Sveučilište u Zagrebu

FILOZOFSKI FAKULTET

Lucia Načinović Prskalo

AUTOMATSKO PREDVIĐANJE I MODELIRANJE HRVATSKIH PROZODIJSKIH OBILJEŽJA NA TEMELJU TEKSTA

DOKTORSKI RAD

Mentor:
dr. sc. Nives Mikelić Preradović, izv. prof.

Zagreb, 2016

University of Zagreb

FACULTY OF HUMANITIES AND SOCIALSCIENCES

Lucia Načinović Prskalo

AUTOMATIC PREDICTION AND MODELLING OF CROATIAN PROSODIC FEATURES BASED ON TEXT

DOCTORAL THESIS

Supervisor:
associate professor Nives Mikelić Preradović, PhD

Zagreb, 2016

O mentorici

Dr. sc. **Nives Mikelić Preradović** doktorirala je u travnju 2008. na Filozofskom fakultetu u Zagrebu disertacijom *Pristupi izradi strojnog tezaurusa za hrvatski jezik* te time stekla akademski stupanj doktora društvenih znanosti, znanstvenog polja informacijske znanosti.

Njen doktorski rad istražio je dva problema, od kojih prvi predstavlja početke razvoja sustava pretvorbe teksta u govor za hrvatski jezik kroz model računalnog naglasno-izgovornog leksikona glagola, imenica i pridjeva. Drugi problem kojim se bavio ovaj rad je izrada računalnog valencijskog leksikona hrvatskih glagola. U disertaciji je potpuno elaboriran model računalnog naglasnog morfološkog analizatora/generatora oblika koji je uklopila u prvi valencijski leksikon hrvatskih glagola - CROVALLEX, koji se sastoji od 1739 glagola s 5118 valencijskih okvira i s 173 sintaktičko-semantičke klase. Leksikon sadrži opsežne sintaktičko-semantičke opise koji su pridonijeli poboljšanju rezultata u drugim doktorskim disertacijama (Žitko, Branko. Model inteligentnog tutorskog sustava zasnovan na obradi kontroliranog jezika nad ontologijom doktorska disertacija. Zagreb: Fakultet elektrotehnike i računarstva, 03.03.2010. 2010. i Agić, Željko. Pristupi ovisnosnom parsanju hrvatskih tekstova / doktorska disertacija. Zagreb: Filozofski fakultet, 09.07. 2012.) CROVALLEX leksikon također predstavlja dobro polazište za izradu višejezičnih valencijskih rječnika kao izvora za strojno prevođenje, pomoć u računalno potpomognutom učenju jezika te pomoć leksikografima u oblikovanju različitih vrsta tekstova.

U znanstveno-nastavno zvanje docentice izabrana je 2009. godine, a u zvanje izvanredne profesorice izabrana je 2013. godine.

Od 2008. godine nositeljica je kolegija na preddiplomskom i diplomskom studiju informacijskih i komunikacijskih znanosti koje je samostalno predložila i uvela: „Uvod u obradu prirodnog jezika“, „Jezični inženjering“, „Diskurs i dijaloški sustavi“, „Odabrana poglavlja obrade prirodnog jezika“ i „Društveno korisno učenje u informacijskim znanostima“ te kolegija „Osnove digitalne obrade teksta i slike“ u suvoditeljstvu s prof. dr. Sanjom Seljan i Hrvojem Stančićem, izv. prof. Također, od 2008. godine suvoditeljica je kolegija „Multimedij i instrukcijski dizajn“ na poslijediplomskom studiju Odsjeka za

informacijske i komunikacijske znanosti, a od 2010. uvela je i drži kolegije „Automatsko sažimanje teksta“ i „Mediji i inteligentno pretraživanje teksta“.

Autorica je dviju knjiga, urednica dva međunarodna zbornika, objavila je 10 poglavlja u knjizi u koautorstvu te niz znanstvenih članaka u časopisima s međunarodnom recenzijom i uredništvom te zbornicima s međunarodnom recenzijom. Redovito sudjeluje na međunarodnim znanstvenim konferencijama objavljujući radove u zbornicima skupova, recenzirajući radove i obnašajući funkcije u međunarodnim konferencijskim odborima (*INFuture, KEOD,TSD*).

Sudjelovala je na nekoliko međunarodnih i nacionalnih projekata: Tipologija znanja i metode obrade obavijesti, Oblikovanje i upravljanje javnim znanjem u informacijskom prostoru, ACCURAT (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation), CESAR (CEntral and South-east europeAn Resources) i Abu-MaTran (Automatic Building of Machine Translation).

Od 2014. hrvatska je koordinatorica međunarodnog projekta Erasmus+ KA2 *Europe Engage – Developing a Culture of Civic Engagement through Service-Learning within Higher Education in Europe* (Uključena Europa - razvoj kulture građanskog angažmana kroz društveno korisno učenje u visokom obrazovanju u Europi).

Registrirana je u registru znanstvenih istraživača u Ministarstvu znanosti, obrazovanja i športa pod brojem 247800.

Zahvala

Prije svega, zahvaljujem se mentorici izv. prof. dr.sc. Nives Mikelić Preradović na podršci, razumijevanju, strpljenju, savjetima i smjernicama tijekom izrade doktorskog rada. Zahvaljujem se i mojim neizravnim mentorima izv. prof. dr. sc. Sandi Martinčić-Ipšić i prof. dr. sc. Ivu Ipšiću koji su me uveli u područje računalne obrade prirodnog jezika te vodili tijekom svih godina doktorskog studija.

Zahvaljujem se svim kolegama Odjela za informatiku Sveučilišta u Rijeci na ugodnoj radnoj atmosferi i podršci. Posebno se zahvaljujem mojoj kolegici i prijateljici Maji Brkić koja je sa mnom dijelila sve sretne i tužne trenutke, kako poslovne tako i životne te me podržavala i hrabrila. Hvala i kolegama Miranu Pobaru na pomoći te Slobodanu Beligi što je uskakao kad je trebalo. Zahvaljujem i kolegici Ivani Nežić s Odsjeka za kroatistiku, Filozofskog fakulteta u Rijeci na ustupljenim materijalima.

Nadasve zahvaljujem onima bez čije svekolike pomoći i odricanja ne bi bilo ni ovog rada - cijeloj mojoj obitelji. Od srca zahvaljujem mami Zdenki, tati Branku, nonetu Martinu, sestri Nataliji i njenoj obitelji na neizmjernoj podršci, ljubavi, stalnom poticanju i ogromnoj požrtvovnosti kako ranije kroz život, tako i za vrijeme nastajanja ovog rada. I naposljetku zahvaljujem mojim najvećim izvorima snage i motivacije, suprugu Franji i sinu Martinu na razumijevanju, strpljenju, odricanju i ljubavi.

Sažetak

Ljudski govor prenosi široki raspon informacija sadržanih u naglasnom sustavu, intonaciji, trajanju, ritmu, stankama, govornoj brzini, a ta se obilježja često nazivaju zajedničkim imenom - prozodija. Za hrvatski jezik dosad nisu provedena opsežna istraživanja na temu predviđanja prozodijskih obilježja i njihova modeliranja. U ovoj se disertaciji istražila primjenjivost metoda predviđanja prozodijskih obilježja i njihova modeliranja na hrvatski jezik te mogućnosti njihova poboljšanja uz uključivanje lingvističkih obilježja i jezičnih specifičnosti karakterističnih za hrvatski jezik kao što je primjerice leksički naglasak.

Hrvatski jezik pripada grupi ograničenih tonskih jezika u kojima tonska kontura realizirana na naglašenoj riječi nosi leksičku informaciju pa je zato preduvjet modeliranju prozodije hrvatskoga jezika postojanje rječnika koji obuhvaća naglaske kako osnovnih tako i izvedenih oblika riječi. U okviru ove disertacije se stoga izradio takav rječnik.

Obzirom da rječnikom ne mogu biti obuhvaćene sve riječi koje se pojavljuju u tekstu, razvio se i sustav za automatsko dodjeljivanje naglasaka riječima koje se ne nalaze u rječniku. Sustav se zasniva na modelu koji se učio na podacima iz izrađenog naglasnog rječnika.

U okviru doktorskog rada provedena je i analiza trajanja slogova hrvatskoga jezika te je izrađen model trajanja slogova.

Tilt intonacijski model primijenjen je za modeliranje F0 konture, a u tu svrhu označen je korpus od 500 rečenica označen Tilt oznakama.

Zbog brojnih uloga prozodije u ljudskoj komunikaciji, njezino predviđanje i modeliranje je važno i može se primijeniti u brojnim područjima obrade prirodnog jezika kao što su automatsko raspoznavanje govora, sinteza govora, automatska identifikacija govornika i jezika, određivanja granica pojedinih tema, određivanja emocionalnih stanja sudionika u komunikaciji, kod sustava za strojno potpomognuto prevođenje, sustava za računalno potpomognuto učenje jezika itd.

Ključne riječi: hrvatski prozodijski sustav, hrvatski naglasni rječnik, automatsko dodjeljivanje naglasaka, analiza trajanja, model trajanja, Tilt intonacijski model

Summary

Human speech conveys a wide range of information on the pitch accent, intonation, duration, rhythm, pauses, speech rate, and these characteristics are often collectively referred to as prosody. Because of the many roles of prosody in human communication, its predicting and modelling is important and can be applied in many areas of natural language processing such as automatic speech recognition, speech synthesis, automatic identification of speakers and languages, determining emotional states etc. Previous to this research no extensive research on the prediction of prosodic characteristics and their modelling had been conducted for the Croatian language. In this doctoral thesis the applicability of the methods for prosodic features predicting and their modelling was tested for Croatian. The possibility of improving their performance with the inclusion of linguistic features and linguistic specificities typical for the Croatian language (for example - lexical stress) was explored.

The Croatian language is a pitch accent language in which the tone contour realized in the prominent words carries lexical information. Therefore a prerequisite for modelling the prosody of Croatian is the existence of the lexicon in which lexical stress of both basic and derived forms of words is marked. Such a lexicon was created by implementing the rules for constructing derived forms of words based on the addition of the appropriate extension and on the place of stress moving if necessary. The entries in the lexicon are comprised of all derived words written without and with its corresponding stress and morph syntactic description (MSD) or part-of-speech tag (POS). Croatian belongs to the group of under-resourced languages and it is therefore considered that the importance of the lexicon will be significant and that it will be greatly applicable in various fields of natural language processing. The lexicon is comprised of 72,366 words in their basic form and over 1.000,00 derived word forms.

Besides the lexicon, the product of the implementation of the rules for constructing derived forms of words is a system for automatic stress assignment for Croatian. The accuracy of the system based on the rules is tested by comparing the results of its implementation to a text to the same text in which the stress to the words was assigned by an expert. The obtained results are very good with the accuracy of 78% if the MSD tags are assigned automatically to the words, and 87,7% if the MSD tags were corrected by hand.

There are words in Croatian that are written independently, but when it comes to their stress, they do not have one, but are prosodically leaning to the next or previous word. Such words are called clitics (proclitics and enclitics). There are cases in Croatian when the stress from the word that usually bears stress moves to the proclitic. Those rules are also implemented in the system and their implementation increased the accuracy of the system to 92,8%.

Sometimes words from the text cannot be found in the lexicon. For such cases, a system for automatic lexical stress assignment to the words was developed. The system consists of two models trained on the data from the above-described lexicon. One model was trained for the place of the stress prediction and the other for the category of the stress prediction (there are four possible stress categories in Croatian). The accuracy of the model for place of the stress prediction measured by tenfold cross-validation is 90,56%, and the accuracy of the model for category of the stress prediction is 86,02%. The accuracy of the models are also tested on the text which was used for the evaluation of the system based on the rules. The achieved accuracy for the place of the stress prediction is 97,4%, for the category of the stress 82,4%, and for both place and category of the stress the achieved accuracy is 80,1%.

The system based on the rules achieved batter accuracy compared to the system for automatic stress assignment based on the models. However, because there were words that were not assigned the stress after the implementation of the system based on the rules, the system for automatic stress assignment based on the models was used as a supplement to the system based on the rules in such cases. Such a hybrid approach achieved the accuracy of 95,3%.

In this doctoral thesis an analysis of syllable duration for Croatian was conducted and duration model developed. It was determined that the position of the syllable within word and sentence has impact to the duration of the syllable. In average, the duration of the syllable increased by 41,4% compared to the reference value if its position was at the beginning of the word and by 37,0% if its position was at the end of the word. If the position of the syllable was at the beginning of the sentence, its duration increased by 71,8% compared to the reference value, and by 104,75% if the syllable was in the end of the sentence. The analysis also showed that the contextual features have impact to the duration of the syllables. The duration of the syllable increased by different percentages according to the category of the consonants that followed after the observed syllable.

There were three categories of features taken into consideration in the duration model that was developed for Croatian - positional, contextual and those related to the stress. First, the accuracy of the duration model was tested after taking into consideration all three categories of the features. Then the accuracy of the model was tested after leaving out one of the category in order to determine how each category of the features contributes to the accuracy of the duration model. It was determined that all three categories impact the accuracy of the model in certain percentage and the greatest impact have features that belong to the positional category.

For intonation modelling of the Croatian language, Tilt intonation model was applied. For that purpose, a database of 500 sentences was labelled with corresponding tilt labels. The best RMSE value that was obtained by comparing the obtained F0 contour to the original is 22,2.

Key words: Croatian prosody, Croatian accent lexicon, automatic lexical stress assignment, duration analysis, duration model, Tilt intonation model

Sadržaj

1. Uvod 1
2. Pregled područja i srodnih istraživanja 6
3. Pregled modela trajanja i intonacije 13
3.1 Pristupi za modeliranje trajanja i F0 konture 14
3.1.1 Pristup temeljen na pravilima 14
3.1.2 Statistički pristup 14
3.2 Pregled pristupa za modeliranje trajanja 15
3.2.1 Klattov model trajanja 15
3.2.2 Model sume produkata (Sums-of-products model) 17
3.2.3 Stabla odlučivanja u modeliranju trajanja 18
3.2.4 Neuronske mreže za modeliranje trajanja 18
3.3 Pristupi za modeliranje intonacije (F0) 19
3.3.1 ToBI intonacijski model 19
3.3.2 INTSINT model 23
3.3.3 Fujisaki intonacijski model 26
3.3.4 Tilt intonacijski model 29
4. Prozodijska obilježja hrvatskoga jezika 30
4.1 Hrvatski naglasni sustav 30
4.1.1 Osobine naglašenog sloga 33
4.1.2 Raspodjela naglasaka 34
4.2 Govorna i jezična riječ 36
4.3 Rečenična intonacija 38
4.3.1 Intonacijska jedinica 38
4.3.2 Intonacijska jezgra 38
4.3.3 Intonacijski početak 40
4.3.4 Intonacijski završetak 41
5. Hrvatski naglasni rječnik 42
5.1 Postupak izrade rječnika 45
5.1.2 Imenice 45
5.1.3 Glagoli 52
5.1.4 Pridjevi 53
5.1.5 Ostale vrste riječi 54
5.2 Dodavanje MSD oznaka 55

5.3 Opis dobivenog rječnika ... 58
5.4 Rezultati primjene pravila na testnom tekstu ... 60
5.4.1 Pravila za prenošenje naglaska s naglasnice na prednaglasnicu ... 63
6. Automatsko dodjeljivanje naglasaka riječima iz teksta ... 64
6.1 Klasifikacijska i regresijska stabla ... 65
6.2 Metodologija ... 66
6.2.1 Jezične značajke ... 66
6.3 Rezultati automatskog dodjeljivanja naglasaka ... 71
6.4 Hibridni pristup automatskog dodjeljivanja naglaska pomoću pravila i modela za naglašavanje ... 74
7. Analiza trajanja i model trajanja slogova ... 76
7.1 Rastavljanje riječi na slogove ... 78
7.2 Korpus ... 80
7.2.2 Fonetski rječnik ... 81
7.2.3 Segmentacija riječi na manje jedinice ... 82
7.2.4 Korpus priča i bajki ... 83
7.3 Priprema podataka ... 86
7.4 Analiza trajanja slogova ... 89
7.4.1 Analiza položajnih čimbenika na trajanje slogova ... 90
7.4.2 Utjecaj kontekstualnih čimbenika na trajanje slogova ... 94
7.5 Model trajanja slogova ... 96
7.5.1 Jezične značajke za učenje modela trajanja ... 98
7.5.2 Rezultati automatskog predviđanja trajanja slogova ... 100
8. Tilt intonacijski model ... 103
8.1 Pregled Tilt modela ... 104
8.1.2 Automatska RFC analiza ... 107
8.2 Primjena Tilt modela na hrvatski jezik ... 109
8.2.1 Postupak nalaženja događaja za model Tilt ... 109
8.2.2 Rezultati Tilt modela ... 112
9. Rasprava ... 113
10. Zaključak ... 120
Popis literature ... 126
Popis slika ... 134
Popis tablica ... 136
PRILOZI ... 138
Prilog 1 Algoritam za izdvajanje genitiva iz teksta ... 139
Prilog 2 Algoritam za izdvajanje nominativa iz rječnika ... 141

Prilog 3 Algoritam za stupnjevito provjeravanje MSD oznaka prilikom dodjeljivanja naglaska pomoću pravila 143
Prilog 4 Algoritam za dodjeljivanje naglaska rječima rastavljenim na slogove 154
Prilog 5 MSD oznake za hrvatski jezik 156
Prilog 6 Naglasni rječnik hrvatskoga jezika 164
Životopis 165

1. Uvod

Prozodija je kompleksna kombinacija jezičnih obilježja pomoću kojih se izražavaju stav i pretpostavke i privlači pažnja u svakodnevnoj komunikaciji. Semantički sadržaj koji se prenosi putem glasovne ili tekstualne poruke još se naziva i denotacija, a emocionalni aspekti i namjera koju govornik želi prenijeti dio su poruke koji nazivamo konotacija. Prozodija ima važnu ulogu u prijenosu denotacije, a ključnu u prijenosu konotacije. Ukratko možemo reći da riječi od kojih je sastavljena rečenica opisuju leksički sadržaj, a prozodija opisuje više aspekata načina na koji su te riječi izgovorene te može i mijenjati sam sadržaj koji se riječima prenosi. U nekim slučajevima različiti leksički naglasak (koji se ubraja u jedno od prozodijskih sredstava) daje riječi različito značenje kao primjerice u *pàs* (kućni ljubimac) i *pâs* (dio tijela). Takvi se slučajevi mogu evidentirati u rječnicima koji se onda mogu koristiti prilikom modeliranja prozodije pa je na taj način prozodija uključena i u područje leksikografije. Prozodija također ima ulogu u različitim sintaksnim tumačenjima rečenice, a koristi se i za isticanje pojedinih riječi i značenja unutar rečenice. Prozodijom se prenose i emocionalna stanja sudionika u komunikaciji, a vrlo važnu ulogu ima i u interpretaciji dijaloga. Zbog brojnih uloga prozodije u ljudskoj komunikaciji, njezino pravilno tumačenje je važno za mnoge postupke u obradi jezika, pa je tako korisna primjerice kod automatskog raspoznavanja govora, sinteze govora, automatske identifikacije govornika i jezika, određivanja granica pojedinih tema, određivanja emocionalnih stanja sudionika u komunikaciji, kod sustava za strojno potpomognuto prevođenje, sustava za računalno

potpomognuto učenje jezika, itd. Nadalje, pravilno dodjeljivanje prozodijskih obilježja i poznavanje prozodijskih obrazaca naročito je važno za pojedina zanimanja, kao što su primjerice radijski i televizijski voditelji, jer se načinom upotrebe prozodije odražava ne samo kulturološki aspekt, već se može mijenjati i samo značenje koje se želi prenijeti. Osim toga, govor voditelja u društvenoj komunikaciji dobiva status modela ispravnog i poželjnog govorenja.

Cilj ovog rada bio je istražiti primjenjivost metoda predviđanja prozodijskih obilježja hrvatskoga jezika i njihovog modeliranja za hrvatski jezik te istražiti mogućnosti njihovog poboljšanja uzimajući u obzir lingvističke značajke i jezične specifičnosti hrvatskoga jezika. Glavne hipoteze od kojih se krenulo u izradu ovog doktorskog rada su:

a) H1: moguće je stvoriti naglasni morfološki analizator/generator koji riječima automatski dodjeljuje naglasak na temelju postojećih modela,

b) H2: prozodijski modeli (modeli trajanja i putanje F0) za hrvatski jezik mogu se izmodelirati na temelju jezičnih značajki teksta, a uključivanje različitih skupina jezičnih značajki u prozodijske modele utječe na točnost prozodijskih modela,

c) H3: jezične značajke specifične za hrvatski jezik poput mjesta i vrste leksičkoga naglaska dodatno utječu na točnost prozodijskih modela za hrvatski jezik.

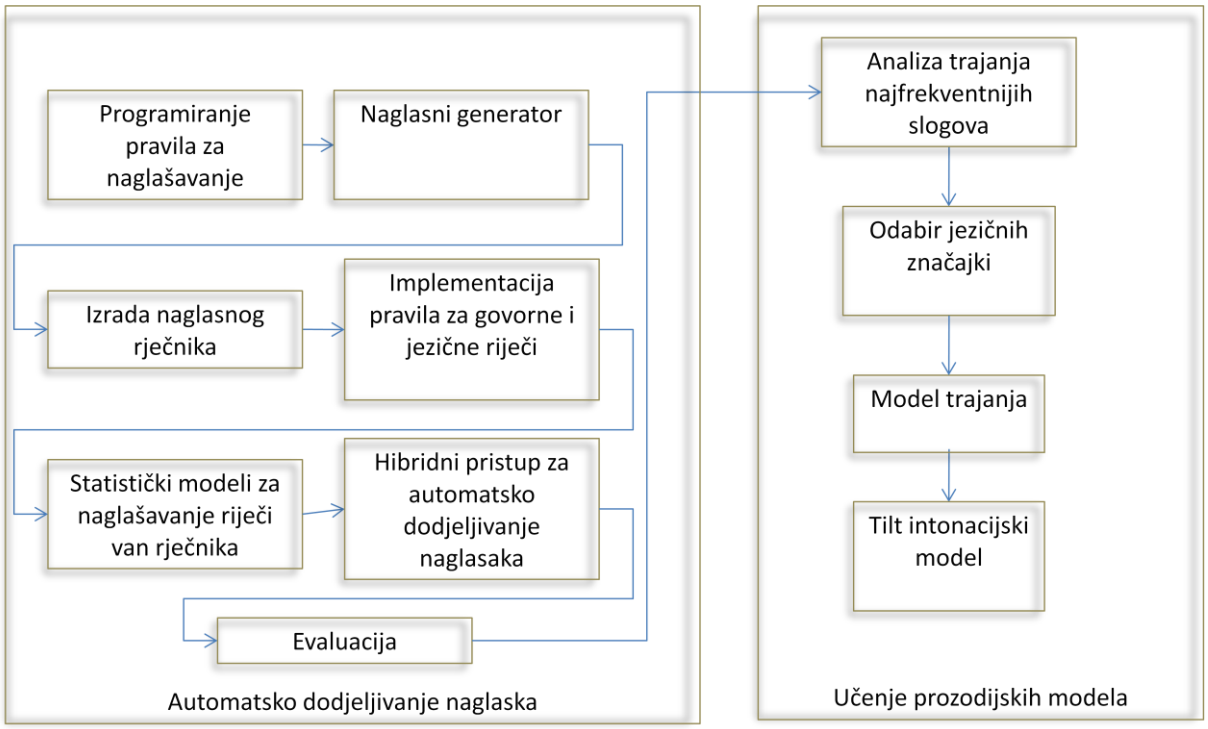
Tehnologije računalne obrade prirodnog jezika i govora koriste različite jezične resurse koji su najčešće dostupni za jezike s velikim brojem govornika, dok su za jezike s manjim brojem govornika, kao što je hrvatski, ti resursi ograničeni.

Dizajn ovog istraživanja i alat za automatsko dodjeljivanje leksičkih naglasaka tekstu na hrvatskom jeziku te prozodijski modeli za hrvatski jezik predstavlja doprinos računalnoj leksikografiji i računalnoj obradi prirodnoga jezika i govora. Pregled provedenog istraživanja u okviru ove doktorske disertacije prikazan je na slici 1.

Predloženi sustav sastoji se od dva osnovna dijela - sustava za automatsko dodjeljivanje naglasaka i sustava za učenje prozodijskih modela. Prvi korak kod nastajanja sustava za automatsko dodjeljivanje naglasaka bio je izgradnja hrvatskog naglasnog rječnika u kojem se nalaze osnovni i izvedeni oblici riječi s pridruženim naglascima. Kako bi se dobio navedeni rječnik, u programski jezik implementirana su pravila za automatsko generiranje

naglašenih oblika riječi, uzimajući njihov osnovni oblik iz leksičke baze i pridružujući im valjane paradigmatske nastavke i naglaske. Osim sustava za automatsko dodjeljivanje naglasaka pomoću pravila, razvio se i model za automatsko dodjeljivanje naglasaka riječima. Model je učen na ranije spomenutom hrvatskom naglasnom rječniku koji je također nastao u okviru ovoga rada. Naposljetku je predložen i treći pristup za dodjeljivanje naglasaka koji kombinira navedena dva pristupa. Pri izradi naglasnog rječnika te kod sustava za automatsko dodjeljivanje naglasaka, uzimaju se u obzir i razlike u hrvatskoj govornoj i jezičnoj riječi i pravila za premještanje naglasaka s naglasnica na prednaglasnice. Na kraju su sustavi za automatsko dodjeljivanje naglasaka evaluirani na tekstu, a njihova se točnost dobila uspoređujući dobiveni naglašeni tekst s tekstom kojemu je naglasak dodijeljen od strane eksperta.

U ovom je doktorskom radu analizirano i trajanje najfrekventnijih slogova ovisno o njihovom položaju i kontekstu kako bi se utvrdio optimalni skup značajki za modeliranje trajanja te je izrađen i model trajanja za hrvatski jezik. Za potrebe doktorskoga rada ručno je izrađen korpus rečenica hrvatskoga jezika označen tilt prozodijskim oznakama. Tilt intonacijski model primijenjen je na hrvatski jezik. Dobiveni modeli moći će se u budućim istraživanjima uključiti u sintezu hrvatskoga govora s ciljem dobivanja prirodnijeg i razumljivijeg sintetiziranog govora, kod automatskog raspoznavanja govora za postizanje boljih rezultata, u području automatske identifikacije govornika i jezika, kod sustava za strojno potpomognuto prevođenje, u sustavima za računalno potpomognuto učenje jezika, kod prepoznavanja emocionalnog stanja u komunikaciji i sličnim područjima u obradi prirodnog jezika.



Slika 1 Pregled provedenog istraživanja

Ovaj doktorski rad strukturiran je u 10 poglavlja. U sljedećem je poglavlju dan pregled područja i srodnih istraživanja te su ukratko opisani postojeći modeli trajanja i intonacije.

U trećem su poglavlju opisani neki od najpoznatijih pristupa za modeliranje trajanja kao što su Klattov model, Model sume produkata, stabla odlučivanja, neuronske mreže te pristupi za modeliranje intonacije - ToBI, INTSINT, Fujisaki i Tilt intonacijski modeli.

U četvrtom poglavlju opisana su prozodijska obilježja hrvatskoga jezika, hrvatski naglasni sustav, osobine naglašenog sloga, dana je raspodjela naglasaka u hrvatskom jeziku, te su opisane osnovne razlike između govorne i jezične riječi.

U petom poglavlju opisan je hrvatski naglasni rječnik koji je dobiven implementacijom pravila za dodjeljivanje naglasaka. Posebnu se pažnju posvećuje postupku dodjeljivanju naglaska imenicama, glagolima i pridjevima te problematici na koju se naišlo prilikom navedenih postupka. Osim toga, u ovom se poglavlju daje i detaljna statistika dobivenog rječnika, postupak dodavanja MSD oznaka u rječnik te se opisuju rezultati primjene pravila za naglašavanje na testnom tekstu.

U šestom poglavlju opisuje se postupak automatskog dodjeljivanja naglaska riječima iz teksta na temelju modela. Nabrajaju se jezične značajke koje se koriste prilikom učenja modela te se navode rezultati primjene modela. U istom je poglavlju predložen i pristup za automatsko dodjeljivanje naglasaka koji kombinira pravila za naglašavanje i naglašavanje pomoću modela.

U sedmom poglavlju opisana je analiza trajanja i model slogova za hrvatski jezik. Opisan je i postupak rastavljanja riječi na slogove i korpus koji se koristio. Prikazani su rezultati analize trajanja slogova s obzirom na položajne i tekstualne čimbenike. Naposljetku se u poglavlju opisuje model trajanja slogova, korištene jezične značajke za učenje modela te navode rezultati.

U osmom poglavlju daje se pregled Tilt intonacijskog modela te se opisuje primjena i rezultati Tilt modela na hrvatski jezik.

Na kraju rada slijedi rasprava vezana uz rezultate u devetom poglavlju i zaključne misli i prijedlozi za budući rad u desetom poglavlju.

2. Pregled područja i srodnih istraživanja

U prozodijska obilježja ili sredstva najčešće se ubrajaju naglasak, intonacija, trajanje emisije fonema (kvantiteta), ritam, stanke i govorna brzina. Slogovi, kao manje jedinice od kojih su sastavljene riječi, različitog su trajanja, jakosti, tona i izgovorne točnosti. U navedenim se osobinama (ili u nekima od njih) jedan slog u riječi ističe nad drugima i zove se naglasak riječi. Intonacija ili govorna melodija je figura koju tvore uzastopne mjere tona (Škarić, 1991) i često se opisuje pomoću F0 konture (fundamentalna frekvencija) koja predstavlja frekvenciju titranja glasnica i osnovni je čimbenik kod percepcije visine tona. Pod trajanjem emisije fonema podrazumijeva se ostvarena dužina fonema u izgovoru i obično se mjeri u milisekundama. Ritam je oblik koji čini ravnomjerni niz jednakih elemenata. Stanke u govoru predstavljaju odsječke govornog vremena bez teksta, a ovisno o ulozi mogu biti primjerice stanke razgraničenja, stanke isticanja, leksičke stanke itd. Govorna se brzina izražava brojem jedinica govora u jednoj jedinici vremena, najčešće brojem izgovorenih slogova u sekundi.

Prema (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991), govor je sastavljen od dva osnovna sloja: glasa i teksta. Podslojevi teksta su nizovi fonema, prozodija riječi i prozodija rečenice, a podslojevi glasa izražajnost i govorni krik. Fonemski sloj se u govoru

realizira izgovornim sredstvima, tj. izgovaranjem glasnika, a preostali slojevi dijele zajednička prozodijska sredstva. Oni u prozodijska sredstva ubrajaju ton i intonaciju, glasnoću i naglasak, stanku, ritam, govornu brzinu, boju glasa, spektralni sustav govornog zvuka, govornu modulaciju, način izgovora glasnika te mimiku i gestu.

Ton opisuju kao učestalost periodičnih titraja zvuka koji nastaje titranjem glasnica, a intonaciju ili govornu melodiju kao figuru koju tvore uzastopne mjere tona. Prozodijska intonacija se u fonemskom sloju nalazi samo na samoglasnicima, tj. sami samoglasnici utječu na ostvarenu intonaciju.

Kontrastne razlike u glasnoći temeljene na jakosti zvuka unutar riječi tvore naglasak riječi, a unutar rečenice rečenični naglasak ili isticanje. Slogovi, kao manje jedinice od kojih su sastavljene riječi, različitog su trajanja, jakosti, tona i izgovorne točnosti. U navedenim se osobinama (ili u nekima od njih) jedan slog u riječi ističe nad drugima i zove se naglasak riječi. Uloga je naglaska riječi isticanje riječi kao jedinice govornog niza pri čemu mjesto naglašenog sloga u riječi nije bitno. Kod nekih je jezika točno određeno mjesto naglašenog sloga u riječi pa je tako primjerice u francuskom to uvijek posljednji slog u riječi, u poljskom pretposljednji, a u češkom prvi slog. Kod drugih je jezika mjesto naglaska slobodno, kao primjerice u engleskom i hrvatskom.

Važno prozodijsko sredstvo čine i stanke u govoru (odsječci govornog vremena bez teksta). Trajanje stanki u govoru zavisi od uloga koje preuzimaju. Pa tako primjerice u čitanju vijesti one imaju ulogu jezične rečenične prozodije i zauzimaju oko 15% od ukupnog trajanja govora, u čitanju umjetničke proze kad stanke osim rečenične prozodije imaju i ulogu izražajnog sredstva oko 30%, a u spontanom govoru gdje postaju i sredstvo krika zauzimaju 40 do 50% govornog vremena (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991).

S obzirom na uloge, stanke mogu biti: stanke razgraničenja, stanke isticanja, leksičke stanke, stanke procesiranja i stanke prekida govora. Stanke razgraničenja označavaju sintaktičku i logičku organizaciju iskaza, a ponekad uklanjaju i dvosmislenost *(„Ako znaš dobro// možeš proći*.“ Ili *„Ako znaš// dobro možeš proći*.“). Njihovo razmjerno trajanje ovisi o veličini organizacijske jedinice govora koju razgraničuju (intonacijske jedinice unutar rečenice, rečenice unutar odlomka ili odlomke unutar diskursa), a apsolutno trajanje o brzini govora. Stanke razgraničenja intonacijskih jedinica očituju se u stankama trajanja jednog sloga, a često se izvode tako da se prethodni slog produlji za trajanje jednog sloga. Stanke razgraničenja rečenica obično su trajanja jedne riječi pri čemu jedan dio otpada na produljenje prethodnog sloga, a drugi na bezglasni dio stanke. Stanke razgraničenja odlomaka obično su

trajanja jedne rečenice. Stanke isticanja događaju se unutar rečenice, a obično ispred riječi koju želimo posebno istaknuti čime se pojačava učinak iščekivanjem. Leksičke stanke javljaju se u govoru kada zamjenjuju neke neizgovorene riječi *(„Šuti//smetaš.“*). Stanke procesiranja nastaju zbog usporenja komunikacijskoga govornog toka ne jednom dijelu lanca radi stvaranja efekta privlačenja pažnje i iščekivanja kod sugovornika. Stanke prekida govora uzrokovane su negovornim razlozima kao što su primjerice kašljanje, kihanje, gutanje, udisanje, ispijanje vode i sl.

Ritam je oblik koji čini ravnomjerni niz jednakih elemenata. Ritam govora razaznajemo kao izmjenjivanje naglašenih i nenaglašenih slogova u govoru (Toporišič, 1991).

Govorna se brzina izražava brojem jedinica govora (glasnika, slogova, riječi, rečenica) u jednoj jedinici vremena (minuti, sekunda), a najčešće brojem izgovorenih slogova u sekundi. Normalna govorna brzina razgovora je 4 do 7 slogova u sekundi, a maksimalna s razumljivom artikulacijom do 13-14 slogova u sekundi. Govorna brzina se povećava skraćivanjem i smanjivanjem broja stanki te skraćivanjem glasnika (više samoglasnika nego suglasnika), a smanjuje povećanjem broja i duljina stanka te produljivanjem glasnika.

Pod bojom glasa obično se podrazumijeva boja koju imaju samoglasnici kad se oduzme njihova razlikovna (fonemska) boja. Glavni čimbenici koji utječu na boju glasa su organske osobine čovjeka te način uporabe glasovnih organa koji je pod utjecajem kulturalnih čimbenika.

Dva su osnovna tipa govora po spektralnom sustavu: harmoničan i šuman. Harmoničan zvuk imaju pretežno samoglasnici, a šuman suglasnici. Prozodijsko sredstvo čini odnos trajanja samoglasnika naspram suglasnika. Kod poetskog i emotivnog govora prevladava harmoničan odsječak, dok kod logičnog govora prevladava šuman odsječak zbog veće razgovjetnosti. Nekada se je kao kriterij za utvrđivanje „melodioznosti“ jezika uzimao odnos broj samoglasnika naspram ukupnom broju fonema. Tako su se primjerice talijanski s 48% samoglasnika i hrvatski s 45% samoglasnika smatrali „melodioznim“, a njemački s 38% samoglasnika „tvrdim“ jezikom. Međutim na „melodioznost“ jezika ne utječe samo broj samoglasnika naspram ostalih fonema već i o broju dugih samoglasnika naspram broju kratkih samoglasnika te o broju suglasničkih skupina u kojima su suglasnici za dvadesetak posto kraći pa se stoga navedeni kriterij za estetsko obilježavanje jezika smatra nepouzdanim (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991).

Govor se u vremenskom slijedu mijenja po jakosti, spektralnim oblicima, spektralnom sastavu i periodičnim titrajima. Prozodijska se modulacija uglavnom svodi na zvučne prijelaze (staccato i legato) i periodične vokalne mijene (vibrato i tremolo). Staccato karakterizira govor kod kojeg su zvukovi odvojeni oštrim prijelazima, česte stanke oštrih rezova, snažni naglasci dok legato ima blage rubove stanke, duže nizove riječi, naglašeni ton je tek lagano naglašen. Vibrato se manifestira kao govor bogatog zvuka dok tremolo karakteriziraju treperave promjene jakosti glasa.

Svaki čovjek ima svoj način izgovaranja kao posljedica različitih fizičkih i psihičkih osobina te utvrđenih izgovornih navika. Način izgovora pokazuje i narav, temperament, umnost, kultiviranost te estetski ideal (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991).

Mimika (pokreti lica) i gesta (pokreti ruku, glave, ramena i ostalih dijelova tijela) utječu na oblikovanje glasovnog materijala, tj. na ton, jakost, stanku, brzinu, ritam, modulaciju, boju, način izgovaranja.

Iako postoje univerzalne zakonitosti vezane za prozodiju u većini jezika, uporaba prozodijskih značajki uvelike je svojstvena pojedinim jezicima. Prema načinu korištenja varijacija tona, jezici se mogu kategorizirati u intonacijske i neintonacijske jezike (Jelaska, 2004). Kod intonacijskih jezika tonske se varijacije koriste za razlikovanje prozodijskih jedinica većih od riječi. Takvi su jezici primjerice engleski i nizozemski. Kod neintonacijskih se jezika tonske varijacije osim u intonacijske svrhe koriste i u okviru riječi s mogućnošću proširenja tonskog uzorka van granica riječi kao što je to slučaj kod hrvatskog jezika kod govorne riječi koja može obuhvaćati dvije ili čak više leksičkih riječi, ali se one spajaju u jednu izgovornu cjelinu s jednim naglaskom. Dakle kod neintonacijskih jezika se ton koristi i za izražavanje leksičkih kontrasta. U tipologiji naglasnih sustava, jezici se mogu kategorizirati kao udarni ili dinamički (engl. stress-accent languages), tonski (engl. tone languages) i ograničeni tonski jezici (engl. pitch-accent languages) (Pletikos, 2008). Kod udarnih jezika naglasak nema tonske obavijesti nego je udar kao naglasno obilježje povezano samo sa slogom dok tonski jezici imaju razlikovnu tonsku visinu na slogovima kojom se razlikuje značenje te riječi (Jelaska, 2004). Naglasni sustav hrvatskoga jezika pripada trećoj skupini, a ograničeni tonski jezici u literaturi se još nazivaju i "tonsko-dinamički" (Škarić, 1991) ili "jezici s visinskim naglaskom" (Jelaska, 2004) kod kojih tonska kontura realizirana na naglašenoj riječi nosi leksičku informaciju. U prozodijskim su riječima kod ograničenih tonskih jezika prisutne fonološke visinske razlike koje su ili označene u leksikonu ili su

uvedene preobličnim leksičkim pravilima. Zbog toga je prilikom modeliranja prozodije hrvatskoga jezika neophodno postojanje leksikona koji obuhvaća označene visinske razlike kako osnovnih tako i izvedenih oblika riječi uvedenih preobličnim leksičkim pravilima. Uz hrvatski, ograničeni tonski jezici su još srpski, bosanski, slovenski, švedski, norveški i litavski (Pletikos, 2008).

U svrhu predviđanja prozodijskih obilježja iz teksta postoje metode temeljene na pravilima, metode temeljene na podacima te one koje kombiniraju ta dva pristupa. Sve one najčešće uzimaju u obzir dva prozodijska sredstva koja najviše utječu na čovjekovu percepciju razumljivosti i prirodnosti: F0 konturu (u kojoj se očituju percepcije visine tona uključujući i naglasak kod hrvatskog jezika) i trajanje (ostvarena dužina fonema).

Najpoznatiji modeli za trajanje su Klattov model, model suma produkata (engl. Sums-of-products model), modeli temeljeni na klasifikacijskim i regresijskim stablima i modeli temeljeni na neuronskim mrežama. Klattov model sastavni je dio MITalk formantne sinteze (Allen, Hunnicut, & Klatt, 1987). Model se sastoji od sekvencijalnih pravila koja uključuju obilježja fonetskog okruženja, naglaske, kraćenje i produljivanje glasova na određenim položajima i sl. Model suma produkata (van Santen, 1994) statistički je model koji se temelji na informacijama o fonetskom i fonološkom okruženju, a u obzir uzima i znanje stručnjaka vezano uz različite čimbenike koji utječu na trajanje segmenata. Za modeliranje trajanja mogu se koristiti i klasifikacijska i regresijska stabla kao primjerice u (Wagner & Katarzyna, 2010). Budući da se pomoću neuronskih mreža mogu naučiti odnosi među kontekstualnim značajkama, one se mogu koristiti u modeliranju trajanja (Shreekanth, Udayashankara, & Chandrika, 2015). U takvim se modelima za odabranu jezičnu jedinicu računa vektor koji se sastoji od informacija o broju fonema u slogu, položaju u tonskoj grupi, naglasku, vrsti riječi i sl.

Među najpoznatije modele za modeliranje F0 konture spadaju ToBI model, Tilt model, Fujisakijev model i model neuronske mreže. ToBI (Tones and Break Indices) (Silverman, i dr., 1992)je fonološki intonacijski model. Koristi lingvistički ili fonološki pristup određujući mali skup diskretnih oznaka kojima se označavaju intonacijska mjesta naglasaka i tonova. Njime se transkribiraju naglasci i način grupiranja rečenica u sintagme. Tilt (Taylor, 2000) je fonetski intonacijski model koji F0 konturu prikazuje kao slijed kontinuirano parametriziranih događaja. Takvi parametri se onda nazivaju Tilt parametri, a određuju se izravno iz F0 konture. Osnovne jedinice Tilt modela su intonacijski događaji - lingvistički značajni dijelovi F0 konture. Fujisaki (Fujisaki & Ohno, 1995) model opisuje F0 konturu kao superpoziciju

komponente za grupiranje riječi u fraze i modeliranje značajki tih fraza i komponente za naglašavanje. Model neuronske mreže korišten je za modeliranje F0 konture u više radova, a primjerice u (Rao, 2012) se koristi neuronska mreža za predviđanje F0 vrijednosti za indijske jezike.

Za hrvatski jezik istraživanja o predviđanju i uključivanju prozodije u sintezu govora iz teksta provedena su u (Lazić, 2006) koji je u svom doktorskom radu opisao okvir za modeliranje strojnih postupaka za izgovaranje teksta pisanoga hrvatskim jezikom i u (Pobar, 2014) koji je u sintezu hrvatskog govora uključio model trajanja govornih jedinica i statistički model putanje F0 pomoću skrivenih Markovljevih modela u uobičajenom okviru statističke parametarske sinteze govora. Pritom nisu uzeti u obzir leksički naglasak ni morfosintaktičke oznake riječi u kojima se govorna jedinica nalazi niti neke druge specifičnosti hrvatskoga govora poput primjerice razlike u govornoj i pisanoj riječi (proklitike i enklitike). Zbog skromnog istraživanja o predviđanju prozodije iz teksta za hrvatski jezik u nastavku slijedi pregled radova na temu predviđanja prozodije za sintezu govora iz teksta za jezike srodne hrvatskom.

Šef i Gams (Šef & Gams, 2003) razvili su sustav za sintezu govora za slovenski jezik koji uključuje modul generiranja prozodije. Modelirali su trajanje na dvije razine: intrinzičnoj i ekstrinzičnoj. Kod intrinzičnog modeliranja u obzir su uzeti čimbenici kao što su tip glasa, okruženje glasa, tip sloga, naglasak sloga i sl. Kod ekstrinzičnog modeliranja čimbenici su brzina izgovora, položaj riječi unutar fraze i broj slogova u riječi. Šef je također istraživao automatsko naglašavanje riječi za sintezu govora u slovenskom jeziku pomoću stabla odlučivanja (Šef, 2006). Marinčič (Marinčič, Tušar, Gams, & Šef, 2009) je analizirao automatsko dodjeljivanje naglasaka u slovenskom jeziku te je uspoređivao ljudske i strojne sposobnosti dodjeljivanja naglasaka. Uz skup pravila o naglašavanju u slovenskom jeziku, koristio je i metode otkrivanja znanja iz podataka i rezultati su bili bolji od onih gdje je koristio samo pravila o naglašavanju.

Češki znanstvenici Romportl i Kala (Romportl & Kala, 2007) opisali su statističko modeliranje F0, modeliranje intenziteta i trajanja za češki jezik. Modeliranje F0 temeljeno je na konkatenaciji automatski dobivenih intonacijskih uzoraka. Za modeliranje intenziteta izdvojili su pravila temeljena na fonetičkim pravilima,a za modeliranje trajanja fonema koristili su stabla odlučivanja. Tihelka (Tihelka, Kala, & Mtousek, 2010) opisuje sustav za sintezu govora češkog jezika koji uključuje prozodijska obilježja koristeći pristup izbora jedinica (*unit selection*).

Sečujski (Sečujski, 2002) je izradio naglasni rječnik srpskog jezika namijenjen sintezi govora na srpskom jeziku u koji su ručno uneseni izvedeni oblici riječi s pripadajućim naglascima.

3. Pregled modela trajanja i intonacije

Pojam trajanja se odnosi na trajanje svih govornih članaka: odlomka, rečenice, intonacijskih jedinica, govornih riječi, slogova i glasnika. Ipak, najviše se istraživanja dosad posvetilo trajanju fonetičkih segmenata, a manje trajanju riječi i slogova (van Santen, 1997)(van Santen, 1994). Razlog tomu je što se stanke (granice) među segmentima, koje su jedno od najvažnijih prozodijskih sredstava trajanja, mogu relativno lako odrediti. Istraživanja u vezi trajanja na razini slogova i glasnika najčešće su bila fokusirana na trajanje slogova u čitanom govoru (Kato, Tsuzaki, & Sagisaka, 1998);(Stergar & Erdem, 2010). Pokazalo se da trajanje samoglasnika ovisi o brojnim čimbenicima, a neki od njih su primjerice okruženje (glasovi prije i poslije samoglasnika), naglasak (kako naglasak u riječi, tako i rečenični naglasak) te položaj (položaj sloga u riječi i u izgovornoj jedinici).

Iako trajanje i F0 nisu potpuno samostalni i mnogi čimbenici koji utječu na F0 utječu i na trajanje, većina se sustava posebno bavi trajanjem, a posebno F0.

3.1 Pristupi za modeliranje trajanja i F0 konture

Dosad su se izdvojila dva glavna pristupa modeliranja trajanja i F0 konture:

1) Pristup temeljen na pravilima i
2) Statistički pristup.

3.1.1 Pristup temeljen na pravilima

Prvi računalni sustavi za sintezu govora temeljili su se na pravilima. Kod implementacije prozodije u sintetizirani govor, pomoću pisanih pravila pokušavaju se predvidjeti prozodijska obilježja pomoću niza "ako-onda" („if-then“) izraza kojima se provjerava prisutnost određenog obilježja kao što je primjerice vrsta riječi. Primjer pravila:"AKO riječ pripada nepromjenjivoj vrsti riječi, ONDA je naglasi". Obzirom da se pravila pišu ručno, tj. određuju ih eksperti sa znanjem jezika, fonetike i lingvistike, obično je predviđanje prozodijskih obilježja kod takvih sustava pouzdano. Međutim, vrijeme koje je potrebno da bi se definirao skup pravila je vrlo dugo i stoga skupo. Najpoznatiji pristup temeljen na pravilima za modeliranje trajanja je sustav kojeg je primijenio Klatt u MTalk sustavu (Allen, Hunnicut, & Klatt, 1987). Za modeliranje F0 konture najpoznatiji takav pristup je korišten u(Pierrehumbert, 1981) gdje je kontura opisana kao niz ciljnih vrijednosti.

3.1.2 Statistički pristup

Stohastičke metode za predviđanje prozodijskih događaja ili stvaranje modela prozodije koriste velike baze podataka govora za učenje prozodijskih modela. Prvi korak kod ovakvog pristupa je označavanje (engl. *labelling*) podataka na kojima će se model učiti. Oznake (engl. *labels*) koje opisuju prozodijske karakteristike pridružuju se ručno. Na temelju takvih prozodijskih informacija, uči se model, tj. procjenjuju se parametri koji predstavljaju vjerojatnost prozodijskih događaja u kontekstu različitih lingvističkih čimbenika. Takav se model onda koristi za predviđanje najvjerojatnijih prozodijskih oznaka na bilo kojem ulaznom tekstu.

Jedna od stohastičkih metoda koja se koristi kod ovakvog pristupa su stabla odlučivanja (CART – classification and regression trees) (Hirschberg, 1995);(Ross & Ostendorf, 1996);(Veilleux, 1994). Kod stabla odlučivanja, najprije se određuje algoritam s

popisom mogućih pitanja ili čimbenika koji se moraju ispitati, a sustav automatski bira koji čimbenik ima najveću sposobnost predviđanja. U model se pomoću stabla odlučivanja može uključiti i Markovljeva pretpostavka (buduće se stanje predviđa na temelju trenutnog stanja i stanja koje je prethodilo trenutnom), što omogućava uključivanje prethodnih klasifikacija u predviđanje prozodijskih događaja. Takav se pristup koristio kod (Ostendorf & Veileux, 1994) za predviđanje granica fraza (intonacijskih jedinica), a kod (Ross & Ostendorf, 1996) za predviđanje naglasaka.

Skriveni Markovljevi modeli još su jedna stohastička metoda koja se može koristiti za predviđanje prozodijskih događaja. Ti su modeli predstavljeni konačnim skupom stanja. Kod ovog su pristupa i proces učenja modela i sam model vjerojatnosni. U (Taylor & Black, 1998) se koriste SMM-i za predviđanje granica fraza, a model je učen na informacijama o vrsti riječi i prethodnoj predviđenoj granici. Ovakav pristup zahtijeva veliku količinu podataka za učenje modela.

3.2 Pregled pristupa za modeliranje trajanja

Kao što je već spomenuto ranije, jedan od dva najvažnija prozodijska obilježja je trajanje. Postoje različiti pristupi modeliranju trajanja, a u ovom su odlomku izdvojeni neki od najznačajnijih.

3.2.1 Klattov model trajanja

Ovaj je model razvijen sedamdesetih i osamdesetih godina 20. st. i sastavni je dio MITalk formantnog sintetizatora govora (Allen, Hunnicut, & Klatt, 1987). Sastavljen je od sekvencijalnih pravila koja uključuju čimbenike fonetskog okruženja, naglaske, skraćivanje i produživanje slogova na određenim položajima i sl. Model polazi od pretpostavke da svaki tip fonetskog segmenta ima inherentno trajanje, svako pravilo povećava ili smanjuje trajanje tog segmenta za određeni postotak, a svaki od segmenata ne može se skratiti na duljinu manju od minimalne. Trajanje *t* svakog glasa računa se prema formuli:

$$t=(t_i-t_m)*P/100+t_m \ .$$

Pri tome je t_i inherentno, a t_m minimalno trajanje glasa. Parametar *P* mijenja se pomoću pravila koja se temelje na učincima koje različite značajke (primjerice fonetsko

okruženje, prisutstvo naglaska i sl.) imaju na trajanje. Pravila za računanje parametra *P* su sljedeća (Taylor, Text-to-Speech Synthesis, 2009):

- Produljenje na kraju rečenice/iskaza: ukoliko je segment samoglasnik u slogu na kraju rečenice onda je $P=1.4P$.
- Skraćivanje nezavršnih fraza: ukoliko se segment ne nalazi u slogu na kraju fraze/iskaza $P=0.6P$.
- Skraćivanje nezavršnih riječi: ako se segment ne nalazi u slogu na kraju riječi, $P=0.85P$.
- Višesložno skraćivanje: ukoliko se samoglasnik nalazi u višesložnoj riječi, $P=0.80P$
- Skraćivanje nepočetnog suglasnika: ukoliko se suglasnik ne nalazi na početnom položaju u riječi, $P=0.85P$.
- Skraćivanje nenaglašenih segmenata: minimalno trajanje nenaglašenih segmenata = minimalno trajanje/2. Za navedeni tip segmenta vrijednost *P* računa se:
  - samoglasnik u srednjem slogu u riječi: $P=0.5P$,
  - ostali samoglasnici: $P=0.7P$,
  - pre-vokalni vibrant, lateral ili spirant: $P=0.1P$,
  - svi ostali: $P=0.7P$.
- Istaknutost: ukoliko se segment nalazi u istaknutom slogu $P=1.4P$.
- Post-vokalni kontekst samoglasnika: samoglasnik se modificira ovisno o suglasniku koji ga slijedi prema sljedećim pravilima:
  - nema suglasnika koji slijedi, položaj na kraju riječi: $P=1.2P$,
  - ispred zvučnog frikativa: $P=1.6P$,
  - ispred zvučnog okluziva: $P=1.2P$,
  - ispred nazala: $P=0.85P$,
  - ispred bezvučnog okluziva: $P=0.7P$,
  - ispred svih ostalih: $P=P$.
- Skraćivanje u klasterima
  - samoglasnik nakon kojeg slijedi samoglasnik: $P=1.2P$,
  - samoglasnik ispred kojeg je samoglasnik: $P=0.7P$,
  - suglasnik okružen suglasnicima: $P=0.5P$,
  - suglasnik nakon kojeg slijedi suglasnik: $P=0.7P$,

- suglasnik kojemu prethodi suglasnik: P=0.7P.

Osim za engleski jezik, Klattov je model primijenjen i na neke druge jezike, primjerice na francuski, opisan u (Bartkova & Sorin, 1987) ili portugalski, opisan u (Simoes, 1990).

3.2.2 Model sume produkata (Sums-of-products model)

Model sume produkata statistički je linearni model, a temelji se na skupu jednadžbi koje se određuju na temelju informacija o fonetskim i fonološkim značajkama jedinica govora i njihovom okruženju. Interakcije kontekstualnih utjecaja se opisuju jednadžbama sastavljenih od suma i produkata. U modelu se polazi od pretpostavke da su operacije zbrajanja i množenja dovoljne za procjenu parametra, čak i ako je frekvencija razdioba vektora značajki u bazi podataka neravnomjerna. Primjer jedne takve jednadžbe za trajanje glasa *e*:

$$trajanje\ (Glas{:}/e/,\ Slijedi{:}Zvučni,\ Lokacija{:}Finalna) =$$

$$\alpha(/e/) + \delta(Finalna) + \beta(Zvučni) \times \gamma(Finalna)\ .$$

U danoj jednadžbi trajanje glasa *e* nakon kojeg slijedi zvučni glas, a nalazi se na finalnom mjestu u rečenici, računa se tako da se intrinzičnom trajanju glasa $e[\alpha(/e/)]$ doda određeni broj trajanja u milisekundama zbog toga jer se nalazi na krajnjoj poziciji u rečenici $[\delta(Final)]$ i naposljetku se dodaje efekt zvučnosti $[\beta(Zvučni)]$promijenjen parametrom $[\gamma(Final)]$ zbog pozicije na kraju rečenice. Model sume produkata prvi je predložio van Santen (van Santen, 1994) i primijenio ga na engleski jezik, a kasnije je primijenjen i na neke druge jezike, primjerice na mandarinski u (Shih & Ao, 1997) te na njemački u (Mobius & van Santen, 1996).

3.2.3 Stabla odlučivanja u modeliranju trajanja

Kao što je već ranije rečeno, trajanje jedinica govora se može predvidjeti pomoću stabla odlučivanja. Neki od čimbenika koji se mogu uzeti u obzir prilikom takvog modeliranja su identitet fonema, naglasak, identitet fonema s lijeve strane, identitet fonema s desne strane i sl. Stabla odlučivanja predstavljaju modele koji se mogu učiti na skupu podataka te koristiti za klasificiranje novih instanci pomoću binarnih pitanja o atributima koje nove instance posjeduju. Počinje se s čvorom korijena te se od njega nastavlja u svakom čvoru niz grane stabla postavljati pitanja o atributima instanci dok se ne dođe do lista stabla.Za svaki čvor, algoritam učenja stabla odlučivanja odabire atribut koji dijeli podatke za učenje na način da se dobije najbolja vrijednost predviđanja za ispravnu klasifikaciju. Stabla odlučivanja primijenjena su primjerice za modeliranje trajanja glasova litvanskog jezika u (Norkevičius & Raškinis, 2008) i za modeliranje trajanja indijskog jezika u (Krishna, Talukdar, Bali, & Ramakrishnan, 2004).

3.2.4 Neuronske mreže za modeliranje trajanja

Neuronske su mreže u(Campbell, 1992) korištene za modeliranje trajanja slogova. Pomoću neuronskih mreža mogu se dobro opisati osnovne interakcije između kontekstualnih utjecaja, tj. pomoću njih se mogu predstaviti pravilni uzorci ponašanja koji su implicitno sadržani u podacima. U modelu se najprije predviđaju trajanja slogova, a zatim se ta trajanja nadopunjuju trajanjem fonema. Za svaki se slog računa vektor koji se sastoji od informacija o dužini sloga, broju fonema u slogu, položaju u tonskoj grupi, naglasku, vrsti riječi i sl. Neuronske su mreže, osim za engleski jezik primijenjene primjerice i za telugu jezik u (Ramesh Bonda & Girija, 2015)te za španjolski u (de Cordoba, Montero, Gutierrez-Arriola, & Pardo, 2001). Osim za sintezu govora, model trajanja pomoću neuronskih mreža primijenjen je i u automatskom raspoznavanju govora, primjerice za estonski u (Alumäe, 2014).

3.3 Pristupi za modeliranje intonacije (F0)

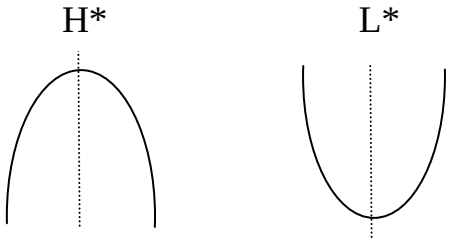
Prozodijski se modeli po načinu generiranja mogu podijeliti u dvije osnovne kategorije: linearni koji generiraju vrijednosti s lijeva na desno kao niz vrijednosti odnosno pomaka (ToBI, Tilt) i hijerarhijski modeli koji generiraju F0 konturu modelirajući parametre zasebno (na razini glasa, sloga, riječi, fraze, rečenice) i zatim kombinirajući modele tih parametara u zajednički model (Fujisaki, 2005).

Prema načinu opisivanja F0 konture, razlikujemo dva osnovna tipa prozodijskih modela: fonološke i fonetske modela. Fonološki modeli koriste skup apstraktnih fonoloških kategorija (ton, stanka i sl.), a svaka kategorija ima svoju lingvističku funkciju (ToBI). Fonetski modeli opisuju F0 konturu koristeći skup kontinuiranih parametara (Tilt, Fujisaki).

U nastavku slijedi pregled odabranih intonacijskih modela.

3.3.1 ToBI intonacijski model

ToBI (**To**nes and **B**reak **I**ndices) (Silverman, i dr., 1992) je fonološki intonacijski model. Koristi lingvistički ili fonološki pristup određujući mali skup diskretnih oznaka (*labela*) kojima se označavaju intonacijska mjesta naglasaka i tonova. Njime se transkribiraju naglasci i način grupiranja rečenica u fraze. ToBI razlikuje dva tipa naglaska: H* (koji predstavlja visoki tip naglaska) ili L* (koji predstavlja niski tip naglaska). Dva tip naglaska prikazani su na slici 2. Kod ToBImodela postoje i četiri granična tona L-L%, L-H%, H-H%, H-L%. Jedan naglasak pridružuje se svakoj naglašenoj riječi, a po jedan granični ton se pridružuje kraju svake intonacijske jedinice.



Slika 2 Dva tipa naglaska u ToBi modelu

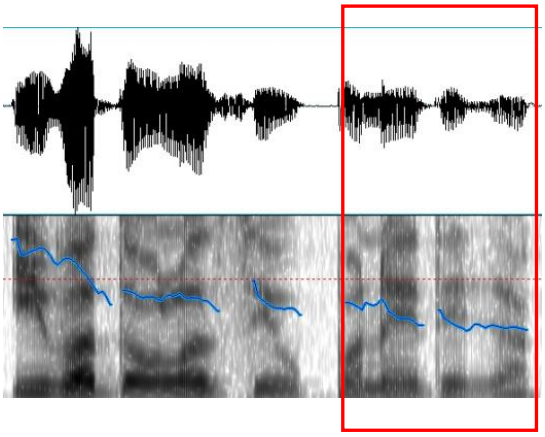
Granični ton

Granični ton predstavlja dio konture F0 između naglašene riječi i desne granice intonacijske jedinice. U primjerima koji slijede, koristimo istu rečenicu na hrvatskom jeziku "Iva dolazi u ponedjeljak." izgovorenu na četiri različita načina – s četiri različita granična tona:

1) L-L% granični ton (slika 3):

- F0 kontura završava niskom vrijednošću na kraju intonacijske jedinice,

- uobičajen je u neutralnim izjavnim rečenicama.

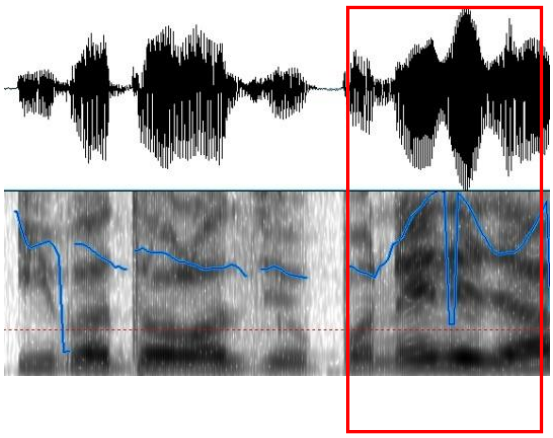


Slika 3 Rečenica s L-L% graničnim tonom

2) H-H% granični ton (slika 4):

- F0 kontura završava visokom vrijednosti frekvencije pri kraju intonacijske jedinice,

- čest u pitanjima s „da/ne“ odgovorima.

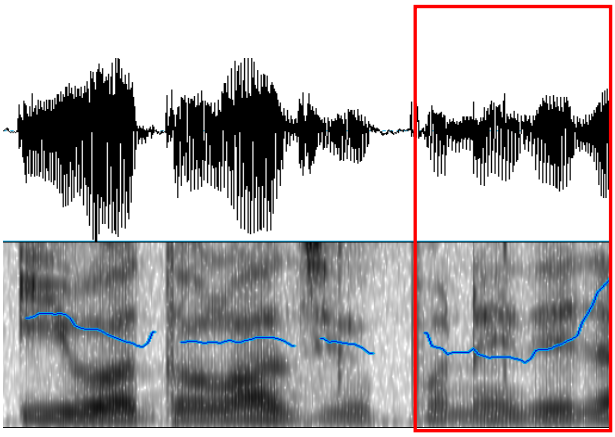


Slika 4 Rečenica s H-H% graničnim tonom

3) L-H% granični ton (slika 5):

- vrijednosti F0 konture su na početku intonacijske jedinice niske, a zatim rastu pri kraju intonacijske jedinice,

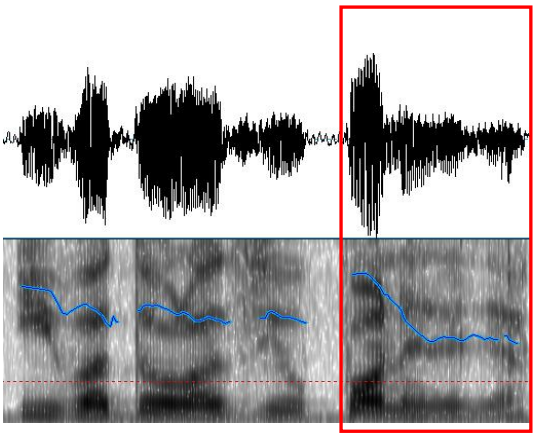
- čest je u rečenicama s nedovršenom misli, sumnjom ili kontradikcijom.



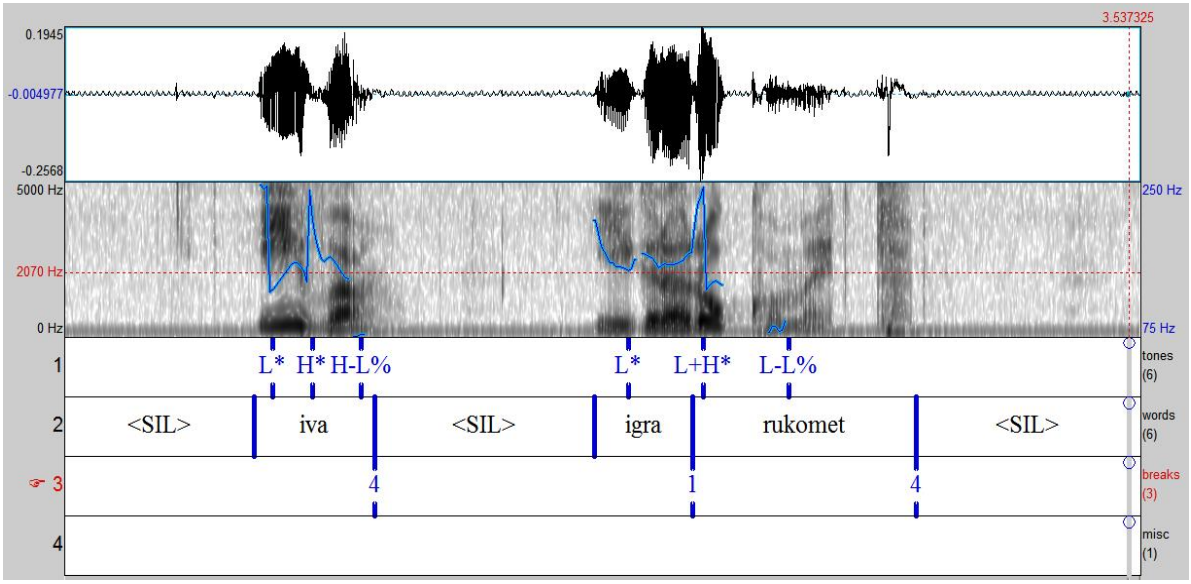
Slika 5 Rečenica s L-H% graničnim tonom

4) H-L% granični ton (slika 6):

- vrijednosti F0 konture su visoke, a pri kraju intonacijske jedinice postepeno padaju.



Slika 6 Rečenica s H-L% graničnim tonom



Slika 7 Primjer rečenice s transkripcijom po modelu ToBI

Transkripcija rečenice po ToBI modelu sastoji se od snimanja govora, prikaza F0 konture i označavanja prozodijskih oznaka pomoću simboličkih oznaka. Te se simboličke oznake obično opisuju u četiri vremenski usklađene trake, tako da se lako mogu uskladiti s odgovarajućom F0 konturom i valnim oblikom govora. Na slici 7 prikazana je transkripcija rečenice pomoću alata Praat (Boersma & Weenink, 2016).

U prvom se prozoru prikazuje valni oblik snimljene rečenice. Horizontalna os predstavlja vrijeme, a vertikalna amplitudu vibracija.

U drugom je prozoru prikazana F0 kontura te iste rečenice. Kao što je već rečeno, F0 kontura predstavlja stopu vibracije glasnica. Prosječna brzina titraja glasnica u govoru muškaraca je 120 Hz, a žena 220 Hz(Škarić, 1991). Prema F0 na slici, može se dakle vidjeti da je rečenicu izgovorila ženska osoba. U drugom se prozoru može vidjeti spektrogram - prikazan u obliku sivo-bijelog uzorka. Slično kao i valni oblik, spektrogram također daje informacije o energiji govornog signala. Kod spektrograma, horizontalna os također predstavlja vrijeme, a vertikalna frekvenciju.

Ispod spektrograma se nalaze četiri trake za upisivanje oznaka po ToBI modelu koje su ranije spomenute. U prvu se traku upisuju oznake za različite prozodijske događaje kao slijed visokih (H) ili niskih (L) tonova označenih dijakritičkim znakovima kao oznaka njihovih intonacijskih funkcija. Tu se upisuju naglasci i granični tonovi. U drugu se traku upisuje transkripcija riječi u rečenici. Zapis riječi poravnat je s mjestom njihova pojavljivanja u prozoru s valnim oblikom i spektrogramom. U treću traku upisuju se oznake za pauze prema

kojima se riječi grupiraju u izgovorne jedinice. Jačina (duljina) pauze označava se brojkama od 0 do 4, pri čemu se s 0 označava najmanja pauza, a s 4 najdulja pauza (Wightman, 2002).

Posljednja traka koristi se za upisivanje komentara ili za označavanje događaja kao što su udasi, kašljanje, smijeh i sl., a zbog jednostavnijeg razlikovanja mogu se pisati unutar znakova „<>“ (npr. <udah>).

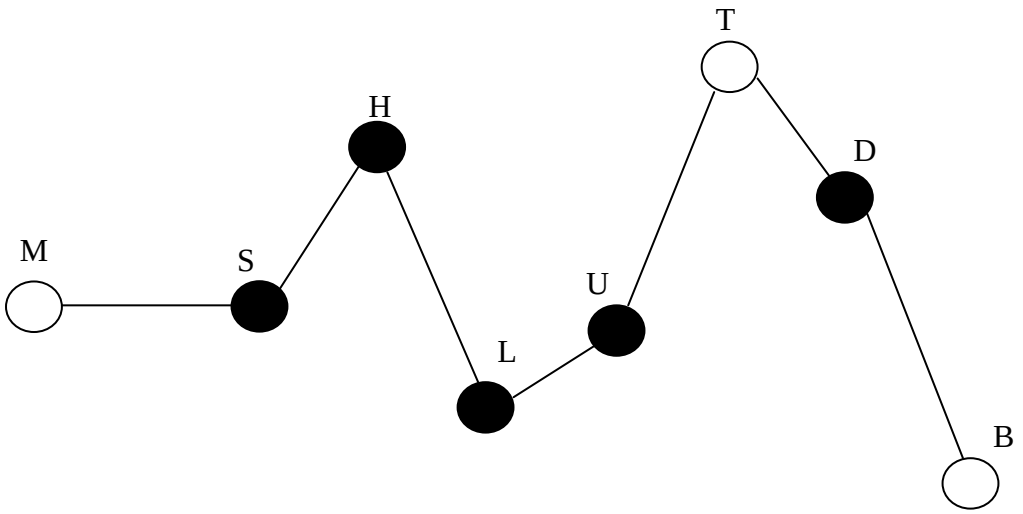
3.3.2 INTSINT model

INTSINT (Hirst, 2001) model opisuje intonaciju skupom tonskih simbola, a budući da ti simboli nisu svojstveni samo jednom jeziku, mogu se primijeniti na različite jezike. Model je izrađen s namjerom da se omogući transkripcijski sustav za intonaciju koji bi se mogao primijeniti na više jezika. Ulaz u INTSINT sustav je niz ciljnih vrijednosti koje se predstavljaju sljedećim skupom simbola:

- **T** – Top (hrv. vrh),
- **M** – Mid (hrv. sredina),
- **B** – Bottom (hrv. dno),
- **H** – Higher (hrv. viši),
- **S** – Same (hrv. isti),
- **L** – Lower (hrv. niži),
- **U** – Up-stepped (hrv. povišen),
- **D** – Down-stepped (hrv. snižen).

Pravila za dodjeljivanje oznaka u INTSINT modelu su:

1) Najviša i najniža ciljna vrijednost u rečenici/iskazu obilježavaju se s oznakom T, odnosno B.
2) Prvoj ciljnoj vrijednosti dodjeljuje se oznaka M.
3) Ostale se ciljne vrijednosti određuju u zavisnosti od prethodne - ciljnoj vrijednosti koja je od prethodne manja od određenog praga dodjeljuje se oznaka S. U ostalim se slučajevima dodjeljuje oznaka H, L, U ili D ovisno o razlici ciljne vrijednosti u odnosu na prethodnu.
4) Za svaku se kategoriju ciljnih vrijednosti potom izračunavaju statističke vrijednosti. Za apsolutne kategorije uzimaju se srednje vrijednosti, a relativne se računaju postupkom linearne regresije uzimajući u obzir prethodnu vrijednost.
5) Svakoj se ciljnoj vrijednosti kojoj je dodijeljena oznaka H ili L može promijeniti oznaka u T, U, B ili D ukoliko se na taj način poboljša statistički model.
6) Koraci 4 i 5 se ponavljaju sve dok se ne izvrši posljednja promjena oznaka.



Slika 8 INTSINT sustav označavanja[1]

[1]Preuzeto iz: (Louw & Bernard, 2004)

Slika 8 prikazuje skup apstraktnih simbola primijenjen kao skup oznaka u INTSINT modelu.

Simboli T, M i B uzimaju se kao tonovi apsolutne vrijednosti koji opisuju govornikov cjelokupni raspon glasa, a tonovi označeni oznakama H, S, L, U i D pridružuju se relativne vrijednosti u odnosu na prethodnu ciljnu vrijednost. Tonovi relativnih vrijednosti U i D se mogu pojaviti više puta dok se H, S, i L pojavljuju jednom. Prema tome, kod INTSINT modela postoje dva parametra koji su zavisni od govornika:

- ključ - definiran je srednjom vrijednošćuF0(Hz)govornika,
- raspon - određen je intervalom između najviše i najniže vrijednosti F0 u rečenici/iskazu.

Ostalim se oznakama vrijednosti pridružuju prema sljedećim formulama (Hirst, 2005):

- **T**: P(i) := ključ + raspon/2,
- **M**: P(i) := ključ,
- **B**: P(i) := ključ - raspon/2.

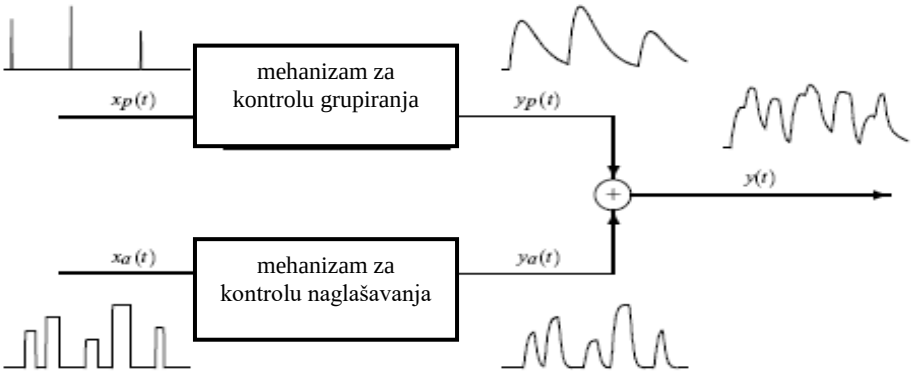
Preostale se vrijednosti računaju u zavisnosti o prethodnim vrijednostima:

- **H**: P(i) := (P(i-1) + T) / 2,
- **U**: P(i) := (3*P(i-1) + T) / 4,
- **S**: P(i) := P(i-1),
- **D**: P(i) := (3*P(i-1) + B) / 4,
- **L**: P(i) := (P(i-1) + B) / 2.

Pri čemu P(i) označava ciljnu vrijednost F0 koje se obično izračunavaju na logaritamskoj skali. Na taj se način iz INTSINT vrijednosti može rekonstruirati F0 kontura.

3.3.3 Fujisaki intonacijski model

Fujisakijev model (Fujisaki & Ohno, 2005) temelji se na pristupu prema kojem se F0 kontura može prikazati pomoću dva osnovna elementa - komponente za grupiranje u fraze koja se sporije mijenja i komponente za naglašavanje koja se mijenja brže. Model čini skup naredbi u obliku impulsa (mehanizam za kontrolu grupiranja) i niz funkcija (mehanizam za kontrolu naglašavanja) koje čine ulaz u linearne filtre te zajedno čine F0 krivulju. Mehanizam za kontrolu grupiranja prikazan je u gornjem dijelu na slici 9, a mehanizam za kontrolu naglašavanja u donjem dijelu. F0 kontura prikazana je kao superpozicija komponente za grupiranje riječi u fraze i modeliranje značajki tih fraza i komponente za naglašavanje.



Slika 9 Princip rada Fujisakijevog modela

Cjelokupna F0 kontura izražena je formulama:

$$\ln \mathrm{F0(t)} = \ln \mathrm{Fmin} + \sum_{\mathrm{i=1}}^{\mathrm{I}} \mathrm{A_{pi}\, G_{pi}\,(t - T_{0i})} + \sum_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{A_{aj}\,(G_{aj}\big(t - T_{1j}\big) - \;\; G_{aj}\big(t - T_{2j}\big)\big)}$$

gdje je $$G_{pi}(t) = \begin{cases} \propto_i^2 \, te^{-\propto_i t} & za\ t \geq 0 \\ 0 & za\ t < 0 \end{cases},$$

$$G_{pi}(t) = \begin{cases} min\big[1 - \big(1 + \beta_j\, t\big)e^{-\beta_j t}, \theta\,\big] & za\ t \geq 0 \\ 0 & za\ t < 0 \end{cases},$$

a izrazi imaju sljedeća značenja:

F_{min} - minimalna vrijednost govornikove F0,

I - broj komponenti mehanizma za grupiranje u fraze,

J - broj komponenti mehanizma za naglašavanje,

A_{pi} - magnituda i-te naredbe mehanizma za grupiranje u fraze,

A_{aj} - magnituda j-te naredbe mehanizma za naglašavanje,

T_{oi} - vrijeme i-te naredbe mehanizma za grupiranje u fraze,

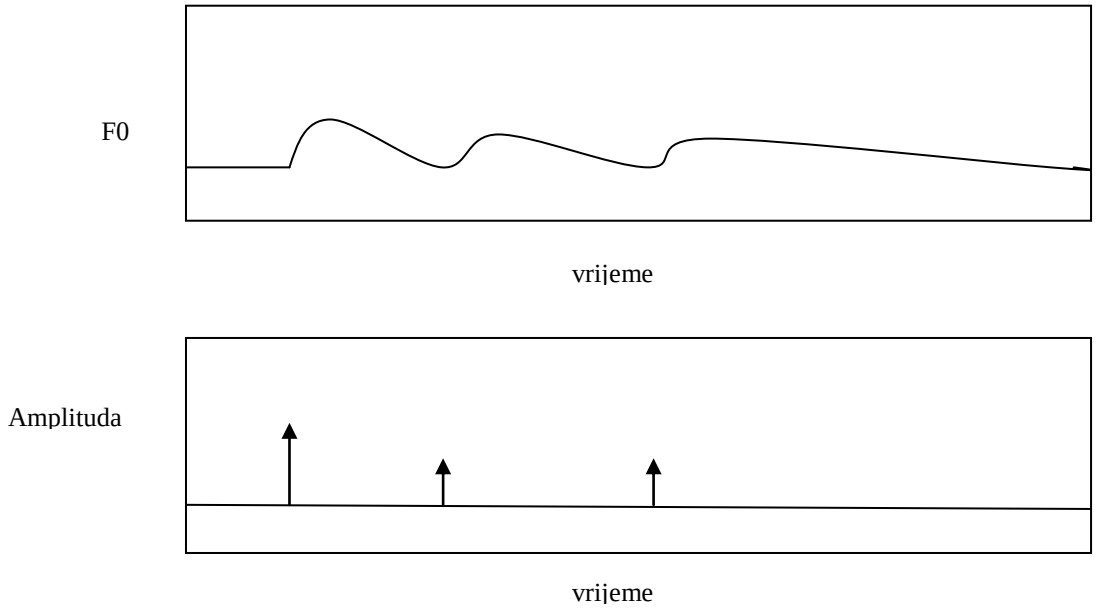
T_{1j} - početak j-te naredbe mehanizma za naglašavanje,

T_{2j} - kraj j-te naredbe mehanizma za naglašavanje,

α_i - prirodna kutna frekvencija i-te naredbe u mehanizmu za grupiranje u fraze,

β_j - prirodna kutna frekvencija j-te naredbe u mehanizmu za naglašavanje,

Θ - parametar kojim je određena maksimalna vrijednost komponente za naglašavanje.

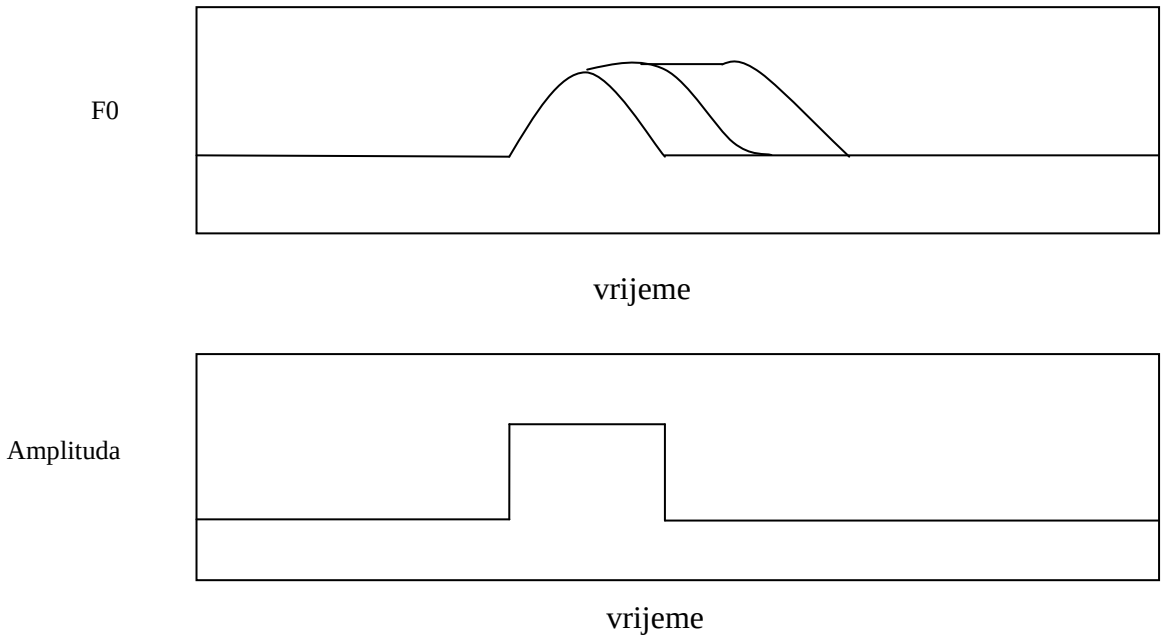


Slika 10 Primjer komponenti mehanizma za grupiranje[2]

Na slici 10 u gornjem je dijelu prikazana F0 kontura, a u donjem ulazni impulsi mehanizma za grupiranje u fraze

Na slici 11 prikazana su tri naglaska različitih trajanja u gornjem okviru, a u donjem je prikazana funkcija ulaza za drugi naglasak.

[2]Preuzeto iz: (Taylor, Text-to-Speech Synthesis, 2009)



Slika 11 Primjer komponenti mehanizma za grupiranje[3]

Mehanizam za grupiranje u fraze se pokreće impulsom koji nakon što se propusti kroz filter, utječe na to da se F0 kontura povisi do lokalne maksimalne vrijednosti i zatim postepeno opada. Sljedeća grupa dodaje se na kraj prethodne i na taj način se stvara uzorak prikazan na slici 10. Vremenska konstanta α regulira kojom brzinom grupa doseže maksimalnu vrijednost.

Mehanizam za naglašavanje pokreće se ulaznom funkcijom. Kada se funkcija propusti kroz filter, stvaraju se odgovori u obliku koji su prikazani na slici 11. Konstanta β obično je znatno viša od konstante α što skraćuje vrijeme odgovora filtra. Na taj način oblik dobiven mehanizmom za naglašavanje, doseže maksimalnu vrijednost mnogo brže nego onaj dobiven kao rezultat mehanizma za grupiranje u grupe.

Fujisakijev model prvotno je dizajniran za generiranje F0 konture za japanski (Fujisaki, Hirose, Halle, & Lei, 1971), a zatim je primijenjen i na neke druge jezike. Primjerice u (Fujisaki, Ohno, & Takashi, 1997) za grčki jezik, u (Fujisaki & Ohno, 1995) za engleski te u (Mixdorff & Fujisaki, 1994) za njemački jezik.

[3]Preuzeto iz: (Taylor, Text-to-Speech Synthesis, 2009)

3.3.4 Tilt intonacijski model

Tilt (Taylor, 2000)je fonetski intonacijski model koji F0 konturu prikazuje kao slijed kontinuirano parametriziranih događaja.

Osnovna jedinica Tilt modela je intonacijski događaj, a osnovni tipovi događaja su naglasak i granični tonovi. Budući da Tilt model opisuje F0 konturu, možemo reći da su naglasci kod Tilt modela dijelovi F0 konture povezani s naglašenim slogovima, a granični tonovi su rastući događaji koji se obično pojavljuju na kraju intonacijske fraze.

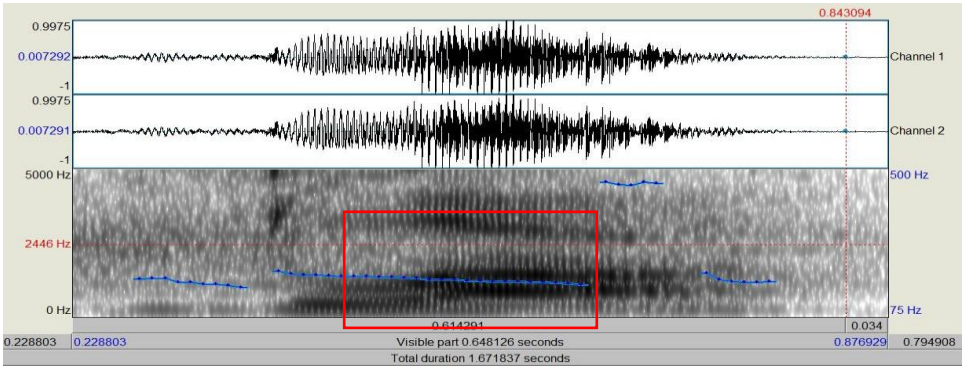
Više o Tilt modelu bit će riječi u 8. poglavlju.

4. Prozodijska obilježja hrvatskoga jezika

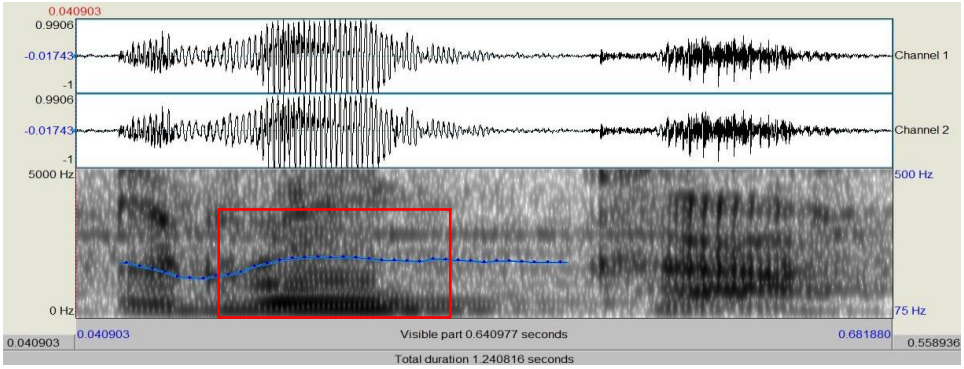
4.1 Hrvatski naglasni sustav

Jedno od najvažnijih prozodijskih obilježja hrvatskoga jezika je njegov naglasni sustav. Riječi se u govoru međusobno razlikuju naglasnim svojstvima koja uključuju silinu izgovora, kretanje tona (dizanje i spuštanje) i trajanje sloga. Upravo se tim svojstvima samoglasnik u naglašenom slogu razlikuje od nenaglašenih samoglasnika. Dakle, prema definiciji "naglasak je istodobni ostvaraj siline, tona i trajanja" (Barić, i dr., 1995.), a isticanje jednog sloga u riječi tim svojstvima nad drugima zove se naglasak riječi. Svojstvo siline očituje se u razlici snage zvučne struje pri izgovaranju slogova, a vezana je uz veći ili manji utrošak zraka iz pluća i veći ili manji potisak kojim zrak putuje kroz govorne organe. Tako se naglašeni slogovi očituju u utrošku veće količine zraka i jačem potisku u odnosu na nenaglašene slogove. Kretanje tona odnosi se na mijenjanje visine tona prilikom izgovaranja naglašenog tona i tona koji slijedi nakon njega. Ako se tijekom izgovora ton podiže, onda je riječ o uzlaznom tonu, a ako se ton spušta, onda se radi o silaznom tonu. Svojstvo trajanja odnosi se na razliku u dužini samoglasnika u naglašenom slogu. Prema navedenim svojstvima, u hrvatskom naglasnom sustavu razlikuju se četiri vrste naglaska te dugi i kratki zanaglasni slogovi. Pa tako postoje dugosilazni naglasak koji se ostvaruje primjerice u riječi

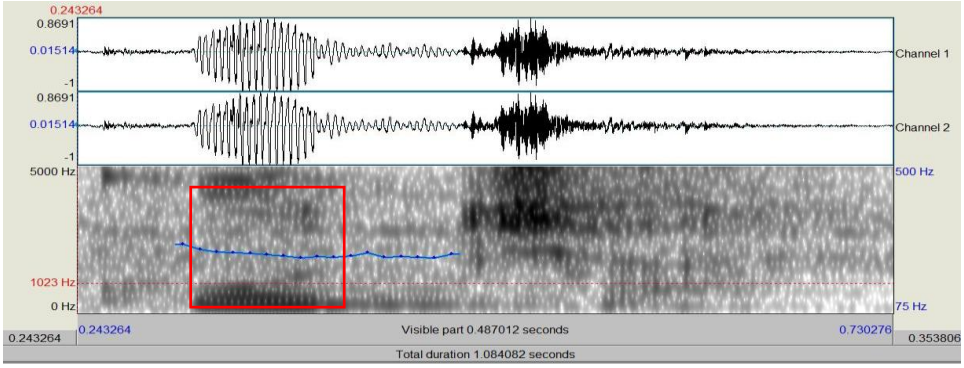
zlâto, dugouzlazni kao primjerice u *rúka*, kratkosilazni kao u riječi *kùća* i kratkouzlazni naglasak kao na primjer u riječ *žèna*. (Barić, i dr., 1995.) Na slici 12 mogu se vidjeti spektrogrami, zvučni valovi i F0 kontura četiri hrvatske riječi (*zlâto,rúka,kùća* i *žèna*) s različitom vrstom naglaska. Dugi zanaglasni slog ostvaruje se primjerice u riječi *djèčāk*. Jedinica za mjerenje dužine izgovora samoglasnika zove se mora.



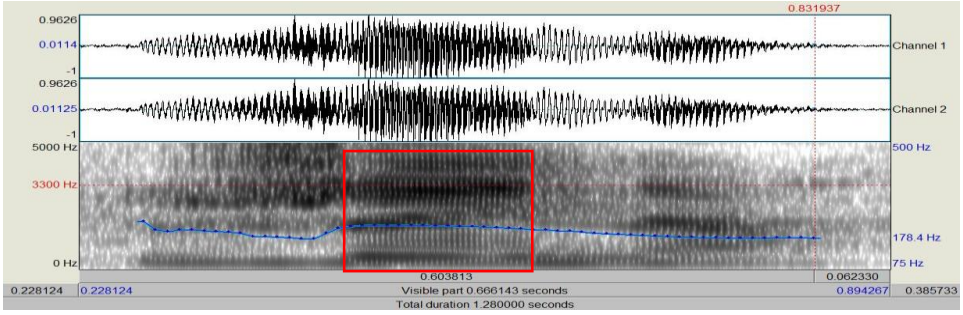
Dugosilazni naglasak u riječi zlâto



Dugouzlazni naglasak u riječi rúka



Kratkosilazni naglasak u riječi kȕća



Kratkouzlazni naglasak u riječi žèna

Slika 12 Četiri hrvatske riječi s različitim naglascima

4.1.1 Osobine naglašenog sloga

Naglašeni visoki slog je jačeg zvuka od prethodnih visokih slogova, a naročito od zanaglasnih niskih. Niski naglašeni slog je također jači od okolnih visokih, ali je ta razlika manja. Visoki naglašeni slogovi najjači su i najviši negdje nakon prve trećine trajanja samoglasnika, a kod niskih se na tom mjestu najviše razlikuje ton od jakosti – jakost je najjača, a ton najniži. Takvo se isticanje kod niskih tonova još naziva i obrnuto ili inverzno isticanje (Škarić, 1991).

4.1.1.1 Visoki naglašeni slog

Kod visokog naglašenog sloga, ton klizi uzlazno do visokog prednaglasnog sloga (ako ga ima) prema vrhu visokog naglašenog sloga i zatim silazi prema kraju naglašenog samoglasnika ili dalje prema niskom zanaglasnom slogu (ako ga ima). Zbog duljeg dvotrećinskog silaznog dijela naziva se još i silazni. Visoki naglašeni dugi slog naziva se dugosilazni ili samo silazni i bilježi se oznakom „ˆ“ iznad samoglasnika, a visoki naglašeni kratki slog naziva se kratkosilazni ili brzi i bilježi se s oznakom „ˋ“.

4.1.1.2 Niski naglašeni slog

Početak je niskog naglašenog tona lagano tonsko uzdignuće na samoglasniku, a to uzdignuće predstavlja tonsko isticanje naglašenosti tog sloga. Nakon početnog dijela, slijedi ulegnuće ili zastoj porasta čime se očituje unutarnja niskost tog sloga. Od prve trećine ton lagano raste tvoreći tonsko isticanje na prijelazu između niskog naglašenog i zanaglasnog visokog sloga. Zbog drugog duljeg uzlaznog dijela obično se takav naglašeni niski slog opisuje kao tonski uzlazni. Dugi niski naglašeni slog se još naziva dugouzlazni ili samo uzlazni i bilježi se s „ˊ“ iznad naglašenog samoglasnika , a kratki niski naglašeni slog se naziva kratkouzlazni ili spori i bilježi se s „ˋˋ“.

4.1.1.3 Trajanje naglašenih slogova

Naglašeni slogovi dulje traju od nenaglašenih slogova u riječi. Unutarnja prozodijska duljina slogova ostvaruje se tako što prozodijski dugi slogovi imaju dulje trajanje samoglasnika nego prozodijski kratki u istim uvjetima. U hrvatskom standardnom jeziku su naglašeni dugi slogovi prosječno za 50% dulji nego dugi nenaglašeni, a kratki naglašeni su za 30% dulji nego kratki nenaglašeni. Dugosilazni naglašeni slog je za 30% dulji nego kratkosilazni, a dugouzlazni za prosječno 22% dulji nego kratkouzlazni. Navedeni prosjeci temelje se na mjerenju povezanog govora, a ne izdvojeno izgovorenih riječi (Babić, i dr. 1991); (Škarić, 1991).

4.1.2 Raspodjela naglasaka

U hrvatskom je standardnom jeziku mjesto naglaska slobodno tj. može se ostvariti na bilo kojem slogu u riječi osim na zadnjem. Naglasak se najčešće ostvaruje na prvom slogu u riječi (u oko 66% riječi u tekstu), zatim na drugom (u oko 23% riječi), pa na trećem (6.7%) i na četvrtom (1.6%) (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991). Osim što mjesto naglaska nije vezano uz jedan slog, ono se može mijenjati i unutar paradigme, pa se tako primjerice u različitim oblicima iste riječi mogu naći sva četiri naglaska (*lònac* - Njd, *lónca* - Gjd, *lônče* - V, *lȍnācā* - Gmn).

4.1.2.1 Prozodijske osobine slogova u riječi

U jednoj govornoj riječi mora i smije biti samo jedan naglašeni slog, a svi ostali nenaglašeni. Ispred naglašenog sloga su svi slogovi visokog tona, a iza prvog zanaglasnog svi su slogovi niskog tona. Naglašeni slog može biti visokog li niskog tona. Prvi zanaglasni slog iza visokog naglašenog sloga je obavezno niska tona ili ga nema, a iza niskog naglašenog sloga obavezan je slog visokog tona. Ispred naglašenog sloga su svi slogovi kratki, a naglašeni i svi zanaglasni mogu biti dugi ili kratki.

Zanaglasna se dužina veže uz naglasak ispred sebe i ne može postojati kao jedina prozodijska jedinica u riječi. Nalazi se ili u osnovnoj riječi ili u tvorbenim nastavcima. U osnovnoj riječi može se naći u riječima naglasnog tipa *kàpūt*, *šèšīr,* ispred suglasničkog skupa koji započinje sonantom (ukoliko njega ne slijedi dugi slog) kao primjerice u riječi *làkōmca* te u govornoj riječi na mjestu naglaska koji se s naglasnice prebacio na prednaglasnicu kao na primjer */ùgrad/* Tvorbeni nastavci u kojima se nalaze zanaglasne dužine su primjerice genitiv množine svih padeža (*jedárā, nȍkātā*), u pojedinim glagolskim oblicima kao što je primjerice prezent (*vȉdīm, vȉdīm*), u dativu lokativu i instrumentalu množine sva tri roda (*dòbrim*).

Prozodijske osobine slogova u riječi potpunije se ostvaruju samo u riječi koja čini intonacijsku jezgru, a u drugim se riječima te osobine manje, više ili potpuno neutraliziraju. Unutar riječi, te se prozodijske osobine ostvaruju samo na samoglasnicima.

4.1.2.2 Raspodjela prozodije u riječima

Slobodna raspodjela prozodije u riječima može se ograničiti sljedećim pravilima (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991);(Barić, i dr., 1995.); (Vukušić, Zoričić, & Grasselli-Vukušić, 2007);

1) Silazni naglasci stoje samo na prvom slogu u riječi,

2) Na jednosložnim riječima stoje samo silazni naglasci,

3) Na posljednjem slogu u riječi nema naglaska,

4) Slogovi su pred naglašenim slogom samo kratki.

Ipak postoje kategorije riječi koje se mogu izuzeti iz pravila:

1) Neke složenice (npr. *poljopr̀ivreda*),

2) Strane jezično neusvojene riječi (npr. *dirigènt*),

3) Strana imena (npr. *Voltêr, Montevidèo*),

4) Kratice koje se izgovaraju imenovanjem početnih slova (npr. */esadê/*),

5) U genitivu množine riječi s nepostojanim *a* i s uzlaznim naglascima u ostalim padežima javljaju se silazni naglasci na nepočetnom slogu (npr. *muškârācā*),

6) Uzlazni naglasak može stajati na završnom slogu ili na jednosložnim riječima ako je otpao jedan slog iza naglaska (npr. *bicìkl*).

4.2 Govorna i jezična riječ

Jezična riječ je dio teksta koji pri sintagmatskim preinakama umetanjem ili premetanjem ostaje nerazdvojna cjelina. Dijelovi jezične riječi su morfemi: značenjska jezgra, prefiksi, infiksi i sufiksi. Jezične se riječi pišu odvojeno jedna od druge. Govornu riječ čine svi slogovi u nizu koji se sintagmatski odnose prema jednom naglašenom slogu te se na njega oslanjaju. To je najčešće značenjska jezgra i jedan ili više morfema koji označavaju jezične, modalne i logičke odnose jezgre. Morfemi su gramatički morfemi koji mogu biti dijelovi jezične riječi ili gramatičke riječi koje su jezično zasebne riječi – klitike ili atoničke riječi koje nemaju svog naglaska i ne mogu se pojaviti kao zasebne govorne riječi. Klitike se mogu prislanjati na riječ koja slijedi pa se nazivaju prednaglasnice ili proklitike, ili se naslanjaju na riječ ispred pa se nazivaju zanaglasnice ili enklitike, a to su pomoćni glagoli, zamjenice, čestica „li“. Sve ostale riječi koje nisu klitike nazivaju se naglasnice ili toničke riječi. U tablici 1 navedene su prednaglasnice i zanaglasnice prema (Barić, i dr., 1995.). Zanaglasnice su uvijek nenaglašene (npr. *vidim ga /vȉdīmga/, predali su nam se /prȅdālisunamse/*) dok su prednaglasnice nenaglašene kad se nalaze ispred riječi s uzlaznim naglascima (npr. *u vòdi, po ljepòti*), a postaju naglašene kada se nalaze ispred riječi sa silaznim naglascima - u tom se slučaju naglasak s naglasnice premješta na prednaglasnicu (npr. /*ùzoru/, /pòvodu/*). Više o pravilima za pomicanje naglaska s naglasnice na prednaglasnicu bit će riječi u sljedećem poglavlju.

Prozodiju govorne riječi čine broj slogova, prozodijska obilježja slogova i međusobni prozodijski odnosi, pa se govorna riječ još naziva i prozodijska. Govorna je riječ jedna prozodijska cjelina bez obzira na to da li se sastoji od jedne ili više jezičnih riječi. A to potvrđuje i preskakivanje naglasaka. Pravilo je da silazni naglasci stoje samo na prvom slogu, a kad se značenjskoj jezgri nadodaje prefiks ili proklitika naglasak se s prvog sloga u jezgri prebacuje naprijed, na prefiks ili prednaglasnicu (npr. *pòznati, nė̀ znati*). Iznimka su složenice gdje se silazni naglasak može naći i u središnjem dijelu riječi, a ponekad se udruživanje s proklitikom stavlja na istu razinu složenice (npr. */poljoprìvreda/*). S trosložnih i višesložnih riječi naglasak ne prelazi (npr. */poopomenama/, /unedogled/*).

Obzirom da se govorna riječ sastoji od jedne ili više jezičnih riječi, ona je prosječno za oko 40% veća nego jezična. U hrvatskom standardnom jeziku prosječna govorna riječ ima 3,12, slogova, a jezična 2,25. Jezične su riječi najčešće jednosložne (43,42%), pa dvosložne (27,3%), trosložne (21,6%), četverosložne (12,5%), peterosložne (3%), šesterosložne (0,85%). Govorne su riječi najčešće trosložne (31,5%), pa dvosložne (28,7%), slijede četverosložne

(22,4 %), peterosložne (8,6%), jednosložne (4,9%), šesterosložne (2,9%)(Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991).

Tablica 1 Prednaglasnice i zanaglasnice[4]

prednaglasnice			
Prijedlozi	jednosložni		svi jednosložni
	dvosložni		među, mimo, nada, poda, pokraj, preko, prema, oko
	trosložni		umjesto, svi složeni s prijedlogom iz- između, iznad, ispod…
veznici			a, i, ni, da, kad (kad može biti i naglašen)
niječna čestica			ne
zanaglasnice			
zamjeničke	nenaglašeni oblici osobnih zamjenica	genitiv	me, te, ga, je, nas, vas, ih
		dativ	mi, ti, mu, joj, nam, vam, im
		akuzativ	me, te, ga (nj), ju (je), nju, nas, vas, ih
	nenaglašeni oblici povratne zamjenice	genitiv	se
		dativ	si
		akuzativ	se
glagolske	nenaglašeni oblici prezenta glagola biti		sam, si, je, smo, ste, su
	nenaglašeni oblici prezenta glagola htjeti		ću, ćeš, će, ćemo, ćete, će
	nenaglašeni oblici aorista glagola biti		bih, bi, bi, bismo, biste, bi
			vezničko-upitno li

[4]Prema izvoru: (Barić, i dr., 1995.)

4.3 Rečenična intonacija

Rečenična jezična intonacija ima neke osobine koje su univerzalne u svim jezicima, a neke su pak specifične za svaki jezik. Neke od univerzalnih osobina rečenične intonacije u svim jezicima su: uzlazni ton na početku i padajući na kraju rečenice; nepotpuni intonacijski pad na kraju iskazuje nedovršenost rečenice, prekid govora, čuđenje, pitanje, oklijevanje ili slično; raščlanjivanje rečenične intonacije na manje intonacijske jedinice(Škarić, 1991).

4.3.1 Intonacijska jedinica

Rečenice se raščlanjuju na intonacijske jedinice – jednu, dvije ili više njih. Raščlanjivanje se ostvaruje u vidu kratke stanke razdvajanja. U pismu se granice intonacijskih jedinica obavezno označavaju zarezom. Međutim, u govoru se ostvaruje znatno veći broj tih jedinica.

4.3.2 Intonacijska jezgra

Intonacijske jedinice se dalje raščlanjuju na intonacijsku jezgru, intonacijski početak i intonacijski završetak, pri čemu je jezgra intonacijski najviši i najrazlikovniji dio intonacijske jedinice. Jezgru u europskim jezicima čini naglašeni slog istaknute riječi u intonacijskoj jedinici, a u hrvatskom uz naglašeni slog i slog iza njega zbog razlikovanja ulaznih od silaznih glasova.

U hrvatskom jeziku razlikujemo šest intonacijskih jezgara: silazna (\), uzlazna (/), silazno-uzlazna (v), silazno-uzlazna-silazna ili obrnuta (\/\), uzlazna i silazna ili složena (/ + \) te ravna (--). Njihova distribucija nije vezana uz gramatičke sintaktičke vrste (Babić, Brozović, Moguš, Pavešić, Škarić, & Težak, 1991).

4.3.2.1 Silazna jezgra

Najbrojnija intonacijska jezgra je silazna koja se nalazi u gotovo svim završnim intonacijskim jedinicama izjavnih i uskličnih rečenica, upitnih s upitnom riječi, te više od pola rečenica s li-pitanjima i pitanjima bez upitne riječi i nezavršnih jedinica.

I unutar silaznih intonacijskih jezgri ima razlike u ostvarenju, pa je tako u uskličnim rečenicama vrh jezgre viši, a silaženje tona usporeno; u upitnim rečenicama bez upitne riječi i u rečenicama s istaknutijom riječi ton je još i viši nego u usklicnima, ali s naglim padom tona;

kod upitnih rečenica s upitnom riječi, a s pitanjem na nekoj drugoj riječi, ton je visok a pad brz.

Kod nezavršnih jedinica, silaženje tona u jezgri intonacijske jedinice nije tako jako naglašeno kao kod završnih rečenica ili kod završetka odlomka gdje je silaženje tona još više naglašeno.

Tonske osobine naglaska riječi u hrvatskom jeziku najpotpunije se ostvaruju u silaznim intonacijskim jezgrama te se stoga kod akustičkog mjerenja naglasaka riječi uzimaju ostvarenja u silaznim jezgrama.

4.3.2.2 Uzlazna jezgra

U uzlaznoj jezgri ton raste u naglasnom i zanaglasnom slogu, a zatim u završetku intonacijske jezgre ostaje ravan ili blago pada. Takva se jezgra može naći samo na posljednjoj riječi intonacijske jedinice, a njezino je opće značenje nezavršnost iskaza. Česte su kod nabrajanja, u pitanjima bez upitne riječi ili s upitnom riječi, ali s upitom na nekoj drugoj riječi *(„Koliko je sati?“*), u upitnim uzrečicama *(„Je li?“*, *„Zar ne?“* i sl.) te u pozdravima i ljubaznim ponudama (npr. *„Dobar dan!“*, *„Izvolite!“* i sl.).

4.3.2.3 Silazno-uzlazno jezgra

Silazno-uzlazna jezgra nalazi se samo na kraju nezavršnih intonacijskih jedinica i to najčešće na pretpretposljednjima, a upućuje na rečenicu s više intonacijskih jedinica. Primjer rečenice sa silazno-uzlaznom jezgrom: *"Kad je ušao i ugledao nepoznatog čovjeka, silno se iznenadio..."*.

4.3.2.4 Silazno-uzlazna-silazna jezgra

Silazno-uzlazna-silazna jezgra je karakteristična samo za balkanske jezike. Zove se još i obrnuta ili inverzna jer je istaknuti slog nižeg, a ne višeg tona od prethodnoga (koji je dio intonacijskog početka). U takvoj je jezgri ton silazan kod naglašenog samoglasnika istaknute riječi, zatim naglo raste i ponovno naglo pada. Te se promjene ostvaruju na tri sloga ako ih ima, a inače na dva ili jednome. Njome se označava da/ne pitanje bez upitne riječi, a-pitanje i li-pitanje. Primjer rečenice sa silazno-uzlazno-silaznom jezgrom: *"Vi dajete <u>nama</u>?"*

4.3.2.5 Složena jezgra

Uzlazna i silazna jezgra zahvaća dvije susjedne riječi koje su podjednako istaknute ili u značenjskoj, idiomatskoj vezi ili vezi koja se ostvaruje pomoću veznika „i“ i „ili“. Složena jezgra razlikuje se od dvije posebne jezgre – uzlazne i silazne tako što nije presječena

stankom razdvajanja intonacijskih jedinica. Primjer rečenice sa složenom jezgrom: *"Naučio je čitati i pisati."*.

4.3.2.6 Ravna jezgra

Ova se vrsta intonacijske jezgre može naći samo na posljednjoj riječi u intonacijskoj jedinici. Rijetka je i zvuči pijevno. Javlja se u pozdravima kao inačica uzlazne jezgre *(„Laku noć!“*, *„Doviđenja!“* i sl.), u usklicima (*„O!“*), u dječjem hvalisanju *(„A ja imam ljepšu torbu!“*) te u nezavršnim intonacijskim jedinicama ispred stanke prekida govora, leksičke stanke i stanke procesiranja.

4.3.2.7 Kombiniranje intonacijskih jezgara u rečenici

Rečenice se obično sastoje od niza intonacijskih jedinica, a najčešće od jedne intonacijske jedinice s uzlaznom jezgrom (ili više njih u slučaju nabrajanja) nakon koje slijedi završna intonacijska jedinica sa silaznom jezgrom.

Kod bezvezničkih složenih rečenica (*„Ja bih još malo, odlično je.“*) te kod vezničkih u inverziji (*„Da je sve u redu, rekao je.“*), uobičajene su dvije uzastopne intonacijske jedinice sa silaznim jezgrama.

Jedinica sa silazno-uzlaznom jezgrom može se nalaziti u nizu ispred barem dvije drugačije jezgre.

4.3.3 Intonacijski početak

Svaki intonacijski početak ima porast tona do prvog naglašenog sloga koji može biti visok, srednji ili nizak. Upravo ton tog prvog naglašenog sloga određuje ton čitave intonacijske jedinice. Od prvog naglašenog sloga ton ide prema jezgri silazno, ravno ili uzlazno.

Intonacijski početak uobičajeno je najduži dio intonacijske jedinice. Može se skratiti pomicanjem jezgre naprijed, a može ga praktički i ne biti u slučajevima kad je istaknuta prva riječ u intonacijskoj jedinici ili kad se jedinica sastoji od jedne riječi.

Najčešći intonacijski početak je silazni, a nakon njega može slijediti bilo koja vrsta intonacijske jezgre.

Nakon ravnog intonacijskog početka mogu slijediti silazna, uzlazna ili silazno-uzlazna jezgra koje su na znatno višem tonu od intonacijskog početka. Takve intonacijske jedinice

daju dojam hladnoće i izražavaju nešto što je govorniku nevažno. Visok ravan intonacijski početak ispred silazne jezgre izražava uskličnost.

Uzlazni intonacijski početak obično je na niskom tonu, može stajati ispred silazne, uzlazne i inverzne intonacijske jezgre. Takav intonacijski početak izražava živahnost i srdačnost.

4.3.4 Intonacijski završetak

Intonacijski završetak ima nakon svih vrsta intonacijskih jezgara silazan ili nizak i ravan ton, osim nakon ravne jezgre kada je visok i ravan. Ako je kraj jezgre nizak, intonacijski završetak se produžuje u ravan nizak ton.

Obzirom da je intonacijska jezgra najčešće na zadnjoj riječi u jedinici, intonacijski je završetak obično vrlo kratak ili ga praktički i nema, a njegovu ulogu preuzima krajnji dio jezgre. Može biti dulji jedino kad se intonacijska jezgra pomakne unaprijed, pa se tada u njemu može još istaknuti poneka riječ.

Viši ton u intonacijskom završetku izražava upitnost i uskličnost ili nedovršenost rečenice u nezavršnih intonacijskih jedinica sa silaznom jezgrom. Na kraju odlomka ton se spušta te najčešće prelazi u šapat. Brzina je govora u intonacijskom završetku to sporija što je ton niži i glasnoća manja.

5. Hrvatski naglasni rječnik

Važnost naglaska kao prozodijske jedinice i uloge isticanja riječi, do sada je u ovom radu više puta istaknuta. Osim te uloge naglasak ima i ulogu isticanja razgraničenja riječi i razlikovnu ulogu u primjerima istovjetnosti dvaju ili više izraza riječi. U (Užarević, 2012) se navodi da je "osnovna uloga naglaska da u govornome kontinuumu izdvaja perceptibilne, auditivno raspoznatljive dionice – riječi kao „značenjske jedinice veće od morfema, a manje od rečenice“ (Garde, 1993.: 18.)." Dok se u pisanom tekstu jasno vide granice riječi, u govoru se granice među riječima naizgled ne mogu percipirati, osim na onim mjestima gdje su prisutne stanke u govoru. Međutim, ukoliko ne bi prepoznali riječi, ljudi ne bi mogli razumjeti poruku koja se govorom prenosi. Stoga je jasno da čovjek posjeduje sposobnost razgraničenja riječi u govoru i to upravo radi postojanja naglaska i ostalih prozodijskih jedinica. Naglasak riječi dakle ima vrlo važnu ulogu u komunikaciji jer slušatelj (primatelj informacije) na temelju naglaska riječi percipira granicu između riječi.

Obzirom da se u hrvatskom jeziku ne pišu naglasci, u pisanim riječima (slovopisu) postoje riječi i oblici koji imaju isti slijed grafema. Takve se riječi nazivaju homogrami. Ukoliko je kod takvih riječi prisutan i jednak slijed prozodema, tj., jednako se i izgovaraju, onda se nazivaju homofoni, a ako se razlikuju naglasno, tj. izgovaraju se različito, onda je riječ o heterofonima. Naglasak ima važnu razlikovnu funkciju kod heterofonskih homograma koji imaju isti grafemski, a različit prozodemski slijed. Pa se tako primjerice vrstom naglaska

razlikuju riječi *lûk* i *lùk*. Dakle, razlikovna se uloga prozodijskih jedinica ostvaruje u mijenjanju gramatičkih i leksičkih značenja oblika jedne riječi kao na primjer kod *róda (N jd. im. róda)* i *ròda (G jd. im. rôd)*, različitih riječi kao primjerice u *grȁd (tuča)* i *grâd (naselje)* te različitih oblika iste riječi kao primjerice u *sèla (G jd.)* i *sȅla (N mn.)*. Zato je vrlo važno prilikom modeliranja prozodije za hrvatski jezik uključiti i leksički naglasak.

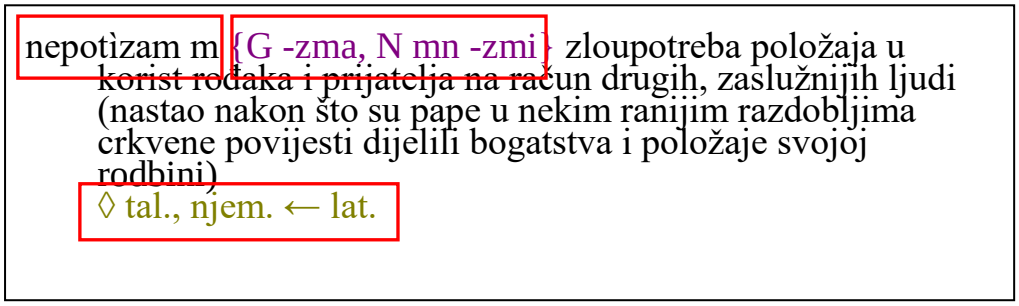
Kao što je rečeno ranije, sustav naglašavanja u hrvatskom jeziku je prilično složen te je gotovo proizvoljan pa je za automatsko dodjeljivanje naglasaka riječi neophodan rječnik svih osnovnih i izvedenih oblika leksema s označenim naglascima. Za hrvatski jezik takav rječnik nije postojao te se je iz gore navedenih razloga pristupilo njegovoj izradi kao dio ovog doktorskog rada.

Hrvatski je jezik izrazito flektivan sa složenim konjugacijskim i sklonidbenim sustavima. Naglasak riječi, međutim, ipak poštuje određena pravila koja se mogu implementirati u naglasni rječnik, koji bi pohranio informacije o mjestu leksičkog naglaska za što više hrvatskih riječi. U svom je doktorskom radu (Mikelić Preradović, 2008) razvila modele i izdvojila pravila i iznimke za automatsko generiranje naglasnih oblika riječi u svim padežima, odnosno licima i oblicima, uzimajući njihov osnovni oblik iz leksičke baze i pridružujući im valjane paradigmatske nastavke i naglaske.

U ovom je poglavlju opisano kako je na temelju tih pravila stvoren naglasni morfološki analizator/generator koji riječima automatski dodjeljuje odgovarajući naglasak.

Dio je pravila iz navedenog doktorskog rada koja se odnose na dvosložne i trosložne glagole implementiran u programskom jeziku Python u sklopu dva završna rada (Cukor, 2009); (Horvacki, 2009), dok su se preostala pravila i iznimke koja se odnose na višesložne glagole te sve imenice i pridjeve implementirala u ovom doktorskom radu.

Kao leksička baza za stvaranje naglasnog rječnika poslužio je *Veliki rječnik hrvatskoga jezika* (Anić, 2009), koji je jedini hrvatski rječnik s označenim naglascima na osnovnom obliku riječi dostupan u računalno pretraživom obliku. Rječnik ukupno sadrži 70.576 leksema s označenim naglascima, ali ne sadrži sve izvedene oblike već samo osnovne, kod nekih imenica genitive i/ili množinu, kod pridjeva određeni oblik te kod glagola prezent 1. lica jednine, glagolski pridjev sadašnji, glagolski pridjev prošli i glagolsku imenicu. Od oznaka vrsta riječi kod većine unosa prisutne su samo osnovne oznake za rod (muški, ženski, srednji). Sve potrebne informacije iz rječnika izdvojile su se korištenjem regularnih izraz (slika 13).



Slika 13 Primjer oblika zapisa natuknice u rječniku i potrebne informacije

5.1 Postupak izrade rječnika

Naglasna svojstva riječi mogu se proučavati s gledišta morfologije i tvorbe riječi, pa onda možemo govoriti o naglasku jednine i množine imenica, o naglasku prezenta, komparativu pridjeva i sl. Takav se leksički naglasak naziva morfološki naglasak.

Promjenjive vrste riječi u koje se ubrajaju imenice, pridjevi, brojevi, zamjenice, prilozi i glagoli imaju više oblika, pa tako možemo govoriti o deklinaciji (sklonidbi), komparaciji i konjugaciji riječi. Nepromjenjive riječi u koje spadaju prijedlozi, veznici, čestice i uzvici imaju jedan oblik.

Pravila koja su izdvojena u (Mikelić Preradović, 2008) odnose se na najbrojnije skupine promjenjivih vrsta riječi - imenice, glagole i pridjeve, a uzimaju u obzir promjenu tih vrsta riječi dodavanjem morfoloških nastavaka i promjenom naglasaka ovisno o stupnju sklonidbe, komparacije i konjugacije. Možemo reći da pravila predstavljaju pregled sklonidbe imenica i pridjeva, komparacije pridjeva i konjugacije glagola po morfološko-naglasnim tipovima.

5.1.2 Imenice

Gramatičke osobine imenica su rod (muški, ženski i srednji), broj (jednina i množina) i padež (nominativ, genitiv, dativ, akuzativ, vokativ, lokativ i instrumental). Ove tri gramatičke osobine imaju posebno gramatičko značenje, ali su sve tri izražene jednim nastavkom. Za tvorbu oblika imenica služe nastavci triju vrsta, koje se prema nastavku u genitivu jednine zovu *vrsta a*, *vrste e*, *vrsta i*. Po *vrsti a* sklanjaju se imenice muškog i srednjeg roda, po *vrsti e* većinom imenice ženskoga roda i neke muškoga, a po *vrsti i* imenice ženskoga roda. Za svaki tip imenica, za svaki rod definirana su pravila za jednosložne, dvosložne, trosložne i višesložne osnove, a pritom se u obzir uzimaju brojna pravila i iznimke za tvorbu imenica i premještanje vrste i mjesta naglaska kao i posebnosti nekih imenica koja slijede vlastiti tip morfološko-naglasne sklonidbe, kao što su primjerice tuđice. Ukupno je u pravilima definirano 196 tipova za muški rod, 73 za ženski te 80 za srednji rod. Sva su ta pravila implementirana s ukupno više od 11 000 linija programskoga kôda u programskom jeziku Python.

5.1.2.1 Opis postupka obrade naglasno-sklonidbenog modela imenica[5]

1) Najprije se odredi broj slogova.
2) Zatim se odredi rod.
3) Potom se za:
    - imenice muškog roda: pomoću oblika za Njd i Gjd iz rječnika odredi pripada li imenica *vrsti a* ili *vrsti e* ili posebnoj osnovi,
    - imenice ženskog roda: pomoću oblika za Njd i Gjd iz rječnika odredi radi li se o *vrsti e* ili *vrsti i*,
    - imenice srednjeg roda: odredi se je li imenica jednakosložna ili nejednakosložna.
4) Zatim se provjerava postoji li u imenici nepostojanog *a* pomoću nominativa i genitiva.
5) U sljedećem koraku odredi se na kojem slogu se nalazi naglasak.
6) Zatim se određuje:
    - za imenice ženskog roda odredi se da li imenica u Njd završava na *–ka/-ga/-ha*, *-ao, -st, -ost, -ica*, itd. te da li nakon oduzimanja nastavka imenica završava na palatal, velar, ili na 2 suglasnika,
    - za imenice muškog roda odredi se imaju li dugu ili kratku množinu (iz informacije o Nmn koju uzima iz leksičke baze),
    - za imenice muškog roda odredi se da li imenica završava na *–c/-ac*, –o, *-ao, -lac*, sonant+*ac* ili +*ak*, nesonant +*ac* ili +*ak*, palatal, velar ili ni na što od tog ,
    - za jednakosložne imenice muškog roda odredi da li imenica završava na *–o, -e* ili ima samo množinski oblik, a za nejednakosložne tip umetka u Gjd *(-n, -t, es-, -v*).
7) Na kraju se dodaju nastavci za sklonidbenu paradigmu i po potrebi promijeni naglasak generiranih oblika prema zadanim pravilima (tj. iznimkama).

[5]prema (Mikelić Preradović 2008)

5.1.2.2 Muški rod

Neke od posebnosti imenica muškoga roda:

1) Kod imenice muškoga roda, ukoliko ona znači nešto živo, onda je akuzativ jd. jednak nominativu jd. Ukoliko pak ona znači nešto neživo, tada je akuzativ jd. jednak genitivu jd. Ako imenica znači i živo i neživo, onda se oblik uzima prema značenju - u stvorenom naglasnom generatoru, za imenice muškog roda se uzima u obzir, tj. generira i jedan i drugi oblik budući da u rječniku ne postoji oznaka za svojstvo živo/neživo.

2) Kod nekih imenica koje završavaju primjerice na *k, g* ili *h* ili na nepčani suglasnik ili skupine suglasnika *st, št, zd, žd* dolazi do promjene (alternante) u nekim nastavcima npr. - *vojnik - vojniče, ražanj - ražnja - ražanja*.

3) Ako osnova završava na suglasničku skupinu, u nom. jedn. se ispred posljednjeg suglasnika umeće kratko *a* - nepostojano *a* (pojavljuje se i u genitivu množine - npr. *nokata*).

4) U nekim se oblicima javlja obezvučena osnova - npr. *hrbat - hrpta*.

5) Imenice muškog roda imaju kratku i/ili dugu množinu - npr. *palac - palci -palčevi*.

6) Neke imenice imaju posebnu osnovu kod kojih sklonidba i promjena naglaska također slijedi vlastita pravila, a takve su primjerice imenice koje završavaju na *-anin* u N jd. i *-ani* u N mn., imena, tuđice.

Sve su posebnosti uzete u obzir prilikom programiranja sklonidbe. Iz rječnika su se sve te posebnosti uspjele detektirati, osim one o imenu - u rječniku ne postoji oznaka da je imenica ime. Neke se posebnosti očituju u nastavku pa je njih bilo lako izdvojiti, a neke u rječniku imaju posebnu oznaku - primjerice hipokoristici imaju oznaku *hip.*, a tuđice oznaku jezika iz kojeg su porijeklom. Na temelju tih oznaka izdvojile su se, primjerice, sve imenice koje su tuđice. Za neke je posebnosti bilo potrebno iz rječnika izdvojiti i dodatne informacije - primjerice, za utvrđivanje prisutnosti nepostojanog *a*, bilo je potrebno iz rječnika izdvojiti i genitiv imenice. U rječniku se nalaze imenice koje imaju ispisan cijeli genitiv kao primjerice *àdūt m {G adúta}*, ali postoje i imenice kojima je genitiv zapisan samo pomoću nastavka - primjerice *alternatívac m {G -vca, N mn -vci}*. Ponekad se taj nastavak samo dodaje na osnovu, a ponekad mijenja određeni broj zadnjih grafema u riječi pa se je stoga definirao algoritam za izdvajanje genitiva iz rječnika. Algoritam se može naći u Prilogu 1. Zbog utvrđivanja postojanja duge množine, bilo je potrebno izdvojiti i nominativ množine koji je u rječniku zapisan na isti način kao i genitiv, pa je i za njega definiran algoritam za izdvajanje. Algoritam se može naći u Prilogu 2.

Primjer pravila za jedan tip imenice, poziv funkcije iz glavnog programa te sama funkcija prikazane su na slikama 14, 15 i 16.

1. **Jednosložna, vrsta i, s ^ naglaskom na 1. slogu, Gjd=Njd (tipa stvar)**

Sklonidbena paradigma se tvori dodavanjem sljedećih nastavaka: N: - /G:-i /D:-i /A:- /V: -i- /L:-i / **I:-i/-u/-ju** /Nmn:-i /Gmn:- ī /Dmn:-ima /Amn:-i /Vmn:-i /Lmn:-ima /I:-ima

Ako imenica u N.jd završava na **ć, đ, lj** u Ijd ima samo nastavak –u.Ako imenica u N.jd završava na **č, š, ž** u Ijd ima samo nastavak –ju.Ako Ijd tvorimo nastavkom –ju, a ispred –ju se nalazi: **b, d, l, m, n, p, t, v, sl, sn, st, zn,** suglasnici i suglasnički skupovi se mijenjaju ovako: **b-blj, d-đ, l-lj, m-mlj,m n-nj, p-plj, t-ć, v-vlj, sl-šl, sn-šnj, st-šć, zn-žnj.**

Naglasna paradigma: Imenica u jd. i mn. ima naglasak Njd.

Iznimke: Imenica u Ljd, te GDLI mn ima dugouzlazni naglasak (/) na 1. slogu.

Slika 14 Primjer pravila za jedan tip imenice[6]

```
## 3.  Jednosložna, vrsta i,  s ^ naglaskom na 1. slogu, Gjd=Njd (tipa stvar)
 elif (u'ŕ' in osn or u'â' in osn or u'ê' in osn or u'ĩ' in osn or u'ô' in osn or u'û' in osn) and ai==1:
        zr3(osn,nosn,dugosilazni1,dugouzlazni1,slog1)
        af.write('\n')
```

Slika 15 Pozivanje funkcije iz glavnog dijela programa

[6] iz (Mikelić Preradović, 2008)

```
## 3.  Jednosložna, vrsta i,  s ^ naglaskom na 1. slogu, Gjd=Njd (tipa stvar)
def zr3(osn,nosn,naglasak,naglasak2,slog):
    nosn2=promjena_zr(nosn)
    zrNj=nosn
    zrGj=nosn+u'i'
    zrDj=nosn+u'i'
    zrAj=nosn
    zrVj=nosn+u'i'
    zrLj=nosn+u'i'
    if (osn[-1] in u'ć') or (osn[-1] in u'đ') or (osn[-2:] in u'lj'):
        zrIj=nosn2+u'u'
    elif(osn[-1] in u'č') or (osn[-1] in u'š') or (osn[-1] in u'ž'):
        zrIj=nosn2+u'ju'
    else:
        zrIj=nosn+u'i'
        zrIj2=nosn2+u'u'
        zrIj3=nosn2+u'ju'
    zrNm=nosn+u'i'
    zrGm=nosn+u'ī'
    zrDm=nosn+u'ima'
    zrAm=nosn+u'i'
    zrVm=nosn+u'i'
    zrLm=nosn+u'ima'
    zrIm=nosn+u'ima'
    af.write(slog(zrNj,naglasak)+'\t')
    af.write(slog(zrGj,naglasak)+'\t')
    af.write(slog(zrDj,naglasak)+'\t')
    af.write(slog(zrAj,naglasak)+'\t')
    af.write(slog(zrVj,naglasak)+'\t')
    af.write(slog(zrLj,naglasak2)+'\t')
    if (osn[-1] in u'ć') or (osn[-1] in u'đ') or (osn[-2:] in u'lj'):
        af.write(slog(zrIj,naglasak)+'\t')
    elif(osn[-1] in u'č') or (osn[-1] in u'š') or (osn[-1] in u'ž'):
        af.write(slog(zrIj,naglasak)+'\t')
    else:
        af.write(slog(zrIj,naglasak)+'\t')
        af.write(slog(zrIj2,naglasak)+'\t')
        af.write(slog(zrIj3,naglasak)+'\t')
    af.write(slog(zrNm,naglasak)+'\t')
    af.write(slog(zrGm,naglasak2)+'\t')
    af.write(slog(zrDm,naglasak2)+'\t')
    af.write(slog(zrAm,naglasak)+'\t')
    af.write(slog(zrVm,naglasak)+'\t')
    af.write(slog(zrLm,naglasak2)+'\t')
    af.write(slog(zrIm,naglasak2)+'\t')
```

Slika 16 Primjer funkcije za jedan tip imenice ženskoga roda

5.1.2.3 Ženski rod

Posebnosti imenica ženskoga roda:

1) Imenice *vrste e* koje završavaju na *k, g* ili *h* imaju sibilariziranu osnovu (npr. *tuga - tuzi*), osim u posebnim slučajevima kao što su primjerice hipokoristici (npr. *baka - baki*), zemljopisna imena (npr. *Krka - Krki*), imena (npr. *Olga - Olgi*), neke pojedinačne imenice poput *kuka - kuki, ovrha - ovrhi* itd. Neke od iznimaka evidentirane su u obliku listi u programskom kôdu, a u budućnosti se planira stvaranje većih popisa zemljopisnih imena i ostalih imena koje bi se uzele u obzir. U pravila su implementirane iznimke hipokorističnih imenica, obzirom da u rječniku postoji oznaka za hipokoristike. Na temelju roda, nastavka i oznake se onda utvrdilo da se radi o skupini imenica koje nemaju sibilariziranu osnovu.

2) Imenice *vrste i* koje završavaju na dva suglasnika, osim na *-st*, te *-št* imaju u nominativu jednine umetnuto nepostojano *a* (npr. *sablazan - sablazni*).

3) Nastavak *-i* u instrumentalu jednine mogu imati sve imenice, a nastavke *-ju* ili *-u* većina. Pritom izbor nastavka *-ju* ili *-u* ovisi od krajnjeg suglasnika osnove. U naglasni rječnik uključene su sva tri oblika.

4) Imenice kojima osnova završava na usnene i zubne suglasnike dobivaju nastavak *-u* na jotiranu osnovu.

5.1.2.4 Srednji rod

Posebnosti imenica srednjega roda:

1) Imenice srednjega roda mogu biti jednakosložne - ukoliko imaju isti broj slogova u nominativu jednine i genitivu jednine, te nejednakosložne ukoliko imaju različiti broj slogova u nominativu i genitivu jednine (npr. *čudo - čudesa*). Svaka od dvije vrste slijedi drukčija pravila za tvorbu izvedenih oblika.

2) Imenice koje završavaju na *-o* kojem prethode dva ili više suglasnika, osim na *-st* i-*zd,* u genitivu množine imaju proširenu osnovu (npr. *jedro - jedara*). Isto tako, imenice koje završavaju na *e*, kojem prethode dva ili više suglasnika (osim *št, šć, žđ*), imaju nepostojano *a* u genitivu množine.

3) Nejednakosložne imenice mogu imati različite umetke u pojedinim oblicima, kao primjerice *uže - užeta, rame - ramena, podne - podneva*.

4) Neke imenice nemaju množinski oblik nego samo zbirni, primjerice *tele -telad, pile - pilad*.

5) Neke imenice imaju samo množinski oblik - primjerice *vrata, leđa.*

Kao kod imenica za muški rod, i za ženski te srednji rod su se sve posebnosti i iznimke pojedinih tipova imenica uzele u obzir prilikom implementiranja pravila za naglašavanje i dodavanje morfoloških naglasaka.

5.1.3 Glagoli

Glagoli su vrsta riječi kod koje određeni oblici podliježu konjugaciji ili sprezanju. Osnovni oblik riječi odnosno infinitiv se prema utvrđenim pravilima za dodavanje tvorbenih nastavaka i promjenu naglaska pretvara u izvedene oblike. Među jednostavne glagolske oblike ubrajaju se prezent, aorist, imperfekt, imperativ, prilog sadašnji, prilog prošli, pridjev radni, pridjev trpni. Složeni se glagolski oblici tvore od infinitnih oblika glagola koji se sprežu i oblika pomoćnih glagola biti, htjeti, bivati.

Pravila koja su izdvojena za stvaranje izvedenih oblika glagola s označenim naglaskom odnose se na razradu jednostavnih oblika prema različitim tipovima glagola. Utvrđivanje tipa glagola vrši se prema infinitivu koji se nađe u rječniku, nastavku za prezent 1. lica jednine te broju slogova. Zatim se još nađe aspekt glagola, tj. odredi se je li glagol svršen ili nesvršen te je li prelazan ili neprelazan. Određivanje aspekta potrebno je zbog toga jer samo nesvršeni glagoli imaju imperfekt, a aorist imaju svršeni glagoli i, rjeđe, nesvršeni. Također, glagolski prilog sadašnji imaju samo nesvršeni glagoli. Glagolski pridjev trpni imaju prijelazni glagoli. Nakon toga se odredi vrsta i mjesto naglaska na infinitivnoj i prezentskoj osnovi. Svi su se navedeni podaci pomoću regularnih izraza izdvojili iz rječnika. Ukupno je pravilima obuhvaćeno 147 grupa i podgrupa glagola, a programski kôd kojim su implementirana pravila broji nešto više od 5000 linija.

5.1.3.1 Opis postupka obrade naglasno-sklonidbenog modela glagola[7]

1) Najprije se odredi broj slogova u infinitivu glagola iz rječnika.
2) Potom se odredi kojim nastavkom završava osnova u 1.l. jd. prezenta (također iz rječnika).
3) Nakon toga pronađe se nastavak za infinitiv u leksičkoj bazi i informacija o aspektu i prijelaznosti glagola.
4) Zatim se odredi:
   - za dvosložni glagol: tip naglaska na samoglasniku na 1. slogu u infinitivu i na 1. slogu u prezentu,
   - za trosložni glagol: redni broj sloga na kojem je naglasak na prezentskoj i infinitivnoj osnovi i potom tip naglaska,
   - za višesložni glagol: redni broj sloga na kojem je naglasak od kraja na prezentskoj i infinitivnoj osnovi i potom tip naglaska.

[7] prema (Mikelić Preradović 2008)

5) Na kraju se dodaju nastavci za konjugacijsku paradigmu i po potrebi promijeni naglasak generiranih oblika prema zadanim pravilima (tj. iznimkama).

5.1.4 Pridjevi

Pridjevi su riječi koje se sklanjaju po padežima, a također su podložni i stupnjevanju. Osim toga postoji i određeni i neodređeni vid pridjeva te posvojni i opisni pridjevi, a svaka od tih kategorija (uz potkategorije) podliježe različitim tipovima morfološko-naglasne sklonidbe. Dakle, za svaki dvosložni i trosložni posvojni pridjev iz rječnika izgenerirani su oblici za sve padeže i za sva tri roda te su za svaki jednosložni, dvosložni i trosložni opisni pridjev izgenerirani svi oblici za sva tri roda kroz sva tri stupnja - pozitiv, komparativ i superlativ neodređenog vida te kroz sve padeže za sva tri roda opisnog pridjeva određenog vida.

Problem na koji sam naišla kod implementiranja pravila za izvođenje svih oblika pridjeva je taj da u rječniku ne postoji genitiv pridjeva koji je potreban za utvrđivanje postojanja nepostojanog *a*. Dakle, ukoliko je prisutno nepostojano *a*, tada pridjev slijedi drukčija pravila tvorbe nego pridjev bez nepostojanog *a*. Obzirom da se programski kôd kasnije koristio kao analizator automatskog dodjeljivanja naglaska riječima iz teksta, problemu se priskočilo na način da su svi pridjevi izgenerirani dva puta - jednom uvažavajući pravila za tvorbu s nepostojanim *a*, a onda uvažavajući pravila za tvorbu bez nepostojanog *a*. Na taj način su dobiveni oblici pridjeva, kako onih s nepostojanim *a*, tako i onih bez nepostojanog *a*. Pritom su se izgenerirali i oblici koji nisu valjani jer, primjerice, ako se pridjev s nepostojanim *a* sklanja prema pravilu sklanjanja za pridjeve bez nepostojanoga *a*, onda se dobije nevaljani oblik (primjerice genitiv od *plitak* je *plitaka* umjesto *plitka*). Kod generatora takva pojavnost može smetati jer oblik nije valjan, ali kod analizatora ne smeta jer se takva riječ neće naći u tekstu.

Budući da su pravila za naglasno-sklonidbene promjene pridjeva definirana samo za jednosložne, dvosložne i trosložne pridjeve, za višesložne oblike nisu se mogli dobiti izvedeni oblici. Ukupno je pravilima obuhvaćeno 97 različitih jednosložnih, dvosložnih i trosložnih tipova pridjeva, a za programiranje pravila bilo je potrebno preko 5000 linija kôda.

5.1.4.1 Opis postupka obrade naglasno-sklonidbenog modela pridjeva[8]

1) Najprije se odredio broj slogova osnovnog oblika iz rječnika.
2) Zatim se prema nastavku odredilo je li pridjev opisni ili posvojni.
3) U sljedećem koraku odredilo se ima li neodređeni oblik opisnog pridjeva nastavak *-eo, -ao, -ok, -tak, -ak*, zvučni suglasnik+*ak* ili ne te završava li posvojni pridjev na *–ji, -nji, -šnji, -ski, -ki, -ev/-ov/-ljev* ili *–in*.
4) Potom se odredio redni broj naglašenog sloga.
5) Slijedi dodatno provjeravanje koje se odnosi na jednosložne pridjeve i dvosložne pridjeve (koji ne završavaju na *-eo*, ni *-ao* ni *-ok*, ni *-tak* ni *-ak*) s ciljem da se odredi da li završavaju na palatal, na *k/g/h/c/z/s/t/d/l/n/p/b/m/v/f/st, –ije* ili ni jedno od tog.
6) Na kraju su se dodali nastavci za sklonidbenu paradigmu, nastavci za komparativ i superlativ i po potrebi promijenio naglasak generiranih oblika prema zadanim pravilima (tj. iznimkama) te se također generirao određeni oblik pridjeva dodavanjem nastavaka pozitivu neodređenog vida pridjeva.

5.1.5 Ostale vrste riječi

Od ostalih vrsta riječi, u hrvatskom jeziku još su prisutne zamjenice, prilozi, prijedlozi, brojevi, uzvici, čestice i veznici. Neke od njih su, kao što je prije rečeno, nepromjenjive vrste - prijedlozi, uzvici, čestice i veznici, a neke su promjenjive - zamjenice, prilozi i brojevi.

Sve riječi koje pripadaju ovdje nabrojanim vrstama preuzele su se iz leksičke baze i pridodale naglasnom rječniku. Za neke su se promjenjive vrste riječi još ručno dodali dodatni oblici, kao što su, primjerice, izvedeni oblici zamjenica kroz padeže (s pripadajućim naglaskom) - *mene, meni, nje, sobom*, itd. koje su pronađene u (Barić, i dr., 1995.). U naglasni su rječnik također ručno dodani i pojedini izvedeni oblici brojeva - primjerice *jednih, dvaju,* itd. koji su također pronađeni u (Barić, i dr., 1995.).

[8]prema (Mikelić Preradović 2008)

5.2 Dodavanje MSD oznaka

POS (engl. part-of-speech) oznakama pridružuju se gramatičke kategorije pojavnicama u tekstu. U jezicima koji nisu morfološki bogati (primjerice u engleskom ili francuskom) dostatno je odrediti vrstu riječi (i obilježiti je POS oznakom). Budući da je hrvatski izrazito flektivan jezik i ima brojne oblike riječi zbog promjena riječi u deklinacijama, konjugacijama i komparacijama, dodjeljivanje POS oznake nije dovoljno kako bi se odredila gramatička kategorija riječi. Za morfološki bogate jezike, uz vrstu riječi, dodaju se i ostale morfosintaktičke oznake (u daljnjem tekstu MSD) poput padeža, broja, itd. Prilikom ispisivanja izvedenih oblika riječi u naglasni rječnik, svakom je obliku dodana i MSD oznaka u obliku koji odgovara MULTEXT-East standardu[9](Erjavec, Krstev, Petkevič, Simov, Tadić, & Vitas, 2003). Iako se nisu mogli dodati svi atributi oznaka jer se neka pravila za izvođenje izvedenih oblika riječi odnose na više vrijednosti pojedinog atributa (primjerice, pravilima nije određeno radi li se o općoj (vrijednost atributa *c* u kategoriji tip morfosintaktičke oznake za imenicu) ili vlastitoj imenici (vrijednost atributa *p* u kategoriji tip morfosintaktičke oznake za imenicu), ipak se smatra da će naglasni rječnik s navedenim MSD oznakama biti višestruko koristan u području jezičnih tehnologija i području računalne obrade prirodnih jezika. Osim što se može koristiti za (barem djelomičnu) morfosintaktičku analizu, može se iskoristiti i u druge svrhe kao što je primjerice poboljšanje rezultata sinteze i prepoznavanja govora, povećanje točnosti morfosintaktičkog označivača teksta i sl.

U ranije navedenim primjerima gdje riječi imaju isti pisani oblik, ali različiti naglasak i različitog su značenja, tj. pripadaju različitim lemama - *róda (N jd. im. róda)* i *ròda (G jd. im. rôd)* ili se radi o istoj riječi, ali različitog oblika i različitog naglaska - *sèla (G jd.)* i *sȅla (N mn.)* problem dvosmislenosti lako se može riješiti pomoću naglasnog rječnika koji uz naglašeni oblik riječi ima i MSD oznaku. Dakako, time nije riješen problem imenica koji imaju isti pisani oblik, a različiti naglasak - *grȁd (tuča)* i *grâd (naselje).* Za rješavanje tog problema, uz morfosintaktičku analizu trebala bi se provesti i semantička analiza jer se dvosmislenost može riješiti samo uz pomoć konteksta. U budućnosti se planiraju identificirati takvi parovi riječi i uz pomoć kolokacija razriješiti problem višeznačnosti.

[9] MULTEXT-East morfosintaktičke specifikacije opisuju oznake riječi koje se koriste kod jezika koji se govore u srednjoj i istočnoj Europi, a među njima se nalazi i hrvatski jezik.

Potpuni popis MSD oznaka po vrstama riječi za hrvatski jezik može se naći u Prilogu 5, a MSD oznake koje se dodane riječima u rječniku (prema MULTEXT-East standardu za hrvatski jezik (Ljubešić, 2013):

- Imenice: dodani svi atributi osim *Type* (*common* i *proper*) - na to mjesto stavljena je oznaka '-'; atribut *Animate* (*yes* i *no*) dodan je samo kod imenica muškog roda kod kojih je poznato da li se pravilo odnosi na živo ili neživo, a kod ostalih je također stavljena oznaka '-' (prikazano u tablici 2).
- Glagoli: dodani atributi - *VForm, Person, Number.*
- Pridjev radni pripada kategoriji *Verb* i takvim su oblicima dodani atributi *VFormp - participle, Number i Gender.*
- Pri izvođenju oblika iz osnove glagola nastale su riječi koje po MULTEXT-East standardu pripadaju drugoj kategoriji:
    - pridjev trpni pripada kategoriji *Adjective* i za takve je oblike dodana oznaka kategorije *Adjective* sa svim atributima osim *Animate,*
    - prilog sadašnji i prošli pripadaju kategoriji *Adverb* i njima je dodan atribut *Type - r (participle).*
- Pridjevi: svi atributi osim *Animate.*
- Sve ostale vrste riječi: dodana oznaka kategorije (*P, R, Q, I, S, C, M*).

Tabela 2 MSD oznake koje su dodane u naglasni rječnik za imenice:

IMENICE						
		muški rod			ženski rod	srednji rod
		živo	neživo	nije poznato		
jednina	nominativ	N-msny	N-msnn	N-msn-	N-fsn-	N-nsn-
	genitiv	N-msgy	N-msgn	N-msg-	N-fsg-	N-nsg-
	dativ	N-msdy	N-msdn	N-msd-	N-fsd-	N-nsd-
	akuzativ	N-msay	N-msan	N-msa-	N-fsa-	N-nsa-
	vokativ	N-msvy	N-msvn	N-msv-	N-fsv-	N-nsv-
	lokativ	N-msly	N-msln	N-msl-	N-fsl-	N-nsl-
	instrumental	N-msiy	N-msin	N-msi-	N-fsi-	N-nsi-
množina	nominativ	N-mpny	N-mpnn	N-mpn-	N-fpn-	N-npn-
	genitiv	N-mpgy	N-mpgn	N-mpg-	N-fpg-	N-npg-
	dativ	N-mpdy	N-mpdn	N-mpd-	N-fpd-	N-npd-
	akuzativ	N-mpay	N-mpan	N-mpa-	N-fpa-	N-npa-
	vokativ	N-mpvy	N-mpvn	N-mpv-	N-fpv-	N-npv-
	lokativ	N-mply	N-mpln	N-mpl-	N-fpl-	N-npl-
	instrumental	N-mpiy	N-mpin	N-mpi-	N-fpi-	N-npi-

5.3 Opis dobivenog rječnika

Dobiveni naglasni rječnik sastoji se od svih oblika imenica, glagola i pridjeva generiranih uvažavajući ranije spomenuta pravila iz (Mikelić Preradović, 2008), svih oblika ostalih vrsta riječi preuzetih iz *Rječnika hrvatskoga jezika* te ručno nadodanih oblika zamjenica i brojeva. Ukupno se u rječniku nalazi 72,366 osnovnih oblika riječi i 1.011,785 izvedenih oblika riječi + izvedeni oblici pridjeva[10]. Veći broj osnovnih oblika riječi nego u samom *Rječniku hrvatskoga jezika* može se objasniti time da su neke riječi u *Rječniku hrvatskoga jezika* bile navedene pod jednom natuknicom u istom retku, a natuknica je sadržavala primjerice imenicu muškoga roda i odgovarajuću imenicu ženskoga roda (primjerice *vjèštāk m (vještàkinja ž)*). Obje takve imenice uzele su se kao osnova iz koje su se izveli svi oblici, pa su se obje imenice ubrojile u osnovni oblik riječi budući da za muški i ženski rod vrijede različita pravila za dodavanje nastavka i promjenu naglaska. Broj osnovnih i izvedenih oblika po vrstama riječi može se vidjeti u tablici 3.

Svaka natuknica u rječniku sastoji se od naglašene riječi, njezine MSD oznake i nenaglašenog oblika riječi. Cjelokupni rječnik nalazi se na CD-u priložen uz doktorski rad (Prilog 6).

[10]točan broj izvedenih oblika pridjeva ne može se odrediti budući da se pri generiranju nekih oblika pridjeva zbog nemogućnosti određivanja prisutstva nepostojanog a, generiraju i oblici koji nisu valjani (problem opisan u poglavlju 5.1.4)

Tabela 3 Broj osnovnih i izvedenih riječi u naglasnom rječniku po vrstama riječi

Broj natuknica u naglasnom rječniku			
	Broj osnovnih oblika	Broj izvedenih oblika	Ukupno
imenice	39.497	636.302	681.799
glagoli	13.220	375.214	388.934
pridjevi	12.846	točan broj ne može se odrediti[11]	točan broj ne može se odrediti
prilozi	5.943	-	5.943
zamjenice	174	230	404
brojevi	244	39	283
uzvici	193	-	193
prijedlozi	111	-	111
veznici	71	-	71
čestice	67	-	67
ukupno	**72.366**	**1.011.785 + izvedeni oblici pridjeva**	**1.077.805+izvedeni oblici pridjeva**

[11]točan broj izvedenih oblika pridjeva ne može se odrediti budući da se pri generiranju nekih oblika pridjeva zbog nemogućnosti određivanja prisutstva nepostojanog a, generiraju i oblici koji nisu valjani (problem opisan u poglavlju 5.1.4)

5.4 Rezultati primjene pravila na testnom tekstu

Provođenje pravila testiralo se na svim primjerima na temelju kojih su nastala pravila. Međutim, htjelo se provjeriti i kakav se rezultat dobije ukoliko se pravila primijene na tekst, a ne samo na zasebne riječi. Za tu svrhu koristio se tekst s označenim naglascima koji je nastao na Odsjeku za kroatistiku pri Sveučilištu u Rijeci, a točnost naglasaka na riječima u rečenici provjerena je od strane eksperta. Tekst se sastoji od ukupno 2.160 riječi. Radi se o transkripciji govora, a naglasci su dodijeljeni prema hrvatskom književnom standardu.

Obzirom da se radi o tekstu, tj. riječi su složene u rečenice, uz pomoć automatskog morfosintaktičkog označivača(Agić, Ljubešić, & Merkler, 2013) dobio se zapis u kojemu uz svaku (nenaglašenu) riječ u tekstu stoji i njezina MSD oznaka. Budući da su se u hrvatski naglasni rječnik koji je gore opisan također dodale MSD oznake, a uz naglašeni oblik riječi postoji i nenaglašeni, provelo se pretraživanje nenaglašenog oblika riječi iz teksta zajedno s MSD oznakom unutar naglasnog rječnika. Nadalje, kada se našla natuknica unutar rječnika i njezina MSD oznaka koja odgovara riječi i njezinoj MSD oznaci iz teksta, onda se nenaglašeni oblik riječi zamijenio ekvivalentnim naglašenim oblikom riječi. Na taj se način kao izlaz dobio niz riječi (zapisane svaka u svoj redak) koji odgovara nizu riječi iz teksta, ali s označenim leksičkim naglaskom. Nakon toga se dobiveni izlaz zapisao u formatu znakovnog niza, a omogućeno je i vraćanje interpunkcijskih znakova te se slova koja su u izvornom tekstu bila zapisana velikim slovima također pretvaraju natrag iz malih u velika. Konačno, krajnji izlaz je tekst ekvivalentan izvornom tekstu, ali s označenim leksičkim naglascima. Navedeni algoritam prikazan je u Prilogu 3.

Budući da je točnost morfosintaktičnog označivača koju navode autori kada se određuje samo osnovna kategorija riječi 97%, a kada se koristi puna oznaka (prema MULTEXT-East standardu) 87%, kada se u rječniku pronašla odgovarajuća riječ krenulo se dalje provjeravati i njenu MSD oznaku, ali postepeno. Najprije se provjerilo odgovara li puna oznaka u rječniku (primjerice N-msny), oznaci iz teksta, a ukoliko ne, onda se jedan atribut izostavio te su se provjeravale jednakosti za preostale atribute (primjerice, N-msn). Granica je postavljena na oznaci za vrstu riječi, tj. POS kategoriji (primjerice, N). Budući da u naglasnom rječniku nisu navedeni svi atributi kategorije (primjerice, u gornjem primjeru "-" u N-msny znači da taj atribut ne postoji), ti su se atributi izostavili pri provjeri. Na opisani se način htjelo dobiti što točniji rezultat usporedbe MSD oznake riječi iz teksta i MSD oznake

riječi iz naglasnog rječnika, ali s time da se dozvoli tolerancija na grešku koja je moguća zbog pogrešne MSD oznake riječi iz teksta koja se dobila automatskim MSD označivačem.

Primjenjujući pravila na tekst na opisani način, od ukupno 2.160 riječi, njih ukupno 1.686 (78%) dobile su pravilni leksički naglasak. Analizom se utvrdilo da neke riječi nisu dobile naglasak jer riječi iz teksta nije bila dodijeljena pravilna MSD oznaka. Nakon što su se sve oznake ručno pregledale i ispravile, rezultat koji je dobiven primjenom pravila je pravilno dodijeljen naglasak na 1.884 riječi (87,7%). Daljnjom analizom utvrđeno je da neke riječi nisu dobile pravilni naglasak zbog premještanja naglaska s naglasnice na prednaglasnicu pa su se stoga u obzir uzela pravila koja definiraju kada naglasak s naglasnice prelazi na prednaglasnicu. U sljedećem odlomku opisana su pravila koja su uzeta u obzir. Nakon primjene pravila, pravilni naglasak dobilo je 2.004 riječi (92,8%). Rezultati naglašavanja pomoću pravila prikazani su u tablici 4 i na slici 17. Većina riječi koja ni tada nije dobila pravilni naglasak, nisu dobila ni pogrešan naglasak nego naglasak nije uopće dodijeljen. Razlog tomu je nepostojanje riječi u rječniku.

Pojedine se riječi nisu našle u rječniku zato što:

- riječi nema u *Rječniku hrvatskoga rječnika* (na temelju kojeg je nastao naglasni rječnik) -primjerice riječ *okej,*
- riječ nije obuhvaćena pravilima za izvođenje izvedenih oblika od osnovnog oblika pa zato u rječniku postoji samo osnovni oblik riječi, ali ne i njezine izvedenice (primjerice *znati*),
- radi se o posebnom tipu riječi koji ima vlastitu sklonidbenu odnosno konjugacijsku paradigmu pa nije obuhvaćen pravilima (primjerice, nepravilni glagoli)
- radi se o imenu ili zemljopisnom imenu čiji osnovni ili izvedeni oblik nije obuhvaćen rječnikom (primjerice *Ljêrkin*).

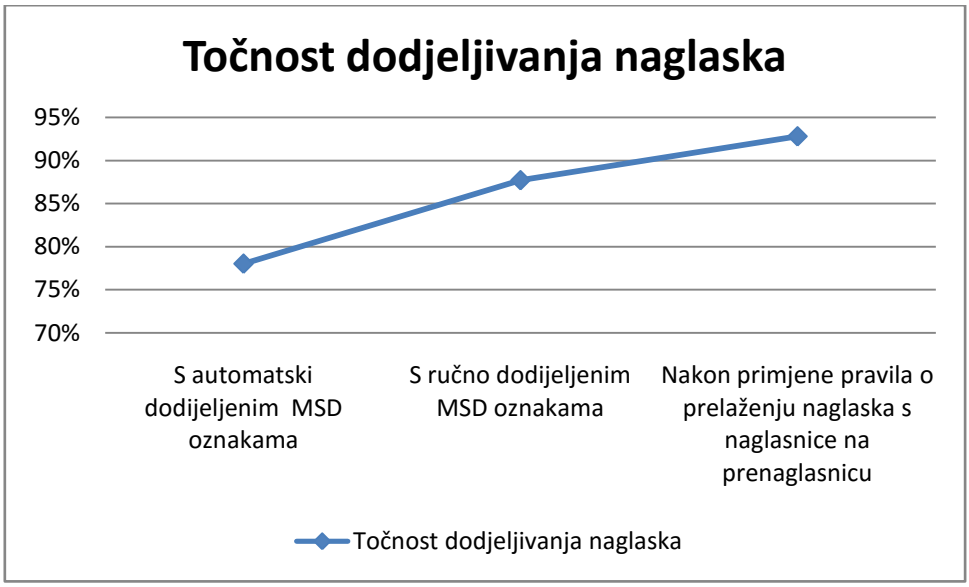
Jasno je stoga da bi se točnost naglašavanja mogla dodatno povećati:

- proširivanjem pravila za izvođenje izvedenih oblika riječi iz osnovne,
- proširivanjem rječnika svim oblicima riječi koje imaju vlastitu paradigmu za sklanjanje odnosno konjugiranje,
- proširivanjem rječnika popisom osnovnih i izvedenih oblika imena i zemljopisnih imena,

- automatskim postupkom dodjeljivanja naglaska na temelju modela koji će biti prikazan u poglavlju 6.

Tabela 4 Točnost automatskog dodjeljivanja naglasaka pomoću pravila

	Točnost dodjeljivanja naglaska
S automatski dodijeljenim MSD oznakama	78%
S ručno dodijeljenim MSD oznakama	87,7%
Nakon primjene pravila o prelaženju naglaska s naglasnice na prenaglasnicu	**92,8%**



Slika 17 Točnost dodjeljivanja naglasaka pomoću pravila

Navedene točnosti predstavljaju postotak riječi koje su automatskim postupkom pravilno naglašene u testnom tekstu, a računaju se prema formuli:

$$\frac{|N_t \cap N_s|}{N_s} * 100$$

gdje presjek $|N_t \cap N_s|$predstavlja broj riječi kojima je pravilno dodijeljen naglasak u odnosu na testni tekst, a N_s ukupni broj riječi u testnom tekstu.

5.4.1 Pravila za prenošenje naglaska s naglasnice na prednaglasnicu

Pravila za prenošenje naglaska s naglasnice s prednaglasnice koja su se uzela u obzir(Barić, i dr., 1995.):

Oslabljeno pomicanje naglaska:

1) prednaglasnica ima kratkouzlazni naglasak umjesto kratkosilaznog i dugosilaznog; naglasak se pomiče s imenica, zamjenica, pridjeva i rednih brojeva u svezi s prijedlozima i veznicima (*i, ni*), a s glagola u vezi s negacijom *ne* (*òd kućē, nì jā*).

Neoslabljeno pomicanje naglaska:

1) s jednosložnih i dvosložnih imenica muškog i ženskog roda koje u N jd. imaju jedan slog s dugouzlaznim naglaskom (primijenjeno samo kad je imenica u nominativu ili akuzativu) (*ü grād*),

2) s 2. i 3. lica jd. aorista na negaciju (*nȅ ukrāde*),

3) s glavnih brojeva *dva, dvije, tri, pet* do *deset, sto, dvoje, troje, oba, obje* (*zà stō gȍdinā*),

4) sa zamjenice i s instrumentalnih oblika zamjenice mnom i tobom (s*à mnōm*),

5) sa zamjenica *mene, tebe, sebe, njega* na prijedlog koji završava suglasnikom kada se umjesto naglašenog oblika zamjenice upotrijebi nenaglašeni: *me, te, se, nj* (*ùzā se*) - prijedlog tada dobiva dugi *ā* na kraju; ako prijedlog završava samoglasnikom, dobiva dugi naglasak (*zá me*).

6. Automatsko dodjeljivanje naglasaka riječima iz teksta

Na temelju rezultata iz prethodnog poglavlja može se zaključiti da se rječnikom ne mogu obuhvatiti sve riječi na koje će se naići u tekstovima. Jedno od rješenja koje se može primijeniti u takvim slučajevima kada se nenaglašena riječ za koju tražimo njezin naglašeni ekvivalent ne nalazi u rječniku je primijeniti neku od statističkih metoda kako bi se našlo najvjerojatnije mjesto i vrsta naglaska u riječi. U tu se svrhu mogu primijeniti primjerice klasifikacijska i regresijska stabla odlučivanja, metoda potpornih vektora (engl. support vector machine - SVM), skriveni Markovljevi modeli, neuronske mreže, naivni Bayesovi klasifikatori i sl. Primjerice u (Taylor, 2009)se u sklopu grafemsko-fonemske pretvorbe za sintezu govora koriste skriveni Markovljevi modeli kako bi se predvidjeli naglašeni slogovi. U (Ciobanu, Dinu, & Dinu, 2014) koriste se SVM za pronalaženje granice među slogovima i predviđanje naglašenih slogova za rumunjski jezik. U (Yarowsky, 1999) koristi se naivni Bayesov klasifikator za predviđanje naglašenih slogova u francuskom i španjolskom. U (Marinčič, Tušar, Gams, & Šef, 2009) koriste se klasifikacijska stabla za određivanje mjesta i vrste naglaska u slovenskom jeziku. U navedenom radu, prvo su izgrađeni model koji za svaki samoglasnik predviđaju je li naglašen ili nije na temelju njegova konteksta (za svaki samoglasnik novi model), a nakon toga se primijenio model koji predviđa vrstu naglaska. Za razliku od hrvatskoga koji ima četiri vrste naglaska, u slovenskom samoglasnik *e* ima tri vrste

naglaska, samoglasnik *o* dvije, a ostali samoglasnici su samo ili naglašeni ili nenaglašeni. Za hrvatski jezik dosad nisu rađena slična istraživanja.

6.1 Klasifikacijska i regresijska stabla

Klasifikacijska i regresijska stabla koriste se za izgradnju stabla odlučivanja za rješavanje problema klasifikacije i regresije. U slučajevima kada se predviđa najvjerojatnija značajka koja pripada određenoj klasi ili kategoriji (kvalitativna varijabla), govorimo o klasifikacijskim stablima, a kada predviđamo vrijednost značajke koja je brojčana ili kontinuirana (kvantitativna varijabla), onda se radi o regresijskim stablima. Za gradnju stabla koristi se skup podataka za učenje $S=\{(z_n,y_n)\ ,\ n\text{-}1,\ldots N\}$ gdje je z_n vektor značajki n-tog uzorka (instance), a y_n zavisna varijabla koja se predviđa. Gradnja stabla počinje od korijenskog čvora kojem su pridruženi svi podaci iz skupa. Sljedeći je korak traženje pitanja (atributa) koje najbolje dijeli podatke u skupu prema određenom kriteriju. Postupak se dalje rekurzivno ponavlja na skupu podataka u svakom čvoru, a postupak se prekida kada podskup određenog čvora ima sve iste vrijednosti izlazne varijable, ili kada daljnje grananje više ne pridonosi poboljšanju rezultata odnosno nije moguće daljnje dijeljenje. Svaki list u stablu predstavlja vrijednost ciljne varijable ako su dane vrijednosti ulaznih varijabli predstavljene putom od korijena stabla do tog lista.

Algoritmi za izgradnju stabala obično vrlo dobro klasificiraju podatke na kojima se uče, međutim, može se dogoditi da neviđene podatke ne klasificiraju s istom točnošću. Tada govorimo o takozvanoj "pretreniranosti" stabala (engl. overfitting) kada model ne odražava stvarne zavisnosti među ulaznim i izlaznim varijablama. Obično se događa kada je model prekompleksan, odnosno kada sadrži previše parametara u odnosu na broj instanci. U postupku gradnje stabla obično se primjenjuje postupak podrezivanja stabla (engl. pruning) kako bi se „pretreniranost“ stabla spriječila. Na taj se način u obzir ne uzimaju grane koje su odgovorne za "pretreniranost". Podrezivanje se obično provodi na način da se stablo evaluira na neviđenim podacima, koji nisu korišteni pri gradnji stabla, i onda se pojednostavni na način da se zanemare dijelovi stabla koji ne klasificiraju dobro te podatke.

6.2 Metodologija

U ovom su istraživanju korištena klasifikacijska stabla za predviđanje mjesta i vrste naglaska u riječi, a kao skup za učenje modela korišten je ranije opisani naglasni rječnik koji sadrži informacije o naglašenom obliku riječi, njezinom nenaglašenom ekvivalentu i MSD oznaku. Algoritam koji se koristio prilikom učenja stabla odlučivanja je J48 koji je implementiran pomoću alata Weka (Witten, Frank, & Hall, 2011).

Slično kao u (Marinčič, Tušar, Gams, & Šef, 2009) najprije se izgradio model za predviđanje mjesta naglaska, a onda model za vrstu naglaska. Za razliku od postupka koji je opisan u navedenom radu gdje se za utvrđivanje mjesta za svaki samoglasnik izgradio model za predviđanje njegova naglaska, ovdje se koristio postupak gradnje modela kojim se predviđa najvjerojatnije mjesto naglaska unutar riječi na način da se tražio najvjerojatniji redni broj sloga unutar riječi koji je naglašen. Nakon toga se u novom modelu našla najvjerojatnija vrsta naglaska (kratkosilazni, kratkouzlazni, dugosilazni, dugouzlazni). Kad se model naučio i kada se evaluirala točnost predviđanja klase za mjesto i vrstu naglaska, primijenio se na neviđeni tekst koji se nije koristio u postupku učenja modela. Radi se o istom tekstu pomoću kojeg su se evaluirali rezultati programa za automatsko naglašavanje pomoću pravila iz poglavlja 5.

Pomoću modela naučenog na cijelom rječniku dobila se predikcija najvjerojatnijeg mjesta naglaska na riječima u testom tekstu. Te su se vjerojatnosti izdvojile i koristile u izgradnji modela na testnom tekstu koji predviđa najvjerojatniju vrstu naglaska. Vjerojatnosti su se u drugi model dodale kao jezične značajke (jedna vjerojatnost za jednu instancu).

6.2.1 Jezične značajke

Kako bi se pomoću modela moglo predvidjeti mjesto i vrsta naglaska u riječi, uz skup podataka na kojem se uči model potrebno je odrediti i jezične značajke koje će se razmatrati u postupku izgradnje stabala. Za učenje prvog modela pomoću kojeg se predviđa mjesto naglaska u riječi koristile su se sljedeće jezične značajke: broj slogova u riječi, fonetske osobine zadnja četiri glasa, fonetske osobine prva tri glasa u riječi i POS riječi. Navedene značajke prikazane su u tablici 5. Za učenje drugog modela pomoću kojeg se predviđa vrsta naglaska korištene su sljedeće jezične značajke: broj slogova u riječi, fonetske osobine zadnja tri glasa u riječi, fonetske osobine prva dva glasa u riječi, POS oznaka riječi, redni broj

naglašenog sloga (dobiven na temelju predviđanja iz prvog modela), identitet naglašenog glasa, fonetske osobine glasa koji se nalazi ispred naglašenog i fonetske osobine glasa koji se nalazi iza naglašenog. Jezične značajke korištene u drugom modelu, prikazane su u tablici 6. Dakle, za svaku riječ kreirao se skup oznaka koje predstavljaju skup jezičnih značajki.

Tablica 5 Jezične značajke uzete u obzir prilikom gradnje modela za predviđanje mjesta naglaska u riječi

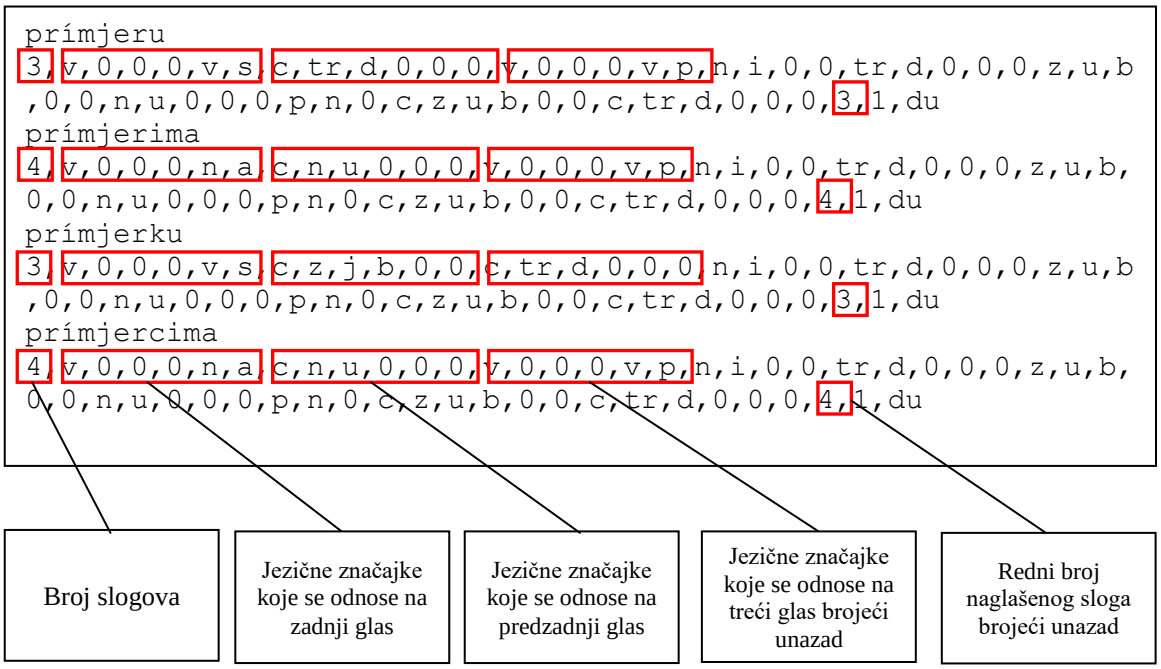
Jezična značajka	Opis značajke	Moguće vrijednosti	Broj atributa
Broj slogova	Broj slogova u riječi	oznake vrijednosti od 1 do 12	1
Zadnji glas u riječi	Fonetske odlike zadnjeg glasa u riječi	prema tablici 7	6
Predzadnji glas u riječi	Fonetske odlike predzadnjeg glasa u riječi	prema tablici7	6
Treći glas od kraja u riječi	Fonetske odlike trećeg glasa od kraja u riječi	prema tablici7	6
Četvrti glas od kraja u riječi	Fonetske odlike četvrtog glasa od kraja u riječi	prema tablici 7	6
Prvi glas u riječi	Fonetske odlike prvog glasa u riječi	prema tablici 7	6
Drugi glas u riječi	Fonetske odlike drugog glasa u riječi	prema tablici 7	6
Treći glas u riječi	Fonetske odlike trećeg glasa u riječi	prema tablici 7	6
POS oznaka	POS oznaka riječi	N, V, A, P, R, S, C, M, Q, I	1

Tabela 6 Jezične značajke uzete u obzir prilikom gradnje modela za predviđanje vrste naglaska u riječi

Jezična značajka	Opis značajke	Moguće vrijednosti	Broj atributa
Broj slogova	Broj slogova u riječi	oznake vrijednosti od 1 do 12	1
Zadnji glas u riječi	Fonetske odlike zadnjeg glasa u riječi	prema tablici 7	6
Predzadnji glas u riječi	Fonetske odlike predzadnjeg glasa u riječi	prema tablici7	6
Treći glas od kraja u riječi	Fonetske odlike trećeg glasa od kraja u riječi	prema tablici7	6
Prvi glas u riječi	Fonetske odlike prvog glasa u riječi	prema tablici 7	6
Drugi glas u riječi	Fonetske odlike drugog glasa u riječi	prema tablici 7	6
POS oznaka	POS oznaka riječi	N, V, A, P, R, S, C, M, Q, I	1
Redni broj naglašenog sloga	Redni broj naglašenog sloga dobio se kao rezultat predviđanja prvog modela	oznake vrijednosti od 1 do 12	1
Redni broj naglašenog sloga brojeći od kraja	Redni broj naglašenog sloga brojeći od kraja	oznake vrijednosti od 1 do 12	1
Identitet naglašenog glasa	Identitet naglašenog glasa	a, e, i, o, u, slogotvorno r	1
Glas ispred naglašenog	Fonetske odlike glasa koji se nalazi ispred naglašenog glasa	prema tablici 7	5
Glas iza naglašenog	Fonetske odlike glasa koji se nalazi iza naglašenog glasa	prema tablici 7	5

Tabela 7 Fonetske značajke glasova

Fonetska značajka		oznaka
Samoglasnik ili suglasnik	samoglasnik	v
	suglasnik	c
Vrsta suglasnika	približnici	p
	treptajnik	tr
	bočnici	b
	nosnici	n
	zapornici	z
	tjesnačnici	tj
	slivenici	s
Vrsta suglasnika po mjestu tvorbe	usnenici	u
	zubnousnenici	z
	desnici	d
	prednepčanici	p
	nepčanici	n
	jedrenici	j
Vrsta suglasnika prema zvučnosti	zvučni	z
	bezvučni	b
Visina samoglasnika	visoki	v
	srednji	s
	niski	n
Samoglasnici prema mjestu tvorbe	prednji	p
	srednji	s
	stražnji	st



Slika 18 Prikaz oznaka za jezične značajke pojedinih riječi

Primjer skupa oznaka za pojedine riječi za model određivanja vrste naglaska prikazan je na slici 18. Iz primjera možemo vidjeti da se navedene riječi razlikuju po sljedećim značajkama: broju slogova, značajkama koje se odnose na fonetske odlike zadnja tri glasa u riječi te rednom broju naglašenog sloga brojeći unazad. Ostale su značajke jednake kod svih riječi u danom primjeru.

6.3 Rezultati automatskog dodjeljivanja naglasaka

Nakon učenja modela za predviđanje najvjerojatnijeg mjesta naglaska u riječi na podacima iz rječnika opisanih u poglavlju 5, postupkom deseterostruke unakrsne validacije (engl. 10-fold cross validation) točnost modela koja se dobila je 90,56%. Za model predviđanja najvjerojatnije vrste naglaska na riječima istim postupkom se dobila točnost od 86,02%.

Nakon što su se riječi iz testnog teksta prikazale pomoću oznaka gore navedenih jezičnih značajki, model za predviđanje najvjerojatnijeg mjesta naglaska primijenio se kako bi se utvrdio redni broj sloga koji je najvjerojatnije naglašen u riječi. Dobivena predviđanja pridružila su se kao jezična značajka svakoj riječi iz testnog jezika i na taj način uključila u testiranje drugog modela koji je odredio predikcije najvjerojatnije vrste naglaska. Dobivene vrijednosti predikcija usporedile su se sa stvarnim vrijednostima kako bi se dobile točnosti primjene modela na testnom tekstu.

Budući da je testni tekst transkripcija govora, tekst sadrži mnogo kratkih riječi, prednaglasnica i zanaglasnica koje nemaju naglasak. Za računanje točnosti modela, one se nisu uzimale u obzir.

Za preostale 1422 riječi model je točno predvidio mjesto naglaska za 1365 riječi (95,99%), vrstu naglaska za 1056 riječi (74,26%), a i mjesto i vrstu naglaska za 1026 riječi (72,15%). Pretpostavlja se da je točnost modela na testnom skupu podataka bolja za predviđanje mjesta naglaska, a lošija za predviđanje vrste naglaska zato što je većina riječi u testnom skupu podataka jednosložna i dvosložna. Na takvim je riječima vjerojatnost utvrđivanja točnog mjesta veća u odnosu na višesložne riječi (jer se većina naglasaka u hrvatskom nalazi na 1. slogu), ali je vjerojatnost utvrđivanja vrste naglaska manja, budući da na riječima koji imaju naglašen prvi slog mogu stajati svi naglasci, a na ostalim slogovima samo uzlazni (prema pravilu da se silazni naglasci nalaze samo na 1. slogu).

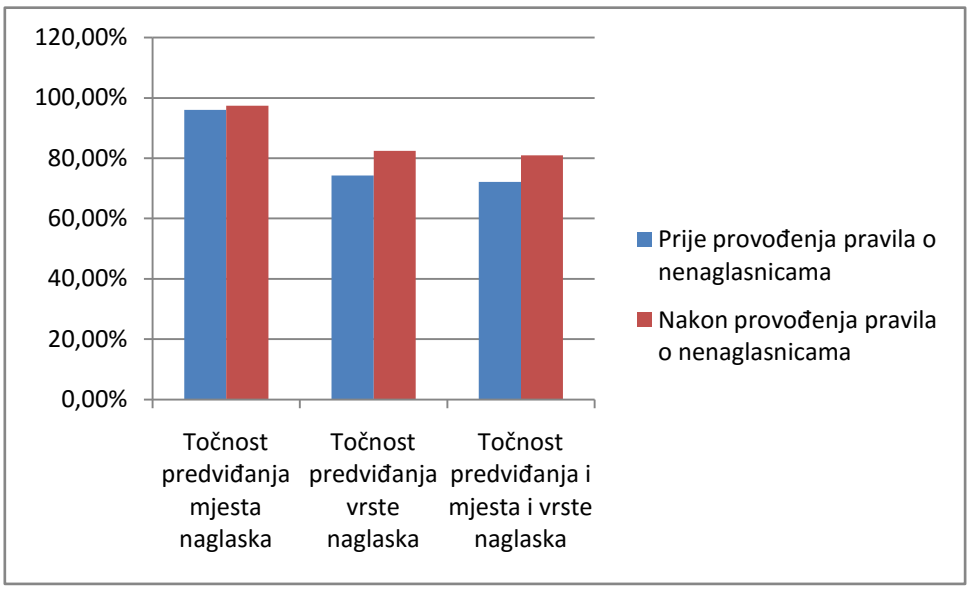
Ako se na prednaglasice i zanaglasnice primijene pravila o njihovom nenaglašavanju, odnosno prelazak naglaska s naglasnice na prednaglasnicu, onda se dobivaju sljedeće točnosti: za mjesto naglaska 97,4%, za vrstu naglaska 82,4% , za točno i mjesto i vrstu 80,1%. Rezultati su prikazani u tablici 8, tablici 9 te slici 19.

Tabela 8 Točnost modela za predviđanja mjesta i vrste naglaska dobivena postupkom deseterostuke unakrsne validacije

Model za predviđanje mjesta naglaska	Model za predviđanje vrste naglaska
90,56%	86,02%

Tabela 9 Točnost predviđanje mjesta i vrste naglaska na testnom skupu

	Točnost predviđanja mjesta naglaska	Točnost predviđanja vrste naglaska	Točnost predviđanja i mjesta i vrste naglaska
Prije provođenja pravila o nenaglasnicama	95,99%	74,26%	72,15%
Nakon provođenja pravila o nenaglasnicama	97,4%	82,4%	**80,97%**



Slika 19 Točnost predviđanje mjesta i vrste naglaska pomoću modela na testnom skupu

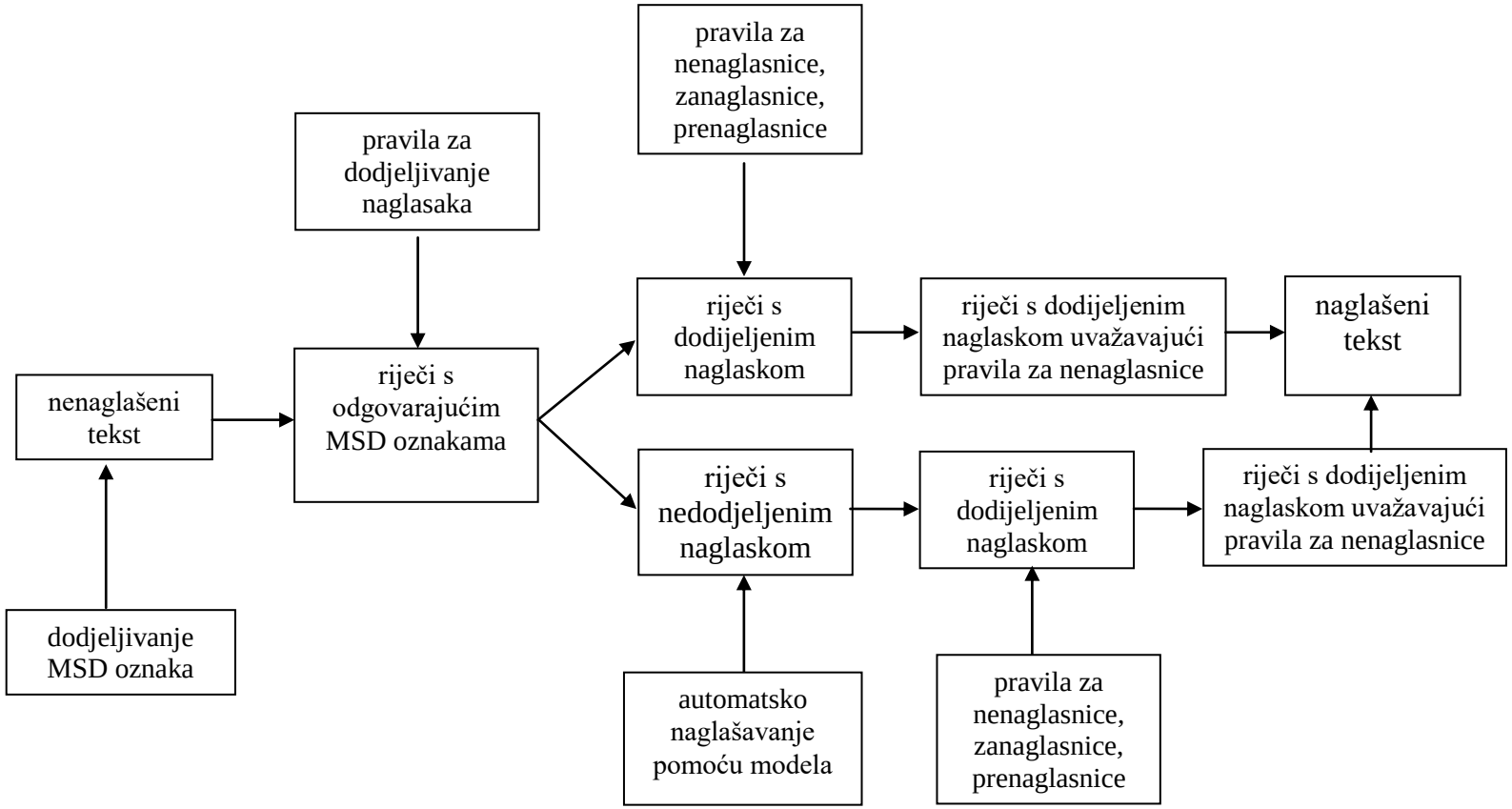
6.4 Hibridni pristup automatskog dodjeljivanja naglaska pomoću pravila i modela za naglašavanje

Kako bi se dobila što veća točnost dodijeljenih naglasaka u tekstu, predlaže se hibridni pristup dodjeljivanja naglasaka koji kombinira dodjeljivanje naglasaka koristeći pravila opisanih u 5. poglavlju i naglašavanja korištenjem modela ranije opisanog u ovom poglavlju.

Obzirom da se nekim riječima prilikom dodjeljivanja naglaska pomoću pravila ne dodijeli naglasak zato što riječ ne postoji u rječniku iz razloga koji su ranije navedeni (riječi nema u *Rječniku hrvatskoga rječnika*, riječ nije obuhvaćena pravilima za izvođenje izvedenih oblika od osnovnog oblika, radi se o posebnom tipu riječi koji ima vlastitu sklonidbenu odnosno konjugacijsku paradigmu, radi se o imenu ili zemljopisnom imenu), predlaže se da se takve riječi identificiraju i da se na njih zatim primjeni automatsko dodjeljivanje naglaska pomoću modela. Iako je glavni izvor za učenje modela rječnik u kojem nema tražene riječi, ipak će se takvoj riječi dodijeliti naglasak jer je model učen na temelju jezičnih značajki ostalih riječi u rječniku pa će se na temelju riječi koje imaju iste jezične značajke kao što ima i riječ kojoj se dodjeljuje naglasak, dodijeliti naglasak i toj riječi.

Ukoliko se opisani postupak primijeni na testni korpus koji je ranije opisan i korišten za evaluaciju dodjeljivanja naglaska u 5. i 6. poglavlju, ukupna točnost dodjele naglaska povisuje se s 92,8% na **95,3%**.

Dakle, u navedenom su se postupku najprije riječi provele kroz pravila opisana u poglavlju 5 kako bi se pomoću njih dodijelio pripadajući naglasak. Svaka je riječ imala i odgovarajuću MSD oznaku. Zatim su se na riječi s dodijeljenim naglaskom primijenila pravila o nenaglašavanju prednaglasnica i zanaglasnica, odnosno prelazak naglaska s naglasnice na prednaglasnicu. Na riječi kojima se nije dodijelio naglasak primijenio se postupak automatskog dodjeljivanja naglaska pomoću modela koji je opisan ranije u ovom poglavlju. Nakon toga su se opet primijenila pravila o nenaglašavanju prednaglasnica i zanaglasnica, odnosno prelazak naglaska s naglasnice na prednaglasnicu. Shematski prikaz predloženog pristupa dan je na slici 20.



Slika 20 pristup automatskog dodjeljivanja naglaska pomoću pravila i modela za naglašavanje

7. Analiza trajanja i model trajanja slogova

Već je ranije napomenuto kako su trajanje i F0 kontura dva najvažnija prozodijska sredstva koja se obično uzimaju u obzir prilikom modeliranja prozodije. Krajnji je cilj pomoću modela što točnije predvidjeti trajanje i vrijednosti F0 konture pojedinih jedinica govora samo na temelju ekvivalentnoga teksta. Modeli se, dakle, uče na podacima koji se sastoje od snimaka govora, odgovarajućih transkripata te označenih različitih jezičnih značajki, a onda se tako naučen model primjenjuje za predviđanje trajanja i vrijednosti F0 konture na novi tekst. Prozodijski bi model trebao predstavljati svojstva prirodnog govora - svojstva trajanja i intonacije.

U ovom je poglavlju opisana analiza trajanja hrvatskih slogova pri čemu su u obzir uzete fonološke, položajne i kontekstualne značajke. Jezgrom sloga smatra se samoglasnik, a na trajanje sloga može utjecati položaj samoglasnika unutar sloga te vrsta suglasnika prisutnih s jedne i/ili druge strane samoglasnika u slogu, kao što je primjerice opisano za engleski jezik u (Flege & Brown, 1982). Na trajanje sloga može utjecati i naglasak. U (Pletikos, 2003) izmjereno je i izračunato prosječno trajanje četiri vrste hrvatskih naglašenih slogova. Trajanje je mjereno na snimljenom govoru triju ispitanika koji su izgovarali zadane riječi (nepovezan

govor). Izmjereno je da glas s kratkosilaznim naglaskom u tim riječima traje prosječno 115ms, a u odnosu na njega, glas s kratkouzlaznim traje 109%, s dugosilaznim 234%, a s dugouzlaznim 243%. Položajne značajke također mogu utjecati na trajanje slogova u riječi. Primjerice, slog na početku riječi ima tendenciju trajati dulje nego kad se nađe na drugim mjestima u riječi (Yang, 1998). Slogovi koji se nalaze ispred ili iza promatranog sloga također mogu utjecati na njegovo trajanje. Primjerice u (Rao, 2012) je za indijski jezik telugu utvrđeno da se trajanje sloga produlji za 25% do 35% ukoliko iza njega slijedi slog koji počinje nazalnim glasom.

7.1 Rastavljanje riječi na slogove

Riječ se može raščlaniti na morfeme i slogove pri čemu morfemi predstavljaju značenjske jedinice, a slogovi izgovorne. "Slog je najmanja i temeljna jedinica izgovora." (Barić, i dr., 1995.)U slogu je jedan glas nositelj sloga, tj. vrhunac sloga, a naziva se još i slogotvorni glas. Ostali su glasovi neslogotvorni. Samoglasnici su slogotvorni glasovi, dok su svi suglasnici neslogotvorni. U hrvatskom jeziku, uz samoglasnike, slogotvorni glasovi još mogu biti i suglasnici *[r̩], [l̩]* i *[n̩]* te nefonemski, neutralni samoglasnik *[ə]*.

Slogotvorno *[r̩]* može se pojaviti na početku riječi ispred suglasnika (npr. u *r̩zati*), između dva suglasnika od kojih prvi nije *j, r, l, lj, n, nj, ć, dž, đ* (npr. u *vr̩t*), iza suglasnika, a ispred *o* koje je nastalo zamjenjivanjem *l* (npr. u *gr̩oce*) te u riječima stranog podrijetla, na kraju, iza suglasnika (npr. u *žanr̩*). Slogotvorni *[l̩]* i *[n̩]* pojavljuju se u sredini, između dva suglasnika ili na kraju, iza suglasnika u riječima stranog podrijetla (npr. u *bicikl̩, njutn̩*). Neutralni samoglasnik *[ə]* obično se izgovara uz neslogotvorne glasove, npr. pri nabrajanju glasova *bə, cə, čə.*

Ovisno o rasporedu glasova u slogu, slogovi mogu biti otvoreni i zatvoreni. Otvoreni slogovi završavaju samoglasnikom (V), a zatvoreni suglasnikom (C). Prema (Škarić, 1991), najzastupljeniji tip sloga u hrvatskom jeziku je CV (60%). Uz navedeni tip, u hrvatskom postoje i ostali tipovi poput V, VC, CVC, CCV, CCCV itd. U ovom je istraživanju korišteno slogovanje prema načelu najvećega pristupa opisanog u (Meštrović, Martinčić-Ipšić, & Matešić, 2015)[12].

Pravila koja su uvažena prilikom rastavljanja na slogove prema navedenom su sljedeća:

P1. jedno mjesto u slogu obavezno zauzima jedan od samoglasnika (silabema);

P2. prije silabema moguć je slijed od najviše 4 suglasnička fonema;

P3. nakon silabema moguć je slijed od najviše 3 suglasnička fonema;

P4. jedan suglasnik pred samoglasnikom pripada uvijek prvom slogu;

P5. suglasnički slijed kojim riječ može početi može stajati i na početku sloga;

[12]Alat za automatsko rastavljenje na slogove dostupan je na http://langnet.uniri.hr/resources.html

P6. medijalni elementi kojima riječ ne može početi dijele se u dva sloga tako da se provjerava može li riječ započeti slijedom koji je početak sloga ili slijedom koji smo dodatno uvrstili kao dopušten slijed;

P7. načelo najvećeg pristupa: ako se primjenom pravila P6 dogodi da dolazi u obzir više mogućnosti, odabire se ono rastavljanje koje će rezultirati najvećim pristupom;

P8. ako se fonem /r/ nalazi između dvaju suglasnika, proglašava se slogotvornim /r̥/, osim ako iza njega slijedi /j/ (npr. u *vrtovi* je /r̥/, a u *vrjednovati* je /r/);

P9. ako se fonem /r/ nalazi na početku riječi, a nakon njega slijedi suglasnik, proglašava se slogotvornim /r̥/, osim ako iza njega slijedi /j/ (npr. u *rzati* je /r̥/, a u *rješenje* je /r/);

P10. ako se fonem /r/ nalazi na kraju riječi iza suglasnika i ako takav slijed odgovara slijedu potvrđenome na početku riječi, proglašava se slogotvornim /r̥/;

P11. ako je slijed *-ije-* od *jata* (osim u *dvije* i *prije*), čitav pripada istome slogu;

P12. slijed koji se bilježi kao *-naest-* (*dvanaest*, *dvanaestica* i sl.) ne rastavlja se na dva sloga;

P13. slijed *-nj-* u nekim se iznimkama tretira kao dvofonemski, a ne kao /*ń*/;

P14. slijed *-dž-* u nekim se slučajevima tretira kao dvofonemski, a ne kao /*ǯ*/.

7.2 Korpus

Govorni korpus koji je korišten za analizu trajanja slogova dio je korpusa VEPRAD (Martinčić-Ipšić & Ipšić, 2003). Korpus VEPRAD uključuje govorne signale vezane uz domenu vijesti, vremenskih prognoza, izvještaja s radijskih emisija, priča i bajki te spontanog govora. Vremenske prognoze čitane su od strane profesionalnih govornika s vijesti s radija, dnevne novosti i priče i bajke također čitaju profesionalni govornici, a spontani je govor izgovaran od strane neprofesionalnih govornika.

Ukupno korpus sadrži oko 15 sati transkribiranog govora izgovaranog u studijskom okruženju i 1 sat spontanog govora. Svaka rečenica/iskaz ima odgovarajuću tekstualnu transkripciju na razini riječi. Korpus ukupno sadrži oko 208.000 riječi, 15.000 jedinstvenih riječi i gotovo 60 muških i ženskih govornika. U tekstovima u bazi najzastupljenijeg govornika, sm04, postoji 21.282 riječi od čega je 5.959 različnica[13] (28%), dok je u materijalima svih govornika 195.594 riječi, od čega je 11.185 ili 5,72% različnica. U tablici 10 prikazana je detaljna statistika korpusa prema (Martinčić-Ipšić, Pobar, & Ipšić, 2011).

Tabela 10 Statistika govornog korpusa VEPRAD

	Broj		Govornici		Riječi		Trajanje
	Snimaka	Iskaza	M	Ž	svih	jedinstvenih	min
Radijske vremenske prognoze	1057	5456	11	14	77322	1462	482
Radijske vijesti	237	3975	1	2	105678	9923	294
Priče	10	2066	1		14689	4160	110
Spontani dijalozi	34	1530	17	17	6664	78	56
Ukupno	1338	13493	28	31	208648	14551	953

Dio korpusa u domeni vremenskih prognoza i vijesti snimljen s nacionalnih radijskih programa sadrži govor 11 muških i 14 ženskih profesionalnih govornika s oko 9.500 rečenica/iskaza i trajanjem od oko 13 sati. U transkripcijama rečenica nalazi se oko 183.000 riječi, od toga 10.227 jedinstvenih. U korpusu vremenskih prognoza nalazi se 1462 jedinstvene riječi što pokazuje da je navedeni korpus pripada uskoj tematskoj domeni.

[13] *različnica (type) je jedinstveni oblik pojavnice iz korpusa.u literaturi*

Spontani govor također je vezan uz tematsku domenu vezanu uz vremenske prognoze, a snimljeni govor izgovaralo je ukupno 17 muških i 17 ženskih govornika, uglavnom studenata Odjela za informatiku, Sveučilišta u Rijeci.

Dio korpusa s tematikom priča i bajki sastoji se od ukupno 10 čitanih priča od strane jednog govornika. Detaljni opis navedenog podkorpusa dan je u poglavlju 7.2.4.

7.2.1 Transkripcija riječi

Transkripcija govora provedena je na razini riječi uz neke dodatne oznake kao što su primjerice *tišina*, *uzdah*, *papir* i slično. Takve su se oznake radi lakšeg razlikovanja pisale između znakova < i > pa se primjerice za tišinu koristila oznaka <sil>, za uzdah <uzdah> i slično.

Prilikom provedbe transkripcije, mogli su se pratiti govorni signali i njihovi spektrogrami, kontura energije, F0 kontura te ponovo preslušavati signal ukoliko je bilo potrebno.

Nakon što su se riječi transkribirale i dodale posebne oznake, takva se datoteka spremila pod istim imenom kao i odgovarajući govorni signal, naravno, s drugom ekstenzijom. Dakle za svaki govorni iskaz postoji odgovarajuća datoteka s istim imenom koja sadrži transkripciju govornog signala. Imena datoteka sastoje se od jednog znaka i 11 brojeva. Prvi znak predstavlja spol govornika, a sljedeća dva broja predstavljaju znakove za jedinstveno označavanje različitih govornika. Sljedeća grupa od 6 brojeva predstavljaju dan, mjesec i godinu snimanja. Nakon toga slijedi serijski broj snimke unutar jednog dana, a zadnja dva znaka predstavljaju broj iskaza.

7.2.2 Fonetski rječnik

Fonetski rječnik govorne baze VEPRAD sastoji se od zapisa riječi u fonetskom obliku koji se temelji na SAMPA simbolima (Bakran & Horga, 1996). Rječnik sadrži skup od 30 standardnih te jednog dodanog fonema - onog za slogotvorno *r*. Rječnik se sastoji od svih riječi koje se pojavljuju u snimkama govora s njihovim fonetskim transkripcijama.

Iako fonetski slijed riječi u hrvatskom jeziku u velikoj mjeri slijedi grafemski, ipak, postoje iznimke. Primjerice grafemski sljedovi *ds, dš,* u govoru poprimaju fonetski oblik *c* odnosno *č* (primjerice *gradski/gracki*). Uz to, u ovisnosti od okruženja, pojedini fonemi mogu

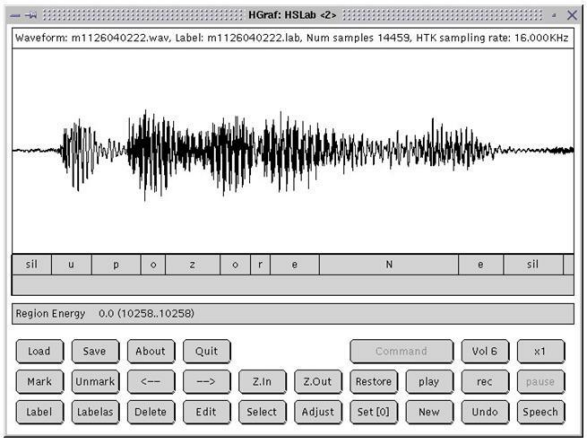
se izgovarati na različiti način. U tom slučaju govorimo o alofonima. Primjerice *n* u riječi *rastanak* i u riječi *banka* izgovaraju se drukčije iako imaju isti grafemski oblik. Pravila za pretvorbu grafemskog zapisa u fonemski poštujući ovakve i slične iznimke za hrvatski jezik izdvojena su u (Načinović, Pobar, Martinčić-Ipšić, & Ipšić, 2009)(Načinović, 2008). Neka od navedenih pravila koja su se mogla primijeniti obzirom na broj glasova i alofona u bazi, primijenila su se prilikom fonetske transkripcije u bazi VEPRAD.

U tekstu se također mogu pojaviti grafemi poput brojeva, skraćenica, datuma i slično koje se onda treba pretvoriti u fonemski zapis. Pravila za normalizaciju takvih slučajeva za hrvatski jezik izdvojena su u (Beliga & Martinčić-Ipšić, 2011) te su također primijenjena na fonetski rječnik.

U dijelu rječnika su također uvedeni različiti simboli za naglašene i nenaglašene samoglasnike, tj. naglašeni samoglasnici su označeni na način da nakon simbola slijedi znak dvotočka.

7.2.3 Segmentacija riječi na manje jedinice

Transkripcije govornih signala na razini riječi, dalje su se segmentirale na manje jedinice, tj. na slogove. Segmentacija se je provela koristeći automatsko poravnanje govornog signala i transkripcija riječi pomoću skrivenih Markovljevih modela. Automatska segmentacija provela se koristeći HTK alat (Young, i dr., 2006) (Slika 21).



Slika 21 Rezultat automatske segmentacije riječi u HTK alatu[14]

[14]Preuzeto iz: (Martinčić-Ipšić, Matešić, & Ipšić, Korpus hrvatskoga govora, 2004)

7.2.4 Korpus priča i bajki

U ovom se istraživanju koristio podskup korpusa kojeg izgovara govornik sm04 iz domene bajki i priča. Domena bajki i priča izabrala se iz razloga što je prozodija u takvom govoru ekspresivnija nego primjerice kod izgovora vijesti ili vremenskih prognoza(Theune, Meijs, Heylen, & Ordelman, 2006). Korpus bajki i priča ukupno sadrži 14.689 riječi, od čega jedinstvenih 4.160. U korpusu ima ukupno 28.325 slogova (dobivenih postupkom opisanim u poglavlju 7.1), od čega je jedinstvenih 1101. Ako se u obzir uzmu različite vrste naglasaka pa se primjerice slog *nà* razlikuje od sloga *nâ*, onda ima ukupno 2099 jedinstvenih slogova. Korpus se sastoji od ukupno 81.453 znakova. Statistika korpusa prikazana je u tabeli 11.

Tabela 11 Statistika korpusa bajki i priča

Broj riječi	14.689
Broj jedinstvenih riječi	4.160
Broj slogova	28.325
Broj jedinstvenih slogova	1.101
Broj jedinstvenih slogova (uzimajući u obzir vrste naglasaka)	2.099
Broj znakova	81.453

U tabeli 12 mogu se vidjeti najčešće riječi koje se pojavljuju u korpusu bajki i priča, u tabeli 13 najčešći slogovi ako se u obzir ne uzmu različite vrste naglasaka, a u tabeli 14 najčešći slogovi ako se slogovi s različitim vrstama naglaska istog glasa uzmu kao različiti slogovi.

Tabela 12 Najčešće riječi u korpusu bajki i priča

	Riječ	Broj pojava		Riječ	Broj pojava
1	i	633	41	još	32
2	je	414	42	sam	31
3	se	353	43	ih	31
4	u	317	44	snjeguljica	30
5	da	264	45	si	30
6	na	236	46	pred	30
7	a	151	47	vrata	29
8	što	142	48	samo	29
9	kad	130	49	kralj	29
10	te	125	50	sada	26
11	ne	106	51	po	26
12	tako	91	52	kraljevna	26
13	mu	89	53	ribar	25
14	nije	84	54	opet	25
15	pa	81	55	odmah	25
16	ga	80	56	ću	24
17	to	77	57	svoje	24
18	sve	73	58	oko	24
19	za	72	59	krojačić	24
20	ali	72	60	kraljica	24
21	bi	70	61	koji	24
22	od	69	62	već	23
23	joj	69	63	sedam	23
24	s	66	64	ništa	23
25	kako	65	65	više	22
26	ona	64	66	onda	22
27	će	56	67	me	22
28	ti	54	68	li	22
29	su	52	69	ono	21
30	on	50	70	bila	21
31	kao	48	71	ondje	20
32	mi	47	72	njih	20
33	iz	47	73	lijepo	20
34	nego	46	74	ju	20
35	reče	40	75	muž	19
36	ni	40	76	im	19
37	žena	39	77	do	19
38	sa	39	78	upita	18
39	dok	38	79	tu	18
40	bijaše	38	80	pod	18

Tabela 13 Najčešći slogovi u korpusu bajki i priča

	Slog	Broj pojava		Slog	Broj pojava
1	o	1115	21	mo	256
2	i	921	22	a	255
3	je	875	23	ri	247
4	u	698	24	do	245
5	na	675	25	ma	244
6	da	544	26	ga	230
7	po	494	27	pa	221
8	ko	458	28	te	217
9	se	457	29	go	217
10	ti	442	30	ra	211
11	ka	413	31	di	211
12	ni	408	32	še	203
13	la	372	33	ca	203
14	ne	348	34	nu	193
15	za	344	35	to	187
16	li	320	36	lje	185
17	vi	290	37	de	181
18	ta	288	38	me	179
19	bi	280	39	što	177
20	ja	259	40	va	176

Tabela 14 Najčešći slogovi u korpusu bajki i priča ako se u obzir uzmu različite vrste naglasaka

	Slog	Broj pojava		Slog	Broj pojava
1	je	678	21	di	183
2	í	637	22	vi	182
3	o	629	23	mo	181
4	na	539	24	ja	179
5	da	492	25	za	174
6	ù	404	26	pa	166
7	se	390	27	lje	164
8	ti	377	28	de	161
9	la	354	29	á	151
10	ko	313	30	ta	151
11	ni	302	31	ra	147
12	ò	280	32	štȍ	143
13	li	247	33	ù	143
14	ne	236	34	pȍ	139
15	ka	205	35	go	139
16	še	197	36	ga	138
17	ri	194	37	lo	135
18	nu	193	38	bi	134
19	ca	193	39	va	131
20	ma	191	40	kȁd	130

7.3 Priprema podataka

Obzirom da se uz analizu trajanja glasova, provela i analiza trajanja slogova, prvo su se datoteke s transkripcijama govornog korpusa trebale provesti kroz algoritam za rastavljanje riječi na slogove premapostupku opisanom u (Meštrović, Martinčić-Ipšić, & Matešić, 2015), a datoteke s odgovarajućim tekstom rastavljenim na slogove spremljene su pod drugim imenom. Rezultat je prikazan u prvom stupcu na slici 15.

Budući da se htio provjeriti i utjecaj vrste naglaska na trajanje, datoteke s tekstualnim transkripcijama govornog korpusa provedene su i kroz postupak automatskog dodjeljivanja naglaska opisanog u poglavlju 5. Sve su se datoteke s naglašenom transkripcijom također spremile pod odgovarajućim nazivom. Rezultat je prikazan u drugom stupcu na slici 15.

Kako bi se dobile transkripcije govornog korpusa koje su rastavljene na slogove i kojima je dodijeljen odgovarajući naglasak, za svaki par datoteka, uparile su se i odgovarajuće linije. Svaka linija sastoji se od jedne riječi kojoj je dodijeljen naglasak odnosno jedne odgovarajuće riječi koja je rastavljena na slogove. Rezultat je prikazan u trećem stupcu na slici 15.

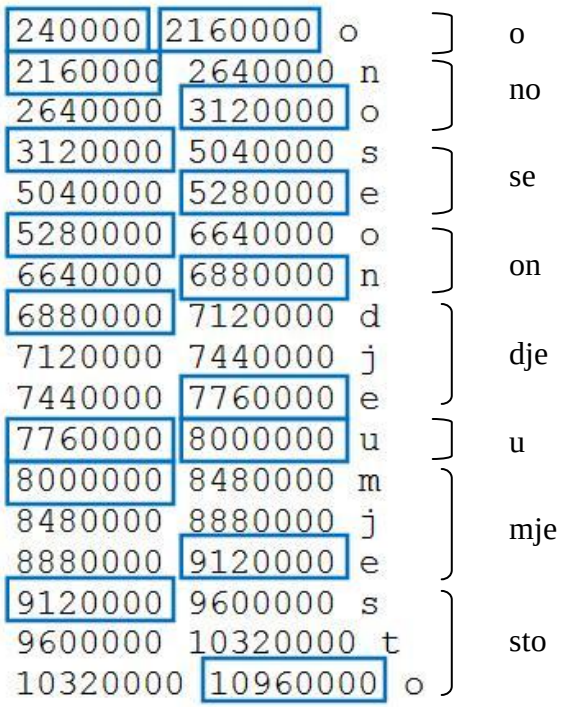
Nakon toga su se algoritmom koji je prikazan u prilogu 4 dodijelili naglasci riječima rastavljenim na slogove. Rezultat je prikazan u četvrtom stupcu u tablici 15.

Tabela 15 Postupak dobivanja naglašenih transkripcija govornog korpusa rastavljenih na slogove

Rezultat nakon rastavljanja na slogove	Rezultat nakon automatskog naglašavanja	Rezultat nakon uparivanja sadržaja odgovarajućih linija odgovarajućih datoteka	Rezultat nakon provođenja algoritma opisanog u Prilogu 4
o=no se on=dje u=mje=sto ža=be stvo=ri kra=lje=vić	òno se óndje ùmjesto žȁbē stvòrī králjević	òno o=no se se óndje on=dje ùmjesto u=mje=sto žȁbē ža=be stvòrī stvo=ri králjević kra=lje=vić	ò=no se ón=dje ù=mje=sto žȁ=be stvò=ri krá=lje=vić

Kao što je ranije rečeno, u bazi VEPRAD postoji segmentacija na razini glasova u obliku .lab datoteka s oznakama izgovorenih glasova i vremenskim trenucima početka i kraja svakog glasa izraženih u milisekundama. Osim trajanja glasova, na temelju tih podataka dobiveno je i trajanje slogova te izrađena analiza trajanja najfrekventnijih slogova ovisno o

njihovim fonološkim značajkama, položaju unutar riječi i rečenice i kontekstu, kako bi se utvrdile najznačajnije značajke koje utječu na trajanje slogova. Na slici 22 prikazan je primjer sadržaja .lab datoteke. Uokvireno se mogu vidjeti počeci i kraj trajanja slogova koji su se dobili na temelju početka i kraja trajanja odgovarajućih glasova u slogu. Na temelju početka i kraja, dobila su se trajanja pojedinih slogova.



Slika 22 Primjer sadržaja .lab datoteke

Kako bi se dobilo trajanja glasova i slogova u naglašenim riječima rastavljenim na slogove, najprije su se morale upariti datoteke u formatu prikazanim u tablici 15 u četvrtom stupcu s odgovarajućom .lab datotekom. Obzirom da su .lab datoteke zapisane tako da je u svakom retku zapisan jedan glas i njegov početak i kraj (u milisekundama), datoteke s naglašenim riječima rastavljenim na slogove također su se morale preoblikovati na način da se svaki glas zapiše u svoj red kako bi se sadržaji mogli upariti s .lab datotekom. Budući da su se htjeli istražiti i položajni i kontekstualni utjecaji na trajanje slogova, u ovom su se koraku dodale oznake *(* za početak rečenice, *)* za kraj rečenice, *P* za početak riječi, *K* za kraj riječi i < za početak sloga, te > za kraj sloga. Rezultat nakon ovog koraka prikazan je na slici 23.

```
(Pò>    240000  2160000 o
<n      2160000 2640000 n
oK      2640000 3120000 o
Ps      3120000 5040000 s
eK      5040000 5280000 e
Pó      5280000 6640000 o
n>      6640000 6880000 n
<d      6880000 7120000 d
j       7120000 7440000 j
eK      7440000 7760000 e
Pù>     7760000 8000000 u
```

Slika 23 Upareni reci lab datoteke i glasovi u riječima s oznakama za početak i kraj rečenice, riječi i sloga

Kad su se uparile .lab datoteke i datoteke s riječima gdje je svaki glas zapisan u svoj red s oznakama za početak i kraj rečenice, riječi i sloga, uzeli su se počeci i krajevi slogova i njihova trajanja kako bi se izračunalo trajanja slogova. Krajnji rezultat je datoteka koja sadrži zapisane slogove svaki u svom redu, a svakom slogu su pridružena trajanja te oznake za početak i kraj sloga, odnosno početak i kraj riječi ako se slog nalazi na početku ili kraju riječi i početak i kraj rečenice ukoliko se slog nalazi na početku i kraju rečenice. Zapis datoteke u krajnjem formatu prikazan je na slici 24.

```
(Pò>    1920000
<noK    960000
PseK    2160000
Pón>    1600000
<djeK   880000
Pù>     240000
<mje>   1120000
<stoK   1840000
Pžä>    2400000
<beK    1200000
Pstvõ>  2880000
<riK    2480000
Pkrá>   4400000
<Le>    1440000
<vićK)  5680000
```

Slika 24 Slogovi s pripadajućim oznakama za početak i kraj sloga, riječi i rečenice i pridruženim trajanjima

7.4 Analiza trajanja slogova

Analizi trajanja slogova pristupilo se sa svrhom provjere utjecaja položaja sloga u riječi i rečenici na trajanje sloga te kontekstualnog okruženja na trajanje sloga.

Slogovi su se prema položaju u riječi podijelili u tri skupine: početni, srednji i završni. U obzir su se uzeli samo slogovi koji se pojavljuju na svim mjestima - u sredini riječi, na početku riječi, na kraju riječi, na početku rečenice i na kraju rečenice. Takvi slogovi i njihov broj pojavljivanja na različitim položajima prikazan je u tablici 16. Pri računanju referentnih vrijednosti trajanja slogova, u obzir su se uzeli samo slogovi na srednjim položajima u riječi za koje se smatra da je utjecaj različitih faktora na trajanje minimalan. Kao referentna vrijednost uzela se prosječna vrijednost trajanja srednjih glasova.

Tabela 16 Slogovi koji se pojavljuju na početnom, srednjem i završnom mjestu u riječi i na početku i na kraju rečenice

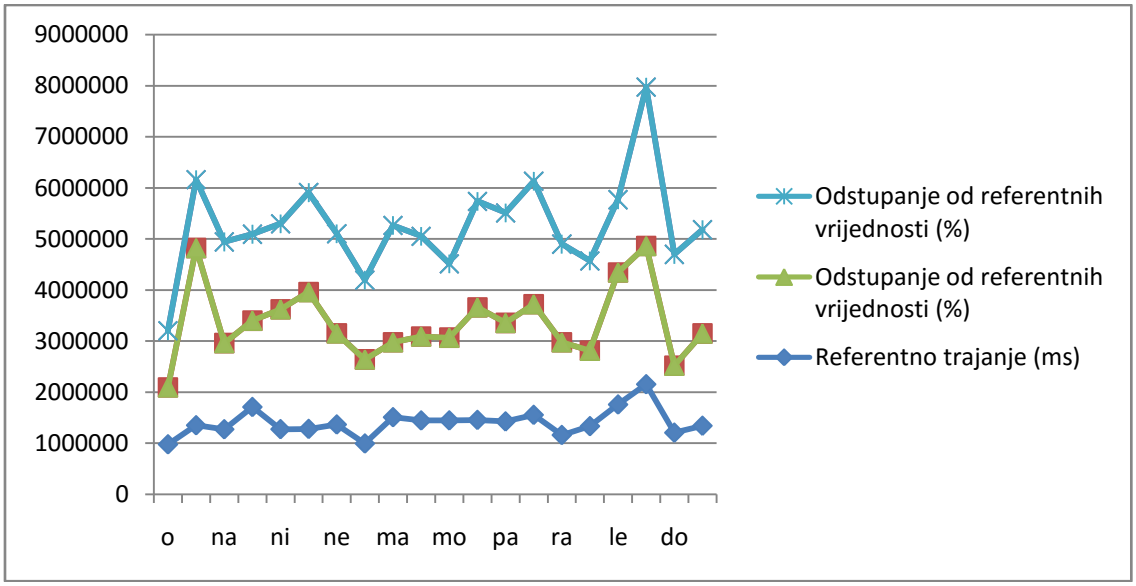
Slog	Broj pojavljivanja u sredini riječi	Broj pojavljivanja na početku riječi	Broj pojavljivanja na početku rečenice	Broj pojavljivanja na kraju riječi	Broj pojavljivanja na kraju rečenice
o	28	153	28	421	55
je	77	8	4	167	9
na	20	50	11	217	59
ko	39	7	1	263	10
ni	87	100	22	110	22
li	70	2	1	172	18
ne	14	44	6	68	18
ri	106	5	1	79	19
ma	49	19	5	116	40
di	84	12	1	83	17
mo	29	64	6	86	6
za	26	49	10	26	3
pa	43	28	3	11	3
ta	42	11	1	92	32
ra	66	25	2	51	19
va	68	2	1	60	21
le	31	3	2	57	13
sa	14	6	1	12	3
do	27	21	2	14	3
mi	21	5	1	26	8

7.4.1 Analiza položajnih čimbenika na trajanje slogova

Analizom trajanja slogova uzimajući u obzir njihove položaje unutar riječi i rečenice pokazalo se da se trajanja slogova u prosjeku povećavaju u odnosu na referentna trajanja ukoliko se nađu na početku ili kraju riječi ili početku ili kraju rečenice. Rezultati u obliku trajanja sloga koja uključuju referentna trajanja, prosječna trajanja na početku riječi te njihova odstupanja od referentnih trajanja i prosječna trajanja na kraju riječi te njihova odstupanja od referentnih trajanja prikazana su u tablici 17 i na slici 25. Kod gotovo svih slogova trajanje se povećalo u odnosu na referentno trajanje ukoliko se nalaze na početku riječi. Najviše se produžilo trajanje sloga *je* - za 156,9% u odnosu na referentno trajanje. U prosjeku su se trajanja slogova povećala za 41,4% u odnosu na referentne vrijednosti ukoliko su se našli na početku riječi. Trajanje slogova ukoliko se nalaze na kraju riječi u prosjeku se produljilo za 37,0% u odnosu na referentna trajanja. Produljenje je također prisutno kod gotovo svih slogova, a od obuhvaćenih slogova, najviše se produljio slog *do* koji se produljio za 80,5% u odnosu na referentno trajanje.

Tabela 17 Referentna trajanja slogova, prosječna trajanja na početku i na kraju riječi te odstupanja od referentnih vrijednosti

Slog	Referentno trajanje (ms)	Prosječno trajanje na početku riječi (ms)	Odstupanje od referentnih vrijednosti (početak riječi)(%)	Prosječno trajanje na kraju riječi (ms)	Odstupanje od referentnih vrijednosti (kraj riječi) (%)
o	977143	1117908	14,4	1111068	13,7
je	1350649	3470000	156,9	1337964	-0,9
na	1272000	1689600	32,8	1980460	55,7
ko	1708718	1691428	-1,0	1692167	-1,0
ni	1274483	2348000	84,2	1674909	31,4
li	1278857	2680000	109,6	1947441	52,3
ne	1365714	1785454	30,7	1951764	42,9
ri	993208	1648000	65,9	1554430	56,5
ma	1508571	1469473	-2,6	2285517	51,5
di	1444762	1646666	14,0	1960481	35,7
mo	1445517	1622500	12,2	1444651	-0,1
za	1455385	2205714	51,6	2073846	42,5
pa	1425116	1931428	35,5	2152727	51,1
ta	1554286	2167272	39,4	2405217	54,7
ra	1160000	1817600	56,7	1921568	65,7
va	1332941	1480000	11,0	1756000	31,7
le	1757419	2586666	47,2	1424561	-18,9
sa	2154285	2706666	25,6	3106666	44,2
do	1205925	1314285	9,0	2177142	80,5
mi	1340952	1808000	34,8	2030769	51,4
prosjek	1400297	1959333	41,4	1899467,4	37,0



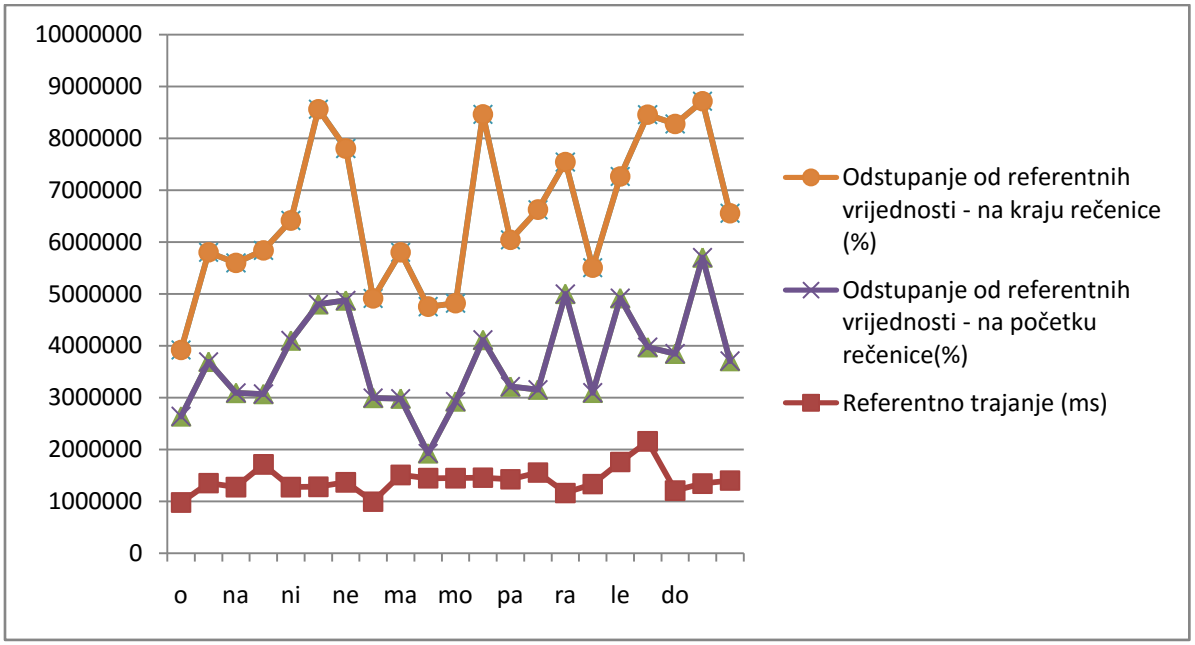
Slika 25 Referentna trajanja slogova, prosječna trajanja slogova na kraju i na početku riječi

Ukoliko se u obzir uzmu položaji slogova unutar rečenice, slogovi se još više produljuju u odnosu na referentne vrijednosti nego kad se u obzir uzmu položaji slogova unutar riječi. Kod gotovo svih slogova koji se nalaze na početnom mjestu u rečenici, trajanje se produljilo, a prosječno se produljilo za 71,8% u odnosu na referentno trajanje. Svim slogovima, bez iznimke, se produljilo trajanje ukoliko su se našli na zadnjem mjestu u rečenici. Prosječno se trajanje produljilo za 104,75% u odnosu na referentno trajanje, a najviše se produljilo trajanje sloga *do* - za 267,1%.

Na temelju rezultata zaključuje se da položaj sloga unutar riječi i rečenice ima veliki utjecaj na trajanje sloga. Najviše na trajanje utječe završni položaj sloga u rečenici koji prosječno povećava trajanje sloga za nešto više od 100%. Početni položaj sloga u rečenici prosječno produljuje njegovo trajanje za oko 70%, a u malo manjim omjerima trajanje sloga produljuju i početno mjesto sloga u riječi te završno mjesto sloga u riječi - prosječno za oko 40%. Rezultati su prikazani u tablici 18 i slici 26.

Tabela 18 Referentna trajanja slogova, prosječna trajanja na početku i na kraju rečenice te odstupanja od referentnih vrijednosti

Slog	Referentno trajanje (ms)	Prosječno trajanje na početku rečenice (ms)	Odstupanje od referentnih vrijednosti (početak rečenice) (%)	Prosječno trajanje na kraju rečenice (ms)	Odstupanje od referentnih vrijednosti (kraj rečenice)(%)
o	977143	1660000	69,9	1280000	31
je	1350649	2340000	73,3	2106666	56
na	1272000	1818181	42,9	2504406	96,9
ko	1708718	1360000	-20,4	2768000	62
ni	1274483	2825454	121,7	2312727	81,5
li	1278857	3520000	175,2	3760000	194
ne	1365714	3506666	156,8	2928888	114,5
ri	993208	2000000	101,4	1915789	92,9
ma	1508571	1472000	-2,4	2818000	86,8
di	1444762	480000	-66,8	2828235	95,8
mo	1445517	1480000	2,4	1893333	31
za	1455385	2656000	82,5	4346666	198,7
pa	1425116	1786666	25,4	2826666	98,3
ta	1554286	1600000	2,9	3467500	123,1
ra	1160000	3840000	231	2538947	118,9
va	1332941	1760000	32	2407619	80,6
le	1757419	3160000	79,8	2344615	33,4
sa	2154285	1817142	-15,6	4480000	108
do	1205925	2640000	118,9	4426666	267,1
mi	1340952	4360000	225,1	3010000	124,5
prosjek	1400297	2304105	71,8	2848236	104,75



Slika 26 Referentna trajanja slogova, prosječna trajanja slogova na kraju i na početku rečenice

7.4.2 Utjecaj kontekstualnih čimbenika na trajanje slogova

Analizom kontekstualnih čimbenika željelo se provjeriti ima li vrsta glasa kojim započinje sljedeći slog utjecaj na trajanje sloga. U obzir su se uzimali isti slogovi kao kod analize položajnih čimbenika na trajanje, ali samo onda kad se nalaze na početnim položajima u riječi obzirom da su kao referentne vrijednosti uzeta trajanja slogova na srednjem položaju u riječi. Ovisno o tome kojom vrstom suglasnika započinje slog koji se nalazi nakon promatranog sloga, sljedeći slog podijeljen je u sedam skupina: one koje započinju približnikom, bočnikom, nosnikom, zapornikom, tjesnačnikom, slivenikom i treptajnikom. Budući da se treptajnik kao prvi glas u slogu nakon promatranog sloga pojavio samo jednom, prilikom analize nije uzet u obzir. Na temelju prosječnih trajanja slogova nakon kojih slijede različite skupine slogova (s obzirom na vrstu prvog glasa u slogu) dobivena su prosječna odstupanja od referentnih trajanja po skupinama. Rezultati su prikazani u tablici 19. Na temelju rezultata može se zaključiti da se trajanje početnih slogova u riječi produljuje u različitom omjeru u odnosu na referentno trajanje ovisno o tome kojoj skupini suglasnika pripada glas kojim započinje sljedeći slog. Trajanje sloga se najviše produljuje ako iza njega slijedi slog koji započinje glasom iz skupine tjesnačnika - za 60,4%, a zatim približnika - za 44,9%. Ukoliko sljedeći slog započinje glasom iz skupine nosnika, zapornika ili slivenika, početni se slog u prosjeku produljuje za oko 30%. Najmanje utjecaja na trajanje prethodnog

sloga ima slog koji započinje glasom iz skupine bočnika te prosječno produljenje tada iznosi 16,2%.

Tabela 19 Utjecaj prvog glasa sloga koji slijedi na trajanje sloga

Slog koji slijedi nakon početnog počinje s:	Prosječno trajanje (ms)	Prosječno odstupanje od referentnog trajanja (%)
Približnikom	1931071	44,9
Bočnikom	1758095	16,2
Nosnikom	1887211	30,0
Zapornikom	1812287	27,3
Tjesnačnikom	2228103	60,4
Slivenikom	1722909	33,6

7.5 Model trajanja slogova

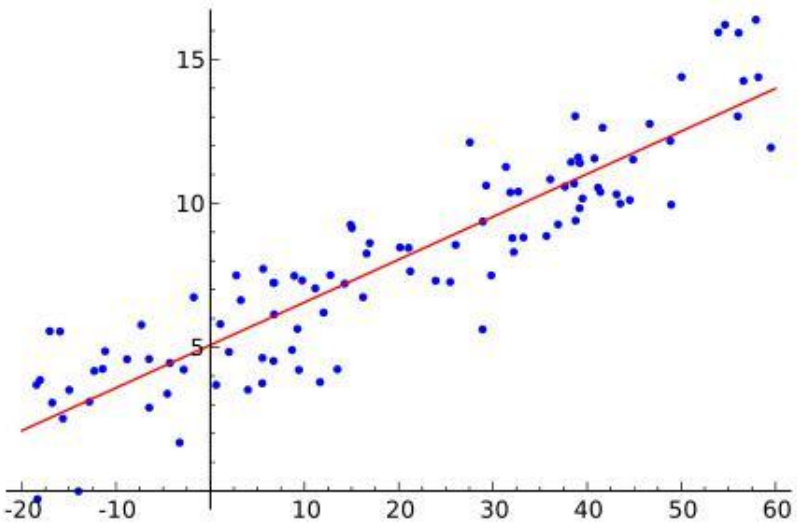
Već je u uvodu napomenuto kako je jedno od najvažnijih prozodijskih sredstava trajanje segmenata govora. Trajanje segmenata govora je glavni faktor zaslužan za dojam ritma i tempa govora (Dutoit, 1997), a govor u kojem sve jedinice govora traju jednako dugo zvuči neprirodno, monotono, neugodno za slušanje te teško razumljivo. Modeliranje trajanja, dakle, ima značajnu ulogu u prirodnosti govora.

Postoji više metoda i algoritama za izgradnju modela trajanja, a neki od njih opisani su u 3. poglavlju. U ovom su se poglavlju primijenili neki od navedenih algoritama za modeliranje trajanja slogova hrvatskoga jezika. Konkretno, primijenio se model linearne regresije, stabla linearnih regresijskih modela izgrađena korištenjem algoritma M5 i metoda potpornih vektora pomoću alata Weka (Witten, Frank, & Hall, 2011).

Modeli trajanja slogova izgrađeni su na temelju podataka o trajanju slogova prirodnog govora iz baze VEPRAD, podskupa korpusa bajki i priča. Iz tekstualnih transkripcija govora izlučene su jezične značajke, a iz segmentacije na razini glasa s odgovarajućim trajanjima, dobivena su trajanja slogova postupkom opisanim ranije u ovom poglavlju.

Skup podataka podijelio se u dva dijela, skup za učenje i skup za testiranje u omjeru 80:20. Modeli su izgrađeni korištenjem podataka za učenje, a njihova učinkovitost testirana je na skupu podataka za testiranje mjerenjem odstupanja predviđanja modela od stvarnih vrijednosti.

Model linearne regresije zasniva se na modeliranju odnosa između skalarne zavisne varijable i jedne ili više nezavisne varijable. U slučaju kad postoji jedna nezavisna varijabla, govorimo o jednostavnoj linearnoj regresiji, a kad postoji više nezavisnih varijabli, onda je riječ o višestrukoj linearnoj regresiji. Na slici 27 prikazana je jednostavna linearna regresija s jednom nezavisnom varijablom.



Slika 27 Jednostavna linearna regresija s jednom nezavisnom varijablom

Pomoću modela linearne regresije, odnosi se modeliraju korištenjem linearne funkcije predviđanja čiji se nepoznati parametri računaju na temelju dostupnih podataka. Jedna od mogućnosti korištenja linearne regresije je predviđanje nepoznate vrijednosti u zavisnosti od neke druge vrijednosti. Kada se koristi u takvom slučaju onda se za predviđanje uči model na podacima u kojima postoje višestruki primjeri zavisnosti varijabli.

Stablo linearnih modela je stablo odlučivanja u kojem su u čvorovima pitanja koja se odnose na jezične značajke, a u listovima linearni regresijski modeli. Za učenje stabla linearnih modela korišten je algoritam M5 (Wang & Witten, 1997). M5 algoritam uzastopno dijeli skup uzoraka tako da se unutar rezultirajućih podskupova maksimizira SDR (engl. standard deviation reduction). SDR se računa po formuli:

$$\mathrm{SDR}(\mathrm{t_i}) \quad = sd(t_i) - \sum_j \frac{|\mathrm{t_j}|}{|\mathrm{t_i}|} * \mathrm{sd}(\mathrm{t_j})$$

gdje je t_i skup uzoraka pridruženih i-tom čvoru, a t_j, j = 1,2,... , skupovi uzoraka koji nastaju dijeljenjem čvora prema odabranoj značajki. Čvorovi se rekurzivno dijele odabirom značajke koja daje najveću vrijednost SDR dok broj uzoraka u čvoru ne padne ispod zadanog praga ili kada je vrijednost izlazne veličine svih uzoraka u čvoru približno ista.

Metoda potpornih vektora (engl. support vector machines) može se koristiti za rješavanje problema klasifikacije, ali i regresije. U ovom se je slučaju koristila za predviđanje trajanja slogova pomoću regresije. Metoda se zasniva na učenju iz podataka formata {(x_1, y_1),

(x_2, y_2), ...(x_n, y_n)}, gdje x_i predstavlja ulazni uzorak, a y_i ciljnu vrijednost za odgovarajuću ulaznu vrijednost. Kada govorimo o regresiji, cilj je pronaći funkciju koja kao rezultat daje izlaz unutar određenih zadanih odstupanja (Smola & Scholkopf, 1998).

7.5.1 Jezične značajke za učenje modela trajanja

Jezične značajke koje su se koristile za učenje modela trajanja su indeks sloga koji predstavlja broj slogova koji se nalazi ispred trenutnog sloga u rečenici/frazi, broj slogova koji se nalaze iza trenutnog sloga u rečenici/frazi, položaj sloga unutar riječi - početak, sredina ili kraj, položaj riječi unutar fraze - početak, sredina ili kraj, broj riječi koje se nalaze ispred i iza trenutne riječi, ukupan broj slogova u rečenici/frazi, ukupan broj slogova u rečenici/frazi, dužina sloga, vrsta sloga obzirom na raspored glasova u slogu, identitet samoglasnika u slogu, vrsta naglaska u slogu, vrsta naglaska u prethodnom i sljedećem slogu, identitet prvog glasa u sljedećem slogu, identitet zadnjeg glasa u prethodnom slogu. Jezične značajke prikazane su u tablici 20. Primjer jezičnih značajki prikazan je na slici 28.

Tabela 20 Jezične značajke korištene u modelu trajanja

Jezična značajka	**Opis značajke**	**Moguće vrijednosti**
Indeks sloga	Redni broj sloga u rečenici	prirodni broj (1-12)
Indeks riječi	Redni broj sloga u rečenici	prirodni broj
Broj riječi	Ukupan broj riječi u rečenici	prirodni broj
Broj slogova	Ukupan broj slogova u rečenici	prirodni broj
Dužina sloga	Ukupan broj glasova u slogu	prirodni broj
Položaj unutar riječi	Položaj sloga unutar riječi	početak - 1, kraj-2, sredina-0
Položaj unutar rečenice	Položaj sloga unutar rečenice	početak - 1, kraj-2, sredina-0
Broj slogova ispred	Broj slogova koji se nalaze ispred trenutnog	prirodni broj
Broj slogova iza	Broj slogova koji se nalaze iza trenutnog	prirodni broj
Broj riječi ispred	Broj riječi koje se nalaze ispred trenutne	prirodni broj
Broj riječi iza	Broj riječi koje se nalaze iza trenutne	prirodni broj
Vrsta sloga	Vrsta sloga obzirom na raspored glasova u slogu	prirodni broj 1-15 (svaki broj predstavlja jednu vrstu)
Identitet samoglasnika	Identitet samoglasnika u slogu	prirodni broj (svaki broj predstavlja jedan samoglasnik)
Vrsta naglaska	Vrsta naglaska u slogu	prirodni broj (svaki broj predstavlja jednu vrstu naglaska)
Vrsta naglaska prethodnog sloga	Vrsta naglaska prethodnog sloga	prirodni broj (svaki broj predstavlja jednu vrstu naglaska)
Vrsta naglaska sljedećeg sloga	Vrsta naglaska sljedećeg sloga	prirodni broj (svaki broj predstavlja jednu vrstu naglaska)
Identitet prvog glasa u sljedećem slogu	Identitet prvog glasa u sljedećem slogu	prirodni broj (svaki broj predstavlja jedan glas)
Identitet zadnjeg glasa u prethodnom slogu	Identitet zadnjeg glasa u prethodnom slogu	prirodni broj (svaki broj predstavlja jedan glas)

7.5.2 Rezultati automatskog predviđanja trajanja slogova

Modeli trajanja su procijenjeni usporedbom predviđenih vrijednosti trajanja sa stvarnim vrijednostima za testni skup podataka. Mjere koje su se koristile su korijen srednje kvadratne pogreške (engl. root mean square error - RMSE) i koeficijent korelacije predviđenih i stvarnih vrijednosti i trajanja. Što je vrijednost RMSE manja, a koeficijent korelacije veći, to je točnost modela veća.

Kako bi se provjerio utjecaj položajnih i kontekstualnih značajki te prisutnost i vrsta naglaska na točnost modela za trajanje, značajke su se podijelile u četiri skupine: one koje se odnose na položaj segmenata, one koje se odnose na kontekst, one koje se odnose na naglasak te skupina sa svim navedenim značajkama.

U položajne značajke ubrojile su se sljedeće značajke:

- Indeks sloga,
- Indeks riječi,
- Broj riječi,
- Broj slogova,
- Položaj unutar riječi,
- Položaj unutar rečenice,
- Dužina sloga.

U kontekstualne značajke ubrojile su se sljedeće značajke:

- Broj slogova ispred,
- Broj slogova iza,
- Broj riječi ispred,
- Broj riječi iza,
- Vrsta naglaska prethodnog sloga,
- Vrsta naglaska sljedećeg sloga,
- Identitet prvog glasa u sljedećem slogu,
- Identitet zadnjeg glasa u prethodnom slogu,
- Broj riječi,
- Broj slogova.

U značajke koje se odnose na naglasak ubrojile su se značajke:

- Vrsta sloga,
- Identitet samoglasnika,
- Vrsta naglaska,
- Vrsta naglaska prethodnog sloga,
- Vrsta naglaska sljedećeg sloga.

Najprije se izračunala vrijednost RMSE sa svim skupinama značajki. Rezultati sva tri algoritma - linearne regresije, stabala linearnih regresijskih modela i metode potpornih vektora dale su približno jednake rezultate. Najbolji rezultati postignuti su pomoću stabala linearnih regresijskih modela s vrijednošću RMSE od 35,3 te koeficijentom korelacije od 0,72. Rezultati su prikazani u tablici 21.

Tabela 21 Rezultati modela trajanja

	Model linearne regresije	Stabla linearnih regresijskih modela	Metoda potpornih vektora
RMSE	38,1	**35,3**	37,2
Koeficijent korelacije	0,69	**0,72**	0,69

Kako bi se provjerio utjecaj značajki po skupinama, iz skupa značajki na kojima se učio model izuzimale su se jedna po jedna skupina značajki i svaki put bi se izmjerila točnost modela u vidu vrijednosti RMSE. Rezultati dobiveni na taj način prikazani su u tablici 22.

Tabela 22 Rezultati modela trajanja dobiveni isključivanjem pojedinih skupina značajki iz skupa značajki za učenje modela

Skupina jezičnih značajki	Model linearne regresije (RMSE)	Stabla linearnih regresijskih modela (RMSE)	Metoda potpornih vektora (RMSE)
bez položajnih značajki	41,2	**39,1**	41,3
bez kontekstualnih značajki	39,2	**35,9**	38,2
bez značajki koje se odnose na naglasak	38,7	**36,2**	37,9
sve značajke	38,1	**35,3**	37,2

Na temelju rezultata može se zaključiti da uključivanje svake od skupine značajki u skup značajki na kojima se model uči, doprinosi većoj točnosti modela trajanja. Pritom prema dobivenim rezultatima najviše utjecaja imaju položajne značajke čijim su se isključivanjem vrijednosti RMSE najviše pogoršale u odnosu na vrijednost RMSE kad se model uči na svim značajkama.

8. Tilt intonacijski model

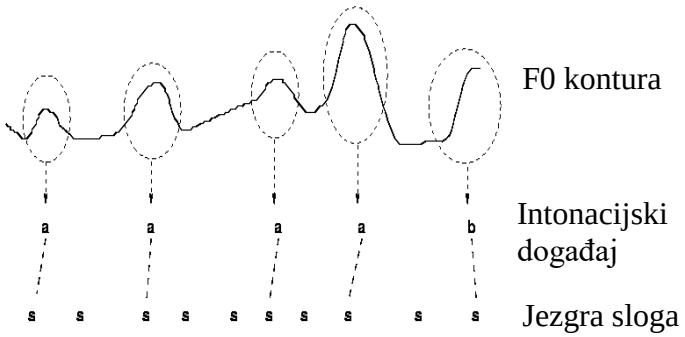
Uz trajanje, jedno od najvažnijih prozodijskih sredstava je intonacija. U ovom će se poglavlju opisati postupak modeliranja intonacije pomoću Tilt modela.

Tilt (Taylor, 2000)je fonetski intonacijski model koji F0 konturu prikazuje kao slijed kontinuirano parametriziranih događaja. Takvi parametri se onda nazivaju tilt parametri, a određuju se na temelju F0 konture.

Tilt model, uz engleski jezik primijenjen je i na neke druge jezike - primjerice na slovenski i španjolski u (Rojc, Aguero, Bonafonte, & Kacic, 2005) te mandarinski u (Thangthai, Thatphithakkul, Wutiwiwatchai, Saychum, & Rugchatjaroen, 2008). Primjena intonacijskog modela na hrvatski jezik provedena je u radu (Načinović, Pobar, Martinčić-Ipšić, & Ipšić, 2011) u kojem se pokušalo provjeriti je li Tilt model prikladan za primjenu na hrvatskom jeziku na pokusnom skupu od 100 rečenica označenih tilt oznakama. Rezultati su bili ohrabrujući te su naveli na zaključak da bi se s većim skupom označenih rečenica vjerojatno dobili i bolji rezultati. Zato se u ovom radu skup ručno označenih rečenica tilt oznakama proširio na ukupno njih 500.

8.1 Pregled Tilt modela

Osnovna jedinica Tilt modela je intonacijski događaj. Osnovni tipovi događaja su naglasak i granični tonovi. Budući da Tilt model opisuje F0 konturu, možemo reći da su naglasci kod Tilt modela dijelovi F0 konture povezani s naglašenim slogovima, a granični tonovi su rastući događaji koji se obično pojavljuju na kraju intonacijske fraze te mogu naznačiti posebne izraze poput upitnosti, nedovršenosti, čuđenja i sl. Moguć je i događaj koji je kombinacija dvaju navedenih događaja, a nalazi se u slučajevima kad se naglasak i granični ton pojavljuju tako blizu jedan drugome da se uzimaju kao jedan. Pomoću kombinacije odabira triju događaja, moguće je opisati različite globalne intonacijske melodije tj., kombinacijom događaja oslikava se rečenična intonacija. Događaji se pridružuju jezgri sloga odnosno samoglasniku. Svaki se događaj pridružuje jezgri sloga, ali se svakoj jezgri ne pridružuje događaj. Intonacijski događaji zaokruženi su na slici 29.



Slika 28 Intonacijski događaji u Tilt modelu

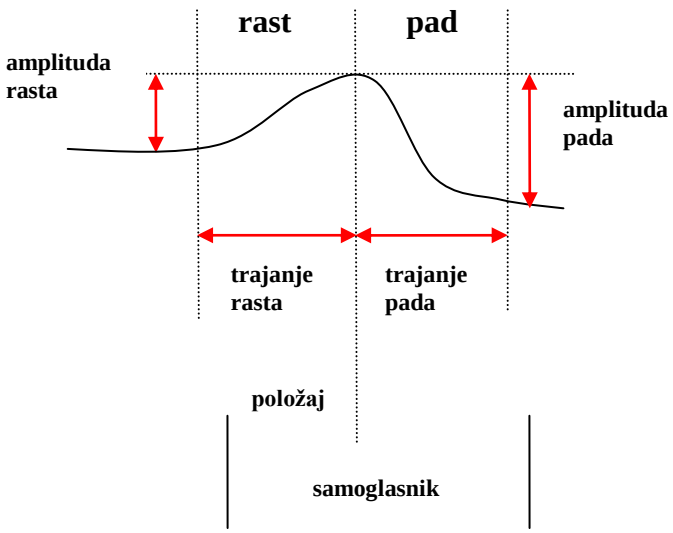
Obzirom da Tilt model opisuje intonacijske pojave koje se očituju u F0 konturi, model se svrstava u fonetske modele. Za razliku od fonoloških modela koji korite klasifikaciju intonacijskih događaja po kategorijama, Tilt model koristi skup kontinuiranih parametara. Tilt parametri dobivaju se na temelju oblika F0 konture. Model se temelji na RFC modelu (Taylor, 1995).

RFC parametri su:

- amplituda rasta (Hz),
- trajanje rasta (sekunde),
- amplituda pada (Hz),

- trajanje pada (sekunde),
- položaj (sekunde),
- visina F0 (Hz).

U RFC modelu, svaki se događaj opisuje pomoću rastućeg oblika, padajućeg ili rastućeg kojeg slijedi padajući. Zatim se svaki događaj parametrizira mjereći amplitude i trajanja rasta i pada događaja. Za rastući oblik nakon kojeg slijedi padajući, određuju se tri mjere - vrh, početak i kraj. Trajanje rasta predstavlja udaljenost od početka događaja do vrha, a trajanje pada je udaljenost od vrha pa do kraja događaja. Amplituda rasta je razlika u F0 vrijednosti vrha događaja i početka događaja, a amplituda pada je razlika F0 vrijednosti vrha događaja i kraja događaja. Dakle, svakom se događaju pridružuju četiri vrijednosti - amplituda rasta, trajanje rasta, amplituda pada i trajanje pada. Ukoliko se događaj sastoji samo od komponente rasta, amplitudi pada i trajanju pada se vrijednosti postavljaju na 0. Isto tako kad se intonacijski događaj sastoji samo od komponente pada, onda se amplitudi i trajanju rasta pridruži vrijednost 0. Uz to svakom se događaju pridružuju po još dva parametra koji određuju vremenski položaj događaja unutar rečenice i visinu F0 događaja. Na slici 30 prikazani su RFC parametri jednog događaja u Tilt modelu.



Slika 29 RFC parametri u Tilt modelu

Zbog lakše interpretacije i korištenja, RFC parametri pretvaraju se u 3 parametra - trajanje, amplituda i tilt parametar. Trajanje je pritom vrijednost sume trajanja rasta i pada, amplituda je suma magnituda amplituda rasta i pada, a tilt parametar je bezdimenzijski broj koji odražava cjelokupan oblik događaja. Uz smanjeni broj parametara u odnosu na RFC model, točnost modela se značajno ne smanjuje (Taylor, 2000).

RFC parametri se prema sljedećim formulama pretvaraju u tilt parametre:

- Tilt-amplituda (Hz): suma magnituda amplituda rasta i pada

$$tilt_{amp} = \frac{|A_{rast}| - |A_{pad}|}{|A_{rast}| + |A_{pad}|}$$

- Tilt-trajanje (sekunde): suma trajanja rasta i pada

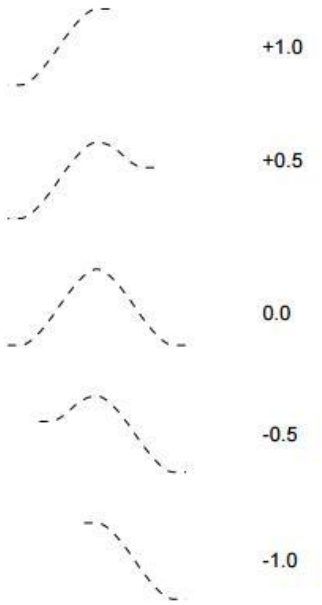
$$tilt_{dur} = \frac{|D_{rast}| - |D_{pad}|}{|D_{rast}| + |D_{pad}|}$$

- Tilt: bezdimenzijski broj koji izražava cjelokupni oblik događaja neovisno o amplitudi i trajanju

$$tilt = \frac{|A_{rast}| - |A_{pad}|}{2|A_{rast}| + |A_{pad}|} + \frac{|D_{rast}| - |D_{pad}|}{2|D_{rast}| + |D_{pad}|}$$

Amplituda događaja predstavlja fonetsku istaknutost događaja - što je amplituda događaja u određenom položaja veća, to je istaknutost veća.

Vrijednost tilt izračunava se na temelju relativnih veličina komponenti rasta i pada u događaju. Vrijednost +1 označava događaj koji raste, a vrijednost -1 indicira čisti pad kao događaj. Bilo koja vrijednost između znači događaj koji se sastoji i od komponenti rasta i pada, vrijednost 0 označava da su veličine komponente rasta i pada jednake. Primjeri intonacijskih događaja s različitim vrijednostima tilt parametra prikazani su na slici 31.



Slika 30 Primjer 5 različitih intonacijskih događaja s različitim vrijednostima tilt parametra

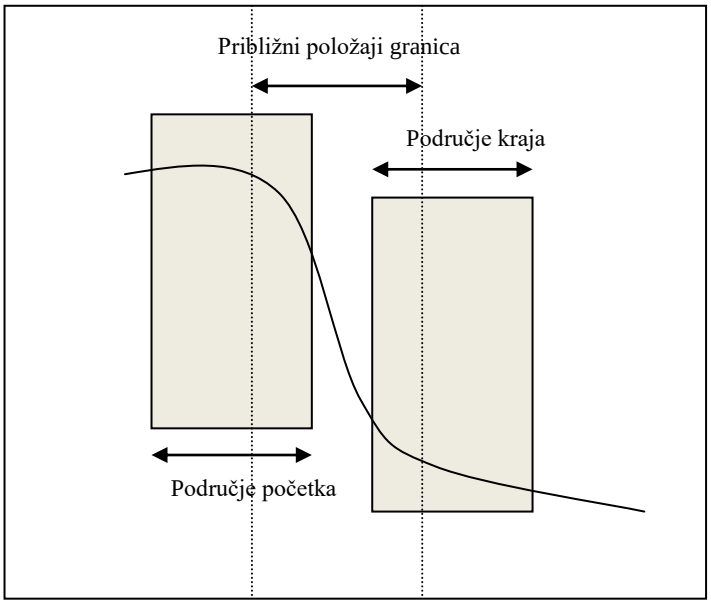
8.1.2 Automatska RFC analiza

U postupku pronalaženja intonacijskih događaja, izvodi se segmentacija iskaza iz koje je moguće izračunati položaje početka i kraja događaja. Tijekom automatske RFC analize izvode se koraci za određivanje točnog položaja početka, vrha i kraja događaja te korištenja određenih vrijednosti za izračunavanje amplitude rasta i pada te trajanja rasta i pada. RFC analiza primjenjuje se samo na dijelove F0 konture koji su izdvojeni kao intonacijski događaji. Na svaki se događaj primjenjuje algoritam za zaglađivanje, a dijelovi bez glasa se rekonstruiraju metodom interpolacije. Zaglađivanje se koristi kako bi se izbjegle izrazito istaknute F0 vrijednosti konture koje su obično rezultat pogreške tijekom postupka određivanja područja događaja te kako bi se otklonile smetnje u F0 konturi koje proizlaze iz prirodnih varijacija u proizvodnji zvuka.

Nakon zaglađivanja, algoritam za traženje vrha koristi se za određivanje je li događaj sastavljen samo od komponente rasta, samo od pada ili kombinacije rasta i pada. Ukoliko se pronađe vrh, onda se događaj svrstava u kombinirani događaj. Položaj vrha, ukoliko je prisutan te početni i završni događaj koji su dobiveni prilikom određivanja događaja, koriste se za definiranje tzv. područja pretraživanja. Ukoliko se radi o događaju koji se sastoji samo od pada ili događaja, onda se područje pretraživanja ograničava na 20% ispred i iza granica koje su se dobile u postupku određivanja događaja. Obično u područje od navedenih 20%

ulazi 10 okvira po 10 ms na početku i 10 okvira nakon granice. Svaki početni okvir u kombinaciji sa svakim krajnjim okvirom uzimaju se kao moguća točka početka odnosno kraja te se F0 kontura sintetizira za svaku kombinaciju početka i kraja. Svaka se od tih kontura uspoređuje sa stvarnim vrijednostima F0 konture u danoj točki te se kontura s najnižom vrijednošću Euklidove udaljenosti uzima kao najbolja (prikazano na slici 32).

Ukoliko se radi o događaju koji je kombinacija dvaju komponenti, provodi se sličan postupak, ali se u tom slučaju provode dvije pretrage - jedna za određivanje rasta, a druga za određivanje pada. Kod određivanja rasta, za određivanje početka koristi se isti postupak kao što je opisan u prethodnom odlomku, a za kraj se uzima vrh. Kod određivanja pada se pak vrh uzima za početak, a postupak pronalaženja kraja je isti kao što je opisano u prethodnom poglavlju. Za svaki se događaj u iskazu provodi isti postupak. Nakon dobivanja RFC parametara, pomoću formula opisanih u poglavlju 8.1, mogu se dobiti tilt parametri.



Slika 31 Područje pretraživanja za događaj koji se sastoji samo od komponente pad[15]

[15]Preuzeto iz: (Taylor, Analysis and Synthesis of Intonation using the Tilt Model, 2000)

8.2 Primjena Tilt modela na hrvatski jezik

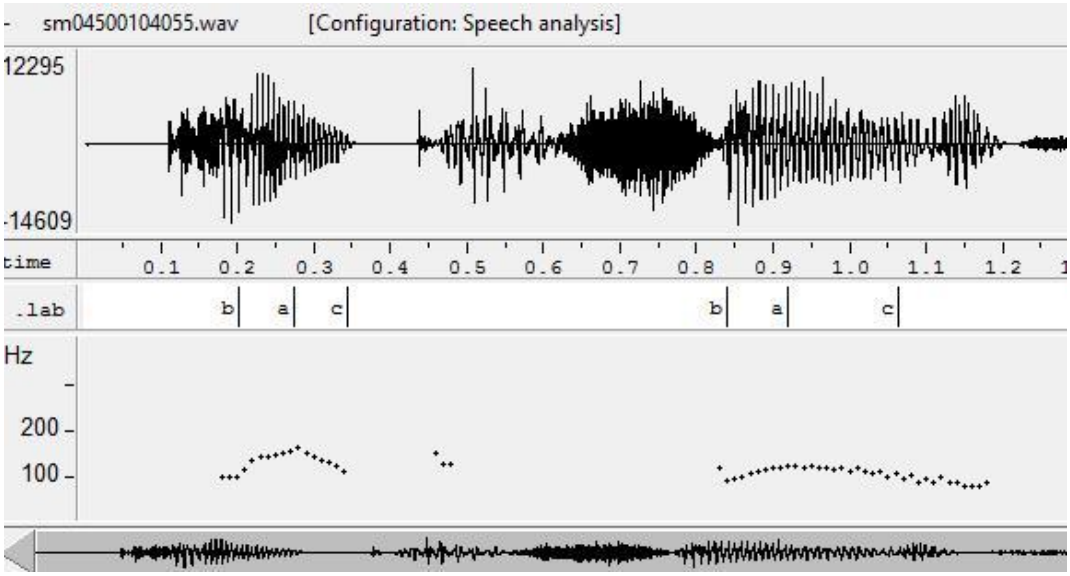
Jedan od razloga zašto se za modeliranje hrvatske intonacije odabrao Tilt intonacijski model je relativno lako označavanje intonacijskih događaja u odnosu na neke druge modele. Isto tako, za hrvatski jezik nije postojao korpus označen oznakama za neki od postojećih intonacijskih modela. Drugi razlog je bio taj da su glavni događaji Tilt modela (uz granične tonove) naglasci (engl. pitch accents), a hrvatski jezik prema tipologiji pripada ograničenim tonskim jezicima (engl. pitch-accent language). Osim toga, granični tonovi također postoje u hrvatskom jeziku budući da su neke od univerzalnih osobina rečenične intonacije u svim jezicima, kao što je ranije rečeno, padajući ton na kraju rečenice, a nepotpuni intonacijski pad na kraju iskazuje nedovršenost rečenice, prekid govora, čuđenje, pitanje, oklijevanje ili slično. Stoga se pretpostavilo da bi intonacijski događaji Tilt modela mogli biti primjenjivi i na hrvatski jezik.

8.2.1 Postupak nalaženja događaja za model Tilt

Za pronalaženje događaja relevantnih za Tilt model i označavanje cijelog korpusa, razvijen je automatski postupak zasnovan na skrivenim Markovljevim modelima. U postupku se koriste 4 zasebna modela kako bi se predvidjele četiri vrste događaja na temelju F0 konture. Skriveni Markovljevi modeli učeni su na skupu označenih rečenica i govornih signala kojima su pridružene odgovarajuće tilt oznake. Korpus koji se koristio isti je govorni korpus bajki i priča koji se koristio u poglavlju 7.

8.2.1.1 Postupak ručnog pridruživanja oznaka

Iz korpusa je nasumično odabran skup od 500 rečenica kojima se ručno pridružila transkripcija intonacije u vidu tilt oznaka. Svakoj su se rečenici pridružile oznake vrha događaja *a*, početka rasta događaja *b*, početak pada događaja *c*, granični ton *g* te oznaka tišine *sil*. Postupak dodjeljivanja oznaka proveden je pomoću alata WaveSurfer (Sjolander i Beskow, 2000). Datoteke s oznakama izvezle su se kao .lab datoteke te su nakon toga korištene u automatskom pronalaženju intonacijskih događaja. Primjer jednog iskaza s dodijeljenim tilt oznakama prikazana je na slici 33.



Slika 32 Primjer govornog signala s pridruženim tilt oznakama

8.2.1.2 Izlučivanje značajki F0

Za izlučivanje F0 značajki korišten je algoritam RAPT (Talkin, 1995). F0 kontura uzorkovana je svakih 10 ms. Dobivena F0 kontura sadržavala je određeni stupanj šumova pa je izlaz izglađen filtrom "medijan triju točaka". U drugom je pokušaju vrijednost F0 konture na mjestima gdje se ona nije mogla odrediti postavljena na 0 Hz, a zatim se koristila linearna interpolacija kako bi se odredile vrijednosti koje nedostaju. Na taj su se način dobila tri različita skupa F0 značajki - izravan izlaz iz RAPT algoritma, zaglađene vrijednosti i interpolirane vrijednosti.

8.2.1.3 Automatsko pronalaženje događaja

Za pronalaženje mjesta vrha događaja, početka rasta, početka pada i mjesta gdje se nalazila tišina naučili su se skriveni Markovljevi modeli. Za svaku vrstu događaja naučio se model s pet stanja. Modeli su učeni Baum-Welch algoritmom koristeći podatke o značajkama F0 konture i položajima ručno označenih događaja. Za svaki od tri skupa F0 značajki iz prethodnog poglavlja izgrađen je zaseban skriven Markovljev model pomoću alata HTK (Young, i dr., 2006).

Kako bi se ograničili valjani sljedovi događaja, definirala se gramatika s dozvoljenim sljedovima događaja.

8.2.1.4 Tilt analiza

Kada su se događaji pronašli, potrebno je odrediti početak, vrh i kraj svakog događaja. U tu svrhu koristio se alat Tilt analiza koja se primjenjuje na one dijelove F0 konture koji su se odredili kao intonacijski događaji. Svaki se događaj opisuje kao rastući ili padajući oblik unutar F0 konture pa se svakom događaju dodjeljuju tilt parametri. Na taj se način dobije model predstavljen tilt parametrima koji su ranije pojašnjeni.

8.2.1.5 Tilt sinteza

Na temelju tilt parametara pomoću sljedećih se formula dobiju RFC parametri:

$$A_{rast} = \frac{A_{doga\ đaj}\,(1+tilt)}{2},$$

$$A_{pad} = \frac{A_{doga\ đaj}\,(1-tilt)}{2},$$

$$D_{rast} = \frac{D_{doga\ đaj}\,(1+tilt)}{2},$$

$$D_{rast} = \frac{D_{doga\ đaj}\,(1-tilt)}{2}.$$

Svaki se događaj dijeli na zasebne komponente rasta i pada, te se zatim svaki od njih sintetizira po sljedećim formulama:

$$F0(t)=A_{abs}+A-2A(t/D)^2, \qquad 0<t<D/2\ ,$$

$$F0(t)=A_{abs}+2A(1-t/D)^2, \qquad D/2<t<D$$

gdje je A amplituda rasta ili pada, D je trajanje rasta ili pada, a A_{abs} je apsolutna vrijednost F0 na početku rasta ili pada koja se dobije na temelju krajnje vrijednosti prethodnog događaja ili spoja.

Spojevi se sintetiziraju koristeći sljedeću formulu:

$$F0(t)=A_{abs}+A(t/D), \qquad 0<t<D$$

gdje je A amplituda spoja, D trajanje spoja, a A_{abs} je apsolutna vrijednost F0 na početku spoja koja se dobije na temelju krajnje vrijednosti prethodnog događaja.

8.2.2 Rezultati Tilt modela

Generirana F0 kontura evaluirana je mjerom korijena srednje kvadratne pogreške (engl. root mean square error - RMSE) koja mjeri razliku između originalne i generirane F0 konture. Za svaki model učen na različitom skupu F0 značajki (izravan izlaz, zaglađeni i interpolirani) određena je zasebna RMSE mjera. Modeli su testirani na 80 rečenica. Rezultati su prikazani u tablici 23.

Tabela 23 RMSE vrijednosti za generirane F0 konture

Model	RMSE (Hz)
izravan izlaz iz RAPT algoritma	**22,2**
zaglađeni izlaz iz RAPT algoritma	26,3
interpolirani izlaz iz RAPT algoritma	24,6
ručno-označeni događaji	20,1

Najbolji rezultati dobili su se modelom koji je naučen na F0 značajkama koje su bile izravan izlaz iz RAPT algoritma.

9. Rasprava

Naglasak riječi u hrvatskom jeziku, osim uloge isticanja riječi, ima i ulogu isticanja razgraničenja riječi i razlikovnu ulogu u primjerima istovjetnosti dvaju ili više izraza riječi. Budući da je sustav naglašavanja u hrvatskom jeziku prilično složen te je gotovo proizvoljan, kako bi se riječima automatski dodijelili naglasci neophodno je postojanje rječnika svih osnovnih i izvedenih oblika leksema s označenim naglascima. Obzirom da za hrvatski takav rječnik nije postojao, on se izradio u sklopu ove disertacije.

Hrvatski je jezik izrazito flektivan sa složenim konjugacijskim i sklonidbenim paradigmama. Dodjeljivanje naglasaka riječi ipak poštuje određena pravila u hrvatskom jeziku, a ta je pravila u svojoj disertaciji (Mikelić Preradović, 2008) izdvojila za generiranje naglašenih oblika imenica, glagola i pridjeva. Koristeći i prilagođavajući navedena pravila, a na temelju njihovog osnovnog oblika koji je preuzet iz leksičke baze *Veliki rječnik hrvatskoga jezika*(Anić, 2009), u ovoj je disertaciji stvoren naglasni morfološki analizator/generator koji riječima automatski dodjeljuje odgovarajući naglasak.

Pravilima za dodavanje morfoloških nastavaka i promjenom naglaska ovisno o stupnju sklonidbe, komparacije i konjugacije obuhvaćene su tri najbrojnije skupine promjenjivih vrsta riječi - imenice, glagoli i pridjevi. Svaka od te tri skupine ima više različitih podskupina i posebnosti koje su se također uzele u obzir prilikom programiranja.

Kod imenica su primjerice za svaki tip kojih ima ukupno 196 za muški, 73 za ženski i 80 za srednji rod, za svaki rod definirana različita pravila za jednosložne, dvosložne, trosložne i višesložne osnove, a moguće su i različite iznimke te imenice koje slijede vlastiti tip morfološko-naglasne sklonidbe kao što su, primjerice, tuđice. Pravila za imenice implementirana su u ovoj disertaciji u programskom kodu s više od 11.000 linija koda u programskom jeziku Python. Pravilima za glagole obuhvaćeno je ukupno 147 grupa i podgrupa glagola, a ta su pravila implementirana s nešto više od 5000 linija koda. Pravilima je ukupno obuhvaćeno 97 različitih jednosložnih, dvosložnih i trosložnih tipova pridjeva, a za programiranje pravila o pridjevima bilo je potrebno preko 5000 linija kôda. Za ostale vrste riječi - zamjenice, priloge, prijedloge, brojeve, uzvike, čestice i veznike, osnovni oblici preuzeli su se iz leksičke baze, a za neke su se još i ručno dodavali dodatni oblici, kao što su primjerice oblici zamjenice kroz padeže te neki od izvedenih oblika brojeva.

Osim što su se izveli svi oblici riječi prema pravilima za dodavanje nastavaka i promjenu nastavaka, riječima se u okviru programskog koda dodala i MSD oznaka. Budući da su pravila definirana na način da za svaki padež, svakog roda, svake vrste riječi postoji barem jedno ili više pravila, onda se uz pravilo dodala i MSD oznaka riječi. Kod nekih se riječi nisu mogli unijeti svi atributi oznake, primjerice kod imenice se nije mogao unijeti atribut koji određuje opću i vlastitu imenicu (type, common i proper) jer u pravilima nema razlike za ta dva oblika. Međutim, za skupine riječi koje pripadaju imenicama, glagolima i pridjevima dodala se većina atributa MSD oznaka. Za ostale vrste riječi dodana je POS oznaka. Smatra se da će MSD oznake pomoći u razrješavanju problema poput onih koji se pojavljuju kod riječi koje imaju isti pisani oblik, ali različiti naglasak i različitog su značenja (*róda (N jd. im. róda)* i *ròda (G jd. im. rôd)*) ili kad se radi o istoj riječi, ali različitog oblika i različitog naglaska (*sèla (G jd.)* i *sȅla (N mn.)*).

Gore opisanim postupkom, dobiven je naglasni rječnik koji se sastoji od svih oblika imenica, glagola i pridjeva generiranih uvažavajući ranije spomenuta pravila iz (Mikelić Preradović, 2008), svih oblika ostalih vrsta riječi preuzetih iz *Rječnika hrvatskoga jezika* te ručno nadodanih oblika zamjenica i brojeva.

Ukupno se u rječniku nalazi 72,366 osnovnih oblika riječi i 1.011,785 izvedenih oblika riječi + izvedeni oblici pridjeva. Svaka natuknica u rječniku sastoji se od naglašene riječi, njezine MSD/POS oznake i nenaglašenog oblika riječi. Obzirom da sličan rječnik za hrvatski ne postoji, smatra se da će nastali naglasni rječnik s respektabilnim brojem natuknica

biti od koristi u više područja računalne obrade prirodnog jezika za hrvatski jezik, a postojanje MSD oznaka uz natuknicu, dodatno povećava njegovu upotrebljivost.

Kako bi se provjerila učinkovitost dobivenog sustava za dodjeljivanje naglasaka temeljenog na pravilima, primjena pravila provela se na nenaglašenom tekstu za koji je postojao istovjetni tekst s označenim naglascima koji su riječima dodijeljene od strane eksperta. Najprije se nenaglašeni tekst rastavio na riječi, a svakoj je riječi dodana i MSD oznaka pomoću automatskog morfosintaktičkog označivača(Agić, Ljubešić, & Merkler, 2013), a kasnije su se automatski pogrešno dodijeljene oznake ispravile ručno. Budući da su se u hrvatski naglasni rječnik koji je gore opisan također dodale MSD oznake, a uz naglašeni oblik riječi postoji i nenaglašeni, provelo se pretraživanje nenaglašenog oblika riječi iz teksta zajedno s MSD oznakom unutar naglasnog rječnika. Kada se našla natuknica unutar rječnika i njezina MSD oznaka koja odgovara riječi i njezinoj MSD oznaci iz teksta, onda se nenaglašeni oblik riječi zamijenio ekvivalentnim naglašenim oblikom riječi. Na taj se način kao izlaz dobio niz riječi koji odgovara nizu riječi iz teksta, ali s označenim leksičkim naglaskom. Točnost dodjeljivanja naglaska provjerila se uspoređujući dobivene naglašene riječi s riječima u tekstu koji je označio jezični ekspert. Točnost koja se dobila nakon automatskog dodjeljivanja MSD oznaka bila je 78%. Nakon što su ručno ispravljene MSD oznake, točnost se povećala na 87,7%. Daljnjom analizom utvrđeno je da neke riječi nisu dobile pravilni naglasak zbog premještanja naglaska s naglasnice na prednaglasnicu pa su se stoga u obzir uzela pravila koja definiraju kada naglasak s naglasnice prelazi na prednaglasnicu. Nakon primjene pravila točnost se povećala na 92,8%. Većina riječi koja ni tada nije dobila pravilni naglasak, nije dobila pogrešan naglasak već naglasak nije uopće bio dodijeljen. Razlog tomu je nepostojanje riječi u rječniku.

Kako bi se doskočilo problemu nedodjeljivanja naglaska riječima zbog njihova nepostojanja u rječniku, pristupilo se izradi sustava za automatsko dodjeljivanje naglaska pomoću modela. Na taj se način postiglo da se i riječima izvan rječnika (dakle, onima koje ne postoje u rječniku), dodijeli naglasak. U postupku učenja modela korištena su klasifikacijska stabla za predviđanje mjesta i vrste naglaska u riječi, a kao skup za učenje modela korišten je ranije opisani naglasni rječnik.

Model za automatsko dodjeljivanje naglasaka se sastoji od dva dijela - prvog za predviđanje mjesta naglaska i drugog za predviđanje vrste naglaska. Za učenje prvog modela koristile su se sljedeće jezične značajke: broj slogova u riječi, fonetske osobine zadnja četiri glasa, fonetske osobine prva tri glasa u riječi i POS riječi. Za učenje drugog modela pomoću

kojeg se predviđa vrsta naglaska korištene su sljedeće jezične značajke: broj slogova u riječi, fonetske osobine zadnja tri glasa u riječi, fonetske osobine prva dva glasa u riječi, POS oznaka riječi, redni broj naglašenog sloga (dobiven na temelju predviđanja iz prvog modela), identitet naglašenog glasa, fonetske osobine glasa koji se nalazi ispred naglašenog i fonetske osobine glasa koji se nalazi iza naglašenog.

Što se tiče rezultata automatskog dodjeljivanja mjesta naglaska u riječi, nakon učenja modela za predviđanje najvjerojatnijeg mjesta naglaska u riječi na podacima iz rječnika, postupkom deseterostruke unakrsne validacije ostvarena je točnost modela od 90,56%. Za model predviđanja najvjerojatnije vrste naglaska na riječima istim postupkom se dobila točnost od 86,02%.

Modeli su se primijenili i na tekstu koji nije korišten za učenje modela, a to je isti tekst koji je korišten i za evaluaciju automatskog dodjeljivanja naglaska pomoću pravila. U tom je slučaju model točno predvidio mjesto naglaska s točnošću od 95,99%, vrstu naglaska s točnošću od 74,26%, a i mjesto i vrstu naglaska s točnošću od 72,15%. Ako se na prednaglasice i zanaglasnice primijene pravila o njihovom nenaglašavanju, odnosno prelazak naglaska s naglasnice na prednaglasnicu, onda se dobivaju sljedeće točnosti: za mjesto naglaska 97,4%, za vrstu naglaska 82,4% , za točno i mjesto i vrstu 80,1%.

Na kraju se još proveo pristup dodjeljivanja naglaska koji kombinira naglašavanje korištenjem pravila i modela naglašavanja. U tom se postupku, nakon što se nad riječima u tekstu provedu pravila, za riječi koje nakon postupka naglašavanja ostanu nenaglašene se dodatno provede i postupak naglašavanja pomoću modela za naglašavanje. Time se točnost automatskog naglašavanja povećala na 95,3%.

Smatra se da je navedena točnost sustava za automatsko naglašavanje prilično visoka te da se sustav može koristiti za automatsko dodjeljivanje naglasaka i koristiti u raznim područjima računalne obrade prirodnog jezika kao što su sinteza i raspoznavanje govora, kod sustava za strojno potpomognuto prevođenje, sustava za računalno potpomognuto učenje jezika i sl.

U sklopu ove doktorske disertacije provela se analiza trajanja slogova na korpusu koji po tematici pripada domeni bajki i priča, a dio je govorne baze hrvatskoga jezika VEPRAD (Martinčić-Ipšić & Ipšić, 2003). Domena bajki i priča odabrala se zbog eksplicitnije prozodije koja je prisutna u takvim korpusima u odnosima na primjerice govorne korpuse koji po

tematici pripadaju vremenskim prognozama ili vijestima. U svrhu rastavljanja transkribiranih riječi iz korpusa na slogove, koristio se algoritam za slogovanje prema načelu najvećega pristupa(Meštrović, Martinčić-Ipšić, & Matešić, 2015). Budući da se želio provjeriti utjecaj vrste naglaska na trajanje, datoteke s tekstualnim transkripcijama govornog korpusa provedene su i kroz postupak automatskog dodjeljivanja naglaska opisanog u ovom radu. U bazi VEPRAD postoji automatska segmentacija riječi na razini glasova u obliku .lab datoteka s oznakama izgovorenih glasova i vremenskim trenucima početka i kraja svakog glasa izraženih u milisekundama. Osim trajanja glasova, na temelju tih podataka dobiveno je i trajanje slogova te izrađena analiza trajanja najfrekventnijih slogova ovisno o njihovim fonološkim značajkama, položaju unutar riječi i rečenice i kontekstu, kako bi se utvrdile najznačajnije značajke koje utječu na trajanje slogova.

Analizom trajanja slogova, uzimajući u obzir njihove položaje unutar riječi i rečenice, pokazalo se da se trajanja slogova u prosjeku povećavaju u odnosu na referentna trajanja ukoliko se nađu na početku ili kraju riječi ili početku ili kraju rečenice. U prosjeku su se trajanja slogova povećala za 41,4% u odnosu na referentne vrijednosti ukoliko su se slogovi našli na početku riječi. Trajanje slogova, ukoliko se nalaze na kraju riječi, u prosjeku se produljilo za 37,0% u odnosu na referentna trajanja. Ukoliko se u obzir uzmu položaji slogova unutar rečenice, slogovi se još više produljuju u odnosu na referentne vrijednosti nego kad se u obzir uzmu položaji slogova unutar riječi. Prosječno trajanje slogova koji se nalaze na početnom mjestu u rečenici produljilo se za 71,8% u odnosu na referentno trajanje. Svim slogovima, bez iznimke, se produljilo trajanje ukoliko su se našli na zadnjem mjestu u rečenici, a prosječno se trajanje u takvim slučajevima produljilo za 104,75% u odnosu na referentno trajanje.

Na temelju rezultata zaključuje se da položaj sloga unutar riječi i rečenice ima utjecaj na trajanje sloga. Najviše na trajanje utječe završni položaj sloga u rečenici koji prosječno povećava trajanje sloga za nešto više od 100%. Početni položaj sloga u rečenici prosječno produljuje njegovo trajanje za oko 70%, a u malo manjim omjerima trajanje sloga produljuju i početno mjesto sloga u riječi te završno mjesto sloga u riječi - prosječno za oko 40%.

Analizom kontekstualnih značajki željelo se provjeriti ima li vrsta glasa kojim započinje sljedeći slog utjecaj na trajanje sloga. Na temelju rezultata zaključilo se da se trajanje početnih slogova u riječi produljuje u različitom omjeru u odnosu na referentno trajanje ovisno o tome kojoj skupini suglasnika pripada glas kojim započinje sljedeći slog. Trajanje sloga se najviše produljuje ako iza njega slijedi slog koji započinje glasom iz skupine

tjesnačnika - za 60,4%, a zatim približnika - za 44,9%. Ukoliko sljedeći slog započinje glasom iz skupine nosnika, zapornika ili slivenika, početni se slog u prosjeku produljuje za oko 30%. Najmanje utjecaja na trajanje prethodnog sloga ima slog koji započinje glasom iz skupine bočnika te prosječno produljenje tada iznosi 16,2%.

Jedan od ciljeva ovog rada bio je primijeniti modele trajanja i intonacije na hrvatski jezik. Za modeliranje trajanja primijenili su se modeli linearne regresije, stabla linearnih regresijskih modela izgrađena korištenjem algoritma M5 i metoda potpornih vektora. Modeli trajanja slogova izgrađeni su na temelju podataka o trajanju slogova prirodnog govora iz baze VEPRAD, podskupa korpusa bajki i priča. Iz tekstualnih transkripcija govora izlučene su jezične značajke, a iz segmentacije na razini glasa s odgovarajućim trajanjima, dobivena su trajanja slogova. Jezične značajke koje su se koristile za učenje modela trajanja su indeks sloga koji predstavlja broj slogova koji se nalazi ispred trenutnog sloga u rečenici/frazi, broj slogova koji se nalaze iza trenutnog sloga u rečenici/frazi, položaj sloga unutar riječi - početak, sredina ili kraj, položaj riječi unutar fraze - početak, sredina ili kraj, broj riječi koje se nalaze ispred i iza trenutne riječi, ukupan broj slogova u rečenici/frazi, ukupan broj slogova u rečenici/frazi, dužina sloga, vrsta sloga obzirom na raspored glasova u slogu, identitet samoglasnika u slogu, vrsta naglaska u slogu, vrsta naglaska u prethodnom i sljedećem slogu, identitet prvog glasa u sljedećem slogu te identitet zadnjeg glasa u prethodnom slogu. Kako bi se provjerio utjecaj položajnih i kontekstualnih značajki te prisutnost i vrsta naglaska na točnost modela za trajanje, značajke su se podijelile u četiri skupine: one koje se odnose na položaj segmenata, one koje se odnose na kontekst, one koje se odnose na naglasak te skupina sa svim navedenim značajkama.

Na temelju rezultata zaključuje se da uključivanje svake od skupine značajki u skup značajki na kojima se model uči, doprinosi većoj točnosti modela trajanja. Pritom najviše utjecaja imaju položajne značajke, ali pokazalo se da točnost modela povećavaju i kontekstualne značajke i značajke koje se odnose na naglasak u hrvatskom jeziku.

Sveukupni rezultati usporedivi su s rezultatima modela trajanja pomoću slogova bez post-obrade kao što je primjerice vrijednost RMSE od 32 za model koji je korišten u (Rao, 2012) a za koji je korišten malo veći korpus. Međutim, u budućem radu će se svakako pokušati dobiti veća točnost modela.

Za primjenu intonacijskog modela na hrvatskom jeziku koristio se Tilt intonacijski model. Tilt (Taylor, 2000) je fonetski intonacijski model koji F0 konturu prikazuje kao slijed kontinuirano parametriziranih događaja. Osnovna jedinica Tilt modela je intonacijski događaj.

Osnovni tipovi događaja su naglasak i granični tonovi. Model se temelji na RFC modelu koji svaki intonacijski događaj opisuje pomoću rastućeg oblika, padajućeg ili rastućeg kojeg slijedi padajući. Zbog lakše interpretacije i korištenja, RFC parametri se zatim pretvaraju u 3 tilt parametra - trajanje, amplituda i tilt parametar. Tilt model odabran je za primjenu na hrvatski jezik jer je označavanje intonacijskih događaja kod ovog modela relativno jednostavno u odnosu na označavanje događaja za neke druge intonacijske modele.

Za hrvatski jezik dosad nije postojao korpus s označenim intonacijskim događajima. U sklopu ovog rada nastao je korpus od 500 iskaza koji se sastoji od 500 govornih signala s odgovarajućim pridruženim tilt oznakama. Kod postupka ručnog označavanja, koristile su se oznake za oznake vrha događaja *a*, početka rasta događaja *b*, početak pada događaja *c*, granični ton *g* te oznaka tišine *sil*. U postupku označavanja, uz govorni signal mogla se pratiti i transkripcija riječi kojima je dodijeljen naglasak. Na taj se način moglo pratiti koji je slog u riječi naglašen te je tako olakšan postupak dodjeljivanja oznaka za intonacijske događaje u Tilt modelu. Za izlučivanje F0 značajki korišten je algoritam RAPT (Talkin, 1995). F0 kontura uzorkovana je svakih 10 ms. Kao izlaz iz algoritma RAPT dobila su se tri različita skupa F0 značajki - izravan izlaz iz RAPT algoritma, zaglađene vrijednosti i interpolirane vrijednosti. Kako bi se ograničili valjani sljedovi događaja, definirala se i gramatika s dozvoljenim sljedovima događaja. Za pronalaženje mjesta vrha događaja, početka rasta, početka pada i mjesta gdje se nalazila tišina naučili su se skriveni Markovljevi modeli.

Naposljetku su se pomoću alata Tilt analiza i Tilt sinteza sintetizirale izgenerirane F0 konture te se izgenerirana F0 kontura usporedila s originalnom kako bi se provjerila točnost modela. Mjera koja se koristila za usporedbu izgenerirane i originalne F0 konture je RMSE, a najbolja dobivena vrijednost RMSE je 22,2. Rezultati su bolji nego u preliminarnom istraživanju koje je provedeno u (Načinović, Pobar, Martinčić-Ipšić, & Ipšić, 2011) kada se za učenje modela koristio skup od 100 rečenica označenih tilt oznakama (prijašnji RMSE rezultat je bio 25,16). Dakle, povećanjem skupa rečenica za učenje, povećala se i točnost modela.

10. Zaključak

Zbog brojnih uloga koje prozodija ima u komunikaciji, njezino pravilno tumačenje važno je za mnoge postupke u području računalne obrade jezika. U ovom se radu željelo ispitati može li se za hrvatski jezik izraditi sustav za automatsko dodjeljivanje naglasaka koji je jedan od najvažnijih prozodijskih sredstava te mogu li se model trajanja i intonacijski model primijeniti na hrvatski jezik. Temeljem provedenog istraživanja, na sve se pitanja može odgovoriti potvrdno, tj. izradio se sustav za automatsko dodjeljivanje naglasaka te su se modeli trajanja i intonacijski model uspješno primijenili na hrvatski jezik.

Cilj ovog rada bio je istražiti primjenjivost metoda predviđanja prozodijskih obilježja hrvatskoga jezika i njihovog modeliranja za hrvatski jezik te istražiti mogućnosti njihovog poboljšanja uzimajući u obzir lingvističke značajke i jezične specifičnosti hrvatskoga jezika. Glavne hipoteze od kojih se krenulo u izradu ovog doktorskog rada su:

a) H1: moguće je stvoriti naglasni morfološki analizator/generator koji riječima automatski dodjeljuje naglasak na temelju postojećih modela,

b) H2: prozodijski modeli (modeli trajanja i putanje F0) za hrvatski jezik mogu se izmodelirati na temelju jezičnih značajki teksta, a uključivanje različitih skupina jezičnih značajki u prozodijske modele utječe na točnost prozodijskih modela,

c) H3: jezične značajke specifične za hrvatski jezik poput mjesta i vrste leksičkoga naglaska dodatno utječu na točnost prozodijskih modela za hrvatski jezik.

Prva je hipoteza potvrđena implementacijom pravila za izvođenje izvedenih oblika riječi iz osnovnih, gdje se dodavanjem odgovarajućeg paradigmatskog nastavka i premještanjem naglaska dobio sustav za automatsko dodjeljivanje naglasaka riječima iz teksta. Implementacijom pravila dobio se i naglasni rječnik čije se natuknice sastoje od naglašenog oblika riječi, nenaglašenog oblika i MSD odnosno POS oznake. Obzirom da je hrvatski jezik jedan od jezika za koje nema puno dostupnih jezičnih resursa, smatra se da će značaj ovog rječnika biti veliki u njegovoj primjeni u području računalne obrade hrvatskog jezika. Rječnik sa osnovnim i izvedenim oblicima hrvatskoga jezika u kojima natuknice imaju pridružen leksički naglasak dosad nije postojao. Uz to, najbrojnijim skupinama promjenjivih vrsta riječi (imenicama, glagolima i pridjevima) dodala se i MSD oznaka. Iako se neki atributi nisu mogli dodati MSD oznaci jer se pomoću pravila na temelju kojih je nastao rječnik, oni nisu mogli odrediti, ipak se smatra da će biti korisni za upotrebu u dobivanju boljih rezultata u različitim područjima jezičnih tehnologija. Budući da rječnik sadrži veliki broj oblika imenica, glagola i pridjeva kojima je dodana (nepotpuna) MSD oznaka, može se upotrijebiti primjerice za dobivanje boljih rezultata postojećeg sustava za automatsko pridruživanje MSD oznake riječima. Rječnik ukupno sadrži 72,366 osnovnih oblika riječi i preko 1.000,00 izvedenih oblika riječi.

Točnost dobivenog sustava za automatsko dodjeljivanje naglaska nastalog na temelju pravila za izvođenje izvedenih, naglašenih oblika riječi iz osnovnih, testirana je uspoređujući rezultat naglašavanja s tekstom u kojem riječi imaju dodijeljeni naglasak od strane eksperta. Rezultati naglašavanja pomoću sustava temeljenog na pravilima su vrlo dobri s točnošću od 78% ukoliko se za dodjeljivanje MSD oznaka riječima u tekstu koristio automatski sustav, a 87,7% ukoliko su se pogreške kod automatski dodijeljenih MSD oznaka ispravile ručno. Nakon primjene pravila o nenaglasnicama odnosno premještanju naglaska s nenaglasnice na prednaglasnicu, točnost se povećala na 92,8%. Navedena točnost modela smatra se primjerenom za korištenje u svrhu utvrđivanja mjesta i vrste naglaska u raznim područjima računalne obrade jezika i jezičnih tehnologija.

Nekim riječima nakon navedenog postupka nisu se dodijelili naglasci jer se riječi nisu našle u naglasnom rječniku. Stoga je za takve slučajeve predložen sustav za automatsko dodjeljivanje naglaska pomoću modela.

Model se učio na podacima iz naglasnog rječnika te je njegova primjena također testirana na istom tekstu kao i sustav za naglašavanje temeljen na pravilima. Model je dostigao točnost od 97,4% točnost za određivanje mjesta naglaska, 82,4% za određivanje vrste naglaska, te 80,1% točnost za oboje - i mjesto i vrstu naglaska. Postupkom deseterostruke unakrsne validacije točnost modela za određivanje mjesta naglaska bila je 90,56%, a za model predviđanja najvjerojatnije vrste naglaska na riječima istim postupkom se dobila točnost od 86,02%.

Postignuta točnost sustava za dodjeljivanja naglaska temeljenog na modelu manja je nego točnost sustava za dodjeljivanje naglaska temeljenog na pravilima. Međutim, obzirom da se nekim riječima nije dodijelio naglasak nakon primjene sustava temeljenog na pravilima, došlo se na ideju da se takvim riječima naglasak dodijeli pomoću sustava temeljenog na modelu. Na taj se način dobio hibridni pristup dodjeljivanja naglaska koji kombinira pristup temeljen na pravilima i pristup temeljen na modelu. Uz to se u navedenim postupcima implementiraju i pravila za premještanje naglaska s naglasnice na prednaglasnice. Postignuta točnost određivanja mjesta i vrste naglaska nakon primjene ovakvog pristupa je 95,3%. Ovim rezultatima dokazana je prva hipoteza (pa i treća ukoliko se naglasni sustav uzme kao dio prozodijskog sustava.

Analizom trajanja hrvatskih slogova pokušao se utvrditi utjecaj kontekstualnih i položajnih značajki na trajanje slogova. Analizom je utvrđeno da položaj sloga unutar riječi i rečenice utječe na trajanje sloga. U prosjeku su se trajanja slogova povećala za 41,4% u odnosu na referentne vrijednosti ukoliko su se našli na početku riječi. Trajanje slogova ukoliko se nalaze na kraju riječi u prosjeku se produljilo za 37,0%. Prosječno trajanje slogova koji se nalaze na početnom mjestu u rečenici produljilo se za 71,8% u odnosu na referentno trajanje, a prosječno trajanje se produljilo za 104,75% u odnosu na referentno trajanje ukoliko se slog našao na posljednjem mjestu u rečenici. Analizom se također utvrdilo da kontekstualne značajke imaju utjecaj na trajanje sloga. Konkretno, provjerilo se koliko vrsta suglasnika kojima počinje slog koji slijedi nakon sloga za koji se izračunava trajanje utječe na trajanje tog sloga. Zaključilo se da se trajanje sloga najviše produljuje ako iza njega slijedi slog koji započinje glasom iz skupine tjesnačnika - za 60,4%, a zatim približnika - za 44,9%. Ukoliko sljedeći slog započinje glasom iz skupine nosnika, zapornika ili slivenika, početni se slog u prosjeku produljuje za oko 30%. Najmanje utjecaja na trajanje prethodnog sloga ima slog koji započinje glasom iz skupine bočnika te prosječno produljenje tada iznosi 16,2%.

Prilikom modeliranja trajanja slogova za hrvatski jezik, jezične su se značajke razvrstale u tri grupe: položajne, kontekstualne i one koje se odnose na naglasak u hrvatskom jeziku. Najprije se izračunala točnost modela (u vidu vrijednosti RMSE) ukoliko su se prilikom učenja modela u obzir uzele sve skupine značajki, a zatim su se iz skupine značajki na kojemu se učio model izuzele jedna po jedna skupina značajki i svaki put izračunala točnost modela kako bi se utvrdio utjecaj pojedinih skupina značajki na točnost modela trajanja. Utvrđeno je da svaka skupina značajki - položajne, kontekstualne i one koje se odnose na naglasak u hrvatskim riječima u određenom omjeru utječe na točnost modela i to na način da se točnost smanji ukoliko se pojedina skupina značajki izuzme iz skupa za učenje modela. Najviše utjecaja od navedene tri skupine imaju položajne značajke. Postignuta vrijednost RMSE modela trajanja od 35,3 usporediva je sa modelom trajanja slogova drugih jezika, ali mjesta za poboljšanje svakako ima. U budućem će se radu pokušati u značajke uvrstiti i fonetske značajke glasova u slogu.

Iz gore opisanog zaključujemo da je dokazana i hipoteza H2, tj. da se prozodijski modeli (modeli trajanja i putanje F0) za hrvatski jezik mogu izmodelirati na temelju jezičnih značajki teksta, a uključivanje različitih skupina jezičnih značajki u prozodijske modele utječe na točnost prozodijskih modela.

Za modeliranje intonacije hrvatskoga jezika odabran je Tilt intonacijski model. Kako bi se model mogao primijeniti na hrvatski jezik, bilo je potrebno govorni korpus nadopuniti tilt oznakama. Tako je nastao ručno označen govorni korpus od 500 iskaza kojima su pridružene tilt oznake. Dosad za hrvatski jezik nije postojao korpus s oznakama za neki od postojećih intonacijskih modela, osim mali probni korpus od 100 iskaza s tilt oznakama koji se koristio u istraživanju: (Načinović, Pobar, Martinčić-Ipšić, & Ipšić, 2011). Tilt model temelji se na principu pronalaženja intonacijskih događaja koji u osnovi odgovaraju leksičkim naglascima u hrvatskom jeziku. Pa se u tom vidu pomoću Tilt modela može zaključiti može li se označavanjem lingvističkih događaja koji odgovaraju naglasku u hrvatskom jeziku, dobiti intonacijski model zadovoljavajuće kakvoće. Uz naglasak, drugi događaj su granični tonovi koji primjerice mogu označavati nedovršenost rečenice, upitnost i slično. U skup oznaka kojima su se označili tilt događaji za hrvatski jezik ubrajaju se oznaka vrha događaja *a*, za početka rasta događaja *b*, za početak pada događaja *c*, granični ton *g* te oznaka tišine *sil*. Za izlučivanje F0 značajki korišten je algoritam RAPT. Za pronalaženje mjesta vrha događaja, početka rasta, početka pada i mjesta gdje se nalazila tišina naučili su se skriveni Markovljevi modeli. Za generiranje F0 konture koristili su se alati Tilt analiza i Tilt sinteza. Izgenerirana

F0 kontura se usporedila s originalnom kako bi se provjerila točnost modela. Mjera koja se koristila za usporedbu izgenerirane i originalne F0 konture je RMSE, a najbolja dobivena vrijednost RMSE je 22,2 koja je zadovoljavajuća. U usporedbi s RMSE vrijednošću koja se dobila u istraživanju provedenom u (Načinović, Pobar, Martinčić-Ipšić, & Ipšić, 2011), dobio se nešto bolji rezultat. Već je prije napomenuto da se u navedenom istraživanju koristio manji korpus s tilt oznakama koji je sadržavao samo 100 iskaza pa je razumljivo da je povećanje broja iskaza u ručno označenom korpusu polučilo bolji uspjeh.

Gore navedeni rezultati potvrđuju hipotezu H3, tj. da jezične značajke specifične za hrvatski jezik poput mjesta i vrste leksičkoga naglaska dodatno utječu na točnost prozodijskih modela za hrvatski jezik.

Prema svemu navedenom, originalni doprinosi ovog rada su:

- naglasni rječnik sa 72,366 osnovnih oblika riječi i više od 1.000,000 izvedenih oblika riječi s označenim naglaskom i pridruženom (nepotpunom) MSD ili POS oznakom,
- sustav za automatsko dodjeljivanje naglaska na temelju pravila,
- sustav za automatsko dodjeljivanje naglaska na temelju modela,
- sustav za automatsko dodjeljivanje naglaska na temelju pristupa koji kombinira naglašavanje pomoću pravila i naglašavanje pomoću modela te pravila za premještanje naglaska s naglasnice na prednaglasnice
- model trajanja slogova za hrvatski,
- Tilt intonacijski model za hrvatski,
- korpus koji se sastoji od 500 iskaza, a kojima su uz govorni signal pridružene tilt oznake događaja.

Budući rad iz ovog područja usmjerit će se na dobivanje još boljih rezultata u sustavima za automatsko dodjeljivanje naglasaka te boljih rezultata u primjeni prozodijskih modela za hrvatski jezik.

Pomoću naglasnog rječnika koji je nastao u ovom radu može se riješiti problem dvosmislenosti (višeznačnosti) kod riječi koje imaju isti pisani oblik, ali različiti naglasak i različitog su značenja kao primjerice u *róda (N jd. im. róda)* i *ròda (G jd. im. rôd)* te kod istih riječi, ali različitog oblika i različitog naglaska kao primjerice u *sèla (G jd.)* i *sȅla (N mn.)*. U

budućem radu nastojat će se identificirati i parovi riječi koje imaju isti pisani oblik, a različiti naglasak kao primjerice u *grȁd (tuča)* i *grâd (naselje)* te uz pomoć kolokacija razriješiti problem višeznačnosti kod takvih slučajeva.

Od budućih planova vezanih uz poglavlja 5 i 6 planira se još dodatno poboljšati točnost automatskog dodjeljivanja naglaska na način da se naglasni rječnik nadopuni svim oblicima riječi koje imaju vlastitu paradigmu za sklanjanje odnosno konjugiranje, popisom osnovnih i izvedenih oblika imena i zemljopisnih imena te proširivanjem pravila za izvođenje izvedenih oblika riječi iz osnovne.

Budući da je točnost sustava za automatsko dodjeljivanje naglaska vrlo visoka (95,3%) te može biti korisna u raznim područjima obrade prirodnog jezika, u budućnosti se planira razviti web aplikacija koja bi omogućila korisnicima da kao ulaz predaju nenaglašeni tekst, a kao izlaz dobiju naglašeni tekst.

Budući rad vezan uz prozodijske modele bit će usmjeren na dobivanje modela primijenjenih na hrvatski jezik s većom točnošću. Pokušat će se izraditi i model trajanja na manjim segmentima od slogova, tj. glasovima kako bi se provjerilo mogu li se na taj način dobiti bolji rezultati.

Tehnologije računalne obrade prirodnog jezika i govora koriste različite jezične resurse koji su najčešće dostupni za jezike s velikim brojem govornika, dok su za jezike s manjim brojem govornika, kao što je hrvatski, ti resursi ograničeni.

Dizajn ovog istraživanja te sustavi i alati koji su razvijeni za hrvatski jezik predstavljaju doprinos računalnoj leksikografiji i računalnoj obradi prirodnoga jezika i govora. Dobiveni sustav za automatsko dodjeljivanje naglaska i prozodijski modeli, moći će se u budućim istraživanjima uključiti u sintezu hrvatskoga govora s ciljem dobivanja prirodnijeg i razumljivijeg sintetiziranog govora, kod automatskog raspoznavanja govora za postizanje boljih rezultata, u području automatske identifikacije govornika i jezika, kod sustava za strojno potpomognuto prevođenje, u sustavima za računalno potpomognuto učenje jezika te kod prepoznavanja emocionalnog stanja u komunikaciji i sličnim područjima u obradi prirodnog jezika.

Popis literature

1. Agić, Ž., Ljubešić, N., & Merkler, D. (2013). Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing (BSNLP 2013)* (str. 48-57). Sofia, Bulgaria: Association for Computational Linguistics.
2. Allen, J., Hunnicut, S., & Klatt, D. (1987). *Text-to-Speech: The MITalk System.* Cambridge: Cambridge University Press.
3. Alumäe, T. (2014). Neural network phone duration model for speech recognition. *Interspeech 2014,* (str. 1204-1208). Singapore.
4. Anić, V. (2009). *Veliki rječnik hrvatskoga jezika.* Novi liber.
5. Babić, S., Brozović, D., Moguš, M., Pavešić, S., Škarić, I., & Težak, S. (1991). *Povijesni pregled, glasovi i oblici hrvatskoga književnog jezika.* Zagreb: Globus, Nakladni zavod.
6. Bakran, J., & Horga, D. (1996). SAMPA za hrvatski. *Govor* , str. 99-104.
7. Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečević, V., i dr. (1995.). *Hrvatska gramatika.* Zagreb: Školoska knjiga.

8. Bartkova, K., & Sorin, C. (1987). A model of segmental duration forspeech synthesis in French. *Speech Communication* , str. 245-260.

9. Beliga, S., & Martinčić-Ipšić, S. (2011). Text Normalization for Croatian Speech Synthesis. *MIPRO 2011,* (str. 382-387). Rijeka.

10. Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer*. Preuzeto 4 2016 iz www.praat.org

11. Campbell, W. N. (1992). Syllable-based segmental durations. *Talking Machines: Theories, Models, and Designs* , str. 43-60.

12. Ciobanu, A. M., Dinu, A., & Dinu, P. L. (2014). Predicting Romanian Stress Assignment. *EACL*, (str. 64-68).

13. Cukor, E. (2009). Naglasno-morfološki generator za glagole: završni rad. Filozofski fakultet u Zagrebu.

14. de Cordoba, R., Montero, J. M., Gutierrez-Arriola, J. M., & Pardo, J. M. (2001). Duration modeling in a restricted-domain female-voice synthesis in Spanish using neural networks. *ICASSP 2001*, (str. 793-796).

15. Dutoit, T. (1997). *An introduction to text-to-speech synthesis.* Springer.

16. Erjavec, T., Krstev, C., Petkevič, V., Simov, K., Tadić, M., & Vitas, D. (2003). The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. *Proceedings of the EACL 2003 Workshop on the Morphological Processing of Slavic Languages*, (str. 25-32). Budimpešta.

17. Flege, J. E., & Brown, W. S. (1982). Effects of Utterance Position on English Speech Timing. *Phonetica 39* , str. 337-357.

18. Fujisaki, H., & Ohno, S. (1995). Analysis and modeling of fundamental frequency contours of English utterances. *Proceedings Eurospeech 95*, (str. 985-988). Madrid.

19. Fujisaki, H., & Ohno, S. (2005). Analysis and Modeling of Fundamental Frequency Contours of English Utterances. *Speech Communication*, *47*, str. 59-70.

20. Fujisaki, H., Hirose, K., Halle, P., & Lei, H. (1971). A generative model for the prosody of connected speech in Japanese. *Ann. Rep. Engineering Research Institute 30*, str. 75-80.

21. Fujisaki, H., Ohno, S., & Takashi, Y. (1997). Analysis and modeling of fundamental frequency contours of Greek utterances. *Eurospeech 97*, (str. 465-468). Rhodes, Greece.

22. Hirschberg, J. (1995). Pitch Accent in Context: Predicting Intonational Prominence from Text. *Artificial Intelligence, 3*, str. 305-340.

23. Hirst, D. (2001). Automatic analysis of prosody for multilingual speech corpora. U E. Keller, G. Baily, A. Monaghan, J. Terken, & M. Huckvale, *Improvements in Speech Synthesis.* Wiley.

24. Hirst, D. (2005). Form and function in the representation of speech prosody. *Quantitative prosody modeling for natural speech description and generation*, (str. 334-347). Beijing.

25. Horvacki, M. (2009). Naglasno-morfološki generator za glagole. Filozofski fakultet u Zagrebu.

26. Jelaska, Z. (2004). *Fonološki opisi hrvatskoga jezika : glasovi, slogovi, naglasci.* Zagreb: Hrvatska sveučilišna naklada.

27. Kato, H., Tsuzaki, M., & Sagisaka, Y. (1998). Acceptability for Temporal Modification of Single Vowel Segments in Isolated Words. *J. Acoust. Soc. Am.* , str. 540-549.

28. Krishna, S. N., Talukdar, P. P., Bali, K., & Ramakrishnan, A. G. (2004). Duration Modeling for Hindi Text-to-Speech Synthesis System. *Proceedings of ICSLP'04*, (str. 789-792). Jeju Island, Korea.

29. Lazić, N. (2006). Modeliranje strojnih postupaka za izgovaranje teksta pisanoga hrvatskim jezikom: doktorska disertacija. Filozofski fakultet u Zagrebu.

30. Louw, J. A., & Bernard, E. (2004). Automatic intonation modeling with INTSINT. *Proceedings of the Pattern Recognition Association of South Africa*, (str. 107-111). Grabouw.

31. Ljubešić, N. (29. 4 2013). *MULTEXT-East Morphosyntactic Specifications, revised Version 4; Croatian Specifications*. Preuzeto 2 2016 iz http://nlp.ffzg.hr/data/tagging/msd-hr.html

32. Marinčič, D., Tušar, T., Gams, M., & Šef, T. (2009). Analysis of Automatic Stress Assignment in Slovene. *Informatica* , str. 35-50.

33. Martinčić-Ipšić, S., & Ipšić, I. (2003). Veprad: a Croatian speech database of weather forecasts. *Information Technology Interfaces ITI 2003* , str. 321-326.

34. Martinčić-Ipšić, S., Matešić, M., & Ipšić, I. (2004). Korpus hrvatskoga govora. *Govor* , str. 135-150.

35. Martinčić-Ipšić, S., Pobar, M., & Ipšić, I. (2011). Croatian Large Vocabulary Automatic Speech Recognition. *Automatika* , str. 147-157.

36. Meštrović, A., Martinčić-Ipšić, S., & Matešić, M. (2015). Postupak automatskoga slogovanja temeljem načela najvećega pristupa i statistika slogova za hrvatski jezik. *Govor* .

37. Mikelić Preradović, N. (2008). Pristupi izradi strojnog tezaurusa za hrvatski jezik: doktorska disertacija. Filozofski fakultet u Zagrebu.

38. Mixdorff, H., & Fujisaki, H. (1994). Analysis of voice fundamental frequency contours of German utterances using a quantitative model. *ICSLP'94*, (str. 2231-2234). Yokohama.

39. Mobius, B., & van Santen, J. (1996). Modeling Segmental Duration in German Text-to-Speech Synthesis. *ICSLP'96* (str. 2395-2398). Philadelphia: IEEE.

40. Načinović, L. (2008). Grafemsko-fonemska pretvorba za hrvatsku sintezu govora; Diplomski rad. Odsjek za informatiku, Filozofski fakultet u Rijeci.

41. Načinović, L., Pobar, M., Martinčić-Ipšić, S., & Ipšić, I. (2011). Automatic Intonation Event Detection Using Tilt Model for Croatian Speech Synthesis. *INFuture2011, The Future of Information Sciences* (str. 383-391). Zagreb: Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb.

42. Načinović, L., Pobar, M., Martinčić-Ipšić, S., & Ipšić, I. (2009). Grapheme-to-Phoneme Conversion for Croatian Speech Synthesis. *MIPRO 2009.* Rijeka: Croatian Society for Information and Communication Technology, Electronics and Microeletronics - MIPRO.

43. Norkevičius, G., & Raškinis, G. (2008). Modeling Phone Duration of Lithuanian by. *Informatica* , str. 271-284.

44. Ostendorf, M., & Veileux, N. (1994). A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location. *Computational Linguistics, 20*, str. 27-54.

45. Pierrehumbert, J. B. (1981). Synthesizing intonation. *J. Acoust. Soc. Am.* , str. 985-995.

46. Pletikos, E. (25. 2 2008). Akustički opis hrvatske prozodije riječi : doktorska disertacija. Zagreb: Filozofski fakultet u Zagrebu.

47. Pletikos, E. (2003). Akustički opis hrvatskih standardnih naglasaka. *Govor* , str. 321-345.

48. Pobar, M. (2014). Sinteza hrvatskoga govora utemeljena na odabiru jedinica i stohastičkim modelima: doktorska disertacija. Fakultet elektrotehnike i računarstva, Sveučilište u Zagrebu.

49. Ramesh Bonda, V. S., & Girija, P. N. (2015). Duration Modeling For Telugu Language with. *International Journal of Innovative Research in Computer* , str. 720-725.

50. Rao, K. S. (2012). *Predicting Prosody from Text for Text-to-Speech Synthesis.* New York: Springer.

51. Rojc, M., Aguero, P. D., Bonafonte, A., & Kacic, Z. (2005). Training the tilt intonation model using the JEMA methodology. *Eurospeech*, (str. 3273-3276). Lisbon, Portugal.

52. Romportl, J., & Kala, J. (2007). Prosody Modelling in Czech Text-to-Speech Synthesis. *Proceedings of the 6th ISCA Workshop on Speech Synthesis*, (str. 200-205). Bonn.

53. Ross, K., & Ostendorf, M. (1996). Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language, 10*, str. 155-185.

54. Sečujski, M. (2002). Akcenatski rečnik srpskog jezika namenjen sintezi govora na osnovu teksta. *DOGS2002* .

55. Shih, C., & Ao, B. (1997). Duration Study for the Bell Laboratories Mandarin Text-to-Speech System. *Progress in Speech Synthesis*, (str. 383-399).

56. Shreekanth, T., Udayashankara, V., & Chandrika, M. (2015). Duration Modelling Using Neural Networks for Hindi TTS System Considering Position of Syllable in a

Word. *Proceedings of the International Conference on Information and Communication Technologies* (str. 60-67). Elsevier.

57. Silverman, K. M., Beckham, M., Pitrelli, J., Ostendorf, M., Pierrehumbert, J., Hirschberg, J., i dr. (1992). TOBI: A Standard Scheme for Labeling Prosody. Banff: Proceedings of the International Conference on Spoken Language 92.
58. Simoes, A. R. (1990). Predicting sound segment duration in connected speech: an acoustical study of brazilian portuguese. *SSW1-1990* , str. 173-176.
59. Smola, A., & Scholkopf, B. (1998). *A Tutorial on Support Vector Regression.* Technical report Neuro COLT NC-TR-98-030.
60. Sproat, R. (1997). *Multilingual Text-to-Speech Synthesis.* Berlin: Springer.
61. Stergar, J., & Erdem, C. (2010). Adapting Prosody in a Text-to-Speech System. *Products and Services; from R&D to Final Solutions* .
62. Šef, T. (2006). Automatic Accentuation of Words for Slovenian TTS System. *Proceedings of the 5th WSEAS International Conference on Signal Processing*, (str. 155-160). 2006.
63. Šef, T., & Gams, M. (2003). SPEAKER (GOVOREC): A Complete Slovenian Text-to Speech System. *INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6* , str. 277-287.
64. Škarić, I. (1991). Fonetika hrvatskoga književnog jezika. U S. Babić, D. Brozović, M. Moguš, S. Pavešić, I. Škarić, & S. Težak, *Povijesni pregled, glasovi i oblici hrvatskoga književnoga jezika: nacrti za gramatiku* (str. 71-378). Zagreb: Globus.
65. Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). *Speech coding and synthesis* , str. 495-518.
66. Taylor, P. (2000). Analysis and Synthesis of Intonation using the Tilt Model. *Journal of the Acoustical Society of America* , str. 1697-1714.
67. Taylor, P. (2009). *Text-to-Speech Synthesis.* Cambridge: Cambridge University Press.
68. Taylor, P. (1995). The rise/fall/connection model of intonation. *Speech Communication* , str. 169-186.
69. Taylor, P., & Black, A. B. (1998). Assigning Phrase Breaks from Part-of-Speech Sequences. *Computer Speech and Language* , str. 99-117.

70. Thangthai, A., Thatphithakkul, N., Wutiwiwatchai, C., Saychum, S., & Rugchatjaroen, A. (2008). T-Tilt: A modified Tilt model for F0 analysis and synthesis in tonal languages. *INTERSPEECH 2008.* Brisbane, Australia.

71. Theune, M., Meijs, K., Heylen, D. K., & Ordelman, R. J. (2006). Generating Expressive Speech for Storytelling Applications. *IEEE transactions on audio, speech and language processing* , str. 1137-1144.

72. Tihelka, D., Kala, J., & Mtousek, J. (2010). Enhancements of viterbi search for fast unit selection synthesis. *Interspeech* , str. 174-177.

73. Toporišič, J. (1991). *Slovenska slovnica.* Maribor: Založba Obzorja.

74. Užarević, J. (14. 9 2012). Bilježenje naglasaka u hrvatskome i dvoznakovni sustav. *Jezik* , str. 126-143.

75. van Santen, J. (1994). Assignment of Segmental Duration in Text-to-Spech Synthesis. *Computer Speech and Language* , str. 95-128.

76. van Santen, J. (1997). Segmental Duration and Speech Timing. *Computing Prosody* , str. 225-250.

77. Veilleux, N. M. (1994). Computational Models of the Prosody/Syntax Mapping for Spoken Language Systems, dissertation. Boston: Boston University of Engineering.

78. Vukušić, S., Zoričić, I., & Grasselli-Vukušić, M. (2007). *Naglasak u hrvatskome književnom jeziku.* Zagreb: Nakladni zavod globus.

79. Wagner, A., & Katarzyna, K. (2010). F0 contour and segmental duration modeling using prosodic features. *Proc. of Speech Prosody 2010.* Chicago.

80. Wang, Y., & Witten, I. (1997). Inducing model trees for continuous classes. *Proceedingsof the Ninth European Conference on Machine Learning*, (str. 128-137).

81. Wightman, C. W. (2002). ToBI or not ToBI. *Speech Prosody* .

82. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining, Practical Machine Learning Tools and Techniques, Third edition.* Morgan Kaufmann.

83. Yang, L.-c. (1998). Contextual Effects on Syllable Duration. *The Third ESCA Workshop on Speech Synthesis.* Jenolan Caves Houses, Australia.

84. Yarowsky, D. (1999). A Comparison of Corpus-Based Techniques for Restoring Accents in Spanish and French Text. *Text, Speech and Language Technology* , str. 99-120.

85. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., i dr. (2006). *The HTK Book (for HTK Version 3.4).* Cambridge: Cambridge University Engineering Department.

Popis slika

Slika 1 Pregled provedenog istraživanja 4

Slika 2 Dva tipa naglaska u ToBi modelu 19

Slika 3 Rečenica s L-L% graničnim tonom 20

Slika 4 Rečenica s H-H% graničnim tonom 20

Slika 5 Rečenica s L-H% graničnim tonom 21

Slika 6 Rečenica s H-L% graničnim tonom 21

Slika 7 Primjer rečenice s transkripcijom po modelu ToBI 22

Slika 8 INTSINT sustav označavanja 24

Slika 9 Princip rada Fujisakijevog modela 26

Slika 10 Primjer komponenti mehanizma za grupiranje 27

Slika 11 Primjer komponenti mehanizma za grupiranje 28

Slika 12 Četiri hrvatske riječi s različitim naglascima 32

Slika 13 Primjer oblika zapisa natuknice u rječniku i potrebne informacije 44

Slika 14 Primjer pravila za jedan tip imenice 48

Slika 15 Pozivanje funkcije iz glavnog dijela programa 48

Slika 16 Primjer funkcije za jedan tip imenice ženskoga roda 49

Slika 17 Točnost dodjeljivanja naglasaka pomoću pravila 62

Slika 18 Prikaz oznaka za jezične značajke pojedinih riječi 70

Slika 19 Točnost predviđanje mjesta i vrste naglaska pomoću modela na testnom skupu 73

Slika 20 pristup automatskog dodjeljivanja naglaska pomoću pravila i modela za naglašavanje 75

Slika 21 Rezultat automatske segmentacije riječi u HTK alatu 82

Slika 22 Primjer sadržaja .lab datoteke 87

Slika 23 Upareni reci lab datoteke i glasovi u riječima s oznakama za početak i kraj rečenice, riječi i sloga 88

Slika 24 Slogovi s pripadajućim oznakama za početak i kraj sloga, riječi i rečenice i pridruženim trajanjima 88

Slika 25 Referentna trajanja slogova, prosječna trajanja slogova na kraju i na početku riječi 92

Slika 26 Referentna trajanja slogova, prosječna trajanja slogova na kraju i na početku rečenice 94

Slika 27 Jednostavna linearna regresija s jednom nezavisnom varijablom 97

Slika 28 Intonacijski događaji u Tilt modelu 104

Slika 29 RFC parametri u Tilt modelu 105

Slika 30 Primjer 5 različitih intonacijskih događaja s različitim vrijednostima tilt parametra 107

Slika 31 Područje pretraživanja za događaj koji se sastoji samo od komponente pad 108

Slika 32 Primjer govornog signala s pridruženim tilt oznakama 110

Popis tablica

Tablica 1 Prednaglasnice i zanaglasnice 37

Tabela 2 MSD oznake koje su dodane u naglasni rječnik za imenice: 57

Tabela 3 Broj osnovnih i izvedenih riječi u naglasnom rječniku po vrstama riječi 59

Tabela 4 Točnost automatskog dodjeljivanja naglasaka pomoću pravila 62

Tablica 5 Jezične značajke uzete u obzir prilikom gradnje modela za predviđanje mjesta naglaska u riječi 67

Tabela 6 Jezične značajke uzete u obzir prilikom gradnje modela za predviđanje vrste naglaska u riječi 68

Tabela 7 Fonetske značajke glasova 69

Tabela 8 Točnost modela za predviđanja mjesta i vrste naglaska dobivena postupkom deseterostuke unakrsne validacije 72

Tabela 9 Točnost predviđanje mjesta i vrste naglaska na testnom skupu 72

Tabela 10 Statistika govornog korpusa VEPRAD 80

Tabela 11 Statistika korpusa bajki i priča 83

Tabela 12 Najčešće riječi u korpusu bajki i priča 84

Tabela 13 Najčešći slogovi u korpusu bajki i priča 85

Tabela 14 Najčešći slogovi u korpusu bajki i priča ako se u obzir uzmu različite vrste naglasaka 85

Tabela 15 Postupak dobivanja naglašenih transkripcija govornog korpusa rastavljenih na slogove 86

Tabela 16 Slogovi koji se pojavljuju na početnom, srednjem i završnom mjestu u riječi i na početku i na kraju rečenice 89

Tabela 17 Referentna trajanja slogova, prosječna trajanja na početku i na kraju riječi te odstupanja od referentnih vrijednosti 91

Tabela 18 Referentna trajanja slogova, prosječna trajanja na početku i na kraju rečenice te odstupanja od referentnih vrijednosti 93

Tabela 19 Utjecaj prvog glasa sloga koji slijedi na trajanje sloga 95

Tabela 20 Jezične značajke korištene u modelu trajanja 99

Tabela 21 Rezultati modela trajanja 101

Tabela 22 Rezultati modela trajanja dobiveni isključivanjem pojedinih skupina značajki iz skupa značajki za učenje modela 102

Tabela 23 RMSE vrijednosti za generirane F0 konture 112

PRILOZI

Prilog 1 Algoritam za izdvajanje genitiva iz teksta

```
sve3=re.compile('{G\s(\w+)}', re.UNICODE)
        g1=sve3.findall(red)
        Gjd1="\n".join(g1)

        sve4=re.compile('{G\s(\w+),.*}', re.UNICODE)
        g2=sve4.findall(red)
        Gjd2="\n".join(g2)

        sve5=re.compile('{G\s-(\w+)}', re.UNICODE)
        g3=sve5.findall(red)
        Gjd_priv1="\n".join(g3)
        Gjd3=''
        Gjd6=''
        Gjd7=''
        Gjd10=''
        Gjd13=''
        Gjd14=''
        Gjd16=''
        Gjd21=''
        if     (Gjd_priv1==u'ēvca')     or     (Gjd_priv1==u'āvca')     or
(Gjd_priv1==u'īvca') or (Gjd_priv1==u'ōvca'):
            Gjd3=osn[:-4]+Gjd_priv1
        elif  (Gjd_priv1[-3:]==u'sci')and  (u'asci'  not  in  Gjd_priv1[-4:])
and osn[-3:]==u'zak':
            Gjd7=osn[:-3]+Gjd_priv1
        elif                          (Gjd_priv1[-4:]==u'ānka')or(Gjd_priv1[-
4:]==u'azak')or(Gjd_priv1[-4:]==u'ānci')or(Gjd_priv1[-
4:]==u'āvca')or(Gjd_priv1[-4:]==u'ārca')or(Gjd_priv1[-
4:]==u'ānca')or(Gjd_priv1[-4:]==u'ārta')or(Gjd_priv1[-
4:]==u'ānta')or(Gjd_priv1[-4:]==u'āvka')or(Gjd_priv1[-
4:]==u'ālca')or(Gjd_priv1[-4:]==u'ājka')or(Gjd_priv1[-
4:]==u'ārka')or(Gjd_priv1[-4:]==u'āmka'):
            Gjd13=osn[-4]+Gjd_priv1
        elif(Gjd_priv1[-5:]==u'ānjka')or(Gjd_priv1[-5:]==u'ānjca'):
            Gjd16=osn[:-5]+Gjd_priv1
        elif (osn[-3:]=='lac') and (Gjd_priv1[-3:]==u'ōca'):
            Gjd14=osn[:-3]+Gjd_priv1
        elif (osn[-1:]=='e') and (Gjd_priv1[-2:]==u'ta'):
            Gjd21=osn+Gjd_priv1
        else:
            Gjdpriv3=''
            if len(Gjd_priv1)==1:
                Gjdpriv3=osn[:-1]+Gjd_priv1
            elif len(Gjd_priv1)==2:
                Gjdpriv3=osn+Gjd_priv1
            elif len(Gjd_priv1)>2:
                nag=[u'à' , u'ȁ' , u'á' , u'â' , u'è' , u'ȅ' , u'é' , u'ê',
u'ì' , u'ȉ' , u'í' , u'î' , u'ò' , u'ȍ' , u'ó' , u'ô' , u'ù' , u'ȕ' , u'ú'
, u'û' , u'ṛ' , u'ř' , u'ŕ' , u'f̂']
                if  (u'à'  in  Gjd_priv1)   or  (u'ȁ'in  Gjd_priv1)  or  (u'á'in
Gjd_priv1)  or  (u'â'in  Gjd_priv1)  or    (u'è'in  Gjd_priv1)  or  (u'ȅ'in
Gjd_priv1)  or  (u'é'in  Gjd_priv1)  or  (u'ê'in  Gjd_priv1)  or  (u'ì'in
Gjd_priv1)  or  (u'ȉ'in  Gjd_priv1)  or  (u'í'in  Gjd_priv1)  or  (u'î'in
Gjd_priv1)  or  (u'ò'in  Gjd_priv1)  or  (u'ȍ'in  Gjd_priv1)  or  (u'ó'in
Gjd_priv1)  or  (u'ô'in  Gjd_priv1)  or  (u'ù'in  Gjd_priv1)  or  (u'ȕ'in
Gjd_priv1)  or  (u'ú'in  Gjd_priv1)  or  (u'û'in  Gjd_priv1)  or  (u'ṛ'in
Gjd_priv1)  or  (u'ř'in  Gjd_priv1)  or  (u'ŕ'in  Gjd_priv1)  or  (u'f̂'in
Gjd_priv1):
```

```
                    for a in nag:
                        Gjd_priv1a=nenag_sam(Gjd_priv1)
                        priv5=nosn.rfind(Gjd_priv1a[0])
                        Gjdpriv6=nosn[:priv5]+Gjd_priv1
                        Gjd3=Gjdpriv6

                elif (u'ā' in Gjd_priv1)or (u'ē' in Gjd_priv1)or (u'ī' in
Gjd_priv1)or(u'ō' in Gjd_priv1)or( u'ū' in Gjd_priv1)or(u'ř' in Gjd_priv1):
                    for b in nag_duz:
                        Gjd_priv2c=nenag_sam(Gjd_priv1)
                        priv6=nosn.rfind(Gjd_priv2c[0])
                        Gjd10=osn[:priv6]+Gjd_priv1
                else:

                    priv1=nosn.rfind(Gjd_priv1[0])
                    Gjdpriv3=nosn[:priv1]+Gjd_priv1
                    nag=[u'à' , u'ầ' , u'á' , u'â' , u'è' , u'ề' , u'é' ,
u'ê', u'ì' , u'ì' , u'í' , u'î' , u'ò' , u'ồ' , u'ó' , u'ô' , u'ù' , u'ừ' ,
u'ú' , u'û' , u'ṛ' , u'ř' , u'ŕ' , u'f̂']

                    for a in nag:
                        #e=osn.find(u'à' , u'ầ' , u'á' , u'â' or  u'ā' or
u'è' or u'ề' or u'é' or u'ê' or u'ē' or u'ì' or u'ì' or u'í' or u'î' or
u'ī' or u'ò' or u'ồ' or u'ó' or u'ô' or u'ō' or u'ù' or u'ừ' or u'ú' or
u'û' or u'ū' or u'ṛ' or u'ř' or u'ŕ' or u'f̂' or u'ř')
                        if a in osn:
                            e=osn.index(a)
                            Gjdp=Gjdpriv3[:e]+osn[e]+Gjdpriv3[e+1:]
                            Gjd6=Gjdp
```

Prilog 2 Algoritam za izdvajanje nominativa iz rječnika

```
nom_mn1=re.compile('{.*\s*N\smn\s(\w+).*}', re.UNICODE)
        nom1=nom_mn1.findall(red)
        nom1="\n".join(nom1)
        n_mn1=nom1

        nom_mn2=re.compile('{.*\s*N\smn\s-(\w+).*}', re.UNICODE)
        nom2=nom_mn2.findall(red)
        nom2="\n".join(nom2)
        nag=[u'à' , u'ȁ' , u'á' , u'â' , u'è' , u'ȅ' , u'é' , u'ê', u'ì' ,
u'ȉ' , u'í' , u'î' , u'ò' , u'ȍ' , u'ó' , u'ô' , u'ù' , u'ȕ' , u'ú' , u'û'
, u'ṛ' , u'ȑ' , u'ŕ' , u'f́']
        nag_duz=[u'ā',u'ē',u'ī', u'ō', u'ū', u'ř']
        nom2a=''
        nom2b=''
        nom2c=''
        nom2d=''
        nom2e=''
        nom2f=''
        nom2g=''
        nom2h=''
        nom2i=''
        if nom2:
            if osn[-1] in u'a' and nom2[0]=='e':
                nom2a=osn[:-1]+nom2
            elif nom2=='i' or nom2=='ovi' or nom2=='evi':
                nom2g=osn+nom2
            elif (nom2[-4:]==u'ēvci') or (nom2[-4:]==u'āvci') or (nom2[-
4:]==u'ōvci')or(nom2[-4:]==u'ālci')or (nom2[-4:]==u'ānci')or (nom2[-
4:]==u'ārci')or (nom2[-4:]==u'ānti')or (nom2[-4:]==u'ānci')or(nom2[-
4:]==u'ārti')or(nom2[-4:]==u'asci'):
                nom2b=osn[:-4]+nom2
            elif (nom2[-5:]==u'ānjci'):
                nom2i=osn[:-5]+nom2
            elif (nom2[-3:]==u'sci') and (u'asci' not in nom2[-4:])and
(u'isci' not in nom2[-4:])and (u'ísci' not in nom2[-4:])and (u'ēsci' not in
nom2[-4:]) and osn[-3:]==u'zak':
                nom2h=osn[:-3]+nom2
            elif (u'à' in nom2)  or (u'ȁ'in nom2) or (u'á'in nom2) or
(u'â'in nom2) or  (u'è'in nom2) or (u'ȅ'in nom2) or (u'é'in nom2) or
(u'ê'in nom2) or (u'ì'in nom2) or (u'ȉ'in nom2) or (u'í'in nom2) or (u'î'in
nom2) or (u'ò'in nom2) or (u'ȍ'in nom2) or (u'ó'in nom2) or (u'ô'in nom2)
or (u'ù'in nom2) or (u'ȕ'in nom2) or (u'ú'in nom2) or (u'û'in nom2) or
(u'ṛ'in nom2) or (u'ȑ'in nom2) or (u'ŕ'in nom2) or (u'f́'in nom2):
                for a in nag:
                    if a in nom2:
                        nom2p=nenag_sam(nom2)
                        priv_br=nosn.rfind(nom2p[0])
                        nom2d=nosn[:priv_br]+nom2

            elif (u'ā' in nom2)or (u'ē' in nom2)or (u'ī' in nom2)or(u'ō' in
nom2)or( u'ū' in nom2)or(u'ř' in nom2):
                for b in nag_duz:
                    if b in nom2:
                        nom2p2=nenag_sam(nom2)
                        priv_br2=nosn.rfind(nom2p2[0])
                        nom2e=osn[:priv_br2]+nom2
            else:
                priv_br3=nosn.rfind(nom2[0])
```

```
        nom2priv7=osn[:priv_br3]+nom2
        nom2f=nom2priv7


Nmn=''
if n_mn1:
    Nmn=n_mn1.strip()
elif nom2a:
    Nmn=nom2a.strip()
elif nom2g:
    Nmn=nom2g.strip()
elif nom2b:
    Nmn=nom2b.strip()
elif nom2h:
    Nmn=nom2h.strip()
elif nom2i:
    Nmn=nom2i.strip()
elif nom2c:
    Nmn=nom2c.strip()
elif nom2d:
    Nmn=nom2d.strip()
elif nom2e:
    Nmn=nom2e.strip()
elif nom2f:
    Nmn=nom2f.strip()
else:
    Nmn=''
```

Prilog 3 Algoritam za stupnjevito provjeravanje MSD oznaka prilikom dodjeljivanja naglaska pomoću pravila

```
if __name__=='__main__':
    import codecs,re,sys,string

    af=codecs.open('naglaseni_tekst_finalno.txt','w','utf-8')

    for red in codecs.open('tekst_format_rijec_pos.txt','r','utf-
8').readlines():

        ispis=''


        sve1=re.compile('(.*)\t.*', re.UNICODE)
        r=sve1.findall(red)
        rijec="\n".join(r)
        rijec1=rijec.lower()

        sve2=re.compile('\.*\t(.*)', re.UNICODE)
        o=sve2.findall(red)
        oznaka1="\n".join(o)



        if oznaka1:

            if oznaka1[0]==u'Z':
                ispis=rijec1.strip()

            elif oznaka1[0] in "P":

                for red6 in
codecs.open('zamjenice_nenag_pos_nag_abecedno.txt','r','utf-
8').readlines():
                    nenag_osn_p=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_p=nenag_osn_p.findall(red6)
                    ne_osn_p="\n".join(ne_p)

                    pos1_p=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_p=pos1_p.findall(red6)
                    pos_p="\n".join(p_p)

                    nag_osn_p=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_p=nag_osn_p.findall(red6)
                    nagl_osn_p="\n".join(nag_p)

                    if rijec1==ne_osn_p and oznaka1[0]==pos_p[0]:
                        ispis=nagl_osn_p

            elif oznaka1[0] in "S":

                for red7 in
codecs.open('prijedlozi_nenag_pos_nag_abecedno.txt','r','utf-
8').readlines():
                    nenag_osn_s=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_s=nenag_osn_s.findall(red7)
                    ne_osn_s="\n".join(ne_s)

                    pos1_s=re.compile('.*\t(.*)\t.*', re.UNICODE)
```

```
                    p_s=pos1_s.findall(red7)
                    pos_s="\n".join(p_s)

                    nag_osn_s=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_s=nag_osn_s.findall(red7)
                    nagl_osn_s="\n".join(nag_s)

                    if rijec1==ne_osn_s and oznaka1[0]==pos_s[0]:
                        ispis=nagl_osn_s

            elif oznaka1[0] in "C":

                for red8 in
codecs.open('veznici_nenag_pos_nag_abecedno.txt','r','utf-8').readlines():
                    nenag_osn_c=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_c=nenag_osn_c.findall(red8)
                    ne_osn_c="\n".join(ne_c)

                    pos1_c=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_c=pos1_c.findall(red8)
                    pos_c="\n".join(p_c)

                    nag_osn_c=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_c=nag_osn_c.findall(red8)
                    nagl_osn_c="\n".join(nag_c)

                    if rijec1==ne_osn_c and oznaka1[0]==pos_c[0]:
                        ispis=nagl_osn_c

            elif oznaka1[0] in "M":

                for red9 in
codecs.open('brojevi_nenag_pos_nag_abecedno.txt','r','utf-8').readlines():
                    nenag_osn_m=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_m=nenag_osn_m.findall(red9)
                    ne_osn_m="\n".join(ne_m)

                    pos1_m=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_m=pos1_m.findall(red9)
                    pos_m="\n".join(p_m)

                    nag_osn_m=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_m=nag_osn_m.findall(red9)
                    nagl_osn_m="\n".join(nag_m)

                    if rijec1==ne_osn_m and oznaka1[0]==pos_m[0]:
                        ispis=nagl_osn_m

            elif oznaka1[0] in "Q":

                for red10 in
codecs.open('cestice_nenag_pos_nag_abecedno.txt','r','utf-8').readlines():
                    nenag_osn_q=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_q=nenag_osn_q.findall(red10)
                    ne_osn_q="\n".join(ne_q)

                    pos1_q=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_q=pos1_q.findall(red10)
                    pos_q="\n".join(p_q)

                    nag_osn_q=re.compile('.*\t.*\t(.*)', re.UNICODE)
```

```
                    nag_q=nag_osn_q.findall(red10)
                    nagl_osn_q="\n".join(nag_q)

                    if rijec1==ne_osn_q and oznaka1[0]==pos_q[0]:
                        ispis=nagl_osn_q

            elif oznaka1[0] in "I":

                for red11 in
codecs.open('uzvici_nenag_pos_nag_abecedno.txt','r','utf-8').readlines():
                    nenag_osn_i=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_i=nenag_osn_i.findall(red11)
                    ne_osn_i="\n".join(ne_i)

                    pos1_i=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_i=pos1_i.findall(red11)
                    pos_i="\n".join(p_i)

                    nag_osn_i=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_i=nag_osn_i.findall(red11)
                    nagl_osn_i="\n".join(nag_i)

                    if rijec1==ne_osn_i and oznaka1[0]==pos_i[0]:
                        ispis=nagl_osn_i

            elif oznaka1[0] in "R":

                for red5 in
codecs.open('prilozi_nenag_pos_nag_sve_abecedno.txt','r','utf-
8').readlines():
                    nenag_osn_r=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_r=nenag_osn_r.findall(red5)
                    ne_osn_r="\n".join(ne_r)

                    pos1_r=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_r=pos1_r.findall(red5)
                    pos_r="\n".join(p_r)

                    nag_osn_r=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_r=nag_osn_r.findall(red5)
                    nagl_osn_r="\n".join(nag_r)

                    if len(pos_r)>1:
                        if rijec1==ne_osn_r and oznaka1[0]==pos_r[0] and
oznaka1[1]==pos_r[1]:
                            ispis=nagl_osn_r
                        elif rijec1==ne_osn_r and oznaka1[0]==pos_r[0]:
                            ispis=nagl_osn_r
                    else:
                        if rijec1==ne_osn and oznaka1[0]==pos[0]:
                            ispis=nagl_osn

            elif oznaka1[0] in "N":

                for red1 in
codecs.open('imenice_nag_pos_nenag.txt','r','utf-8').readlines():
                    nenag_osn=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne=nenag_osn.findall(red1)
                    ne_osn="\n".join(ne)

                    pos1=re.compile('.*\t(.*)\t.*', re.UNICODE)
```

```
                    p=pos1.findall(red1)
                    pos="\n".join(p)

                    nag_osn=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag=nag_osn.findall(red1)
                    nagl_osn="\n".join(nag)

                    if len(oznaka1)>4:
                        if rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2] and oznaka1[3]==pos[3] and oznaka1[4]==pos[4]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2] and oznaka1[3]==pos[3]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0]:
                            ispis=nagl_osn
                    elif len(oznaka1)>3:
                        if rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2] and oznaka1[3]==pos[3]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0]:
                            ispis=nagl_osn
                    elif len(oznaka1)>2:
                        if rijec1==ne_osn and oznaka1[0]==pos[0] and
oznaka1[2]==pos[2]:
                            ispis=nagl_osn
                        elif rijec1==ne_osn and oznaka1[0]==pos[0]:
                            ispis=nagl_osn
                    else:
                        if rijec1==ne_osn and oznaka1[0]==pos[0]:
                            ispis=nagl_osn

            elif oznaka1[0] in "V":

                for red2 in
codecs.open('glagoli_nag_pos_nenag_abecedno_samo_v.txt','r','utf-
8').readlines():
                    nenag_osn_v=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_v=nenag_osn_v.findall(red2)
                    ne_osn_v="\n".join(ne_v)

                    pos1_v=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_v=pos1_v.findall(red2)
                    pos_v="\n".join(p_v)

                    nag_osn_v=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_v=nag_osn_v.findall(red2)
                    nagl_osn_v="\n".join(nag_v)


                    if len(oznaka1)>5:
                        if rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3] and oznaka1[4]==pos_v[4] and
oznaka1[5]==pos_v[5]:
                            ispis=nagl_osn_v
```

```
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3] and oznaka1[4]==pos_v[4]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0]:
                            ispis=nagl_osn_v
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3] and oznaka1[4]==pos_v[4]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0]:
                            ispis=nagl_osn_v
                    elif len(oznaka1)>3:
                        if rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2] and oznaka1[3]==pos_v[3]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0]:
                            ispis=nagl_osn_v
                    elif len(oznaka1)>2:
                        if rijec1==ne_osn_v and oznaka1[0]==pos_v[0] and
oznaka1[2]==pos_v[2]:
                            ispis=nagl_osn_v
                        elif rijec1==ne_osn_v and oznaka1[0]==pos_v[0]:
                            ispis=nagl_osn_v
                    else:
                        if rijec1==ne_osn_v and oznaka1[0]==pos_v[0]:
                            ispis=nagl_osn_v

            elif oznaka1[0] in "A" and rijec1[0] in
u'abcčćdđefghijklmnABCČĆDĐEFGHIJKLMN':

                for red3 in
codecs.open('pridjevi_nag_pos_abecedno_a_nj.txt','r','utf-8').readlines():
                    nenag_osn_a=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_a=nenag_osn_a.findall(red3)
                    ne_osn_a="\n".join(ne_a)

                    pos1_a=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_a=pos1_a.findall(red3)
                    pos_a="\n".join(p_a)

                    nag_osn_a=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_a=nag_osn_a.findall(red3)
                    nagl_osn_a="\n".join(nag_a)

                    if len(oznaka1)>6:
```

```
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4] and oznaka1[5]==pos_a[5] and oznaka1[6]==pos_a[6]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4] and oznaka1[5]==pos_a[5]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    elif len(oznaka1)>5:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4] and oznaka1[5]==pos_a[5]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3] and
oznaka1[4]==pos_a[4]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    elif len(oznaka1)>3:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2] and oznaka1[3]==pos_a[3]:
                            ispis=nagl_osn_a
```

```
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    elif len(oznaka1)>2:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1] and oznaka1[2]==pos_a[2]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    elif len(oznaka1)>2:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a
                        elif rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a
                    else:
                        if rijec1==ne_osn_a and oznaka1[0]==pos_a[0]:
                            ispis=nagl_osn_a

            elif oznaka1[0] in "A" and rijec1[0] in
u'oprsštuvzžOPRSŠTUVZŽ':

                for red4 in
codecs.open('pridjevi_nag_pos_abecedno_o_z.txt','r','utf-8').readlines():
                    nenag_osn_a2=re.compile('(.*)\t.*\t.*', re.UNICODE)
                    ne_a2=nenag_osn_a2.findall(red4)
                    ne_osn_a2="\n".join(ne_a2)

                    pos1_a2=re.compile('.*\t(.*)\t.*', re.UNICODE)
                    p_a2=pos1_a2.findall(red4)
                    pos_a2="\n".join(p_a2)

                    nag_osn_a2=re.compile('.*\t.*\t(.*)', re.UNICODE)
                    nag_a2=nag_osn_a2.findall(red4)
                    nagl_osn_a2="\n".join(nag_a2)

                    if len(oznaka1)>6:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and oznaka1[3]==pos_a2[3]
and oznaka1[4]==pos_a2[4] and oznaka1[5]==pos_a2[5] and
oznaka1[6]==pos_a2[6]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3] and oznaka1[4]==pos_a2[4] and oznaka1[5]==pos_a2[5]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3] and oznaka1[4]==pos_a2[4]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3]:
```

```
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    elif len(oznaka1)>5:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and oznaka1[3]==pos_a2[3]
and oznaka1[4]==pos_a2[4] and oznaka1[5]==pos_a2[5]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3] and oznaka1[4]==pos_a2[4]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and oznaka1[3]==pos_a2[3]
and oznaka1[4]==pos_a2[4]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and
oznaka1[3]==pos_a2[3]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2] and oznaka1[3]==pos_a2[3]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a2[1] and oznaka1[2]==pos_a2[2]:
                            ispis=nagl_osn_a2
```

```
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]
and oznaka1[1]==pos_a2[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    elif len(oznaka1)>4:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0] and
oznaka1[1]==pos_a[1]:
                            ispis=nagl_osn_a2
                        elif rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2
                    else:
                        if rijec1==ne_osn_a2 and oznaka1[0]==pos_a2[0]:
                            ispis=nagl_osn_a2

            if ispis:
                output=ispis
                if rijec[0] in u'ABCČĆDĐEFGHIJKLMNOPRSŠTUVZŽ' and output[0]
not in u'àáâãèéêëìíîïòóôõùúûüŗřŕř':
                    if output[0] in u'a':
                        prvo_slovo=u'A'
                    elif output[0] in u'b':
                        prvo_slovo=u'B'
                    elif output[0] in u'c':
                        prvo_slovo=u'C'
                    elif output[0] in u'č':
                        prvo_slovo=u'Č'
                    elif output[0] in u'ć':
                        prvo_slovo=u'Ć'
                    elif output[0] in u'd':
                        prvo_slovo=u'D'
                    elif output[0] in u'đ':
                        prvo_slovo=u'Đ'
                    elif output[0] in u'e':
                        prvo_slovo=u'E'
                    elif output[0] in u'f':
                        prvo_slovo=u'F'
                    elif output[0] in u'g':
                        prvo_slovo=u'G'
                    elif output[0] in u'h':
                        prvo_slovo=u'H'
                    elif output[0] in u'i':
                        prvo_slovo=u'I'
                    elif output[0] in u'j':
                        prvo_slovo=u'J'
                    elif output[0] in u'k':
                        prvo_slovo=u'K'
                    elif output[0] in u'l':
                        prvo_slovo=u'L'
                    elif output[0] in u'm':
                        prvo_slovo=u'M'
                    elif output[0] in u'n':
                        prvo_slovo=u'N'
                    elif output[0] in u'o':
                        prvo_slovo=u'O'
                    elif output[0] in u'p':
                        prvo_slovo=u'P'
                    elif output[0] in u'r':
                        prvo_slovo=u'R'
                    elif output[0] in u's':
                        prvo_slovo=u'S'
```

```
                    elif output[0] in u'š':
                        prvo_slovo=u'Š'
                    elif output[0] in u't':
                        prvo_slovo=u'T'
                    elif output[0] in u'u':
                        prvo_slovo=u'U'
                    elif output[0] in u'v':
                        prvo_slovo=u'V'
                    elif output[0] in u'z':
                        prvo_slovo=u'Z'
                    elif output[0] in u'ž':
                        prvo_slovo=u'Ž'
                    output=prvo_slovo+output[1:]

                elif rijec[0] in u'ABCČĆDĐEFGHIJKLMNOPRSŠTUVZŽ' and
output[0] in u'àáâãèéêēìíîīòóôõùúûūṛŕŕř':
                    if output[0] in u'à':
                        prvo_slovo1=u'À'
                    elif output[0] in u'á':
                        prvo_slovo1=u'Á'
                    elif output[0] in u'â':
                        prvo_slovo1=u'Â'
                    elif output[0] in u'ã':
                        prvo_slovo1=u'Ã'
                    elif output[0] in u'è':
                        prvo_slovo1=u'È'
                    elif output[0] in u'é':
                        prvo_slovo1=u'É'
                    elif output[0] in u'ê':
                        prvo_slovo1=u'Ê'
                    elif output[0] in u'ē':
                        prvo_slovo1=u'Ē'
                    elif output[0] in u'ì':
                        prvo_slovo1=u'Ì'
                    elif output[0] in u'í':
                        prvo_slovo1=u'Í'
                    elif output[0] in u'î':
                        prvo_slovo1=u'Î'
                    elif output[0] in u'ī':
                        prvo_slovo1=u'Ī'
                    elif output[0] in u'ò':
                        prvo_slovo1=u'Ò'
                    elif output[0] in u'ó':
                        prvo_slovo1=u'Ó'
                    elif output[0] in u'ô':
                        prvo_slovo1=u'Ô'
                    elif output[0] in u'õ':
                        prvo_slovo1=u'Õ'
                    elif output[0] in u'ù':
                        prvo_slovo1=u'Ù'
                    elif output[0] in u'ú':
```

```
            prvo_slovo1=u'Ú'
        elif output[0] in u'û':
            prvo_slovo1=u'Û'
        elif output[0] in u'ū':
            prvo_slovo1=u'Ū'
        elif output[0] in u'ŗ':
            prvo_slovo1=u'Ŗ'
        elif output[0] in u'ř':
            prvo_slovo1=u'Ř'
        elif output[0] in u'ŕ':
            prvo_slovo1=u'Ŕ'
        elif output[0] in u'ŷ':
            prvo_slovo1=u'Ŷ'
        elif output[0] in u'ǔ':
            prvo_slovo1=u'Ǔ'
        output=prvo_slovo1+output[1:]

else:
    output=rijec


af.write(output+'\n')
```

Prilog 4 Algoritam za dodjeljivanje naglaska rječima rastavljenim na slogove

```
#!/usr/bin/env python
#-*- coding: utf-8 -*-
if __name__=='__main__':
import codecs,re,sys,string,os,glob
sys.stdout = codecs.getwriter('UTF-8')(sys.stdout)


list_of_files = glob.glob('nag_rast_casm04*.txt')

for fileName in list_of_files:
data_list = codecs.open(fileName, "r","utf-8" ).readlines()
af=codecs.open("a"+fileName,"w","utf-8")
for red in data_list:
n=re.compile('(.*)\t+.*', re.UNICODE)
m=n.findall(red)
naglaseno="\n".join(m)
naglaseno=naglaseno.strip()

r=re.compile('.*\t+(.*)', re.UNICODE)
p=r.findall(red)
rastavljeno="\n".join(p)
rastavljeno=rastavljeno.strip()

if rastavljeno:

if u'à' in naglaseno or u'á' in naglaseno or u'à' in naglaseno or u'â' in
naglaseno or u'è' in naglaseno or u'è' in naglaseno or u'é' in naglaseno or
u'ê' in naglaseno or u'ì' in naglaseno or u'ì' in naglaseno or u'í' in
naglaseno or u'î' in naglaseno or u'ò' in naglaseno or u'ò' in naglaseno or
u'ó' in naglaseno or u'ô' in naglaseno or u'ù' in naglaseno or u'ù' in
naglaseno or u'ú' in naglaseno or u'û' in naglaseno or u'ŗ' in naglaseno or
u'ŕ' in naglaseno or u'ŕ' in naglaseno or u'f' in naglaseno or u'ř' in
naglaseno:
for i in naglaseno:

if i==u'á' or i==u'à' or i==u'à' or i==u'â' or i==u'è' or i==u'è' or
i==u'é' or i==u'ê' or i==u'ì' or i==u'ì' or i==u'í' or i==u'î' or i==u'ò'
or i==u'ò' or i==u'ó' or i==u'ô' or i==u'ù' or i==u'ù' or i==u'ú' or
i==u'û' or i==u'ŗ' or i==u'ŕ' or i==u'ŕ' or i==u'f' or i==u'ř':
b=naglaseno.find(i)
else:
b=u'-'

if u'=' in rastavljeno:
c=rastavljeno.find(u'=')
else:
c=u'!'

if b==u'-':
rast_nag=rastavljeno
elif b!=u'-' and c==u'!':
if b>1:
rast_nag=rastavljeno.replace(rastavljeno[b],naglaseno[b],1)
else:
rast_nag=naglaseno

if b!=u'-' and c!=u"!":
```

```
if b<c:
rast_nag=rastavljeno.replace(rastavljeno[b],naglaseno[b],1)
else:
e=rastavljeno[c+1:]
d=e.find(u'=')
if d==-1:
rast_nag=rastavljeno[:c].strip()+e.replace(rastavljeno[b+1],naglaseno[b],1)
.strip()
else:
f=len(rastavljeno[:c+1])+d
if b<f:
rast_nag=rastavljeno[:c].strip()+e.replace(rastavljeno[b+1],naglaseno[b],1)
.strip()
else:
g=rastavljeno[f+1:]
h=g.find(u'=')
if h==-1:
rast_nag=rastavljeno[:f].strip()+g.replace(rastavljeno[b+2],naglaseno[b],1)
else:
j=len(rastavljeno[:f+1])+h
if b<j:
rast_nag=rastavljeno[:f].strip()+g.replace(rastavljeno[b+2],naglaseno[b],1)
.strip()
else:
l=rastavljeno[j+1:]
n=l.find(u'=')
if n==-1:
rast_nag=rastavljeno[:j].strip()+l.replace(rastavljeno[b+3],naglaseno[b],1)
.strip()
else:
p=len(rastavljeno[:j+1])+n
if b<p:
rast_nag=rastavljeno[:j].strip()+l.replace(rastavljeno[b+3],naglaseno[b],1)
.strip()
#print rastavljeno[:j+1]
#print l.replace(rastavljeno[b+3],naglaseno[b],1).strip()
else:
rast_nag=rastavljeno
af.write(rast_nag+'\n')
```

Prilog 5 MSD oznake za hrvatski jezik[16]

Tablica mogućih kategorija

CATEGORY (en)	Value (en)	Code (en)	Attributes
CATEGORY	Noun	N	5
CATEGORY	Verb	V	6
CATEGORY	Adjective	A	7
CATEGORY	Pronoun	P	11
CATEGORY	Adverb	R	2
CATEGORY	Adposition	S	1
CATEGORY	Conjunction	C	2
CATEGORY	Numeral	M	6
CATEGORY	Particle	Q	1
CATEGORY	Interjection	I	0
CATEGORY	Abbreviation	Y	0
CATEGORY	Residual	X	1

[16]Preuzeto iz: MULTEXT-East Morphosyntactic Specifications, revised Version 4, 3.8. Croatian Specifications(Ljubešić, 2013)

Imenice

Specifikacija za imenice			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	noun	N
1	Type	common	c
		proper	p
2	Gender	masculine	m
		feminine	f
		neuter	n
3	Number	singular	s
		plural	p
4	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		vocative	v
		locative	l
		instrumental	i
5	Animate	no	n
		yes	y

Glagoli

Specifikacija za glagole			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Verb	V
1	Type	main	m
		auxiliary	a
		copula	c
2	VForm	infinitive	n
		participle	p
		present	r
		future	f
		imperative	m
		aorist	a
		imperfect	e
3	Person	first	1
		second	2
		third	3
4	Number	singular	s
		plural	p
5	Gender	masculine	m
		feminine	f
		neuter	n
6	Negative	no	n
		yes	y

Pridjevi

Specifikacija za pridjeve			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Adjective	A
1	Type	general	g
		possessive	s
		participle	p
2	Degree	positive	p
		comparative	c
		superlative	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		vocative	v
		locative	l
		instrumental	i
6	Definiteness	no	n
		yes	y
7	Animate	no	n
		yes	y

Zamjenice

Specifikacija za zamjenice			
P	Attribute (en)	Value (en)	Code (en)
0	CATEGORY	Pronoun	P
1	Type	personal	p
		demonstrative	d
		indefinite	i
		possessive	s
		interrogative	q
		relative	r
		reflexive	x
2	Person	first	1
		second	2
		third	3
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		vocative	v
		locative	l
		instrumental	i
6	Owner_Number	singular	s
		plural	p
7	Owner_Gender	masculine	m
		feminine	f
		neuter	n
8	Clitic	no	n
		yes	y
9	Referent_Type	personal	p
		possessive	s
10	Syntactic_Type	nominal	n
		adjectival	a
11	Animate	no	n
		yes	y

Prilozi

Specifikacija za priloge			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Adverb	R
1	Type	general	g
		participle	r
2	Degree	positive	p
		comparative	c
		superlative	s

Prijedlozi

Specifikacija za prijedloge			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Adposition	S
1	Case	genitive	g
		dative	d
		accusative	a
		locative	l
		instrumental	i

Veznici

Specifikacija za veznike			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Conjunction	C
1	Type	coordinating	c
		subordinating	s
2	Formation	simple	s
		compound	c

Brojevi

Specifikacija za brojeve			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Numeral	M
1	Form	digit	d
		roman	r
		letter	l
2	Type	cardinal	c
		ordinal	o
		multiple	m
		special	s
3	Gender	masculine	m
		feminine	f
		neuter	n
4	Number	singular	s
		plural	p
5	Case	nominative	n
		genitive	g
		dative	d
		accusative	a
		vocative	v
		locative	l
		instrumental	i
6	Animate	no	n
		yes	y

Čestice

Specifikacija za čestice			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Particle	Q
1	Type	negative	z
		interrogative	q
		modal	o
		affirmative	r

Uzvici

Specifikacija za uzvike			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Interjection	I

Kratice

Specifikacija za kratice			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Abbreviation	Y

Ostalo

Specifikacija za ostalo			
P	**Attribute (en)**	**Value (en)**	**Code (en)**
0	CATEGORY	Residual	X
1	Type	foreign	f
		typo	t
		program	p

Prilog 6 Naglasni rječnik hrvatskoga jezika

Dobiveni naglasni rječnik hrvatskoga jezika nalazi se na CD-u u prilogu ovoj disertaciji.

Životopis

Lucia Načinović Prskalo rođena je 15. siječnja 1983. u Rijeci. Osnovnu školu i opću gimnaziju pohađala je u Labinu gdje i danas živi. Diplomirala je 2008. godine na Filozofskom fakultetu u Rijeci na dvopredmetnom studiju informatike i engleskog jezika i književnosti. U prosincu 2009. godine upisala je poslijediplomski doktorski studij informacijskih i komunikacijskih znanosti na Odsjeku za informacijske i komunikacijske znanosti pri Filozofskom fakultetu u Zagrebu.

Od veljače 2009. godine zaposlena je na Odjelu za informatiku Sveučilišta u Rijeci gdje aktivno sudjeluje u izvođenju nastave na preddiplomskim i diplomskim studijima Odjela za informatiku.

Bila je stručni suradnik na znanstvenom projektu MZOŠ-a "Govorne tehnologije" pod vodstvom prof. dr. sc. Ive Ipšića, a trenutno je suradnik na projektu "LangNet" pod vodstvom izv. prof. dr. sc. Sande Martinčić-Ipšić. Njezino područje istraživanja je računalna obrada prirodnog jezika. Recenzirala je više konferencijskih radova te je član programskog odbora međunarodne konferencije. Objavila je više znanstvenih radova na znanstvenim skupovima i kod međunarodnih nakladnika.

Popis objavljenih radova:

- Načinović, L.; Martinčić-Ipšić, S. Prosodic Modelling for Croatian Speech Synthesis. Towards Solving the Social Science Challenges with Computing Methods. Mileva Boshkoska, Biljana (ur.). Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien : Peter Lang, 105-118. 2015.
- Načinović Prskalo, L.; Martinčić-Ipšić, S. An Overview of Prosodic Modelling for Croatian Speech Synthesis, Proceedings of Information Technologies and Information Society (pp. 105-112), Novo Mesto, Slovenia. 2013.
- Načinović, L.; Perak, B.; Meštrović, A.; Martinčić-Ipšić, S. Identifying Fear Related Content in Croatian Texts, Language Technologies (pp. 153-156), Ljubljana, Slovenia. 2012.
- Načinović, L.; Pobar, M.; Martinčić-Ipšić, S.; Ipšić, I. Automatic Intonation Event Detection Using Tilt Model for Croatian Speech Synthesis, Information Sciences and e-Society (pp. 383-391), Zagreb, Croatia. 2011.

- Strčić, V., Načinović, L., Ipšić, I. A review on Creation and Structure of Emotional Multimodal Databases. Proceedings of the CECIIS Central European Conference on Information and Intelligent Systems 21st International Conference 2010 (pp. 19-24). Varaždin, Croatia. 2010.
- Načinović, L., Martinčić-Ipšić, S., Ipšić, I. Intonation Modeling for Croatian Speech Synthesis. Proceedings of the 33rd International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2010 (pp. 253-257). Opatija, Croatia. 2010.
- Načinović, L., Martinčić-Ipšić, S., Ipšić, I. Statistical Language Models for Croatian Weather-domain Corpus. Proceedings of the InFuture2009 conference (pp. 333-340). Zagreb, Croatia. 2009.
- Načinović, L., Pobar, M., Ipšić, I., Martinčić-Ipšić, S. Grapheme-to-Phoneme Conversion for Croatian Speech Synthesis. 32nd International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO 2009 (pp. 318 -323). Opatija, Croatia. 2009.