UNIVERZITET U NOVOM SADU
FILOZOFSKI FAKULTET
STUDIJSKI PROGRAM: PSIHOLOGIJA

AUTOMATSKO ODREĐIVANJE VRSTA RIJEČI U MORFOLOŠKI SLOŽENOM JEZIKU

Ispitivanje uloge fonotaktičkih informacija u obradi i produkciji infleksione morfologije

DOKTORSKA DISERTACIJA

Mentor: Prof. dr Petar Milin Kandidat: Mr Strahinja Dimitrijević

Novi Sad, 2015. godine

UNIVERZITET U NOVOM SADU
FILOZOFSKI FAKULTET

KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj: RBR	
Identifikacioni broj: IBR	
Tip dokumentacije: TD	Monografska dokumentacija
Tip zapisa: TZ	Tekstualni štampani materijal
Vrsta rada: VR	Doktorska disertacija
Ime i prezime autora: AU	Mr Strahinja Dimitrijević, dipl. psiholog
Mentor / Ko-mentor: MN	Dr Petar Milin, vanredni profesor
Naslov rada: NS	Automatsko određivanje vrsta riječi u morfološki složenom jeziku
Jezik publikacije: JZ	Srpski
Jezik izvoda: JI	Srpski/engleski
Zemlja publikovanja: ZP	Republika Srbija
Uže geografsko područje: UGP	AP Vojvodina
Godina: GO	2015
Izdavač: IZ	Autorski reprint
Mesto i adresa: MA	Novi Sad, dr Zorana Đinđića 2
Fizički opis rada: FO	Poglavlja 5/ stranica 130/ slika 13/ tabela 19/ referenci 295/ priloga 17
Naučna oblast: OB	Psihologija
Naučna disciplina: DI	Kognitivna psihologija, računarska lingvistika
Predmetna odrednica /Ključne reči: PO	Fonotaktičke informacije, infleksiona morfologija, mašine sa vektorima podrške, učenje zasnovano na memoriji, kognitivna vjerodostojnost
UDK	

Čuva se: ČU	Biblioteka Filozofskog fakulteta u Novom Sadu
Važna napomena: VN	Nema
Izvod: IZ	Istraživanje je imalo za cilj da provjeri u kojoj mjeri se naš kognitivni sistem može osloniti na fonotaktičke informacije, tj. moguće/dozvoljene kombinacije fonema/ grafema, u zadacima automatske percepcije i produkcije riječi u jezicima sa bogatom infleksionom morfologijom. Da bi se dobio odgovor na to pitanje, sprovedene su tri studije. U prvoj studiji, uz pomoć mašina sa vektorima podrške (SVM), obavljena je diskriminacija promjenljivih vrsta riječi. U drugoj studiji, produkcija infleksionih oblika riječi izvedena je pomoću učenja zasnovanog na memoriji (MBL). Na osnovu rezultata iz druge studije, izveden je eksperiment u kojem se tražila potvrda kognitivne vjerodostojnosti modela i korišćenih informacija. Diskriminacija promjenljivih vrsta riječi obavljena je na osnovu dozvoljenih sekvenci dva i tri grafema/fonema (tzv. bigrama i trigrama), čije su frekvencije javljanja unutar pojedinačnih gramatičkih tipova izračunate u zavisnosti od njihovog položaja u riječima: na početku, na kraju, unutar riječi, svi zajedno. Maksimalna tačnost se kretala oko 95% i dobijena je na svim bigramima, uz pomoć RBF jezgrene funkcije. Ovako visok procenat tačne diskriminacije ukazuje da postoje karakteristične distribucije bigrama za različite vrste promjenljivih riječi. S druge strane, najmanje informativnim su se pokazali bigrami na kraju i na početku riječi. MBL model iskorišćen je u zadatku automatske infleksione produkcije, tako što je za zadatu riječ, na osnovu fonotaktičkih informacija iz posljednja četiri sloga, generisan traženi infleksioni oblik. Na uzorku od 89024 promjenljivih riječi uzetih iz Frekvencijskog rečnika dnevne štampe srpskog jezika, koristeći metod izostavljanja jednog primjera i konstantu veličinu skupa susjeda (k = 7), ostvarena je tačnost oko 92%. Identifikovano je nekoliko faktora koji su uticali na ovu tačnost, kao što su: vrsta riječi, gramatički tip, način tvorbe i broj

	primjera u okviru jednog gramatičkog tipa, broju izuzetaka, broj fonoloških alternacija itd. U istraživanju na subjektima, u zadatku leksičke odluke, za riječi koje je MBL pogrešno obradio utvrđeno je duže vrijeme obrade. Ovo ukazuje na kognitivnu vjerodostojnost učenja zasnovanog na memoriji. Osim toga, potvrđena je i kognitivna vjerodostojnost fonotaktičkih informacija, ovaj put u zadatku razumijevanja jezika. Sveukupno**,** nalazi dobijeni u ove tri studije govore u prilog teze o značajnoj ulozi fonotaktičkih informacija u percepciji i produkciji morfološki složenih riječi. Rezultati, takođe, ukazuju na potrebu da se ove informacije uzmu u obzir kada se diskutuje pojavljivanje većih jezičkih jedinica i obrazaca.
Datum prihvatanja teme od strane NN Veća: DP	16.06.2009.
Datum odbrane: DO	
Članovi komisije: (Naučni stepen / ime i prezime / zvanje /fakultet) KO	**Predsjednik:** Dr Dušica Filipović Đurđević Vanredni profesor Filozofski fakultet, Novi Sad **Član:** Dr Aleksandar Kostić Redovni profesor Filozofski fakultet, Beograd **Član:** Dr Petar Milin (mentor) Vanredni profesor Filozofski fakultet, Novi Sad

UNIVERSITY OF NOVI SAD
FACULTY OF PHILOSOPHY

KEY WORDS DOCUMENTATION

Accession number: ANO	
Identification number: INO	
Document type: DT	Monograph documentation
Type of record: TR	Textual printed material
Contents code: CC	PhD thesis
Author: AU	M.Sc. Strahinja Dimitrijević, dipl. psiholog
Menthor, co-Menthor: MN	Prof. Petar Milin, PhD
Title: TI	Automatic parts of speech determination in a morphologically complex language
Language of text: LT	Serbian
Language of abstract: LS	Serbian/English
Country of publication: CP	Republic of Serbia
Locality of publication: LP	AP Vojvodina
Publication year: PY	2015
Publisher: PB	Autor's reprint
Publication place: PL	Novi Sad, dr Zorana Đinđića 2
Physical description: PD	Chapters 5/ pages 130/ figures 13/ tables 19 / references 295/ appendices 17
Scientific field: SF	Psychology
Scientific discipline: SD	Cognitive psychology, computational linguistics
Subject / Key words CX	Phonotactic information, inflectional morphology, support vector machines, memory based learning, cognitive plausibility
UC	

Holding data: HD	In the library of the Faculty of Philosophy, Novi Sad
Note: N	No
Abstract: AB	The study was aimed at testing the extent to which our cognitive system can rely on phonotactic information, i.e., possible/ permissible combinations of phonemes/ graphemes, in the tasks of automatic processing and production of words in languages with rich inflectional morphology. In order to obtain the answer to this question, three studies have been conducted. In the first study, by applying the support vector machines (SVM) the discrimination of part of speech (PoS) with more than one possible meaning (i.e., ambiguous PoS) was performed. In the second study, the production of inflected word forms was done with memory based learning (MBL). Based on the results from the second study, a behavioral experiment was conducted as the third study, to test cognitive plausibility of the MBL performance. The discrimination of ambiguous PoS was performed using permissible sequences of two and three characters/sounds (i.e., bigrams and trigrams), whose frequency of occurrence within individual grammatical types was calculated depending on their position in a word: at the beginning, at the end, and irrespective of position in a word. Maximum accuracy achieved was approximatelly 95%. It was obtained when bigrams irrespective of position in a word were used. SVM model used RBF kernel function. Such high accuracy suggests that brigrams' probability distribution is informative about the types of flective words. Interestingly, the least informative were bigrams at the end and at the beginning of words. The MBL model was used in the task of automatic production of inflected forms, utilizingphonotactic information from the last four syllables. In a sample of 89024 flective words, taken from the Frequency dictionary of Serbian language (daily press), achieved accuracy was 92%. For this result the MBL used leave-one-out method and

	nearest neighborhood size of 7 (k = 7). We identified several factors that have contributed to the accuracy; in particular, part of speech, grammatical type, formation method and number of examples within one grammatical type, number of exceptions, the number of phonological alternations, etc. The visual lexical decision experiment revealed that words that the MBL model produced incorrectly also induced elongated reaction time latencies. Thus, we concluded that the MBL model might be cognitively plausibile. In addition, we reconfirmed informativeness of phonotactic information, this time in human conmprehension task. Overall, findings from three undertaken studies are in favor of phonotactic information for both processing and production of morphologically complex words. Results also suggest a necessity of taking into account this information when discussing emergence of larger units and language patterns.
Accepted by the Scientific Board on: ASB	16.06.2009.
Defended on: DE	
Thesis Defend Board: (Degre / name /surname/ title / faculty) DB	President: Dr Dušica Filipović Đurđević Associate professor Faculty of Philosophy, Novi Sad Member: Dr Aleksandar Kostić Professor Faculty of Philosophy, Beograd Member: Dr Petar Milin (mentor) Associate professor Faculty of Philosophy, Novi Sad

SADRŽAJ

SPISAK SLIKA

SPISAK TABELA

PRILOZI

Postoji mnogo dragih ljudi koji su pomogli da započnem i završim ovaj, za mene, veliki i važan posao, zbog čega sam im neizmjerno zahvalan.

Najveću zahvalnost dugujem mentoru, profesoru Petru Milinu, koji mi je nesebično prenosio svoje iskustvo i znanje i vjerovao u mene i kada ja nisam. Privilegija je imati takvog mentora.

Veliku zahvalanost dugujem profesorima, Aleksandru Kostiću, koji mi je otkrio zanimljivi svijet nauke i proučavanja jezika, i Jovanu Saviću, koji već dugo nije među nama, a koji me je usmjerio težim ali interesantnijim i jedino ispravnim putem. Biću srećan ako budem u prilici da utičem na svoje studente, onako kako su moji profesori uticali na mene.

Posebno se zahvaljujem Emanuelu Kulersu (Emmanuel Keuleers) na strpljenju i pomoći oko implementacije *modela zasnovanog na memoriji*. Zahvaljujem se i Siniši Subotiću, koji mi je pomogao da razjasnim neka od statističko-metodoloških pitanja; Fermín Moscoso del Prado Martínu koji je, zajedno sa profesorom Milinom, predložio modifikaciju *proste Gud-Turingove korekcije za učestalosti*; Dušici Filipović Đurđević na korisnim sugestijama, Igoru Simanoviću i Sanji Josifović Elezović na pomoći oko sređivanja teksta, te Martinu Sivelu (Martin Sewell) za ustupljenu sliku koja se odnosi na *minimalizaciju strukturnog rizika*.

Želim da se zahvalim Vasiliju Gvozdenoviću, Nadi Uletilović i Aleksandri Kukoljac na podršci. Zahvaljujem se Ministarstvu nauke i tehnologije Republike Srpske na materijalnoj pomoći u realizaciji ovog istraživanja, kao i članovima laboratorija za eksperimentalnu psihologiju u Novom Sadu i Beogradu, jer sam se uvijek osjećao dobrodošlo i kao član tima.

Hvala sestri i roditeljima na razumijevanju. Žao mi je što im, zbog obaveza, nisam mogao posvetiti onoliko vremena koliko sam želio.

Bez svakodnevne podrške koju sam dobijao od supruge Sonje Stančić, vjerovatno bi ovaj posao ostao nedovršen. Srećan sam što je dio mog života.

AUTOMATSKO ODREĐIVANJE VRSTA RIJEČI U MORFOLOŠKI SLOŽENOM JEZIKU

Ispitivanje uloge fonotaktičkih informacija u obradi i produkciji infleksione morfologije

Sažetak

Istraživanje je imalo za cilj da provjeri u kojoj mjeri se naš kognitivni sistem može osloniti na fonotaktičke informacije, tj. moguće/dozvoljene kombinacije fonema/grafema, u zadacima automatske percepcije i produkcije riječi u jezicima sa bogatom infleksionom morfologijom.

Da bi se dobio odgovor na to pitanje, sprovedene su tri studije. U prvoj studiji, uz pomoć mašina sa vektorima podrške (SVM), obavljena je diskriminacija promjenljivih vrsta riječi. U drugoj studiji, produkcija infleksionih oblika riječi izvedena je pomoću učenja zasnovanog na memoriji (MBL). Na osnovu rezultata iz druge studije, izveden je eksperiment u kojem se tražila potvrda kognitivne vjerodostojnosti modela i korišćenih informacija.

Diskriminacija promjenljivih vrsta riječi obavljena je na osnovu dozvoljenih sekvenci dva i tri grafema/fonema (tzv. bigrama i trigrama), čije su frekvencije javljanja unutar pojedinačnih gramatičkih tipova izračunate u zavisnosti od njihovog položaja u riječima: na početku, na kraju, unutar riječi, svi zajedno. Maksimalna tačnost se kretala oko 95% i dobijena je na svim bigramima, uz pomoć RBF jezgrene funkcije. Ovako visok procenat tačne diskriminacije ukazuje da postoje karakteristične distribucije bigrama za različite vrste promjenljivih riječi. S druge strane, najmanje informativnim su se pokazali bigrami na kraju i na početku riječi.

MBL model iskorišćen je u zadatku automatske infleksione produkcije, tako što je za zadatu riječ, na osnovu fonotaktičkih informacija iz posljednja četiri sloga, generisan traženi infleksioni oblik. Na uzorku od 89024 promjenljivih riječi uzetih iz Frekvencijskog rečnika dnevne štampe srpskog jezika, koristeći metod izostavljanja jednog primjera i konstantu veličinu skupa susjeda (k = 7), ostvarena je tačnost oko 92%. Identifikovano je nekoliko faktora koji su uticali na ovu tačnost, kao što su: vrsta riječi, gramatički tip, način tvorbe i broj primjera u okviru jednog gramatičkog tipa, broj izuzetaka, broj fonoloških alternacija itd.

U istraživanju na subjektima, u zadatku leksičke odluke, za riječi koje je MBL pogrešno obradio utvrđeno je duže vrijeme obrade. Ovo ukazuje na kognitivnu vjerodostojnost učenja zasnovanog na memoriji. Osim toga, potvrđena je i kognitivna vjerodostojnost fonotaktičkih informacija, ovaj put u zadatku razumijevanja jezika.

Sveukupno**,** nalazi dobijeni u ove tri studije govore u prilog teze o značajnoj ulozi fonotaktičkih informacija u percepciji i produkciji morfološki složenih riječi. Rezultati, takođe, ukazuju na potrebu da se ove informacije uzmu u obzir kada se diskutuje pojavljivanje većih jezičkih jedinica i obrazaca.

Ključne riječi: fonotaktičke informacije, infleksiona morfologija, mašine sa vektorima podrške, učenje zasnovano na memoriji, kognitivna vjerodostojnost.

PART OF SPEECH DISAMBIGUATION IN A MORPHOLOGICALLY COMPLEX LANGUAGE

The role of phonotactic information in processing and production of inflectional morphology

Abstract

The study was aimed at testing the extent to which our cognitive system can rely on phonotactic information, i.e., possible/permissible combinations of phonemes/graphemes, in the tasks of automatic processing and production of words in languages with rich inflectional morphology.

In order to obtain the answer to this question, three studies have been conducted. In the first study, by applying the support vector machines (SVM) the discrimination of part of speech (PoS) with more than one possible meaning (i.e., ambiguous PoS) was performed. In the second study, the production of inflected word forms was done with memory based learning (MBL). Based on the results from the second study, a behavioral experiment was conducted as the third study, to test cognitive plausibility of the MBL performance.

The discrimination of ambiguous PoS was performed using permissible sequences of two and three characters/sounds (i.e., bigrams and trigrams), whose frequency of occurrence within individual grammatical types was calculated depending on their position in a word: at the beginning, at the end, and irrespective of position in a word. Maximum accuracy achieved was approximatelly 95%. It was obtained when bigrams irrespective of position in a word were used. SVM model used RBF kernel function. Such high accuracy suggests that brigrams' probability distribution is informative about the types of flective words. Interestingly, the least informative were bigrams at the end and at the beginning of words.

The MBL model was used in the task of automatic production of inflected forms, utilizingphonotactic information from the last four syllables. In a sample of 89024 flective words, taken from the Frequency dictionary of Serbian language (daily press), achieved accuracy was 92%. For this result the MBL used leave-one-out method and nearest neighborhood size of 7 (k = 7). We identified several factors that have contributed to the accuracy; in particular, part of speech, grammatical type, formation method and number of examples within one grammatical type, number of exceptions, the number of phonological alternations, etc.

The visual lexical decision experiment revealed that words that the MBL model produced incorrectly also induced elongated reaction time latencies. Thus, we concluded that the MBL model might be cognitively plausibile. In addition, we reconfirmed informativeness of phonotactic information, this time in human conmprehension task.

Overall, findings from three undertaken studies are in favor of phonotactic information for both processing and production of morphologically complex words. Results also suggest a necessity of taking into account this information when discussing emergence of larger units and language patterns.

Keywords: phonotactic information, inflectional morphology, support vector machines, memory based learning, visual lexical decision.

AUTOMATSKO ODREĐIVANJE VRSTA RIJEČI U MORFOLOŠKI SLOŽENOM JEZIKU

Ispitivanje uloge fonotaktičkih informacija u obradi i produkciji infleksione morfologije

STRAHINJA DIMITRIJEVIĆ

1. UVOD

U svakom jeziku postoje ograničenja koja se odnose na mogućnost da se određene foneme nađu jedna uz drugu. Sekvence (nizovi) fonema koje se javljaju u jeziku, u okviru fonološko prihvatljivih riječi, predstavljaju fonotaktičke informacije (Crystal, 2008). U kojoj mjeri se naš kognitivni sistem može osloniti na takvu vrstu informacija prilikom diskriminacije vrsta riječi i produkcije infleksionih oblika, u jeziku sa bogatom infleksionom morfologijom kakav je srpski, pitanje je na koje se željelo odgovoriti u ovoj studiji.

Informacioni potencijal fonotaktičkih informacija ispitan je na tri načina: (1) statistički, primjenom najmoćnije tehnike klasifikacije – *mašina sa vektorima podrške* (eng. *support vector machines*; Vapnik, 1995, 1998); (2) računarskim modelovanjem zadatka produkcije infleksionih oblika s osloncem na *učenju zasnovanom na memoriji* (eng. *memory based learning – MBL*; Daelemans & Van den Bosch, 2005), u čijoj osnovi leži zaključivanje po analogiji; (3) eksperimentalnom provjerom kognitivne plauzabilnosti, tj. vjerodostojnosti rezultata modelovanja i informacija na koje se MBL oslanjao prilikom izvršenja zadatka produkcije.

Studija predstavlja logičan nastavak istraživanja koja su provedena na srpskom jeziku, u kojima je u zadacima automatske obrade riječi provjeravana informativnost leksičkih i gramatičkih informacija, datih izolovano ili u kontekstu (Dimitrijević, 2011; Dimitrijević, Milin i Kostić, 2008; Ilić i Kostić, 2002; Milin, 2004, 2005; Sečujski i Kupusinac, 2009, itd.). Ona je, takođe, zasnovana na velikom broju psiholingvističkih nalaza koji govore o značaju fonotaktičkih informacija (Albright, 2007; Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Levelt & Wheeldon, 1994; Luce & Pisoni, 1998; Luce & Large, 2001; Pierrehumbert, 2003a, 2003b; Pitt & McQueen, 1998; Vitevitch & Luce, 1998, 1999, 2004; Vitevitch, 2003; Vitevitch, Luce, Pisoni, & Auer, 1999, itd.), kao i o značaju analogije u obradi jezika (Bybee, 2010; Daelemans & Van den Bosch, 2005; Daelemans, Berck, & Gillis, 1997; Eddington, 2002a, 2002b; Hahn & Nakisa, 2000; Keuleers, 2008;

Keuleers & Daelemans, 2007; Keuleers et al., 2007; Krott, Baayen, & Schreuder, 2001; Krott, Schreuder, Baayen, & Dressler, 2007; Milin, Keuleers, & Filipović Đurđević, 2011, itd.).

Istraživani problem pripada domenu kognitivne psihologije odnosno psiholingvistike, ali se značajnim dijelom tiče i *obrade prirodnog jezika* (eng. *natural language processing – NLP*) i *računarske lingvistike* (eng. *computational linguistics*). U okviru NLP-a, automatska obrada riječi je jedan od najčešće ispitivanih problema (Manning & Schuetze, 2000), a kod jezika sa bogatom infleksionom morfologijom obrada riječi predstavlja i najveći izazov i centralni problem.[1]

Ključno pitanje u konstrukciji *tagera* (eng. *tagger*),[2] tj. algoritama (softvera) kojim se automatski pridružuju dodatne informacije jezičkim elementima (fonemama, morfema, riječima ili većim jezičkim cjelinama) u jezičkom korpusu (Baker, Hardie, & McEnery, 2006), predstavlja izbor informacija na koje će se tager oslanjati prilikom izvršavanja tih zadataka (Voutilainen, 1999). Donošenje konačnog zaključka o vrsti informacija na osnovu kojih se najefikasnije razrješava problem automatske obrade riječi nije jednostavno. Pored vrste informacija, postoje i drugi faktori koji utiču na efikasnost primijenjenih modela: broj i "dubina" oznaka koje se pripisuju,[3] način računanja tačnosti, vrsta materijala, jezik koji se obrađuje itd. (Manning & Schuetze, 2000; Tufis, Dienes, Oravecz, & Varadi, 2000). Zbog toga se nalazi o maksimalnoj tačnosti automatske obrade riječi, koja se kreće između 95% i 97% (vidjeti u: Güngör, 2010; Manning, 2011; Manning & Schuetze, 2000, itd), moraju uzeti sa rezervom. U prilog tome govori i istraživanje Džeroskog i sar. (Džeroski, Erjavec, & Zavrel, 2000) koje pokazuje da postoje značajne razlike u tačnosti obrade *nepoznatih riječi* (eng. *unknown words*), tj. riječi koje algoritam nije imao priliku da vidi prilikom učenja, i *poznatih riječi* (eng. *known word*).

[1] Kod jezika koji nemaju bogatu infleksionu morfologiju, kao što je npr. engleski jezik, centralni problem je obrada rečenice, tj. sintakse (Manning & Schuetze, 2000).
[2] Automatsko pridruživanje lingvistički relevantnih informacija naziva se *tagiranje* (eng. *tagging*; Baker, Hardie, & McEnery, 2006).
[3] Neki autori u analize uključuju i tzv. "plitke" oznake, kao što su znaci interpunkcije, što značajno povećava efikasnost tagera.

Oslanjajući se na dvije oznake koje prethode meti, tj. koristeći Markovljeve modele drugog reda, jednog od najčešće upotrebljavanih pristupa u automatskoj obradi riječi (Charniak, 1993; Church, 1988; DeRose, 1988; Güngör, 2010; Manning & Schütze, 2000, itd), dostignuta je tačnost od oko 55% za nove i nepoznate tekstove, dok se tačnost na poznatim riječima kretala između 92 i 95% (Džeroski et al., 2000). Sličan rezultat je dobijen i pri obradi *višeznačnih riječi* (eng. *lexical ambiguity*), tj. riječi kojima može biti pripisano više od jedne oznake, na osnovu informacija o zajedničkoj vjerovatnoći javljanja dvije odnosno tri vrste riječi. U ovim slučajevima tačnost se kretala oko 57% (Dimitrijević i sar., 2008).

Postoji veliki broj različitih informacija koje se mogu koristiti u zadacima automatske obrade riječi, ali do sada nije dat njihov iscrpan i sistematičan pregled. Razlozi za to su dvojaki. S jedne strane, one se u oblasti NLP-a ne koriste na principijelan način (Milin, 2004), jer postoji tendencija uključivanja većeg broja različitih informacija u nadi da će se tako ostvariti bolji rezultat. S druge strane, cilj istraživanja koja tretiraju problem automatske obrade riječi je, uobičajeno, praktični rezultat – efikasnost tagera, a ne principijelni – razmatranje zadatka, mogućih funkcija i/ili mehanizama obrade, vrsta informacija i sl. U ovoj studiji je data drugačija perspektiva – fokus je na jezičkim informacijama, a navedeni tageri su u funkciji ilustracije efikasnosti obrade riječi koja se zasniva na tim informacijama.

1.1. Strategije u zadacima automatske obrade riječi

Maning i Šice (Manning & Schuetze, 2000) razlikuju dva izvora informacija koje sistemi za obradu riječi tipično koriste: (a) informacije dobijene na nivou pojedinačnih riječi i (b) informacije dobijene na osnovu konteksta u kojem se riječi realizuju. Milin (2004), s druge strane, u svojoj sistematizaciji polazi od dvije dihotomije: (1) informacije o tipovima (gramatičke) – informacije o riječima (leksičke); (2) diskretne informacije (pravila) – probabilističke informacije (na osnovu vjerovatnoća javljanja). Kombinovanjem ovih tipova informacija dobijaju se

četiri moguće strategije automatske obrade riječi: (a) leksičko-diskretna, (b) gramatičko-diskretna, (c) gramatičko-probabilistička i (d) leksičko-probabilistička strategija (Milin, 2004).[4] Ove strategije, odnosno tipovi informacija, mogu biti lokalne ili kontekstualne, kako to razlikuju i Maning i Šice (Manning & Schuetze, 2000). Pored toga, prikupljanje gramatičkih kontekstualnih informacija se može obaviti na dva načina: (a) s osloncem na gramatiku (eng. *top-down* ili *grammar-driven*) ili (b) s osloncem na tekst (eng. *bottom-up* ili *data-driven*; Milin, 2004). Uzevši u obzir ove različite načine generisanja informacija, klasifikacija strategija koja se zasniva na dvije dihotomije (Milin, 2004) postaje znatno složenija.

Leksičko-diskretna strategija: Leksičko-diskretna strategija podrazumijeva pronalaženje zadatog oblika riječi u elektronskom rječniku, nakon čega mu se pripisuju moguće odrednice i sve odgovarajuće oznake. Najveći izazov za primjenu ovog pristupa je višeznačnost riječi, jer takvim riječima pripisuje više od jednog atributa. U većini jezika višeznačne riječi su prisutne u tolikom broju da ovaj pristup čine praktično neupotrebljivim. Iz tog razloga se, na primjer, pomoću leksičko-diskretne strategije tačno lematizuje tek 55% proznog teksta na srpskom jeziku (Ilić i Kostić, 2002).

Gramatičko-diskretna strategija: Gramatičko-diskretna strategija zasniva se na primjeni diskretnih fonoloških, morfoloških ili sintaksičkih pravila, posebno ili u kombinaciji. Ovu strategiju karakteriše balansiranje između uspješnosti u razrješenju leksičkih nejasnoća i broja pravila. Povećanjem broja pravila povećava se efikasnost, ali raste i rizik od konflikata između tih pravila. Taj konflikt se može riješiti: intervencijom jezičkog stručnjaka, gradacijom postojećih pravila na osnovu njihovih vjerovatnoća ili uvođenjem novih diskretnih pravila. Jedan od najefikasnijih tagera, *EngCG tager* (eng. *Constraint Grammar Parser*), polazeći od gramatičkih pravila, tačno pripisuje 99,7% oznaka riječima koje se obrađuju (Voutilainen, Heikkilä, & Anttila, 1992). Međutim, za 3% do 7% riječi, on vraća

[4] Između ovih pristupa ponekad se ne mogu povući oštre granice. Takođe, oni nisu isključivi, pa se u praktičnoj primjeni često kombinuju, s ciljem dobijanja što veće tačnosti. Zbog toga na ove podjele treba gledati kao na pomoćno sredstvo, koje olakšava sistematizaciju do sada prikupljenog znanja.

više od jedne oznake, što znači da u tim slučajevima problem višeznačnosti nije riješen.[5]

Primjenom jednog takvog ekspertskog sistema, koji se oslanja na direktnu implementaciju gramatičkih pravila srpskog jezika, postiže se tačnost morfološke anotacije od 93.25% (Sečujski i Kupusinac, 2009).[6]

Lokalna leksičko-probabilistička strategija: Problem višeznačnosti se u okviru lokalnog leksičko-probabilističkog pristupa rješava tako što se odredi vjerovatnoća javljanja svih oblika višeznačne riječi, a kao tačno rješenje uzme najučestaliji (najvjerovatniji) oblik.[7] Oslanjajući se samo na informacije o vjerovatnoćama odrednica, tačno se lematizuje oko 95% teksta na srpskom jeziku (Ilić i Kostić, 2002; Milin 2004). Ako se uzme u obzir i oko 6% *neotvorenih* ili *neobrađenih* riječi (Dimitrijević, 2007; Milin, 2004), tj. riječi koje nisu pronađene u rječniku iz kojeg se preuzimaju odgovarajuće vjerovatnoće, onda se uspješnost ovog pristupa kreće između 88% i 90%, što je približno tačnosti koja je dobijena na višeznačnim riječima (Dimitrijević, 2007). Najčešće greške koje se javljaju su greške na prilozima i rječcama, tako što se prilozi zamjenjuju pridjevima, zamjenicama i veznicima, a rječce – veznicima, glagolima i prilozima (Dimitrijević, 2007; Milin, 2004).

I u slučaju detaljnijeg morfološkog anotiranja, koje slijedi nakon lematizacije, ovaj pristup daje statistički bolji rezultat nego kada se koristi lokalni

[5] Brilov transformacioni algoritam (Brill, 1992, 1994, 1995) se takođe oslanja na diskretna pravila, s tim da ta pravila uči koristeći unaprijed definisane obrasce. Prvi korak podrazumijeva pridruživanje inicijalnih oznaka svim riječima (npr. pripisivanje najvjerovatnije oznake iz frekvencijskog rječnika). U drugom koraku algoritam uči transformaciona pravila, koja, na primjer, mogu biti u obliku: "Ako je riječi pripisana oznaka *a*, a nalazi se u kontekstu *C*, promijeniti oznaku *a* u *b*" (Brill, 1992). Prilikom učenja pravila algoritam se, pored konteksta, može oslanjati i na leksička svojstva riječi koja se obrađuju ili svojstva riječi iz okruženja. U fazi anotiranja se koriste pravila kod kojih je razlika između broja tačnih i broja pogrešnih korekcija dovoljno velika da bi se mogla smatrati pouzdanom.

[6] Ovaj pristup se pokazao efikasnijim od *Markovljevih modela* (eng. *Markov models*) i morfološke anotacije zasnovane na primjeni transformacionih pravila (Sečujski i Kupusinac, 2009). O ovim pristupima biće više riječi u dijelu teksta koji slijedi.

[7] U grupu *tupavih* (eng. *dummy*) *tagera* mogli bi se svrstati pristupi koji se oslanjaju isključivo na vjerovatnoće jezičkih elemenata, bilo da je riječ o leksičkim ili vjerovatnoćama na nivou gramatičkih tipova. Zbog svoje jednostavnosti, prvi ovakav tager, uzimajući najvjerovatniju oznaku za datu riječ u tekstu kao tačnu, postavlja donju granicu efikasnosti tagera od 90% tačno obrađenih riječi (Cherniak, 1997; Cherniak, Hendrickson, Jacobson, & Perkowitz, 1993).

gramatičko-probabilistički pristup, koji se oslanja na nejednake vjerovatnoće morfoloških tipova (Milin, 2004).

Kontekstualna leksičko-probabilistička strategija: Kao najefikasniji pristup automatskoj obradi riječi pokazao se kontekstualni leksičko-probabilistički pristup, koji se oslanja na zavisne vjerovatnoće pojedinačnih riječi i vrsta riječi koje ih okružuju. Njegovom primjenom se postiže uspjeh od 95.81% tačno obrađenih riječi, što je statistički značajno bolji rezultat od onog koji se dobija lokalnim leksičko-probabilističkim pristupom (Milin, 2005). Može se pretpostaviti da bi se proširenjem ovog pristupa zavisnim leksičkim vjerovatnoćama ili, eventualno, detaljnijom specifikacijom gramatičkog konteksta (specifikacijom na nivou gramatičkih oblika) postigla (naj)veća tačnost u obradi riječi. Međutim, sistematska provjera ove ideje do sada nije napravljena, jer ona podrazumijeva stabilne zavisne vjerovatnoće na nivou oblika riječi sa detaljnom obradom, za šta ne postoje dostupni jezički korpusi odgovarajuće veličine. Stabilne zavisne vjerovatnoće su preduslov za uspješno obavljanje ovakvih zadataka jer se njihova distribucija ne razlikuje od distribucije u jeziku, što za posljedicu ima dobijanje pouzdanih rezultata. Kako se radi o veoma finoj rezoluciji vjerovatnoća, koje bi bile generisane na nivou parova riječi ili pojedinačnih riječi i detaljnog gramatičkog konteksta, neophodni su veliki jezički uzorci.

Kontekstualna gramatičko-probabilistička strategija s osloncem na gramatiku: Za razliku od kontekstualnog leksičko-probabilističkog pristupa, kontekstualna gramatičko-probabilistička strategija koristi procijenjene vjerovatnoće slaganja dvaju ili više morfoloških tipova, tzv. *sintagmatske strukturalne informacije* (eng. *syntagmatic structural information*; Manning & Schuetze, 2000). Ove informacije se mogu prikupiti na dva načina. Jedan od načina je da se utvrde vjerovatnoće javljanja prihvaćenih i/ili alternativnih gramatičkih pravila u jezičkom uzorku, kako bi se formirala njihova rang lista. Greška u zadatku morfološke anotacije riječi srpskog jezika, koja se dobija primjenom ovako dobijenih pravila, kreće se oko 10% (Sečujski i Kupusinac, 2009). Međutim, u zadatku lematizacije, kontekstualni gramatičko-probabilistički pristup baziran na gramatici nije ništa

efikasniji od pristupa koji se isključivo oslanja na vjerovatnoće oblika riječi (Milin, 2005).

Kontekstualna gramatičko-probabilistička strategija bazirana na tekstu: Drugi način prikupljanja gramatičkih kontekstualnih informacija je s osloncem na tekst. Ovaj pristup podrazumijeva računanje zavisnih vjerovatnoća na nivou morfoloških (gramatičkih) tipova. Najčešće se koriste informacije o gramatičkom statusu jedne ili dvije prethodne riječi (Brants, 2000; Cherniak, Hendrickson, Jacobson, & Perkowitz, 1993; Church 1988; Kupiec, 1992, itd.), dok su modeli, koji se na njih oslanjaju, poznati kao *n-gram* modeli ili *Markovljevi lanci* (Manning & Schuetze, 2000). Mada se navodi da se njihova tačnost, pod povoljnim uslovima, kreće između 95% i 97% anotiranih riječi (Güngör, 2010; Manning, 2011; Manning & Schuetze, 2000), vidjeli smo da postoje istraživanja koja govore suprotno. Efikasnost postignuta u zadatku lematizacije višeznačnih riječi srpskog jezika kreće se oko 57% tačno obrađenih riječi (Dimitrijević, 2007; Dimitrijević i sar., 2008), što odgovara rezultatima koji su dobijeni prilikom morfosintaksičke obrade nepoznatog teksta na slovenačkom jeziku pomoću trigram tagera, koji koristi skrivene Markovljeve modele drugog reda (eng. *hidden Markov models*), tj. informacije o dvije riječi koje prethode meti (Džeroski et al., 2000).[8] Pri korišćenju kontekstualnih informacija na nivou vrsta riječi u lematizaciji, nešto bolji rezultat u odnosu na modele prvog i drugog reda (oko 63% tačno obrađenih riječi) dobija se kada se riječ-meta nalazi u sredini niza, tj. kada algoritam raspolaže informacijom o vrsti riječi koja prethodi i vrsti riječi koja slijedi riječ-metu (Dimitrijević, 2007; Dimitrijević i sar., 2008). Polazeći od ovog rezultata, aproksimirana gornja granica uspješnosti kontekstualnog gramatičko-probabilističkog pristupa na kontinuranom tekstu kreće

[8] Kod skrivenih Markovljevih modela (eng. *hidden Markov models – HMM*) poznata je samo sekvenca riječi, dok su kod (*vidljivih*) *Markovljevih modela* (eng. *visible Markov models – VMM* ili *Markov models – MM*) poznate i sekvenca riječi i sekvenca oznaka. Rezultati istraživanja (Merialdo, 1994; Elworthy, 1994) ukazuju da su i manji anotirani korpusi bolji izvor za generisanje vjerovatnoća, nego veća količina neobrađenog teksta, zbog čega upotreba HMM modela ima opravdanja samo ako se ne raspolaže anotiranim korpusom (Manning & Schuetze, 2000).

se oko 78% (Dimitrijević, 2007).[9] U jednom novijem istraživanju, sa drugačijim ciljem i načinom ocjenjivanja greške, ali s osloncem na skrivene Markovljeve modele, Sečujski i Kupusinac (2009) postižu tačnost koja je za približno 11% viša od tačnosti o kojoj je izvijestio Dimitrijević (2007). Razlog ove neusaglašenosti treba tražiti, između ostalog, i u strukturi materijala: Dimitrijević koristi slučajni uzorak višeznačnih riječi iz književnog djela, dok Sečujski i Kupusinac koriste kontinuirani tekst sastavljen prvenstveno iz novinskih članaka. Konačno, dok je cilj istraživanja Dimitrijevića ocjenjivanje gornje, teorijske granice greške, zbog čega i koristi samo višeznačne riječi, Sečujski i Kupusinac predstavljaju jedan sistem i njegove praktične domete.

U cilju postizanja što veće tačnosti u zadacima automatske obrade jezika veoma često se kombinuju različite vrste informacija. Pri tome, izbor informacija se ne bazira na kriterijumima koji se oslanjaju na rezultate fundamentalnih istraživanja. To je, na primjer, bio slučaj i u zadatku automatskog pripisivanja morfosintaksičkih oznaka (eng. *part of speech tagging*) uz pomoć mašina sa vektorima podrške (Nakagawa, Kudo, & Matsumoto, 2001). Mašine sa vektorima podrške (Vapnik, 1995, 1998) trenutno su jedan od najboljih dostupnih klasifikatora (Baayen, 2011; Joachims, 1998; Meyer, Leisch, & Hornik, 2003; Steinwart & Christmann, 2008, Van Gestel et al., 2004), koji postiže tačnost veću od 97% (Giménez, & Màrquez, 2003, 2004; Nakagawa et al., 2001). SVM predstavljaju realizaciju ideje da nelinearno odvojivi podaci postaju linarno odvojivi u prostoru sa više dimezija, tj. da postoji definisana optimalna hiperravan koja ih maksimalno diskriminiše (Cortes & Vapnik, 1995).[10] Mašine sa vektorima podrške imaju dvije prednosti u odnosu na ostale modele koji se koriste u zadacima automatske obradi riječi; s jedne strane, veliki broj varijabli uključenih u analizu ne predstavlja im izazov, dok su, s druge strane, otpornije na problem *preučavanja* (eng. *overfitting*; Mayfield, McNamee,

[9] Aproksimacija procenta tačno obrađenog kontinuiranog teksta računata je na sljedeći način: (broj jednoznačnih riječi + broj višeznačnih riječi x proporcija tačnosti sistema za lematizaciju) / (broj jednoznačnih riječi + broj višeznačnih riječi).

[10] O ovome će biti više riječi u dijelu koji se bavi problemom diskriminacije vrsta riječi na osnovu fonotaktičkih informacija.

Piatko, & Pearce, 2003; Nakagawa et al., 2001). U pomenutom zadatku, tager se oslanjao na veći broj različitih informacija: POS kontekst (dvije oznake ispred i dvije oznake iza *riječi-mete*), lokalni leksički kontekst (dvije riječi ispred i dvije iza) i informacije do četiri karaktera na početku i na kraju riječi-mete, kao i informacije o postojanju brojeva, velikih slova, crtice i sl. u riječima koje se obrađuju. Ovaj tager je, koristeći polinomialnu jezgrenu funkciju,[11] postigao tačnost od 97.1%, što je za oko pola procenta više od procenta tačno obrađenih riječi pomoću *TnT tagera* (Brants, 2000), koji se bazira na Markovljevim modelima drugog reda (Nakagawa et al., 2001). Kao naročito važne informacije pokazale su se informacije o karakterima na početku i kraju riječi, velikim slovima, brojevima i crticama u riječi-meti. Kada se ove informacije isključe, efikasnost tagera drastično pada, pa za nepoznate riječi, na *uzorku za uvježbavanje* veličine milion riječi, iznosi 30%, a za poznate riječi oko 75%.[12]

Pored spomenutih pristupa, u upotrebi su i brojni drugi modeli: *model maksimalne entropije* (eng. *maximum entropy model*; Ratnaparkhi, 1996), *neuralne mreže* (eng. *neural networks*, npr. *Net-Tager*; Schmid, 1994a); stabla odlučivanja (eng. *decision tree*; Schmid, 1994b); *genetički algoritmi* (eng. *genetic algorithms*; Araujo, 2002); *dinamičke Bajesove mreže* (eng. *dynamic Bayesian networks*; Peshkin & Savova, 2003; Peshkin, Pfeffer, & Savova, 2003), *logičko programiranje* (eng. *logical programming*; Cussens, 1998); *algoritmi relaksiranog indeksiranja* (eng. *relaxation labeling*; Padro, 1996); *latentno semantičko indeksiranje* (eng. *latent semantic mapping*; Bellegarda, 2008); *robusna minimalizacija rizika* (eng.

[11] Polinomialne funkcije su funkcije predstavljene polinomima, na primjer: $f(x) = x^3+x$ ili $f(x) = x^2+2xy-3y^2+3$. Polinom je izraz koji povezuje nekoliko monoma (cijelih izraza u kojima je množenje jedina dozvoljena operacija, npr. xy, $0.25x^2$, itd.) operacijama sabiranja i množenja. U polinomima su dozvoljene i operacije stepenovanja i oduzimanja, koje se mogu posmatrati kao specijalni slučajevi prethodne dvije operacije, npr. $x \cdot x = x^2$, odnosno, $2 + (-1)x = 2-x$.

[12] Oslanjajući se na mašine sa vektorima podrške, razvijen je i *SVMtool*, poseban alat za morfosintaksičko tagiranje, za kojeg autori kažu da predstavlja jednostavan, fleksibilan i efiksan tager (Giménez & Màrquez, 2004). Koristeći ovaj alat dobijena je prosječna tačnost od 97.16% za engleski i 96.89% za španski jezik (Giménez & Màrquez, 2003, 2004). Nešto slabiji rezultat dobijen je za telugu jezik, kojim se govori u Indiji i kojeg, kao i srpski jezik, karakteriše da se piše onako kako se čita. Postignuta tačnost na uzorku za obučavanje veličine 20000 i test-uzorku od 5000 riječi u pripisivanju deset oznaka iznosila je 95% (Sindhiya, Anand, & Soman, 2009).

robust risk minimization; Ando, 2004); *uslovna slučajna polja* (eng. *conditional random fields*; Lafferty, McCallum, & Pereira, 2001), pristupi zasnovani na *teoriji rasplinutih skupova* (eng. *fuzzy set theory*; Kim & Kim, 1996) itd.

Iako navedene pristupe karakteriše približno ista efikasnost, oni produkuju različite greške, zbog čega su u upotrebi i hibridni tageri. Kod hibridnih tagera, izlaz (eng. *output*) jednog algoritma može da se koristi kao ulaz (eng. *input*) u drugi (npr. da se prije transformacionih pravila u inicijalnoj anotaciji koriste oznake dobijene uz pomoć Markovljevih modela) ili se, nakon pripisivanja oznaka od strane više tagera, krajnje rješenje bira nekom od *glasačkih strategija* (eng. *voting strategy*; npr. u slučaju pripisivanja različitih oznaka od strane više tagera, uzima se ona oznaka koja je pripisana od tagera koji se smatra najefikasnijim; Güngör, 2010).

Opšta primjedba koja se može uputiti većini sistema za obradu riječi jeste da se ne oslanjaju na rezultate psiholingvističkih istraživanja i zanemaruju pitanje kognitivne plauzabilnosti informacija i korišćenih modela, već su usmjereni na ostvarivanje praktičnih ciljeva i koristi (Milin, 2004). Najznačajnije dileme u obasti računarske lingvistike i obrade prirodnog jezika, kao što je, na primjer, debata o *obradi oslonjenoj na pravila* nasuprot *obradi oslonjenoj na vjerovatnoće* (Atwell, 1987; Jurafsky, 2003; Jurafsky & Martin, 2000; Manning & Schuetze, 2000, itd.), izlaze iz domena ovih oblasti i tiču se kognitivnih osnova i načina obrade prirodnog jezika, generalno (npr. Albright & Hayes, 2003; Baayen, Feldman, & Schreuder, 2006; Hay & Baayen 2005; Keuleers et al., 2007; Milin, Filipović Đurđević, & Moscoso del Prado Martín, 2009; Tremblay & Baayen, 2010, itd.). To je slučaj i sa pitanjem da li se u predikciji jezičkog ponašanja treba oslanjati na pravila (Albright & Hayes, 2003) ili na analogiju (Daelemans & Van den Bosch, 2005; Hare, Elman, & Daugherty 1995; Rumelhart & McClelland 1986; Skousen, 1989, 1992, itd.).

Ako bi se sistemi za obradu riječi oslanjali na nalaze fundamentalnih istraživanja jezika, oni bi trebalo da uključe princip zaključivanja po analogiji, tj. zaključivanje na osnovu primjera koji su rezultat prethodnog iskustva. Na ovo ukazuju brojna istraživanja koja su pokazala značaj analogije u obradi riječi (više u: Bybee, 2010), ali i zaključak da su modeli obrade jezika zasnovani na memoriji

(Daelemans & Van den Bosch, 2005; Daelemans, Zavrel, Van der Sloot, & Van den Bosch, 2010) neizostavan faktor u opisu obrade morfološki složenih riječi (Milin, Kuperman, Kostić, & Baayen, 2009). Ovi autori smatraju da je, ako se želi dobiti adekvatan opis obrade morfološki složenih riječi, neophodno kombinovati pristup morfologiji zasnovan na modelu *riječ-paradigma* (eng. *word-paradigm morphology*; više u Blevins, 2006, 2013) i informaciono-teorijski pristup (Kostić, 1991, 1995; Kostić, Marković, & Baucal, 2003; Moscoso del Prado Martin, Kostić, & Baayen, 2004) sa modelom obrade jezika zasnovanim na memoriji.

Prema modelu riječ-paradigma, svaka riječ, tj. svi oblici jedne lekseme uskladišteni su u mentalnom leksionu. Pri tome, riječi se "grupišu" u okviru *infleksionih paradigmi* (skup svih oblika jedne lekseme, npr. svi oblici lekseme – *kuća*), a paradigme u okviru *infleksionih klasa* (skup leksema koje dijele isti set nastavaka u njihovim paradigmama, npr. sve imenice ženskog roda koje se u nominativu jednine završavaju na *–a*). S druge strane, informaciona teorija predstavlja dobar alat za opis ovakvog pristupa morfologiji (Milin et al., 2009).

1.2. Analogija i obrada jezika

Indukcija i analogija kao principi/mehanizmi usvajanja i obrade jezika prisutni su u modernoj lingvistici od njenih početka, tj. od početka XX vijeka (Bloomfield, 1933; De Saussure, 1916/1966). Analogija se definiše kao upotreba novih riječi u postojećim obrascima (eng. *existing pattern*), utemeljenim isključivo na uskladištenim primjerima (Baayen, 2003; Boas, 2003; Bybee, 2010; Bybee & Eddington, 2006; Eddington, 2000; Krott et al., 2001; Skousen, 1989, 1992, itd.). U najvećem broju slučajeva, analogije u jeziku zasnivaju se na semantičkoj i fonološkoj sličnosti sa postojećim oblicima u memoriji (Bybee, 2010).

Analogija ima važnu ulogu i u mnogim modelima obrade riječi. Na primjer, mnogi modeli zasnovani na dva principa, tzv. *pristupi dvostrukog puta* (eng. *dual-route approaches*) oslanjaju se na gramatička pravila u obradi pravilnih oblika riječi,

dok se obrada nepravilnih oblika obavlja na osnovu sličnosti sa postojećim formama u memoriji (Clahsen, 1999; Marcus, Brinkmann, Clahsen, Wiese & Pinker 1995; Pinker, 1991, 1999; Pinker & Prince, 1988, 1994; Prasada & Pinker, 1993). Tako se, npr. prošlo vrijeme pravilnih oblika glagola u engleskom jeziku formira na osnovu pravila, dodavanjem nastavka *–ed*: *walk – walked*. S druge strane, formiranje prošlog vremena nepravilnih glagola oslanjaće se na sličnost sa primjerima u leksičkoj memoriji, zbog čega oblici nepravilnih glagola *sting – stung, swing – swung, cling – clung* itd. mogu poslužiti kao osnova za formiranje prošlog vremena za fonološki sličnu pseudoriječ *spling*, za koju će, po analogiji, biti generisano prošlo vrijeme *splung* (Albright & Hayes, 2003; Bybee & Moder, 1983; Prasada & Pinker, 1993).

Skousen (2002a), međutim, smatra da i svi oni modeli koji uopšte ne koriste pravila nove forme grade po analogiji sa sličnim primjerima u memoriji (Daelemans & Van den Bosch, 2005; Hare et al., 1995; Rumelhart & McClelland, 1986; Skousen, 1989, 1992; itd.). Tako, ovaj autor razlikuje dva tipa modela: *neuralne mreže* i *modele zasnovane na egzemplarima*.[13]

Neuralne mreže polaze od primjera (egzemplara), koji se koriste za učenje i sistematizaciju karakteristika, koje se distribuiraju kroz čitavu mrežu. Konekcionistički modeli paralelno-distribuiranog procesiranja (eng. *parallel-distributed connectionist models*), koji svi implementiraju neuralne mreže, koriste kompeticiju između ortografski i/ili fonološki sličnih riječi, tj. efekat susjedstva (eng. *neighborhood effect*), kao značajan izvor informacija za učenje i klasifikaciju (više o tome u: Plaut & Gonnerman, 2000; Seidenberg & Gonnerman, 2000). Slično tome, model *naivnog diskriminatornog učenja* (eng. *naive discrimination learning – NDL*; Baayen, Milin, Filipović Đurđević, Hendrix, & Marelli, 2011) takođe naglašava uticaj susjedstva u fazi učenja sistematskih odnosa između ortografskog ili fonološkog ulaza (tipično, nizova od dva ili tri karaktera/glasa) i leksičkog izlaza (koji nije mentalna predstava, već sistem simboličkih kontrasta koji kombinuje

[13] Egzemplar: primjerak, uzorak, pojedinačni otisak, jedan jedini primjerak u zbirci (Vujaklija, 1970).

semantičke, fonološke, ortografske pa čak i gestualne jezičke elemente, onako kako je to definisao Arenof (Aronoff, 1994).[14]

Nasuprot neuralnih mreža, kod modela zasnovanih na egzemplarima Skousen (2002a) pravi razliku između *analoškog modeliranja* (eng. *analogical modeling*; Skousen, 1989, 1992, 2002b) i pristupa koji se oslanjaju na *najbliže susjede* (eng. *nearest neighbor*), kao što je učenje zasnovano na memoriji (Daelemans & Van den Bosch, 2005). Uprkos sličnostima, između ova dva pristupa postoje teorijske i implementacijske (algoritamske) razlike, što rezultira i razlikama u uspješnosti izvršavanja specifičnih zadataka.

Analoško modeliranje je opšta teorija predikcije ponašanja koja se može primijeniti i na jezik (Skousen, 2002b, 2009). Ovaj pristup pripada klasi modela koji se zasnivaju na primjerima (eng. *instance-based* ili *exemplar-based* model). Za pomenutu klasu modela karakteristčno je da ne postoji razlika između pravilnih i nepravilnih oblika jezičkih fenomena. Kao što je već rečeno, analoško modeliranje koristi informacije o egzemplarima koji ne pripadaju skupu najbližih susjeda i po tome se razlikuje od učenja zasnovanog na memoriji (uporediti: Daelemans, 2002; Skousen, 2002b). Vjerovatnoća da neki egzamplar bude uzet kao model zavisi od više faktora: sličnosti egzemplara sa zadatim kontekstom (eng. *proximity*);[15] broja egzemplara koji ga okružuju, a imaju slične karakteristike (eng. *gang effect*); postojanje egzemplara koji se ponašaju na drugačiji način, a sličniji su zadatom kontekstu (eng. *heterogenity*; Skousen, 1995).

Metod najbližih susjeda (eng. *nearest neighbor classifiers methods*) je razvijen radi učenja klasifikacionih pravila u oblasti prepoznavanja složaja (eng. *pattern recognition*; Cover & Hart, 1967; Fix & Hodges, 1951). Novi objekat klasifikuje se tako što se: (a) uporedi sa objektima i njihovim klasama koje su uskladištene u memoriji, (b) nađu udaljenosti između nepoznatog objekta i svih

[14] Ovi kontrasti omogućavaju, po Arenofu (Aronoff, 1994), tok komunikacije, odnosno, nužne i dovoljne znake koji, dok razmjenjujemo informacije – komuniciramo, razlikuju npr. stolicu od drugih entiteta i/ili događaja.

[15] Tako, npr. ako se želi predvidjeti izgovor slova *c* u riječi *ceiling* na osnovu prva tri slova koja ga slijede (*e*, *i*, *l*), ova slova predstavljaju zadati kontekst (Skousen, 2002).

objekata u memoriji, (c) pronađe objekat ili objekti koji su najbliži objektu koji se klasifikuje (*najbliži susjed/i*; eng. *nearest neighbor* – NN),[16] (d) klasifikuje objekat u klasu kojoj pripada najbliži susjed/i. Odluka o pripadnosti nekog objekta određenoj klasi može se donijeti na osnovu klasne pripadnosti najbližeg susjeda (*1-NN* klasifikator) ili većeg broja najbližih susjeda (*k-NN* klasifikator), kada se novi objekat svrstava u najfrekventniju klasu. Metod najbližih susjeda, sa različitim algoritamskim rješenjima za klasifikaciju, danas ima široku primjenu u oblasti vještačke inteligencije. On pripada grupi *lijenih* (eng. *lazy*) metoda, jer pamti naučene primjere i odlaže generalizaciju sve do momenta kada treba da se klasifikuje nova instanca, zbog čega je i vrijeme klasifikacije duže. Metod najbližih susjeda poslužio je kao osnova za razvoj modela učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005).

Učenje zasnovano na memoriji može se smatrati operacionalizacijom pristupa koji se oslanjaju na analogiju. Riječ je o modelu usvajanja i obrade jezika. MBL polazi od pretpostavke da se pri izvršenju kognitivnih zadatka ne koriste mentalna pravila ili druge apstraktne reprezentacije (eng. *representations*) izvedene iz iskustva, već da se o novoj situaciji zaključuje direktno na osnovu njene sličnosti sa događajima iz prošlosti (više u: Daelemans & Van den Bosch, 2005; Daelemans et al., 2010).[17] MBL se pokazao uspješnim u različitim zadacima automatske obrade jezika, od morfo-fonoloških analiza i klasifikacija, do analize teksta i diskursa. Prvi put su ga primijenili Dalemans i saradnici (Daelemans, Zavrel, Berck, & Gills, 1996) u zadatku automatskog pripisivanja sintaksičkih kategorija. Primjeri koje je algoritam koristio prilikom tagiranja uzeti su iz ručno anotiranog korpusa, pri čemu se svaki primjer sastojao od riječi i konteksta u kom se ta riječ nalazi te odgovarajuće oznake. Uspješno je obrađeno oko 96% riječi (Daelemans et al.,

[16] Mjera udaljenosti među objektima može biti, npr. Euklidska ili Mahalanobisova distanca itd.

[17] Nazivi koji se još koriste u označavanju ovog prostupa su: *similarity-based*, *example-based*, *exemplar-based*; *analogical*, *case-based*, *instance-based learning/approach* i *lazy learning* (Daelemans et al., 2010). Iako se često koriste kao sinonimi, između nekih od ovih pojmova, ipak, postoje određene razlike. Tako je, npr., *lazy learning* (*lijeno učenje*) širi pojam od učenja zasnovanog na memoriji i označava klasu metoda vještačke inteligencije koje dijele neka zajednička svojstva, zbog čega ga neki autori i koriste kao nadređeni pojam gore navedenim pojmovima (Aha, 1997).

1996).[18] U najvećem broju slučajeva upotreba MBL se vezuje za domen primijenjenih nauka (inženjerstvo), mada je bio primjenjivan i u lingvističkim i psiholingvističkim istraživanjima (više u: Daelemans et al., 2010).

Automatska obrada teksta s osloncem na učenju zasnovanom na memoriji testirana je i na fenomenima infleksije, npr. u građenju prošlog vremena u engleskom jeziku (Keuleers, 2008), građenju množine u holandskom (Keuleers & Daelemans, 2007; Keuleers et al., 2007) i njemačkom (Hahn & Nakisa, 2000), markiranju roda u španskom (Eddington, 2002a), deminutiva u španskom (Eddington, 2002b) i holandskom (Daelemans et al., 1997), u predikciji infiksa u složenicama holandskog (Krott et al., 2001) i njemačkog jezika (Krott et al., 2007) itd. Konačno, generisanje alomorfnih varijanti instrumentala jednine imenica muškog roda u srpskom jeziku, takođe je ispitivano primjenom MBL-a (Milin et al., 2011).

S obzirom na to da ovaj pristup ne pravi razliku prilikom skladištenja pravilnih i nepravilnih oblika, MBL pripada klasi modela jednog puta, koji pretpostavljaju jedan, jedinstven mehanizam produkcije infleksionih oblika. Ovaj mehanizam se, u prvom redu, oslanja na fonološku sličnost (vidi i: Rumelhart & McClelland, 1986), ali i na druge, nefonološke informacije (Keuleers et al., 2007). Za modele jednog puta koji se oslanjaju na analogiju, poseban izazov predstavljaju *nekanonički* (nestandardni) oblici riječi, kao što su prezimena u engleskom jeziku (npr. množina prezimena *Foot* je *the Foots*, a ne *the Feet*, kako bi trebalo da bude, ako se zaključuje po analogiji *foot – feet*), neasimilovane tuđice (od riječi *fireman*/*vatrogasac* množina je *firemen*, dok je za riječ *talisman* množina *talismans*) itd. Ipak, modeli jednog puta i ovakve probleme mogu riješiti na principijelan način i bez uvođenja dodatnih pretpostavki o prirodi fenomena i/ili dodatnih mehanizama. Dovoljno je da se, npr. u jezicima sa "dubokom" ortografijom, pored fonoloških, uvedu i ortografske informacije (Keuleers et al., 2007).

[18] U okviru prirodne obrade jezika, pristup koji se bazira na učenju zasnovanom na memoriji označava se i kao *obrada jezika zasnovana na memoriji* (eng. *memory-based language processing*; Daelemans & Van den Bosch, 2005).

U odnosu na druge pristupe iz skupa modela jednog puta, kao što je analoški model (Skousen, 1989, 1992) ili *minimalno generalizovano učenje* (eng. *minimal generalization learning*; Albright & Hayes, 2003),[19] neki nalazi govore u prilog učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005), dok uporedna istraživanja, za sada, ne daju konačne i jednoznačne ocjene (Daelemans, 2002; Eddington, 2002b). Ipak, činjenice pokazuju da je MBL algoritamski jednostavniji i računarski manje zahtjevan model od klasičnog analoškog modela (Daelemans, 2002). Minimalno generalizovano učenje (Albright & Hayes, 2003), koje se oslanja na morfološka i fonološka probabilistička pravila u produkciji infleksionih oblika, predstavlja poseban slučaj učenja zasnovanog na memoriji. Ovaj pristup, štaviše, postiže slabije rezultate u zadacima automatske obrade jezika (Keuleers & Sandra, 2008).

[19] Minimalno generalizovano učenje se, takođe, zasniva na pretpostavci da je osnovni mehanizam infleksione produkcije sličnost, ali ne sličnost novog objekta sa primjerima u memoriji, nego *strukturirana sličnost* (eng. *structured similarity*), koja omogućava otkrivanje konteksta u kojem se dati infleksioni oblik veoma često javlja. Na primjer, u kontekstu s_ŋ, koji dobro opisuje grupu nepravilnih glagola, kao što je glagol swing-swung, promjena fonema [I] u fonem [Λ] prilikom tvorbe prošlog vremena je veoma česta (u: Keuleers & Sandra, 2008).

1.3. Cilj istraživanja

Istraživanjima u oblasti prirodne obrade jezika, koja u fokusu imaju problem automatske obrade riječi, može se staviti primjedba da ne uzimaju u obzir nalaze fundamentalnih istraživanja obrade jezika, kao i to da je sve podređeno postizanju uspjeha i zadovoljenju praktičnih potreba. To je razlog zašto se, s jedne strane, često u analizama nađu i npr. znaci interpunkcije, jednoznačne oznake i sl., što značajno povećava tačnost korišćenih tagera, dok, s druge strane, postoji neprincipijelno uključivanje velikog broja različitih informacija na koje se ovi sistemi za obradu riječi oslanjaju. Iako veliki broj autora navodi da automatska obrada riječi dostiže tačnost i do 98% (vidjeti, npr. Güngör, 2010; Manning, 2011; Manning & Schuetze, 2000), često je stvarna efikasnost daleko niža. O ovome govore rezultati obrade nepoznatih (Džeroski et al., 2000) i višeznačnih riječi (Dimitrijević i sar., 2008). U oba slučaja korišćeni su Markovljevi modeli, a postignuta tačnost kretala se između 55% i 57%.

Na osnovu prethodno rečenog može se zaključiti da su načini izbora informacija na koje se sistemi za obradu riječi oslanjaju problematični i da se ti sistemi ne zasnivaju na principima, već najčešće na *ad hoc* rješenjima. S obzirom na to da problem automatske obrade riječi nije razriješen, potrebno je nastaviti sistematski provjeravati moguću ulogu različitih jezičkih informacija u obavljanju ovakvih zadataka. Iz tog razloga, u ovoj studiji ispitivane su fonotaktičke informacije i njihov potencijal za povećanje tačnosti i robusnosti sistema za automatsku obradu riječi.

Termin *fonotaktičke informacije* (eng. *phonotactic information*) odnosi se na sekvence fonema koje se javljaju u jeziku, u okviru fonološki ispravnih riječi, pri čemu vjerovatnoće javljanja tih sekvenci nisu jednake (Crystal, 2008). Ove sekvence, tj. nizovi fonema, tradicionalno se razmatraju u kategorijama *dozvoljeno* vs. *nedozvoljeno* (Vitevitch & Luce, 2004). Tako, na primjer, u srpskom jeziku, niz od dvije foneme *fl* je moguć (dozvoljen), dok niz *žz* nije.

Značaj fonotaktičkih sekvenci u kognitivnoj obradi jezika potvrđen je u većem broju istraživanja. Na primjer, djeca ove informacije koriste za određivanje granica među riječima i segmentaciju govora (Cairns, Shillcock, Chater, & Levy, 1997; Mattys & Jusczyk, 2000; Mattys, Jusczyk, Luce, & Morgan, 1999; Morgan & Saffran, 1995), ali i za učenje novih riječi (Storkel, 2001, 2004; Storkel & Morrisette, 2002; Storkel & Rogers, 2000). Pokazano je i da bebe preferiraju pseudoriječi koje sadrže visoko frekventne nizove fonema (Jusczyk, Luce, & Charles-Luce, 1994). Kod odraslih, vjerovatnoće fonotaktičkih informacija utiču na uspjeh u ponavljanju pseudoriječi (Vitevitch & Luce, 1998, Vitevitch, Luce, Charles-Luce, & Kemmerer, 1997), na identifikaciju fonema (Pitt & McQueen, 1998), na brzinu izgovaranja (Levelt & Wheeldon, 1994) i na brzinu prepoznavanja izgovorenih riječi (Luce & Large, 2001; Vitevitch, 2003; Vitevitch & Luce, 1998, 1999; Vitevitch, Pisoni, Kirk, Hay-McCutcheon, & Yount, 2002). Dalje, fonotaktičke informacije utiču i na procjenu sličnosti pseudoriječi sa riječima (eng. *wordlikeness*; Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Frisch, Large, & Pisoni, 2000; Vitevitch et al., 1997). Zbog toga, mnogi autori smatraju da fonotaktičke informacije imaju ključnu ulogu u organizaciji mentalnog leksikona (Albright, 2007; Bailey & Hahn, 2001; Bybee, 1985, 2001; Coleman & Pierrehumbert, 1997; Luce & Pisoni, 1998; Pierrehumbert, 2003a, 2003b; Vitevitch & Luce, 1998, 2004; Vitevitch et al., 1999, itd.).

Pored vjerovatnoća sekvenci fonema, sličnost fonološkog niza sa riječima, iskazana preko broja riječi koje se od datog niza razlikuju za *n* elementarnih promjena, u koje spadaju brisanje, dodavanje ili zamjena postojećih fonema/grafema, takođe utiče na kognitivnu obradu jezika (Bailey & Hahn, 2001; Luce, 1986). Pritom, u najvećem broju istraživanja razmatrana je elementarna promjena samo jednog znaka, takozvano Kolhartovo *N* (vidjeti: Coltheart, Davelaar, Jonasson, & Besner, 1977; kao i brojne druge izvore: Grainger, Muneaux, Farioli, & Ziegler, 2005; Luce & Pisoni, 1998; Yao, 2011, Yates, Locker, & Simpson, 2004, itd.). U zadacima vizuelne obrade riječi (eng. *visual word recognition*) utvrđeno je da se riječi sa većim brojem susjeda (eng. *neighborhood density*) obrađuju brže i sa

manje grešaka, bilo da se radi o ortografskoj (Andrews, 1989, 1992; Coltheart et al., 1977; Forster & Shen, 1996; Sears, Hino, & Lupker, 1995) ili fonološkoj sličnosti (Yates et al., 2004). S druge strane, u zadacima auditivne obrade riječi (eng. *auditory word recognition*) veći skup ortografski sličnih susjeda facilitira obradu (Ziegler, Muneaux, & Grainger, 2003), dok, istovremeno, veći skup fonoloških susjeda inhibira obradu (Cluff & Luce, 1990; Garlock, Walley, & Metsala, 2001; Goldinger, Luce, & Pisoni, 1989; Vitevitch & Luce, 1998, 1999; Ziegler et al., 2003).

Novija istraživanja pokazuju da se efekat susjedstva može objasniti principima diskriminativnog učenja (Milin, Ramscar, Cho, Baayen, & Feldman, 2014; više o diskriminativnom učenju u: Ramscar & Yarlett, 2007 i Baayen et al., 2011). Upotrebom modela naivnog diskriminativnog učenja pokazano je da se uticaj susjedstva javlja tokom učenja, a ne u momentu donošenja odluke. Učenje se, opet, odvija uvijek kada je znak (eng. *cue*) prisutan i može voditi (a) jačanju veze s relevantnim ishodom (eng. *outcome*), u situaciji kada je i taj ishod takođe prisutan, ili (b) slabljenju veze, tj. korekciji greške, ako ishod nije prisutan. Susjedne riječi su značajne, upravo, za korigovanje uspostavljenih veza (eng. *error-driven recalibration*), jer dijele mnoge diskriminativne znakove sa metom. Ovo učenje, dakle, dovodi do slabljenja aktivnih veza sa metom. Međutim, upravo korigovanje greške pruža više učenja i ostvaruje "bogatije iskustvo" sa tom istom metom (Milin, et al., 2014). Jednostavno rečeno, susjedne riječi obezbjeđuju i pozitivne i negativne efekte za metu. Ovi složeni dinamički odnosi ne mogu dati jednostavne predikcije o čemu i govore brojni kontradiktorni nalazi ranijih istraživanja (na primjer, kada je riječ o uticaju skupa ortografski sličnih susjeda, facilitacija je dobijena u: Andrews, 1989, 1992; Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Forster & Shen, 1996; Huntsman & Lima, 2002; Sears et al., 1995; a inhibicija u: Carreiras, Perea, & Grainger, 1997; Johnson & Pugh, 1994; itd.).

Uzevši u obzir prethodnu diskusiju, u ovom radu ispitana je mogućnost korišćenja fonotaktičkih informacija u automatskoj obradi riječi u morfološki složenom jeziku. Informativnost fonotaktičkih nizova testirana je u zadatku

automatske diskriminacije vrsta riječi i zadatku automatske produkcije infleksionih oblika riječi.

U prvom zadatku, vjerovatnoće mogućih sekvenci fonema poslužile su kao osnova za klasifikaciju gramatičkih tipova (npr. imenice muškog roda u nominativu jednine; pozitiv pridjeva ženskog roda u genitivu množine, infinitiv itd.) po pripadajućim vrstama riječi. Klasifikacija je obavljena pomoću mašina sa vektorima podrške, koje spadaju u najuspješnije algoritme za klasifikaciju (Baayen, 2011; Joachims, 1998; Meyer et al., 2003; Steinwart & Christmann, 2008; Van Gestel et al., 2004).

U drugom zadatku ispitivana je mogućnost automatske produkcije infleksionih oblika riječi pomoću modela učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005). Infleksioni oblik nove riječi generisan je na osnovu fonotaktičke sličnosti sa riječima u memoriji, tj. na osnovu analogije sa grupom najbližih/najsličnijih susjeda. Ovaj pristup odabran je s obzirom na dugu tradiciju objašnjavanja različitih fenomena jezičke produkcije pomoću mehanizama analogije (vidjeti: De Saussure, 1916/1966; Bloomfield, 1933; Harris, 1951; 1957 itd). Povrh toga, neki autori smatraju da je učenje zasnovano na memoriji bitna komponenta u obradi morfološki složenih riječi te da je nužno poređenje i/ili kombinovanje ovog i drugih relevantnih pristupa istom problemu (na primjer: Milin et al., 2009).

Na kraju, kognitivna plauzabilnost (vjerodostojnost) učenja zasnovanog na memoriji provjeravana je i eksperimentalno, u zadatku leksičke odluke. Stimulusi (riječi) za ovaj ekperiment odabrani su u zavisnosti od toga da li ih je prethodno primijenjeni model generisao tačno ili pogrešno. Drugim riječima, eksperimentalno su testirani ishodi – rezultati koje je generisao MBL model.

Iako se problem ovog istraživanja značajnim dijelom dotiče problematike koja se razmatra u oblasti računarske lingvistike i obrade prirodnog jezika, cilj istraživanja nije bio razvoj računarskog sistema za automatsku obradu riječi. Dobijeni rezultati značajni su, prije svega, za kognitivnu psihologiju i psiholingvistiku.

Prvo, oni će ukazati u kojoj mjeri se naš kognitivni sistem može osloniti na fonotaktičke informacije u različitim zadacima razumijevanja i produkcije morfološki složenih riječi.

Drugo, rezultati ovog istraživanja pokazaće domet učenja zasnovanog na memoriji u zadatku automatske produkcije infleksionih oblika riječi, u jeziku sa bogatom infleksionom morfologijom.

Konačno, istraživanje će pokazati da li su teorije i modeli koji se u tumačenju jezičkih fenomena oslanjaju na mehanizme analoškog zaključivanja, kognitivno plauzibilni ili ne. S obzirom na to da se učenje zanovano na memoriji oslanja na fonotaktičke informacije, indirektno će biti provjerena i kognitivna vjerodostojnost ovih informacija. Pritom, cilj je bio ispitivanje kognitivne plauzibilnosti fonotaktičkih informacija i analoškog zaključivanja sistema za učenje zasnovanog na memoriji (o sličnim pitanjima vidjeti i u: Boden, 1987; 2006).

2. FONOTAKTIČKE INFORMACIJE I PROBLEM DISKRIMINACIJE VRSTA RIJEČI

U prvoj studiji ispitivana je tačnost diskriminacije vrsta riječi[20] na osnovu fonotaktičkih informacija. U tu svrhu korišćene su kombinacije dva odnosno tri fonema/grafema, koje se mogu javiti u jeziku,[21] tzv. bigrami i trigrami.[22] Vjerovatnoće bigrama i trigrama izračunate su na nivou gramatičkih tipova (npr. imenica muškog roda u nominativu množine i sl.), a sistematski je variran položaj bigrama i trigrama u riječima.

Klasifikacija je obavljena upotrebom *mašina sa vektorima podrške* (eng. *support vector machines – SVM*; Vapnik, 1995, 1998). Ova tehnika je odabrana zato što su SVM pouzdan, matematički dobro zasnovan i efikasan alat za klasifikaciju, koji nema posebnih pretpostavki o distribucijama prediktorskih varijabli (Steinwart & Christmann, 2008). Ovo je bio ključni preduslov jer je najveći broj vrijednosti za bigrame i trigrame bio nula, tj. ulazne matrice su bile prazne (eng. *sparse matrix*). To praktično znači da se većina fonotaktičkih jedinica javila u malom broju riječi.[23] Distribucije se nisu značajno promijenile ni nakon što su u postupku *korekcije učestanosti* (eng. *smoothing*), između ostalog, nultim vrijednostima pripisane male pozitivne vrijednosti. Značajno odstupanje distribucija od normalne onemogućilo je primjenu postupaka iz porodice generalnih linearnih modela, koji se zasnivaju na pretpostavci o normalnosti distribucija prediktorskih varijabli, kao što je npr.

[20] S obzirom na to da su kao osnova za diskriminaciju vrsta riječi poslužile frekvencije bigrama i trigrama izračunate na nivou gramatičkih tipova, sintagme *diskriminacija/razdvajanje vrsta riječi* i *klasifikovanje gramatičkih tipova* u ovom radu su korišćeni kao sinonimi. Naime, ove sintagme označavaju isti proces, sagledan iz dva ugla: razdvajanje vrsta riječi na osnovu fonotaktičkih informacija specifikovanih na nivou gramatičkih kategorija, odnosno, svrstavanje tih gramatičkih kategorija na osnovu izračunatih frekvencija u pripadajuće vrste riječi.

[21] Srpski jezik se odlikuje plitkom ortografijom, tj. grafemski i fonološki kod su izomorfni, pri čemu je fonološki kod primaran (Feldmann & Turvey, 1983). U daljem tekstu će se zbog toga navoditi samo fonološki kod, iako su podaci, na kojima su vršene analize, dobijeni na osnovu korpusa pisanog jezika.

[22] Termini bigram i trigram mogu dovesti do konfuzije pošto se koriste, kako za nizove karaktera (ili fonema), tako i za nizove riječi, pa čak i za veće jezičke jedinice. Osim toga, neki autori upozoravaju da su ovi termini i pogrešni, jer miješaju starogrčki i latinski, te da bi dosljedno trebalo koristiti termin *digram* (vidi i: Shannon, 1951) i sl.

[23] Distribucija vjerovatnoća za ovakvu situaciju pripada klasi *velikog broja rijetkih događaja* (eng. *Large Number of Rare Events* – LNRE).

diskriminativna analiza. S druge strane, svođenje sirovih frekvencija na kategoričke varijable (npr. prisutan-odsutan bigram/trigram) nije preporučljiva, iz više razloga. Dihotomizacija kontinuiranih varijabli ima brojne negativne posljedice, kao što su: gubitak informacija o pojedinim slučajevima, gubitak veličine efekta i statističke snage, moguće previđanje nelinearnih relacija među varijablama, manja pouzdanost mjerenja itd. (više u: MacCallum, Zhang, Preacher, & Rucker, 2002). Ovaj postupak se predlaže samo u slučajevima analize ekstremnih vrijednosti, kada je cilj istraživanja provjera ponašanja dihotomiziranih mjera te u slučaju kada je varijabla, po svojoj prirodi, kategorička (vidi: DeCoster, Iselin, & Gallucci, 2009). Kao što se može primijetiti, ovo ne vrijedi za distribucije bigrama i trigrama u ovoj studiji. Iz navedenih razloga, opravdan izbor za klasifikaciju vrsta riječi na osnovu bigrama i trigrama bile su mašine sa vektorima podrške.

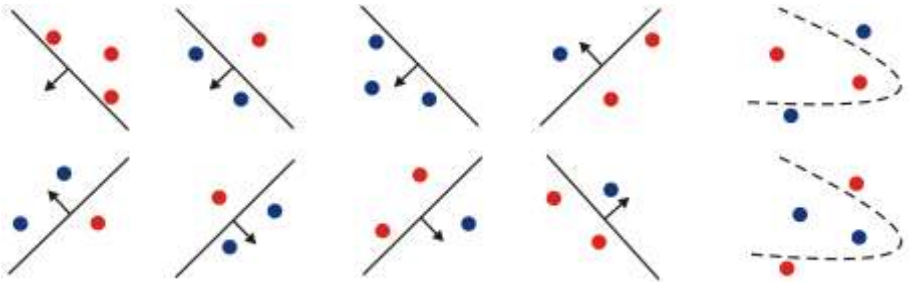
2.1. Mašine sa vektorima podrške

Mašine sa vektorima podrške predstavljaju grupu metoda *nadgledanog učenja* (eng. *supervised learning*),[24] koje se mogu koristi u različitim zadacima klasifikacije (npr. Boser, Guyon, & Vapnik, 1992; Cortes & Vapnik, 1995) i regresije (Drucker, Burges, Kaufman, Smola, & Vapnik,1997). SVM su zasnovane na *teoriji statističkog učenja* (eng. *statistical learning theory*) i *Vapnik-Červonenkisovoj dimenziji* (eng. *Vapnik–Chervonenkis dimension*; Vapnik, 1979, 1995, 1998; Vapnik & Chervonenkis; 1968, 1971). Teorija statističkog učenja ima za cilj da pruži okvir za proučavanje problema statističkog zaključivanja i generalizacije znanja stečenog na nekom skupu podataka – nezavisnim podacima dobijenim iz stabilnih (fiksnih), ali nepoznatih distribucija, tj. znanja stečenog učenjem (Vapnik, 1995, 1998). Na učenje se može gledati kao na proces aproksimacije ciljne funkcije, koja opisuje odnos ulaza i izlaza, npr. pronalaženje funkcije koja opisuje vezu između

[24] U toku učenja algoritam popravlja svoje performanse. Kada se radi o *nadgledanom učenju* (eng. *supervised learning*), pored podataka na osnovu kojih se vrši generalizovanje (podaci za uvježbavanje), algoritam raspolaže i informacijom o željenom izlazu.

karakteristika primjera na kojima se učenje obavlja (ulaz) i klasa kojima primjeri pripadaju (izlaz).

Teorija statističkog učenja sugeriše da je neophodno ograničiti *prostor hipoteza*, tj. skup svih dopustivih aproksimativnih funkcija – hipoteza, s tim da te funkcije treba da imaju dovoljan *kapacitet* (eng. *capacity*), koji se tiče maksimalnog broja primjera za učenje, koji dati skup hipoteza treba da *razdvoji* (eng. *shattered*). Ovo svojstvo se naziva *VC kapacitet* ili *VC dimenzija* (Vapnik, 1979, 1995, 1998; Vapnik & Chervonenkis; 1968, 1971). VC kapacitet predstavlja i mjeru kompleksnosti prostora hipoteza (Burgess, 1999). Tako je, na primjer, VC dimenzija skupa prâvih u dvodimenzionalnom prostoru – tri. Naime, *n* tačaka (primjera za učenje) može se rasporediti na 2^n načina u dvije kategorije (npr. [crveno, plavo]). U slučaju tri primjera za učenje, dovoljna je prava za njihovo tačno klasifikovanje, bez obzira kako se oni raspoređuju po kategorijama (Slika 1a). Međutim, četiri primjera za učenje, koji se mogu rasporediti na 16 (2^4) različitih načina po kategorijama, ne mogu uvijek biti razdvojeni pravom (Slika 1b), zbog čega je nužan klasifikator veće kompleksnosti (Ivanciuc, 2007).

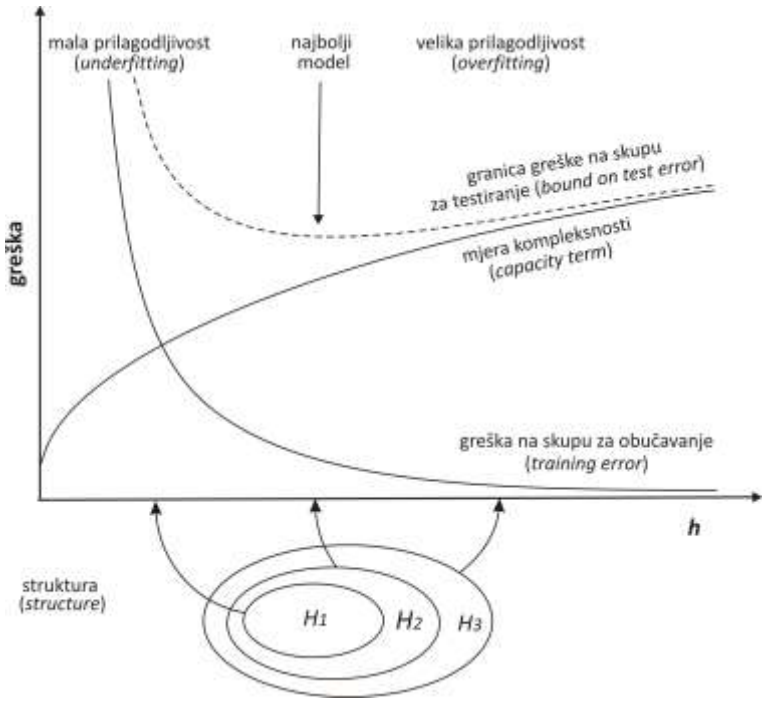


a. b.

Slika 1. VC dimenzija skupa pravih u dvodimenzionalnom prostoru. a. Različiti rasporedi tri primjera u dvije kategorije, razdvojivi pomoću jedne prava. b. Distribucije četiri primjera u dvije kategorije koje nisu razdvojive jednom pravom (Kecman, 2001).

Za uspješno učenje ključan je izbor prostora hipoteza odgovarajuće kompleksnosti. Pretpostavimo da se zamišljeni podaci mogu opisati polinomialnim funkcijama sa jednom promjenljivom. Ovakve funkcije mogu se podijeliti u *ugniježdene* (eng. *nested*) podskupove, zavisno od kompleksnosti, tj. stepena

polinoma (Burgess, 1999; Cherkassky & Mulier, 1998). U strukturi H_1 naći će se linearne funkcije (funkcije iskazane polinomom stepena jedan) $f(x) = a_1x + a_0$; u strukturi H_2 funkcije drugog stepena $f(x) = a_2x^2 + a_1x + a_0$; a u strukturi H_3 polinomialne funkcije trećeg stepena $f(x) = a_2x^2 + a_2x^2 + a_1x + a_0$. Odgovarajuća hipoteza/model za ovaj zamišljeni primjer učenja dobija se za strukturu H_2 (vidi Sliku 3). Naime, strukturu H_1 karakteriše niska VC dimenzija, tj. model je nedovoljno kompleksan, zbog čega pravi značajnu grešku već na skupu za učenje (eng. *underfitting*). S druge strane, visoka VC dimenzija prostora hipoteza H_3 omogućava veliku prilagodljivost modela podacima za trening, što za posljedicu može imati njegovu manju uspješnost predikcije na novim podacima (pretjerano pristajanje, eng. *overfitting*).

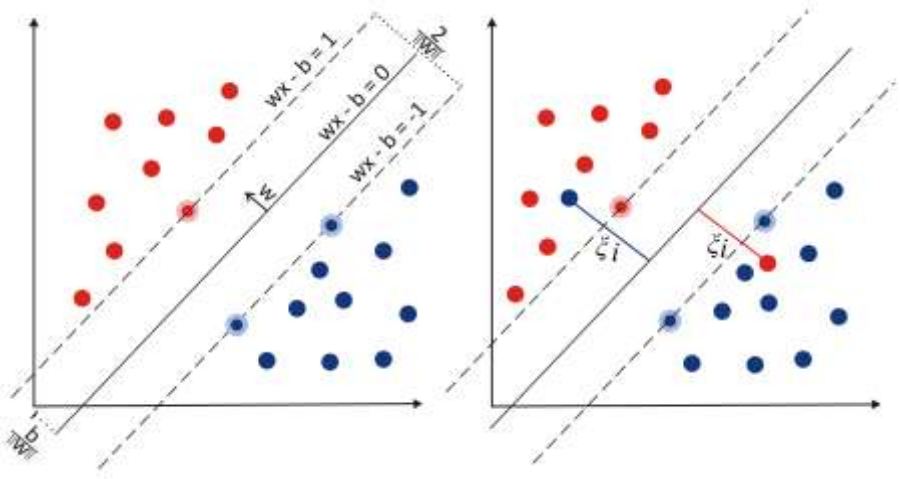


Slika 2. Minimalizacija strukturalnog rizika (Sewell, 2006)

Da bi se našla optimalna hipoteza, tj. model, potrebno je balansirati između kapaciteta (VC dimenzije) i utvrđene greške na uzorku za učenje, stoga algoritam slijedi princip *minimizacije strukturalnog rizika* (eng. *structural risk minimization*).

Pomenuti princip podrazumijeva da se model bira iz prostora sa što manjom VC dimenzijom, pri čemu greška na primjerima za učenje ne smije biti (pre)velika (Slika 2; Cortes & Vapnik, 1995). Mnogi autori smatraju da SVM postižu visoku efikasnost upravo zato što poštuju ovaj princip (Cortes & Vapnik, 1995; Vapnik, 1995, 1998).

Ideja koja leži u osnovi SVM jeste pronalaženje optimalne *hiperravni* (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1995), čija jednačina predstavlja sam model. Optimalna hiperravan obezbjeđuje najbolju generalizaciju stečenog znanja. Da bi se ovo realizovalo, potrebno je pronaći *maksimalnu marginu*, koja razdvaja skupove podataka (Slika 3a). Hiperravan je koncept iz geometrije, koji se odnosi na bilo koji podprostor dimenzije *n-1* u nekom prostoru R^n (Guggenheimer, 1977; Prasolov & Tikhomirov, 2001). Tako, na primjer, u jednodimenzionalnom prostoru hiperravan je predstavljena tačkom, u dvodimenzinalnom prostoru – pravom, a u trodimenzionalnom prostoru hiperravan je površ (ravan). U svim navedenim slučajevima hiperravan dijeli posmatrani prostor na odgovarajući broj dijelova (klasa ili grupa). Pritom, margina, tj. udaljenost hiperravni od najbližeg primjera za učenje, mora biti maksimalna.



a. b.

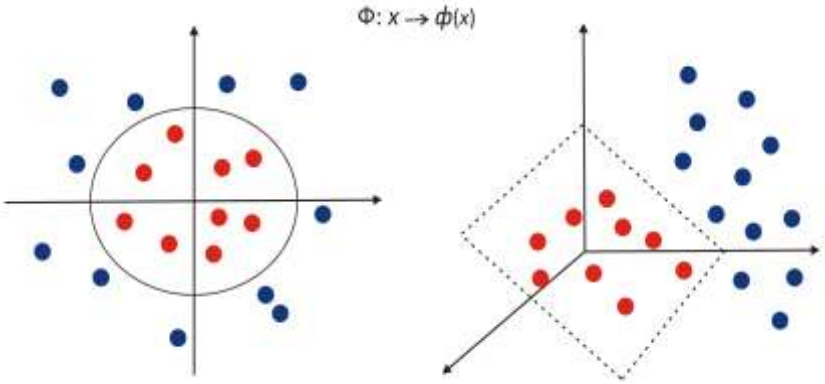
Slika 3. Shematski prikaz optimalne hiperravni. *a*. Linearno razdvojivi podaci; *b*. linearno neodvojivi podaci.

Optimalna hiperravan određena je *podržavajućim* (*potpornim*) *vektorima*, koji predstavljaju podskup podataka (objekata) iz skupa za uvježbavanje. Na osnovu informacija o udaljenosti objekta od hiperravni, donosi se odluka o grupi kojoj taj objekat pripada. Jasno je da zadatak klasifikacije teži, što je maksimalna dostignuta margina manja.

Ako se pretpostavi da su objekti linearno razdvojivi, jednačina hiperravni određena je parametrom w (vektor koji je ortogonalan na hiperravan; Slika 3a) i parametrom b i glasi $(w \cdot x_i - b) = 0$, pri čemu je podatak x, iz skupa za uvježbavanje veličine n, predstavljen d-dimenzionalnim vektorom $x_i = (x_1, x_2, ..., x_d)$, kojem je pridružena jedna od vrijednosti iz skupa $y_i \in \{-1,1\}$, tj. klasa kojoj objekat pripada. Tada, za sve elemente skupa za uvježbavanje vrijedi $(w \cdot x_i - b) \geq 1$ ili $(w \cdot x_i - b) \leq -1$ (Slika 3a), što je ekvivalentno $c_i (w \cdot x_i - b) \geq 1$, $1 \leq i \leq n$ (Olson & Dulen, 2008), pri čemu je c_i konstanta, koja može imati vrijednost 1 ili -1. Kako bi pronašli maksimalnu marginu koja razdvaja klase, potrebno je maksimirati vrijednost izraza $2 / ||w||$ (Slika 3a), što odgovara minimalizaciji vrijednosti izraza $(1/2)||w||^2$, pri čemu je $c_i(w \cdot x_i - b) \geq 1$, $1 \leq i \leq n$).

SVM se mogu primijeniti i u slučaju kada objekti nisu linerano razdvojivi i to na dva načina. Prvi način je tolerisanje manjeg broja grešaka prilikom faze učenja, ali i kasnije, prilikom klasifikacije. To se može postići uvođenjem promjenljivih ξ_i (Slika 3b), tako da se dobija $c_i(w \cdot x_i - b) \geq 1 - \xi_i$, $1 \leq i \leq n$. U tom slučaju, neophodno je da parametri w i b budu određeni tako da se dobije minimum za izraz $(1/2)||w||^2 + C\Sigma\xi_i$, gdje je $i = 1, ..., n$; $\xi_i \geq 0$ (Vapnik, 1995). Parametar C određuje "cijenu" greške u klasifikaciji objekata (eng. *error cost*). On pravi balans između dozvoljenih grešaka u procesu učenja i maksimalne širine margine. Manja vrijednost ovog parametra podrazumijeva i manju osjetljivost na greške, što za posljedicu ima maksimiranje širine optimalne hiperravni (Ivanciuc, 2007). Velikim "kaznama" za greške u klasifikaciji dobija se tačniji model u procesu učenja, uz rizik slabije generalizacije na novim podacima.

Drugi način je da se koordinate objekata iz *ulaznog prostora*, pomoću nelinerane funkcije *preslikavanja karakteristika* $\phi(x)$ (eng. *feature functions*), preslikaju u višedimenzionalni *prostor karakteristika* (eng. *feature space*). Novi prostor, u kome je trenirajući skup linearno razdvojiv (Slika 4), uobičajeno ima više dimenzija od ulaznog (Zhang, 2001).



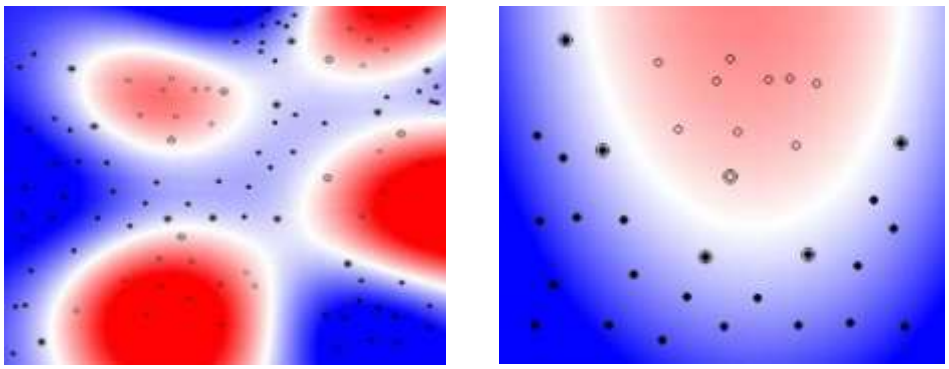
Slika 4. Preslikavanje linearno neodvojivih podataka (ulazni prostor) u prostor sa više dimenzija (prostor karakteristika)

S jedne strane, prostor karakteristika može imati veoma veliki broj dimenzija, čak i beskonačan, dok je, s druge strane, nalaženje odgovarajuće funkcije $\phi(x)$ složen proces, koji se zasniva na pokušajima i pogreškama (Ivanciuc, 2007). S obzirom na to da se ovaj proces odvija u dvije faze (transformacija ulaznog prostora u prostor karakteristika i linearno razdvajanje podataka, on može biti i računarski veoma zahtjevan). Zbog toga nije praktična direktna upotreba funkcije preslikavanja karakteristika $\phi(x)$, koja podrazumijeva transformisanje i kombinovanje originalnih koordinata objekata i njihovo preslikavanje u prostor sa više dimenzija. Ovaj problem se može izbjeći uz pomoć *jezgrene funkcije* (eng. *kernel*) $K(x_i, x_j)$, koja predstavlja skalarni proizvod $K(x_i, x_j) = \phi(x_i)\ \phi(x_j)$ i koristi se kao mjera sličnosti između x_i i x_j. Više nije neophodno eksplicitno specifikovati novi prostor niti poznavati funkciju $\phi(x)$ jer se jezgrena funkcija može direktno izračunati u originalnom (ulaznom) prostoru. Ovakav pristup se ne oslanja na pojedinačne reprezentacije ulaznih podataka, kao što bi to bio slučaj prilikom preslikavanja u novi, višedimenzionalni prostor karakteristika, već na skup udvojenih poređenja

(eng. *pairwise comparisons*) $k(x_i, x_j)$ datih u matrici n x n (Vert, Tsuda, & Schölkopf, 2004).

Postoji veliki broj jezgrenih funkcija, kao što su: *linearna*, *radijalna* (eng. *gaussian radial basis function – RBF*), *eksponencijalno-radijalna* (eng. *exponential radial basis function*), polinomialna (eng. *polynomial*), *ANOVA kernel*, *sigmoidalni kernel* (eng. *neural/sigmoid or tanh*) itd. (Ivanciuc, 2007).

Radijalna (RBF) funkcija je najčešće korišćena funkcija pri implementaciji SVM (Slika 5). Osim toga što se može primijeniti i u slučaju kada je odnos između podataka i njima pripadajućih klasa nelinearan, pod određenim uslovima linearni i sigmoidalni kernel se mogu tretirati kao specijalni slučajevi radijalne funkcije (Keerthi & Lin, 2003; Lin & Lin, 2003). Prednost radijalne u odnosu na ostale nelinearne jezgrene funkcije je i manji broj parametara koji utiče na njegovu kompleksnost. Ipak, ako je broj karakteristika (varijabli) veliki, opravdana je upotreba linearnog kernela, jer se nelinearnim preslikavanjem ne dobija značajno bolji rezultat, a kompleksniji je za primjenu zbog većeg broja parametara koje je potrebno odrediti: linearna kernel funkcija traži ocjenu jednog – C, a radijalna kernel funcija ocjenu dva parametra – γ, C (Hsu, Chang, & Lin, 2010).



a. b.

Slika 5. Mašine sa vektorima podrške sa: *a*. RBF jezgrenom funkcijom; *b*. polinomialnom jezgrenom funkcijom sa eksponentom dva (Moore, 2001, 2003).

Veoma popularan i relativno jednostavan način određivanja parametara jezgrenih funkcija je *mrežno pretraživanje* (eng. *grid search*; Hsu & Lin, 2002).

Ovaj postupak se predlaže, prije svega, zbog osjećaja nepovjerenja prema metodama koje ne vrše iscrpnu pretragu (Hsu et al., 2010).

Mrežno pretraživanje bi se, npr. za RBF kernel obavljalo u dva koraka. Prvi korak bi podrazumijevao sistematsko variranje parametara *C* i *γ*, gdje bi se ovi parametri eksponenicijalno povećavali u određenom rasponu, npr. $C = 2^{-5}; 2^{-3}; ..., 2^{15}$, a $\gamma = 2^{-15}; 2^{-13}; ..., 2^{3}$. Kada se identifikuje oblast u kojoj se dobija najbolji rezultat, npr. $C(2^{1}, 2^{5})$ i $\gamma(2^{-7}, 2^{-3})$, pristupa se "finom" pretraživanju, gdje se parametri linearno povećavaju za 0.25, C $(2^{1}, 2^{1,25},..., 2^{5})$ i γ $(2^{-7}, 2^{6,75},..., 2^{-3})$. Nakon što se nađu najbolji parametri *C* i *γ*, ponovo se trenira cijeli *uzorak za uvježbavanje* (eng. *training set*), kako bi generisao najbolji klasifikator (Hsu et al., 2010; Olson & Dulen, 2008). Ovakav pristup može biti nedovoljno dobar u slučaju kada postoji više hiljada atributa koji su uključeni u analizu (Olson & Dulen, 2008). U tom slučaju, prije nego što se primijene SVM, potrebno je prvo napraviti selekciju atributa.

Mrežno pretraživanje treba kombinovati sa *kros-validacijom* (eng. *cross-validation*) kako bi se izbjeglo *preučavanje* (eng. *overfitting*), tj. pretjerano prilagođavanje modela podacima na kojima se testira. Pridavanje velike pažnje slučajnim varijacijama podataka ima za posljedicu lošiju generalizaciju modela, tj. manje efikasnu primjenu na nove podatke. Ovaj problem je moguće prevazići postupkom kros-validacije, čija osnova leži u dijeljenju podataka na dva poduzorka, na skup za učenje/treniranje i skup za testiranje.[25]

[25] Postoji više različitih varijanti kros-validacije – *KV*: *izostavljanje jednog primjera* (eng. *leave-one-out*), *izostavljanje p primjera* (eng. *leave-p-out*), *unakrsna validacija sa k preklapanja* (eng. *k-fold CV*), *podjela na dva dijela* (eng. *holdout*), *balansirana nekompletna KV* (eng. *balanced incomplete CV*), *Monte Karlo KV* (eng. *Monte-Carlo CV*), *generalizovana KV* (eng. *generalized CV*), *testiranje ponovljenim učenjem* (eng. *repeated learning-testing*), *LOO bootstrap*, *.632 bootstrap* itd. (Arlot & Celisse, 2010). Koji postupak je odgovarajući zavisi od brojnih faktora: *aproksimirane greške* (eng. *bias*), varijanse, kompleksnosti izračunavanja itd. (Arlot & Celisse, 2010).

2. 2. Metod

2.2.1. Materijal: Materijal neophodan za provjeru mogućnosti diskriminacije vrsta riječi na osnovu fonotaktičkih informacija, tj. bigrama i trigrama, bio je raspoređen u osam matrica. Za bigrame su formirane četiri matrice: matrica sa frekvencijama/učestalostima javljanja bigrama na početku riječi (npr. bigram *#p*, pri čemu # označava prazno polje), matrica sa frekvencijama javljanja bigrama na kraju riječi (npr. *p#*), matrica sa bigramima, bez obzira na njihovu poziciju u riječi (npr. *pa*), i matrica koja objedinjuje prethodne tri vrste informacija. Ovakve matrice su formirane i za trigrame, tj. u jednoj matrici su se nalazile informacije o učestalosti javljanja trigrama na početku riječi (npr. *#pa*), u drugoj frekvencije trigrama na kraju riječi (npr. *pa#*), u trećoj trigrami, bez obira na njihovu poziciju (npr. *pap*), dok su u četvrtoj objedinjene informacije iz prethodne tri matrice. Dio matrice učestalosti za bigrame, bez obzira na poziciju bigrama u riječima, dat je u Tabeli 1.

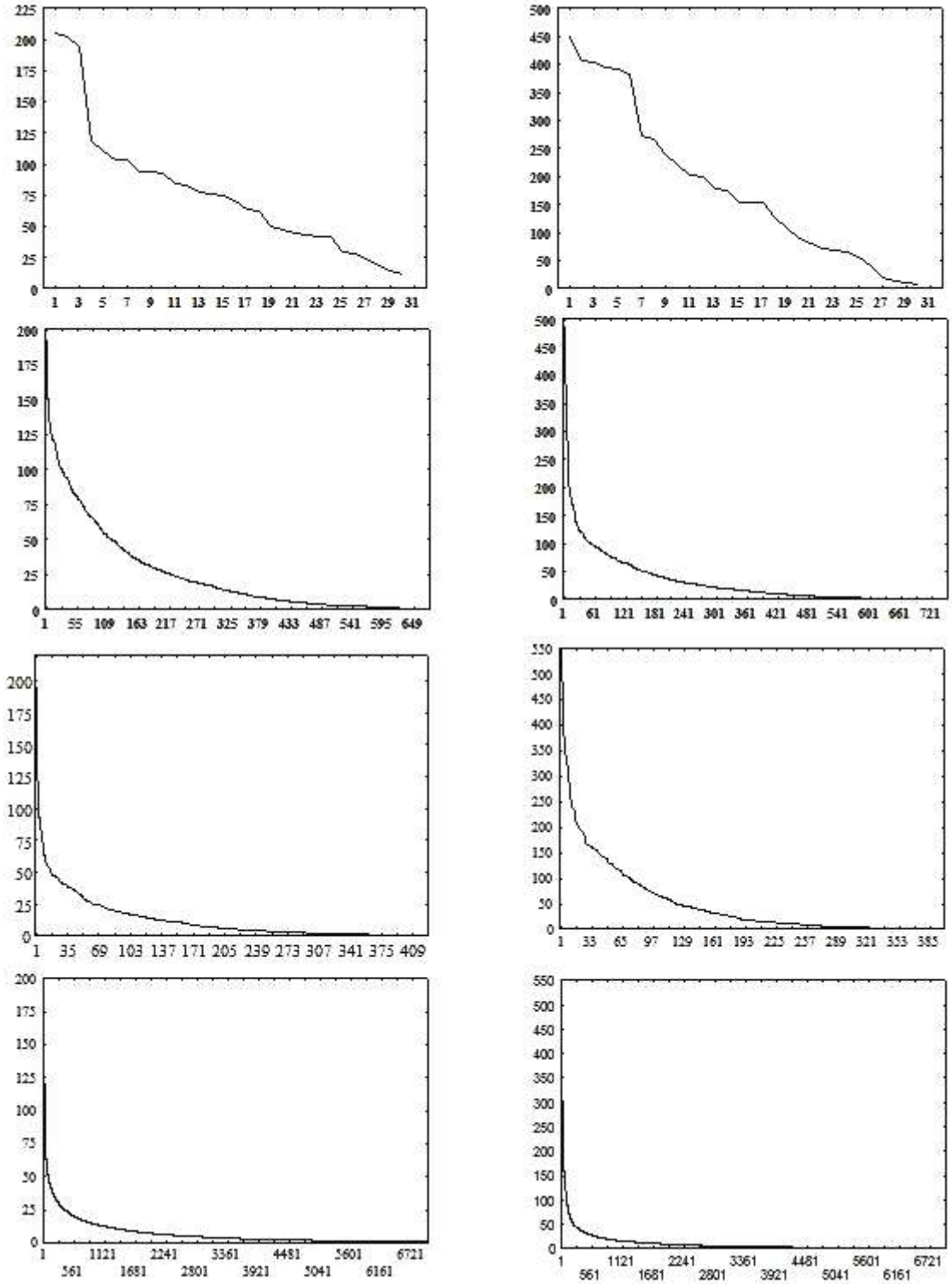
Tabela 1. *Dio matrice učestalosti za bigrame, bez obzira na njihovu poziciju u riječima*

Šifra	ra	ke	je	no	an	re	ar	st
.....	...	...	...	...	...	...	...	...
100213	1136	4	174	441	391	1107	535	1496
100221	5101	30	336	794	1410	789	2768	1215
100222	934	5	216	237	964	505	445	750
100223	171	0	31	5	22	399	49	347
100231	16	2	2	0	14	1	5	0
100232	25	0	77	12	58	0	10	20
100233	16	0	0	0	1	0	0	6
.....	...	...	...	...	...	...	...	...

Napomena: Prvi broj u šifri određuje vrstu riječi (1 – imenica), četvrti broj padež (2 – genitiv), peti broj (1 – jednina, 2 – množina, 3 – pluralia tantum), a šesti rod (1 – muški, 2 – ženski, 3 – srednji). U tekstu će, zbog jednostavnosti, biti korišćene numeričke šifre. Detaljan opis sistema anotacije dostupan je na web-stranici Korpusa srpskog jezika: http://www.serbian-corpus.edu.rs.

Matrice su sadržavale po 1293 reda, dok je broj kolona zavisio od broja bigrama odnosno trigrama. U redovima su se nalazili gramatički tipovi, i to za: imenice (74), pridjeve (131), zamjenice (594) i glagole (494). Za računanje frekvencija poslužio je poduzorak dnevne štampe *Korpusa savremenog srpskog jezika* (Kostić, 2001), veličine milion riječi.

Formiranje matrica imalo je nekoliko faza. Prvo su identifikovani svi bigrami i trigrami koji se javljaju u poduzorku dnevne štampe Korpusa srpskog jezika (Kostić, 2001). Zatim su eliminisani oni koji su sadržavali interpunkcijske znakove, brojeve i sl., kao i sve greške – kombinacije fonema koje se ne javljaju u srpskom jeziku (npr. *rr*). U sljedećem koraku utvrđene su frekvencije za zadržane bigrame i trigrame. Nakon što su izračunate frekvencije, broj bigrama i trigrama, izuzev bigrama na početku i na kraju riječi, redukovan je tako što su zadržani oni čija je prosječna frekvencija po jednom gramatičkom tipu (suma frekvencija/broj različitih gramatičkih tipova u kojima se dati bigram ili trigram javio) veća od 25. Ova vrijednost je određena tako da se sačuva što više ulaznih, sirovih podataka, uz istovremeno smanjivanje fragmentiranosti matrice (kontrolisanje ukupnog broja praznih ćelija). Njen izbor se zasnivao na grafičkoj procjeni broja "važnih" bigrama i trigrama. Postupak je bio sličan Katelovom postupku određivanja značajnih faktora u faktorskoj analizi (eng. *scree test*; Catell, 1966). Ako se vrijednosti prosječnih frekvencija po jednom gramatičkom tipu predstave grafikonom, tako što se sortiraju u opadajućem redoslijedu, od najveće ka najmanjoj, izabrana mjera je predstavljala, približno, tačku u kojoj se oblik krive mijenja i ona prelazi u horizontalnu liniju (Slika 6). Pretpostavka je da će podaci ispod ove tačke, slično faktorima u faktorskoj analizi, najviše doprinijeti objašnjenju varijanse u skupu bigrama/trigrama (više u: Catell, 1966). Na taj način su zadržani visoko frekventni bigrami i trigrami, bez obzira na broj gramatičkih tipova u kojima se pojavljuju. Takođe, zadržani su i bigrami i trigrami koji nemaju visoku frekvenciju, ali se javljaju u manjem broju oblika.



Slika 6. Racio frekvencije i broja gramatičkih tipova za bigrame i trigrame prije redukcije. Prvi red: lijevo – bigrami na početku riječi [#x], desno – bigrami na kraju riječi [x#]; drugi red: lijevo – bigrami bez obzira na poziciju [xy], desno – svi bigrami [#x, x#, xy]; treći red: lijevo – trigrami na početku riječi [#xy], desno – trigrami na kraju riječi [#xy]; četvrti red: lijevo – trigrami bez obzira na poziciju u riječi [xyz], desno – svi trigrami [#xy, xy#, xyz]. Na *x*-osi se nalazi rang, a na *y*-osi prosječna vrijednost frekvencije po gramatičkoj kategoriji.

U Tabeli 2 dat je broj bigrama i trigrama prije i poslije redukcije, kao i procenat zadržanih bigrama i trigrama (%), a za zadržane bigrame i trigrame prosječan racio frekvencije i broja gramatičkih tipova u kojima se određeni niz fonema javlja ($M_{f/r}$), te standardne devijacije ($SD_{f/r}$).

Tabela 2. *Deskriptivne mjere za bigrame i trigrame, prije i poslije redukcije*

	bigrami				trigrami			
	#x	x#	xy	svi	#xy	xy#	xyz	svi
prije redukcije	30	30	672	732	410	396	6987	7793
poslije redukcije	30	30	221	273	62	179	374	615
%	100	100	32.89	37.24	15.12	45.20	5.35	7.89
$M_{f/r}$	73.80	179.34	60.86	76.64	48.72	107.06	42.79	62.09
$SD_{f/r}$	49.16	143.67	31.98	67.38	28.26	89.75	23.10	59.78

Napomena: Oznake *x, y* i *z* se odnose na bilo koje foneme, pri čemu su moguće samo kombinacije dva odnosno tri fonema, koje se javljaju u srpskom jeziku. Kada se oznaka # nalazi ispred *x* označava početak riječi, a iza *x* odnosno *y*, označava kraj riječi.

2.2.2. Priprema materijala za analizu: Postojanje velikog broja ćelija u matricama u kojima je frekvencija jednaka nuli (eng. *sparse data*) ne predstavlja (preveliki) izazov za SVM (Farquad, Ravi, & Bapi, 2010). Međutim, rezultati istraživanja pokazuju da je ovaj postupak efikasniji, ako se iz analize isključe one karakteristike koje se rijetko javljaju, npr. jednom (Nakagawa et al., 2001). Osim toga, prije primjene SVM autori predlažu linearno skaliranje ulaznih podataka u rasponu [-1, 1] ili [0, 1] (Hsu et al., 2010). Iz tog razloga su frekvencije bigrama i trigrama preračunate pomoću modifikovane *proste Gud-Turingove korekcije za učestalosti* (eng. *simple Good-Turing discounting*; Gale & Sampson, 1995).

Pretpostavimo da je *r* frekvencija javljanja neke varijable (npr. bigrama ili trigrama) u uzorku veličine *N*, a N_r broj varijabli koje imaju tu frekvenciju.[26] Ako je

[26] Prilikom opisivanja Gud-Turingove korekcije za učestalosti, uobičajeno se za označavanje frekvencija koriste oznake kao što su *r* ili *c* (umjesto standardne oznake *f*). Ova praksa je uvedena, vjerovatno, i iz razloga da se naglasi da se radi o rangovanju frekvencija.

r malo (ili nula) tada *r/N* nije dobra procjena vjerovatnoće javljanja datih varijabli u populaciji, zbog čega je predloženo da se umjesto *r* koristi *r**, koje se računa po formuli: $r^* = (r+1)N_{r+1} / N_r$ (Good, 1953). Na primjer, za bigrame čija je frekvencija $r = 0$, dobila bi se nova vrijednost *r** (različita od nule), tako što bi broj bigrama n_1, koji su se javili jednom, podijelili sa brojem bigrama n_0, koji se nikada nisu javili. Bolja procjena se dobija ako se umjesto empirijski utvrđenih vrijednosti broja varijabli (N_r), koje imaju istu učestanost, koriste *poravnate* (eng. *smooth*) vrijednosti (Good, 1953). Jedan od načina *poravnavanja* (eng. *smoothing*) jeste korišćenje vrijednosti Z_r umjesto N_r, pri čemu je $Z_r = 2N_r / (r'' - r')$, a r'' i r' iznosi susjednih frekvencija (s lijeva i s desna) frekvencije *r*, koje imaju ne-nulte vrijednosti $N_{r'}$ i $N_{r''}$ (Gale & Sampson, 1995). Ipak, za male vrijednosti *r* bolje je koristiti N_r nego Z_r. S obzirom na to da se koriste procijenjene vrijednosti *r** ne može se očekivati da će suma ovako dobijenih proporcija biti jedan, zbog čega ih je potrebno normalizovati, tako da se dobija $p_{r\ nor.} = (1- p_0)p_r / \Sigma p_r$, za $r \geq 1$, pri čemu je $p_0 = N_1 / N$ (Gale & Samson, 1995).

Dobijene proporcije p_0 modifikovane su tako što su množene vjerovatnoćom reda i vjerovatnoćom kolone u kojima se nalaze i podijeljene sumom svih proizvoda vjerovatnoća redova i kolona koje sadrže "prazne" ćelije. Na taj način je izbjegnuto uniformno pripisivanje jedne vrijednosti svim varijablama čija je frekvencija nula za dati gramatički tip, već je ova vrijednost zavisila od ukupne frekvencije varijable (suma javljanja bigrama ili trigrama kroz sve gramatičke tipove), ali i frekvencije gramatičkog tipa (suma javljanja svih bigrama ili trigrama u jednom gramatičkom tipu). Ovaj postupak su, u ličnoj komunikaciji, predložili Milin i Moscoso del Prado Martín.

2.2.3. Uzorak za uvježbavanje i test-uzorak: Veličina uzorka za uvježbavanje iznosila je 75% ukupnog uzorka ili 969 gramatičkih kategorija, dok je test-uzorak bio veličine 25% ukupnog uzorka ili 324 gramatičke kategorije. Veličine uzorka određene su arbitrarno. Struktura uzorka za uvježbavanje i uzorka za testiranje data je u Tabeli 3.

Tabela 3. *Struktura uzorka za uvježbavanje i test-uzorka*

Vrsta riječi	N	uzorak za uvježbavanje		test-uzorak	
		n	%	n	%
imenice	74	54	5.57	20	6.17
pridjevi	131	103	10.63	28	8.64
zamjenice	594	438	45.20	156	48.15
glagoli	494	374	38.60	120	37.04
UKUPNO	1293	969	100.00	324	100.00

2.2.4. Statistička analiza: Opravdanost korišćenja fonotaktičkih informacija (bigrama i trigrama) u zadatku diskriminisanja vrsta riječi provjerena je uz pomoć mašina sa vektorima podrške *C-SVM,* sa *linearnom jezgrenom funkcijom* (*linearni kernel*). *C-SVM* je namijenjena klasifikaciji objekata, pri čemu je funkcija minimalizovanja greške: $(1/2)w^Tw + C\Sigma\xi_i$, pod uslovom da je $y_i(w^T\phi(x_i) + b) \geq 1-\xi_i$, $1 \leq i \leq n$ i $\xi_i \geq 0$ (Vapnik, 1995). Linearna jezgrena funkcija podrazumijeva podešavanja (prilagođavanja) jednog – C parametra. Ovaj parametar određuje "kaznu" za pogrešno klasifikovane objekte. On je procijenjen pomoću mrežnog pretraživanja, koje preporučuje veći broj autora (npr. Olson & Dulen, 2008; Hsu et al., 2010). Prvi korak je podrazumijevao variranje veličine *C* u rasponu $[2^{-5}, 2^{15}]$, sa povećavanjem eksponenta za dva. Kada je nađena oblast u kojoj je postignut najbolji rezultat, parametar *C* je variran u datom rasponu, sa povećanjem eksponenta za 0.25, uz korišćenje kros-validacije sa uzorkom podijeljenim na pet jednakih dijelova. Na kraju, za optimalnu vrijednost parametra *C* urađena je nova klasifikacija na uzorku podjeljenom po omjeru 75% : 25%, pri čemu je veći dio bio uzorak za uvježbavanje, a manji test-uzorak.

Zbog potpunijeg uvida u proces klasifikacije, napravljeno je i nekoliko dodatnih analiza na pojedinim matricama. Jedna od njih se odnosila na ispitivanje odnosa između tačnosti klasifikacije i broja trigrama na osnovu kojih se klasifikacija vrši. Iako prilozi nisu uključeni u analize, jer se javljaju samo u deset gramatičkih

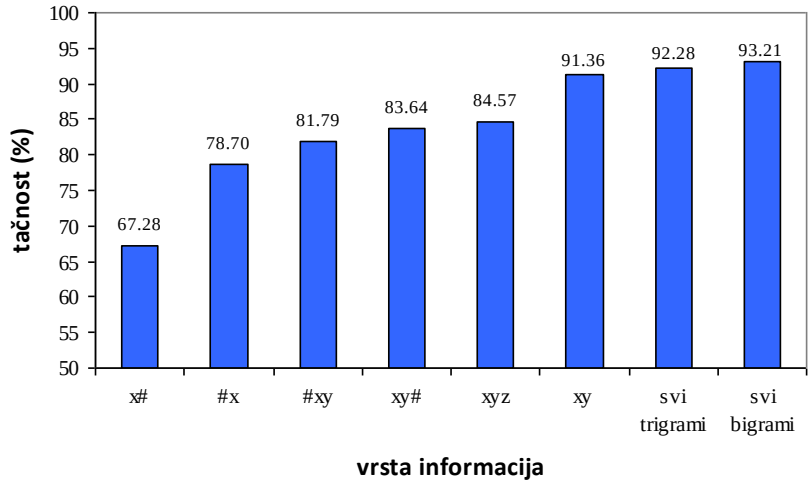
tipova, zbog čega nisu u dovoljnom broju zastupljeni u test-uzorku, provjereno je da li njihovo uvođenje, kao pete vrste riječi, otežava klasifikovanje i u kojoj mjeri.

Osim toga, radi potpunosti primijenjenih tehnika, primijenjene su i SVM sa *radijalnom jezgrenom funkcijom* (*RBF*). Optimalni parametri C i γ, na osnovu kojih je dobijen najveći procenat tačno klasifikovanih vrsta riječi, određeni su na sličan način kao i u slučaju linearne funkcije. Parametar C je sistematski variran u rasponu $[2^{-5}, 2^{15}]$, a γ u rasponu $[2^{-15}, 2^{3}]$. Nakon što je identifikovana oblast u kojoj se dobija najbolji rezultat, obavljeno je "fino" pretraživanje sa stopom linearnog povećanja parametara za 0.25. Na kraju, za optimalne vrijednosti C i γ napravljena je klasifikacija na uzorku podijeljenom po omjeru 75% : 25%, gdje je veći poduzorak bio za uvježbavanje, a manji je predstavljao test-uzorak.

2.3. Rezultati i diskusija

Mašine sa vektorima podrške implementirane su uz pomoć statističkog paketa Statistica 7.0. Značajnost razlika između procenata tačno diskriminisanih vrsta riječi za različite setove bigama i trigrama provjeravane su Mek Nemarovim testom (*McNemar*) za proporcije dobijene na istom uzorku (Sheskin, 2000). Izračunavanje ovog statistika i odgovarajućih *p* vrijednosti urađeno je pomoću statističkog okruženja R (verzija 2.13.1; R Development Core Team, 2011).

Najveći procenat tačno razdvojenih vrsta riječi, 93.21%, dobijen je u slučaju klasifikacije koja se obavljala na svim bigramima, tj. kada su u analizu bile uključene informacije o bigramima na početku [*#x*], na kraju riječi [*x#*], te informacije o vjerovatnoćama bigrama [*xy*], bez obzira na njihovu poziciju u riječi (Slika 7, Prilog 1). Za jedan procenat manje efikasna bila je diskriminacija koja se oslanjala na sve trigrame [#xz, xyz, xz#] (Slika 7, Prilog 2) odnosno za dva procenta ona koja se oslanjala samo na bigrame [xy], bez dodatnih informacija o njihovom položaju u riječima (Slika 7, Prilog 1).



Slika 7. Tačnost diskriminacije vrsta riječi na osnovu bigrama i trigrama. Oznake *x, y* i *z* se odnose na bilo koje foneme, pri čemu su moguće samo kombinacije dvije odnosno tri foneme koje se javljaju u srpskom jeziku. Kada se oznaka # nalazi ispred *x* označava početak riječi, iza *x* ili *y* označava kraj riječi.

Razlike između tri najuspješnije klasifikacije gramatičkih tipova u odgovarajuće vrste riječi nisu statistički značajne (Tabela 4).

Tabela 4. *Značajnost razlika između tri najuspješnije klasifikacije*

	McNemar's χ^2	df	p
[#x, x#, xy] : [#xy, xy#, xyz]	0.1905	1	.663
[#x, x#, xy] : [xy]	1.3889	1	.239
[#xy, xy#, xyz] : [xy]	0.1739	1	.677

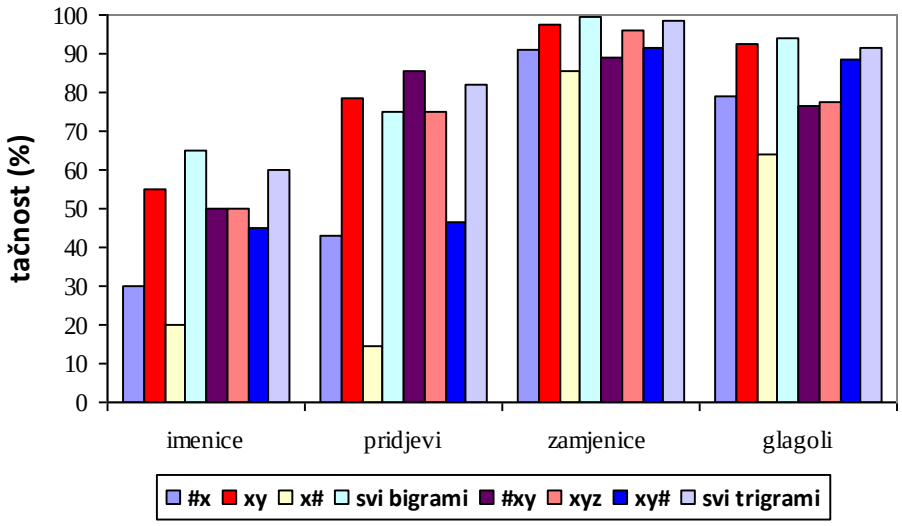
S druge strane, najmanji procenat tačno obrađenih vrsta riječi (67.28%) dobijen je na osnovu informacija o vjerovatnoći bigrama na kraju riječi [x#] (Slika 7, Prilog 1). Ovo je za oko 10% manji procenat u odnosu na klasifikacije u kojima su korišćene informacije o vjerovatnoćama bigrama na početku riječi [#x]. Razlika između ove dvije, najmanje efikasne klasifikacije, statistički je značajna (*McNemar's* χ^2 (1) = 13.6, $p < .001$).

Između najmanje efikasne i tri najuspješnije klasifikacije, pored klasifikacija na osnovu bigrama na početku riječi [#x], nalaze se i klasifikacije dobijene na

osnovu trigrama na početku [#xy] i na kraju riječi [xy#], te trigrama bez obzira na njihovu poziciju [xyz]. Tačnost ovih klasifikacija kreće se u rasponu od 78% do 85% (Slika 7, Prilog 2). Izvedeno je deset uporednih provjera značajnosti razlika između broja tačno klasifikovanih gramatičkih tipova, kada se obrada oslanjala na navedene četiri vrste fonotaktičkih informacija, te bigrame [xy], bez obzira na njihovu poziciju. Zbog većeg broja pojedinačnih poređenja korigovan je nivo α, kako bi se smanjila greška *I vrste* (greška odbacivanja tačne nulte hipoteze H_0), primjenom Benjamini-Hochberg korekcije za multipla testiranja hipoteza (Benjamini & Hochberg, 1995). Razlike između bigrama na početku riječi [#x], trigrama na početku [#xy] i na kraju riječi [xy#], te trigrami bez obzira na njihovu poziciju [xyz] nisu statistički značajne (Prilog 3).[27] Ipak, pri konačnom donošenju suda o efikasnosti klasifikacije na osnovu trigrama [xyz], treba uzeti o obzir da je za ovu vrstu fonotaktičkih informacija zadržan najmanji procenat od početnog skupa (od svih trigrama koji su se javili u uzorku). Ovo je za posljedicu moglo imati nešto manji procenat tačno klasifikovanih gramatičkih tipova nego što bi bio slučaj da je zadržan veći broj trigrama.

Veliki procenat pogrešno klasifikovanih gramatičkih tipova na osnovu bigrama i trigrama sa kraju riječi govori u prilog diskusije Bajina i sar. (Baayen et al., 2011), koji ističu da u jezicima s bogatom infleksionom morfologijom sufiksi nemaju značajnu ulogu u obradi riječi jer više gramatičkih tipova može da dijeli isti sufiks. Tako se, na primjer, infleksioni sufiks *–ima* javlja u dativu, instrumentalu i lokativu množine imenica, zamjenica, ali i dužih oblika pridjeva. To je i razlog što se, za ovu vrstu fonotaktičkih informacija, imenice i pridjevi, u najvećem broju slučajeva, mijenjaju zamjenicama, tj. vrstom riječi iz iste, imenske, grupe riječi (Prilog 1c, 2c). Ujedno, najveće greške, između 80 i 85%, utvrđene su na pridjevima i imenicama, kada se klasifikacija obavljala na osnovu bigrama na kraju riječi [x#] (Slika 8, Prilog 1).

[27] Za izračunavanje *korigovanog* α korišćen je FDR online kalkulator (http://www.sdmproject.com/utilities/?show=FDR).



Slika 8. Tačnost diskriminacije u zavisnosti od vrste riječi. Oznake *x, y* i *z* se odnose na bilo koje foneme, pri čemu su moguće samo njihove kombinacije koje se javljaju u srpskom jeziku. Kada se oznaka # nalazi ispred *x* označava početak riječi, a iza *x* ili y označava kraj riječi.

Ovo ukazuje na to da sufiksi ne mogu biti od koristi kao "nosioci" specifičnih značenja, već je njihova uloga drugačija: oni mogu biti nosioci korisne redundanse, tako što kontrolišu/smanjuju nivo neizvjesnosti i sl. O ulozi sufiksa u markiranju sintaksičkih funkcija i značenja riječi na sličan način govori i Kostić (2004) u okviru diskusije o razvoju jezičkih struktura. Usložnjavanje sistema (sve veći broj leksema i njihovih relacija) u okviru ograničenih kognitivnih kapaciteta ima za posljedicu stalnu reorganizaciju jezika, pri čemu je osnovna strategija na kojoj se bazira ta reorganizacija dijeljenje i uvođenje, prije svega, novih podklasa. Kostić ističe da ove promjene, iz ugla deskriptivne lingvistike, izgledaju nemotivisano. Međutim, sa pozicije razvoja jezika, kao dinamičkog kompleksnog sistema (više o tome u Beckner et al., 2009), značajni su procesi samoregulacije koja ne poznaje "…naše taksonomije i naše kriterijume koherentnosti", već "…teži optimalnoj distribuciji informacionog opterećenja" (Kostić, 2004, str. 51).

Najveći procenat grešaka utvrđen je na imenicama, na kojima se greška kretala u rasponu od 35% do 80%, a najmanji na zamjenicama, od 0.64% do 15% (Slika 8). Najveći procenat tačno obrađenih imenica, zamjenica i glagola dobijen je kada su u klasifikaciji korišćeni svi bigrami [#x, x#, xy], dok je najbolji rezultat za

pridjeve postignut na trigramima, kada je algoritam raspolagao informacijom o poziciji početka riječi [#xy] (Slika 8, Prilog 2). Međutim, pitanje je da li je ovakav rezultat posljedica veće diskriminativnosti pridjeva na osnovu početnih trigrama ili razlog leži u strukturi-test uzorka, s obzirom da su pridjevi zastupljeni u relativno malom broju slučajeva (8.6%), zbog čega treba biti oprezan pri izvođenju konačnog zaključka.

Radi boljeg uvida u mogućnost diskriminacije vrsta riječi na osnovu fonotaktičkih informacija, obavljeno je nekoliko dopunskih analiza. Prva takva analiza napravljena je za slučaj kada se u zadatak klasifikacije uključi još jedna vrsta riječi – prilozi, što ovaj zadatak čini kompleksnijim. Ovo je podrazumijevalo da se postojećoj matrici, u kojoj se nalaze svi bigrami [#x, x#, xy], pridruži još deset gramatičkih tipova, koji se javljaju u okviru priloga, sa pripadajućim distribucijama bigrama. Pomoću linearne jezgrene funkcije, sa parametrom $C = 0.281$, tačno je klasifikovano 92.03% vrsta riječi, pri čemu niti jedan od pet priloga koji su se našli u test-uzorku nije tačno obrađen. S obzirom na to da uključivanje nepromjenljivih vrsta riječi ne bi značajnije uticalo na tačnost, dobijeni rezultat se može tertirati kao pokazatelj tačnosti klasifikacije svih vrsta riječi na osnovu bigrama.

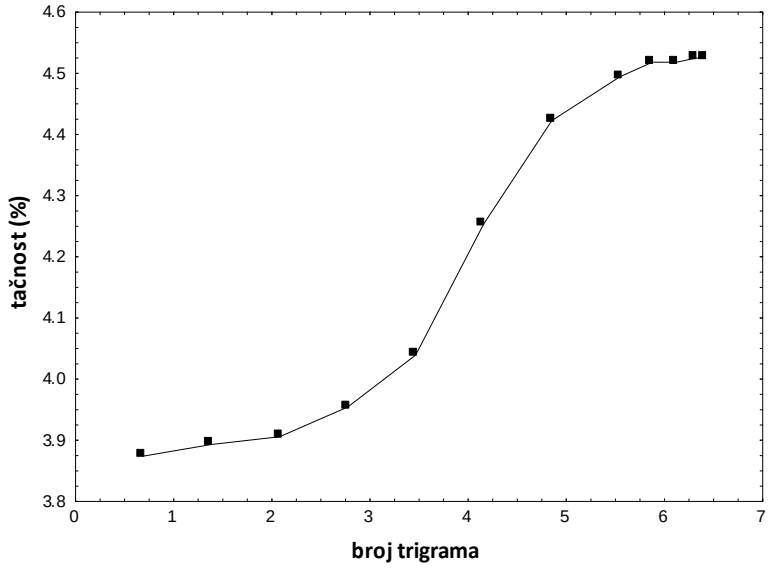
Takođe, na uzorku svih trigrama ispitano je kako se tačnost klasifikacije mijenja u funkciji veličine skupa atributa na osnovu kojih se klasifikacija vrši. S obzirom na to da je osnovna analiza obavljena na 615 trigrama, što je oko 8% svih trigrama identifikovanih u uzorku dnevne štampe Korpusa savremenog srpskog jezika (Kostić, 2001), na ovaj način se može utvrditi da li je dobijeni procenat tačno diskriminisanih vrsta riječi (92.28%), ujedno, i najveća tačnost koja se može dobiti kada se analiza oslanja na sve trigrame [#xz, xyz, xz#], tj. skup trigrama koji obuhvata trigrame na početku riječi, na kraju riječi i trigrame bez obzira na poziciju na kojoj se nalaze u riječima. Ako bi se na manjem uzorku dobio približan procenat tačne klasifikacije kao na uzorku od 615 trigrama, to bi ukazivalo da je dosegnuta maksimalna moguća tačnost i da dalje povećanje uzorka ne bi doprinijelo efikasnijoj obradi. Da bi se to provjerilo, broj trigrama je sistematski variran (Tabela 5). Početni uzorak se sastojao od dva trigrama, da bi uzorci, zatim, povećavani geometrijskom

progresijom, tj. svaki sljedeći uzorak bio je duplo veći od prethodnog, i tako do uzorka sa 256 trigrama. Od te veličine, u svakom sljedećem koraku broj trigrama je povećavan za 100, tj. naredni uzorak je imao 356, zatim 456 i, na kraju, 556 trigrama. U slučaju da je nastavljeno sa uvećanjem uzoraka geometrijskom progresijom, u rasponu između 256 i 615 trigrama bio bi formiran samo jedan uzorak veličine 512 trigrama, što ne bi bilo dovoljno za pouzdano zaključivanje o maksimalnoj mogućoj tačnosti.

Tabela 5. *Odnos broja trigrama i tačnosti diskriminacije vrsta riječi*

	Broj trigrama											
	2	4	8	16	32	64	128	256	356	456	556	615
C	0.5	0.5	0.5	1	0.281	0.625	0.625	0.125	0.375	0.531	0.125	0.125
tačnost (%)	48.15	49.07	49.69	52.16	56.79	70.37	83.33	89.51	91.67	91.67	92.28	92.28

Za ovako definisan broj trigrama, tačnost klasifikacije se kretala u rasponu od 48% do 92% tačno obrađenih vrsta riječi (Tabela 5). Kada se obje ose prevedu na logaritamsku skalu sa osnovom *e*, odnos tačnosti i broja trigrama uključenih u klasifikaciju daju relativno pravilnu sigmoidalnu funkciju (Slika 9). Na osnovu dobijenih rezultata (Slika 9) može se zaključiti da je na uzorku od 615 trigrama dosegnut gornji plato efikasnosti primijenjenog postupka na ovoj vrsti materijala. To znači da bi dalje povećanje broja trigrama vodilo, eventualno, marginalnom povećanju tačnosti. U prilog ovakvom zaključku govori nekoliko činjenica: (a) na cijelokupnom uzorku trigrama uključenim u analizu (615 trigrama) nije zabilježen porast tačnosti diskriminacije u odnosu na uzorak veličine 556 trigrama, (b) ne postoji razlika u procentu tačnosti na uzorcima veličine 356 i 456 trigrama i (c) zabilježeno je neznatno povećanje tačnosti od 0.6%, kada je broj trigrama sa 456 povećan na 556.



Slika 9. Funkcija tačnosti diskriminacije u zavisnosti od broja trigrama. Tačnost i broj trigrama prikazani su na ln-skali.

Na osnovu dobijenih rezultata može se zaključiti da svi trigrami u istoj mjeri ne doprinose tačnosti diskriminacije vrsta riječi, iz čega slijedi da nije nužno uključivanje svih mogućih fonotaktičkih informacija ovog tipa u slične analize. Zbog različite diskriminacione moći fonotaktičkih informacija, postavlja se pitanje kriterijuma za njihovu selekciju. Čini se da bi prosječna frekvencija po jednoj gramatičkoj kategoriji mogla biti adekvatna mjera za izbor bigrama/trigrama. Ipak, za donošenje konačnog zaključka potrebno je provjeriti i druge mjere (npr. samo frekvenciju ili broj gramatičkih oblika u kojima se javio određeni bigram/trigram i sl.) te njihove odgovarajuće vrijednosti za selekciju fonotaktičkih informacija. Možemo, zato, pretpostaviti da *informaciona dobit* (eng. *information gain*), tj. redukcija entropije, ali i sama entropija, mogu poslužiti u dijagnostičke svrhe: kao mjere informativnosti fonotaktičkih jedinica na osnovu kojih bi se vršio odabir onih koje mogu doprinijeti uspješnom obavljanju određenog zadatka.[28]

Pored linearne jezgrene funkcije, veoma često se koristi i preporučuje radijalni kernel (RBF), zbog čega su napravljene i dvije dodatne analize, radi

[28] Ova pitanja prevazilaze ciljeve ove studije i zbog toga nisu detaljnije razmatrana.

provjere uspješnosti ove jezgrene funkcije. Za tu svrhu su poslužili bigrami na početku riječi [#x], zbog malog broja atributa, te svi bigrami uzeti zajedno [#x, x#, xy], na kojima je u prethodnim klasifikacijama postignut najbolji rezultat (Prilog 4). Upotrebom SVM sa RBF jezgrenom funkcijom dobijena je za nekoliko procenata bolja klasifikacija od klasifikacije sa linearnim kernelom. Oslanjajući se na informacije o bigramima na početku riječi, broj tačno klasifikovanih vrsta riječi, kada su parametri $C = 32$ i $\gamma = 0.03125$ je 81.79%. Razlika između ovog rezultata i rezultata dobijenog uz pomoć linearnog kernela (78.80%) nije statistički značajna (*McNemar's* $\chi 2$ (1) = 2.7, $p = .100$). S druge strane, razlika između procenta tačno klasifikovanih riječi na svim bigramima uz pomoć RBF funkcije, koji za $C = 14.25$ i $\gamma = 0.03125$, iznosi 95.06%, i onog dobijenog linearnom jezgrenom funkcijom (93.21%) je granično statistički značajna (*McNemar's* $\chi 2$ (1) = 4.17, $p = .041$).

Rezultati dobijeni u ovoj studiji pokazali su da se na osnovu frekvencija fonotaktičkih informacija, izračunatih na nivou gramatičkih tipova, mogu uspješno odrediti vrste riječi kojima ti gramatički tipovi pripadaju. Koristeći RBF jezgrenu funkciju, najveći procenat tačno klasifikovanih gramatičkih tipova (95%) dobijen je na skupu svih bigrama. To ukazuje da postoji složaj bigrama koji dobro diskriminiše jednu vrstu promjenljivih riječi od druge. Pri tome, pokazalo se da nisu svi bigrami odnosno trigrami podjednako informativni.

Četiri najmanje efikasne klasifikacije dobijene su za slučajeve bigrama i trigrama na početku i kraju riječi. Brojni su faktori koji utiču na to da ove vrste fonotaktičkih informacija imaju slabiji potencijal za razdvajanje gramatičkih tipova. O tim faktorima, u okviru diskusije zašto morfeme, kao najmanje jezičke jedinice koje imaju značenje, nemaju značajnu ulogu u diskriminaciji riječi, govore i Bajin i saradnici (Baayen et al., 2011).

Prvo, infleksioni sufiksi se mogu javiti u više gramatičkih tipova u okviru jedne infleksione klase, s minimalnim promjenama u značenju. Na primjer, sufiks *–a*

se, kod imenica muškog roda koje označavaju živa bića, javlja u genitivu i akuzativu jednine (*dječak-a*).[29]

Drugo, postoje sufiksi koji se koriste za produkciju različitih gramatičkih tipova u okviru jedne infleksione klase, pri čemu izražavaju semantički različite stvari. Tako se sufiks *–om* koristi u tvorbi lokativa množine imenica muškog i ženskog roda koje se završavaju na *–a* (*borb-om* : *sudij-om*). Pored toga, ovaj sufiks se javlja u dativu i lokativu množine određenih oblika pridjeva muškog (*bijel-om* putu) i srednjeg roda (*širok-om* polju), te instrumentalu jednine pridjeva ženskog roda (*širok-om* ulicom).

Treće, sufiksi *–a, –o, –e, –i, –u* služe za gradnju gramatičkih tipova u različitim infleksionim klasama (više u: Petrović & Gudurić, 2010). Pored ovih sufiksa, i veće fonotaktičke kombinacije, kao što su *–im* (*drž-im* : *širok-im*) ili *–ju* (*ču-ju* : *ju* – kraći oblik akuzativa zamjenice ona), se javljaju i kod glagola, ali i u imenskoj grupi riječi.

Četvrto, pojedine morfeme, npr. *ov–* i *ev–*, upotrebljavaju se u konstrukciji različitih gramatičkih tipova, tako što se umeću između osnove i završetka (tzv. augmentacija, eng. *augmentation*). Ove dvije morfeme se mogu naći u oblicima nominativa, genitiva i akuzativa množine imenica muškog roda (npr. *rat-ov-i, rat-ov-a, rat-ov-e; zmaj-ev-i, zmaj-ev-a, zmaj-ev-e*).

Peto, prefiksi, generalno, ne mogu značajnije doprinijeti razlikovanju infleksionih oblika. Oni se nalaze na početku riječi (npr. *po-tonuti*) i prisutni su u svim inflektivnim varijantama date riječi. Osim toga, jedan prefiks može biti korišćen za tvorbu novih značenja u okviru različitih vrsta riječi (npr. imenica: *po-majka*, prilog: *po-dalje*, glagol: *po-jesti*, pridjev: *po-visok* itd.; detaljnije u: Stevanović, 1977), što dodatno umanjuje njihovu prediktivnu vrijednost u zadatku diskriminacije gramatičkih tipova.

[29] Infleksiona klasa je skup riječi koje dijele istu infleksionu paradigmu (npr. imenice koje se završavaju na *–a*). Infleksiona paradigma predstavlja skup infleksionih varijanti jedne riječi, koje se dobijaju na osnovu standardnih morfoloških transformacija (npr. infleksiona paradigma za riječ *borba* su oblici ove riječi dobijeni deklinacijom; Milin et al., 2009)

Uticaj ovih faktora odrazio se i na uspješnost obrade jezičkog materijala u ovom istraživanju. Tako, npr. kod svih vrsta riječi, najslabiji rezultat dobijen je na osnovu bigrama na kraju riječi [x#]. Ovo je naročito izraženo kod imenica i pridjeva, jer ove dvije vrste riječi dijele veliki broj zajedničkih infleksionih sufiksa. U tim slučajevima, imenice i pridjevi se, uglavnom, mijenjaju zamjenicama, koje su najtačnije obrađena vrsta riječi. Razlog za to treba tražiti u činjenici da se u slučaju zamjenica radi o zatvorenom skupu riječi, koje se distribuiraju kroz veliki broj gramatičkih tipova, što olakšava njihovu klasifikaciju.

Treba istaći da u ovoj studiji naglasak nije bio na dostizanju maksimalne moguće tačnosti, tj. gornje granice tačnosti koja se može ostvariti u zadatku diskriminacije promjenljivih vrsta riječi na osnovu fonotaktičkih informacija. Ako se ovo postavi kao cilj, potrebno je sistematski varirati broj bigrama i trigrama, kako bi se odredio njihov optimalan broj. Optimalan uzorak bigrama i trigrama bio bi najmanji uzorak bigrama i trigrama na kojem se dobija maksimalan procenat tačno klasifikovanih gramatičkih tipova. U slučaju optimalnog broja bigrama i trigrama, dalje povećanje njihovog broja ne bi dovelo do povećanja tačnosti (o problemu stabilnosti jezičkih distribucija i dostizanje optimalne veličine uzorka za različite zadatke u obradi jezika više u: Dimitrijević, Kostić, & Milin, 2009; Kostić, Ilić, i Milin, 2008).

Analize nisu rađene na svim bigramima i trigramima (osim bigrama na početku i kraju riječi), već samo na osnovu onih kod kojih je odnos ukupne frekvencije po broju gramatičkih tipova u kojima se javljaju veći od 25. Na taj način su zadržane one fonotaktičke informacije koje se ne javljaju rijetko, što je jedan od uslova efikasnije klasifikacije pomoću SVM (Nakagawa et al., 2001). Iako je vrijednost od 25 bigrama/trigrama po gramatičkom tipu određena aproksimativno na osnovu jednostavne grafičke procjene, provjera tačnosti klasifikacije u zavisnosti od broja trigrama (broj trigrama je variran od dva do 615) pokazuje da je, bar za ovu vrstu informacija, ta mjera odgovarajuća. Osim toga, i ovo ukazuje da sve fonotaktičke informacije nemaju isti značaj za obradu jezika. To je u skladu sa nalazima dobijenim na drugim jezicima, kao što su engleski i holandski. Naime,

utvrđeno je da se u 80% govora, od desetak hiljada slogova koji postoje, javlja tek oko 500 slogova (Levelt, Roelofs, & Majers, 1999). Ipak, bilo bi korisno istražiti i uspješnost pojedinih bigrama/trigrama u diskriminaciji te utvrditi podskup onih koji imaju najveću diskriminatornu moć. Na taj način bi se dobile dodatne informacije o bigramima i trigramima, kao i mogućnost da se popravi tačnost diskriminacije bazirane na fonotaktičkim informacijama.

Ovi problemi, naravno, prevazilaze ciljeve rada. Zapravo, oblast mašinskog učenja podijeljena je na dvije velike grupe (ili porodice) algoritama: diskriminativne i generativne. Diskriminativni algoritmi vrše mapiranje prediktivnih na kriterijumske varijable (koje se često nazivaju i ulazne, tj. izlazne varijable), sa ciljem klasifikacije i/ili regresije. SVM jesu najbolji, odnosno najefikasniji predstavnik diskriminativnih algoritama. S druge strane, generativni algoritmi za cilj imaju definisanje združene distribucije vjerovatnoća – $\Pr(X_1, X_2,..., X_n)$, za dati domen koji je definisan varijablama $X_1, X_2,...,X_n$. Združena funkcija vjerovatnoće, dalje, omogućava generisanje novih uzoraka podataka za dati domen. U generativne modele spadaju različiti *grafički modeli* (eng. *graphical generative models*), *Bejzove mreže* (eng. *Bayesian networks*), *Markovljeva slučajna polja* (eng. *Markov random fields*) i drugi (više o tome u: Jebara, 2004). Upravo su Dela Pietra i saradnici koristili Markovljeva slučajna polja da bi izdvojili relevantne n-grame u produkciji riječi engleskog jezika (Della Pietra, Della Pietra, & Lafferty, 1997; ovaj model kritički razmatraju i unapređuju Klimova & Rudas, 2014).

Kada se radi o odabiru alata za obavljanje zadataka iz domena automatske obrade jezika, SVM su razumljiv izbor jer predstavljaju veoma efikasan alat za klasifikaciju i predikciju, bez posebnih preduslova; na primjer, preduslova koji se tiču raspodjela/distribucija podataka na kojima se obrada vrši. Ipak, može se staviti primjedba da SVM ne reflektuju način na koji čovjek obrađuje jezik (Baayen, 2011). Iz tog razloga je, u narednoj studiji, u zadatku automatske produkcije infleksionih oblika, korišćeno učenje zasnovano na memoriji (Daelemans & Van den Bosch, 2005) – pristup za koji se može reći da odražava mogući kognitivni mehanizam obrade jezika (Baayen, 2011).

3. FONOLOŠKA SLIČNOST RIJEČI I PROBLEM PRODUKCIJE INFLEKSIONIH OBLIKA

U prethodnoj studiji ispitana je mogućnost diskriminacije promjenljivih vrsta riječi na osnovu fonotaktičkih informacija, tačnije, na osnovu distribucija bigrama i trigrama u različitim gramatičkim tipovima. Ovaj problem odnosi se na domen automatske obrade, tj. razumijevanja jezika. Međutim, može se postaviti pitanje uloge, ali i načina korišćenja ovih informacija u zadatku produkcije jezika. Iz tog razloga je, u drugoj studiji, u zadatku automatske produkcije infleksionih oblika, provjerena mogućnost primjene modela koji se oslanja na fonološku sličnost sa primjerima iz iskustva.

Infleksiona produkcija podrazumijeva transformacije u okviru jedne vrste riječi, kojima se dobijaju novi oblici riječi sa minimalnim i predvidivim promjenama u značenju (Keuleers & Sandra, 2008). Zadatak, koji je postavljen pred računarski model, sastojao se u produkciji traženog gramatičkog oblika za datu riječ: na primjer, za zadatu leksemu *apoteka*, potrebno je generisati genitiv jednine – *apoteku*. Pri tome se model oslanjao na fonotaktičke informacije iz zadnja četiri sloga osnovnog oblika (odrednice/leme) riječi. Takođe, za svaku riječ koja se obrađuje, bio je zadat i ciljani gramatički oblik, koji odgovara nekoj sintaksičkoj funkciji (npr. funkciji subjekta) i gramatičkom značenju koju ta riječ može da ima u rečenici (npr. za imenice su to značenja roda, broja i padeža).

Za izvršavanje ovog zadatka korišćeno je *učenje zasnovano na memoriji* (eng. *memory based learning – MBL*; Daelemans & Van den Bosch, 2005). Za razliku od SVM, MBL predstavlja model koji se zasniva na analogiji i čije glavne postavke odražavaju kognitivno utemeljena jezička znanja i mehanizme obrade prirodnog jezika (Baayen, 2011; Keuleers, 2008; Keuleers & Dealemans, 2007; Keuleers & Sandra, 2008; Milin et al., 2011, itd.).

U ovakvim zadacima uobičajeno je oslanjanje na slogove i njihove sastavne elemente: *nastup*/*ulaz*, *jezgro* i *rub*/*koda* (Keuleers et al., 2007). Međutim, postoje različita gledišta na strukturu sloga: prema jednom, unutar sloga ne postoje posebni

konstituenti; prema drugom, postoje tri dijela koja čine jedan slog: nastup/ulaz (eng. *onset*), jezgro (eng. *nucleus*) i rub/koda (eng. *coda*); dok se prema trećem slog sastoji od nastupa/ulaza i rime (eng. *rhyme*), koja obuhvata jezgro i rub itd. (više u: Blevins, 1995). Iako i dalje postoje kontroverze oko ovog pitanja, drugo gledište, po kome postoje tri dijela koja čine jedan slog, široko je prihvaćeno (Davis, 1988; Dell, 1986, 1988; Haugen, 1956; Hockett, 1955; Kager, 1999; Kristal, 1988; Prince & Smolensky 1993; Zec, 2000, 2007, itd). U prilog ovakve strukture sloga govore i podaci o greškama u govoru, koji ukazuju da, kada se jedan dio riječi zamijeni dijelom druge riječi, često radi o istoj poziciji u slogu, tj. jezgro se mijenja jezgrom, rub s rubom itd. (na primjer: "stick neff" umjesto "stiff neck/ukočen vrat"; Yip, 2003). Oslanjajući se na rezultate sličnih istraživanja, Del (Dell, 1986, 1988) razvija računarski model produkcije riječi, koji je zasnovan na ovakvoj, tripartitnoj, strukturi slogova.[30]

Kada se radi o takvoj, tripartitnoj, strukturi sloga, jezgro je obavezni dio i po pravilu ga čini vokal, a rjeđe silabički konsonant (eng. *syllabic consonant*; npr. p<u>r</u>st). Slogovi mogu, ali ne moraju, sadržavati periferne dijelove, tj. nastup i rub. Nastup može činiti jedan konsonant (npr. <u>p</u>as) ili konsonantski klaster (npr. <u>gl</u>as), dok rub, po pravilu, čini konsonant. Ako slog ima rub (npr. pa<u>s</u>) zove se zatvoreni slog, u suprotnom, ako se završava vokalom, tj. ako ne sadrži rub (npr. knj<u>i</u>-g<u>a</u>), onda je otvoreni slog. U ovoj studiji, korišćeni modeli je, pored informacija o posljednja četiri sloga, na raspolaganju imao i informaciju o izostanku konsonanata na početku i kraju sloga.

[30] Pored analize govornih grešaka, druga istorijska osnova savremenih studija i modela produkcije govora su studije u kojima je mjereno vrijeme reakcije potrebno da se imenuje objekat ili izgovori riječ (Levelt, 1999). Nastavak ove tradicije predstavlja tzv. WEAVER, odnosno WEAVER++ (Word-form Encoding by Activation and VERification) računarski model (više u: Roelofs, 1997; 2000). Za razliku od Delovog modela (Dell, 1986, 1988), u ovom modelu unutrašnja struktura slogova (nastup, jezgro i rub) nije unaprijed definisana.

3.1. Učenje zasnovano na memoriji

Model učenja zasnovanog na memoriji sastoji se od dvije komponente, od kojih je jedna zadužena za učenje (eng. *learning component*) i podrazumijeva skladištenje materijala u memoriji, dok je druga zadužena za produkciju (eng. *performance component*) i bazira se na sličnosti novog objekta sa objektima u memoriji (Daelemans & Van den Bosch, 2005). Da bi se opisao jedan ovakav model potrebna su najmanje tri elementa: (a) baza znanja koja sadrži egzemplare sa pridruženim klasama,[31] (b) funkcija sličnosti dva egzemplara i (c) funkcija na osnovu koje se donosi odluka o klasi novog egzemplara (Keuleers & Daelemans, 2007).

Egzemplari su, kada je riječ o učenju zasnovanom na memoriji, uskladišteni bez ikakvih transformacija, restrukturiranja ili apstrahovanja zajedničkih karakteristika. Takođe, ne postoje ni razlike u skladištenju pravilnih i nepravilnih oblika, zbog čega se na model učenja zasnovanog na memoriji može gledati kao na pristup jednog puta u obradi jezika.

Sličnost jednog objekta sa objektima u memoriji može se predstaviti nekom od *mjera sličnosti* (eng. *similarity metric*). Osnovna mjera sličnosti je *mjera preklapanja* (eng. *overlap metric*), koja predstavlja ukupnu sumu razlika između karakteristika dva objekta:[32]

$$\Delta(X, Y) = \sum\delta(x_i, y_i),\ i = 1, 2,..., n; \qquad (1)$$

gdje je $\Delta(X, Y)$ – distanca između dva objekta *X* i *Y*, sa *n* karakteristika, δ – distanca na svakoj od tih osobina, pri čemu je za numeričke vrijednosti $\delta(x_i, y_i) = |(x_i - y_i) / (max_i - min_i)|$, dok je u ostalim slučajevima $\delta(x_i, y_i) = 0$, ako je $x_i = y_i$ odnosno $\delta(x_i, y_i) = 1$, ako je $x_i \neq y_i$ (Daelemans et al., 2010).[33]

[31] Objekti u memoriji se još označavaju kao*: instance*, *primjeri*, *iskustvo* (Keuleers & Daelemans, 2007).
[32] Detaljan matematički opis algoritamskih rješenja učenja zasnovanog na memoriji dali su Delemans i Van den Boš (Daelemans & Van den Bosch, 2005) i Dalemans i sar. (Daelemans et al., 2010).
[33] Ova mjera preklapanja se kod *k*-NN algoritma naziva IB1 (Aha, Kibler, & Albert, 1991).

Mjere preklapanja daju informacije o tome da li među objektima postoje razlike. Međutim, postoje slučajevi kada objekti mogu biti manje ili više slični (npr. riječi *pas* i *par* su sličnije nego riječ *pas* i *asocijacija*), što zahtijeva i drugačije *mjere sličnosti*. Tako, npr. *Levenštejnova distanca* (Levenshtein, 1966) daje informaciju o potrebnom broju umetanja, brisanja i zamjene elemenata da bi se jedna riječ transformisala u drugu, dok se *Dajs koeficijent* (Dice, 1945) oslanja na informacije o zajedničkom broju *n*-grama u riječima koje se porede (Majumder, Mitra, & Chaudhuri, 2002; Kondrak, Marcu, & Knight, 2003).[34] Postoje i druge, sofisticiranije mjere distance, koje se koriste u slučaju nominalnih vrijednosti (Hendrickx & Van den Bosch, 2005; Daelemans et al., 2010), što i jeste najčešći slučaj u zadacima obrade jezika. Jedna od takvih mjera je MVDM (eng. *modified value difference metric;* Stanfill & Waltz, 1986; Cost & Salzberg, 1993), koja se oslanja na zajedničko javljanje (eng. *co-occurrence*) atributa i klasa:

$$\delta(v_1, v_2) = \sum |P(C_i \mid v_1) - P(C_i \mid v_2)|, \; i = 1, 2, ..., n, \qquad (2)$$

pri čemu su v_1,v_2 vrijednosti atributa, a C_i ciljana klasa. Ako se u prethodnu jednačinu uvede logaritam prosječnih iznosa zavisnih vjerovatnoća $P(C_i \mid v_1)$ i $P(C_i \mid v_2)$, dobija se *Džefrijeva divergencija* (*Jeffrey divergence*; Daelemans et al., 2010). Za razliku od MVDM, koja predstavlja geometrijsku udaljenost između dva vektora, *Džefrijeva divergencija* je simetrična varijanta Kulbak-Leiblerove distance (Kullback, 1959; Kullback & Leibler, 1951), kojom se izražava udaljenost između dvije distribucije.

U prethodnim slučajevima računa se broj (ne)podudaranja vrijednosti pojedinih karakteristika objekata koji se porede. Međutim, neke od karakteristika mogu biti informativnije za klasifikaciju od drugih, tj. mogu biti bolji prediktori.

[34] Dajs koeficijent se računa po formuli 2c / (a + b), gdje je *c* broj zajedničkih n-grama, a *a* i *b* broj n-grama u prvoj, odnosno drugoj riječi (Majumder, Mitra, & Chaudhuri, 2002). Tako, za riječi *pas* (bigrami: *p, pa, as, s*) i *par* (bigrami: *p, pa, ar, r*) *Dajs koeficijent* bi iznosio $\delta(x_i, y_i) = 2x2 / (4+4) = 0.5$; a za riječi *pas* i *asocijacija* (bigrami: *a, as, so, oc, ci, ij, ja, ac, ci, ij, ja, a*) *Dajs koeficijent* je $\delta(x_i, y_i) = 2x1 / (4+12) = 0.125$. Drugi način računanja Dajs koeficijenta je da se ne uzimaju u obzir bigrami na početku i na kraju riječi (Kondrak, Marcu, & Knight, 2003). Levenštejnova distanca za prvi primjer je 1 (zamjena *s* sa *r*), a za drugi 10.

Informacija o tome se može dobiti tako što se posmatra svaka karakteristika izolovano i mjeri njen doprinos tačnom određivanju klase kojoj objekt pripada. *Informaciona dobit* (eng. *information gain*) predstavlja očekivanu redukciju entropije, kao posljedicu podjele skupa objekata po osnovu određenog atributa:

$$w_i = H(C) - \sum P(v_i) \times H(C \mid v_i),\ v \in V_i,\ i = 1, 2,..., n, \qquad (3)$$

pri čemu je: *C* – skup oznaka za klase, *H*(C) – entropija skupa *C*, koja se računa po formuli: $H(C) = -\sum P(c_i)\log_2 P(c_i)$, $c_i \in C$, $i = 1, 2,..., n$; V_i skup vrijednosti za atribut *i*, a $H(C|v_i)$ – entropija vrijednosne distribucije različitih klasa. Pretpostavimo da postoji klasa/skup od deset elemenata: pet plavih i pet crvenih, koji se razlikuju po nekim karakteristikama (npr. veličini, obliku i sl.). Entropija početnog skupa je maksimalna (H(C) = 1), jer je podjednak broj elementa iz obje klase (podjednak broj plavih i crvenih objekata). Nakon što se, po osnovu neke karakteristike (npr. veličine: manji – veći), razdvoje elementi iz početnog skupa, računa se entropija koja je jednaka sumi ponderisane entropije klase manjih objekata i ponderisane entropije klase većih objekata (u jednačini 3 ova entropija je predstavljena izrazom $\sum P(v_i) \times H(C \mid v_i)$). Za svaku klasu, ponder $P(v_i)$ predstavlja procenat elemenata iz početnog skupa koji se našao u toj klasi. Ako, npr. postoji potpuno razdvajanje početnog skupa na plave i crvene, ponderi su 0.5. Entropije klasa, u tom slučaju, su minimalne, tj. imaju vrijednost 0 (zato što se u skupu malih objekata nalaze elementi jedne, a u skupu velikih objekata elementi druge boje), pa je i ukupna entropija 0. To bi značilo da je veličina objekata "idealna" karakteristika za klasifikaciju na plave i crvene. Može se zaključiti da, što je informaciona dobit (w_i) veća, data karakteristika je bolji prediktor, tj. da je informativnija. Informaciona dobit se računa za sve karakteristike. Međutim, informaciona dobit ima tendenciju da precjenjuje važnost karakteristika sa velikim brojem vrijednosti, zbog čega je predložena njena normalizacija (Quinlan, 1993; Daelemans et al., 2010), tako da se dobija tzv. *relativna informaciona dobit* (eng. *gain ratio*).

$$w_i = [H(C) - \sum P(v_i) \times H(C|v_i)] / s_i(i),\ v \in V_i,\ i = 1, 2,..., n, \qquad (4)$$

pri čemu $s_i(i)$ predstavlja entropiju skupa vrijednosti jednog atributa:

$$s_i(i) = -\sum P(v_i) \times \log_2 P(v_i),\ v \in V_i,\ i = 1, 2, ..., n. \quad (5)$$

Relativna informaciona dobit se može koristiti kao težinski koeficijent (ponder, eng. *weight*) u mjeri preklapanja (jednačina 1), tako da se dobija:

$$\Delta(X,Y) = w_i \sum \delta(x_i, y_i),\ i = 1, 2, \ldots, n.^{35} \quad (6)$$

Pokazano je, međutim, da je i ova mjera pristrasna prema atributima koji mogu imati veći broj vrijednosti pa je predložen način računanja težinskih faktora koji se oslanja na χ^2 statistik (White & Liu, 1994). Osim ovih, postoje i druge mjere kvaliteta ovakvih atributa, kao što su: *G statistik* (White & Liu, 1994), *Gini indeks* (Breiman, Friedman, Olsen, & Stone, 1984), *mjera udaljenosti* (eng. *distance measure*; Mantaras, 1989), *j-mjera* (eng. *j-measure*; Smyth & Goodman, 1991), *težina dokaza* (eng. *the weight of evidence*; Michie, 1989), mjera bazirana na MDL (eng. *mimial description lenght*; Kononenko, 1995) itd.

Najjednostavniji metod klasifikacije novog objekta jeste da mu se pripiše najfrekventija klasa iz skupa najbližih susjeda (eng. *majority voting method*). U slučaju da postoje dvije klase sa istim brojem "glasova", odluka o klasi kojoj pripada novi objekat može se donijeti tako što se parametar *k* uveća za jedan, što za posljedicu ima povećanje broja susjeda. Ako i dalje postoji isti broj primjera po klasama, uzima se klasa koja je frekventnija u uzorku, a ako i to nije dovoljno, može se izabrati klasa koja se prva javi u uzorku ili klasa koja se dobije slučajnim izborom. U slučajevima kada je *k* malo, a podaci *rasuti* (eng. *sparse data*), donošenje odluke o klasi postaje lokalno, uzimajući u obzir samo grupu nafrekventnijih egzemplara. Tada je postupak nepouzdan (Daelemans et al., 2010). Postoji nekoliko načina prevazilaženja ovog problema, a oni se zasnivaju na ponderisanju primjera iz skupa najbližih susjeda, u zavisnosti od udaljenosti objekata koji se klasifikuje. Jedna mogućnost je da se susjedima linearno pripišu

[35] *k*-NN algoritam, koji se oslanja na ovaj težinski koeficijent, zove se IB1-IG (Daelemans & Van den Bosch, 1992).

vrijednosti u rasponu [0, 1], pri čemu se najbližem susjedu pridružuje vrijednost jedan, a najdaljem nula (eng. *inverse-linear*; Dudani, 1976);[36] druga mogućnost je da se nađe recipročna vrijednost (eng. *inverse distance weight*; Dudani, 1976), dok se treći način, sa eksponencijalnim udaljavanjem (eng. *exponential decay function),* koji je predložio Zavrel (1997), zasniva na pretpostavci da će vjerovatnoća generalizacije opadati ekponencijalno sa distancom objekata u psihološkom prostoru (Shepard, 1987).[37] U uporednim analizama nekih od navedenih pristupa, kao najefikasniji se pokazao inverzno-linearni pristup (Zavrel, 1997; Xu, 2011).

Prethodna diskusija se odnosi na slučajeve kada su egzemplari uskladišteni u memoriji bez transformacija, kao jednodimenzionalni vektori, i kada klasifikacija novog objekta podrazumijeva "prolazak" kroz sve primjere u memoriji. Ovakav pristup može biti neekonomičan, jer zahtijeva velike memorijske kapacitete i dugo vrijeme obrade, zbog čega se vrši njegova optimizacija. Optimizacija se može napraviti na tri načina: (a) spajanjem više svojstava u apstraktnije (tj. generalnije/opštije) svojstvo; (b) uređenjem (eng. *editing*) egzemplara u memoriji i (c) (djelimičnim) reorganizovanjem egzemplara u stabla odluke (Van den Bosch, 1999).

Postoji nekoliko tehnika za generalizaciju atributa i korišćenje drugih formi za predstavljanje egzemplara u memoriji, kao što su npr. *prototipovi* (Chang, 1974), *pravila* (Domingos, 1995), *hiperpravougaonici* (eng. *hyperrectangles*; Salzberg, 1991), *hibridni modeli* (Dasarathy & Sheela, 1979; Wettschereck, 1994), *familije* (eng. *family expressions*; Van den Bosch, 1999a, 1999b) itd. Ovakvi postupci predstavljaju proširenje "klasičnog" učenja zasnovanog na memoriji (Van den Bosch, 2000).

Drugi način optimizacije zasniva se na pretpostavci da je korisno zadržati u memoriji one egzemplare koji na bilo koji način doprinose uspješnoj klasifikaciji,

[36] Postoji i modifikovana verzija, po kojoj se linearno pripisuju vrijednosti u rasponu [1/2, 1] (MacLeod, Luk, & Titterington, 1987).

[37] Pored ovih, postoje i drugi načini ponderisanja egzemplara u memoriji, npr. prema njihovoj tipičnosti (Zhang, 1992). Ponderisanje egzemplara se može koristiti, ne samo u procesu donošenja odluke, nego i u procesu određivanja mjera distance.

dok se oni, koji nemaju pozitivnu ulogu u tom procesu, mogu i odbaciti. To je moguće izvesti na dva načina: eliminacijom egzemplara čije izostavljanje neće uticati na performanse klasifikatora iz daljeg procesa klasifikacije ili eliminacijom egzemplara, kojima se u procesu klasifikacije pripisuju drugačije klase, od većine njihovih najbližih susjeda (Van den Bosch, 1999a).

Treći način optimizacije jeste predstavljanje egzemplara *stablima odlučivanja* (eng. *decision tree*), koja, osim prikaza znanja, služe i za donošenje odluka na osnovu tog znanja (Daelemans, Van den Bosch, & Weijters, 1997).

Na osnovu rezultata istraživanja, u kojima je provjeravana korisnost različitih postupaka optimizacije, neki autori smatraju da nema razloga da se koriste ovi postupci i da je, u zadacima prirodne obrade jezika, optimalna strategija oslanjanje na "čisto" učenje zasnovano na memoriji, koje podrazumijeva upotrebu IB1-IG algoritma (Van den Bosch, 1999a, 1999b, 2000).

3.2. Metod

3.2.1. Uzorak: Efikasnost modela provjerena je na uzorku koji je sadržavao 89024 različitih oblika riječi. Riječi u uzorku su pripadale grupi promjenljivih riječi (imenice, pridjevi, zamjenice i glagoli), a uključeni su i prilozi, koji mogu biti promjenljivi, kada se radi o komparaciji.[38] Uzorak je formiran na osnovu *Frekvencijskog rečnika dnevne štampe*, koji je dio *Frekvencijskog rečnika savremenog srpskog jezika* (Kostić, 1999), tako što su iz ovog poduzorka uklonjeni svi nestandardni oblici riječi, kao što su npr. riječi sa crticom, brojevima i sl. Za svaki oblik riječi, koji je ujedno i ciljani oblik koji se želi dobiti primjenom modela, postojala je informacija o lemi (osnovnom obliku riječi) i šifra. Ova šifra je precizno određivala gramatički status ciljanog oblika. Jedna riječ se mogla javiti samo jednom u određenom gramatičkom obliku. Ilustracija korišćenih informacija za formiranje uzorka, za tri slučajno odabrana oblika riječi, data je u Tabeli 6.

[38] Prilozi za način, količinu i neki prilozi za mjesto imaju, pored pozitiva, i komparativ i superlativ (Stanojčić & Popović, 1997).

Tabela 6. *Informacije iz* Korpusa savremenog srpskog jezika, *koje su korišćene za formiranje uzorka*

Lema	Oblik riječi	Šifra*
.....	.....	.....
tekst	tekstom	100611
pogodan	pogodne	201212
igrati	igram	521110
.....	.....	.....

*Šifra 100611 označava imenicu muškog roda u instrumentalu jednine; šifra 201212 označava pozitiv pridjeva ženskog roda u genitivu jednine, a šifra 521110 glagol u prezentu prvog lica jednine.

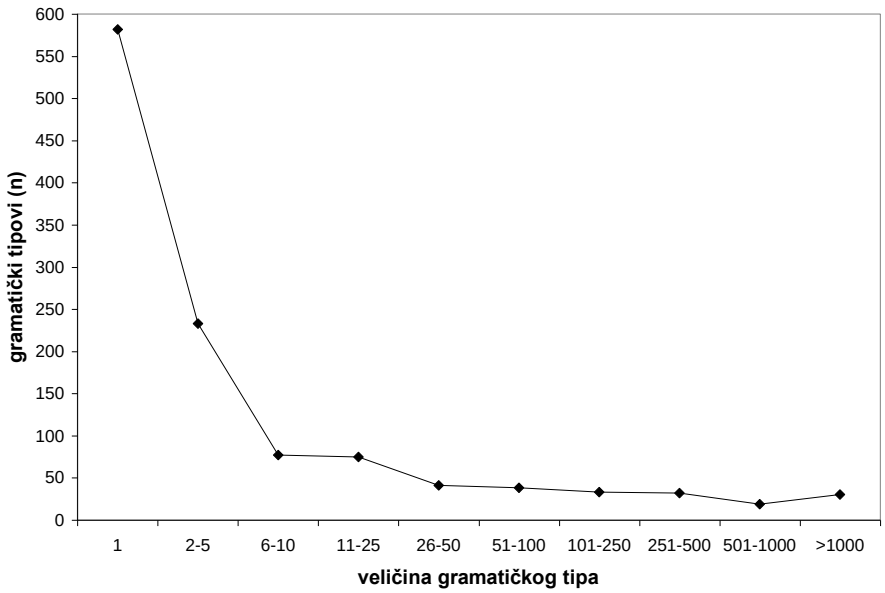
Riječi u uzorku su distribuirane u 1160 različitih gramatičkih tipova. Ovako veliki broj gramatičkih tipova je posljedica razvijene infleksione morfologije srpskog jezika. Tako, na primjer, samo kombinovanje roda, broja i padeža kod imenica daje 42 različita morfološka atributa ove vrste riječi.

U Tabeli 7 data je struktura uzorka prema broju slučajeva za svaku vrstu riječi.

Tabela 7. *Distribucija oblika riječi u uzorku u zavisnosti od vrste riječi*

Vrsta riječi	Oblici riječi (n)	Procenat	Gramatički tipovi (n)
imenice	41062	46.12	70
pridjevi	25189	28.30	117
zamjenice	906	1.02	534
glagoli	20618	23.16	434
prilozi	1249	1.40	5
Ukupno	89024	100.00	1160

Slika 10 prikazuje distribuciju gramatičkih tipova u zavisnosti od njihove veličine, tj. broja različitih riječi koje su se javile u odgovarajućem gramatičkom obliku.



Slika 10. Distribucija veličina gramatičkih tipova.

Veličine gramatičkih tipova kretale su se u rasponu od jedan do 6783, koliko je primjera bilo za imenice muškog roda u nominativu jednine. Druga po veličini je kategorija imenica muškog roda u genitivu jednine sa 3318 egzemplara; zatim slijede imenice ženskog roda u nominativu jednine sa 2720, imenice muškog roda u akuzativu jednine sa 2263, imenice ženskog roda u genitivu jednine sa 2145 egzemplara itd.

Od 1160 gramatičkih tipova polovina je sadržavala jednu riječ, a 20% (233 gramatička tipa) od dvije do pet. Veličine ostalih 30% (345) gramatičkih tipova varirale su od šest, do 6783 različitih riječi.

3.2.2. Instrument: Model namijenjen produkciji infleksionih oblika srpskog jezika implementiran je korišćenjem *TiMBL* softvera – *Tilburg memory based learner* (Daelemans et al., 2010), koji omogućava primjenu nekoliko različitih algoritama učenja zasnovanih na memoriji, te klasifikaciju na osnovu *metode najbližih susjeda* (eng. *k-nearest neighbor classification*), prilagođenu zadacima koji se obavljaju na jezičkom materijalu.

Prilikom implementacije modela, kao mjera sličnosti objekta, korišćena je MVDM metrika (eng. *modified value difference metric*; Cost & Salzberg, 1993). Ova mjera obezbjeđuje informaciju o stepenu sličnosti atributa tako što posmatra zajedničko javljanje (eng. *co-occurrence*) vrijednosti atributa i ciljnih klasa. Simulacija sa MVDM metrikom je izvedena za slučaj kada je $k = 7$. Za razliku od k-NN algoritma, gdje se *k* odnosi na broj najbližih susjeda, kod učenju zasnovanog na memoriji ova vrijednost se odnosi na *k*-tu najbližu distancu (Daelemans & Van den Bosch, 2005; Daelemans et al., 2010). Zbog toga, broj egzemplara u skupu najbližih susjeda može biti i veći od sedam. Ova vrijednost je uzeta iz razloga što se 7-NN model pokazao robusnijim od drugih modela u simulaciji rezultata dobijenih u produkciji množine imenica holandskog jezika i prošlog vremena za *pseudo-glagole* (eng. *novel forms*) engleskog jezika, (Keuleers, 2008; Keuleers & Dealemans, 2007; Keuleers & Sandra, 2008), ali i zadatku generisanja alomorfa u srpskom jeziku (Milin et al., 2011). Doprinos pojedinih atributa procesu donošenja odluke o novom objektu iskazana je pomoću koeficijenta relativne informacione dobiti (Quinlan, 1993; Daelemans et al., 2010).

Prilikom implementacije korišćena je varijanta modela u svom najjednostavnijem, osnovnom obliku. To znači da je svaki od *k* egzemplara iz skupa najbližih susjeda imao istu težinu (eng. *zero decay weighting*), tj. oni nisu dodatno ponderisani u zavisnosti od stepena sličnosti (veličine distance) sa objektom koji se klasifikuje. Nije vršena ni dodatna optimizacija procesa, u smislu poboljšanja performansi izvođenja, tj. nisu uvođene restrikcije, kao što je, na primjer, predstavljanje egzemplara u memoriji apstraktnijim svojstvima (npr. prototipovima) ili selekcija egzemplara na osnovu njihove korisnosti, koje bi model činile kompleksnijim.

3.2.3. Procedura: U prvom koraku formirani su egzemplari[39] koji predstavljaju znanje pohranjeno u memoriji, neophodno za automatsku jezičku produkciju.

Uobičajen postupak poravnanja egzemplara (eng. *alignment method*), kojim se postiže da svi egzemplari u memoriji imaju isti broj elemenata, oslanja se na slogove i njihove elemente: nastup/ulaz, jezgro i rub/koda (Keuleers et al., 2007). Prilikom dekompozicije leme na slogove označeni su slučajevi kada slog nije imao konsonantski početak ili završetak. Odsustvo konsonanta na početku ili kraju sloga označeno je sa =. Ovaj znak je, takođe, korišćen da se označi izostanak fonema na određenoj poziciji u vektoru. Svaki član vektora bio je odvojen zarezom od elemenata koji ga okružuju.

Egzemplari su formirani za svaku riječ iz uzorka, a predstavljeni su jednodimenzionim vektorima sa 14 elemenata. U prvih dvanaest ćelija vektora smještena su posljednja četiri sloga, dobijena dekompozicijom leme; na trinaestom mjestu nalazi se šifra, tj. detaljna gramatička specifikacija riječi, a na četrnaestom mjestu *infleksiona klasa.* U Tabeli 8 dat je primjer strukture tri egzemplara.

Tabela 8. *Struktura egzemplara*

Lema	Oblik riječi	Egzemplari
.....	.....	.....
polagan	polaganih	=, =, =, p, o, =, l, a, =, g, a, n, 201221, 9876543210ih
poseban	posebnih	= ,=, =, p, o, =, s, e, =, b, a, n, 201221, 987654320ih
opasan	opasnih	=, =, =, =, o, =, p, a, =, s, a, n, 201221, 987654320ih
.....	.....	.....

Infleksiona klasa predstavlja način na koji se dolazi do traženog oblika riječi. Slovima je označen infleksioni nastavak, dok se brojevi od nula do devet, odnose na foneme iz osnovnog oblika riječi, tako što je nulom označen posljednji, jedinicom

[39] Termin *egzemplar* se prevodi kao: primjerak, uzorak, pojedinačni otisak, jedan jedini primjerak u zbirci (Vujaklija, 1970). S obzirom na to da u uzorku nije moguće da se jedna riječ više puta javi u jednom obliku, ovaj termin dobro opisuje objekte u uzorku.

pretposljednji fonem itd. Na primjer, ispravan oblik genitiva množine muškog roda za pridjev *polagan* dobija se primjenom infleksione klase *9876543210ih*, što znači da se na osnovni oblik riječi dodaje nastavak *ih*, tj. polagan + *ih* = polaganih. Izostanak neke od cifara u okviru infleksione klase znači da gramatički oblik ne sadrži fonemu koja se u osnovnom obliku riječi nalazila na mjestu označenom tim brojem. Na primjer, ispravan oblik genitiva množine muškog roda za pridjev *opasan* dobija se primjenom infleksione klase *987654320ih*. Dakle, iz osnovnog oblika riječi izostavljena je fonema na pretposljednjem mjestu, a zatim je dodat nastavak *ih*: opas(a)n = opasn + *ih* = opasnih.

Prilikom testiranja modela korišćen je postupak *izostavljanja jednog primjera* (eng. *leave-one-out*). U simulaciji, lema i oblik riječi (prve dvije kolone u Tabeli 8) nisu uzimane u obzir.

Pretpostavimo da se obrađuje riječ *polagan*, a da je zadatak da se produkuje genitiv množine muškog roda, tj. traženi oblik je *polaganih*. Informacija od koje model polazi je informacija o fonološkoj strukturi riječi i informacija o gramatičkom obliku koji se testira:

=, =, =, p, o, =, l, a, =, g, a, n, 201221

U prvoj fazi modeliranja formira se skup susjeda, maksimalne udaljenosti *k*, što, u ovom slučaju, predstavlja broj dozvoljenih transformacija u odnosu na testiranu riječ. Ovo se postiže kroz sljedeće korake:

1. Izračunaju se pojedinačne distance (broj transformacija) između riječi koja se obrađuje i ostalih riječi u memoriji, na osnovu fonoloških karakteristika i infleksionog koda;
2. Sortiraju se svi slučajevi u memoriji prema veličini distance (neke od riječi mogu imati istu distancu);
3. Oni slučajevi koji imaju distancu *k* ili manju čine skup susjeda testiranog oblika riječi (K). U tom skupu susjeda najmanji broj mogućih slučajeva je *k*, a može se desiti da skup ima i više elemenata. Neka ukupan broj susjeda nosi oznaku K.

Pretpostavimo da je dobijen sljedeći skup susjeda:

=,	=,	=,	p,	o,	=,	s,	e,	=,	b,	a,	n,	201221,	987654320ih
=,	=,	=,	p,	o,	=,	m,	e,	=,	š,	a,	n,	201221,	9876543210ih
=,	=,	=,	=,	o,	=,	t,	e,	=,	r,	a,	n,	201221,	9876543210ih
=,	=,	=,	i,	=,	z,	l,	a,	=,	g,	a,	n,	201221,	9876543210ih
=,	=,	p,	r,	i,	=,	k,	a,	=,	z,	a,	n,	201221,	9876543210ih
=,	=,	=,	=,	o,	=,	p,	a,	=,	s,	a,	n,	201221,	987654320ih
=,	=,	=,	n,	e,	=,	d,	a,	=,	v,	a,	n,	201221,	987654320ih

Sljedeća faza podrazumijeva donošenje odluke o infleksionoj klasi za traženi oblik. Ovo se postiže tako što se izračunaju vjerovatnoće javljanja svake infleksione klase u skupu najbližih susjeda (f/K), pri čemu je infleksiona klasa sa najvećom vjerovatnoćom očekivana klasa za testirani oblik

Za pretpostavljeni primjer, model bi trebao da primijeni infleksionu klasu *9876543210ih*, čija je vjerovatnoća u skupu susjeda 4/7. S obzirom na to da je ovo najfrekventnija klasa u skupu susjeda (klasa *987654320ih* se javila u tri od sedam slučajeva) simulirani oblik tražene riječi će biti *polaganih* (klasa *9876543210ih*, uslovno rečeno, predstavlja sljedeću operaciju: "na lemu dodaj infleksioni sufiks *ih*"), što, u ovom slučaju, daje tačan oblik.

Tabela 9. *Raspoložive informacije nakon primjene* učenja zasnovanog na memoriji *u zadatku automatske produkcije infleksionih oblika*

Lema	Šifra	*p* testiranog oblika	Testirani oblik	*p* simuliranog oblika	Simulirani oblik	Tačnost (0-tačno; 1-netačno)
...	...	...	...	...	...	...
dopisnik	100311	1	dopisniku	1	dopisniku	0
svetac	100721	0	svecima	0.428	svetcima	1
apoteka	100212	0.875	apoteke	0.875	apoteke	0
pravilan	201422	1	pravilne	1	pravilne	0
radioamater	100411	0	radioamatera	1	radioamater	1
...	...	...	...	...	...	...

U Tabeli 9 dat je dio izlazne matrice, koja je sadržavala sedam kolona i 89024 redova. U prvoj koloni se nalazila lema, u drugoj šifra, u trećoj *vjerovatnoća testiranog oblika* u skupu susjeda, u četvrtoj traženi oblik riječi (oblik koji predstavlja tačno rješenje u simulaciji), u petoj *vjerovatnoća simuliranog oblika* u skupu susjeda, u šestoj simulirani oblik riječi (oblik koji je rezultat primjene modela), a u sedmoj tačnost rješenja (0 – tačno, 1– netačno).

Vjerovatnoća testiranog oblika predstavlja vjerovatnoću javljanja infleksione klase u skupu susjeda, koju model treba da primijeni da bi došao do tačnog rješenja. Vjerovatnoća simuliranog oblika predstavlja vjerovatnoću najučestalije klase u skupu susjeda, koju model tretira kao tačno rješenje.

3.3. Rezultati i diskusija

Razultati dobijeni u zadatku automatske produkcije infleksionih oblika riječi srpskog jezika primjenom učenja zasnovanog na memoriji prikazani su u dva dijela. Uspješnost ovog pristupa prvo je proanalizirana u zavisnosti od vrste riječi, broja egzemplara po gramatičkom tipu, te materijala koji se obrađuje. Zatim su rezultati predstavljeni za svaku vrstu riječi posebno, zavisno od gramatičkih tipova u okviru tih vrsta riječi, npr. padeža za imenice ili vremena za glagole. Ovakvom detaljnijom analizom, dobio se dublji uvid u faktore koji utiču na tačnost automatske produkcije infleksionih oblika.

3.3.1. Uspješnosti učenja zasnovanog na memoriji u obradi srpskog jezika

Oslanjajući se isključivo na fonotaktičke informacije i informaciju o gramatičkom statusu riječi, u zadatku infleksione produkcije, učenje zasnovano na memoriji uspješno generiše 89% riječi (Tabela 10). Ovaj procenat se može smatrati i većim, s obzirom na to da u uzorku nisu markirani slučajevi kada su dva oblika ravnopravna, tj. tačna (npr. dubletni oblici imenica *vukovi – vuci*, kraći ili duži oblici

pridjeva *plav – plavi*, priloga *kad – kada*; zamjenica *kog – koga*, ekavica i ijekavica *cvet – cvijet* itd), te postojanje eventualnih grešaka u jezičkom uzorku. Na slučajnom poduzroku od 537 riječi koje su pogrešno obrađene (oko 5.5% svih pogrešno obrađenih riječi) utvrđeno je oko 29% ovakvih slučajeva. Ako se u obzir uzme ova korekcija, procenat uspješno produkovanih infleksionih oblika kreće se oko 92.2%.

Tabela 10. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika, zavisno od vrsta riječi*

Vrsta riječi	Tačno		Total
	n	%	
imenice	38190	93.01	41062
pridjevi	21820	86.63	25189
zamjenice	549	60.60	906
glagoli	17519	84.97	20618
prilozi	1188	95.12	1249
UKUPNO	79266	89.04	89024

Najveći broj grešaka model je pravio na zamjenicama (Tabela 10).[40] Ova vrsta riječi se, s obzirom na broj gramatičkih tipova i broj egzemplara/primjera, značajno razlikovala od ostalih. U uzorku zamjenica bilo 906 primjera distribuiranih u 534 različita gramatička tipa, što u prosjeku daje 1.7 primjera po tipu. Za razliku od zamjenica, prosječan broj primjera po gramatičkom tipu kod imenica je 586.6, kod pridjeva 215.3 i kod glagola 47.5.

Pretpostavka da veličina gramatičkog tipa, tj. broj raspoloživih primjera u okviru jednog gramatičkog tipa, utiče na uspješnost infleksione produkcije, provjerena je na imenicama, pridjevima i glagolima. Iz analize su isključene zamjenice, jer se kod ove vrste riječi broj egzemplara po gramatičkom tipu kretao u rasponu od 1 do 11, te prilozi, u okviru kojih je bilo samo pet gramatičkih tipova.

[40] Zbog specifičnosti svake vrste riječi treba biti veoma oprezan pri poređenju uspješnosti modela na različitim vrstama riječi.

Izuzev slučaja gramatičkih tipova sa jednim egzemplarom kod imenica,[41] kod kojih je ostvaren stopostotni učinak, najveći procenat tačnosti dobijen je za tipove sa više od 100 egzemplara (Tabela 11). Može se primijetiti opadanje procenta tačne produkcije za gramatičke tipove sa preko 1000 elemenata. Međutim, u uzorku je bilo tek 30 gramatičkih tipova (2.6%) koji su imali više od 1000 egzemplara pa ovaj rezultat treba uzeti s rezervom.

Tabela 11. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod različitih vrsta riječi, u zavisnosti od veličine gramatičkog tipa*

Veličina gramatičkog tipa (broj egzemplara po tipu)	Imenice (%)	Glagoli (%)	Pridjevi (%)
1	100.00	79.63	50.00
2-5	88.89	79.34	59.38
6-10	82.76	58.01	49.65
11-25	84.54	65.17	47.88
26-50	82.11	75.13	48.80
51-100	79.68	80.92	57.22
101-250	92.59	81.96	91.23
251-500	92.68	88.38	90.04
501-1000	94.92	95.15	90.96
>1000	92.94	88.10	86.82

S druge strane, zamjenice su pogodne za detaljniju analizu odnosa tačnosti i veličine gramatičkih tipova, kada broj egzemplara po tipu ne prelazi 10, tj. kada se radi o malom broju primjera za učenje. Naime, u okviru ove vrste riječi postoji više gramatičkih tipova sa istim brojem elemenata u rasponu od 0 do 10. Najlošiji rezultat je dobijen kada je veličina gramatičkog tipa jedan, nakon čega, s porastom

[41] Kod imenica su 23 gramatička tipa imala deset i manje egzemplara, od čega je bilo sedam tipova samo sa jednim egzemplarom.

broja egzemplara po tipu, dolazi do značajnog povećanja efikasnosti modela (Tabela 12).[42]

Tabela 12. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod zamjenica, u zavisnosti od veličine gramatičkog tipa*

Veličina gramatičkog tipa (broj egzemplara po tipu)	Gramatički tip (N)	% tačno klasifikovanih oblika
1	407	38.08
2	43	76.74
3	29	74.71
4	19	89.47
5	10	78.00
6	6	83.33
7	8	76.79
8	5	80.00
9	3	92.59
10	3	73.33
11	1	45.45
	534	60.60

Rezultati pokazuju da efikasnost u zadatku automatske produkcije infleksione morfologije zavisi od vrste riječi (Tabela 10) i veličine gramatičkih tipova (Tabele 11 i 12). Može se pretpostaviti da, pored veličina gramatičkog tipa i vrste riječi, tačnost zavisi i od vrste gramatičkog tipa koji se obrađuje. Da bi se o tome donio nedvosmislen zaključak, potrebno je u okviru jedne vrste riječi provjeriti efikasnost modela na različitom materijalu, kada je veličina gramatičkog tipa konstantna. Za takvu analizu su, ponovo, bile pogodne zamjenice, zbog velikog broja tipova sa samo jednim egzemplarom. Dobijeni rezultati potvrđuju da, bez obzira što je veličina gramatičkog tipa fiksna, uspješnost varira zavisno od vrste

[42] Kako je postojao samo jedan gramatički tip sa 11 egzemplara (pokazne zamjenice muškog roda u aoristu jednine), dobijeni procenat tačno produkovanih oblika nije pouzdan pokazatelj uspješnosti modela na gramatičkom tipu ove veličine.

zamjenica (Tabela 13). Ovakav zaključak dodatno potkrepljuju i rezultati dobijeni na većim gramatičkim tipovima. Tako je, na primjer, na imenicama ženskog roda u dativu jednine (453 egzemplara) ostvarena tačnost od 92.7%, dok je na imenicama ženskog roda u instrumentalu množine (451 egzemplar) tačnost infleksione produkcije iznosila 98%. Slično, kod imenica srednjeg roda u genitivu jednine (1149 egzemplara) tačno je produkovano 94,5% infleksionih oblika, dok je kod imenica muškog roda u akuzativu množine (1143 egzemplara) to postignut u 86.5% slučajeva.

Tabela 13. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod različitih gramatičkih tipova zamjenica sa po jednim egzemplarom*

Oblik zamjenice	Gramatički tip (N)	% tačno klasifikovanih oblika
lične	54	9.26
prisvojne	261	31.80
pokazne	3	33.33
odnosne	12	83.33
upitne	17	88.24
neodređene	14	64.29
odrične	26	84.62
opšte	20	50.00
	407	38.08

U okviru prethodnih analiza identifikovana su dva opšta faktora koji utiču na efikasnost infleksione produkcije, i to: vrsta materijala koji se obrađuje (vrste riječi i gramatički tipovi) te veličina gramatičkih tipova, tj. broj primjera u okviru jednog gramatičkog tipa koje je algoritam imao na raspolaganju za učenje.

Kako bi se dobio detaljniji uvid u faktore koji utiču na tačnost automatske produkcije infleksionih oblika uz pomoć MBL, analiza je napravljena posebno za svaku od promjenljivih vrsta riječi.

3.3.2. Uspješnost učenja zasnovanog na memoriji u zavisnosti od vrste riječi

Prilozi: Najveći procenat tačne klasifikacije ostvaren je na prilozima, preko 95% (Tabela 10). Potrebno je, međutim, istaći da je u okviru priloga bilo skoro 90% gramatičkih tipova koji su se odnosili na oblike priloga koji nemaju komparaciju (nepromjenljivi oblici riječi) ili pozitiv, a da je uspješnost primijenjenog modela na ovim oblicima bila preko 97%. Na preostalim oblicima priloga (komparativ i superlativ), kojih je u uzorku bilo manje od 10%, model je pravio znatno veću grešku, koja je iznosila oko 25%.

Imenice: Kod imenica, model je najslabiji uspjeh imao kod vokativa, generalno (oko 65% tačno produkovanih infleksionih nastavaka), pri čemu je najlošiji rezultat postignut kod vokativa jednine (oko 55%) i vokativa ženskog roda (61%). Sličan rezultat dobijen je i za oblike muškog (65%) i srednjeg roda (67%; Tabela 14). S druge strane, model je pravio tek 10% greške na oblicima množine imenica u vokativu.

Tabela 14. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod imenica, zavisno od gramatičkog broja i gramatičkog roda*

PADEŽ	Uspješnost (%)	BROJ		ROD			Ukupno egzemplara
		Jednina	Množina	Muški rod	Ženski rod	Srednji rod	
Nominativ	97.28	98.83	91.24	96.92	97.98	97.61	13347
Genitiv	92.00	94.25	86.99	90.77	93.60	92.62	9494
Dativ	93.01	92.78	93.54	91.57	94.23	96.00	2159
Akuzativ	88.72	87.57	91.58	78.41	97.26	96.97	7581
Vokativ	64.29	55.78	89.80	65.25	61.54	66.67	196
Instrumental	92.09	92.32	91.59	86.97	96.76	95.69	3666
Lokativ	91.69	92.26	90.33	89.20	92.13	96.17	4619
	93.01	94.02	90.26	90.63	95.59	95.75	41062

Postoji više faktora koji su mogli uticati na ovakav rezultat: (a) mali broj egzemplara, (b) javljanje fonoloških alternacija (palatalizacija, gubljenje samoglasnika *a*), (c) postojanje dubletnih oblika, (d) različiti nastavci za njegovo formiranje itd.[43]

Akuzativ je drugi padež na kojem je model imao uspješnost ispod 90%. Najveći izazov za model, u okviru akuzativa, imali su oblici muškog roda (tačnost od 78% u odnosu na 97% kod imenica ženskog i srednjeg roda). Pri tome je greška za imenice muškog roda u akuzativu jednine iznosila oko 26%, dok je greška za isti padež u množini bila oko 13%. U najvećem broju slučajeva greške su nastale zbog različitih nastavaka za imenice koje označavaju živa bića (*student-a*) i imenice koje označavaju nežive stvari (*računar-ø*). Naime, u okviru akuzativa javlja se genitivno-akuzativni i nominativno-akuzativni sinkretizam (Blagus Bartolec, 2006): kod imenica muškog roda, koje označavaju nešto neživo, akuzativ je jednak nominativu, dok je kod imenica koje označavaju živa bića akuzativ jednak genitivu. Zbog toga je, na primjer, kao rezultat obrade za akuzativ dobijen oblik *poskok* umjesto *poskoka*, *pravca* umjesto *pravac* itd. Međutim, postoje brojni izuzeci od spomenutog pravila (npr. kod naziva biljaka, zbirnih imenica koje označavaju bića (*čopor-ø*), imena predmeta i stvari koja su izvedena od njihovih pronalazača (*cepelin-ø*), naziva geografskih mjesta (*Sveti Stefan-ø*) itd; Blagus Bartolec, 2006; Stanojčić i Popović, 1997), što dodatno otežava produkciju.

Problemi u produkciji akuzativa množine imenica muškog roda javljaju se i zbog postojanja duge množine, koja se tvori dodavanjem infiksa *–ov* ili infiksa *–ev* i odgovarajućeg nastavka za akuzativ (na primjer, model je pogriješio na imenici *sin* (izlaz je glasio *sine* umjesto *sinove*), imenici *prst* (*prstove* umjesto *prste*) itd.).

[43] Otvoreno je pitanje da li bi i u kojoj mjeri informacije o prozodiji (akcentu i dužini ili kratkoći samoglasnika; Stanojčić i Popović, 1997) pomogle u produkciji infleksione morfologije. Naime, modeli jezičke produkcije (Caramazza, 1997; Dell, 1986; Levelt et al., 1999, itd.) pretpostavljaju da generisanje riječi obuhvata nekoliko kognitivnih procesa. Jedan od procesa je i *kodovanje riječi* (eng. *word form encoding*), koje obuhvata *fonološko kodovanje* i *metrički okvir riječi* (eng. *metrical frame of a word*). Dalje, metrika riječi podrazumijeva najmanje dvije vrste informacija (Schiller, 2006) – informacije o broju fonotaktičkih segmenata (u ovom slučaju slogova) i mjestu naglaska u riječi (Dell, Chang, & Griffin, 1999; Levelt et al., 1999). Zbog toga bi bilo interesenatno vidjeti da li bi u identičnom zadatku iz ove studije, ako bi jedna od karakteristika grupisanja riječi u skup najbližih susjeda bile i informacije o mjestu i vrsti akcenta, došlo do efikasnije produkcije infleksionih oblika.

Umetanje ovih infiksa ispred odgovarajućeg nastavka za padež, tj. augmentacija, karakteristična je za padeže imenica muškog roda u množini. Pored akuzativa, javlja se i u nominativu i genitivu množine. U svim padežima u kojima se pojavljuje, augmentacija je značajan izvor grešaka (npr. kod nominativa: *rojovi* umjesto *rojevi*, *prestupovi* umjesto *prestupi*, *štiti* umjesto *štitovi* itd.; kod genitiva: *brojova* umjesto *brojeva*, *mravova* umjesto *mrava*; *roka* umjesto *rokova* i sl.).

Oblici kod kojih je greška iznad 10% su još i genitiv množine (86%), te instrumental i lokativ muškog roda. U značajnom broju slučajeva ove greške su posljedica jezičkih fenomena karakterističnih za taj gramatički tip, kao što je pojava alomorfije[44] u instrumentalu jednine imenica muškog roda (npr. pogrešno je dobijen oblik *ciljom* umjesto *ciljem* i sl.), ili gore spomenute augmentacije kod genitiva množine. Pored ovih, postoje i druge greške, kao što je npr. izostavljanje slova (*seljkom* umjesto *seljakom*), problem sa nepostojanim *a* (*kadarom* umjesto *kadrom*) ili kombinacija više faktora (*smeštjom* umjesto *smeštajem*) itd.[45]

Kao ilustracija zadataka koji mogu predstavljati izazov u procesu automatske produkcije infleksionih oblika u okviru jednog gramatičkog tipa, može poslužiti nominativ množine imenica muškog roda. Na ovom obliku imenica model je postigao tačnost od 87%. U uzorku od 1490 slučajeva, bilo je 30 imenica muškog roda koje tvore nominativ množine dodavanjem nastavka (infleksionog sufiksa) *–i* sa umetkom (infiksom) *–ev* i 142 imenica koje tvore množinu dodavanjem nastavka *–i* i umetka *–ov*. Ostali slučajevi su se odnosili na imenice čija množina se dobija sa nastavkom *–i*.

MBL je u 90% pogrešno klasifikovao oblike imenica muškog roda koji tvore množinu sa nastavkom *–ev+i*. U okviru ovih grešaka model je pogrešno pripisao nastavak *–i* u 11 slučajeva (npr. *kursi* umjesto *kursevi*), a u 16 slučajeva nastavak *–ov-i* (npr. *žuljovi, lešovi* umjesto *žuljevi*, *leševi* itd.).

[44] Više o alomorfiji u srpskom jeziku u Jovanović i sar. (Jovanović, Filipović Đurđević, & Milin, 2008).
[45] U okviru imenica, s obzirom na njihovo značenje, model je pravio grešku od oko 15% na onim imenicama koje oblicima množine označavaju pojedinačne predmete (pluralia tantum).

Greška na imenicama čija se množina dobija uz pomoć nastavka *–ov+i* iznosila je oko 23%. U okviru ovih grešaka oko 42% otpada na oblike riječi u kojima se pri tvorbi nominativa množine javljaju fonološke alternacije (uglavnom nepostojano *a*). S druge strane, oblici riječi sa fonološkim alternacijama su u tačnim rješenjima prisutni tek u oko 3% slučajeva. U gotovo svim slučajevima pogrešno obrađenih imenica muškog roda koje tvore množinu sa nastavakom *–ov+i* greška se ogleda u pokušaju da se traženi oblik napravi sa infleksionim sufiksom *–i* (npr. *ritami*, *pojami* umjesto *ritmovi*, *pojmovi* itd.).

Kod netačno obrađenih imenica muškog roda, kod kojih se množina pravi dodavanjem infiksa *–i* (134 oblika), u jednom slučaju pogrešno je dodat nastavak *–evi* (*konjevi*), a u 28 slučajeva (oko 21%) nastavak *–ovi* (npr. *pasovi*, *mravovi* itd.). Ostale greške vezane su, uglavnom, za glasovne alternacije koje se javljaju pri tvorbi množine, kao što su: nepostojano *a* (*pucanji* umjesto *pucnji*), palatalizacija (*hirurgi*, *potoki* umjesto *hirurzi*, *potoci*) ili njihova kombinacija (*čuperaki* umjesto *čuperci*) itd.

Ovi primjeri jasno pokazuju da fonološke alternacije predstavljaju izazov u zadatku automatske produkcije infleksione morfologije pomoću TiMBL-a. Takođe, vidljivo je da model ima tendenciju da sebi "olakša" obradu, koristeći frekventnije nastavke za formiranje traženog oblika. Ovo je u skladu sa zaključkom da jezička produkcija bar jednim dijelom zavisi od frekvencije gramatičkog oblika (eng. *type frequency*; Bybee, 2001, 2010). Po Bajbijevoj, što je viša frekvencija gramatičkog oblika, veća je vjerovatnoća da će on biti primijenjen na nove slučajeve.

Pridjevi: Kod pridjeva, slično kao kod imenica, najveće greške su bile ponovo kod vokativa i akuzativa, s tim da je uspješnost modela na akuzativu pridjeva manja nego kod imenica i iznosi oko 83% (Tabela 15).

I na svim ostalim kategorijama pridjeva dobijeni su slabiji rezultati nego kod imenica. Razlog za to leži u činjenici da se većina pridjeva javlja u dva oblika – kratkom (neodređeni vid) i dugom (određeni vid pridjeva), npr. *hrabar – hrabri*, a u uzorku je bio markiran samo jedan oblik. Takođe, ovu vrstu riječi karakteriše javljanje pokretnih samoglasnika (*širokim – širokima* poljima), koji su potpuno

ravnopravni (Stanojčić i Popović, 1997). Tako npr. ako je u uzorku naveden kratki oblik pridjeva *mlad*, a kao rješenje dobijen dugi oblik *mladi*, algoritam je to bilježio kao grešku.[46]

Tabela 15. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod pridjeva, zavisno od stepena poređenja*

PADEŽ	Uspješnost (%)	KOMPARACIJA			Ukupno egzemplara
		Pozitiv	Komparativ	Superlativ	
Nominativ	84.86	87.16	67.62	40.71	7218
Genitiv	89.07	91.60	46.63	34.18	6479
Dativ	89.04	91.95	40.00	31.82	1168
Akuzativ	82.87	84.64	63.19	38.21	4973
Vokativ	66.67	66.67	-	-	30
Instrumental	88.52	90.73	52.63	50.91	2300
Lokativ	89.61	91.92	61.43	30.38	3021
	86.63	88.91	59.28	38.01	25189

Ovakve nedosljednosti pri obradi dešavale su se samo kod pridjeva u pozitivu, jer komparativ i superlativ karakteriše pridjevska promjena; oni imaju oblike određenog pridjevskog vida, tj. duže oblike (Stanojčić i Popović, 1997). U okviru grešaka u pozitivu, u oko 37.5% slučajeva produkovani oblik može se smatrati tačnim rješenjem, jer odgovara dužem ili kraćem obliku pridjeva, sa ili bez pokretnog samoglasnika. Ovakvi slučajevi su bili karakteristični za pridjeve u muškom rodu (oko 96%), rijetko su se javljali u srednjem (oko 4%), a nije ih bilo kod pridjeva ženskog roda. Naime, kod pridjeva ženskog roda padežni oblici određenog i neodređenog vida se ne razlikuju; pokretni samoglasnik može da se javi u dativu, instrumentalu i lokativu množine. Velika razlika između broja slučajeva koji su se javili kod pridjeva muškog i srednjeg roda, u značajnoj mjeri je odraz strukture uzorka. U uzorku pridjeva u pozitivu bilo je 43.3% pridjeva muškog,

[46] Slična situacija se može javiti kod priloga *kad – kada*; te zamjenica *kog – koga* itd.

41.4% ženskog i 15.3% srednjeg roda.[47] Osim toga, promjena pridjeva muškog roda je kompleksnija u odnosu na pridjeve srednjeg roda, što je, takođe, uticalo na veći broj grešaka. Tako, na primjer, pridjevi muškog roda imaju dva nastavka za nominativ jednine (*širok – široki* put), dok je kod pridjeva srednjeg roda isti nastavak, ali različit naglasak (*širòko* polje – *šìrok-ō* polje). Slično je i sa akuzativom jednine, koji se kod pridjeva muškog roda javlja u dva oblika, zavisno da li se imenica označava živa bića ili nežive stvari. Kada se u obzir uzme i neodređeni i određeni vid pridjeva, dobijaju se četiri moguća nastavka za građenje akuzativa jednine pridjeva muškog roda: *–ø, –a, –i, –og* (*lijep – lijepi* put; *lijepa – lijepog* konja itd).

Uspješnost obrade pridjeva u pozitivu raste sa 88.9% na 93.1%, ako se kao tačna rješenja računaju i slučajevi koji su označeni kao pogrešni (37.5%), zato što ne odgovaraju markiranom obliku u uzorku, iako postoji više alternativnih rješenja za traženi oblik. Tada je procenat tačno obrađenih pridjeva gotovo isti kao procenat tačno obrađenih imenica. Takođe, tačnost na cijelom uzorku pridjeva, uključujući komparativ i superlativ, porasla bi sa 86.6% na 90.5% automatski tačno generisanih infleksionih oblika.

Što se tiče komparacije pridjeva, najveći procenat grešaka dobijen je na superlativu, preko 60%, dok se kod komparativa greška kretala oko 40%.[48] Kod određenog broja pridjeva, kao što su: *jak, brz, dug* itd., pri poređenju se javlja jotovanje pa su za ove pridjeve oblici komparativa: *jači, brži, duži*. Na ovakvim pridjevima model je u produkciji komparativa pogriješio 113 puta, što čini oko trećine svih grešaka (34%), a tačno ih je obradio u 44 slučaja. Sličan odnos je i kod pridjeva čiji se komparativ pravi umetanjem nastavka *–š* (*lak, lijep* i *mek*), koji su tačno obrađeni u šest, a pogrešno u 16 slučajeva. Kod pridjeva koji imaju supletivne oblike komparativa (*dobar, zao, mali, velik*) približno je jednak broj tačnih (32 oblika) i pogrešnih rješenja (34 oblika).

[47] U okviru grešaka dobijenih na pozitivu, 1563 pogrešno obrađenih pridjeva su muškog roda (oko 15% svih pridjeva muškog roda u pozitivu), 356 srednjeg (oko 10% svih pridjeva u srednjem rodu) i 710 ženskog roda (oko 7% svih pridjeva u ženskom rodu).
[48] U poduzorku pridjeva bilo je 94.16% oblika pozitiva, 3.21% komparativa i 2.63% superlativa.

U oko 60% netačno obrađenih oblika superlativa, pored ostalih grešaka, javljala se i greška izostavljanja ili stavljanja na pogrešno mjesto prefiksa *naj-*, npr. *znajdraviji*, *ponajgodnije*, *gnajnusnije* (umjesto: *najzdraviji*, *najpogodnije*, *najgnusnije*). Očigledno, način reprezentovanja znanja u memoriji pomoću egzemplara koji sadrže informacije o posljednja četiri sloga (a ne cijeloj riječi) nije odgovarajući za obradu superlativa. Ovo je posljedica načina tvorbe superlativa, koji se gradi dodavanjem prefiksa *naj–* na komparativ. Zbog toga je potrebno da u memoriji budu uskladišteni i oblici superlativa, a ne samo osnovni oblici pridjeva, predstavljeni nominativom jednine u muškom rodu, npr. *slobodan*, *lijep*, *crven* i sl. Ovakav način skladištenja je nužan, kako bi model imao adekvatne primjere za učenje.

Glagoli: Ukupna greška dobijena na glagolima je nešto veća nego kod pridjeva i iznosi 15% (Tabela 16).

Tabela 16. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod glagola, zavisno od glagolskih oblika*

Oblik glagola	% tačno klasifikovanih oblika	Ukupno egzemplara
infinitiv	99.05	3174
prezent	75.39	6161
aorist	85.42	576
imperfekat	38.46	13
futur	95.77	638
imperativ	65.89	302
radni pridjev	94.09	7229
trpni pridjev	60.05	1762
glagolski prilozi	73.79	763
	84.97	20618

Najslabiji rezultat je dobijen u slučaju imperfekta, ali je za ovu gramatičku kategoriju u uzoku bilo tek 13 egzemplara distribuiranih u šest klasa. Kada se

izuzme imperfekt, uslovno bi se mogle napraviti četiri kategorije, u zavisnosti od tačnosti klasifikacije. Najslabiji uspjeh postignut je kod glagolskih oblika trpnog pridjeva i imperativa, i kreće se od 60 do 65%. Nešto bolji rezultat dobijen je na oblicima glagolskih priloga i prezentu (oko 75%). U sljedeću kategoriju bi se mogli svrstati oblici aorista (oko 85%), dok je model bio najuspješniji kod oblika radnog pridjeva, futura i infinitiva.

Zadržimo se na razlici između glagolskih pridjeva, radnog i trpnog, koji su prosti glagolski oblici (sastoje se od jedne riječi), čija je osnovna funkcija građenje složenih gramatičkih oblika. Više faktora je moglo uticati da se javi razlika u tačnosti automatske produkcije infleksionih oblika između ova dva oblika glagola.

Prvo, postoji velika razlika u broju egzemplara u okviru ovih gramatičkih kategorija. U uzroku je bilo 7229 oblika glagolskog pridjeva radnog i 1762 oblika glagolskog pridjeva trpnog.

Drugo, dok se glagolski pridjev radni gradi dodavanjem jednog tipa nastavaka na infinitivnu osnovu: *–o, –la, –lo, –li, –le, –la,* glagolski pridjev trpni pravi se od infinitivne i prezentske osnove dodavanjem trojakih nastavaka: (a) *–n, –na, –no, –ni, -ne, –na*, (b) *–en, –ena, –eno, –eni, –ene, –ena*, (c) *–t, –ta, –to, –ti, –te, –ta* (Stevanović, 1975).

Treće, kod trpnog glagolskog pridjeva postoji veći broj, uslovno rečeno, složenijih fonoloških alternacija, kao što su: jotovanje (npr. *naseliti – naseljen, premjestiti – premješten*) i umetanje samoglasnika *–v* ili *–j* (*čuti – čuven*; *ispiti – ispijen*). S druge strane, kod glagolskog pridjeva radnog, u slučaju da se infinitvna osnova završava suglasnikom, kod oblika muškog roda u jednini javlja se nepostojano *a* (npr. *tresti – tresao*).

Zamjenice: Već je istaknuto da je najlošiji rezultat ostvaren na zamjenicama. Razlog za to vjerovatno leži u činjenici da gramatički tipovi u okviru ove vrste riječi nisu imali više od 11 elemenata, te da nešto manje od 40% svih tipova ima samo jedan egzemplar. Najmanji procenat tačne infleksione produkcije, od svih posmatranih (pot)kategorija u okviru ove vrste riječi, dobijen je kod ličnih zamjenica

(Tabela 17).[49] U okviru ličnih zamjenica bilo je 54 različitih oblika. Model je tačno obradio oblike *ja* (prvo lice u nominativu jednine), *on* (treće lice u nominativu jednine) i oblik *njima*, za dativ množine sva tri roda.

Tabela 17. *Uspješnost* učenja zasnovanog na memoriji *u produkciji infleksionih oblika kod zamjenica*

Oblik zamjenica	% tačno klasifikovanih oblika	Ukupno egzemplara
lične	9.26	54
prisvojne	31.80	261
pokazne	76.19	231
odnosne	76.09	92
upitne	94.59	37
neodređene	90.38	104
odrične	75.00	48
opšte	63.29	79
	60.60	906

Postoji nekoliko faktora koji su uticali na ovako veliku grešku u obradi ličnih zamjenica. Pretpostavimo da se traži nominativ množine prvog lica, tj. oblik *mi*. Algoritam će uzeti u obzir sve egzemplare koje se odnose na zamjenicu *ja*, jer ona predstavlja osnovni oblik za prvo lice, iz kojeg se izvode ostali oblici. U odnosu na polazni egzemplar =,=,=,=,=,=,=,=,=,j,a,411120 (šifra označava oblik lične zamjenice u nominativu množine) ovi egzemplari se razlikuju za jednu informaciju, tj. šifru koja predstavlja detaljnu gramatičku specifikaciju riječi. Međutim, pored ovih egzemplara, algoritam će uključiti i sve riječi čija je distanca manja od sedam jedinica u odnosu na polazni egzemplar, kao što su, npr. riječi koje na kraju imaju *ja*: *arija*, *boja*, *sudija* i mnoge druge. Može se zaključiti da niti jedan egzemplar sa zamjenicom *ja* nije adekvatan za traženi oblik *mi* (ni u jednom slučaju ne postoji

[49] Osim toga, uspješnost koja je znatno ispod prosjeka utvrđena je i kod prisvojnih zamjenica (Tabela 18).

infleksiona klasa *98765432mi* koja znači: posljednje i pretposljednje slovo zamijeni sa *mi*, a predstavljala bi tačno rješenje). Osim toga, postoji veliki broj riječi koje su "dovoljno blizu" zamjenici *ja*, koje je algoritam uzeo u obzir prilikom donošenja odluke o produkciji nominativa množine prvog lica.

U našem primjeru, u najvećem broju slučajeva obuhvaćene su riječi kod kojih se radi o osnovnom obliku, npr. nominativu jednine (npr. =,=,=,=,=,=,=,b,o,=,j,a,100112,9876543210 i sl.). To dalje znači da je algoritam tačnim rješenjem smatrao osnovni oblik (tj. oblik bez promjene, na što upućuje infleksiona klasa 9876543210 – koja znači: ništa ne mijenjaj u odnosu na lemu) pa smo tako, umjesto *mi*, kao rješenje dobili *ja*. Iz tog razloga je model tačno obradio oblike *ja* i *on*.

Povećanje uzorka, kao i smanjenje vrijednosti parametra *k* (broja dozvoljenih transformacija), ne bi značajnije uticalo na efikasnost produkcije infleksionih oblika. U oba slučaja, u skupu susjeda ponovo bi najbrojniji bili slučajevi u kojima se traži osnovni oblik pa bi i rješenja bila slična. Jedno od mogućih rješenje jeste da u memoriji budu reprezentovani oblici u nominativu za tri lica i dva broja (*ja*, *ti*, *mi*, *vi*, *on*, *ona*, *ono*, *oni*, *one*).

U ovoj studiji identifikovano je nekoliko faktora koji utiču na uspješnost učenja zasnovanog na memoriji u zadatku automatske produkcije infleksionih oblika. Različiti procenat tačno produkovanih oblika dobijen je na različitim vrstama riječi, kao i na različitim gramatičkim tipovima unutar vrsta riječi. Kolika će uspješnost u produkciji u okviru jedne gramatičke kategorije biti, zavisi od broja infleksionih nastavaka za tvorbu traženog oblika, kao i broja fonoloških alternacija koje se javljaju pri tvorbi. Postoji zakoniti odnos: broj nastavaka i mogućih alternacija obrnuto je srazmjeran tačnosti.

Broj primjera u okviru gramatičkog tipa (veličina gramatičkog tipa) takođe predstavlja faktor koji utiče na uspješnost infleksione produkcije. Na osnovu rezultata može se zaključiti da je tačnost produkcije infleksionih oblika veća što je veći gramatički tip. Drugim riječima, što je veći broj primjera za učenje u okviru

nekog gramatičkog tipa, to se tačnije produkuju infleksioni oblici. Ovakav zaključak je opravdan, iako je utvrđeno da za gramatičke kategorije sa preko 1000 egzemplara dolazi do pada tačnosti. Razlog za to je taj što se u skupu gramatičkih kategorija sa više od 1000 egzemplara nalaze kategorije kao što su genitiv i akuzativ imenica i pridjeva, u okviru kojih se mogu javiti ravnopravni infleksioni oblici, koji u analizi nisu uzimani kao ispavni (u uzorku je markiran samo jedan oblik kao tačan). Pored toga, u automatskoj produkciji infleksionih oblika genitiv i akuzativ su bili među najzahtjevnijim padežima za obradu. Na kraju, ne treba ni zaboraviti da je u uzorku bilo tek 30 gramatičkih tipova (2.6%) koji su imali više od 1000 egzemplara, što može uticati na pouzdanost zaključivanja.

Uzorak u ovoj studiji je formiran na osnovu raspoloživog *Frekvencijskog rečnika dnevne štampe*, koji je dio *Frekvencijskog rečnika savremenog srpskog jezika* (Kostić, 1999). Uzorak je činilo 28.971 različitih riječi, koje su se javile u 89.024 različita oblika. Ovakva struktura uzorka je, svakako, uticala i na krajnji procenat uspješne produkcije, koji bi, vjerovatno, bio nešto veći da je uzorak bio veći. Ipak, to nije bilo od presudnog značaja za kvalitet rezultata i izvedenih zaključaka.

Prvo, cilj ove studije nije bio praktične prirode – postizanje maksimalne efikasnosti tagera u zadatku automatske produkcije infleksionih oblika pomoću MBL-a, već provjera načelne mogućnosti da se taj zadatak obavi na osnovu fonotaktičkih informacija, kao i identifikovanje faktora koji imaju ulogu u tom zadatku.

Drugo, i dalje bi ostao problem reprezentacije primjera u memoriji, koji se očitovao kod superlativa i zamjenica.

Treće, povećanjem uzorka povećao bi se broj primjera po gramatičkom tipu, ali se ne bi značajnije promijenili odnosi među gramatičkim tipovima, s obzirom na njihovu veličinu. Model bi i dalje imao tendenciju da koristi frekventnije nastavke za formiranje traženog oblika. S obzirom na to da se obrada odvijala na cijelom uzorku, a ne pojedinačnim, izolovanim, gramatičkim kategorijama, i konačan rezultat bi, vjerovatno, bio sličan.

Četvrto, na sličnim zadacima, pokazano je da je već na malim uzorcima (nekoliko hiljada egzemplara) uspješnost MBL-a relativno visoka (Daelemans et al, 1996). Ipak, ne treba zaboraviti da, što je uzorak na kojem se algoritam obučava veći, uticaj vrijednosti različitih parametara, ali i vrste algoritma (vrste mašinskog učenja) manji (Banko & Brill, 2001), a dobijeni rezultati pouzdaniji (Daelemans et al, 1996).

4. ESPERIMENTALNA PROVJERA KOGNITIVNE VJERODOSTOJNOSTI FONOTAKTIČKIH INFORMACIJA I UČENJA ZASNOVANOG NA MEMORIJI

U prethodnim studijama ispitana je mogućnost korišćenja fonotaktičkih informacija u zadacima automatske obrade vrsta riječi i produkcije infleksionih oblika. U prvoj studiji, diskriminacija promjenljivih vrsta riječi obavljena je uz pomoć mašina sa vektorima podrške (Vapnik, 1995, 1998), dok je u drugoj studiji, za produkciju infleksionih oblika riječi korišćeno učenje zasnovano na memoriji – MBL (Daelemans & Van den Bosch, 2005). Za razliku od SVM, koje predstavljaju matematički alat i za koje ne važi pretpostavka o psihološkoj relevantnosti, u osnovi MBL-a leži jasan mehanizam koji bi mogao biti kognitivno plauzabilan (vjerodostojan); tj. mogao bi odražavati način kognitivne organizacije jezičkih informacija i procese prisutne pri obradi prirodnog jezika (Baayen, 2011; Keuleers, 2008; Keuleers & Dealemans, 2007; Keuleers & Sandra, 2008; Milin et al., 2011, itd.).

Upravo zato što MBL možda odražava način kognitivnog funkcionisanja, na osnovu rezultata dobijenih u zadatku automatske produkciji infleksionih oblika postavljen je eksperiment, kako bi se provjerilo da li stimulusi na kojima je MBL pravio grešku predstavljaju veći izazov i u kognitivnoj obradi, tj. razumijevanju, u odnosu na stimuluse koje je tačno obradio.

Ako se pokaže da su iste imenice zahtjevnije za obradu i za MBL (pogrešno generisan infleksioni oblik) i za subjekte (duže vrijeme reakcije, veći broj grešaka), mogu se izvesti dva zaključka: (1) to direktno govori u prilog teze da je MBL kognitivno vjerodostojan model, tj. da se obrada infleksione morfologije može obavljati na osnovu analogije sa primjerima u memoriji; (2) s obzirom na to da se MBL oslanjao na fonotaktičke informacije u zadatku automatske produkcije infleksionih oblika, to, indirektno, ukazuje i na kognitivnu relevantnost tih informacija. To bi, dalje, značilo da su fonotaktičke informacije dovoljno

informativne da se pri obradi morfološki složenih riječi naš kognitivni sistem može osloniti na njih.

Provjera kognitivne vjerodostojnosti modela i informacija na koje se on oslanjao obavljena je pomoću zadatka vizuelne leksičke odluke. Iako bi direktna analogija sa zadatkom iz prethodne studije bio zadatak produkcije zadatog infleksionog oblika za datu riječ, bilo da subjekti daju odgovor u pisanoj ili usmenoj formi, zadatak leksičke odluke je adekvatna procedura za dobijanje odgovora na postavljeni problem, koja omogućava generalizaciju dobijenih rezultata. Naime, modeli zasnovani na analogiji predstavljaju modele predikcije ponašanja, generalno, a samim tim i modele jezičkog ponašanja (Skousen, 2002b, 2009), bilo da je riječ o produkciji ili razumijevanju jezičkog materijala. Dakle, može se izvesti pretpostavka kako bi model učenja zasnovan na analogiji obavio zadatak razumijevanja (koji bi bio nalik zadatku leksičke odluke). Prvo, model bi za zadati infleksioni oblik najprije pronašao skup najbližih susjeda, oslanjajući se na isti tip informacija kao i u zadatku produkcije. Drugo, najfrekventniji gramatički tip u skupu najbližih susjeda bio bi pripisan zadatom obliku. S obzirom na to da su mehanizam/način obrade i vrsta informacija isti u oba zadatka, ukoliko nalazi dobijeni u studiji modeliranja infleksione produkcije budu i eksperimentalno potvrđeni u zadatku precepcije/razumijevanja (leksička odluka), generalizacija bi imala empirijsko uporište. Konačno, ovakvo kombinovanje istraživačkih postupaka omogućava izvođenje širih implikaciji o kognitivnoj vjerodstojnosti učenja zasnovanog na memoriji, kao i samih fonotaktičkih informacija.

Za stimuluse izabran je skup nasumično odabranih imenica muškog roda u nominativu množine. Ove imenice izabrane su iz dva razloga. Prvo, postojanje više nastavaka, kao i nekoliko alternacija koje se dešavaju pri tvorbi ovog oblika, zadatak produkcije čine zahtjevnijim. Naime, nominativ množine imenica muškog roda tvori se na tri načina: dodavanjem nastavka za oblik *–i* (npr. mrav – mrav-*i*), dodavanjem infiksa *–ov* i nastavka *–i* (npr. top – top-*ov*-*i*) i dodavanjem infiksa *–ev* i nastavaka *–i*

(npr. mač – mač-*ev*-*i*).[50] Prilikom tvorbe nominativa množine mogu se javiti fonološke alternacije: zadnjonepčani suglasnici *k, g, h,* ispred *i* prelaze u *c, z, s* (pašnjak – pašnjaci, hirurg – hirurzi, tepih – tepisi), sonant *l* u vokal *o* (čitalac – čitaoci) te nepostojano *a* (orkestar – orkestri). Drugo, imenice muškog roda odabrane su i zato što one čine dovoljno veliki skup potencijalnih stimulusa za psiholingvistički eksperiment. U uzorku koji je korišćen u analizi uspješnosti učenja zasnovanog na memoriji, bilo je 1490 imenica muškog roda u nominativu množine, pri čemu je, primjenom učenja zasnovanog na memoriji, tačno obrađeno 1296 (87%), a pogrešno 194 imenica (ili 13%).

4.1. Metod

4.1.1. Nacrt: Eksperiment je bio jednofaktorski. Faktor, *tip klastera*, imao je dva nivoa: (a) imenice u nominativu množine muškog roda koje je MBL pogrešno obradio i (b) imenice u nominativu množine muškog roda kojima je MBL pripisao tačan infleksioni oblik, tj. ispravno ih obradio. Faktor je bio ponovljen po subjektima.

Pored tipa klastera, postojao je i dodatni faktor koji se odnosio na leksikalnost stimulusa, a koji je uveden zbog eksperimentalne kontrole. Formiran je odgovarajući skup pseudoriječi koje su korišćene samo za kontrolu i nisu bile od značaja za ispitivani problem. Ovaj faktor nije uključen u analizu vremena reakcije i analizu grešaka.

S obzirom na to da se relativno mali broj stimulusa nalazio u klasteru pogrešnih rješenja (194 imenice), a da je ujednačavanje stimulusa u klasterima obavljeno po nekoliko osnova (vrste nastavka, broja alternacija, dužine riječi i sl.) stimulusi nisu ujednačeni po frekvencijama oblika i odrednice. Zbog toga su učestalost oblika riječi muškog roda u nominativu množine i varijansa učestalosti

[50] Neki oblici mogu imati i dubletne oblike: *vukovi* i *vuci*. Takođe, kod nekih imenica oblik množine se gradi dodavanjem nastavka na okrnjenu osnovu: *čobanin – čobani* (Stanojčić i Popović, 1997).

leme, koja se ne može pripisati varijansi oblika nominativa (rezidualna varijansa), tretirane kao kontinuirani prediktori.[51]

Rezidualna varijansa je korišćena zbog visoke korelacije učestalosti oblika i odrednice $r = .658$, $p < .01$. Nakon određivanja reziduala, korelacija između učestalosti odrednice i novodobijene varijable – rezidualne varijanse iznosila je $r = .627$, $p < .001$. Ovo je dalo potvrdu da je nova varijabla zadržala visoku i pozitivnu korelaciju s originalnom varijablom (frekvencijom leme), a da istovremeno više nije bila u problematičnoj korelaciji sa frekvencijom oblika, koja je bila glavna kovarijabla u istraživanju.

Postupak ovakvog dekoreliranja uobičajen je u psiholingvističkim istraživanjima, a koristi se kako bi se dobio "nemaskirani" pojedinačni efekat prediktora, u slučajevima kada je korelacija između njih veća od 0.3 (Bürki & Gaskell, 2012) ili .05 (Kuperman, Bertram, & Baayen, 2008). Iako postupuak rezidualizacije prediktora ne rješava u potpunosti problem multikolinearnosti (više u: Wurm & Fisicaro, 2014), u ovom slučaju je ipak predstavljao optimalni postupak prilagođavanja varijabli preduslovima analize linearnih modela. Naime, dok učestalost oblika riječi ima facilitirajući efekat, učestalost leme može imati različit uticaj na obradu riječi (Milin et al., 2009), zbog čega je potrebno ispitati njihove pojedinačne efekte. Iz tog razloga, problem visoke linearne povezanosti među prediktorima, u ovom slučaju, nije bilo umjesno rješavati postupcima kao što su: analiza glavnih komponenti radi postizanja ortogonalnog odnosa među prediktorima (više u: Baayen et al, 2006; Wurm & Fisicaro, 2014),[52] izostavljanje jednog prediktora iz dalje analize, pravljenje novog prediktora kombinovanjem postojećih itd. (više u: Tabachnick & Fidell, 2007; Wurm & Fisicaro, 2014, itd.).[53]

[51] Učestalost odrednice predstavlja sumu učestalosti oblika za datu odrednicu. Tako, npr. za odrednicu *front*, učestalost odrednice je zbir pojavljivanja ove riječi u svim padežnim oblicima, dok se učestalost oblika odnosi samo na broj javljanja u nominativu množine.

[52] S obzirom na to da se radi o dva prediktora i visokoj korelaciji među njima, došlo bi do agregacije prediktora na prvoj glavnoj komponenti, pa ne bi mogao biti ispitan njihov pojedinačni uticaj.

[53] Centriranje podataka (oduzimanje aritmetičke sredine od pojedinačnih vrijednosti) nije dovelo do smanjenja povezanosti između frekvencije oblika i frekvencije odrednice.

Zavisne varijable u nacrtu su bile: (a) vrijeme reakcije, definisano kao vrijeme proteklo od prikazvanja stimulusa do davanja odgovora u zadatku vizuelne leksičke odluke, izraženo u milisekundama i (b) broj grešaka.

4.1.2. Stimulusi: U eksperimentu je korišćeno 240 stimulusa, 120 riječi (imenice muškog roda u nominativu množine) i 120 pseudoriječi. Iz svakog od klastera (klaster tačnih i klaster pogrešnih rješenja) izabrano je po 60 riječi. Struktura stimulusa u svakom od klastera je bila sljedeća: 43 stimulusa su bili oblici koji se završavaju na infleksioni sufiks *–i* (71.67%), a 17 stimulusa (28.33%) oblici koji nominativ množine tvore sa nastavkom *–ov+i*.

U oba klastera polovina stimulusa su imenice, kod kojih se u nominativu množine javljaju jedna ili dvije alternacije, kao na primjer: *krivac – krivci* (nepostojano *a*) ili *primjerak – primjerci* (*k* ispred *i* u *c*, nepostojano *a*). Nominativ množine ovih imenica dobija se dodavanjem nastavka *–i*. Isti tip alternacija bio je jednak u obje grupe stimulusa.

Stimulusi u klasterima su bili ujednačeni i po broju slova i slogova. Prosječan broj slova oblika nominativa množine u klasteru pogrešnih rješenja iznosio je 7.18, a prosječan broj slogova 3.23. U klasteru tačnih rješenja, prosječan broj slova bio je 7.12, dok je prosječan broj slogova bio 3.17.

Kako klasteri nisu mogli biti ujednačeni i po frekvenciji oblika i odrednice, ovi prediktori korišćeni su kao kontrolni kovarijati u statističkoj obradi. Prosječna frekvencija oblika za klaster pogrešnih rješenja iznosila je 22.77, a prosječna frekvencija odrednice 199.5. Za klaster tačnih rješenja frekvencija oblika bila je 25.28, a prosječna frekvencija odrednice 257.85. Frekvencije oblika i odrednica uzete su iz Frekvencijskog rječnika savremenog srpskog jezika (Kostić, 1999).

Pseudoriječi su formirane uz pomoć generatora za pravljenje pseudoriječi (Keuleers & Brysbaert, 2010) i njihova struktura je odgovarala riječima srpskog jezika u nominativu množine muškog roda.

4.1.3. Subjekti: U eksperimentu je učestvovao 41 student druge i četvrte godine, studijske grupe za psihologiju, Filozofskog fakulteta Univerziteta u Banjoj Luci. Iz dalje analize isključena su dva subjekta, koji su napravili više od 20% grešaka u toku eksperimenta. Konačni uzorak je činilo 39 ispitanika, 8 mladića i 31 djevojka. Svim subjektima maternji jezik je srpski. Subjekti su imali normalan ili korigovan vid do normalnog.

4.1.4. Aparatura: Eksperiment je realizovan uz pomoć softvera *SuperLab 4.5 for Windows* (Cedrus Corporation, 2010) i odgovarajuće serijalne kutije (eng. *response box*) RB-530 za prikupljanje odgovora. Korišćen je računar, čija je brzina procesora 2.0 GHz, a RAM memorija 0.99GB. Stimulusi su izlagani na 17 inčnom LG Flatron monitoru, u rezoluciji 1024x768 piksela, sa frekvencijom osvježavanja ekrana 85Hz i vremenom osvježavanja (eng. *refresh rate*) od 11.78ms.

4.1.5. Procedura: Svaki ispitanik je prvo prošao vježbu, koja se sastojala od osam stimulusa. Ovi stimulusi nisu ponavljani u eksperimentu niti su razmatrani u analizi. Prosječno vrijeme trajanja eksperimenta bilo je oko 15 minuta po ispitaniku. Stimulusi su bili napisani velikim crnim boldovanim slovima na bijeloj podlozi, a prikazivani su na centru monitora. Korišćen je font tipa ariel, veličine 25. Maksimalna ekspozicija stimulusa je iznosila 1500 ms, nakon čega je odgovor tretiran kao greška. U tim slučajevima, ispitanici su na ekranu dobijali uputstvo da odgovaraju brže. Fiksaciona tačka je izlagana 1500 ms, dok je interval između dva stimulusa (eng. *intertrial interval*) iznosio 3500 ms. U slučaju pogrešnog odgovora, stimulusi nisu ponavljani. Uz informacije o načinu odgovaranja, ispitanicima je bilo sugerisano da se trude da rade što je moguće brže, ali da pri tom što manje griješe.

4.1.6. Statistička analiza: Analiza vremena reakcija izvedena je pomoću linearnih mješovitih modela u statističkom okruženju R (verzija 2.13.1; R Development Core Team, 2011), sa odgovarajućim modulima: *lme4* (Bates & Maechler, 2009) i

languageR (Baayen, 2008, 2010). Za analizu grešaka, u okviru navedenih modula, korišćena je binomna link-funkcija (Faraway, 2004, 2005; Jaeger, 2008).

4.2. Rezultati i diskusija

Vremena reakcije su pročišćena tako što su iz analiza isključena dva ispitanika, koja su imala više od 20% grešaka na svim stimulusima. Odbačeno je i 11 stimulusa (9.2% ukupnog broja stimulusa), na kojima je više od 20% subjekata napravilo grešku: osam stimulusa iz *klastera pogrešnih rješenja* i tri stimulusa iz *klastera tačnih rješenja* (Prilog 5). Konačno, izostavljena su i vremena reakcije ispod 415 ms (šest slučajeva ili 0.13%). Ukupno, iz analiza je isključeno 13.72% podataka.

Da bi se ispunili preduslovi za primjenu analize linearnih modela, prije svega normalna (simetrična) raspodjela zavisne varijable, vremena reakcije su transformisana u logaritamske vrijednosti. Frekvencije oblika i leme su, takođe, pretvorene u log-vrijednosti, kako bi se ujednačila disperzija vrijednosti numeričkih (kontinuiranih) prediktora. Pripadnost klasteru (klaster tačnih i klaster pogrešnih rješenja) tretirana je kao fiksni faktor, a ispitanici i stimulusi kao slučajni faktori. Testiran je veći broj modela, da bi se utvrdilo koji od njih je najviše potkrijepljen podacima (eng. *fit*; Prilog 6).

Konačni model uključivao je i interakciju tipa klastera i reziduala frekvencije leme:

$$model <\text{-}\ lmer(RT \sim klaster * rezidual\ f_{\text{odr.}} + f_{\text{oblika}} + (1|subjekt) + (1|stimulus))$$

Utvrđeni su efekti svih faktora uključenih u model, kao i efekat interakcije tipa klastera i reziduala frekvencije leme (Tabela 18), u prilog čemu govore *t*-vrijednosti veće od dva (Baayen, Davidson, & Bates, 2008) i aproksimirane *p*-vrijednosti manje od .05, simulirane uz pomoć *MCMC* metoda (eng. *Markov chain Monte Carlo*) sa 10000 iteracija.

Uključivanje slučajnih efekata subjekata (χ^2 = 1275.1, *df* = 3, *p* < .001) i slučajnih efekata stimulusa (χ^2 = 808.35, *df* = 3, *p* < .001) značajno popravlja prilagođenost modela podacima. Kvadrat korelacije između transformisanih logaritamskih vrijednosti vremena reakcije i ocijenjenih vrijednosti iznosi .44. Ova mjera predstavlja procenat varijanse vremena reakcije (44%) koji je objašnjen odabranim modelom.

Tabela 18. *Parametri mješovitog modela, koji fituje vrijeme reakcije dobijeno na imenicama muškog roda u nominativu množine u zadatku leksičke odluke*

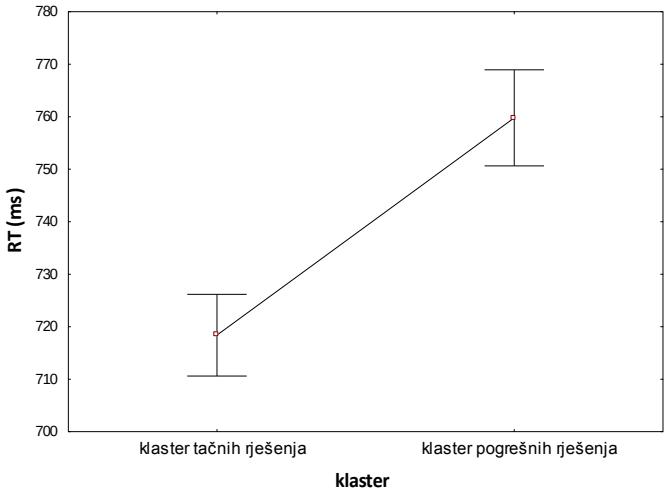
	Koeficijent	Standardna greška	t-vrijednost	Pr(>|t|)
intercept	6.6767	0.0250	267.49	.000
TAČNI	-0.0421	0.0162	-2.60	.009
rezidual frekvencije odrednice	-0.0621	0.0123	-5.05	.000
frekvencija oblika	-0.0369	0.0056	-6.65	.000
klaster*rezidual frekvencije leme	0.0545	0.0164	3.32	.001

Izvršene su i dvije naknadne analize, kako bi se provjerilo da li je izabrani model plauzabilan. Prva analiza je obavljena na svih 120 stimulusa, tj. u analizu su bila uključena i vremena reakcija za 11 stimulusa na kojima je greška bila iznad 20%, kao što su: *mitinzi*, *slivovi*, *kalemovi* itd., iz klastera pogrešnih rješenja, i *povici*, *parlamentarci*, *skverov*i iz klastera tačnih rješenja (Prilog 5). Parametri i nivoi značajnosti dobijeni u ovoj analizi (Prilog 7) ne razlikuju se od rezultata koji su dobijenim primjenom odabranog modela na selektovanim podacima. Ovi nalazi ukazuju da odbacivanje stimulusa koji su subjektima bili teški za obradu i na kojima su pravili značajan broj grešaka (više od 20%) nije uticalo na izbor odgovarajućeg modela.

U drugoj provjeri iz analize su izostavljeni i oni subjekti i stimulusi na kojima se greška zadržanog modela kreće izvan opsega +/-2.5 jedinice standardne devijacije u distribuciji reziduala (više o kritici statističkog modela u: Baayen &

Milin, 2010). Na ovaj način je pokazano da dalje isključivanje podataka, koji najviše odstupaju od predloženog modela, ne mijenja značajnost utvrđenih efekata (Prilog 8).

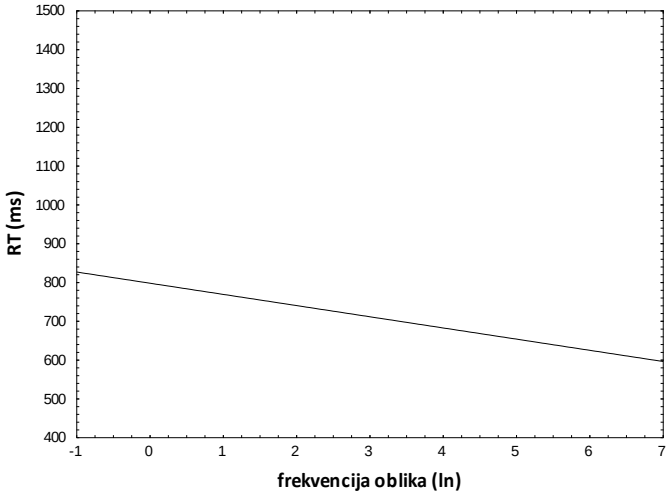
Po važnosti za ovaj rad, izdvaja se pitanje brzine obrade stimulusa u zavisnosti od toga da li pripadaju klasteru tačnih ili klasteru pogrešnih rješenja. Utvrđeno je da je potrebno više vremena da se obrade stimulusi na kojima je model pravio greške, od stimulusa koji pripadaju skupu tačnih rješenja (Slika 11). Razlika u prosječnim vremenima reakcija između ova dva klastera je statistički značajna (Tabela 18).



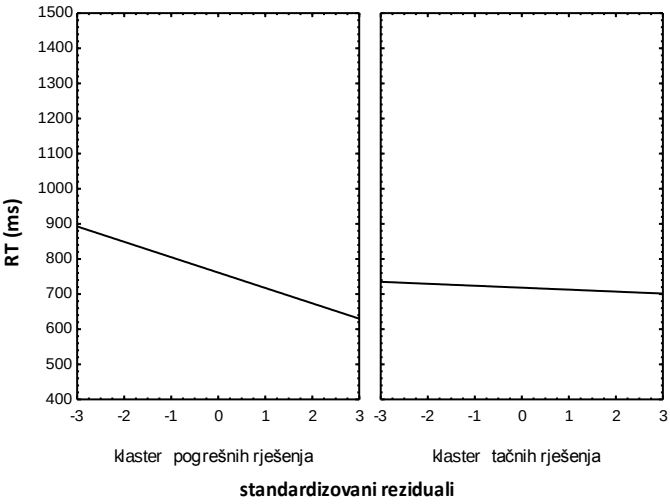
Slika 11. Prosječno vrijeme reakcije za klastere tačnih i pogrešnih rješenja.

Ustanovljen je i značajan facilitatorni efekat frekvencije oblika, tj. viša frekvencija podrazumijeva kraće vrijeme obrade, bez obzira o kojem klasteru je riječ (Tabela 18, Slika 12).

Uticaj (reziduala) frekvencije odrednice je, takođe, dosegao statističku značajnost. Ova kovarijabla, dodatno, ima efekat na vrijeme reagovanja i u interakciji sa tipom klastera (Tabela 18). Priroda interakcije postaje sasvim jasna kada se prouči slika koja prikazuje efekat učestanosti odrednice, odvojeno po nivoima faktora tip klastera (Slika 13).



Slika 12. Efekat frekvencije oblika na vrijeme reagovanja.



Slika 13. Interakcija tipa klastera i frekvencije odrednice.

Učestanost odrednice ima facilitatorni efekat samo za "teške" imenice, tj. imenice na kojima MBL griješi. Ovakav nalaz prilično ubjedljivo upućuje na pretpostavku o kompenzatorskim efektima: kada se kognitivni sistem nađe pred teškim zadacima "traže" se dopunski mehanizmi, koji bi olakšali obradu, dok, s druge strane, u slučaju lakših/lakih zadataka, sistem "štedi" resurse. Ovakvi kompenzacioni procesi omogućavaju kognitivnom sistemu da ostane u opsegu optimalnog funkcionisanja što, nadalje, posredno potvrđuje pretpostavku onih autora

koji se zalažu za opisivanje i posmatranje jezičkog sistema kao *kompleksnog adaptivnog sistema* (eng. *complex adaptive system*; Beckner et al, 2012).

Čini se da interakcija učestanosti odrednice i tipa klastera indirektno ukazuje i na problem veličine uzorka za učenje, a time i na problem veličine korišćenog korpusa. Ako se pogleda profil za klaster pogrešnih rješenja, primjetno je da visoka frekvencija lema kompenzuje težinu materijala koji se obrađuje, zbog čega je RT kraće nego kod imenica iz klastera tačnih rješenja. Zato je za očekivati da bi veliki korpus, sa brojnim primjerima i njihovim stabilno ocijenjenim učestanostima, obezbijedio i veću tačnost u zadatku automatske produkcije infleksionih oblika.

Pored analize vremena reakcije, napravljena je i analiza grešaka. Ona je izvedena na svim stimulusima, uz pomoć binomne link-funkcije (Faraway, 2004, 2005; Jaeger, 2008). Iz analize su isključena dva subjekta, kod kojih je procenat grešaka na svim stimulusima bio iznad 20%.

Na osnovu dobijenih rezultata (Tabela 19) može se zaključiti da su greške koje su subjekti pravili u zadatku leksičke odluke u funkciji frekvencije leme i frekvencije odrednice: što su ove frekvencije više, subjekti su manje griješili, i obratno, što su frekvencija odrednice i oblika niže, broj grešaka je bio veći.

Tabela 19. *Parametri mješovitog modela, koji fituje greške dobijene na imenicama muškog roda u nominativu množine u zadatku leksičke odluke*

	Koeficijent	Standardna greška	z-vrijednost	Pr(>|z|)
intercept	-2.1683	0.2739	-7.92	.000
TAČNI	-0.5076	0.3007	-1.69	.091
rezidual frekvencije odrednice	-0.4248	0.2082	-2.04	.041
frekvencija oblika	-0.6045	0.1156	-5.23	.000
klaster*rezidual frekvencije leme	0.2976	0.2951	1.01	.314

S druge strane, iako nije utvrđen statistički značajan efekat klasterske pripadnosti na broj grešaka, dobijena razlika jeste na granici značajnosti (p = .091). U prilog zaključku da broj grešaka zavisi od tipa klastera kojem pripadaju imenice, govori i postojanje glavnog efekta klastera na broj grešaka, χ^2 = 17.97, C = .065, p = .000.[54] Pored toga što je na imenicama iz klastera pogrešnih rješenja (tzv. "teških" imenica) bilo više grešaka, stiče se i utisak da su subjekti u obradi ovih imenica "žrtvovali" brzinu obrade zarad tačnosti. U prilog ovakvog zaključka govori i podatak da je ukupno bilo 335 grešaka, što predstavlja 7,2% svih odgovora.

Rezultati dobijeni u ovom eksperimentu potvrđuju kognitivnu vjerodostojnost učenja zasnovanog na memoriji. Pokazano je da je za imenice muškog roda u nominativu množine iz klastera pogrešnih rješenja vrijeme reagovanja duže i da je na njima pravljen veći broj grešaka nego na imenicama iz klastera tačnih rješenja. To znači da su imenice, koje su predstavljale izazov modelu u zadatku automatske produkcije infleksione morfologije, zahtjevnije za obradu i subjektima u zadatku leksičke odluke, tj. zadatku razumijevanja. U skladu sa tim je i nalaz da je za obradu riječi iz klastera pogrešnih rješenja potreban dodatni izvor informacija, koji kognitivni sistem nalazi u učestalosti riječi, kako bi se zadržalo optimalno funkcionisanje. Ujedno, nalazi ovog dijela studije posredno ukazuju i na kognitivnu plauzabilnost fonotaktičkih informacija na koje se učenje zasnovano na memoriji oslanjalo u zadatku automatske produkcije infleksionih oblika.

[54] Razlog zašto nije dobijen statisički značajan efekat klasterske pripadnosti na broj grešaka u slučaju kada je ova varijabla testirana zajedno sa frekvencijama oblika i leme je posljedica toga što je u model uključena i interakcija klasterske pripadnosi i reziduala frekvencije leme. Ova, testirana interakcija je, najvjerovatnije, maskirala efakat klasterske pripadnosti na broj grešaka.

5. OPŠTA DISKUSIJA

U ovoj studiji ispitivana je mogućnost automatske obrade i produkcije infleksione morfologije na osnovu fonotaktičkih informacija, tj. kombinacija fonoloških jedinica koje se javljaju u jeziku (Crystal, 2008).

Informativnost ovih jedinica testirana je na tri načina: (1) pomoću jednog od najboljih algoritama za klasifikaciju, tzv. mašina sa vektorima podrške (eng. *support vector machines – SVM*; Vapnik, 1995, 1998); (2) u zadatku automatske produkcije infleksione morfologije, pomoću TiMBL-a (eng. *Tilburg memory based learner*; Daelemans et al., 2010), koji se oslanja na zaključivanje po analogiji i primjere uskladištene u memoriji; (3) eksperimentalno, provjerom rezultata dobijenih u zadatku automatske produkcije infleksione morfologije u zadatku leksičke odluke.

Ad 1.) U prvom dijelu studije obavljena je diskriminacija promjenljivih vrsta riječi na osnovu vjerovatnoća dva odnosno tri fonema/grafema (bigrami i trigrami), pri čemu su te vjerovatnoće izračunate na nivou gramatičkih tipova. Korišćene su sve kombinacije bigrama odnosno trigrama: bigrami/trigrami na početku riječi, na kraju riječi, bez obzira na poziciju unutar riječi, te ove vrste informacija uzete zajedno.

Značajno ograničenje izbora alata za klasifikaciju predstavljale su distribucije vjerovatnoća bigrama i trigrama, koje značajno odstupaju od normalnih. S druge strane, redukcija vjerovatnoća bigrama i trigrama na kategorije (javio se – nije se javio) imala bi za posljedicu značajan gubitak varijanse, što bi rezultovalo manjom tačnošću klasifikacije. Iz tih razloga je klasifikacija obavljena uz pomoć SVM (Vapnik, 1995, 1998), koje predstavljaju robustan alat za analizu podataka.

Potrebni parametri za ovu analizu određeni su pomoću mrežnog pretraživanja. Ovaj postupak, iako nešto zahtjevniji od automatskog, podrazumijeva punu kontrolu istraživača u procesu određivanja odgovarajućih parametara, do kojih se dolazi sistematskim variranjem njihovih veličina i provjerom tačnosti odgovarajuće klasifikacije.

Najveći broj analiza obavljen je korišćenjem linearne jezgrene funkcije za koju je potrebno odrediti samo jedan parametar *C*, koji određuje kaznu za pogrešno klasifikovane objekte (Hsu et al., 2010; Olson & Dulen, 2008, itd).

Ad 2.) Primjedba koja se može uputiti mašinama sa vektorima podrške jeste da u njihovoj osnovi ne leži jasan mehanizam koji bi mogao da predstavi način na koji čovjek izvršava slične zadatke. Za razliku od SVM (Vapnik, 1995, 1998), to nije slučaj sa MBL (Daelemans & Van den Bosch, 2005). Učenje zasnovano na memoriji predstavlja računarsku implementaciju pristupa baziranog na primjerima, čije glavne postavke odražavaju, ne samo jezička znanja, nego i moguću (pretpostavljenu) kognitivnu arhitekturu i procese prisutne pri obradi prirodnog jezika, tj. daju prikaz načina na koji čovjek (možda) stiče i koristi jezička znanja (Baayen, 2011; Keuleers, 2008; Keuleers & Dealemans, 2007; Keuleers & Sandra, 2008; Milin et al., 2011, itd.).

U drugom dijelu empirijskog istraživanja, učenje zasnovano na memoriji je primijenjeno u zadatku automatske produkcije infleksionih oblika. Zadata je određena riječ, a algoritam je trebalo da, na osnovu fonotaktičkih informacija iz zadnja četiri sloga, generiše traženi infleksioni oblik. U uzorku je bilo 89024 promjenljivih riječi – egzemplara, a korišćen je metod izostavljanja jednog primjera (eng. *leave-one-out*).

Ad 3.) Na osnovu rezultata dobijenih u zadatku automatske produkcije infleksione morfologije pomoću MBL-a, osmišljen je eksperiment kako bi se provjerilo da li su primijenjeni model i informacije na koje se ovaj model oslanjao kognitivno vjerodostojni.

Formirane su dvije grupe stimulusa, jedna od tačnih, a druga od pogrešnih oblika, kako ih je automatski produkovao sistem za učenje. Grupe su bile ujednačene po broju slova, slogova i fonoloških alternacija, a razlikovale su se po frekvencijama oblika i odrednica.

Cilj je bio da se ispita da li su ispitanicima, u zadatku leksičke odluke, za obradu zahtjevniji stimulusi na kojima je model griješio, od onih koje je tačno obradio.

Osnovni zaključak studije jeste da su fonotaktičke jedinice srpskog jezika dovoljno informativne (nose dovoljnu količinu informacija) da se obave složeni jezički zadaci, kao što su razvrstavanje gramatičkih tipova u odgovarajuće vrste riječi i produkcija infleksionih oblika.

Uspješnost diskriminacije vrsta riječi, na osnovu fonotaktičkih informacija specifikovanih na osnovu gramatičkih tipova, koja je dobijena na svim bigramima [#x, x#, xy] uz pomoć radijalne jezgrene funkcije kretala oko 95%. S druge strane, u zadatku infleksione produkcije uz pomoć učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005), na svim promjenljivim vrstama riječi uzetim zajedno, ispravno je generisano oko 92% infleksionih oblika.[55] Ovakav procenat tačno obrađenih riječi je nešto manji od tačnosti koja se dobija uz pomoć najefikasnijih tagera koji se koriste u ovakvim ili sličnim zadacima jezičke obrade (više u: Güngör, 2010; Manning, 2011; Manning & Schuetze, 2000, itd.). Pri tom, ne treba zaboraviti činjenicu da su fonotaktičke informacije, uslovno rečeno, jednostavan, ako ne i najjednostavniji oblik informacija koji se može koristiti u ovakvim i sličnim zadacima, kao i da je učenje zasnovano na memoriji (Daelemans & Van den Bosch, 2005) primijenjeno na sve promjenljive riječi, što, inače, nije praksa; obično se ovaj metod primjenjuje na jedan jezički fenomen, npr. generisanje alomorfije (Milin et al., 2011), množine (Hahn & Nakisa, 2000; Keuleers & Daelemans, 2007; Keuleers et al., 2007) i slično. Ukupno gledano, dobijeni nalazi ne iznenađuju, s obzirom na to da je u brojnim psiholingvističkim istraživanjima utvrđena značajna uloga fonotaktičkih informacija u usvajanju, percepciji i produkciji jezika, o čemu je bilo riječi u uvodnom dijelu ovog rada.

Značaj fonotaktičkih informacija potvrđen je i u eksperimentalnom dijelu istraživanja, u kojem je pokazana kognitivna vjerostojnost učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005), što implicira i vjerodostojnost informacija na koje se model oslanjao u zadatku infleksione produkcije. Utvrđeno je

[55] Radi se o korigovanoj tačnosti, kada se u obzir uzmu i nemarkirani ranopravni oblici u uzorku (kao npr. u slučaju dugih ili kratkih oblika pridjeva i zamjenica, dubletnih oblika imenica, oblika ijekavice i ekavice itd.) koji su, u obradi, okarakterisani kao pogrešni. Tačnost bez ovakvog korigovanja, iznosila je oko 89% ispravno produkovanih infleksionih oblika.

da je u zadatku leksičke odluke za obradu stimulusa iz klastera pogrešnih rješenja potrebno više vremena te da subjekti više griješe na ovakvim stimulusima, u odnosu na riječi iz klastera tačnih rješenja.

Kao što je već konstatovano, direktna analogija sa zadatkom automatske infleksione produkcije uz pomoć MBL bila bi produkcija (usmena ili pisana) infleksionih oblika za zadate riječi. Međutim, i zadatak leksičke odluke je procedura koja omogućava provjeru kognitivne vjerodostojnosti MBL-a i korišćenih informacija, a rezultati dobijeni tim postupkom dopuštaju generalizaciju nalaza i na procese razumijevanja jezika.

Pored toga, postupak provjere kognitivne vjerodostojnosti razlikovao se od "standardnog" načina upotrebe ovakvih modela u sličnim zadacima. Uobičajeno je da se u bihejvioralnim studijama prvo utvrde parametri neophodni za izvršavanje nekog jezičkog zadataka, koji se zatim unose, tj. zadaju modelu. Ovdje je prvo primijenjen model, a zatim je, na osnovu dobijenih rezultata, osmišljen i izveden ekperiment, kako bi se utvrdilo da li su riječi, koje model nije mogao tačno da obradi, teže i subjektima za obradu nego tačno obrađene riječi. Drugim riječima, model obrade poslužio je kao osnova za generisanje eksperimentalnog nacrta i, što je važnije, empirijskih pretpostavki.

Osim potvrde dobijene u eksperimentu, o kognitivnoj relevantnosti fonotaktičkih informacija indirektno govore i rezultati dobijeni prilikom diskriminacije vrsta riječi uz pomoć SVM-a. Iako SVM ne odslikavaju vjerovatni kognitivni mehanizam obrade jezika (Baayen, 2011), ovaj statistički alat kao osnovu za izračunavanje koristi matrice udvojenih poređenja elemenata (eng. *pairwise comparisons*; Vert, Tsuda, & Schölkopf, 2004), tj. oslanja se na informacije o međusobnoj sličnosti ulaznih podataka, što bi, uz činjenicu da postoje karakteristične distribucije fonotaktičkih informacija za pojedine vrste riječi, moglo implicirati "blizinu" bigrama i trigrama na kognitivnom planu.

S druge strane, visok procenat tačne produkcije infleksione morfologije dobijen na svim promjenljivim riječima (oko 92%) ukazuje na robusnost učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005). Ovaj rezultat je

važan jer učenje zasnovano na memoriji nije samo još jedan, potencijalno efikasan, algoritam za obradu riječi. Na ovaj model se može gledati i kao na moguću operacionalizaciju načina kako stičemo i koristimo jezička znanja. S obzirom na rezultate brojnih istraživanja, danas nije pitanje da li zaključivanje na osnovu primjera i analogije ima ulogu u obradi morfološki složenih riječi,[56] već da li se ta obrada realizuje isključivo na osnovu tog mehanizma (Daelemans & Van den Bosch, 2005; Hare et al., 1995; Keuleers, 2008; Keuleers & Daelemans, 2007; Keuleers et al., 2007; Milin et al., 2009; Rumelhart & McClelland, 1986; Skousen, 1989, 1992, 2002a, 200b, 2009, itd.).[57]

Značajne implikacije dobijenih nalaza tiču se i međusobnog odnosa obrade (recepcije) i produkcije jezika. Može se primijetiti da je u zadatku diskriminacije vrsta riječi na osnovu fonotaktičkih informacija tačnost bila manja za vrste riječi u okviru kojih se javlja manji broj gramatičkih tipova. U prosjeku, najmanji procenat tačno obrađenih vrsta riječi, bez obzira na tip bigrama i trigrama na osnovu kojih se obavlja razdvajanje, utvrđen je na imenicama, u okviru kojih su identifikovana 74 gramatička tipa, zatim pridjevima (131), glagolima (494), dok je najveći procenat tačno diskriminisanih vrsta riječi dobijen na zamjenicama, u okviru kojih je bilo 594 gramatičkih tipova. Situacija je potpuno drugačija kada se radi o zadatku infleksione produkcije, u kojem je najveća tačnost ostvarena na imenicama, a najmanja na zamjenicama. Pri tome, tačnost u zadatku automatske infleksione produkcije je u vezi i sa veličinom infleksione klase, tj. brojem različitih riječi u okviru jednog gramatičkog tipa. Tako je, npr. najveća tačnost u produkciji infleksionih oblika za imenice i glagole dobijena na klasama koje su imale između 500 i 1000 primjera, a za pridjeve kada se veličina klase kretala između 100 i 250 primjera. Nasuprot tome,

[56] Radi podsjećanja, i suprotstavljeni pristup analoškom modelu, tzv. pristupi dvostrukog puta (eng. *dual-route approaches*) postuliraju obradu zasnovanu na analogiji sa sličnim primjerima u memoriji kod nepravilnih oblika, dok se obrada pravilnih oblika obavlja na osnovu gramatičkih pravila (Clahsen, 1999; Marcus et al., 1995; Pinker, 1991, 1999; Pinker & Prince, 1988, 1994; Prasada & Pinker, 1993, itd.).

[57] Modeli zasnovani na analogiji nemaju primjenu samo u okviru obrade morfološki složenih riječi, već je to pristup obradi jezika generalno, ali i model predikcije cjelokupnog ponašanja (više u: Daelemans et al., 2010; Nosofsky, 1992; Nosofsky, Kruschke, & McKinley, 1992; Nosofsky & Palmeri, 1997; Skousen, 2002b, 2009, itd.).

na zamjenicama je pokazano da je zaključivanje najnepouzdanije u slučajevima kada klasa ima samo jedan primjer, dok već kod klasa sa dva primjera dolazi do značajnog povećanja tačnosti.

Ovakvi rezultati se mogu tumačiti u svjetlu širih implikacija Zipfovog zakona i *principa najmanjeg napora* (eng. *principle of least effort*; Zipf, 1935, 1949). Zipfov zakon govori o tome da je umnožak učestalosti riječi i njihovog ranga konstantan, tj. da se mali broj riječi javlja veoma često, dok se veliki broj riječi javlja rijetko. Ovakav odnos, pretpostavlja se, posljedica je principa najmanjeg napora, po kome i onaj ko šalje signal (govornik) i onaj ko signal prima (slušalac) nastoje da minimalizuju svoj napor u produkciji i razumijevanju jezika (Manning & Schutze, 2000). Da bi se to postiglo, potrebno je da govornik ima mali vokabular zajedničkih riječi (riječi sa više značenja), a slušalac veliki vokabular jednoznačnih riječi, koje se rijetko javljaju. Maksimalan kompromis između ove dvije potrebe ogleda se u obrnuto-recipročnom odnosu frekvencije i ranga riječi o kojem govori Zipfov zakon (Manning & Schutze, 2000).

Ako bi se princip najmanjeg napora generalizovao na vrste riječi i broj gramatičkih tipova u okviru njih, to znači da bi se u zadatku diskriminacije vrsta riječi na osnovu fonotaktičkih informacija specifikovanih na nivou gramatičkih tipova, koji je analogan razumijevanju jezika, trebale bolje diskriminisati vrste riječi sa većim brojem gramatičkih tipova i ograničenim brojem fonotaktičkih informacija[58] i obratno. Slično prethodnom, u zadatku automatske produkcije infleksionih oblika, veća tačnost bi se trebala dobiti na vrstama riječi sa manjim brojem gramatičkih tipova i većom učestalošću primjera u njima.[59] Ovakve pretpostavke su u skladu sa rezultatima koji su dobijeni u ovom radu. Iako nalazi ne dopuštaju direktno izvođenje ovakvih zaključaka, oni su indirektna potvrda teze

[58] Analogno velikom vokabularu jednoznačnih riječi, koje se rijetko javljaju. U konkretnom slučaju, to je veliki broj gramatičkih tipova koji se javljaju kod zamjenica, u okviru kojih postoji jasno specifikovana distribucija fonema, zbog toga što se radi o zatvorenom skupu riječi (tj. konačnom broju zamjenica).
[59] Analogno malom vokabularu višeznačnih riječi.

Ramskara i Bajina (Ramscar & Baayen, 2013) da su produkcija i precepcija jezika "u ravnoteži".

Pored navedenih, može se izdvojiti još nekoliko značajnih nalaza dobijenih u ovoj studiji. Jedan takav nalaz tiče se dubljeg uvida u obradu "teških" imenica, tj. imenica koje TiMBL u zadatku infleksione produkcije nije tačno obradio. U okviru eksperimentalnog dijela istraživanja pokazano je da obradu ovih imenica, kao i obradu imenica iz klastera tačnih rješenja, facilitira frekvencija oblika. Međutim, pri obradi imenica iz klastera pogrešnih rješenja facilatornu ulogu ima i frekvencija leme koja kompenzuje veće opterećenje koje stoji pred kognitivnim sistemom prilikom obrade ovih imenica. Pojednostavljeno rečeno, da bi kognitivni sistem mogao da izađe na kraj sa zahtjevnijim stimulusima, potrebno je da na raspolaganju ima odgovarajući resurs. Taj resurs predstavlja veliki rječnik, bilo da je riječ o vokabularu pojedinca ili velikim jezičkim korpusima, koji su osnova za različite zadatke iz oblasti automatske obrade jezika.

Sljedeći nalaz koji vrijedi spomenuti jeste potvrda o nedovoljnoj količini informacija koje nose bigrami na kraju riječi, potrebnih za uspješno obavljanje diskriminacije vrsta riječi. Najlošiji rezultat u diskriminaciji promjenljivih vrsta riječi dobijen je u slučaju kada se diskriminacija oslanjala na informacije o zadnjem grafemu/fonemu u riječima. Najveće greške utvrđene su kod pridjeva i imenica, kod kojih je greška iznosila oko 80%. S druge strane, iako je procenat tačne dikriminacije kod zamjenica (85%) i glagola (oko 65%) znatno veći, on je niži u odnosu na tačnost razdvajanja na osnovu drugih tipova bigrama i trigrama. Razlog manje efikasne diskriminacije vrsta riječi na osnovu bigrama na kraju riječi posljedica je njihove manje informativnosti jer ih dijeli veći broj gramatičkih tipova i veći broj vrsta riječi, o čemu govore i Bajin i saradnici (Baayen et al., 2011).

Iako u fokusu ovog istraživanja nije bio problem reprezentacije jezičkih informacija u memoriji, nije moguće ne dotaći se i ovog pitanja. Pretpostavka od koje se pošlo je da su fonotaktičke informacije (bigrami, trigrami i slogovi) reprezentovane u kognitivnom sistemu. Takođe, u dijelu koji se tiče infleksione

produkcije, pretpostavljeno je da su egzemplari, na osnovu kojih je model učio, uskladišteni u memoriji.

S obzirom na visoku efikasnost i procenat tačno diskriminisanih vrsta riječi (95%) te procenat tačne infleksione porodukcije (92%) na osnovu fonotaktičkih informacija, opravdano je zapitati se da li je ova vrsta informacija "dovoljna" za obavljanje zadataka iz domena morfologije. Ako jeste, to dovodi u pitanje postavke lingvističkih modela koji za centralnu jedinicu postavljaju morfem (eng. *morpheme-based morphology*; Bresnan, 1982; Di Sciullo & Williams, 1987; Lieber, 1992; Scalise, 1986; Selkirk, 1982). Međutim, ovo predstavlja neznatni problem za *morfologiju baziranu na leksemama* (eng. *lexeme-based morphology*; Anderson, 1992; Aronoff, 1976, 1994, Halle & Marantz, 1993; Stump, 1991; Zwicky, 1989) odnosno morfologiju zasnovanu na riječima (eng. *word-paradigm morphology*; više u Blevins, 2006, 2013).

Međutim, ključno je da ne postoji potreba da se, zbog pitanja morfologije, pretpostavlja postojanje mentalnog leksikona u kojem su uskladištene reprezentacije morfema i/ili riječi. Dobijeni nalazi sugerišu da je za objašnjenje morfologije dovoljno da model sadrži ortografiju/fonologiju i semantiku. Ovakav zaključak je u skladu sa pretpostavkom Bajbijeve, koja smatra da je veza između riječi u kognitivnom sistemu fonološke i semantičke prirode (više u: Bybee, 1985, 2001, 2010). U jednom takvom modelu, informacije o sekvencama fonema/grafema su nužne kako bi bili integrisani empirijski nalazi koji govore o značajnoj ulozi fonotaktičkih ograničenja i veličine skupa sličnih susjeda na obradu jezika. Pri tome, treba imati na umu da oba ova faktora nisu nezavisna od riječi. S druge strane, veze među riječima nisu samo bazirane na sličnosti fonema/grafema koje dijele (koje su im zajedničke), nego su i semantičke prirode. O ovome zapravo govore i to ističu svi modeli "jednog puta" i modeli koji zagovaraju direktno mapiranje forme i značenja (Baayen et al., 2011; Bybee, 1985, 1999, 2010; Plaut & Gonnerman, 2000; Rumelhart & McClelland, 1986; Seidenberg & Gonnerman, 2000, itd.). Konačno, modeli moraju uključivati, makar implicitno, znanje o svijetu (Dimitrijević, 2007; Milin 2005). U slučaju automatske produkcije infleksionih oblika pomoću MBL to

implicitno znanje se ogleda u informaciji o traženom gramatičkom tipu. Oznaka za traženi gramatički oblik bila je jedna od odlika po kojima su formirani skupovi najbliži susjeda. Ona nije nosilac samo gramatičkog značenja, nego i informacije o funkciji koju traženi oblik ima, tj. funkciji koju bi obavljao u rečenici (npr. funkciji subjekta), u kojoj je i sadržano "znanje o svijetu". Iz svih navedenih razloga, gotovo da nije moguće zamisliti model obrade jezika koji ne bi uključivao i semantiku. Svi modeli uključuju i/ili podrazumijevaju semantiku, a razlikuju se po tome kako specifikuju formu, da li uključuju morfologiju i sl.

U korist modela morfologije koji je zasnovan na ortografiji i semantici govore i konekcionistički modeli (Harm & Seidenberg, 1999, 2004; Rumelhart & McClelland, 1986; Seidenberg & Gonnerman, 2000), kao i model obrade koji se oslanja na *naivno diskriminativno učenje* (eng. *naive discriminative learning – NDL*; više u: Baayen et al., 2011).[60] Ovaj model je inspirisan teorijom Reskorle i Vagnera o diskriminativnom učenju i ukorijenjen je u tradiciju kognitivne psihologije, za razliku od modela zasnovanih na analogiji, koji vode porijeklo iz lingvistike. Morfološka obrada se obavlja na osnovu bigrama i trigrama, tj. model osim fonotaktičkih informacija ne sadrži reprezentaciju složenih riječi. U obradi morfologije postiže gotovo identičan rezultat kao i SVM, koje su najefikasniji statistički alat za klasifikaciju (Baayen, 2011; Joachims, 1998; Meyer et al., 2003; Steinwart & Christmann, 2008, Van Gestel et al., 2004). Međutim, potrebno je da se ovaj model prilagodi i primijeni i u zadacima produkcije morfološki složenih riječi te postane osjetljiv na semantičke veze između riječi (Baayen et al., 2011).

Prostor za dalja istraživanja postoji i u provjeri efikasnosti učenja zasnovanog na memoriji u zavisnosti od načina specifikovanja fonotaktičkih informacija. Uobičajen postupak za dobijanje egzemplara iste dužine jeste raščlanjivanje riječi na slogove i njihove elemente: nastup/ulaz, jezgro i rub/koda, (Keuleers et al., 2007). Međutim, postoje dokazi koji govore protiv tvrdnje da

[60] Iako slični, postoji nekoliko značajnih razlika između NDL i *triangl modela* (eng. *triangle model*; Harm & Seidenberg, 1999, 2004; Seidenberg & Gonnerman, 2000), konekcionističkog modela čitanja, koji se oslanja na ortografiju, fonologiju i semantiku (više u Baayen et al., 2011).

fonotaktička ograničenja (eng. *phonotactic constraints*) ovise isključivo od strukture sloga (na primjer: Blevins, 2003; 2006). U prilog tome govore i rezultati dobijeni u ovoj studiji, u zadatku diskriminacije vrsta riječi uz pomoć SVM-a, ali i efikasnost NDL-a (Baayen et al., 2011), koji se također oslanja na bigrame i trigrame.

Pored problema koji se tiče strukture fonotaktičkih informacija, postoje i "funkcionalna" pitanja, koja se odnose na način kako se odlučujemo za određenu veličinu skupa primjera na osnovu kojih donosimo odluku o novom obliku. U tom kontekstu, bilo bi korisno provjeriti da li bi se i kako uspješnost učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005) mijenjala u zavisnosti od vrijednosti *k*, tj. vrijednosti dopuštene distance. Iako se 7-NN model pokazao robusnim u produkciji infleksionih oblika, simulacije su rađene na pojedinačnim gramatičkim tipovima, npr. množini imenica ili prošlom vremenu glagola (Keuleers, 2008; Keuleers & Dealemans, 2007; Keuleers & Sandra, 2008), a ne na svim oblicima promjenljivih vrsta riječi uzetim zajedno. Ovo bi bilo važno, ne samo iz razloga da se nađe optimalna vrijednost *k* te postigne maksimalna efikasnost modela, već i zbog uvida u prirodu ovakvog zadatka. Naime, učenje zasnovano na memoriji (Daelemans & Van den Bosch, 2005) pravi jasnu razliku u produkciji poznatih (eng. *retreival*) i nepoznatih infleksionih oblika (eng. *generalization*; Keuleers & Dealemans, 2007). U prvom zadatku model pretražuje uskladištene primjere, dok se samo u slučaju predikcije nepoznatih infleksionih oblika zaključivanje vrši na osnovu analogije sa prethodnim iskustvom. Različitim zadacima odgovaraju i različite vrijednosti *k*: za produkciju poznatih infleksionih oblika optimalna vrijednost je jedan, dok za drugi tip zadatka zavisi od toga da li se radi o produkciji riječi ($k = 3$) ili pseudo-riječi ($k = 7$), što se smatra "čistim" zadatkom generalizacije (Keuleers & Dealemans, 2007). Osim toga, ako bi se za manje vrijednosti *k*, tj. za suboptimalne vrednosti, ustanovile značajne razlike u tačnosti produkcije, to bi ukazivalo na postojanje donje granice veličine skupa susjeda potrebne za uspješno obavljanje zadatka iz domena obrade jezika. Ovo je samo jedan od koraka kako bi se došlo do odgovora na pitanja, kao što su: kolika je vrijednost optimalnog *k*, da li je

ona ista za sve jezičke fenomene ili ne, mijenja li se u zavisnosti od vrste zadatka, zašto je baš ta vrijednost optimalna i sl.?

5.1. Zaključak

Centralni problem obrade jezika sa bogatom infleksionom morfologijom je obrada riječi. Kao osnova za obradu riječi mogu poslužiti različite vrste informacija, od fonotaktičkih, leksičkih, do strukturalnih sintagmatskih informacija, pojedinačno ili u kombinaciji. Ključno pitanje obrade riječi je, upravo, izbor informacija na koje se ta obrada oslanja. Iako postoji veliko interesovanje za ovu oblast, još je prisutan veliki broj otvorenih problema. Ti problemi često izlaze iz okvira obrade riječi i tiču se obrade jezika generalno, ali i samog funkcionisanja kognitivnog sistema.

Jedna od takvih dilema je da li se morfološki složene riječi obrađuju na osnovu analogije sa primjerima u memoriji (Daelemans & Van den Bosch, 2005; Hare et al., 1995; Keuleers et al., 2007; Rumelhart & McClelland, 1986; Skousen, 1989, 1992, itd.) ili su potrebna dva principa, jedan za obradu pravilnih, a drugi za obradu nepravilnih oblika (Clahsen, 1999; Marcus et al., 1995; Pinker, 1991, 1999; Pinker & Prince, 1988, 1994; Prasada & Pinker, 1993, itd.). Trenutno, u prednosti su modeli koji zagovaraju jedinstven princip obrade. Ovo potkrepljuju i neurološki nalazi, koji ne nude dovoljno dokaza o potrebi postuliranja dva fundamentalno različita mehanizma obrade morfološki složenih riječi (više u: Woollams & Patterson, 2012, str. 348).

Ako se prihvati stanovište o jedinstvenom principu obrade morfološki složenih riječi, otvara se pitanje o kakvom se mehanizmu radi. Rezultati dobijeni u ovoj studiji, u zadatku automatske produkcije infleksione morfologije, govore u prilog učenja zasnovanog na memoriji (Daelemans & Van den Bosch, 2005), u čijoj osnovi leži metod najbližih susjeda. Efikasnost učenja zasnovanog na memoriji demonstrirana je ranije u zadacima obrade infleksije (Daelemans et al., 1997; Eddington, 2002a, 2002b; Hahn & Nakisa, 2000; Keuleers, 2008; Keuleers & Daelemans, 2007; Keuleers et al., 2007; Krott et al., 2001; Krott et al., 2007; Milin

et al., 2011, itd.). Ipak, ne treba zaboraviti da postoje i drugačiji načini operacionalizacije zaključivanja po analogiji, kao što su, npr. Skounsenov analoški model (Skousen, 2002b, 2009), neuralne mreže (Rumelhart & McClelland, 1986) itd. S druge strane, postoje i modeli jednog pristupa koji se ne zasnivaju na analogiji, kao što je model naivnog diskriminativnog učenja (Baayen et al., 2011).

Rezultati dobijeni u prvom dijelu ove studije u zadatku diskriminacije promjenljivih vrsta riječi, kao i rezultati dobijeni upotrebom NDL modela (Baayen et al., 2011), sugerišu da su bigrami, tj. moguće kombinacije dva fonema/grafema, minimalne fonotaktičke jedinice na koje se možemo osloniti u obradi morfološki složenih riječi. Međutim, postoje i modeli koji koriste specifikovane fonotaktičke informacije na nivou slogova (Dell, 1986, 1988; Keuleers et al., 2007; Roelofs, 1997, 2000). Ipak, može se reći da se percepcija i produkcija morfološki složenih riječi može obaviti na osnovu fonoloških/ortografskih i semantičkih informacije, te da nema potrebe uvođenja posebnog domena zaduženog za obradu morfologije (vidi u: Baayen et al., 2011; Davis, 2004; Harm & Seidenberg, 1999, 2004; Seidenberg & Gonnerman, 2000, itd.).

Na kraju, može se zaključiti da još uvijek postoji mnogo pitanja vezanih za obradu morfološki složenih riječi, koji čekaju razrješenje. Odgovor na ta pitanja treba tražiti u integraciji nalaza dobijenih u bihejvioralnim i neurološkim studijama i računarskim simulacijama. Pri tome, jeziku treba prići kao kompleksnom adaptivnom sistemu, čije se strukture "…pojavljuju iz isprepletenih obrazaca iskustva, socijalne interakcije i kognitivnih mehanizama" (Beckner et al, 2012, str. 2), koji se značajno razlikuje od statičnog sistema gramatičkih principa koji zagovaraju pristalice generativnog pristupa. Gramatika nije predstavljena setom apstraktnih pravila, koja su samo indirektno povezana sa jezičkim iskustvom, već u formi *obazaca* (eng. *paterns*) koji *izrastaju* (eng. *emerge*), tj. koji su direktna posljedica upotrebe jezičkih informacija uskladištenih u kognitivnom sistemu (Bybee, 2001, 2006; Beckner et al., 2012). To podrazumijeva da je gramatičko znanje – proceduralno znanje, što se reflektuje i na fonologiju. Ona, na taj način,

postaje dio znanja za produkciju i percepciju gramatičkih konstrukcija, umjesto da je "…isključivo apstraktni, psihološki sistem" (Bybee, 2001, str. 8).

Ovi i slični stavovi imaju dalekosežne implikacije, jer ukazuju na brojna ograničenja strukturalnih teorije jezika i jezičkog ponašanja. Teorija informacija, teorija minimalnog ulaganja, različite teorije učenja, riječ-paradigma teorije klasične lingvistike i modeli koji iz njih proizilaze pokazuju da funkcionalni pristup jezičkim fenomenima olako odbačen. Rezultati dobijeni u ovoj studiji, na primjeru razumijevanja i produkcije složenih riječi u jeziku sa bogatom infleksionom morfologijom, kao što je srpski jezik, pokazuju da funkcionalne veze između ortografije/fonologije, s jedne strane, i značenja (semantike), s druge, omogućavaju uspjeh u leksičkom učenju, kao i da su karakteristike ovakvog, na memoriji zasnovanog učenja, kognitivno vjerodstojne. U krajnjoj liniji, ovi nalazi dovode u pitanje velike i skupe postulate o mentalnom leksikonu i o mentalnim predstavama koje sadrže složenu, a nedovoljno specifikovanu hijerarhiju najrazličitih lingvističkih opisa.

LITERATURA

Aha, D. W. (1997). *Lazy learning*. Dordrecht, NL: Kluwer Academic Publishers.

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, *6*(1), 37-66. http://dx.doi.org/10.1007/BF00153759

Albright, A. (2007). *Gradient phonological acceptability as a grammatical effect*. Retrieved from http://www.mit.edu/~albright/papers/Albright-GrammaticalGradience.pdf

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational / experimental study. *Cognition, 90*(2), 119-161. http://dx.doi.org/10.1016/ S0010-0277(03)00146-X

Anderson, S. R. (1992). *A-morphous morphology*. Cambridge, UK: Cambridge University Press.

Ando, R. K. (2004). Exploiting unannotated corpora for tagging and chunking. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics* (*ACL-2004*), Short paper, Barcelona, ES. http://dx.doi.org/10.3115/1219044.1219057

Andrews, S. (1989). Frequency and neighborhood size effects on lexical access: Activation or search? *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *15*, 802-814.

Andrews, S. (1992). Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy? *Journal of Experimental Psychology*: *Learning, Memory, and Cognition, 18*(2), 234-254. http://dx.doi.org/10.1037//0278-7393.18.2.234

Araujo, L. (2002). Part-of-speech tagging with evolutionary algorithms. In A. Gelbukh, (Ed.), *Computational Linguistics and Intelligent Text Processing*, *3rd Internacional Conference, CICling-2002, Mexico City* (pp. 230-239). Berlin, DE: Springer. http://dx.doi.org/10.1007/3-540-45715-1_21

Aronoff, M. (1976). *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Aronoff, M. (1994). *Morphology by itself*. Cambridge, MA: MIT Press.

Arlot, S., & Celisse, A. (2010). Survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40-79. http://dx.doi.org/10.1214/09-SS054

Atwell, E. (1987). Constituent-likelihood grammar. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analyses of English*: *A corpus-based approach* (pp. 57-65). London, UK: Longman.

Baayen, R. H. (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probability theory in linguistics* (pp. 229-287). Cambridge, UK: MIT Press.

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Baayen, R. H. (2010). *LanguageR*: *Data sets and functions with analyzing linguistic data. A practical introduction to statistics* (R package version 1.2) [Computer software]. Available from http://CRAN.R-project.org/package=languageR.

Baayen, R. H. (2011). Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics. 11*(2), 295-328. http://dx.doi.org/10.1590/S1984-63982011000200003

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Baayen, R. H., Feldman, L. F., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290-313. http://dx.doi.org/10.1016/j.jml.2006.03.008

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12-28.

Baayen, R. H., Milin, P., Filipović Đurđević, D., Hendrix, P., & Marelli, M. (2011). An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review, 118*, 438-481. http://dx.doi.org/10.1037/a0023851

Bailey, T. M., & Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language*, *44*(4), 568-591. http://dx.doi.org/10.1006/jmla.2000.2756

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology*: *General*, *133*(2), 283-316. http://dx.doi.org/10.1037/0096-3445.133.2.283

Baker, P., Hardie, A., & McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh, UK: Edinburgh University Press.

Banko, M., & E. Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (ACL '01), 26-33, http://dx.doi.org/10.3115/1073012.1073017

Bates, D., & Maechler, M. (2009). *lme4*: *Linear mixed-effects models using S4 classes* [Computer software manual]. Available from http://CRAN.R.project.org/package=lme4

Beckner et al., (2012). Language is a complex adaptive system: Position paper. *Language Learning*, *59*(1), 1-26. http://dx.doi.org/10.1111/j.1467-9922.2009.00533.x

Bellegarda, J. R. (2008). A novel approach to part-of-speech tagging based on latent analogy. In *International Conference on Acoustics, Speech, and Signal Processing* (I*CASSP-2008*), 4685-4688, Las Vegas, NV. http://dx.doi.org/10.1109/ICASSP.2008.4518702

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*, 289-300.

Blagus Bartolec, G. (2006). Od neživoga do živoga. *Rasprave Instituta za hrvatski jezik i jezikoslovlje*, *32*(1), 1-23.

Blevins, J. (1995). The syllable in phonological theory. In J.A. Goldsmith (Ed.), *The handbook of phonological theory* (pp. 206-244). Cambridge, UK: Blackwell.

Blevins, J. (2003). The independent nature of phonotactic constraints: An alternative to syllable-based approaches. In Féry, C., Van de Vijver, R. (Eds.), *The syllable in optimality theory* (pp. 375-403). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511497926.016

Blevins, J. (2006). Word-based morphology. *Journal of Linguistics*, *42*(3), 531-573. http://dx.doi.org/10.1017/S0022226706004191

Blevins, J. (2013). Word-Based morphology from Aristotle to modern WP (Word and paradigm models). In K. Allan (ed.), *Oxford handbook of the history of linguistics* (pp. 375-396). Oxford, UK: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199585847.013.0017

Bloomfield, L. (1933). *Language*. New York, NY: Holt, Rinehard and Winston.

Boas, H. (2003). *A constructional approach to resultatives*. *Stanford monographs in linguistics*. Stanford, CA: CSLI Publications.

Boden, M. A. (1987). *Artificial intelligence and natural man*. New York, NY: Basic.

Boden, M. A. (2006). *Mind as machine*: *A history of cognitive science, II*. Oxford, UK: Oxford University Press.

Boser, B. E, Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifier. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 144-152, Pittsburg, PA.

Brants, T. (2000). TnT – A statistical part-of speech tagger. *Proceedings of the 6th Applied Natural Language Conference* (*ANLP-2000*), 224-231, Seattle, WA. http://dx.doi.org/10.3115/974147.974178

Breiman, L., Friedman J. H., Olsen R. A., & Stone C. J. (1984). *Classification and regression.* Belmont, CA: Wadsworth International Group.

Bresnan, J. (1982). *On the mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Brill, E. (1992). A simple rule-based part of speech tagger. *Proceedings of the 3rd Applied Natural Language Conference* (*ANLP-1992*), 152-155, Trento, IT. http://dx.doi.org/10.3115/974499.974526

Brill, E. (1994). Some advances in rule-based part of speech tagging. *Proceedings of the 12th National Conference on Artificial Intelligence*, (*AAAI-1994*), 1, 722-727, Seattle, WA.

Brill, E. (1995). Transformation-based error-driven learning and Natural language processing: A case study in part of speech tagging. *Computational Linguistics, 21*, 543-565.

Bybee, J. L. (1985). *Morphology*: *A study of the relation between meaning and form*. Amsterdam, NL: John Benjamins.

Bybee, J. L. (2001). *Phonology and language use*. Cambridge, UK: Cambridge University Press.

Bybee, J. L. (2010). *Language, usage and cognition*. New York, NY: Cambridge University Press.

Bybee, J. L., & Eddington, D. (2006). A usage-based approach to Spanish verbs of 'becoming'. *Language, 82*(2), 323-55. http://dx.doi.org/10.1353/lan.2006.0081

Bybee, J. L., & Moder, C. L. (1983). Morphological classes as natural categories. *Language*, *59*(2), 251-270. http://dx.doi.org/10.2307/413574

Burges, C. J. C. (1999). Geometry and invariance in kernel based methods. In B. Schölkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods - Support vector learning* (pp. 89-116). Cambridge, UK: MIT Press.

Bürki, A., & Gaskell, M. G. (2012). Lexical representation of schwa words: Two mackerels, but only one salami. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 617-631. http://dx.doi.org/10.1037/a0026167

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus based approach to speech segmentation. *Cognitive Psychology*, *33*(2), 111-153. http://dx.doi.org/10.1006/cogp.1997.0649

Caramazza, A., (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*(1), 177-208. http://dx.doi.org/10.1080/026432997381664

Carreiras, M., Perea, M., & Grainger, J. (1997). Effects of the orthographic neighborhood in visual word recognition: Cross-task comparisons. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *23*(4), 857-871. http://dx.doi.org/10.1037//0278-7393.23.4.857

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245-276. http://dx.doi.org/10.1207/s15327906mbr0102_10

Cedrus Corporation (2010). SuperLab v. 4.5. [computer software]. San Pedro, CA.

Chang, C. L. (1974). Finding prototypes for nearest neighbor classifiers. *IEEE Transactions on Computers*, *23*(11), 1179-1184.

Charniak, E. (1993). *Statistical language learning.* Cambridge, MA: MIT Press.

Charniak, E. (1997). Statistical techniques for natural language parsing. *AI Magazine*, *18*, 33-44.

Charniak, E., Hendrickson, C., Jacobson, N., & Perkowitz, M. (1993). Equations for part-of-speech tagging. *Proceedings of the 11th National Conference on Artificial Intelligence (AAAI-1993)*, 784-789, Washington, USA: AAAI Press/MIT Press.

Cherkassky, V., & Mulier, F. (1998). *Learning from data – concepts, theory and methods*. New York, NY: JohnWiley & Sons,

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. *Proceedings of the 2nd Applied Natural Language Processing Conference (ANLAP-1988)*, 136-143, Austin, TX.

Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology*: *Human Perception and Performance*, *16*(3), 551-563. http://dx.doi.org/10.1037/0096-1523.16.3.551

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and Brain Sciences, 22*(6), 991-1013. http://dx.doi.org/10.1017/S0140525X99002228

Coleman, J. S., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology* (pp 49-56). Somerset, NJ: Association for Computational Linguistics.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI* (pp. 535-555). London, UK: Academic Press.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273-297. http://dx.doi.org/10.1007/BF00994018

Cost, S., & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning, 10*(1), 57-78. http://dx.doi.org/10.1007/BF00993481

Cover, T. M., & Hart, P. E. (1967). Nearast neighbor pattern classification. *Institute of Eletrical Electronics Engineers Transactions on Information Theory, 13*(1), 21-27. http://dx.doi.org/10.1109/TIT.1967.1053964

Crystal, D. (2008). *An encyclopedic dictionary of language and languages*. Oxford, UK: Blackwell Publishing.

Cussens, J. (1998). Using prior probabilities and density estimation for relational classification. In *8th International Conference on Inductive Logic Programming* (*ILP -1998*), 106-115, Madison, WI. http://dx.doi.org/10.1007/BFb0027314

Daelemans, W. (2002). A comparison of analogical modeling to memory-based language processing. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling*: *An exemplar-based approach to language* (pp. 157-180). Amsterdam, NL: John Benjamins.

Daelemans, W., Berck, P., & Gillis, S. (1997). Data mining as a method for linguistic analysis: Dutch diminutives. *Folia Linguistica, 31*(1-2), 57-75. http://dx.doi.org/10.1515/flin.1997.31.1-2.57

Daelemans, W., & Van den Bosch, A. (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., Van den Bosch, A., & Weijters, A. (1997). IGTree: Using trees for compression and classification in lazy learning algorithms. *Artificial Intelligence Review, 11*(1-5), 407-423.

Daelemans, W., Zavrel, J., Berck, P., & Gills, S. (1996). MBT: A memory-based part of speech tagger-generation. *Proceedings of the 4th Workshop on Very Large Corpora* (*WVLC-4*), 14-27, Copenhagen, DK.

Daelemans, W., Zavrel, J., Van der Sloot, K., & Van den Bosch, A. (2010). *TiMBL*: *Tilburg memory based learner, version 6.3, reference guide.* ILK Research Group Technical Report Series no. 10-01. Retrieved from http://ilk.uvt.nl/downloads/pub/papers/ilk.1001.pdf.

Dasarathy, B. V., & Sheela, B. V.(1979). A composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5), 708-713. http://dx.doi.org/10.1109/PROC.1979.11321

Davis, S. (1988). *Topics in syllable geometry*. New York, NY: Garland.

DeCoster, J., Iselin, A. R., & Gallucci, M. (2009). A conceptual and empirical examination of justifications for dichotomization. *Psychological Methods*, *14*(4), 349-366. http://dx.doi.org/10.1037/a0016956

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283-321. http://dx.doi.org/10.1037//0033-295X.93.3.283

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*(2), 124-142. http://dx.doi.org/10.1016/0749-596X(88)90070-8

Dell, G.S., Chang, F., Griffin, M. Z. (1999). Connectionist models of language production: lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517-542. http://dx.doi.org/10.1016/S0364-0213(99)00014-2

DeRose, S. J. (1988). Grammatical category disambiguation by statistical optimization. *Computational Linguistics, 14*(1), 31-39.

De Saussure, F. (1916/1966). *Course in general linguistics*. New York, NY: McGraw-Hil.

Della Pietra, S., Della Pietra, V. J., & Lafferty, J. D. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(4), 380-393. http://dx.doi.org/10.1109/34.588021

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, *26*(3), 297-302. http://dx.doi.org/10.2307/1932409

Dimitrijević, S. (2007). *Kognitivne strategije u obradi jezika*: *Primjena kontekstualnih jezičkih informacija u zadatku automatske lematizacije* (Nepublikovani magistarski rad). Filozofski fakultet, Univerzitet u Banjoj Luci, Banja Luka.

Dimitrijević, S., Milin, P., i Kostić, A. (2008). Primjena kontekstualnih jezičkih informacija na nivou vrsta riječi u zadatku automatske lematizacije. *XIV Naučni skup – Empirijska istraživanja u psihologiji* (pp. 19-20), Beograd, RS: Filozofski fakultet, Univerzitet u Beogradu.

Dimitrijević, S., Kostić, A., & Milin, P. (2009). Stability of the syntagmatic probability distributions. *Psihologija*, *42*(1), 107-119.

Dimitrijević, S. (2011). Kontekst i vrsta riječi kao faktori tačnosti automatske lematizacije. *Radovi, 14,* 67-87.

Di Sciullo, A. M., & Williams, E. (1987). *On the Definition of Word*. Cambridge, MA: MIT Press.

Domingos, P. (1995). Rule induction and instance-based learning: A unified approach. *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (*IJCAI- 95*). 1226-1232, San Mateo, CA.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155-161.

Dudani, S. A. (1976). The distance-weighted k-nearest neighbor rule. *In IEEE Transactions on Systems, Man, and Cybernetics, SMC-6*(4), 325-327. http://dx.doi.org/10.1109/TSMC.1976.5408784

Džeroski, S., Erjavec, T., & Zavrel, J. (2000). Morphosyntactic tagging of Slovene: Evaluating taggers and tagsets. *Proceedings of the 2nd Conference on Language Resources and Evaluation* (*LREC-2000*), 1099-1104, Athens, GR: European Language Resources Association (ELRA).

Eddington, D. (2000). Analogy and the dual-route model of morphology. *Lingua*, *110*(4), 281-289. http://dx.doi.org/10.1016/S0024-3841(99)00043-1

Eddington, D. (2002a). Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics, 9*(1), 49-75. http://dx.doi.org/10.1076/jqul.9.1.49.8482

Eddington, D. (2002b). A comparison of two analogical models: Tilburg memory-based learner versus analogical modeling. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling*: *An exemplar-based approach to language* (pp. 141-156). Amsterdam, NL: John Benjamins.

Elworthy, D. (1994). Does Baum-Welch re-estimation help taggers? *Proceedings of the 4th Applied Natural Language Processing Conference* (*ANLP 4*), 53-58, Stuttgart, DE. http://dx.doi.org/10.3115/974358.974371

Faraway, J. J. (2004). *Linear models with R.* Boca Raton, FL: A CRC Press Company.

Faraway, J. J. (2005). *Extending the linear model with R*: *generalized linear, mixed effects and nonparametric regression models*. Boca Raton, FL: A CRC Press Company.

Farquad, M. A. H., Ravi, V., & Bapi, R. S. (2010). Support vector machine based hybrid classifiers and rule extraction thereof: Application to bankruptcy prediction in banks. In E. S., J. Olivas, D. M. Guerrero, M. M. Sober, & J. R. M. Benedito (Eds.), *Handbook of research on machine learning applications and trends*: *Algorithms, methods, and techniques* (pp. 404-426). Hershey, PA: IGI Global.

Feldman, L. B, & Turvey, M. T. (1983). Word recognition in Serbo-Croatian is phonologically analytic. *Journal of Experimental Psychology*: *Human Perception and Performance, 9*(2), 288-298. http://dx.doi.org/10.1037//0096-1523.9.2.288

Fix, E., & Hodges, J. L., (1951). *Discriminatory analysis-nonparametric discrimination; consistency properties*. Tehnical Report Project 21-49-004, Report No. 4. Randolph Field, TX: School of Aviation Medicine.

Forster, K. I., & Shen, D. (1996). Neighborhood frequency and density effects in visual word recognition. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition, 22*, 696-713.

Frisch, S., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on processing non-words. *Journal of Memory and Language, 42*, 481-496.

Gale, W., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics, 2*(3), 217-237. http://dx.doi.org/10.1080/09296179508590051

Garlock, V., Walley, A., & Metsala, J. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language, 45*(*3*), 468-492. http://dx.doi.org/10.1006/jmla.2000.2784

Giménez, J., & Màrquez, L. (2003). Fast and accurate part-of-speech tagging: The SVM approach revisited. *Proceedings of International Conference Recent Advances in Natural Language Processing* (*RANLP*-2003), 153-163, Borovets, BG. http://dx.doi.org/10.1075/cilt.260.17gim

Giménez, J., & Màrquez, L. (2004): SVMTool: A general POS tagger generator based on support vector machines. *Proceedings of the 4th International Conference on Language Resources and Evaluation* (*LREC-2004)*, 43-46, Lisbon, PT.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1989). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language, 28*(5), 501-518. http://dx.doi.org/10.1016/0749-596X(89)90009-0

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*(3-4), 237-264. http://dx.doi.org/10.2307/2333344

Grainger, J., Muneaux, M., Farioli, F., & Ziegler, J. C. (2005). Effects of phonological and orthographic neighbourhood density interact in visual word recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *58*(6), 981-998. http://dx.doi.org/10.1080/02724980443000386

Guggenheimer, H. (1977). *Applicable geometry*: *Global and local convexity.* Huntington, WV: Krieger.

Güngör, T. (2010). Part-of-speech tagging. In Indurkhya, N. & Damerau, F. J. (Eds), *Handbook of natural language processing* (pp. 205-236). Boca Raton, FL: Taylor and Francis Group, LLC.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, *41*(4), 313-360. http://dx.doi.org/10.1006/cogp.2000.0737

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. J. Keyser (Eds.), *The View from Building 20*: *Essays in Linguistics in Honor of Sylvain Bromberger*, (pp. 111-176). Cambridge, MA: MIT Press,

Harm, M., & Seidenberg, M.S. (1999). Reading acquisition, phonology, and dyslexia: Insights from a connectionist model. *Psychological Review*, *106*(3), 491-528. http://dx.doi.org/10.1037//0033-295X.106.3.491

Harm, M., & Seidenberg, M.S. (2004). Computing the meanings of words in reading: Division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662-720. http://dx.doi.org/10.1037/0033-295X.111.3.662

Harris, Z. S. (1951). *Methods in structural linguistics*. Chicago, IL: University of Chicago Press.

Harris, Z. S. (1957). Co-occurrence and transformation in linguistic structure. *Language*, *33*(3), 283-340. http://dx.doi.org/10.2307/411155

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences, 9*(7), 342-348. http://dx.doi.org/10.1016/j.tics.2005.04.002

Hare, M., Elman, J. L., & Daugherty, K. G. (1995). Default generalization in connectionist networks. *Language and Cognitive Processes, 10*(6), 601-630.

Haugen, E. (1956). The syllable in linguistic description. In M. Halle, H. Lunt, H., & H. McLean. (Eds.), *For Roman Jakobson*: *Essays on the occasion of his sixtieth birthday*, (pp. 213-221). The Hague, NL: Mouton.

Hendrickx, I., & Van den Bosch, A. (2005). Hybrid algorithms with instance-based classification. In J. Gama, R. Camacho, P. Brazdil, A. Jorge, & L. Torgo, L. (Eds.), *Machine Learning*: *ECML 2005*. 16th European Conference on Machine Learning, Porto, Portugal (pp. 158-169). Berlin, DE: Springer. http://dx.doi.org/10.1007/11564096_19

Hockett, C. (1955). *A manual of phonology*. Chicago, IL: University of Chicago Press.

Huntsman, L. A., & Lima, S. D. (2002). Orthographic neighbors and visual word recognition. *Journal of Psycholinguistic Research*, *31*(3), 289-306. http://dx.doi.org/10.1023/A:1015544213366

Hsu, C. H, Chang, C. C., & Lin, C. J. (2010). *Practical guide to support vector classication*. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/ guide.pdf.

Hsu, C.W., & Lin, C.J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks 13*(2), 415-425.

Ilić, N., i Kostić, A. (2002). Problem homografije pri automatskoj lematizaciji. *VIII Naučni skup – Empirijska istraživanja u psihologiji* (pp. 36-37). Beograd, RS: Filozofski fakultet, Univerzitet u Beogradu.

Ivanciuc, O. (2007). Applications of support vector machines in chemistry. In K. B. Lipkowitz, & T. R. Cundari (Eds), *Reviews in Computational Chemistry, 23,* (pp. 291-400). Hoboken, NJ: John Wiley & Sons, Inc.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446. http://dx.doi.org/10.1016/j.jml.2007.11.007

Jebara, T. (2004). *Machine learning*: *Discriminative and generative*. New York, NY: Springer Science+Business Media, LLC.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European Conference on Machine Learning (ECML)*. http://dx.doi.org/10.1007/BFb0026683

Johnson, N. F., & Pugh, K. R. (1994). A cohort model of visual word recognition. *Cognitive Psychology, 26*(3), 240-346. http://dx.doi.org/10.1006/cogp.1994.1008

Jovanović, T., Filipović Đurđević, D. i Milin, P. (2008). Kognitivna obrada alomorfije u srpskom jeziku. *Psihologija, 41*(1), 86-102.

Jusczyk, P. W., Luce, P. A., & Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory & Language, 33,* 630-645.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistics comprehension and production. In R. Bod, J., Hay, & S. Jannedy, S. (Eds.), *Probabilistic Linguistics* (pp. 29-95). Cambridge, UK: MIT Press.

Jurafsky, D. & Martin, J. H. (2000). *Speech and language processing*: *An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kager, R. (1999). *Optimality theory*. Cambridge, UK: Cambridge University Press.

Kecman, V. (2001). *Learning and soft computing*: *Support vector machines, neural networks, and fuzzy logic models.* Cambridge, UK: The MIT Press.

Keerthi, S. S., & Lin, C. J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation, 15*(7), 1667-1689. http://dx.doi.org/10.1162/089976603321891855

Keuleers, E. (2008). *Memory-based learning of inflectional morphology*. (Unpublished doctoral thesis). Faculteit Lettern en Wijsbergeerte, Universiteit Antwerpen, Antwerpen.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods, 42*(3), 627-633. http://dx.doi.org/10.3758/BRM.42.3.627

Keuleers, E., & Dealemans, W. (2007). Memory-based learning models of inflectional morphology: A methodological case study. *Lingue e Linguaggio, 6,* 151-174.

Keuleers, E., & Sandra, D. (2008) *Similarity and productivity in the English past tense*, (Manuscript submitted for publication).

Keuleers, E., Sandra, D., Daelemans, W., Gillis, S., Durieux, G., & Martens, E. (2007). Dutch plural inflection: The exception that proves the analogy. *Cognitive Psychology*, *54*(4), 283-318. http://dx.doi.org/10.1016/j.cogpsych.2006.07.002

Kim, J., & Kim, G. C. (1996). Fuzzy network model for part-of-speech tagging under small training data. *Natural Language Engineering*, *2*(*2*), 95-110. http://dx.doi.org/10.1017/S1351324996001258

Klimova, A., & Rudas, T. (2014). *Iterative scaling in curved exponential families*. arXiv:1307.3282.

Kondrak, G., Marcu, D., & Knight, K. (2003). Cognates can improve statistical translation models. *Proceedings of the of Human Language Technology Conference of the North American* (*HLT-NAACL-2003*). 46-48, Edmonton, CA. http://dx.doi.org/10.3115/1073483.1073499

Kononenko, I. (1995). On biases in estimating multi-valued attributes. *Proceedings of the International Joint Conference on Aartificial Intelligence* (*IJCAI - 95*). 1034-1040, Montreal, CA.

Kostić, A. (1991). Informational approach to processing inflected morphology: Standard data reconsidered. *Psychological Research*, *53*(1), 62-70. http://dx.doi.org/10.1007/BF00867333

Kostić, A. (1995). Informational load constrains on processing inflected morphology. In L. B. Feldman (Ed.), *Morphological Aspects of Language Processing*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Kostić, A. (2004). Kognitivna ograničenja i obrada jezika. U A. Kostić, D. Todorović, S. Marković (Eds.), *Jezik i opažanje. Tri studije iz eksperimentalne psihologije* (pp. 7-51). Beograd, SR: Filozofski fakultet, Univerzitet u Beogradu.

Kostić, A., Ilić, S., & Milin, P. (2008). Aproksimacija verovatnoća i optimalna veličina jezičkog uzorka. *Psihologija*, *41*(1), 35-51.

Kostić, A., Marković i Baucal, A. (2003). Inflectional morphology and word meaning: Orthogonal or co-implicative cognitive domains? In H. Baayen & R. Schreuder, (Eds.), *Aspects of Morphological Processing*. Berlin, DE: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110910186.1

Kostić, Đ. (1999). *Frekvencijski rečnik savremenog srpskog jezika*. Beograd, RS: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu, <htpp://www. serbian-corpus.edu.yu/>.

Kostić, Đ. (2001). *Korpus srpskog jezika*. Beograd, RS: Institut za eksperimentalnu fonetiku i patologiju govora i Laboratorija za eksperimentalnu psihologiju, Univerzitet u Beogradu, <htpp://www.serbian-corpus.edu.yu/>.

Kristal, D. (1988). *Enciklopedijski rečnik moderne lingvistike.* Beograd, RS: Nolit.

Krott, A., Baayen, H., & Schreuder, R. (2001). Analogy in morphology: Modeling the choice of linking morphemes in Dutch. *Linguistics, 39*(1), 51-93. http://dx.doi.org/10.1515/ling.2001.008

Krott, A., Schreuder, R., Baayen, R. H., & Dressler, W. U. (2007). Analogical effects on linking elements in German compounds. *Language and Cognitive Processes*, *22*(1), 25-57. http://dx.doi.org/10.1080/01690960500343429

Kullback, S. (1959). *Information theory and statistics*. New York, NY: Wiley.

Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*(1), 79-86. http://dx.doi.org/10.1214/aoms/1177729694

Kuperman, V., Bertram, R., & Baayen, R. H. (2008). Morphological dynamics in compound processing. *Language and Cognitive Processes*, *23*(7-8), 1089-1132. http://dx.doi.org/10.1080/01690960802193688

Kupiec, J. (1992). Robust part-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, *6*(3), 225-242. http://dx.doi.org/10.1016/0885-2308(92)90019-Z

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 11th International Conference on Machine Learning* (*ICML-2001*), 282-289, Williamstown, AU.

Levelt, W. J. (1999). Models of word production. *Trends in Cognitive Sciences*, *3*(6), 223-232.http://dx.doi.org/10.1016/S1364-6613(99)01319-4

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioural and Brain Sciences, 22*(1), 1-38. http://dx.doi.org/10.1017/S0140525X99001776

Levelt, W. J. M., & Wheeldon, L. R. (1994). Do speakers have access to a mental syllabary? *Cognition, 50*(1-3), 239-269. http://dx.doi.org/10.1016/0010-0277(94)90030-2

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Sovjet Physics Doklady*, *10*(8), 707-710.

Lieber, R. (1992). *Deconstructing morphology*: *Word formation in syntactic theory*. Chicago, IL: University of Chicago Press.

Lin, K. M., & Lin, C. J. (2003). A study on reduced support vector machines, *IEEE Transactions on Neural Networks, 14*(6), 1449-1559.

Luce, P. (1986). Neighborhoods of words in the mental lexicon. In *Research on Speech Perception Technical Report No. 6.* Bloomington: Indiana University, Department Of Psychology, Speech Research Laboratory.

Luce, P. A., & Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language & Cognitive Processes*, *16*(5-6), 565-581. http://dx.doi.org/10.1080/01690960143000137

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, *19*(1), 1-36. http://dx.doi.org/10.1097/00003446-199802000-00001

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19-40. http://dx.doi.org/10.1037//1082-989X.7.1.19

MacLeod, E. S. J, Luk, A, & Titterington, D. M. (1987). A re-examination of the distance-weighted k-nearest neighbor classification rule. *IEEE Transactions on Systems, Man, and Cybernetics, 17*(4), 689-696. http://dx.doi.org/10.1109/TSMC.1987.289362

Majumder, P., Mitra M., & Chaudhuri B. (2002). N-gram: A language independent approach to IR and NLP. *Proceedings in the International Conference on Universal Knowledge and Language* (*ICUKL- 2002*), Goa, IN.

Manning, C. D., & Schuetze, H. (2000). *Foundations of statistical natural language processing*. Cambridge, UK: MIT Press.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In Gelbukh, A. (Ed.), *Proceedings in the 12th International Conference Computational Linguistics and Intelligent Text Processing, CICLing*, 171-189. Berlin, DE: Springer.

Mantaras, L. R. (1989). ID3 revisited: A distance-based criterion for attribute selection. *Proceedings of International Symposium Methodologies for Intelligent Systems*, 342-350, Charlotte, NC.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*(3), 189-256. http://dx.doi.org/10.1006/cogp.1995.1015

Mattys, S. L., & Jusczyk, P.W. (2000). Phonotactic cues for segmentation of fluid speech by infants. *Cognition*, *78*, 91-121.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*(4), 465-494. http://dx.doi.org/10.1006/cogp.1999.0721

Mayfield, J., McNamee, P., Piatko, C., & Pearce, C. (2003). Lattice-based tagging using support vector machines. *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM-2003)*, 303-308, New Orleans, LA. http://dx.doi.org/10.1145/956919.956921

Meyer, D., Leisch, F., Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*(1-2), 169-186. http://dx.doi.org/10.1016/S0925-2312(03)00431-4

Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, *20*(2), 155-172.

Michie, D. (1989). Personal models of rationaly. *Journal of Statistical Planning and Inference, Special Issue on Foundations and Philosophy of Probabily and Statistics*, *21*, 381-399.

Milin, P. (2004). *Probabilistički pristup određivanju gramatičkog statusa reči i kognitivne strategije u obradi jezika* (Nepublikovana doktorska disertacija). Filozofski fakultet, Univerzitet u Beogradu; Beograd.

Milin, P. (2005). Istraživanja jezičkih fenomena pomoću računarskih simulacija obrade prirodnog jezika. *XI Naučni skup – Empirijska istraživanja u psihologiji* (pp 39-40). Beograd, RS: Filozofski fakultet, Univerzitet u Beogradu.

Milin, P., Filipović Đurđević, D., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian. *Journal of Memory and Language*, *60*(1), 50-64. http://dx.doi.org/10.1016/j.jml.2008.08.007

Milin, P., Keuleers, E., & Filipović Đurđević, D. (2011). Allomorphic responses in Serbian pseudo-nouns as a result of analogical learning. *Acta Linguistica Hungarica*, *58*(1), 65-84. http://dx.doi.org/10.1556/ALing.58.2011.1-2.4

Milin, P., Kuperman, V., Kostić, A., & Baayen, H. R. (2009). Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In J. P. Blevins, & J. Blevins (Eds.), *Analogy in Grammar*: *Form and Acquisition* (pp. 214-252). Oxford, UK: Oxford University Press

Milin, P., Ramscar, M., Cho, K., Baayen, R. H., & Feldman, L. B. (2014). *Processing partially and exhaustively decomposable words*: *An amorphous approach based on discriminative learning*. Manuscript under review.

Moore, A. W. (2001, 2003). Support vector machine. *Tutorial slides*. Retrieved from http://www.cse.msu.edu/~cse802/Papers/svmjain.ppt#452,65, Diffusion Kernel.

Morgan, J. L., & Saffran, J. R. (1995). Emerging integration of sequential and suprasegmental information in preverbal speech segmentation. *Child Development*, *66*(4), 911-936. http://dx.doi.org/10.2307/1131789

Moscoso del Prado Martin, F., Kostic, A., & Baayen, R. H. (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*, *94*(1), 1-18. http://dx.doi.org/10.1016/j.cognition.2003.10.015

Nakagawa, T., Kudo, T., & Matsumoto, Y. (2001). Unknown word guessing and part-of-speech tagging using support vector machines. *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium* (NLPRS-2001), 325-331, Tokyo, JP.

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification and recognition. In E. G. Ashby (Ed.), *Multidimensional models of perception and cognitio*n (pp. 363-393). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition*, *18*(2), 211-233. http://dx.doi.org/10.1037//0278-7393.18.2.211

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*(2), 266-300. http://dx.doi.org/10.1037//0033-295X.104.2.266

Olson, D. L., & Dulen, D. (2008). *Advanced data mining techniques.* Berlin, DE: Springer-Verlag.

Padró, L. (1996). Pos tagging using relaxation labelling. *Proceeding of the 16th International Conference on Computational Linguistics* (COLING-1996), 877-882, Copenhagen, DK. http://dx.doi.org/10.3115/993268.993320

Peshkin, L., Pfeffer, A., & Savova, V. (2003). Bayesian nets in syntactic categorization of novel words. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL-2003*), 79-81, Edmonton, CA. http://dx.doi.org/10.3115/1073483.1073510

Peshkin, L., & Savova, V. (2003). Why build another part-of-speech tagger? A minimalist approach. *Proceedings of International Conference Recent Advances in Natural Language Processing* (*RANLP-2003*). Borovets, BG.

Pierrehumbert, J. B. (2003a). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probability theory in linguistics* (pp. 177-228). Cambridge, UK: MIT Press.

Pierrehumbert, J. B. (2003b). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(2-3), 115-154. http://dx.doi.org/10.1177/00238309030460020501

Pinker, S. (1991). Rules of language. *Science*, *253*(5019), 530-535. http://dx.doi.org/10.1126/science.1857983

Pinker, S. (1999). *Words and rules*. London, UK: Phoenix.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*(1-2), 73-193. http://dx.doi.org/10.1016/0010-0277(88)90032-7

Pinker, S., & Prince, A. (1994). Regular and irregular morphology and the psychological status of rules of grammar. In S. D. Lima, R. L. Corrigan, & G. K. Iverson (Eds.), *The reality of linguistic rules* (pp. 321-351). Amsterdam, NL: John Benjamins. http://dx.doi.org/10.1075/slcs.26.21pin

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, *39*(3), 347-370. http://dx.doi.org/ 10.1006/jmla.1998.2571

Plaut, D. C., & Gonnerman, L.M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4-5), 445-485. http://dx.doi.org/10.1080/ 01690960050119661

Prasada, S. & Pinker, S. (1993). Generalizations of regular and irregular morphological patterns. *Language and Cognitive Processes*, *8*, 1-56.

Prasolov, V. V., & Tikhomirov, V. M. (2001). Geometry. *Translations of Mathematical Monographs, 200.* American Mathematical Society.

Prince, A., & Smolensky, P. (1993). *Optimality theory*: *Constraint interaction in generative grammar*. New Brunswick, CA: Rutgers University.

R Development Core Team (2011). R: A language and environment for statistical computing (Version 2.13.1.) [Computer Software]. R Foundation for Statistical Computing, Austria: Vienna. Available from http://www.R-project.org.

Ramscar, M., & Baayen, R. H. (2013) Production, comprehension, and synthesis: A communicative perspective on language. *Frontiers in Psychology. 4*:*233*. http://dx.doi.org/10.3389/fpsyg.2013.00233.

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science*, *31*(6), 927-960. http://dx.doi.org/10.1080/03640210701703576

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. *Proceedings of the 1st Conference on Empirical Methods in Natural Language Processing* (*EMNLP-1996*), *1*, 133-142. Philadelphia, PA.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249-284.

Roelofs, A. (2000). WEAVER++ and other computational models of lemma retrieval and word-form encoding. In L. Wheeldon (Ed.), *Aspects of language production* (pp. 71-114). Sussex, UK: Psychology Press.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel Distributed Processing. Explorations in the Microstructure of Cognition: Volume. 2. Psychological and Biological Models* (pp. 216-271). Cambridge, UK: MIT Press.

Quinlan, J. R. (1993). *C4.5*: *Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Salzberg, S. (1991). A nearest hyperrectangle learning method. *Machine Learning*, *6*(3), 277-309. http://dx.doi.org/10.1007/BF00114779

Scalise, S. (1984). Generative morphology. *Studies in generative grammar*, *18*. Dordrecht, NL: Foris. http://dx.doi.org/10.1515/9783110877328

Schmid, H. (1994a). Part-of-speech tagging with neural networks. *Proceeding of the 15th International Conference on Computational Linguistics* (*COLING-1994*), 172-176. Kyoto, JP. http://dx.doi.org/10.3115/991886.991915

Schiller, N.O. (2006). Lexical stress encoding in single word production estimated by event-related brain potentials. *Brain Research*, *1112*(1), 201-212. http://dx.doi.org/ 10.1016/j.brainres.2006.07.027

Schmid, H. (1994b). Probabilistic part-of-speech tagging using decision trees. *Proceeding of the 2nd International Conference on New Methods in Language Processing* (*NEMLP-1994*), 44-49. Manchester, UK.

Sears, C. R., Hino, Y., & Lupker, S. J. (1995). Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology*: *Human Perception and Performance*, *21*(4), 876-900. http://dx.doi.org/10.1037//0096-1523.21.4.876

Seidenberg, M. S., & Gonnerman, L.M. (2000). Explaining derivational morphololy as the convergence of codes.*Trends in Cognitive Sciences*, *4*, 353-361.

Sečujski, M., i Kupusinac, A (2009). Poređenje postupaka automatske morfološke anotacije tekstovana srpskom jeziku. *XVII Telekomunikacioni forum TELFOR*. Beograd, RS: Društvo za telekomunikacije. Retrieved from http://2009.telfor.rs/files/radovi/ 09_44.pdf.

Selkirk, E. (1982). *The syntax of words*. Cambridge, MA: MIT Press.

Sewell, M. (2006). *Structural risk minimization*. Retrieved from http://www. svms.org/srm/srm.pdf.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50-64. http://dx.doi.org/10.1002/j.1538-7305.1951.tb01366.x

Shepard, R.N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317-1323. http://dx.doi.org/10.1126/science.3629243

Sheskin, J. D. (2000). *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall /CRC.

Sindhiya, B. A., Anand, P., & Soman , K. P. (2009). SVM based approach to Telugu parts of speech tagging using SVMTool. *International Journal of Recent trends in Engineering, 1*(*2*), 183-185.

Skousen, R. (1989). *Analogical modeling of language*. Dordrecht, NL: Kluwer Academic Publishers.

Skousen, R. (1992). *Analogy and structure*. Dordrecht, NL: Kluwer Academic Publishers.

Skousen, R. (1995). Analogy: A non-rule alternative to neural networks. *Rivista di Linguistica, 7*, 213-232.

Skousen, R. (2002a). Introduction. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language* (pp. 1-8). Amsterdam, NL: John Benjamins.

Skousen, R. (2002b). An overiew af Analogical modeling. In R. Skousen, D. Lonsdale, & D. Parkinson (Eds.), *Analogical modeling: An exemplar-based approach to language* (pp. 11-48). Amsterdam, NL: John Benjamins.

Skousen, R. (2009). Expanding analogical modeling into general theory of language prediction. In R. J.P. Blevins, & J. Blevins (Eds.), *Analogy in grammar, form and acquisition* (pp. 164-184). New York: NY: Oxford University Press Inc.

Smyth, P. & Goodman, R.M. (1991). Rule induction using information theory. In. G. Piatetsky-Shapiro and W.Frawley (Eds.), *Knowledge discovery in databases* (pp. 159-176). Cambridge, UK: MIT Press.

Stanfill, C., & Waltz, D. (1986). Toward memory-based reasoning. *Communactions of the ACM, 29*(12), 1213-1228. http://dx.doi.org/10.1145/7902.7906

Stanojčić, Ž., i Popović, LJ. (1997). *Gramatika srpskog jezika. Udžbenik za I, II, III i IV razred srednje škole.* Beograd, RS: Zavod za udžebnike i nastavna sredstva.

Stevanović, M. (1975). Savremeni srpskohrvatski jezik. Beograd, RS: Naučna knjiga.

Steinwart, I., & Christmann, A. (2008). *Support vector machine*s. New York, NY: Springer.

Storkel, H. L. (2001). Learning new words: Phonotactic probabilities in language development. *Journal of Speech, Language, and Hearing Research*, *44*(6), 1321-1337. http://dx.doi.org/10.1044/1092-4388(2001/103)

Storkel, H. L. (2004). The emerging lexicon of children with phonological delays. *Journal of Speech, Language, and Hearing Research*, *47*(5), 1194-1212. http://dx.doi.org/10.1044/1092-4388(2004/088)

Storkel, H. L., & Morrisette, M. L. (2002). The lexicon and phonology: Interactions in language acquisition. *Language, Speech, and Hearing Services in Schools*, *33*(1), 24-37. http://dx.doi.org/10.1044/0161-1461(2002/003)

Storkel, H. L., & Rogers, M. A. (2000). The effect of probabilistic phonotactics on lexical acquisition. *Clinical Linguistics & Phonetics*, *14*, 407-425.

Stump, G. (1991). A paradigm-based theory of morphosemantic mismatches. *Language*, *67*(4), 675-725. http://dx.doi.org/10.2307/415074

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston, MA: Allyn and Bacon.

Tremblay, A., & Baayen, R. H. (2010). Holistic processing of regular four-word sequences: A behavioral and erp study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language acquisition and communication* (pp. 151-173). London, UK: Continuum International.

Tufis, D., Dienes, P., Oravecz, C., & Varadi, T. (2000). Principled hidden tagset design for tiered tagging of Hungarian. *Proceeding of the 2nd International Conference on Language Resources and Evaluation* (*LREC-2000*), 1421-1426. Athens, GR.

Van den Bosch, A. (1999a). Careful abstraction from instance families in memory-based language learning. *Journal for Experimental and Theoretical Artificial Intelligence*, *11*(3), 339-368. http://dx.doi.org/10.1080/095281399146454

Van den Bosch, A. (1999b). Instance-family abstraction in memory-based language learning. In I. Bratko, & S. Džeroski (Eds.), *Machine Learning*: *Proceedings of the Sixteenth International Conference* (*ICML'99*), 39-48, Bled, SI.

Van den Bosch, A. (2000). Using induced rules as complex features in Memory-based language learning. *Proceedings of the 4th Conference on Computational Language Learning and the 2nd Learning Language in Logic Workshop* (*CoNLL-2000 and LLL-2000*), 73-78, Lisbon, PT.

Van Gestel, T., Suykens J., Baesens B., Viaene S., Vanthienen J., Dedene G., De Moor B., Vandewalle J. (2004). Benchmarking least squares support vector machine classifiers. *Machine Learning*, *54*(1), 5-32. http://dx.doi.org/10.1023/B:MACH.0000008082.80494.e0

Vapnik, V. N. (1979). *Estimation of dependences based on empirical data* [in Russian]. Moscow, RU: Nauka. (1982) New York, NY: Springer-Verlag, English translation, 1982.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag New York.

Vapnik, V. N. (1998). *Statistical learning theory*. New York, NY: John Wiley and Sons, Inc.

Vapnik, V. N., & Chervonenkis, A. Y. (1968). On the uniform convergence of relative frequencies of events to their probabilities. *Soviet Mathematics Doklady*, *9*, 915-918.

Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, *16*(2), 264-280. Translated by B. Seckler. http://dx.doi.org/10.1137/1116025

Vert, J.-P., Tsuda, K., & Schölkopf, B. (2004). A primer on kernel methods. In B. Schölkopf, K. Tsuda, & J.-P. Vert (Eds.), *Kernel methods in computational biology* (pp. 35-70). Cambridge, UK: MIT Press, A Bradford Book.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics*, *17*(6), 487-499. http://dx.doi.org/10.1080/0269920031000107541

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325-329. http://dx.doi.org/10.1111/1467-9280.00064

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374-408. http://dx.doi.org/10.1006/jmla.1998.2618

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers 36*(3), 481-487. http://dx.doi.org/10.3758/BF03195594

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language & Speech*, *40*, 47-62.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain & Language*, *68*(1-2), 306-311. http://dx.doi.org/10.1006/brln.1999.2116

Vitevitch, M. S., Pisoni, D. B., Kirk, K. I., Hay-McCutcheon, M., & Yount, S. L. (2002). Effects of phonotactic probabilities on the processing of spoken words and nonwords by postlingually deafened adults with cochlear implants. *Volta Review*, *102*, 283-302.

Voutilainen, A. (1999). Orientation. In H. van Halteren (Ed), *Syntactic wordclass tagging* (pp. 3-7). Dordrecht, NL: Kluwer Academic Publishers.

Voutilainen, A., Heikkilä, J., & Anttila, A. (1992). *Constraint grammar of English. A performance-oriented introduction*, *21*, Helsinki, FI: University of Helsinki.

Vujaklija, M. (1970). *Leksikon stranih reči izraza*. Beograd, RS: Prosveta.

Wettschereck, D. (1994). A hybrid nearest-neighbor and nearest-hyperrectangle algorithm. In F. Bergadano, L. de Raedt (Eds.), *Machine Learning*: *ECML-94, European Conference on Machine Learning, Catania, Italy* (pp. 323-338). Berlin, DE: Springer. http://dx.doi.org/10.1007/3-540-57868-4_67

White, A. P., & Liu, W. Z. (1994). Bias in information-based measures in decision tree induction. *Machine Learning*, *15*(3), 321-329. http://dx.doi.org/10.1007/BF00993349

Woollams A. M., & Patterson K. (2012). The neural basis of morphology: A tale of two mechanisms?. In M. J. Spivey, K. McRae, & M. F. Joanisse (Eds.), *Cambridge handbook of psycholinguistics*. (pp. 333-352). New York NY: Cambridge University Press.

Wurm, L.H., & Fisicaro, S.A. (2014). What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language*, *72*, 37-48. http://dx.doi.org/10.1016/j.jml.2013.12.003

Xu, J. (2011). An empirical comparison of weighting functions for multi-label distance weighted k-nearest neighbour method. *The First International Conference on Artificial, Intelligence, Soft Computing and Applications* (*AIAA2011*), 13-20, Tirunelvelli, IN.

Yao, Y. (2011). *The Effects of Phonological Neighborhoods on Pronunciation Variation in Conversational Speech* (Unpublished PhD dissertation) University of California, Berkeley.

Yates, M., Locker, L., & Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, *11*(3), 452-457. http://dx.doi.org/10.3758/BF03196594

Yip, M. (2003). Casting doubt on the onset-rime distinction. *Lingua*, *113*(8), 779-816. http://dx.doi.org/10.1016/S0024-3841(02)00130-4

Zavrel, J. 1997. An empirical re-examination of weighted voting for k-NN. In W. Daelemans, P. Flach, and A. Van den Bosch, (Eds.), *Proceedings of the 7th Belgian-Dutch Conference on Machine Learning*, 139-148, Tilburg, NL.

Zec, D. (2000). O strukturi sloga u srpskom jeziku. *Južnoslovenski filolog,* LVI/1-2, 435-448. Srpska akademija nauka i umetnosti i Institut za srpski jezik SANU. Beograd.

Zec, D. (2007). The Syllable. In P. de Lacy (Ed.) *Handbook of phonological theory*, (pp. 161-194). Cambridge, UK: Cambridge University Press.

Ziegler, J. C., Muneaux, M., & Grainger, J. (2003). Neighborhood effects in auditory word recognition: Phonological competition and orthographic facilitation. *Journal of Memory and Language*, *48*(4), 779-793. http://dx.doi.org/10.1016/S0749-596X(03)00006-8

Zipf, G. K. (1935). *The psycho-biology of language*: *An introduction to dynamic philology.* Boston, MA: Houghton Mifflin Co.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*: *An introduction to human ccology*. New York, NY: Hafner.

Zhang, B. (2001). *Is the maximal margin hyperplane special in a feature space?* Technical Report HPL-2001-89, Hewlett-Packards Labs. Retrieved from http://www.svms.org/hyperplane/Zhan01.pdf.

Zhang, J. (1992). Selecting typical instances in instance-based learning. *Proceedings of the 9th International Conference on Machine Learning* (*ML-1992*), 470-479, Aberdeen, UK. http://dx.doi.org/10.1016/B978-1-55860-247-2.50066-8

Zwicky, A. (1989). Inflectional morphology as a (sub)component of grammar. In W. Dressler, H. Luschützky O. Pfeiffer, & J. Rennison (Eds.) *Contemporary morphology*, (pp. 216-236). Berlin, DE: Walter de Gruyter. http://dx.doi.org/10.1515/9783110874082.217

PRILOZI

Prilog 1. *Rezultati klasifikacije vrsta riječi na osnovu bigrama, pomoću* mašina sa vektorima podrške

Prilog 1a. *Parametar C, broj vektora i tačnost klasifikacije vrsta riječi na osnovu bigrama, pomoću* mašina sa vektorima podrške

		[#x]	[xy]	[x#]	[#x, x#, xy]
Parametar C		3.000	0.281	0.375	0.281
Broj podržavajućih vektora		411	357	627	330
Broj graničnih podr.vektora		142	86	373	58
Broj podržavajućih vektora po klasi	imenice	49	43	46	41
	pridjevi	84	48	102	42
	zamjenice	126	115	256	110
	glagoli	152	151	223	137
Tačnost (%)	uzorak za učenje	83.59	98.56	70.07	99.17
	test uzorak	78.70	91.36	67.28	93.21
	ukupno	82.37	96.75	69.37	97.68

Prilog 1b. *Tačnost klasifikacije na osnovu bigrama na početku riječi* [#x] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	6	30.00	14	70.00	30.00	10.00	10.00	50.00
pridjevi	12	42.86	16	57.14	14.29	42.86	32.14	10.71
zamjenice	142	91.03	14	8.97	0.00	0.64	91.03	8.33
glagoli	95	79.17	25	20.83	1.67	5.83	13.33	79.17

Prilog 1c. *Tačnost klasifikacije na osnovu bigrama na kraju riječi* [x#] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	4	20.00	16	80.00	20.00	0.00	55.00	25.00
pridjevi	4	14.29	24	85.71	3.57	14.29	75.00	7.14
zamjenice	133	85.26	23	14.74	0.00	0.64	85.26	14.10
glagoli	77	64.17	43	35.83	0.00	0.00	35.83	64.17

Prilog 1d. *Tačnost klasifikacije na osnovu bigrama* [xy], *bez obzira na njihovu poziciju u riječima, sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	11	55.00	9	45.00	55.00	10.00	10.00	25.00
pridjevi	22	78.57	6	21.43	3.57	78.57	0.00	17.86
zamjenice	152	97.44	4	2.56	0.00	0.00	97.44	2.56
glagoli	111	92.50	9	7.50	2.50	0.00	5.00	92.50

Prilog 1e. *Tačnost klasifikacije na osnovu svih bigrama* [#x, x#, xy] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	13	65.00	7	35.00	65.00	5.00	5.00	25.00
pridjevi	21	75.00	7	25.00	7.14	75.00	3.57	14.29
zamjenice	155	99.36	1	0.64	0.00	0.00	99.36	0.64
glagoli	113	94.17	7	5.83	0.83	0.00	5.00	94.17

Prilog 2. *Rezultati klasifikacije vrsta riječi na osnovu trigrama, pomoću* mašina sa vektorima podrške

Prilog 2a. *Parametar C, broj vektora i tačnost klasifikacije vrsta riječi na osnovu trigrama, pomoću* mašina sa vektorima podrške

		[#xy]	[xyz]	[xy#]	[#xy, xy#, xyz]
Parametar C		1.750	0.500	0.500	0.125
Broj podržavajućih vektora		416	481	502	446
Broj graničnih podr.vektora		187	244	207	217
Broj podržavajućih vektora po klasi	imenice	41	34	38	35
	pridjevi	66	47	83	48
	zamjenice	152	195	211	185
	glagoli	157	205	170	178
Tačnost (%)	uzorak za učenje	86.79	91.74	87.72	95.56
	test uzorak	81.79	84.57	83.64	92.28
	ukupno	85.54	89.95	86.70	94.74

Prilog 2b. *Tačnost klasifikacije na osnovu trigrama na početku riječi* [#xy] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	10	50.00	10	50.00	50.00	15.00	10.00	25.00
pridjevi	24	85.71	4	14.29	7.14	85.71	3.57	3.57
zamjenice	139	89.10	17	10.90	0.00	6.41	89.10	4.49
glagoli	92	76.67	28	23.33	0.83	4.17	18.33	76.67

Prilog 2c. *Tačnost klasifikacije na osnovu trigrama na kraju riječi* [xy#] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	9	45.00	11	55.00	45.00	15.00	25.00	15.00
pridjevi	13	46.43	15	53.57	3.57	46.43	39.29	10.71
zamjenice	143	91.67	13	8.33	0.64	0.00	91.67	7.69
glagoli	106	88.33	14	11.67	0.00	0.00	11.67	88.33

Prilog 2d. *Tačnost klasifikacije na osnovu trigrama, bez obzira na njihovu poziciju u riječima* [xyz], *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	10	50.00	10	50.00	50.00	15.00	15.00	20.00
pridjevi	21	75.00	7	25.00	3.57	75.00	3.57	17.86
zamjenice	150	96.15	6	3.85	0.00	0.00	96.15	3.85
glagoli	93	77.50	27	22.50	0.00	0.00	22.50	77.50

Prilog 2e. *Tačnost klasifikacije na osnovu svih trigrama* [#xy, xy#, xyz] *sa matricom konfuzije*

Vrsta riječi	tačno		pogrešno		Pripisana vrsta riječi (%)			
	N	%	N	%	imenice	pridjevi	zamjenice	glagoli
imenice	12	60.00	8	40.00	60.00	0.00	20.00	20.00
pridjevi	23	82.14	5	17.86	0.00	82.14	7.14	10.71
zamjenice	154	98.72	2	1.28	0.00	0.00	98.72	1.28
glagoli	110	91.67	10	8.33	0.00	0.00	8.33	91.67

Prilog 3. *Značajnost razlika između tačnosti klasifikacija (df = 1)*

	McNemar's χ^2	p	korigovano p
[#x] : [xy]	27.119	.0001	.0005
[#xy] : [xy]	16.981	.0001	.0005
[xy#] : [xy]	10.105	.0015	.005
[xyz] : [xy]	9.587	.002	.005
[#x] : [xyz]	4.563	.033	.065
[#x] : [xy#]	2.885	.089	.149
[#x] : [#xy]	1.191	.275	.372
[#xy] : [xyz]	1.085	.298	.372
[#xy] : [xy#]	0.329	.566	.629
[xy#] :[xyz]	0.062	.804	.804

Prilog 4. *Primjena radijalnog kernela (RBF) u klasifikaciji vrsta riječi*

Prilog 4a. *Primjena radijalnog kernela (RBF) u klasifikaciji vrsta riječi u slučaju bigrama na početku riječi* [#x]

C \ γ	2^{-15}	2^{-13}	2^{-11}	2^{-9}	2^{-7}	2^{-5}	2^{-3}	2^{-1}	2^{1}	2^{3}
2^{-5}	48.15	48.15	48.15	48.15	48.15	58.33	63.27	70.99	72.22	70.37
2^{-3}	48.15	48.15	48.15	48.15	58.33	62.96	74.07	76.24	72.53	70.37
2^{-1}	48.15	48.15	48.15	58.33	62.65	71.61	77.47	79.63	77.78	75.00
2^{1}	48.15	48.15	58.33	62.96	72.53	75.93	79.01	80.25	79.32	77.16
2^{3}	48.15	58.33	62.96	72.53	75.93	77.47	80.25	80.56	79.63	77.16
2^{5}	58.33	62.96	72.53	75.93	76.24	81.17	80.86	79.94	79.01	78.09
2^{7}	62.96	72.53	75.93	76.85	78.09	81.79	79.63	70.68	49.07	50.93
2^{9}	72.53	75.93	76.54	78.09	80.86	74.07	73.15	54.63	26.85	45.37
2^{11}	75.93	76.54	77.78	81.48	64.51	66.05	70.06	41.98	43.52	9.26
2^{13}	76.54	77.78	80.86	70.06	56.48	34.26	34.57	41.98	44.14	9.26
2^{15}	78.09	79.01	59.88	45.99	45.68	37.65	34.57	41.36	44.14	9.26

Prilog 4b. *Primjena radijalnog kernela (RBF) u klasifikaciji vrsta riječi u slučaju svih bigrama* [#x, x#, xy]

C \ γ	2^{-15}	2^{-13}	2^{-11}	2^{-9}	2^{-7}	2^{-5}	2^{-3}	2^{-1}	2^{1}	2^{3}
2^{-5}	48.15	48.15	48.15	53.09	56.79	62.96	67.59	48.15	48.15	48.15
2^{-3}	48.15	48.15	53.39	56.79	63.89	80.86	77.16	51.85	48.76	48.76
2^{-1}	48.15	53.70	56.79	65.43	84.88	88.27	83.95	69.14	58.64	57.72
2^{1}	53.70	56.79	65.43	86.11	90.74	91.98	86.73	72.84	59.88	57.72
2^{3}	56.79	65.74	87.04	90.12	93.52	94.14	86.73	72.84	59.88	57.72
2^{5}	65.74	87.04	90.12	93.52	94.14	93.52	86.73	72.84	59.88	57.72
2^{7}	87.04	90.12	93.21	93.21	94.14	93.52	86.73	72.84	59.88	57.72
2^{9}	90.12	93.21	93.21	92.28	93.83	93.52	86.73	72.84	59.88	57.72
2^{11}	93.21	93.21	92.28	91.67	93.83	93.52	86.73	72.84	59.88	57.72
2^{13}	93.21	91.98	91.36	91.67	93.83	93.52	86.73	72.84	59.88	57.72
2^{15}	91.98	91.05	91.36	91.67	93.83	93.52	86.73	72.84	59.88	57.72

Prilog 5. *Stimulusi korišćeni u eksperimentu*

Klaster pogrešnih rješenja		Klaster tačnih rješenja	
manevri	madraci	oblaci	alati
metalci	prsti	izlozi	eseji
tepisi	osvrti	opanci	navijači
zlodusi	vitamini	nalozi	poeni
mitinzi	magacini	sokaci	predmeti
okruzi	vlasi	dvorci	slikari
centri	nervi	jezici	prozori
pucnji	domaćini	oblici	novinari
odsjeci	bioskopi	muškarci	dueti
potoci	mravi	rudnici	studenti
industrijalci	ateljei	izdajnici	romani
ministri	delegati	trgovci	profesori
neuspjesi	rodoljubi	*parlamentarci*	otmičari
orkestri	vitezovi	pojedinci	startovi
metalurzi	*kalemovi*	radnici	kvartovi
primjerci	*ćilimovi*	vatrogasci	*skverovi*
svjedoci	limunovi	prodavci	sokovi
valjci	lakovi	biolozi	štrajkovi
samostalci	plikovi	krivci	brodovi
odjeljci	jablanovi	ratnici	gradovi
odjeci	timovi	ustupci	horovi
usjeci	likovi	pomoćnici	vozovi
razmaci	*slivovi*	zvučnici	drumovi
treninzi	kerovi	talenti	sportovi
hirurzi	golubovi	konopci	plodovi
dušeci	stanovi	zaključci	ratovi
pašnjaci	planovi	blizanci	golovi
prvaci	sinovi	rođaci	radovi
koraci	štitovi	uzvici	frontovi
kilometri	satovi	*povici*	pragovi

Napomena: Italikom su označeni stimulusi koji su isključeni iz dalje analize, jer je na njima grešku napravilo više od 20% ispitanika.

Prilog 6. *Pokazatelji prilagođenosti testiranih modela*

	AIC	BIC
RT~ klaster + rezidual f*leme* + f*oblika* + (1|subjekt) + (1|stimulus)	-1730	-1686
RT~ klaster * rezidual f*leme* * f*oblika* + (1|subjekt) + (1|stimulus)	-1708	-1638
RT~ klaster + rezidual f*leme* * f*oblika* + (1|subjekt) + (1|stimulus)	-1721	-1670
RT~ klaster * f*oblika* + rezidual f*leme* + (1|subjekt) + (1|stimulus)	-1722	-1672
RT~ klaster * rezidual f*leme* + f*oblika* + (1|subjekt) + (1|stimulus)	-1733	-1682

Prilog 7. *Parametri mješovitog modela, koji fituje podatke dobijene za nominativ množine imenica muškog roda na svim stimulusima*

	Koeficijent	Standardna greška	t-vrijednost	Pr(>|t|)
intercept	6. 71357	0.02607	257.45	.000
TAČNI	-0.04839	0.01856	-2.61	.009
rezidual frekvencije leme	-0.05976	0.01379	-4.33	.000
frekvencija oblika	-0.04558	0.00641	-7.11	.000
klaster*rezidual frekvencije leme	0.04881	0.01872	2.61	0.009

Prilog 8. *Parametri mješovitog modela, koji fituje podatke dobijene za nominativ množine imenica muškog roda, bez stimulusa i subjekata na kojima su reziduali izvan opsega +/- 2.5 SD*

	Koeficijent	Standardna greška	t-vrijednost	Pr(>|t|)
intercept	6.6639	0.0251	265.66	.000
TAČNI	-0.0411	0.0164	-2.51	.009
rezidual frekvencije leme	-0.0666	0.0124	-5.35	.000
frekvencija oblika	-0.0364	0.0056	-6.47	.000
klaster*rezidual frekvencije leme	0.0607	0.0166	3.65	.001