5	neproziran
podići optužnicu	3.0	neproziran
životno djelo	3.0	neproziran
izgubiti utakmicu	3.0	neproziran
uložiti žalbu	3.0	neproziran
politička scena	3.0	neproziran
osvojiti glas	3.0	neproziran
dati glas	3.0	neproziran
žuti karton	3.0	neproziran
staklenički plinovi	3.0	neproziran
posebna skrb	3.0	neproziran
dati ostavku	3.0	neproziran
otvoreno more	3.0	neproziran
medicinska sestra	3.0	neproziran
trgovački lanac	3.0	neproziran
društvena mreža	3.0	neproziran
robna kuća	3.0	neproziran
cijena iznositi	3.0	neproziran
prvi korak	3.0	neproziran
crno tržište	3.0	neproziran
upisati pobjedu	3.0	neproziran
otvoreno pismo	3.0	neproziran
mali poduzetnik	3.0	neproziran
internetski portal	3.0	neproziran
nemati pojma	3.0	neproziran
nemati veze	3.0	neproziran
televizijska kuća	3.0	neproziran
nemoguća misija	3.0	neproziran

podnijeti prijavu	3.0	neproziran
plavi dizel	3.0	neproziran
ostaviti trag	3.0	neproziran
platiti cijenu	3.0	neproziran
bilježiti rast	3.0	neproziran
slobodno tržište	3.0	neproziran
voditi brigu	3.0	neproziran
optuženička klupa	3.0	neproziran
ubaciti poen	3.0	neproziran
istraživanje pokazivati	3.0	neproziran
diskografska kuća	3.0	neproziran
primiti gol	3.0	neproziran
skrenuti pažnju	3.0	neproziran
cijena pasti	3.0	neproziran
slobodno bacanje	3.0	neproziran
plastična operaija	3.0	neproziran
izboriti plasman	3.0	neproziran
služiti kaznu	3.0	neproziran
seksualna orijentacija	3.0	neproziran
policijske snage	3.0	neproziran
zabilježiti pad	3.0	neproziran
tiskovna konferencija	3.0	neproziran
vatreno oružje	3.0	neproziran
elektronska pošta	3.0	neproziran
poduzeti mjere	3.0	neproziran
propustiti priliku	3.0	neproziran
seksualni život	3.5	proziran
biometeorološka prilika	3.5	proziran
ljubavni život	3.5	proziran
podnijeti ostavku	3.5	proziran
sudski vještak	3.5	proziran
zauzeti mjesto	3.5	proziran
privući pozornost	3.5	proziran
duhovna obnova	3.5	proziran
zabiti gol	3.5	proziran
osvojiti nagradu	3.5	proziran

komunalna usluga	3.5	proziran
vojne snage	3.5	proziran
kućni ljubimac	3.5	proziran
svinjska gripa	4.0	proziran
obiteljsko gospodarstvo	4.0	proziran
misno slavlje	4.0	proziran
svjetski prvak	4.0	proziran
teretno vozilo	4.0	proziran
ljudski život	4.0	proziran
mali nogomet	4.0	proziran
snositi trošak	4.0	proziran
zapaliti svijeću	4.0	proziran
operacijski sustav	4.0	proziran
promet odvijati	4.0	proziran
radni stol	4.0	proziran
narodna nošnja	4.0	proziran
dati život	4.0	proziran
pružiti pomoć	4.0	proziran
koncentracijski logor	4.0	proziran
telefonska linija	4.0	proziran
brokerska kuća	4.0	proziran
umjetničko djelo	4.5	proziran
zabilježiti rast	4.5	proziran
turistička atrakcija	4.5	proziran
izvršna vlast	4.5	proziran
arbitražni sporazum	4.5	proziran
pristupni pregovori	4.5	proziran
investicijski fond	4.5	proziran
radni odnos	4.5	proziran
crveni karton	4.5	proziran
simfonijski orkestar	4.5	proziran
investicijska banka	4.5	proziran
zatresti mrežu	4.5	proziran
doživotni zatvor	4.5	proziran
oglasna ploča	4.5	proziran
smještajni kapacitet	4.5	proziran

međunarodno pravo	4.5	proziran
biciklistička staza	5.0	proziran
igrati nogomet	5.0	proziran
glavni urednik	5.0	proziran
smanjiti rizik	5.0	proziran
samoubilački napad	5.0	proziran
financijska pomoć	5.0	proziran
rodni grad	5.0	proziran
dizelski motor	5.0	proziran
motorno vozilo	5.0	proziran
mobilna telefonija	5.0	proziran
nogometni klub	5.0	proziran
ruralno područje	5.0	proziran
pjevati pjesmu	5.0	proziran
seksualno zlostavljanje	5.0	proziran
stručnjak tvrditi	5.0	proziran
popiti kavu	5.0	proziran
nemati sumnje	5.0	proziran
kardiovaskularna bolest	5.0	proziran
medicinska oplodnja	5.0	proziran
humanitarna svrha	5.0	proziran
zlatna medalja	5.0	proziran
nevladina organizacija	5.0	proziran
stečajni upravitelj	5.0	proziran
televizijska emisija	5.0	proziran
raskinuti ugovor	5.0	proziran
spasiti život	5.0	proziran
plastična vrećica	5.0	proziran
navijačka skupina	5.0	proziran
policija uhititi	5.0	proziran
toplinska izolacija	5.0	proziran
gospodarski kriminal	5.0	proziran
plaćati porez	5.0	proziran
komunalni otpad	5.0	proziran
krvni tlak	5.0	proziran
počiniti samoubojstvo	5.0	proziran

poslati pismo	5.0	proziran
riješiti problem	5.0	proziran
ostvariti cilj	5.0	proziran
policijska postaja	5.0	proziran
organizacijski odbor	5.0	proziran
saborski zastupnik	5.0	proziran
maslinovo ulje	5.0	proziran

Dodatak B
Upute za označivače

Označavanje doslovnosti višerječnih izraza

Ovaj upitnik sastoji se od 65 izraza za koje Vi morate odlučiti koliko doslovno značenje imaju. Možete li shvatiti značenje cijelog izraza razmatrajući njegove dijelove potpuno doslovno ili izraz nosi neko "posebnije" značenje? Razmislite o doslovnom značenju pojedine sastavnice i provjerite je li to isto značenje nosi u cijelom izrazu u danom kontekstu. Riječi u izrazu ne moraju nužno biti susjedne.
Označite ponuđene izraza na ljestvici od 1 do 5 gdje je:

1 => izraz koji se u danom kontekstu uopće ne može shvatiti doslovno

5 => izraz koji se u danom kontekstu može shvatiti skroz doslovno

Nekoliko primjera slijedi u nastavku.

Primjer 1:
Izraz: medeni mjesec
Kontekst: Dan poslije vjenčanja otputovali su na MEDENI MJESEC u Italiju gdje će obići sve veće talijanske gradove.
Ocjena: 1
Razlog: Medeni mjesec označava bračno putovanje, nema veze s medom, niti traje doslovno mjesec dana.

Primjer 2:
Izraz: ratni zločin
Kontekst: Karadžiću se sudi za genocid i RATNE ZLOČINE tijekom rata u BiH.
Ocjena: 5
Razlog: Ratni zločin je zločin počinjen tijekom rata i to se može shvatiti iz riječi "ratni"

i "zločin" bez ikakvog dodatnog znanja.

Primjer 3:
Izraz: plava kosa
Kontekst: Upravo uoči ovog showa, Mare je svoju PLAVU KOSU obojila u tamno smeđu.
Ocjena: 3
Razlog: Radi se o kosi, ali plava u ovom smislu znači žuta pa se može reći da je izraz negdje između potpuno doslovnog i nimalo doslovnog.

Primjer 4:
Izraz: voditi borbu
Kontekst: Njena majka s bakom i djedom otada VODI BORBU za skrbništvo nad problematičnom kćeri.
Ocjena: 2
Razlog: Ne radi se baš doslovno o borbi, niti se nešto vodi u smislu upravljanja kretanjem nekoga ili nečega, ali ipak je na neki način povezano s pravim značenjem: "boriti se za nešto".

Primjer 5:
Izraz: odati priznanje
Kontekst: Barack Obama u nedjelju je u Bijeloj kući ODAO PRIZNANJE Bruceu Springsteenu (60) za izniman doprinos američkoj kulturi.
Ocjena: 3
Razlog: Odati ovdje nije u svom doslovnom, dominatnom značenju: prokazati, potkazati, oktriti, izdati. Priznanje se može shvatiti doslovno.

NB: U navednim primjerima ne postoji "točan odgovor", zadatak je subjektivan i ovdje pitamo za Vaše mišljenje.
NB2: Neke riječi, poput "odati", "voditi" imaju više (različitih) značenja. U ovom upitniku, kad Vas pitamo za doslovno značenje, mislimo na ono dominatno značenje, obično ono koje Vam prvo padne na pamet kad čujete samu riječ ili koje je prvo navedeno na Hrvatskom jezičnom portalu (hjp.novi-liber.hr).

Model za određivanje semantičke kompozicionalnosti višerječnih izraza hrvatskoga jezika

Sažetak

Automatsko određivanje semantičke kompozicionalnosti višerječnih izraza važno je za niz primjena obrade prirodnog jezika poput strojnog prevođenja i pretraživanja informacija. U ovom radu rješavanju tog problema pristupa se upotrebom distribucijskih semantičkih modela i modela distribucijske semantičke kompozicije. Izgrađena je zbirka od dvjesto hrvatskih višerječnih izraza s ručnim ocjenama semantičke kompozicionalnosti na kojima su modeli vrednovani. Od distribucijskih semantičkih modela odabran je model latentne semantičke analize, a od modela distribucijske semantičke kompozicije evaluirani su multiplikativni i aditivni modeli. Aditivni modeli s korelacijom od $0,45$ nadmašuju multiplikativni model koji postiže korelaciju od $-0,19$. Rezultati su obećavajući, statistički značajni i u rangu rezultata relevantnih radova.

Ključne riječi: višerječni izrazi, distribucijska semantika, distribucijski semantički modeli, modeli distribucijske semantičke kompozicionalnosti, latentna semantička analiza, semantička kompozicionalnost (prozirnost)

Model for Determining Semantic Compositionality of Croatian Multi-Word Expressions

Abstract

Automatic identification of semantic compositionality of multi-word expression is very important for many tasks in natural language processing e.g. machine translation and information retrieval. In this thesis that issue is addressed using distributional semantic models and distributional models of semantic composition. Dataset consisting of 200 multi-word expressions was annotated with semantic compositionality scores and it was used to evaluate the model. Distributional semantic model was built using Latent Semantic Analysis (LSA). Several models of semantic composition were evaluated. Results show that additive models outperform multiplicative model. Results are promising, statistically significant and comparable to the relevant related work.

Keywords: multi-word expressions, distributional semantics, distributional semantic models, compositional distributional semantics, latent semantic analysis, semantic compositionality (transparency)