UNIVERZA V LJUBLJANI
FILOZOFSKA FAKULTETA
ODDELEK ZA SLOVENISTIKO

ALENKA LAHARNAR

Korpus govorjenega jezika in njegova uporabnost za analize govora

Diplomsko delo

Ajdovščina, 2014

UNIVERZA V LJUBLJANI
FILOZOFSKA FAKULTETA
ODDELEK ZA SLOVENISTIKO

ALENKA LAHARNAR

Korpus govorjenega jezika in njegova uporabnost za analize govora

Diplomsko delo

Mentor: doc. dr. Hotimir Tivadar

Študijski program:
Slovenistika − E

Ajdovščina, 2014

Izvleček

Korpus govorjenega jezika in njegova uporabnost za analize govora

Diplomsko delo obravnava slovensko fonetiko v povezavi s korpusi govorjenega jezika. Teoretični del predstavi govorjeni jezik in fonetiko, pravila za transkribiranje in označevanje govora ter transkripcijske standarde, uporabljene pri tujejezičnih korpusih. Nadalje izpostavi korpus govorjenega jezika GOS in njegovi najpomembnejši komponenti za analizo govora, transkripcijo in konkordančnik, ter nekaj manjših projektov, tj. govornih zbirk. V analitičnem delu pokaže na nekatera problematična področja slovenske fonetike v povezavi s pojavitvami v korpusu GOS, ki se kažejo v načinu pogovornega zapisa in v pomanjkljivostih konkordančnika, po drugi strani pa ponuja možnosti za analizo govora, ki je lahko izvedena le na dovolj veliki bazi kakovostno transkribiranega govora.

Ključne besede: fonetika, korpus govorjenega jezika, transkribiriranje

Abstract

Speech Corpus and its Usefulness in Speech Analysis

The following thesis deals with Slovene phonetics in relation with speech corpora. The theoretical part presents in more detail the characteristics of spoken language and phonetics, principles of transcription and speech labelling, and describes transcription standards applied in foreign language corpora. Furthermore, the GOS Corpus of Spoken Slovene and its two most important components when dealing with speech analysis (i.e. transcription and the concordancer) and some minor projects (i.e. speech databases) are considered in more detail. In the analytical part, the thesis, on one hand, highlights some problem areas of the Slovene phonetics in connection with the occurrences in the GOS speech corpus, which surfaced from the pronunciation-based transcription and the concordancer's weaknesses, and, on the other hand, sheds light on the possibilities for speech analyses, which can only be done when having a big enough high quality transcription database.

Key words: phonetics, speech corpus, transcription

KAZALO VSEBINE

KAZALO TABEL

KAZALO SLIK

1 UVOD

Jezik se najpogosteje udejanja v obliki govora, ki pa je od pisnega jezika zahtevnejši, kar zadeva dokumentacijo, transkripcijo in analizo. Kot meni Tivadar (2004: 33), le korpus realnih besedil govorjenega jezika lahko pripomore k lažji in hitrejši obravnavi določenih fonetičnih problemov. Jezikovne tehnologije so omogočile boljšo raziskanost govora, zato nas ne sme presenetiti Krekova primerjava (2012: 1) digitalne revolucije z Gutenbergovim izumom tiskarskega stroja. Šele razvoj računalniške tehnologije je v zadnjih desetletjih povečal število raziskav govorjenega jezika – to je torej postalo bistveno lažje. Pestrost evropskih jezikov je velika dragocenost, a jeziki, ki niso imeli pisne norme (na primer dalmatinski)[1], postajajo omejeni le na govorno rabo, ki pa se še dodatno krči. V tej luči je raziskovanje govorjenega jezika ključnega pomena.

Živimo v informacijski družbi in vplivu informacijske tehnologije se tudi jezikoslovje ni moglo izogniti. Besedilni korpusi so postali pomemben vir informacij za jezikoslovne raziskave, spletni pregledovalniki besedil in spletni slovarji pa nepogrešljiv del večine osebnih računalnikov. Jezikoslovje se je usmerilo k uporabnim aplikacijam, ki služijo tako navadnim uporabnikom kot tudi strokovnjakom z vseh področij; področje uporabnega jezikoslovja je izrazito interdisciplinarno.

V tujini je področje korpusnega jezikoslovja večinoma dobro razvito (Anglija, Češka), tudi v Sloveniji pa se je v zadnjih letih to področje razširilo na kar nekaj novih projektov. Eden takih je govorni korpus GOS, ki je največja zbirka govorjenih besedil pri nas. Ob tem pa ne smemo pozabiti na manjše projekte, predvsem govorne zbirke, ki predstavljajo pomemben vidik raziskovanja govorjenega jezika, pa čeprav gre večinoma za bolj specifična besedila.

Pri vsem tem pa se zdi, da je fonetika oziroma glasoslovje, ki je eden od temeljev govora, nekako potisnjena na rob. Korpusi govorjenega jezika in fonetične raziskave so v vzročno-posledičnem razmerju, slednje pa so brez ustreznega konkordančnika in transkripcije toliko težje izvedljive.

[1] Dalmatinski jezik je omejen le na prenos govorne oblike, kar je vzrok za vse manjšo rabo.

2 GOVORJENI JEZIK

Preden začnemo obravnavati korpuse govorjenega jezika, ne moremo mimo osnovnega vprašanja, kaj govorjeni jezik sploh je. Najlažje ga definiramo s prenosnikom, saj je nasprotje pisnega. Po Toporišiču (1992: 53−54) je slušno uresničeno besedilo, ki je lahko prvotno govorno ali pa besedilno obnavljalno (brano, recitirano, deklamirano, na pamet naučeno, pripovedovano). Govorni dogodek je tvorjenje besedila iz jezikovnih prvin z namenom, da se obvestilo o nečem prenese od govorečega ogovorjenemu, lahko gre pa tudi za obnavljanje kakega besedila. Prav tako kot je obširna definicija govorjenega jezika, pa so tudi široke možnosti analize te vrste jezika: te segajo vse od spontanosti v jeziku, posebnih stavčnih struktur, narečnih prvin do fonetike in prozodije.

V strokovni literaturi se pojavljata sorodna izraza: govorjeno besedilo in (govorjeni) diskurz. V nekaterih definicijah sta pomensko prekrivna, sicer pa prvi sodi na področje korpusnega jezikoslovja, drugi na področje analize diskurza. Ker je osnova te diplomske naloge korpusno jezikoslovje, bomo v nadaljevanju vedno uporabljali termin govorjeno besedilo. Zemljarič Miklavčič (2008a: 24−25) pravi, da je govorjeno besedilo v materialnem smislu akustični dogodek. Med seboj se ta besedila razlikujejo – tista z enakimi lastnostmi spadajo v isto besedilno vrsto. Podoben je še izraz tip besedila, ki pa razlikuje govorjena besedila glede na posamezen kriterij, na primer na monologe in dialoge.

Pri vsem tem pa ne smemo pozabiti na spontanost v govoru, katerega zapis je pogosto kaotičen, neurejen in netekoč, kar je posledica časovnega pritiska ter situacijskega konteksta. Ko govor prenesemo v pisno obliko, izgubimo poleg številnih značilnosti govorjenega jezika tudi njegovo avtentičnost, zato si je potrebno prizadevati za ohranitev čim večjega števila specifičnih značilnosti, kot so akustične in neverbalne lastnosti. Pri raziskovanju govorjenega jezika nas v prvi vrsti zanima njegova spontanost, torej brana besedila za take raziskave niso relevantna (Zemljarič Miklavčič 2008b: 93). Poznamo različne oblike spontanega govora: razlike nastanejo zaradi demografskih značilnosti govorcev (na primer regija, starost), odnosov med sogovorci (formalni, neformalni govor), okoliščin (zasebni, javni govor) ali prenosnika (neposredno, mediji). Slovenščina ne pozna enotnega govorjenega jezika, kar povzroča mnoge težave, saj je težko določiti skupne značilnosti. Zemljarič Miklavčič (2008b: 94) je razlike spontanega govora razdelila v dve vrsti: na vertikalni ravni so razlike med govorci različnih regij (besedna, glasovna, deloma skladenjska raven), na horizontalni ravni pa so vzrok za razlike okoliščine (višja ali nižja stopnja formalnosti).

Razmerje med govorjenim in pisnim knjižnim jezikom – predvsem govorom v javnosti − vse do danes ostaja nedorečeno. Vzrok za tako stanje lahko pripisujemo zgodovinskemu dejstvu, da je bil knjižni jezik do konca drugega tisočletja večinoma določen z njegovo pisno podobo, kar je posledično pripeljalo do manjše enotnosti govorjene podobe. Slovenščina je realizirana (in s tem tudi zapovedana na vseh javnih področjih) na državni ravni šele dobri dve desetletji, kar se kaže v slabši kakovosti javnega govora, posledice pa so vidne tudi v normativnih priročnikih, ki prikazujejo normativno pravorečno podobo. Pravorečje je namreč odvisno tudi od pisne podobe jezike, tj. branega govora (Tivadar 2012a: 203−206).

2.1 Govorjeni jezik, fonetika in fonologija

Kot smo omenili že v začetku tega poglavja, se lahko pri analizi govora osredotočimo na različne komponente govorjenega jezika. Ta diplomska naloga bo osredotočena na fonetiko, zato jo bom v naslednjih podpoglavjih natančneje opisala.

Fonetika je mednarodni izraz za glasoslovje in pomeni jezikoslovni nauk o izrazni slušni podobi jezika. Predmet raziskovanja je izrazna podoba glasov, glasovnih zvez, besed, besedil itd., raziskuje pa dvoje: razločevalno oblikovnost izrazne podobe jezika (fonologija) in njeno tvarnost (Toporišič 2003: 41). Fonetika je lahko akustična ali artikukacijska, za našo obravnavo je pomembnejša slednja, saj raziskuje glasove glede na mesto in način izgovorjave. Z izrazom fonem opisujemo glas, ki loči pomen besed ali morfemov v nasprotju z drugimi glasovi, torej gre za pomensko razločevalno glasovno enoto. Slovenski knjižni jezik ima devetindvajset fonemov. Slovenska slovnica opredeljuje še fonemske variante oziroma alofone, saj imajo lahko fonemi več pojavnih oblik z določeno skupno značilnostjo na izgovorni ali slušni ravni; navadno se pojavljajo v posebnem glasovnem okolju. Za našo analizo pomembna izraza sta tudi fonetična pisava, to je zapisovanje po izgovoru, in fonetični prepis, ki je lahko tonemski, na primer [*jɛ ̯us̓ta ̯u ɔp šéstix*], ali netonemski, [*jɛ ̯ustà ̯u ɔp šéstix*] (Toporišič 1992: 41−42). Pri raziskavi se bom osredotočila na samoglasnike in naglaševanje ter na zvočnike in nezvočnike z variantami posameznih fonemov.

2.1.1 Samoglasniki

Prozodične lastnosti samoglasnikov v slovenščini so trajanje, jakost in ton. Po dolžini trajanja so dolgi in kratki, po jakosti krepki (naglašeni) in šibki (nenaglašeni) ter po tonu visoki (cirkumflektirani) in nizki (akutirani). Naglas je jakostna in tonska izrazitost, izrazitost po trajanju pa kolikost (kvantiteta) (Toporišič 2003: 59–60).

Le naglašeni samoglasniki so lahko dolgi ali kratki: ozka *e* in *o* sta vedno dolga in naglašena, *i*, *a* in *u* ter široka *e* in *o* so lahko dolgi ali kratki, slednji so lahko naglašeni ali nenaglašeni. Med naglašene spada tudi zveza *ə* + *r*, pisana z *r*, ki je vedno dolga. Trajnost samoglasnikov se v slovenščini razlikuje od narečja do narečja – večina jih pozna razliko med dolgim in kratkim naglašenim samoglasnikom, le ponekod pride do odstopanj (na primer rovtarsko narečje ima vse naglašene *i*-je in *u*-je kratke − Toporišič 2003: 60).

Pri obravnavi samoglasnikov in naglaševanja torej silijo v ospredje naslednja dejstva (Tivadar 2004: 31):

- samoglasniški sistem temelji na nasprotju med širokim in ozkim *e*-jevskim oziroma *o*-jevskim samoglasnikom, ki sta težko predvidljiva;
- svobodno mesto naglasa;
- trajanje jakostno naglašenih samoglasnikov (*bràt* : *brát* in *kùp* : *kúp*).

Slovenščina ima glede na artikulacijske in akustične lastnosti osem naglašenih fonemov (visoka *a* in *u*, sredinske *e, ɛ, ə, o* in *ɔ* ter nizki *a*), kot pa smo že omenili, je posebnost dvojnost *e*-jevskih in *o*-jevskih fonemov (Tivadar 2004: 35).

Tabela 1: Samoglasniški trikotnik (Toporišič 2003: 48).

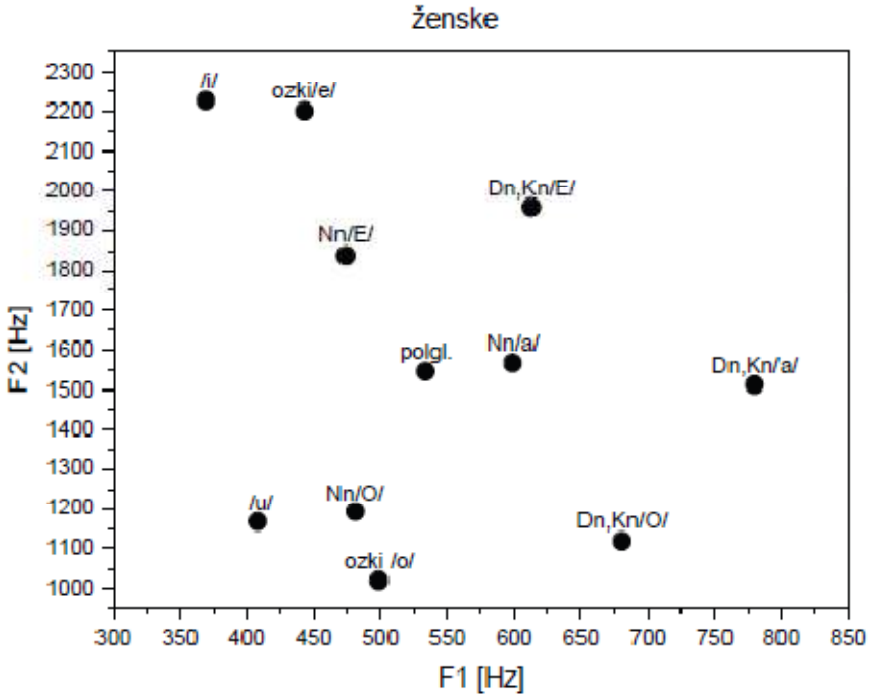
	Sprednji	Nesprednji	
		Srednji	Zadnji
Visoki	*i*		*u*
Sredinski	*e* *ɛ*	*ə*	*o* *ɔ*
Nizki		*a*	

Nekatere novejše raziskave samoglasnikov se osredotočajo na njihovo kakovost in trajanje ter jih analizirajo s pomočjo sodobnih fonetičnih programov. Tivadar (2010: 111−112) obe lastnosti samoglasnikov določa s pomočjo formantne analize[2] v programu Praat, kjer ločeno analizira izgovorjene samoglasnike treh moških govorcev in štirih ženskih govork (vsi govorci so medijsko prepoznavni). Tako ugotavlja naslednje značilnosti:

- slovenski samoglasniki so jasno izgovorjeni;
- podoba slovenskih samoglasnikov je glede na govorce istega spola enotna;
- F1 in F2 določata slovenske samoglasnike;
- nenaglašena *e* in *o* nista tako široka kot naglašena široka *ɛ* in *ɔ*;
- nenaglašeni *e, o* in *a* so centralizirani, torej so blizu polglasniku, in tvorijo svoj samoglasniški trikotnik;
- ozka *e* in *o* nista dolga glede na trajanje dolgo naglašenih *ɛ* in *ɔ* (polglasnik − približno 50 ms, *i* in *u* − od 50 do 75 ms, ozka *e* in *o* − od 65 do 85 ms, široka *ɛ* in *ɔ* − od 80 do 100 ms, *a* − več kot 100 ms);
- nenaglašeni samoglasniki so kratki in praviloma trajajo pod 50 ms;

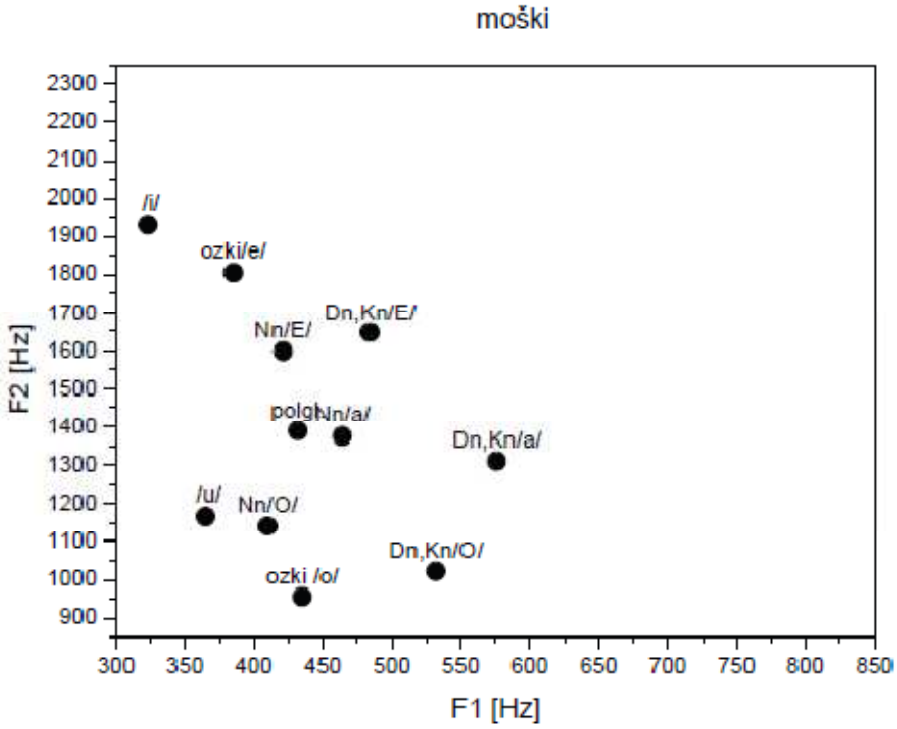
[2] Formantna analiza je analiza frekvence nihanja glasilk pri izreki določenega glasu (osnovna enota je Hz, to je dogodek na sekundo). F0 je osnovni ton (glasilki nihata hitreje, zato je višji), za razlikovanje med glasovi pa zadoščata prva (F1) in druga (F2) formantna frekvenca. Samoglasniki imajo največjo odprtostno stopnjo (Toporišič 2003: 43).

- obstaja razlika med moškimi in ženskimi samoglasniki.



Slika 1: Povprečne vrednosti F2 v odvisnosti od povprečnih vrednosti F1 pri ženskih govorkah (Tivadar 2010: 114).[3]

[3] Fonema *ɛ* in *ɔ* sta na slikah zapisana kot *E* in *O*.



Slika 2: Povprečne vrednosti F2 v odvisnosti od povprečnih vrednosti F1 pri moških govorcih (Tivadar 2010: 114).

V knjižni slovenščini poznamo več fonemskih variant: namesto širokega *e* pred *j*, ki je na koncu besede ali pred soglasnikom, izgovarjamo srednji *e*, podobno se zgodi s širokim *o* pred *u̯* – fonetično ju zaznamujemo s strešico spodaj (*e̯* in *o̯*). Kratka in naglašena *i* ter *u*, ki nista na začetku ali na koncu besede, izgovarjamo malce nižje, zato zvenita širša (*sìt, kùp*), *a* pa v enakih razmerah zveni višje in bolj polglasniško (*brʌ̀t*) (Toporišič 2003: 50). Pri tem pa velja omeniti še malce novejšo teorijo, ki jo je osnoval Jurgec (2011: 243) in se nanaša na deveti samoglasnik; to naj bi bil omenjeni srednji nizki samoglasnik *ʌ*. Teorijo argumentira z naslednjimi dokazi: govorci ne slišimo razlikovalne kvantitete, razlike med obema nizkima samoglasnikoma so fonetične in teoriji naj bi pritrjevali tudi fonološki vzorci, ki pa jih je avtor omejil le na govorce osrednje Slovenije. Tivadar (2012b: 595) meni, da je pri tem nujno opredeliti, ali gre za lokalno ali za splošnoslovensko značilnost. Na območju osrednje Slovenije je namreč v pogovornem jeziku izgovor samoglasnikov pogosto reduciran, kar je moteče za govorce z drugih narečnih območij, poleg tega združevalna vloga knjižnega jezik ne deluje popolnoma. Pri vsem tem se velja še vprašati, koliko dodatnih fonemov (sedaj fonemskih variant) bi torej morali vpeljati v knjižno slovenščino, če bi se ravnali po tej teoriji.

Nekatera narečja, med drugim tudi govor Ljubljane, namesto kratkih *i*, *u* in *a* izgovarjajo polglasnik – kdaj je knjižen in kdaj ne, vidimo v primerih, ko pri pregibanju izginja (*pes psa*) ali pa pride do premene. Od knjižnega jezika zelo odstopata široka *e* in *o*. Več narečij pozna dvoglasniški izgovor, ponekod pa je uveljavljen kratek izgovor. Polglasnik ponekod izgovarjajo blizu *a*-ja, drugje pa sta namesto njega *e* ali *a* (Toporišič 2003: 50–52).

Ozka *é* in *ó* sta pogostejša kot široka, saj ju načeloma izgovarjamo v vseh prevzetih in na novo sestavljenih besedah (*Krémelj, zébra, Lóndon, Móskva*). Dolgi široki *ê* poznajo nekatere skupine besed: korenske besede (*sêlo, têčem*), širok je lahko vsak naglašen *e* pred *j*, ki mu sledi samoglasnik (*orhidêja*), vsak *e* pred *r* v neslovanskih prevzetih besedah (*afêra, bolêro*), pri številnikih v nekaterih sklonih (*pêtim*) ... Prav tako dolgi široki *ô* poznajo nekatere korenske besede (*vôjska, klôpotec*), širok je vsak naglašeni *o* pred *v*, ki mu sledi samoglasnik (*sôva, vdôva*), pripone *-ôba, -ôča, -ôta* (*grenkôba, čistôča*) ... Kratka naglašena *è* in *ò* ter nenaglašena *e* in *o* sta vedno široka (Toporišič 2003: 52–56).

Polglasnik imajo nekateri koreni (*čebéla, pekèl, meglà, mezgà* ...); če tak polglasnik ni naglašen in pri pregibanju ne izginja, se počasi zamenjuje z *e* (na primer *mêgla*). Prav tako imajo polglasnik osnove brez samoglasnika (*pes psa, ves vsa*), pojavlja se tudi ob vsakem *r*, ki ni ob samoglasniku (*črv, žanr*) in še v nekaterih drugih redkejših primerih (Toporišič 2003: 56–59).

Kratki naglašeni samoglasniki so prisotni v nekaterih korenih (*bòb, fànt, otròk, šmènt* ...), priponskih morfemih (na primer *-ìč, -è, -òt; fantìč, fantè, čofòt*), predponskih morfemih (na primer *pò-, prèd-; pònaročíti, prèdnaročíti*), medmetih (*àh*) ... (Toporišič 2003: 60–63).

2.1.2 Naglaševanje

Pomembna komponenta analize govora je naglaševanje. Ko naglašujemo, izgovarjamo z naglasom, ki pa ga lahko definiramo takole: »/N/aglas je [...] fonetično-fonološka lastnost zloga glede na druge zloge v nizu, v slovenskem (knjižnem) jeziku in v jezikih z nepredvidljivim oz. poljubnim mestom naglasa izraziteje vezana na besedo« (Tivadar 2004: 35). Naglas povezuje posamezne zloge v besedo; če je zlog en sam, loči posamezne vrste besed (na primer samostalnik *pòd* od predloga *pod*), v večzložnih besedah pa je lahko razločevalno mesto naglasa (*leží* proti *lêži*). Slovenski jezik pozna dve vrsti naglaševanja:

jakostno in tonemsko. Jakostno naglašene samoglasnike izgovarjamo z večjo močjo kot nenaglašene, tonemsko naglaševanje pa loči dva naglasna tipa – samoglasnike izgovarjamo enkrat višje (cirkumflektirani oziroma padajoči samoglasniki), drugič nižje (akutirani oziroma rastoči samoglasniki). Vloga obeh tonemov (cirkumfleksa in akuta) je oblikorazločevalna, pa tudi pomenskorazločevalna (*pọ́t* v pomenu cesta proti *pọ̑t* v pomenu znoj). Obe vrsti naglaševanja sta knjižni, odvisni pa sta od narečne rabe. Besede brez naglasa imenujemo breznaglasnice ali naslonke, saj se v govorni verigi naslanjajo na naglašene besede; te so na primer *in*, *nad*, *se* ter *nam*. Poznamo še večnaglasnice, kot so *kinodvorána*, *slovénsko-italijánski* in *sívozelèn* (Toporišič 2003: 63–66).

V slovenščini naglasno mesto ni vezano na določen zlog, pač pa je lahko celo v isti besedi na različnih zlogih (na primer *stvár stvarí*); torej je naglasno mesto določeno z vsako besedo posebej in se ga tako tudi naučimo. Nekatera pravila pa nam lahko določanje naglasnega mesta precej olajšajo (Toporišič 2003: 66–71):

- pregibne besede z enozložno osnovo imajo naglas na osnovi, na primer *králj králja*, poznamo pa nekatere izjeme, kot so *gôra goré gôri* (to je mešani naglasni tip, ki pa se v knjižni slovenščini vse redkeje pojavlja);

- besede, pri katerih je koren nezložen, imajo naglas na zlogu za njim, na primer *téga* in *prišèl*; izjeme so redke (*pójdem*);

- predpona je načeloma nenaglašena, naglas se nahaja desno od nje (*nakúp*, *odnêsti*, *prevèč* ...); izjeme so redke (*nóčem*, *pôtok*, *nékaj*, *nárazen – narazèn* ...);

- predslonka je nenaglašena, na primer *na dopúst*;

- končnica je naglašena le v nekaterih primerih, in sicer pri nezložni osnovi, na primer *psá psôma*, če je v njej polglasnik edini samoglasnik, na primer *temà* (te besede spadajo v končniški naglasni tip);

- nekateri morfemi imajo obvezen naglas – to so oblikospreminjevalni morfemi oziroma končnice (na primer *gorá* in *sinù*), oblikotvorni morfemi (na primer *nesó* in *prepeváje*) ter besedotvorni končaji oziroma priponska obrazila (na primer pri glagolu v nedoločniku in sedanjiku (*okopávati okopávam*), pri samostalniku v vseh treh spolih (*hudír*, *klepetúlja* in *pecívo*) ter pri pridevnikih (*bakrén* in *kamnít*)).

Samoglasniške foneme v slovenščini delimo glede na naglašenost oziroma nenaglašenost in glede na dolgost oziroma kratkost na tri samoglasniške sisteme (ločevalna znamenja veljajo za jakostno naglaševanje).

Tabela 2: Trije samoglasniški sestavi (Toporišič 2003: 71).

NAGLAŠENI		NENAGLAŠENI
Dolgi	Kratki	
í *ú* *é* *ó* *ê* *ô* *á*	*ì* *ù* *ə* *è* *ò* *à*	*i* *u* *ə* *e* *o* *a*

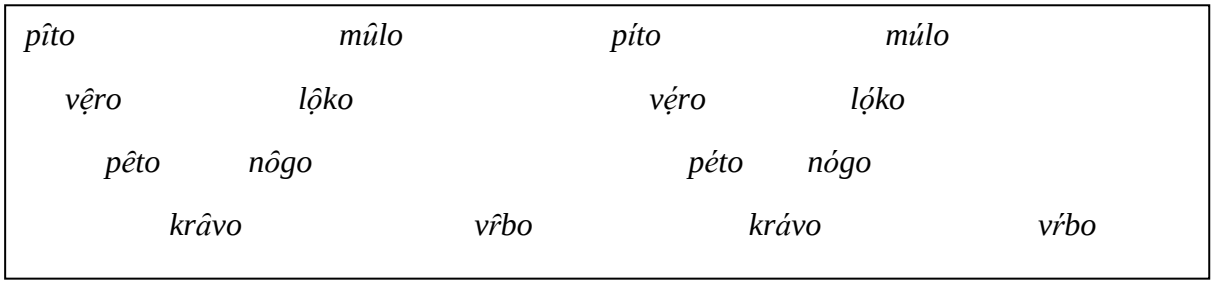
Ločevalna znamenja za mesto naglasa, kolikost in kakovost pri jakostnem naglaševanju so ostrivec, krativec in strešica.

Tabela 3: Ločevalna znamenja pri jakostnem naglaševanju (Tivadar 2004: 36).[4]

Naglasno znamenje	Funkcija naglasnega znamenja	Primeri
krativec	kračina, poleg tega pri *e* in *o* še odprtost/širokost	*obràt, sìt, kùp; mèt, prihòd*
ostrivec	dolžina, poleg tega pri *e* in *o* še zaprtost/ozkost	*obráta, míza, júha, smŕt; Slovénija, dóber*
strešica	dolžina in odprtost/širokost pri *e* ter *o*	*dôbra júha, vôda; žêlja, rêči; tvôje mnênje*

Pri tonemskem naglaševanju ločimo dva sistema, akutiranega in cirkumflektiranega ter štiri različna znamenja: akut, cirkumfleks, brevis in dvojni brevis (Toporišič 2003: 72−73).

[4] Gre za aktualizirano tabelo iz Slovenske slovnice (Toporišič 2003: 72−73), kjer je nakazana pomembnost naglasa in kakovosti samoglasnikov za slovensko fonetiko.



Slika 3: Dva samoglasniška sistema pri tonemskem naglaševanju (Toporišič 2003: 72)

Tabela 4: Ločevalna znamenja pri tonemskem naglaševanju (Toporišič 2003: 73).

Znamenje	Vloga
akut	mesto naglasa, dolžina, nizki ton (naslednji nenaglašeni zlog je visok)
cirkumfleks	mesto naglasa, dolžina, visoki ton (naslednji nenaglašeni zlog je nizek)
brevis	mesto naglasa, kračina, nizki ton (naslednji nenaglašeni zlog je visok)
dvojni brevis	mesto naglasa, kračina, visoki ton (naslednji nenaglašeni zlog je nizek)

Pri tonemskem naglaševanju zaznamujemo ozkost *e*-ja in *o*-ja s piko pod posamezno črko (*pẹ̑t, pọ̑t*), širine pa ne zaznamujemo posebej.

2.1.3 Zvočniki

Ko obravnavamo soglasnike, je smiselno ločiti zvočnike od nezvočnikov. Prvi namreč ne vplivajo na zvenečnost pred njimi stoječih nezvočnikov. Mednje spadata nosnika *m* in *n* (edina nosnika v slovenskem jeziku), jezičnika *l* in *r* ter drsnika *v* in *j*. Zvočnik *v* ima še tri druge oblike: *u̯* izgovarjamo podobno kot *u*, podobno izgovarjamo tudi *w*, a z bolj približanima ustnicama, *ʍ* pa je enak *w*-ju, le da je nezveneč. Prav tako ima zvočnik *n* več oblik: pred palatalnimi *k, g* in *h* naredimo zaporo na mehkem nebu, pred *j* na koncu besede in pred soglasnikom iste besede pa ga izgovarjamo s privzdignjenim srednjim delom jezika – podobno izgovarjamo zvočnik *l* pred *j* (palatalizirani *l* − Toporišič 2003: 73–77).

Tabela 5: Zvočniške variante (Toporišič 2003: 76).[5]

Osnovna varianta		Posebna varianta		
Fonetični zapis	Zgledi	Fonetični zapis	V položaju pred	Zgledi
m	*sam, mamka*	*ɱ*	*f, v*	*simfonija, sem videl*
n	*Ana, Ančka*	*ŋ*	*k, g, h,*	*Anka, angel, Anhovo*
		n'	*j# ali jC*	*konj, konjski*
		ɱ	*f, v*	*konfin, konvoj*
v	*siva*	*u̯*	*# ali C za V*	*siv, sivka, bo vzela, bo vstala, bo ušla*
		w	*zven. C in ne za V*	*vzeti, vreti, odvreči, odvzeti, barv, brvjo*
		ʍ	*nezven. C in ne za V*	*vsak, v temi, predvsem*
l	*pila, glagol*	*l'*	*j# ali C*	*polj, poljski*

Zvočniško varianto *u̯* zapisujemo s črko *v* (*siv, sivka*), poleg tega pa še s črko *l* (*bral, polh, bralca, sol, žal*) in v redkih primerih z *u* (*nauk, bo ušla, audi*). Zvočnik *j* lahko zapišemo še s črko *i* (*bo imela, bonboniera, pacient*), zvočnika *l* in *n* pa z *lj* in *nj* (*poljski, konjski*).

2.1.4 Nezvočniki

Kot rečeno, nezvočniki vplivajo na zvenečnost pred njimi stoječih glasov. Skupina nezvočnikov v slovenščini obsega kar petnajst fonemov, in sicer zapornike (*p, b, t, d, k* in *g*), pripornike (*f, h, s, z, š* in *ž*) in zlitnike (*c, č* in *dž*).

[5] Znak # pomeni konec besede, C soglasnik, V pa samoglasnik.

Nekateri zaporniki imajo v določenih glasovnih okoliščinah poseben izgovor. Glasova *t* in *d* pred *n* izgovarjamo z nosno odporo in ju imenujemo favkalna, enako velja za *p* in *b* pred *m*. Zapornika *t* in *d* pred *l* z obstransko odporo imenujemo obstranska (Toporišič 2003: 79–80).

Tabela 6: Zaporniške variante (Toporišič 2003: 80).

Osnovna varianta		Posebna varianta						
Fonet. zapis	Zgledi	Favkalna		Zobnoustnična		Obstranska		Zgledi
		Fon. zapis	V pol. pred	Fon. zapis	V pol. pred	Fon. zapis	V pol. pred	
p	*papa, zob*	*P*	*m*					*zob me boli*
				p	*f/v*			*ob fantu, rob vaze*
b	*baba, čep drži*	*B*	*m*					*zobmi*
				ʙ	*v*			*obveza, ob vodi*
t	*tata, tod*	*T*	*n*					*tnalo, sod nagni*
						t	*l*	*tla*
d	*deda, svatba*	*D*	*n*					*dno*
						d	*l*	*dleto*

Pripornik *f* je zobnoustnični, *s* in *z* sta zobna, *š* in *ž* zadlesnična, pripornik *h* pa je mehkoneben. Zlitnike *c*, *č* in *dž* izgovarjamo kot ustrezne zapornike in pripornike (namesto *t* in *s*, *t* in *š* ter *d* in *ž*), a kot enotne glasove, kar dokazuje tudi dolžina izgovorjave. Sicer pa uporabljamo nezveneče nezvočnike pogosteje kot zveneče. Slednji so možni le v isti besedi pred samoglasnikom, zvočnikom ali zvenečim nezvočnikom (*sadu, sadje* − Toporišič 2003: 80–81).

Poznamo pa še nekatere narečne posebnosti: del zahodne Slovenije namesto *g* izgovarja zveneči *h* oziroma *ɣ*, rožansko narečje namesto *g* izgovarja grlni *k*, na Krasu in Slovenci v

Italiji poznajo v nekaterih besedah *ć*. Nekatera slovenska narečja poznajo izgovor zvenečih nezvočnikov tudi pred premorom (*ped, pod*), medtem ko tudi severovzhodnoštajerski *si[d] muc* ni knjižen. Končna *-d* in *-b* ponekod izgovarjajo s *-s* in *-f* (*gras, zof*), *ft* namesto *pt* (*ftič*), *kl* in *gl* namesto *tl* in *dl* (*klačiti, gleto*), *h* namesto *k* (*hmet, h Tavčarju*), *š* namesto *šč* (*kleše* − Toporišič 2003: 83).

2.2 Govor in jezikovne tehnologije

V uvodu sem že nakazala, kako pomembna je povezava korpusov govorjenega besedila, ki jih omogočajo jezikovne tehnologije, in analize govora. Za to vrsto tehnologije v preteklosti ni bilo ravno značilno, da bi se z njo ukvarjali jezikoslovci, pa vendar je prehod v digitalno dobo prinesel tudi to spremembo. Zanimanje jezikoslovcev za jezikovne tehnologije narašča, pri izdelavi uporabnih jezikovnih aplikacij pa je velikega pomena tudi njihovo sodelovanje s strokovnjaki s področja računalništva.

Navadno se termin jezikovne tehnologije nanaša na računalniške aplikacije, povezane z jezikom. Te so na primer črkovalnik, slovnični pregledovalnik, avtomatsko povzemanje, rudarjenje podatkov ... Ko govorimo o govornih tehnologijah, sta najpomembnejši sintetizator govora in aplikacija, ki deluje obratno, to je razpoznavalnik govora. Med aplikacije jezikovnih tehnologij lahko štejemo tudi elektronske slovarje in korpuse besedil s programskim vmesnikom, konkordančnikom, ki omogoča avtomatsko iskanje po besedilih (Verdonik 2008: 232). Povezave do opisanih aplikacij so na spletni strani Slovenskega društva za jezikovne tehnologije, njihove predstavitve pa v zbornikih konference Jezikovne tehnologije.

Jezikovni viri so potrebni za razvoj algoritmov, ki tvorijo tehnologijo kot tako. Ti viri vsebujejo znanje o jeziku, na njihovi podlagi pa lahko algoritmi sestavljajo vzorce. Delimo jih na pisne jezikovne vire in na govorne jezikovne vire. Slednji zajemajo posnetke govora – lahko gre za vnaprej napisana in prebrana besedila, za javno govorno nastopanje ali za spontani govor. Nadaljnja obdelava temelji na transkripciji, ortografski ali fonemski, in na dodanih različnih jezikoslovnih ter akustičnih oznakah (Verdonik 2008: 232–233).

Znotraj jezikovnih tehnologij se je uveljavil poseben termin za tiste, ki se ukvarjajo z govorom, to so govorne tehnologije, kamor sodijo področja sinteze govora, razpoznave

govora, strojnega simultanega prevajanja govora, govornih sistemov dialoga in govornih virov ter raziskav. Kot posledica variabilnosti jezikovne rabe pri govornih tehnologijah prevladuje statistični pristop: na podlagi vzorcev se izračuna najbolj verjetna jezikovna uresničitev, dejansko uresničitev pa primerjamo s poznanimi vzorci. Tako je uspešnost tehnologije zelo odvisna tudi od zajetih vzorcev (Verdonik 2008: 234–235).

2.3 Govorni korpus kot osnova za analizo govora

Jezikovne tehnologije torej omogočajo izgradnjo govornih korpusov. Ta je namenjena tako preteklosti kot prihodnosti. Po eni strani ohranjamo vir kulturne dediščine našim zanamcem, po drugi strani pa je njihova izdelava namenjena razvoju raznih govornih tehnologij in računalniških aplikacij. Mnoge jezikovne skupnosti imajo zgrajene govorne korpuse, nekatere celo več, kar je ponudilo nove možnosti za raziskovanje govorjenega jezika (Zemljarič Miklavčič idr. 2009: 423). Za gradnjo korpusov govorjenega jezika je ključnega pomena spontanost govora – ta lahko variira od nepripravljenega do pripravljenega govora. Brana besedila so v tem smislu le oralizacija pisnega besedila, kar mora biti ustrezno označeno v korpusu. Enota govornega jezika je govorjeno besedilo, ta izraz pa pogosto zamenjuje tudi termin diskurz. Govorni korpus se lahko oblikuje samostojno ali kot podkorpus referenčnega korpusa.

Posebej pomembna pa je povezava govornih korpusov in možnosti fonetičnih raziskav, o čemer so že razmišljali nekateri avtorji. Gorjanc (2005: 8) meni, da so govorni korpus transkribirani posnetki spontanega govora, pri tem so zanimive slovnično-leksikalne lastnosti, prozodične pa le, če je ustrezno označen. Korpusi za zajetje fonetično-fonoloških raziskav in govornih tehnologij nastajajo kot študijski posnetki in zajemajo le izbrane stavke – te korpuse imenujemo govorne zbirke. Podobno razmišljajo tudi Atkins in drugi (1992: 2), namreč da ti korpusi niso namenjeni fonetičnim raziskavam, ampak le zajetju posebnosti govorne komunikacije.

Kot opisujeta Verdonik in Zwitter Vitez (2012: 71–72), je za raziskave govora potrebnih čim več informacij o gradivu in čim več iskalnih možnosti, kot so dostop do pogovornega in standardiziranega zapisa (pomembno zaradi interpretacije zadetkov), natančne kontekstne informacije (regija, leto snemanja, tip diskurza, prvi jezik udeležencev ...), omejevanje

gradiva glede na izbrane kriterije, predvajanje zvoka, shranjevanje rezultatov ter možnost frekvenčnih seznamov za statistično analizo. Pri vsem tem pa je fonetika ostala nekako v ozadju, saj je, razen dostopa do pogovornega zapisa, ne omenja nobena od teh postavk. Za raziskovalce fonetike bi bili še kako zanimivi podatki, ali je prevladujoče naglaševanje jakostno ali tonemsko, ali se naglas Dolenjca po večletnem bivanju v Novi Gorici lahko spremeni ter, če se, kako in ali Ljubljančani pogosteje izgovorjajo *məglà* ali *mə̀gla*. Vsi ti podatki bi lahko služili za osnovo mnogim raziskavam govora.

3 TRANSKRIBIRANJE IN OZNAČEVANJE KOT PODLAGA ZA ZAPIS GOVORA

S transkribiranjem in označevanjem govorjenih besedil se je vzpostavilo več priporočil, kakšno metodologijo dela uporabiti pri tej vrsti dela. Tako so se oblikovala priporočila TEI, to je organizacije s področja govornih tehnologij, priporočila EAGLES, ki jih je oblikovala evropska iniciativa z namenom oblikovanja skupnih standardov za izdelavo jezikovnih virov, ter priporočila NERC, ki so nastala v okviru gradnje govorne komponente korpusa COBUILD. Ta tri priporočila so podrobneje opisana v naslednjih poglavjih.

Nadalje so opisane nekatere specifičnosti transkribiranja in označevanja govorjenih besedil, ki so se uveljavile v slovenskem prostoru. Ločimo jih predvsem glede na namen besedil, saj je vrsta transkripcije odvisna od kasnejše rabe.

3.1 Priporočila za označevanje govorjenih besedil TEI

TEI (*Text Encoding Initiative*) je organizacija, ki deluje na področju govornih tehnologij. Leta 1991 je v publikaciji *Text Encoding Initiative, Spoken text Working group, final report* objavila priporočila za označevanje govorjenih besedil TEI.

Transkribirano govorjeno besedilo je razdeljeno na dve osnovni enoti: telo in glavo besedila. Osnovni enoti sheme za segmentacijo govorjenih besedil sta besedilo in izjava. Besedilo je transkripcija niza govorjenega besedila, ki ga je mogoče imeti za samostojno enoto in ga obravnavati kot zaključeno besedilo. Podenota je izjava oziroma segment diskurza, določen po skladenjskih, fonoloških ali prozodičnih načelih, podenoto izjave pa predstavlja segment. Ponujen je še izraz divizija (odsek), ki je hierarhična stopnja med besedilom in izjavo (Zemljarič Miklavčič 2008a: 94–95).

Transkriptorji označujejo tudi akustične in druge dogodke, kot so premori, neverbalni glasovi, mimika obraza in gibi, zvoki v ozadju ... Premor se lahko pojavi med dvema izjavama (pripišemo ga vsem govorcem) ali znotraj posamezne izjave (pripišemo ga samo govorcu). Določamo lahko še dolžino premora. Neverbalni glasovi so lahko označeni opisno (na primer *kašljanje*) ali pa zapisani v ortografski obliki kot del besedila (*khm*). Oznaki sta

vedno pripisana še identifikacija govorca in opis dogodka. V nekaterih primerih se neverbalni glasovi prekrivajo z izjavo, torej tvorijo samostojno izjavo, na primer če govorec sodeluje v pogovoru le z *mhm* – v takih primerih sta v veljavi dve praksi zapisovanja: kot izjava ali kot niz neverbalnih glasov. Kinezični dogodki (pokimati, odkimati, skomigniti z rameni) so v besedilih označeni in pripisani posameznemu govorcu. Ti dogodki nimajo statusa izjave. Pri transkribiranju pa naletimo tudi na nekomunikacijske zvoke, ki so pomembni pri interpretaciji besedila. V besedilu so te vrste zvoka, na primer zvonjenje telefona, primerno označene in so ključnega pomena za razumevanje komunikacije. Če je potrebno, je tak dogodek lahko pripisan posameznemu govorcu (Zemljarič Miklavčič 2008a: 97−98).

Določene transkripcijske sheme dopuščajo tudi pripis trajanja posameznih dogodkov: izjav, premorov, neverbalnih glasov in kinezičnih ter nekomunikacijskih dogodkov. Tudi časovno zaporedje dogodkov mora čim natančenje predstavljati sekvence zvoka oziroma govora. Do težav pride, ko na primer govori več govorcev hkrati. Prekrivni govor mora biti vedno označen, čeprav na različne načine (z zvezdicami, z oglatimi oklepaji, lahko je indeksiran ali vertikalno poravnan). Nekatera transkripcijska orodja (na primer Transcriptor) olajšajo delo, saj sama merijo časovni potek in tako dopuščajo zapisovanje prekrivnega govora (Zemljarič Miklavčič 2008a: 99–100). Določene so tudi oznake za hitrost govora (na primer *rall* – rallentando oziroma počasneje), za glasnost (na primer *cresc* – crescendo oziroma glasneje), za višino (na primer *monot* – monotono in *high* – visok višinski razpon), za napetost izgovora (na primer *sl* – nerazločno in *pr* – zelo napeto) in za kvaliteto glasu (na primer *whisp* – šepet).

Dodatno lahko transkriptorji zapisujejo še fonetične oznake, ki v transkripcijah sploh ne obstajajo ali pa so zelo natančne. Bolj običajna je uporaba ortografske transkripcije z nekaterimi prozodičnimi oznakami, na primer za glasnost, hitrost govorjenja, kvaliteto glasu in tonski potek. Zapis govora zahteva nekakšen kompromis med dejanskimi zvoki in konvencionalnimi ortografskimi znaki, še posebej ko imamo opraviti z neformalnimi ali narečnimi oblikami jezika. Pozornost je treba nameniti tudi okrajšavam in kraticam: lahko so izgovorjene kot besede ali pa črkovane, in kot take jih tudi zapisujemo. Raba velikih začetnic in ločil mora biti strogo določena. Sheme običajno dopuščajo tudi prostor za uredniške opombe (na primer oznake za nerazumljive odseke), ki se zdijo pomembne za pravilno interpretacijo transkribiranega besedila. Vsako besedilo oziroma del besedila mora biti označeno tako, da ga je mogoče umestiti v širši kontekst oziroma izvorno besedilo. Besedilo

je lahko označeno še oblikoskladenjsko in/ali skladenjsko, kar razširi uporabnost korpusa v smislu jezikoslovnih analiz (Zemljarič Miklavčič 2008a: 100–102).

Obstaja nekaj kritik priporočil TEI: preveč pozornosti naj bi namenjala označevanju neverbalnih dogodkov (Payne 1993), podobno pa meni tudi Sinclair, sicer eden vodilnih pri gradnji COBUILD, da so avtorji priporočil predstavili svetu, kar naj bi videl samo računalnik (Zemljarič Miklavčič 2008a: 102).

3.2 Priporočila EAGLES in NERC

Evropska iniciativa EAGLES (*Expert Advisory Group on Language Engineering Standards*) je nastala leta 1993, glavni namen nastanka pa je bil oblikovanje skupnih standardov za izdelavo jezikovnih virov. Ena od petih delovnih skupin je bila zadolžena za področje govorjenega jezika. Pred priporočili EAGLES pa sta že bila uveljavljena dva transkripcijska modela: NERC in TEI.

Projekt NERC (*Network of European Reference Corpora*) je v končno poročilo vključil transkripcijsko konvencijo, ki je nastala v okviru gradnje govorne komponente korpusa COBUILD. Transkripcijski sistem NERC ima štiri stopnje (Zemljarič Miklavčič 2008a: 103–104):

- prva stopnja – ortografski zapis z minimalnim številom ločil, zamenjave govorcev niso označene;
- druga stopnja – razširjena ortografska transkripcija z osnovnimi podatki o govorcih, njihovih menjavah in z oznakami za neverbalne dogodke;
- tretja stopnja – vključene so vse informacije druge stopnje, dodatno pa še intonacijske in interakcijske informacije (označene tonske enote in prekrivni govor);
- četrta stopnja – vključene so vse informacije tretje stopnje, dodatno pa še intonacijske, akustične in fonetične informacije, transkripcija je povezana s spektrogramom.

Po besedah delovne skupine NERC je za referenčne govorne korpuse priporočljiva uporaba druge stopnje, torej ortografska transkripcija z osnovnimi informacijami o govorcih, prekrivnem govoru in neverbalnih dogodkih. Taka vrsta transkripcije je primerna za

jezikoslovne študije, ki pa ne potrebujejo podatkov o intonaciji in fonetičnih lastnostih glasov (Zemljarič Miklavčič 2008a: 104).

Fonemska in fonetična transkripcija sta navadno zapisani s simboli IPA (*International Phonetic Alphabet*), ki je v jezikoslovju najpogosteje uporabljani transkripcijski standard, priporočen tudi s strani TEI in NERC. Zapleten sistem simbolov IPA je težko berljiv za računalnik, zato je bil razvit sistem SAMPA (*Speech Assessment Methodology Phonetic Alphabet*), ki je v sistem ASCII prenesel simbole IPA na način, primeren za fonetično transkribiranje številnih evropskih jezikov – adaptacija je bila narejena tudi za slovenščino (Zemljarič Miklavčič 2008a: 106).

Po priporočilih EAGLES so stopnje označevanja govornih korpusov naslednje (Zemljarič Miklavčič 2008a: 105):

- ortografska raven (standardni zapis);
- citatno-fonemska raven (besede so zapisane s fonemi, zapis izhaja iz besede v izolaciji);
- širša fonetična raven (zapis fonemov ob upoštevanju izgovora besed v kontekstu);
- ožja fonetična raven (zapis tudi z variantami fonemov);
- akustično-fonetična raven (ločuje vse segmente govora).

Znotraj vsake transkripcijske ravni so označeni oziroma zapisani pojavi, na primer neverbalni glasovi govorcev, nekomunikacijski dogodki in prozodične lastnosti govora.

Posebnosti ortografskega transkribiranje po načelih EAGLES so (Zemljarič Miklavčič 2008a: 106):

- reducirane oblike besed;
    - uporaba besed, kot se pojavljajo v standardnem slovarju,
    - če je treba, lahko zapišemo tudi druge oblike reduciranih besed, ki jih ne najdemo v standardnem slovarju,

    - zapis takih oblik besed je priporočljiv, kadar imajo visoko frekvenco ponavljanja in kadar vključujejo izbris celotnega zloga;
- dialektalne oblike (v transkripciji morajo biti označene);
- številke (lahko transliterirane kot besede);
- kratice, okrajšave in črkovanja;
    - kratice zapišemo kot kratice in jih označimo,
    - kratice, izgovorjene kot besede, zapišemo kot besede,
    - črkovanje v govoru mora biti označeno;
- medmeti (označeni in zapisani način, kot ga najdemo v slovarju).

Za označevanje pojavov, ki spremljajo govor, EAGLES določa naslednje prozodične oznake (Zemljarič Miklavčič 2008a: 107):

- glasovni polverbalni dogodki (oklevanja, polverbalna komunikacijska sredstva – *mhm*);
- glasovni neverbalni dogodki (kašljanje, kihanje, tleskanje);
- nekomunikacijski dogodki (akustični dogodki, ki se zgodijo med komunikacijo, na primer zapiranje vrat);
- identifikacija govorca (povzema identifikacijo po TEI s kometarjem, da obstajajo tudi manj zapletene oblike označevanja);
- menjava govorcev;
- prekrivni govor;
- izpusti branega besedila (v primeru, da obstaja pisna predloga govorjenega besedila);
- samopopravki (eksplicitni, na primer z besedo *oziroma*, ali implicitni, kot je na primer ponavljanje besed);
- besedni fragmenti (oklevanja, napačni začetki);

- nerazumljivi fragmenti.

V priporočilih EAGLES ostaja nedorečena problematika ločil. V priporočilih NERC ortografska transkripcija določa delno uporabo ločil (piko, kjer transkriptor čuti stavčno mejo, in veliko začetnico na začetku stavka). Nasprotno pa priporočila SpeechDat popolnoma odsvetujejo rabo ločil. Stališče skupine EAGLES je le rahlo nakazano v ugotovitvi, da postavljanje stavčnih mej v govorjenem besedilu ni enostavno in da je zato raba ločil v ortografski transkripciji zelo otežena.

V zvezi z digitalnimi verzijami vsakega posnetka EAGLES določa, naj bodo te v korpus vključene kot posebne komponente. Januarja 1996 je bil na konferenci delovne skupine med primarne naloge uvrščen tudi razvoj primerne računalniške opreme, ki bo omogočala hkratni dostop do zvočnega posnetka (Zemljarič Miklavčič 2008a: 108).

V slovenskem prostoru je priporočilom EAGLES sledila Verdonik (2006: 69−71), ki je besede zapisala tako, da so sledile knjižnemu standardu, ne pa dejanski izgovorjavi, pri čemer je bilo nekaj izjem, na primer zapis *mogli* v pomenu *morati*, *pol* v pomenu *potem* in *more* namesto oblike *mora*. Besedam, ki niso izgovorjene po knjižnem standardu, je dodala fonetični prepis v SAMPA-abecedi in oznako *+pron=** (izgovorjava je razvidna iz fonetičnega prepisa) ali *+pron=izg*. (slaba izgovorjava), na primer *kakšne [k/a:SnE][+pron=izg.]* in *tudi [t/u:t][+pron=*]*. Pri tem pa se pojavita naslednji vprašanji: kaj je pričakovani govorni standard in ali bi bilo mogoče podobno transkripcijo takih besed izvesti na obsežnejšem gradivu (Zemljarič Miklavčič 2008a: 120).

3.3 Transkripcijske prakse v slovenskem prostoru

V preteklosti so se v slovenskem prostoru uveljavili trije poglavitni tipi transkribiranja, ki so se razvili na podlagi namena transkribiranih besedil. Tako ločimo transkripcije za dialektološko rabo, za besediloslovne in pragmatične raziskave ter za jezikovnotehnološko prakso (Verdonik in Zwitter Vitez 2011: 38).

Pri izdelavi diplomske naloge sem si pomagala z vnašalnim sistem ZRCola[6], ki je bil razvit v dialektološki sekciji Inštituta za slovenski jezik Frana Ramovša na ZRC SAZU za potrebe Slovanskega lingvističnega atlasa (OLA; ena od članic Mednarodne komisije za izdelavo OLA s sedežem v Moskvi je tudi red. prof. dr. Vera Smole – Weiss 2004: 145). Fonetična transkripcija OLA ima za razliko od IPA dodana različna diakritična znamenja, kot so krožec pod zvočnikom za silabem ali Husovo tradicijo za šumevce (*č, ž, š*), piko pod *e* ali *o* za ozki izgovor, dvopičje za vokalom za dolgi izgovor oziroma resico za kratek (Karničar 2008: 159).

Za besediloslovne in pragmatične raziskave govorjene slovenščine se večinoma uporablja ortografska transkripcija. Te vrste transkripcije sicer nimajo skupnega standarda, sta pa Verdonik in Zwitter Vitez (2011: 38–39) na podlagi nekaterih raziskav ugotovili naslednje temeljne značilnosti:

- uporabljan je slovenski knjižni črkopis brez dodatnih znakov za posebne glasove, na primer za diftonge, izjemoma se ponekod pojavlja znak za polglasnik;
- pri teh vrstah zapisa so zabeleženi pojavi moderne vokalne redukcije in raznih pogovornih ter narečnih prvin govorjenega jezika (na primer *beu auto, tko je blo, mislm* ...), različne pa so rešitve posameznih avtorjev pri zapisu pogovornih oblik (*js* in *jes* za jaz, zapis polglasnika kot diskurznega označevalca – *ee* ali kot polglasnik ...).

V jezikovnotehnološki praksi pa se najbolj uporablja poknjižen zapis govorjene slovenščine, na podlagi katerega pa niso več razvidni pogovorni in narečni pojavi. Prisoten je le zapis določenih odstopanj, ki se v govorjenem jeziku najpogosteje pojavljajo (kratki nedoločnik, *pol* v pomenu *potem*).

Največja težava pri zapisovanju govora je, v kolikšni meri naj se ta zapis prilagaja govoru. Pri ortografski transkripciji besede zapisujemo v skladu z ustaljenim knjižnim zapisom, pri tem pa so številne informacije izgubljene ali celo potvorjene (na primer beseda *kle* je zapisana kot *tukaj*). Iskanje bi ponudilo neresnične oziroma neavtentične rezultate. Sklep vsega tega je, da se mora transkripcija bolj prilagoditi govorjenemu jeziku (Zemljarič Miklavčič 2008a: 134).

[6] Besedilo je bilo pripravljeno z vnašalnim sistemom ZRCola (http://ZRCola.zrc-sazu.si), ki ga je na Znanstvenoraziskovalnem centru SAZU v Ljubljani (http://www.zrc-sazu.si) razvil dr. Peter Weiss.

Fonemska transkripcija je dosti bolj prilagojena govorjenemu jeziku kot ortografska, a je hkrati tudi bolj zahtevno in zamudno opravilo, ki ga lahko opravijo le strokovnjaki. Posledično je zaradi zapletenih znakov oteženo tudi iskanje po korpusu, kar pa lahko rešimo z dodatnim zapisom – vkodirati je potrebno dodatni zapis besed, ki ustreza ustaljenemu knjižnemu zapisu in tako olajša iskanje po korpusu (Zemljarič Miklavčič 2008a: 135).

Za zapis na fonetični ravni uporabljamo standardni fonetični računalniški zapis simbolov za slovenski knjižni zapis MRPA. Zapis upošteva fonetično abecedo z osmimi samoglasniki (dolgi/kratki, naglašeni/nenaglašeni), šestimi zvočniki (z alofoni) in šestnajstimi nezvočniki (z alofoni). Pomanjkljivost te transkripcije je njena zamudnost; transkripcija večje količine govorjenih besedil bi bila preveč zamudna, za namene referenčnega korpusa pa, po mnenju avtorice, tudi nepotrebna. Smiselno jo je narediti na manjšem delu korpusa in v sodelovanju s fonetiki (Zemljarič Miklavčič 2008a: 136)

3.4 Računalniški simbolni fonetični zapis slovenskega govora

Na Slovenskem je bila v preteklosti večina razprav o transkripcijskih standardih namenjena slovenskim dialektom, manj pa nedialektalnemu pogovornemu jeziku. Računalniški fonetični zapis temelji na mednarodni fonetični abecedi IPA.[7] Za ohranitev jezikovne identitete in jezikovnega napredka posameznega jezika je potrebno delovanje na področju jezikovnih tehnologij, kar velja tudi za slovenščino, ki kot jezik z majhnim številom govorcev poenotenje fonetičnega zapisa vsekakor potrebuje (Zemljak idr. 2002: 159).

Mednarodna fonetična abeceda IPA je standard pri označevanju govora na ožji fonetični ravni in je ustvarjena po načelu »en glas en znak«, in sicer v vsakem jeziku za enak glasovni pojav, težijo pa k zapisu brez diakritičnih znamenj (Karničar 2008: 158). Za lažjo računalniško berljivost teh posebnih fonetičnih simbolov pa je pretvorjena v mednarodno uveljavljene fonetične simbole MRPA (*Machine Readable Phonetic Alphabet* − Zemljak idr. 2002: 160, Zemljarič Miklavčič 2008a: 119).

Sistem upošteva fonetične značilnosti slovenskega govora, kot so opisane v Slovenski slovnici iz leta 1976; torej osem samoglasnikov (dolgi in kratki, naglašeni in nenaglašeni),

[7] Slovensko adaptacijo IPA so ustvarili raziskovalci z ljubljanske in mariborske univerze. Namen tega sodelovanja je bil poenotiti raziskovalno delo na področju govornih in jezikovnih tehnologij oziroma standardizirati fonetične oznake slovenskih govorjenih besedil (Zemljak 2002: 159).

šest zvočnikov (z alofoni) in šestnajst nezvočnikov (z alofoni). Zemljak (2002) navaja le primere iz slovnice, ni pa primerov transkripcij realnega govora (Zemljarič Miklavčič 2008: 119).

Knjižni jezik pozna osem samoglasnikov – *i*, *e*, *E*, *a*, *O*, *o*, *u* in *ə*. Kvantitativno jih razdelimo na dolge (ti so samo naglašeni) in kratke (lahko so naglašeni ali nenaglašeni).

Tabela 7: Dolgi samoglasniki (Zemljak idr. 2002: 161).[8]

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
i:	*iː*	*pila*	*''p**i:**la*
e:	*eː*	*pet*	*''p**e:**t*
E:	*ɛː*	*teta*	*''t**E:**ta*
a:	*aː*	*mama*	*''m**a:**ma*
O:	*ɔː*	*voda*	*''v**O:**da*
o:	*oː*	*pot*	*''p**o:**t*
u:	*uː*	*suša*	*''s**u:**Sa*

Tabela 8: Kratki naglašeni samoglasniki (Zemljak idr. 2002: 161).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
i	*i*	*sit*	*''s**i**t*
E	*ɛ*	*zelen*	*zE''l**E**n*
a	*a*	*čas*	*''tS**a**s*
O	*ɔ*	*potop*	*pO''t**O**p*
u	*u*	*kruh*	*''kr**u**x*
@	*ə*	*tema, vrt*	*''t**@**ma, ''v**@**rt*

[8] Dolgi glasovi so označeni z dvopičjem (:), simbol " pa označuje začetek naglašenega zloga.

Tabela 9: Kratki nenaglašeni samoglasniki (Zemljak idr. 2002: 161).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
i	*i*	*prijava*	*pr**i**''ja:va*
E	*ɛ*	*zelen*	*z**E**''lEn*
a	*a*	*tema*	*''t@m**a***
O	*ɔ*	*potop*	*p**O**''tOp*
u	*u*	*konzul*	*''ko:nz**u**l*
@	*ə*	*veter, vrtiček*	*''vet**@**r, ''v@r''tič**@**k*

Soglasniški sistem pozna šest zvočnikov, šest zapornikov, šest pripornikov in tri zlitnike. Zvočniki so glasovi srednje odprtostne stopnje in ne vplivajo na zvenečnost nezvočnikov pred ali za seboj. Razen [*W*] in [*w*], ki pa nista fonema, nimajo ustreznih nezvenečih parov (Toporišič 2003: 77).

Tabela 10: Zvočniki (Zemljak idr. 2002: 162).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
m	*m*	*miza*	*''**m**i:za*
n	*n*	*noč*	*''**n**o:tS*
l	*l*	*leto*	*''**l**e:tO*
r	*r*	*riba*	*''**r**i:ba*
v	*v*	*voda*	*''**v**O:da*
j	*j*	*jeza*	*''**j**e:za*

Nekateri zvočniški fonemi ločijo alofone (Zemljak idr. 2002: 162):

- alofon fonema *m* je [*F*] in se pojavlja pred *f* ali *v*;
- alofoni fonema *n* so [*N*] pred *k*, *g* in *x*, [*F*] pred *f* ali *v* ter [*n'*] pred soglasnikom in na koncu besede;

- alofona fonema *l* sta [*l'*] pred soglasnikom in na koncu besede ter [*U*] za samoglasnikom in hkrati na koncu besede ali pred soglasnikom;
- alofoni fonema *v* so [*w*] pred zvenečimi soglasniki ne za samoglasnikom, [*W*] pred nezvenečimi soglasniki in ne za samoglasnikom ter [*U*] za samoglasnikom in hkrati na koncu besede ali pred soglasnikom;
- alofon fonema *j* je [*I*] za samoglasnikom in hkrati na koncu besede ali pred soglasnikom.

Tabela 11: Alofoni zvočnikov (Zemljak idr. 2002: 163).

Simbol fonema po MRPA	Simbol fonema po IPA	Simbol alofona po MRPA	Simbol alofona po IPA	Ortografski zapis	Fonetični prepis
m	*m*	*F*	*ɱ*	*nimfa, sem vesela*	*"ni:**F**fa, s@**F** vE"se:la*
n	*n*	*N*	*ŋ*	*Anka, angel, Anhovo*	*"a:**N**ka, "a:**N**gEl, "a:**N**xovo*
		F	*ɱ*	*konfin, konvoj*	*ko**F**"fi:n, ko**F**"vo:I*
		n'	*n^j*	*konj, konjski*	*"kO**n'**, "ko:**n'**ski/"kO:**n'**ski*
l	*l*	*l'*	*l^j*	*polj, poljski*	*"po:**l'**, "po:**l'**ski*
l	*l*	*U*	*u za soglasnikom*[9]	*cev, cevka* *tkal, tkalca*	*"ce:**U**, "ce:**U**ka* *"tka:**U**, "tka:**U**tsa*
v	*v*	*w*	*w*	*vzeti*	*"**w**ze:ti*
v	*v*	*W*	*ʍ*	*vsak*	*"**W**sa:k*
j	*j*	*I*	*I*	*lajna*	*"la:**I**na*

Nezvočnike, tj. glasove najmanjše odprtostne stopnje, lahko delimo glede na način tvorbe, in sicer na zapornike, pripornike in zlitnike. Prvi imajo v določenih položajih naslednje alofone (Zemljak idr. 2002: 165):

[9] Gre za zvočniško varianto − dvoglasniški *u̯*, ki jo lahko pišemo s črko *v* (*cev*, *cevka*) ali *l* (*tkal*, *tkalca*).

- alofona fonema *p* sta [*p_n*] pred *m* in [*p_f*] pred *f* ali *v*;
- alofona fonema *b* sta [*b_n*] pred *m* in [*b_f*] pred *f* ali *v*;
- alofona fonema *t* sta [*t_l*] pred *l* in [*t_n*] pred *n*;
- alofona fonema *d* sta [*d_l*] pred *l* in [*d_n*] pred *n*.

Tabela 12: Zaporniki (Zemljak idr. 2002: 163).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
p	*p*	*pipa*	*''**p**i:**p**a*
b	*b*	*beda*	*''**b**e:da*
t	*t*	*teta*	*''**t**E:**t**a*
d	*d*	*delo*	*''**d**e:lO*
k	*k*	*kolo*	***k**O''lo:*
g	*g*	*goba*	*''**g**o:ba*

Tabela 13: Alofoni zapornikov (Zemljak idr. 2002: 164).

Simbol fonema po MRPA	Simbol fonema po IPA	Simbol alofona po MRPA	Simbol alofona po IPA	Ortografski zapis	Fonetični prepis
p	*p*	*p_n*	*p̃*	*zob me boli*	*''zo:**p_n** mE bO''li:*
p	*p*	*p_f*	*p*^{f}	*ob fantu, rob vaze*	*O**p_f** ''fa:ntu, ''ro:**p_f** ''va:zE*
b	*b*	*b_n*	*b̃*	*z zobmi*	*z zO**b_n**''mi:*
b	*b*	*b_f*	*b*^{f}	*obveza*	*O**b_f**''ve:za*
t	*t*	*t_l*	*t*^{l}	*tla*	*''**t_l**la*
		t_n	*t̃*	*tnalo*	*''**t_n**na:lO*
d	*d*	*d_l*	*d*^{l}	*dleto*	*''**d_l**le:tO*
		d_n	*d̃*	*dno*	*''**d_n**nO*

Slovenski knjižni jezik pozna šest pripornikov, od teh sta zveneča le [*z*] in [*Z*]. Fonem *x* ima tudi alofon [*G*].

Tabela 14: Priporniki (Zemljak idr. 2002: 164).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
f	*f*	*figa*	*''**f**i:ga*
s	*s*	*soba*	*''**s**O:ba*
z	*z*	*zima*	*''**z**i:ma*
S	*ʃ*	*šoba*	*''**S**o:ba*
Z	*ʒ*	*žoga*	*''**Z**o:ga*
x	*x*	*hiša*	*''**x**i:Sa*
G	*ɣ*	*h gori*	***G** ''gO:ri*

Zlitniki so sičnik [*ts*] in šumevca [*TS*] ter [*dZ*].

Tabela 15: Zlitniki (Zemljak idr. 2002: 165).

Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
ts	*ts*	*cula*	*"**ts**u:la*
tS	*tʃ*	*čelo*	*"**tS**E:lO*
dZ1	*dʒ*	*džungla*	*"**dZ**u:Ngla*
dz2	*dz*	*Kocbek*	*"ko:**dz**bEk*

Nezvočnike lahko delimo tudi glede na zvenečnost. V slovenskem knjižnem jeziku se zvenečnost nezvočnikov spreminja glede na (ne)zvenečnost sosednjih fonemov. Nezveneči fonemi iz Tabele 16 ostanejo nezveneči v vseh položajih razen pred zvenečimi nezvočniki – tu postanejo nezveneči nezvočniki zveneči fonemi. Fonemi *ts*, *tS*, *f* in *x* imajo alofone [*dz*], [*dZ*], [*v*] in [*G*] (Zemljak idr. 2002: 165).

Tabela 16: Nezveneči nezvočniki (Zemljak idr. 2002: 165).[10]

		Besedna enota	
Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
t>d	*t>d*	*svatba*	*"sva:**d**ba*
s>z	*s>z*	*glasba*	*"gla:**z**ba*
S>Z	*ʃ>ʒ*	*izvršba*	*iz"v@r**Z**ba*
k>g	*k>g*	*prerokba*	*prE"ro:**g**ba*
ts>dz	*ts>dz*	*Kocbek*	*"ko:dzbEk*
tS>dZ	*tʃ>dʒ*	*enačba*	*E"na:dZba*
f>v	*f>v*	*Afganistan*	*av"ga:nistan*
x>G	*x>ɣ*	*h gori*	*g "gO:ri*
p>b	*p>b*	*snop daj*	*"snO**b** "daj*
t>d	*t>d*	*tat dobi*	*"ta:**d** "gre:/"grE*
s>z	*s>z*	*glas doni*	*"gla:**z** dO"ni:*
S>Z	*ʃ>ʒ*	*iščeš žogo*	*"i:StSE**Z** "Zo:gO*
k>g	*k>g*	*vsak dan*	*"Wsa:**g** "da:n*
ts>dz	*ts>dz*	*stric ga je*	*"stri:**dz** ga jE*
tS>dZ	*tʃ>dʒ*	*proč gre*	*"prO**dz** "grE*
f>v	*f>v*	*škof ga je*	*"SkO**v** ga jE*
X>G	*x>ɣ*	*strah ga je*	*s"tra:**G** ga jE*

Spremembe zvenečih nezvočnikov so deloma odvisne od položaja (ali nastopajo v besedni ali govorni enoti)[11]. V besedni enoti ostanejo zveneči pred samoglasniki, zvočniki in zvenečimi

[10] Že od strukturalizma naprej je vidna usmeritev, da je primarna varianta zveneča.

nezvočniki, v govorni verigi pa prihaja do izgube zvenečnosti pred samoglasniki, zvočniki, nezvenečimi nezvočniki in v izglasju – zveneči ostanejo le pred zvenečimi nezvočniki (Zemljak idr. 2002: 166).

Tabela 17: Zveneči nezvočniki (Zemljak idr. 2002: 166).

		Besedna enota	
Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
b	*b*	*biba*	*"**b**i:**b**a*
d	*d*	*dedi*	*"**d**e:**d**i*
g	*g*	*gol*	*"**g**o:l*
z	*z*	*zet*	*"**z**Et*
Z	*ʒ*	*žaga*	*"**Z**a:ga*
dZ	*dʒ*	*džezva*	*"**dZ**e:zva*

Tabela 18: Zveneči nezvočniki pred zvočniki (Zemljak idr. 2002: 167).

		Besedna enota	
Simbol fonema po MRPA	Simbol fonema po IPA	Ortografski zapis	Fonetični prepis
b	*b*	*objem*	*O"**b**jEm*
d	*d*	*dedje*	*"de:**d**jE*
g	*g*	*glava*	*"**g**la:va*
z	*z*	*zrelost*	*"**z**re:lOst*
Z	*ʒ*	*lažna*	*"la:**Z**na*

[11] »Izraz besedna enota se nanaša na nesestavljene besede ter na predložne in predponske zveze, govorna enota pa na vse druge zveze.« (Zemljak 2002: 165).

Tabela 19: Zveneči nezvočniki pred zvočniki (Zemljak idr. 2002: 167).

	Besedna enota pred nezvočniki			
	Nezveneči		Zveneči	
Simbol fonema po MRPA	Ortografski zapis	Fonetični prepis	Ortografski zapis	Fonetični prepis
b	*zobca*	*"zo:ptsa*	*obdati, ob dedu*	*Ob"da:ti, Ob "de:du*
d	*godčev*	*"go:ttSEU*	*odgnati, od gore*	*Od"gna:ti, Od "gO:re*
g	*bogcu*	*"bo:ktsu*	*Bogdan*	*"bo:gdan*
z	*grozd*	*"grOst*	*zdaj, čez deda*	*"zdaj, tSEz "de:da*
Z	*Nežka*	*"ne:Ska*	*žgati*	*"Zga:ti*

Tabela 20: Zveneči nezvočniki pred zvenečimi nezvočniki (Zemljak idr. 2002: 167).

	Govorna enota pred zvenečimi nezvočniki	
Simbol fonema po MRPA	Ortografski zapis	Fonetični prepis
b	*golob gre*	*gO"lo:b "gre:/gO"lo:b "grE*
d	*divjad gre*	*diw"ja:d "gre:/diw"ja:d "grE*
g	*rog zveni*	*"ro:g zve"ni:*
z	*rez boli*	*"re:z bO"li:*
Z	*gož gre*	*"go:Z "gre:/"go:Z "grE*
dZ	*govoreč glas*	*gOvO"rEdZ "gla:s/gOvO"re:dZ "gla:s*

4 TRANSKRIPCIJSKI STANDARDI KORPUSOV GOVORJENEGA JEZIKA

V nadaljevanju so podrobneje opisani nekateri korpusi in transkripcijski standardi, ki so jih uporabili ustvarjalci evropskih ter slovenskih korpusov govorjenega jezika pa tudi govornih zbirk za slovenščino. Izbira transkripcije je odvisna od namena korpusa ali govorne zbirke, specifičnih lastnosti jezika in jezikovne situacije. Opisani so korpus London-Lund, ki ima v listkovni obliki tudi prozodično transkripcijo, korpus Lancaster/IBM, ki ga med drugim uporabljajo za študij fonetike, korpus BNC, ki je z uporabo priporočil TEI požel nekatere kritike, in slovenski korpus GOS, edini korpus govorjene slovenščine. V nadaljevanju so opisane še nekatere govorne zbirke, osredotočila sem se le na tiste, ki sledijo določenim transkripcijskim načelom.

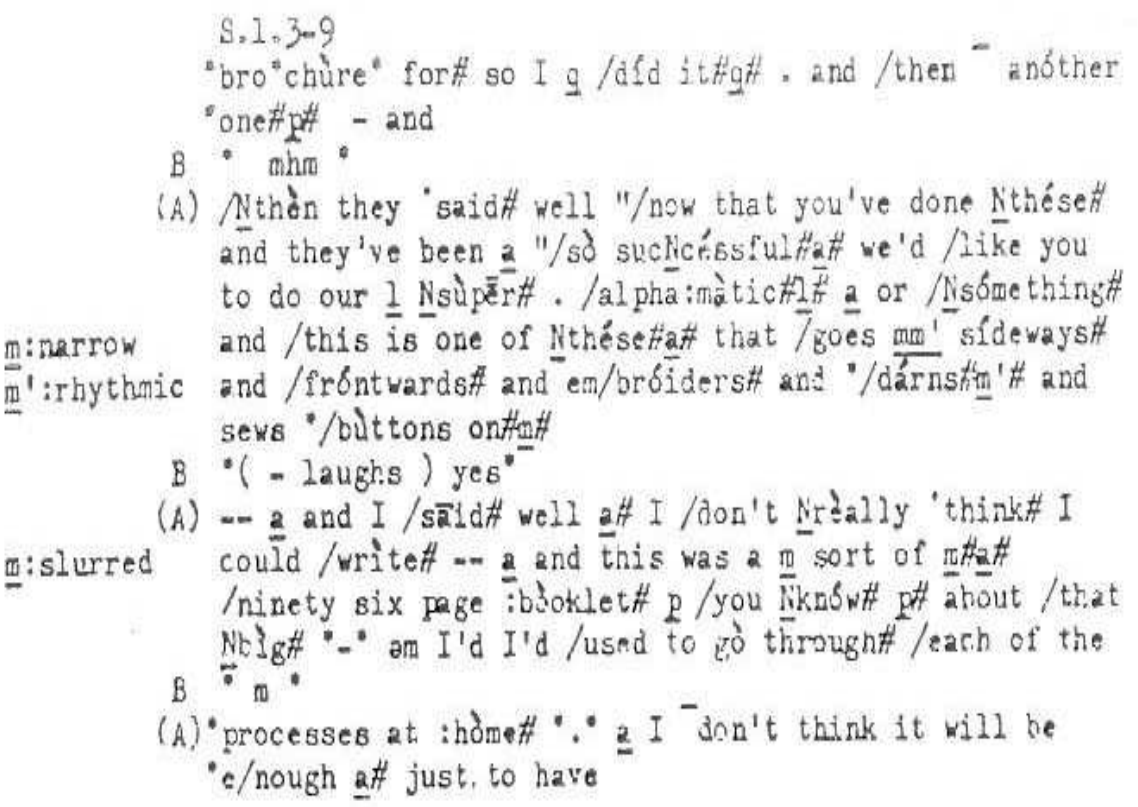
4.1 Prozodična transkripcija korpusa London-Lund

Korpus govorjene angleščine London-Lund Corpus (LLC) je za to diplomsko nalogo pomemben z vidika prve računalniške zbirke govorjenih besedil. Kot pove ime, je nastal na podlagi dveh projektov: korpusa SEU (Survey of English Usage, 1959), ki sodi še v obdobje predračunalniških govornih zbirk, in SSE (Survey of Spoken English, 1975), ki je bil zgrajen v švedskem mestu Lund, kjer so skušali zbrani in transkribirani material prenesti v obliko, primerno za računalniško rabo. Govorni korpus SEU vsebuje dvesto enako dolgih besedil, od tega sto govornih in sto pisnih, skupno pa milijon besed, in je označen tako prozodično kot tudi oblikoskladenjsko. Leta 1980 je bil del korpusa SEU prenesen v računalniško obliko. Gradivo LLC obsega sedeminosemdeset besedil originalne verzije, kateri je bilo dodanih še trinajst govorjenih besedil SEU-korpusa; LLC torej obstaja v treh različnih oblikah – listkovno kot del korpusa SEU, na CD-romu in v knjižni obliki (Greenbaum in Svartvik 1990, Zemljarič Miklavčič 2008a: 30–31).

LLC je najstarejši govorni korpus, ki naj bi služil kot vir za slovnični opis britanske angleščine. Sestavljajo ga monologi in dialogi. Prve so razdelili na spontani govor in na vnaprej pripravljeni govor, ki je lahko prosto govorjen ali zapisan; dialog pa je lahko zasebna konverzacija ali javna diskusija. Pogovor je še nadalje razdeljen na pogovor iz oči v oči, ki je

lahko posnet brez vednosti govorca ali pa z njegovo vednostjo, in pa na telefonski pogovor (Greenbaum in Svartvik 1990).

Ko govorimo o transkripcijskih standardih LLC, moramo razlikovati med prozodično transkripcijo in paralingvitično transkripcijo v SEU-korpusu od reducirane transkripcije LLC in računalniško obdelanih glasno branih sedemnajstih besedil. Obe vrsti transkripcij uporabljajo raziskovalci po vsem svetu za raziskovanje govorjene angleščine in primerjave med pisnim in govorjenim jezikom. Prozodično transkribiranje je zelo podaljšalo njegovo nastajanje. V prozodični transkripciji SEU-korpusa so označene meje tonskih enot s podenotami, lokacija zloga, ki je v intonacijski enoti najbolj poudarjen (v angleščini poudarek temelji na spremembi v višini, dolžini in glasnosti), različne dolžine premorov, različne stopnje poudarka (normalen ali močnejši). Oznake so še za različne stopnje hitrosti govora in glasnosti, razlike v kvaliteti glasu, paralingvistične prvine, kot sta šepet in krik. Z zvezdicama zgoraj je označen sočasni govor, komentarji so v oklepajih, nerazločne besede v dvojnih oklepajih, začetek tonske enote je označen s poševnico, konec tonske enote z veliko zvezdico, v oglatih oklepajih pa podenote. Angleščina pozna pet različnih vrst tonskega poteka (padajoči, naraščajoči, nevtralen, padajoči naraščujoči, naraščujoči padajoči) in še primere z dvojno intonacijo, jedro imenujemo, kjer pride intonacija najbolj do izraza (na primer na Sliki 1 so označeni *díd*, *anóther*, *thèn*, *sāid*). Oznaka *'booster'* označuje, kakšna je višina zloga za naglašenim (na Sliki 1 sta primera *alpha:màtic* in *:hòme*, kjer je zlog za dvopičjem višji od prejšnjega zloga). Simboli za poudarek so pozicionirani pred naglašenim zlogom, ki je lahko kjer koli v besedi. Pomembno vlogo ima še vezaj, če je presledek na primer pred njim, nakazuje pavzo v govoru (Greenbaum in Svartvik 1990, Zemljarič Miklavčič 2008a: 109).



Slika 4: Primer prozodične transkripcije korpusa London-Lund v listkovni obliki (Greenabaum in Svartvik 1990).

4.2 Večstopenjska transkripcija korpusa Lancaster/IBM

Naslednji obravnavani korpus je pomemben zaradi njegove rabe: uporabljajo ga namreč pri študiju fonetike. Projekt gradnje govornega korpusa britanske angleščine (*Spoken English on Computer*, SEC) se je začel leta 1984 s sodelovanjem Univerze v Lancastru in znanstvenega centra za raziskave govora podjetja IBM. Korpus vsebuje le nekaj več kot dvainpetdeset tisoč besed, a je po nastanku predstavljal za področje jezikovnih tehnologij velik napredek, saj naj bi bil v pomoč pri raziskavah tehnologij za sintezo in analizo govora. Dejansko so v IBM korpus uporabljali za raziskave govora, na Univerzi v Lancastru pa pri študiju fonetike. Največji delež besedil predstavljajo transkribirani posnetki radijskih oddaj BBC. Če je bilo besedilo nerazumljivo ali prezapleteno za transkribiranje (na primer prekrivni govor), ga preprosto niso uvrstili v korpus; podobno se je zgodilo z narečnimi besedili (Zemljarič Miklavčič 2008a: 32).

Korpus uporabljajo na Univerzi v Lancastru kot pomoč pri študiju fonetike, za kar pa dialogi, sicer večinski del govorjenih besedil, niso najbolj primerni. Korpus Lancaster/IBM je zanimiv tudi z vidika petih različnih verzij: zvočni posnetki, transkripcija brez ločil, z ločili (ortografska transkripcija), prozodična transkripcija in oblikoskladenjska verzija. V letih od 1992 do 1994 so korpus nadgradili, prvotni verziji so dodali zvočne posnetke in program za sinhronizacijo zvoka ter transkripcij, in ga preimenovali v MARSEC (*Machine Readable Spoken English Corpus* − Zemljarič Miklavčič 2008a: 33, 35).

Transkripcija brez ločil in prozodična transkripcija sta nastali neposredno po zvočnih posnetkih, transkripcija z ločili (t. i. ortografska transkripcija) na podlagi transkripcije brez ločil, oblikoskladenjsko označevanje pa na podlagi transkripcije z ločili. Vse verzije razen zadnje, narejena je bila polavtomatsko, so bile narejene ročno (Zemljarič Miklavčič 2008a: 112).

Transkripcija brez ločil in brez velikih začetnic skuša posnemati govor, kjer ločila večinoma nadomeščajo tonski poteki in premori. Označene so menjave govorcev. Ta transkripcija je bila podlaga za ortografsko in prozodično transkripcijo. Gradivo je večinoma nespontani govor, saj gre za radijske oddaje in (brana) univerzitetna predavanja (Zemljarič Miklavčič 2008a: 112).

good morning more news about the Reverend Sun Myung Moon founder of the Unification church who's currently in jail for tax evasion he was awarded an honorary degree last week by the Roman Catholic University of la Plata in Buenos Aires Argentina

Slika 5: Transkripcija brez ločil (Taylor in Knowles 1988).

Good morning. More news about the Reverend Sun Myung Moon, founder of the Unification church, who's currently in jail for tax evasion: he was awarded an honorary degree last week by the Roman Catholic University of la Plata in Buenos Aires, Argentina.

Slika 6: Transkripcija z ločili (Taylor in Knowles 1988).

Prozodično sta korpus označila dva strokovnjaka (iz Univerze v Lancastru in iz laboratorija IBM). Osnovna enota prozodične transkripcije je bila tonska enota, označena z dvema

navpičnima črtama. Tonski potek (visok, nizek, padajoč, rastoč, nizek padajoče-rastoč – vseh možnosti je štirinajst) je označen le na naglašenih zlogih (Zemljarič Miklavčič 2008a: 113).

001 SPOKEN ENGLISH CORPUS TEXT A01
In Perspective: Rosemary Hartill
Broadcast notes: Radio 4, 07.45a.m., 24th November, 1984
Transcribed by BJW on 6.1.86

↓good ˋmorning || ↑ˋmore -news about the ¯Reverend _Sun -Myung ˋMoon | _founder of the Unifiˋcation -Church | who's ˇcurrently in -jail | for ˋtax evasion || ↓he was a¯warded an _honorary deˇgree last -week | by the ¯roman _catholic Uni_versity of la ˋPlata | in -Buenos ¯Aires | Argenˋtina || in anˋnouncing the aˋward in New ˇYork | the ˇRector of the uni-versity | ↓Dr ¯Nicholas Argenˊtato | deˇscribed Mr -Moon as | a _prophet of our ˋtime || ↑ˇnext week | a ¯delegation of ˋnine protestant -ministers | ˇfrom Argentina | ¯visits the _Autumn As,sembly | of the _British -Council of ˋChurches || ↓it's _meant as a ¯symbol of reconciliˋation | between ,Christians | ,following the ,Falklands ˋwar || _Protestants howˉever ¯are | a _tiny miˋnority in Argen-tina | and the ¯delegation ˇwon't be including | a _Roman ˋCatholic || the assembly will ˇalso be dis-cussing | the _UK immiˇgration -laws | -Hong ,Kong | ˋteenagers in the ¯Church | ↑and of _course | -church ˋunity -schemes || in the ↑ˇFree Churches | there's some reˉnewed ˋgrumbling | ↓about -anglican amˋbivalence to the -British Council of -Churches || ↓though the ¯Anglicans still ¯talk | ↓about _doing as -much as ˊpossible with other -churches | ↑ˋsome Free ˇChurch people | -feel that in ˇpractice | the _Anglicans go it aˋlone whenever they -can || an ¯article in this week's -Baptist ˇTimes | asks ˋwhat -bishop wants to conˇfer | ↑when _he can have a ,camera | and a ,microphone | ,all to himˋself || ↓when the ¯Church of -England's General ˇSynod | can ˋnow get so much atˇtention from the -press | the ¯role of the ˋBritish Council of ˇChurches | ¯seems to _fade into the ˋbackground || ↑of course ˋwhat concerns -church ˋleaders | ¯isn't neces,sarily | ¯what -worries ordinary ˋchurchgoers | even ¯less the _general ˋpublic || ˇI can't recall | -ever _having -had a single ˋletter on the -British Council of -Churches and its -problems || -going by ˇmy postbag | ˋmost people are -worried about the -problem of ˊpain | the eˋxistence of ,God | and ˋmiracles || ↑they _want to know ,whether for ,instance | in a scienˇtific -age | ¯Christians can _really be_lieve | in the ,story of the ,feeding of the _five ˋthousand | as deˋscribed || or _was the ,miracle that those in the crowd ˇwith -food | _shared it with -those who had ˋnone || ↓is the ¯aid beginning to -flow into Ethi,opia | any ,less a -miracle | than the _five loaves and two ˋfishes || ↑well just ,recently | a ˇday conference on -miracles | was con,vened by the Reˇsearch -Scientists | -Christian ˋFellowship || it was ˋchaired | by Pro,fessor -Sir _Robin ˋBoyd | a ¯Fellow of the _Royal Soˋciety | and the ˇkey -question was | ¯when is a _miracle ˋnot a -miracle || I sup↑¯pose the ˬpopular notion of a -miracle is | an e_vent | unex_plained by ¯science or -natural ˋlaws || but the ˋscientists them,selves | ¯weren't having ˋany of -that || on ˇthis defi-nition they say | very ˋfew events | can ˋconfidently be -called ˇmiracles | because we have ˇno i-dea | ˋwhat natural -laws | may be dis_covered in the ˋfuture || more imˇportant how-ever | is that the ¯biblical _writers themˇselves | ¯thought that e_vents that ˇfollowed natural -laws | could ¯still be re_garded as miˋraculous || ↑take the _crossing of the -Red -Sea by the ˋIsraelites for -instance || ˋthis was made ¯possible by a strong ˋwind || the -wind itˇself | follows _patterns of -natural ˋlaw || the ˋmiracle | is that it _happened just ¯when | and _where it was ˋneeded || so it's the ˋtiming that can be mi-raculous || but the ˋreal -problem | the ,scientists ,say | with deˉfining ,miracles | as e_vents unex-plained by natural ˇlaw | is that it _springs from an ¯understanding of ˬnature | as a maˋchine | which _runs itˋself || ↑it's the _notion of an ˇimpersonal | ˇuncaring -universe | reˋlentlessly ,following ,laws of ,cause and efˋfect || ˇthat's a po-sition | ˇrecently ex-pounded | by _Don ˋCupitt | in his -BBˇC -TV -series | The _Sea of ˋFaith || ↑ˋthis underˇstanding say the -scientists | ˇis in -fact | -unscienˋtific || ↓and the -reason ˊis they -say | that ˇnatural laws | do ˇnot -cause | or dic_tate eˋvents || they're -merely deˇscriptions | of _what we exˋpect to -happen | on the ,basis of ,previous exˋperience || ↑the re_search -scientists ˊsay | that to -claim that ˇall events -follow these -patterns | is a philoˇsophical po-sition | and ˋnot | a scienˋtific -one || they ˋcall that philo-sophical position ˊnaturalism | and ˋthey say it is ˇquite -contrary | to the _christian -view of the ˋworld || the ˇchristian view they -argue | is to -see e_vents that can be ˇcovered by natural -laws | as _God's ˋusual ac-tivity || we get ˋso -used to the ˇusual -patterns | that we for,get how a_mazing they ˋare || but at ˇtimes they -say | ¯God _does ˇunusual things ˬtoo || ↓the ˋusual | ˋand the ˇunusual | are ˇboth God's -doing | and it's ˋmiracles | which con,front us with ,God most ˋclearly || ↑Don _Cupitt was pre-sumably ˇone of the -people | the _conference -members had in ˇmind | when they _emphasised ˇtheir -confidence in the -miracles | -as ˇstated in the -Bible | and deˋplored the -tendency of ˇsome | to _try to reˋduce them | to -merely meˇchanical e-vents | by sug_gesting -scientific explaˋnations || it's a ˇuseful re-minder | that ˇsome -scientists | -find _Don ˊCupitt | -unscien,tific || the deˉbate | goes ˬon ||

Slika 7: Prozodična transkripcija (Taylor in Knowles 1988).

4.3 Transkripcija govorne komponente Britanskega nacionalnega korpusa

Stomilijonski Britanski nacionalni korpus (BNC) je bil zgrajen v letih med 1990 in 1994. Gradnja je bila razdeljena med institucije in partnerje – za besedila njegove govorne komponente so poskrbeli v založbi Longman. Po velikosti ga sicer preseže korpus Bank of English, a ima BNC druge prednosti, kot sta na primer boljša dostopnost in večja uravnoteženost. Govorna komponenta obsega približno deset odstotkov BNC, torej podkorpus obsega deset milijonov besed. Razmerje med govornimi in pisnimi besedili je bilo določeno na podlagi ekonomske logike, saj sta zbiranje in transkripcija dosti dražja od preproste priključitve pisnih besedil v korpus. Pri zajemanju besedil so prvič uporabili metodo demografske klasifikacije govorcev in jo kombinirali s kontekstualno metodo zbiranja. S pomočjo statistične metode so določili reprezentativni vzorec govorcev britanske angleščine glede na spol, starost, regijsko pripadnost in socialni razred − tako je nastal demografski del govornega korpusa BNC (imenovan tudi konverzacijski podkorpus), ki obsega 4.206.058 besed, to je približno 40 % govorne komponente korpusa BNC (Zemljarič Miklavčič 2008a: 39–40).

Gradivo korpusa BNC je bilo transkribirano v zvezi s priporočili TEI, kar so zaradi lažje narave dela lahko opravljali nelingvisti. Posledično so pri pregledu transkripcij ugotovili veliko napak oziroma subjektivnih interpretacij zvokov, kot sta *əm* in *hm* z nestandardnimi zapisi. Adaptacije transkripcij po načelih TEI so bile narejene naknadno v Računalniškem centru Univerze v Oxfordu – transkriptorji torej niso zapisovali v zahtevanem sistemu, ampak so bile oznake dodane kasneje (Zemljarič Miklavčič 2008a: 115).

Priporočila TEI so pri nekaterih strokovnjakih naletela na velik odpor: težko so berljiva in težko zapisljiva. Izjava je po BNC definirana kot intonacijska enota. Transkriptorji naj bi ločila (piko, vejico, vprašaj in klicaj) vnašali po intuiciji in tako besedilo segmentirali v enote – stavke. Vnašali so jih le na mesta, kjer je bilo to skladenjsko ustrezno (Zemljarič Miklavčič 2008a: 116).

Osnovna enota je bila menjava govorcev (*speak turn*). Večina nestandardnih oblik je bila zapisana ortografsko. Nekomunikacijske zvoke so zapisali v transkripcijo le, če so bili relevantni pri razumevanju besedila (Zemljarič Miklavčič 2008a: 116–117).

<u who=PS04Y><s n=01296><w ITJ>**M**m <pause> <w ITJ>yes <pause dur=7><w PNP>I <w VVD>told <w NP0>Paul <pause><w CJT>that <w PNP>he <w VM0>can <w VVI>bring<w AT0>a <w NN1>lady <w AVP>up <pause> <w PRP>at<w NN1>Christmas-time<c PUN>**.**</u><u who=PS04U><s n=01297><w VBZ>**I**s <w PNP>he <w XX0>not<w VVG>going <w AV0>home <w AV0>then<c PUN>**?**</u><u who=PS04Y><s n=01298><w ITJ>**N**o <pause dur=8> <w CJC>and<w UNC>erm <pause dur=7> <w PNP>**I**<w VBB>**'m**<w VVG>leaving <w AT0>a <w NN1>turkey <w PRP>in<w AT0>the <w NN1>freezer<event desc="kettle boils"><s n=01299><w NP0>Paul <w VBZ>is <w AV0>quite<w AJ0>good <w PRP>at <w NN1-VVG>cooking<pause><w AJ0>standard <w NN1>cooking<c PUN>**.**</u>

Slika 8: Primer transkribiranega govora v BNC (http://users.ox.ac.uk/~lou/wip/silfitalk.pdf, 22. 11. 2012).[12]

4.4 Govorni korpus GOS

4.4.1 Izhodišča za definiranje pravil transkribiranja

Korpus GOS obsega milijon besed, zato je bilo potrebno transkribiranje omejiti na le najnujnejše podatke, kar bi pomenilo tudi večjo homogeniziranost transkripcij. Zato naj bi transkripcije vključujevale (Verdonik in Zwitter Vitez 2011: 39–40):

- pomembne informacije, pridobljene ob zbiranju gradivu, ki so nujne kot iskalni pogoj;
- besedilnozvrstno razvrstitev diskurzov, ki loči različne tipe govora;
- pragmatično pomembne informacije o strukturi diskurza;

kar naj bi:

- omogočalo hitrejše transkribiranje;
- nazorneje predstavilo dejansko podobo govorjenega jezika;
- omogočalo avtomatsko iskanje po različnih besednih oblikah, ki imajo enako oblikoslovno in semantično vlogo, a različne glasovne podobe;

[12] Odebeljeno so poudarjene velike začetnice in ločila.

- vsebovalo le najpomembnejše informacije o nebesednih in nejezikovnih zvokih, ki so pomembni za uporabnikovo boljše razumevanje diskurza.

Transkripcije ne vključujejo podatkov o mimiki, kretnjah, gestah in drugih nejezikovnih zvokih, saj je korpus GOS zasnovan kot jezikovni govorni vir. Kot prednost korpusa pa lahko izpostavimo dvojni nivo zapisa, pogovorni in standardizirani, kar uporabniku omogoča boljši vpogled v dejansko govorno podobo slovenščine. Pogovorni zapis poteka veliko hitreje, kot bi potekal fonetični zapis z abecedo SAMPA, uporabnikom pa omogoča boljše razumevanje transkripcij, saj ne vsebuje zapletenih fonetičnih oznak. Kljub temu so z omejenim naborom črk poskušali čim bolje predstaviti dejansko jezikovno podobo. Standardiziran zapis je namenjen boljšim iskalnim možnostim. Če določenega leksema ni v knjižni normi, so ga transkriptorji ohranili v pogovorni obliki (Verdonik in Zwitter Vitez 2011: 40–41).

Na osnovnem nivoju je delovalo kar dvajset transkriptorjev, ki so posnetke posneli in transkribirali na pogovornem nivoju. Te transkripcije so pregledali bolje usposobljeni transkriptorji, naknadno pa še validator. Standardizacijo je opravljala ena sama oseba, kar je zagotovilo visoko stopnjo homogeniziranosti standardiziranega zapisa (Verdonik in Zwitter Vitez 2011: 44–45).

4.4.2 Orodja za transkribiranje

Računalniško orodje za transkribiranje mora biti uporabniku prijazno, poleg tega pa podpirati dolge zvočne datoteke, šumnike, vnos metaoznak in podatkov o govorcih ter besedilnovrstnih podatkov o diskurzu. Poleg tega mora omogočati segmentacijo zvočnega signala na poljubne enote in enostavno vnašanje informacij o povezavah med deli transkripcije (Verdonik in Zwitter Vitez 2011: 41).

V slovenskem prostoru je bilo za izdelavo večine specializiranih govornih korpusov uporabljeno transkripcijsko orodje Transcriber. Nekatera orodja so na spletu prosto dostopna, a imajo določene pomanjkljivosti. Tako ima na primer program Exmaralda slabo povezavo med transkripcijo in zvočnim posnetkom – funkcije za pomikanje po zvočnem signalu niso dovolj natančne, kar upočasni transkribiranje. Program Praat ima sicer številne funkcije za akustično analizo, ki pa pri samem transkribiranju niso prijazne uporabnikom, poleg tega pa tudi ne podpira vnosa podatkov o govorcih in metaoznak. Transcriber je pri nas nabolj

uporabljano orodje: odlikujejo ga dobra povezava med posnetkom in transkripcijo, enostavna segmentacija posnetka, vnos metaoznak in (sicer omejen) vnos nekaterih podatkov o govorcih in diskurzu. Njegova največja pomanjkljivost pa je, da ne omogoča zapisa hkratnega govora več kot dveh govorcev (Verdonik in Zwitter Vitez 2011: 42–43).

Na podlagi vseh prednosti in pomanjkljivosti so se ustvarjalci korpusa GOS odločili za Transcriber, saj je uporabniku najbolj prijazen in je namenjen transkribiranju televizijskih informativnih oddaj, prednost pa je tudi dejstvo, da je bil za transkribiranje slovenskega govora že večkrat preizkušen. Kot rečeno, je njegova največja pomanjkljivost ta, da ne omogoča zapisa hkratnega govora več kot dveh govorcev (Verdonik in Zwitter Vitez 2011: 43).

4.4.3 Zapis govora

Pri zapisu govora je bilo predpostavljenih več ciljev, in sicer »hitro in enostavno transkribiranje, dejanska podoba diskurza, avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a z različnimi glasovnimi podobami« (Verdonik in Zwitter Vitez 2011: 57). A vseh teh ciljev z eno samo rešitvijo ni bilo mogoče izpolniti.

4.4.3.1 Pogovorni zapis

Cilj pogovornega zapisa je transkripcija, ki čim bolj sledi dejanskemu govornemu stanju v čim bolj berljivi obliki. Govor je zapisan s slovenskim črkopisom, zapis pa ni skladen z normo le na mestih, kjer izgovorjena beseda odstopa od standardne izreke.

Glasovi, ki niso izgovorjeni, niso zapisani (*tud*, *mam*, *čevli*). Polglasnik ni zapisan pred zvočniki *m*, *n*, *r*, *l* (*zloml*, *hitr*), pri enoglasovnih predlogih in členkih, čeprav so izgovorjeni z njim (*s*), ter pri enozložnih besedah (*nč*). V dvo- ali večzložnih besedah (razen pred zvočniki) je lahko zapisan z *e* (*kešni*). Podobno zapisujemo redukcije pri oblikah pomožnega glagola *biti* in zaimka *si/se* (*najraj b vidu, da s* ...) ter redukcije za prihodnjik (*čev* za *če bom*, *nav* za *ne bo*). Izjema pri zadnjem pravilu so zveze s polnopomenskimi glagoli, ki so zapisane kot dve besedi (*rekla m* za *rekla bom*). Premene po zvenečnosti v transkripciji niso upoštevane (*tud dobr, grandž scena*). Dvoustnični *v*, ki ni nosilec zloga, je lahko zapisan s črko *v* (*prov,*

davn, gledov), pa tudi s črko *l*, če je tako v normi (*kosil, mel*); če je *u* nosilec zloga ali je predlog *v* izgorjen samoglasniško, pa s črko *u* (*pršu, u tem*).

Diftongi in drugi pokrajinsko specifični fonemi, ki jih ni v knjižnem jeziku, so, odvisno od izgovorjave, zapisani z najbližjimi ustreznimi črkami, na primer *rejs, tujdi, puole, rjekla, nea, duobru* ... Težje so zapisljivi bolj specifični fonemi: *u* s preglasom (panonsko narečje) je lahko zapisan z *u* ali z *i*, zveneči *h* (primorsko narečje) z *g* ali s *h*, mehkonebni *r* (koroško narečje) z *r* ... Medmeti so zapisani z nizom črk, ki najbolje ustreza dejanski izgovorjavi, podaljšani polglasnik ter zvočnik *m* ali *n* s kombinacijami kot mašila pa z nizom treh črk (*eee, een, mmm* ...).

Začetki izjav so zapisani z malo začetnico. Označeni so nekateri pomembnejši prozodični pojavi: ? (vprašaj) označuje vprašanje, ! (klicaj) se pojavlja ob vzkliku ali ukazu in ... (tropičje) za mejo med izjavami istega govorca. Posebne oznake veljajo tudi za petje: če je zapetih le nekaj besed ali krajši verz, je besedilo neoznačeno in preprosto transkribirano, če pa je zapeta cela pesem, je označena kot premor oziroma prazna izjava. Nerazumljiv govor je označen z oznako [neraz]. Z oznako [glas] so označeni pojavi, ki so pomembni za razumevanje vsebine diskurza, kot so jok, zehanje, vzdih, medtem ko zvoki, ki ne nosijo sami po sebi sporočila, niso označeni. Pri smehu nista označeni glasnost in dolžina, pač pa je označeno, kdo se smeji: [smehgo], če se smeji le govorec, [smehna], če se smeji naslovnik oziroma občinstvo, in [smehob], če se smejijo vsi. Zvoki, ki so zunanjega izvora in ne nastanejo z govorili in vplivajo na potek diskurza, so označeni z oznako [ok]. Taki zvoki, ki pragmatično niso pomembni, niso označeni.

Ostale jezikovne strukture so transkribirane različno. Korpus ne loči prirednih in podrednih zloženk – vse zloženke so zapisane skupaj in brez vezaja. Domača imena so zapisana tako, kot jih izgovarjamo, a po pravopisu z veliko začetnico (*Brežice*). Večbesedna stvarna in zemljepisna lastna imena so dodatno označena z zavitimi oklepaji (*{Osnovna šola Ivana Cankarja}*). Citatne besede so zapisane tako, kot jih izgovarjamo, enako velja za tuja lastna imena, le da imajo veliko začetnico (*Hjuston, {Nju Jork}*). Kjer se v besedilu pojavi več prvin tujega jezika, je označeno z [tujjez]; če gre le za posamezno besedo, to ni posebej označeno. Osebni podatki so anonimizirani, in sicer označeni z [ime], [priimek], [ulica], [podjetje] ... Kratice so zapisane tako, kot so izgovorjene (*erteve, trr*), podobno velja tudi za naslove spletnih strani (*veveve pika ... pika si*). Številke so vedno zapisane z besedo (Verdonik in Zwitter Vitez 2011: 58–62).

4.4.3.2 Standardizirani zapis

Razloga, da so celotno gradivo korpusa GOS naknadno transkribirali še v standardni zapis, sta boljše iskalne možnosti in avtomatsko oblikoslovno označevanje. Pri tem so odpravili glasoslovne premene, saj je ciljna oblika zapisa standardna različica istega leksema, ostalih jezikovnih ravni niso spreminjali. Leksemi, ki jih v knjižni normi ni, so ostali zapisani pogovorno (Verdonik in Zwitter Vitez 2011: 62).

Za tematiko fonetike so najbolj zanimive spremembe zapisa na glasoslovni ravni, in sicer pri redukcijah (*bl* v *bolj*, *kolk* v *koliko*, *tolk* v *toliko* ...) in pri premenah (*a uš* v *a boš*, *druzga* v *drugega*, *jejli* v *jedli*, *jeskat* v *iskat/i* ...).

4.4.4 Konkordančnik za govorni korpus GOS

V preteklosti je bil za jezikovnotehnološko rabo korpusa značilen zapis govora v knjižni normi, kar pa popači resnično jezikovno podobo. Ta vrsta transkripcije ne vključuje redukcij glasov in neknjižnih besednih oblik ter besed. Uporabnik korpusa tako ne dobi pravega pogleda v dejansko jezikovno podobo, zato so se ustvarjalci korpusa GOS odločili za dva nivoja zapisa: pogovornega in standardiziranega (Verdonik in Zwitter Vitez 2011: 40).

Prednost obravnavanega korpusa je v povezavi zvoka in zapisa, vsako konkordanco je namreč mogoče poslušati v kontekstu. Konkordančnik omogoča specifično iskanje za bolj zahtevne uporabnike, pa tudi za manj zahtevne, saj lahko iskanje zlahka usvoji vsak. Posebnost je še dvojni zapis govora – pogovorna in standardizirana različica (Verdonik idr. 2010: 12).

V tujih govornih korpusih so situacije različne. Britanski nacionalni korpus je prosto dostopen, njegova govorna podsekcija pa zajema le transkripcije brez izvirnih zvočnih posnetkov, obsega pa deset milijonov besed. Iskanje je omogočeno z vmesnikom XAIRA, rezultat pa nudi še podatke o pogostosti pojavitev, slovničnih strukturah, kolokatorjih in primerih konkordanc. Večina korpusov ne omogoča poslušanja (ruski, francoski, češki, italijanski in poljski), nasprotno pa nemški nacionalni korpus (Verdonik idr. 2010: 13).

Konkordančnik služi uporabnikom korpusa, zato je pomembno, komu točno je namenjen. Potencialne skupne uporabnikov korpusa GOS so raziskovalci, zaposleni v izobraženju in poklici v stiku z govorom (na primer pisci, tolmači, prevajalci, poklicni govorci ...).

Konkordančnik omogoča dva osnovna tipa prikaza rezultatov: v obliki konkordančnega niza ali v obliki seznama. Iskalne funkcije torej delimo na iskanje po konkordančnem nizu (enostavno ali napredno) in na iskanje po seznamu. Enostavno iskanje omogoča tako iskanje po pogovornem zapisu (samo po besedah), kot tudi iskanje po standardiziranem zapisu (privzeto po lemah, med narekovaji po besedah − Verdonik idr. 2010: 13).

Pogovorni zapis je v veljavnem slovarskem črkopisu (ne v fonetičnem zapisu), ki upošteva veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Nabor črk je omejen, zato se glede nanj skušajo čim bolj približati dejanski glasovni podobi. Iščemo po kanalu leme – iskalna beseda se pretvori v osnovno obliko in prikažejo se rezultati za vse oblike te besede; če vpišemo iskalni niz med narekovaji, iščemo po kanalu besede, to je vpisani obliki (Verdonik idr. 2010: 13).

Pri standardziranem zapisu so odpravljene glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ciljna oblika je knjižna različica istega leksema – na drugih jezikovnih ravneh ni sprememb. Če določenega leksema ni v knjižni normi, je ohranjen v obliki, ki se pojavi v govoru. V tem primeru iščemo le po kanalu besede, to je oblike, ki smo jo vpisali (Verdonik idr. 2010: 13).

Rezultate iskanja lahko filtriramo po tipih diskurzov ali govorcev – funkcija namreč omogoča iskanje med metapodatki o diskurzih in govorcih (spol, starost, izobrazba, regionalna pripadnost – lahko je enotna ali razpršena, prvi jezik). Iskanje po diskurzu omogoča delitev po regiji snemanja in letu snemanja (Verdonik idr. 2010: 14).

Zahtevejši uporabniki lahko uporabijo napredno iskanje, ki poleg funkcij, ki jih omogoča že enostavno iskanje, omogoča še dve novi: iskanje po bližini (v okolici petih besed) in iskanje po oblikoslovnih oznakah (le standardiziran zapis lahko omejimo dodatno omejimo z besedno vrsto in ostalimi oblikoslovnimi oznakami (Verdonik idr. 2010: 14).

Pri načinu iskalne funkcije seznam lahko uporabljamo nadomestne znake, na primer * za poljubno število znakov, ? za en znak (na primer vpišemo *''bed*''* v narekovajih, da dobimo samo te oblike rezultatov, kot so *beda*, *bedarija* in *bedak*). Pri tem lahko iščemo po

pogovornem ali standardiziranem zapisu, pri slednjem lahko filtriramo po oblikoslovnih oznakah, po tipih diskurza in govorcev pa zaenkrat še ne. Rezultati iskanja se ne pojavijo v obliki konkordanc, temveč v obliki seznama besed s podatki o frekventnosti in njeni standardni obliki (Verdonik idr. 2010: 15). Če nas na primer zanima beseda *beda* in jo iščemo po seznamu, nam ta izpiše pogovorni in standardizirani zapis, ki je v tem primeru enak, oblikoslovne lastnosti standardizirane oblike (torej samostalnik, občno ime, ženski spol, ednina, imenovalnik) in število pojavitev, tj. ena.

Oba iskalna načina sta med seboj povezana: če kliknemo na besedo v seznamu, se v zavihku konkordančni niz izdela ustrezen konkordančni niz s primeri rabe v kontekstu (Verdonik idr. 2010: 15).

Zadetki se prikažejo v obliki seznama konkordanc ali v obliki seznama besed s podatki o frekvenci in standardni obliki pri iskanju po seznamu. Rezultati iskanja po konkordančnem nizu so lahko prikazani v pogovornem zapisu ali označeni glede na tip diskurza (JI, JR, NN, NZ). Desno od vsake konkordance je zvočnik, ki omogoča poslušanje posnetka. Za več podatkov kliknemo na posamezno konkordanco. Tako dobimo podatke o razširjenem kontekstu v pogovornem zapisu in podatke o diskurzu in govorcu, dodatno pa še o razširjenem kontekstu, standardnem zapisu tega konteksta, korpusne oznake, shranimo v odložišče ... Rezultate lahko razvrstimo in/ali izvozimo. Lahko jih razvrstimo glede na jedrno besedo(-e) ali glede na levo ali desno sobesedilo (Verdonik idr. 2010: 15).

V tujini podobni referenčni korpusi obsegajo okoli deset milijonov besed ali več (na primer angleški, poljski, nizozemski, portugalski). Torej se lahko vprašamo, ali je beseda referenčni za govorni korpus slovenščine sploh primerna – zajeti vzorec namreč ni dovolj velik. Za rabo v jezikoslovju pa bi rabili večjo opremljenost besedila, na primer skladenjsko označevanje, fonetični zapis ... (Verdonik idr. 2010: 15).

4.5 Govorne zbirke

Pomembno poglavje v razvoju korpusnega jezikoslovja predstavljajo tudi govorne zbirke, ki so namenjene različnim raziskavam s področja jezikovnih oziroma govornih tehnologij. Sicer ne gre za prave govorne korpuse, a sledijo načelom načrtnega zbiranja, shranjevanja in transkribiranja posnetkov govora. Pri tem ne gre za zbiranje besedil iz različnih virov, ampak

so v ospredju posnetki govora, ki so opisani ter grafemsko ali kako drugače označeni. Pomembno vlogo pri snovanju infrastrukture, potrebne za razvoj govorne tehnologije, ima Center za jezikovne tehnologije na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru: infrastrukturo namreč sestavljajo baza izgovorjav SNABI, baza izgovorjav SpeechDat II, fonetični slovar ONOMASTICA s fonetičnimi transkripcijami slovenskih lastnih imen in korpus besed.

4.5.1 Slovenska nacionalna baza izgovorjav (SNABI)

Govorna zbirka SNABI (Slovenska nacionalna baza izgovorjav) je bila ustvarjena za gradnjo sistemov za avtomatsko prepoznavanje govora po telefonu. Predstavljajo jo trije segmenti: besede, mmc in lingua. Govorni signal je bil segmentiran, označen in fonetično transkribiran, vsebuje pa več kot petnajst tisoč posnetkov govornega signala (Zemljarič Miklavčič 2008a: 52; Kačič 2007: 10). Ročno označevanje govora je bilo izvedeno na fonemskem nivoju, poleg tega pa je bil za zagotavljanje ustrezne natančnosti večkrat zaporedoma izveden postopek avtomatskega označevanja govora in ročnega preverjanja označenega govora (Kačič in Horvat 1998: 102). Na podlagi zbirke SNABI je Melita Zemljak (1998: 68–78) analizirala foneme in alofone, ki so zapisani s SAMPA fonetično abecedo, alofoni so zapisani v oglatih oklepajih. Avtorica je v zbirki zaznala več primerov moderne vokalne redukcije, na primer *urej**E**va:lnik* proti *urej**@**va:lnik*, *p**O**zdra:wleni* proti *p**@**zdra:wleni* in *z**a**tSe:tek* proti *z**@**tSe:tek*, pa tudi vokalne harmonije, na primer *p**O**pu:st* proti *p**u**pu:st*. Pri nenaglašenih širokih *e* in *o* je v nekaterih narečjih pogosta tendenca k ožanju, zato se na primer pojavlja namesto kratkega nenaglašenega širokega *e* ozki *e* (*deve:t*, *moto:r*). Pogost je tudi pojav srednjih *e* in *o* (pred *j* in pred *w*, ki sta na koncu besede ali pred soglasnikom). Sicer pa je avtorica ugotovila, da se glasovi razlikujejo glede na to, ali se pojavljajo v tekočem govoru (mmc) ali v besedah, ali se pojavljajo na začetku, na koncu ali sredi besede oziroma stavka.

4.5.2 Govorna zbirka GOPOLIS

Govorna zbirka GOPOLIS (GOvorjena POizvedovanja o Letalskih Informacijah) je služila za razvoj sistema za razpoznavo govora in krmilnika dialoga pri govornih poizvedbah o letalskih informacijah. Vsebuje petnajst ur telefonskih pogovorov, večja pomanjkljivost pa so

neurejene avtorske pravice (Zemljarič Miklavčič 2008a: 52–53). Posneti dialogi so bili ročno zapisani, stavke so posebej razvrstili glede na temo dialoga, naknadno pa še na natančnejše skupine. Slovar vsebuje množico geselskih člankov, posamezno geslo pa je sestavljeno iz grafemskega zapisa, ožjefonetičnega prepisa, besedne vrste, skladenjskih lastnosti, semantičnih in pragmatičnih lastnosti ter polja za dopolnilni opis gesla. Pri snemanju pogovorov so pazili na pestrost starosti govorcev, izogibali so se govorcev z zelo izrazitim narečnim govorom. Zbirka vsebuje predvsem brani in ne spontani govor (govorcem oziroma bralcem so prikazali stavek, ki so ga skušali bolj ali manj spontano prebrati), kar je njena velika pomanjkljivost. Grafemski zapis predstavlja osnovni zapis posnetega govora – ta je zaradi omejenosti s črkovnimi znaki in posledične neskladnosti z realno zvočno podobo slabša izbira. Zato pa fonetika ponuja bolj ustrezne rešitve, in sicer foneme in njihove glasovne uresničitve (Dobrišek idr. 1998: 105–108).

Tabela 21: Izbor SAMPA računalniških simbolov za standardne fonemske zapise in prepise slovenskega govorjenega jezika (Dobrišek idr. 1998: 108).

IPA	SAMPA	Primer	IPA	SAMPA	Primer
Zaporniki					
p	p	/p/ipa	b	b	/b/eda
t	t	/t/Eta	d	d	/d/elO
k	k	/k/aSa	g	g	/g/oba
Priporniki					
f	f	/f/iga	s	s	/s/Oba
z	z	/z/ima	ʃ	S	/S/ola
ʒ	Z	/Z/oga	x	x	/x/iSa
Zlitnika					
ts	ts	/ts/ula	tʃ	tS	/tS/elO
Zvočniki					
m	m	/m/iza	n	n	/n/ega
j	j	/j/eza	l	l	/l/etO
r	r	/r/iba	v	v	/v/Oda
Samoglasniki					
i	i	s/i/t	e	e	p/e/t
ɜ	E	z/E/lEn	a	a	k/a/dar
ə	@	p/@/rst	ɔ	0	p/O/tOp
o	o	p/o/t	u	u	kr/u/x

Poleg simbolov za standardne glasove slovenskega jezika ponuja SAMPA še nekaj dodatnih simbolov, ki omogočajo ožji fonetični prepis, ter dodatne simbole za akustično-fonetični zapis govora (simbol za označitev premora v govoru, simbol za zvenečo in simbol za nezvenečo zaporo pred odporami zapornikov − Dobrišek idr. 1998: 108).

Tabela 22: Dodatni SAMPA računalniški simboli za ožji fonetični zapis in dodatni simboli za akustično-fonetični zapis slovenskega govorjenega jezika (Dobrišek idr. 1998: 108).

IPA	SAMPA	Primer	IPA	SAMPA	Primer
Dodatna zlitnika					
dz	dz	0[dz]i:U	dʒ	dZ	p0[dZ]ga:ti
Glasovne različice nekaterih zvočnikov					
m	m	[m]i:za	ɱ	F	ni:[F]fa
ŋ	N	sa:[N]kE	n	n	[n]e:ga
nʲ	n'	vo:[n']	j	j	[j]e:za
I	I	pO:kO[I]	l	l	[l]e:tO
lʲ	l'	Zu:[l']	v	v	[v]O:da
U	U	wze:[U]	w	w	[w]ze:U
ʍ	W	[W]sa:kli			
Dolgi samoglasniki					
iː	i:	p[i:]la	eː	e:	p[e:]t
ɜː	E:	p[E:]ta	aː	a:	s[a:]m
ɔː	0:	p[O:]kOI	oː	o:	p[o:]t
uː	u:	s[u:]Sa			
Dodatni akustično-fonetični simboli					
_	[_]ne:ga_	-	_Wsa:[-]k_		
=	_v0:[=]da_				

4.5.3 Baza izgovorjav SpeechDat

SpeechDat (1997) je bil evropski projekt gradnje govornih zbirk za razvoj sistemov telefonskega govornega dialoga za delo v realnem okolju. Baza SpeechDat II za slovenski jezik je bila osnovana na podlagi govora tisoč govorcev (Zemljarič Miklavčič 2008a: 53; Kačič 2007: 10). Predpisani vsebinski okvir za posamezni jezik vsebuje izolirane številke, zaporedja številk, naravna števila, denarne zneske, črkovane besede, uro, datum, odgovor

da/ne, imena mest, pogoste ukaze pri izvajanju posameznih aplikacij, fraze, fonetično uravnotežene besede in stavke. Snemanje je bilo izvedeno v okviru sedmih regij in desetih različnih narečnih področjih. Hkrati s snemanjem je potekalo tudi označevanje baze, to je slušno preverjanje pravilnosti izgovorjave, vnašanje podatkov iz vrnjenih vprašalnikov govorcev in označevanje motenj v govornem signalu. Skupina na FERI bazo uporablja pri razpoznavanju tekočega govora v okviru raziskav na področju avtomatskega večjezikovnega razpoznavanja govora in pri razvoju demonstracijskega sistema obratnega telefonskega imenika (Kačič in Horvat 1998: 102–103). Vsakemu od triinštiridesetih posnetkov je dodana ortografska transkripcija z dodanimi podrobnostmi (slišni govorni in negovorni dogodki: vzdihovanje, glasno dihanje, presluh in motnje na linijah). Besede so zbrane v fonetičnem leksikonu. Fonetična transkripcija, narejena avtomatsko s pretvornikom in naknadno ročno pregledana, je zapisana s SAMPA fonetično abecedo (Kaiser in Kačič 1998: 56).

4.5.4 Glasoslovni slovar Onomastica

Glasoslovni slovar Onomastica je slovar slovenskih lastnih imen. Zgrajen je bil v okviru mednarodnega projekta EU Copernius. Zajema vsa slovenska lastna imena glede na evidenco Urada za statistiko RS iz leta 1995 in ima dodane podatke o vrsti lastnega imena in fonetično transkripcijo (Kačič 2007: 12). Med lastna imena so bila uvrščena osebna imena, priimki, imena mest, krajev, imena ulic in imena podjetij. Slovar vsebuje 282 000 enot, vsaka posamezna enota pa vsebuje ortografski zapis, do pet fonetičnih transkripcij, informacije o etimološkem razvoju besede in njenem pomenu, ime osebe, ki je transkripcijo izvedla, ter njen komentar. Vse fonetične transkripcije so bile ročno preverjene. Slovar je bil namenjen razvoju aplikacij na področju avtomatske sinteze imen oziroma njihovega avtomatskega razpoznavanja (na primer za avtomatski telefonski imenik − Kačič in Horvat 1998: 103). Avtomatska fonetična transkripcija je bila izvedena s pomočjo oseminosemdesetih splošnih pravil za prevorbo grafemov v foneme (uspešnost je bila 91,3-odstotna) in triintridesetih dodatnih pravil za pretvorbo posameznih skupin imen (uspešnost je bila 94,85-odstotna). Pred tem so vsa imena ročno naglasili in razvili sistem za avtomatsko zlogovanje (uspešnost tega sistema je bila 94,85-odstotna, kar je odličen rezultat − Kačič 1998: 41).

Leksikon imen loči tri nivoje transkripcije – kvaliteto III imajo zapisi, narejeni s pomočjo avtomatske fonetične transkripcije, kvaliteto II ročno preverjeni zapisi in kvaliteto I še

dodatno preverjeni zapisi. Za definicijo pravil, ki se navezujejo na Slovenski pravopis, so uporabili metaznake in z njimi lahko opisali skupine glasov, na primer samoglasnike ali soglasnike. Transkripcije je dodatno pregledal strokovnjak s področja fonetike in glede na analizo napak so definirali dodatna pravila. Opisani postopek so ustvarjalci zbirke ponavljali tako dolgo, dokler ni bila ustreznost transkripcije dovolj velika. Pri definiranju pravil so upoštevali enaintrideset oznak fonetične abecede iz nabora, podanega na Elektrotehniški in računalniški konferenci ERK'97, in oznake, ki ponujajo možnost pomenskega razločevanja (Kačič 1998: 43−44).

Pravilnost transkribiranja je bila zmanjšana za mesto naglasa – ne obstaja namreč pravilo, ki bi ga univerzalno določalo. Sama transkripcija je pogosto odvisna od naglasnega mesta; posledično je prihajalo do napak pri širokem in ozkem *o*-ju ter pri širokem in ozkem *e*-ju, zato je bilo naglaševanje za to govorno zbirko izvedeno ročno. Dodatno težavo pa so predstavljala tuja osebna imena in priimki oziroma imena in priiimki, ki izhajajo iz tujih jezikov. Napačno transkribiranih enot je bilo pri slovenskih priimkih le 1,8 odstotka. Analiza je pokazala, da je bila napaka najpogostejša pri zaporedju fonemov *au* (kar 1,3 odstotka), ki je v slovenščini sicer zelo redko (Kačič 1998: 44-45).

Za avtomatsko zlogovanje so ustvarjalci korpusa razvili algoritem, in sicer na osnovi nosilca zloga – to je lahko samoglasnik ali diftong *ə+r* – in njegovih mej. Zlogovanje je bilo izvedeno po avtomatski fonetični transkripciji, njena napaka je bila 2,5-odstotna, večinoma v zaporedjih *nj* in *lj* (kar 1,25 odstotka − Kačič 1998: 45).

4.5.5 Govorni pomočnik Radiotelevizije Slovenije

Ko govorimo o zborni izreki, ne moremo mimo govora na naši nacionalni televiziji. Strokovnjaki s področja govora in fonetike poskrbijo, da so govorci primerno strokovno podučeni o pravorečni normi, ki je pomembna za širjenje govorne kulture v slovenskem prostoru – za to ima take vrste medij, kot je javna televizija, po mojem mnenju kar velike zasluge. Na izobraževalnem portalu spletne strani Radiotelevizije Slovenije (RTV SLO) lahko zasledimo spletno aplikacijo, imenovano Govorni pomočnik. Namenjena je sodelavcem RTV SLO, predvsem napovedovalcem, pa tudi interesentom, ki iščejo pomoč pri pravilni izgovorjavi. V naboru besedišča so tiste besede in besedne zveze, pri katerih pogosto

prihaja do težav pri izgovorjavi, ga pa ustvarjalci programa sproti posodabljajo. V bazi so tuja lastna imena, na primer lastna osebna in lastna krajevna imena, kot so Akbar Hašemi Rafsandžani (*ákbar hašémi rafsandžáni*), Sergej Ivanov (*sergêj ivanôv*, s komentarjem, da je pri osebnem imenu naglašen drugi *e*, da pa sicer pogosto napačno naglašujemo na prvem), Sölden (*zéldn*, s komentarjem, da je izgovor z *z*-jem, kot navaja tudi nemški slovar Duden) in Cerklje ob Krki (*cʹərklje ob krki*). Prednost te aplikacije sta zapis izgovorjave in možnost poslušanja posnetka, ponekod pa lahko zasledimo tudi razlago, zakaj so se ustvarjalci baze odločili za določeno izgovorjavo. Aplikacija je naravnana uporabniško prijazno, dostop do nje je preprost, primerna je tudi za laično uporabo, saj je zapisana s klasičnimi fonetičnimi znaki, njihovo razumevanje pa olajšajo tudi komentarji.

5 ANALIZA GRADIVA GOVORNEGA KORPUSA GOS

Gradivo korpusa GOS je obširno in od raziskovalca, ki ga zanimajo fonetične posebnosti, zahteva veliko časa in potrpljenja. Konkordančnik sicer do neke mere omogoča iskanje besed po končnicah ali predponah, kar pa za potrebe fonetičnih raziskav ne zadostuje. Z iskalno funkcijo seznam, vpisom dela besede, ki ga iščemo (na primer predpono *pre**, kjer zvezdica nadomešča izpuščene znake) in iskanjem po standardiziranem zapisu lahko pridemo do velikega nabora rezultatov, ki pa jih je treba še dodatno selekcionirati, saj iskalnim zahtevam ne ustrezajo vsi (na prvi strani zadetkov se med drugim pojavijo *dobro*, *malo* in *lepa*, torej osnovne oblike pridevnikov). Pri tako velikem številu zadetkov je selekcija rezultatov zamudno opravilo.

Pri iskanju primernih besed sem si pomagala s Slovenskim pravopisom in s priročnikom Slovenska zborna izreka, kjer sem iskala tipične primere izgovorjave določenih fonemov in jih tako vnašala v konkordančnik ter iskala odstopanja od pravorečne norme. Pri analizi gradiva sem se osredotočila na kakovost samoglasnikov, na naglasna mesta pri jakostnem naglaševanju, na soglasnike oziroma izgovor fonemov v govorni verigi, na zapis diskurznih označevalcev in na izgovor črke *l* oziroma tako imenovano elkanje.

5.1 Glasovi v slovenščini in zgledi iz korpusa GOS

Glasove v slovenskem jeziku sem podrobneje teoretično opisala že v podpoglavju 2.1, v analizi pa se bom osredotočila na njihove konkretne pojavitve. Pri izdelavi tabele sem si pomagala s Toporišičevim poglavjem o zapisovanju glasov, ki je v Slovenskem pravopisu (2003: 139). V prvem stolpcu so napisane in opisane posamezne glasovne pojavitve po abecedi, v drugem njihov zapis z določeno črko in glasovno okolico, če je potrebno, v tretjem pa konkretni zgledi iz korpusa GOS.

Tabela 23: Glasovi v slovenščini, njihov zapis in zgledi iz korpusa GOS.[13]

Glas	**Črka**	**Zgledi**
[*a*] kateri koli	*a*	*angel, viadukt*
[*ˈa:*] dolgi naglašeni	*á*	*vláda, izrázil, jábolko*
[*ˈa/ˈʌ*] kratki naglašeni	*à*	*bràt, čàs, fànt, gràh, pràg, ràk, zràk, mràz*
[*a*] nenaglašeni	*a*	*francóščina, téma*
[*b*]	*b*	*babica, obramba*
	p pred zvenečim nezvočnikom	*lep dan*
[*c*]	*c*	*cesar, Bistrica*
[*dz*]	*c* pred zvenečim nezvočnikom	/
[*č*]	*č*	*počivaj, čuden*
	dž pred samoglasnikom	*Madžarska*
	c pred nezvenečim šumevcem	/
[*dž*]	*dž*	*džamija, džip*
	č pred zvenečim nezvočnikom	*proč zel, enačba, več zelenja*
	c pred *dž, ž*	/
[*d*]	*d*	*produkt, od Ljubljane*
	t	*svatba, pot domov*
[*e*] kateri koli	*e*	*seniku, lune, egoisti*
[*ˈe:*] dolgi naglašeni ozki	*é*	*želéti, péljem, vzéti, debélega, debéla*
	é pred *r,* navadno približan *i*	*véra, večér*
[*ˈe̯:*] dolgi naglašeni srednji	*ê* pred *j* na koncu besede	*idêj*
	pred soglasnikom	*idêjni*
[*ˈe̯*] kratki naglašeni srednji	*è* pred *j* na koncu besede	*jèj*
	è pred drugim soglasnikom	*dèjte*
[*e̯*]	*e* pred *j* na koncu besede	/
	e pred drugim soglasnikom	*mejnik*
[*ˈɛ:*] dolgi naglašeni široki	*ê*	*sêja, mêja, vêja, idêja, afêra, Amêrika, šofêr, dêrbi, zelêna, študênta, življênje, zaposlênost*
[*ˈɛ*] kratki naglašeni široki	*è*	*žèp, lèv, mèč, rèp, študènt, deklè, odcèp*
[*ɛ*] naglašeni široki	*e*	*obetáven, brême*
[*ə*] polglasnik	*e*	*danes, pripovedovalec*

[13] Znak ' pomeni naglašenost, dvopičje (:) dolžino, *o̯* in *e̯* pa srednjost.

	ø pred *l, r, m, n* v ustreznem sklopu	*rdeč*
[*'ə*] naglašeni	*è*	*dèž, vèndar, predvsèm*
	ø[14] pred *l, r, m, n* v ustreznem sklopu	*vŕt, pŕt, smŕt, vŕh, gŕme, čŕn*
[*ə*] nenaglašeni	*e*	*dánes, Slovénec, Bôvec*
	ø pred *l, r, m, n* v ustreznem sklopu	*vétrnica, méter, minístrski, kóprski, decêmber*
[*f*]	*f*	*piflarji, fakulteta*
[*g*]	*g*	*globalnega, pogost*
	k pred zvenečim nezvočnikom	*kdaj, k frizerju*
[*h*]	*h*	*hinavščina, Pohorje*
[*ɣ*]	*h* pred zvenečim nezvočnikom	*h gojenim*
[*i*] kateri koli	*i*	*mislim, ideal*
[*'i:*] dolgi naglašeni	*í*	*rasíst, bolník*
[*'i*] kratki naglašeni	*ì*	*dìm, mìš, nìt, rìt, ptìč, uvrstìt*
[*i*] nenaglašeni	*i*	*môški, finále*
[*j*]	*j*	*jablana, troje*
	i v polcitatnih besedah	*material*
	i na začetku domače besede, če se prejšnja končuje na samoglasnik	*pa itak, zelo izkušen*
	ø[15] med *i* in samoglasnikom	*pacienta, socialno*
[*k*]	*k*	*kolona, žakelj*
	g pred nezvenečim nezvočnikom	*bog kot, poleg tega*
	pred premorom	*nog*
	pred samoglasnikom	*poleg une*
	pred zvočnikom naslednje besede	*predlog je, poleg magistrata*
[*l*]	*l*	*obilo, zalet*
	lj na koncu besede ali pred soglasnikom	*hmelj, bruseljsko*
[*l̄*]	*lj* na koncu besede ali pred soglasnikom	*hmelj, bruseljsko*
[*l'*]	*lj* na koncu besede ali pred soglasnikom	*hmelj, bruseljsko*
[*m*]	*m*	*omara, zamudil*
[*ɱ*] zobnoustnični	*m*	*ampak, simfoniji, sem vedela*
[*n*]	*n*	*zakon, napiše*
	nj ne pred samoglasnikom	*konj, kranjska*

[14] Znak ø pomeni polglasnik, ki ni zapisan s posebno črko.
[15] Znak ø pomeni *j*, ki ni zapisan s posebno črko.

[n̄]	*nj* ne pred samoglasnikom	*konj, kranjska*
[*n*'] mehčani	*nj* ne pred samoglasnikom	*konj, kranjska*
[ŋ] mehkonebni	*n* pred *k, g, h*	*anketa, bančni*
	nj pred *k, g, h*	*manjkajo, pomanjkljivost, primanjkljaj*
[*o*] kateri koli	*o*	*opus, zanos, obiskati*
[ˈ*o*ː] dolgi naglašeni ozki	*ó*	*svój, bós, štós*
[ˈo̯ː] dolgi naglašeni srednji	*ô* pred *u̯*	*fluôrov, lôvski*
[ˈo̯] kratki naglašeni srednji	*ò* pred u̯	*kòl, navzdòl*
[*o*] nenaglašeni srednji	*ô* pred *u̯*	*beljákov*
[ˈɔː] dolgi naglašeni široki	*ô*	*člôvek, glôboka, hôtel*
[ˈɔ] kratki naglašeni široki	*ò*	*globòk, stròp*
[ɔ] nenaglašeni široki	*o*	*procès, oklèp*
[*p*]	*p*	*paziti, opera*
	b pred nezvenečim nezvočnikom	*absolutno, poslabšala*
	pred premorom	*klub, navkljub*
	pred samoglasnikom	*kljub izgubi, ob isti*
	pred zvočnikom naslednje besede	*ob močnem, kljub nuji*
[*r*]	*r*	*traktor, praska*
[*s*]	*s*	*spači, okus*
	z pred nezvenečim nezvočnikom	*brez prepiranja, čez travo*
	pred premorom	*dokaz, kviz*
	pred samoglasnikom	*iz ust, brez enologa*
	pred zvočnikom naslednje besede	*brez navdiha, iz leta*
[*š*]	*š*	*poboljšati, seštevku*
	ž pred nezvenečim nezvočnikom	*bržkone, delež tega*
	pred premorom	*čudež, glaž*
	pred samoglasnikom	*delež od*
	pred zvočnikom naslednje besede	*dež mora*
	s pred nezvenečim šumevcem	*sčasoma*
	z pred nezvenečim šumevcem	*iz Škotske, brez šip*
[*t*]	*t*	*trese, tabela, odpret*
	d pred nezvenečim nezvočnikom	*odcep, redko*

	pred premorom	*izhod, Kad*
	pred samoglasnikom	*od iraškega, na dočmi*
	pred zvočnikom naslednje besede	*nad vsemi, pod najbolj*
[*u*] kateri koli	*u*	*sum, presunil*
[*ˈu:*] dolgi naglašeni	*ú*	*kúhar, pústi*
[*ˈu*] kratki naglašeni	*ù*	*mùh*
[*u*] nenaglašeni	*u*	*očétu*
	v/l v ustreznem zvočniškem okolju	*bŕv*
	v/l ob nenaglašenih *ə, i* in *e* (večinoma v oblikoslov. kategorijah)	*okúsil*
[*ʋ*] zobnoustnični	*v*	*veseljem, opravičila*
	u v polcitatnih besedah	*/*
	f pred zvenečim nezvočnikom	*Afganistan, šef gledal*
[*u̯*] "dvoglasniški"	*v* kadar se premenjuje z *v*	*siv, živ, življenje*
	l kadar se premenjuje z *l*	*delal, cel*
	u v redkih domačih besedah	*nauk, poudarek*
	v besednih zvezah	*bo uspela*
	v polcitatnih besedah	*/*
	če je predpona	*malo ušla*
[*w*] zveneči dvoustniški nedvoglasniški	*v* ne ob samoglasniku	*vzajemno, v golfu*
	v kot predlog pred samoglasnikom	*v interesu*
	l na koncu besede in za *r*	*umrl*
[*ʍ*] nezveneči dvoustniški nedvoglasniški	*v* ne ob samoglasniku in pred nezvenečim nezvočnikom	*vpis, predvsem, vtičnice*
	u če je predpona	*uspela*
[*z*]	*z*	*zelo, ozrimo*
	s pred zvenečim nezvočnikom	*glasbeni*
[*ž*]	*ž*	*žago, požreti*
	š pred zvenečim nezvočnikom	*/*
	z pred zvenečim šumevcem	*brez žensk, iz župe*
	s pred zvenečim šumevcem	*danes že*

Kot rečeno, iskanje zgledov je v korpusu zelo oteženo, saj konkordančnik le delno omogoča iskanje zapisov posameznih glasov ali glasov v določeni glasovni okolici. Tako sem iskala samo primere zgledov, navedenih v strokovni literaturi, ali pa sem navedla zglede, ki so se

pojavili naključno med konkordancami. Posledično nisem našla v korpusu zgledov za spodaj opisane glasove, dopisala pa sem zglede iz Slovenskega pravopisa (2003: 139):

- glas [*dz*], ki ga zapišemo s *c* pred zvenečim nezvočnikom (na primer *Kocbek*);
- glas [*č*], ki ga zapišemo s *c* pred nezvenečim šumevcem (na primer *stric šepa*);
- glas [*dž*], ki ga zapišemo s *c* pred *dž* ali *ž* (na primer *stric Džon*);
- glas [*e̯*], ki ga zapišemo z *e* pred *j* na koncu besede (na primer *Tonej*);
- zobnoustnični [*v*], ki ga v polcitatnih besedah zapišemo z *u* (na primer *Dachaua*);
- dvoglasniški [*u̯*], ki ga v polcitatnih besedah zapišemo z *u* (na primer *Dachau*);
- glas [*ž*], ki ga zapišemo s *š* pred zvenečim nezvočnikom (na primer *izvršba*).

5.2 Samoglasniki in njihove variante v korpusu GOS

O samoglasnikih sem pisala v podpoglavju 2.1.1, njihove glasovne variante so razvidne tudi iz Tabele 23, v tem poglavju pa so opisno opredeljeni nekateri tipični fonetični problemi oziroma odstopanja od pravorečne norme v slovenskem jeziku.

Dolgi in široki *e* [*ê*] je v slovenskem knjižnem jeziku izgovorjen dosti redkeje kot dolgi ozki *e* [*é*], zato si zasluži posebno obravnavo. Pravorečno izgovorjavo si lahko olajšamo tudi z nekaterimi pravili. Naglašeni *e* pred *j*, ki mu sledi samoglasnik, je dolg in širok, a predvsem v ljubljanski in gorenjski regiji je mogoče zaslediti nekaj odstopanj: beseda *idêja* je tako izgovorjena s polglasnikom, torej [*idə̀ja*] (deset govorcev iz ljubljanske regije in ena govorka iz kranjske), enako *mêja* kot [*mə̀ja*] (štirje govorci iz ljubljanske regije in en iz kranjske) in *vêja* kot [*və̀ja*] (ena govorka ljubljanske regije). Dolg in širok je tudi naglašeni *e* pred *r* v prevzetih besedah, kot je *afera*, ki jo govorci dosledno pravorečno izgovarjajo [*afêra*]. Kot nasprotje je v neprevzetih besedah naglašeni *e* pred *r* ozek in izgovorjen blizu *i*-ja: tako govorci besedo vera v besedni zvezi »*ala ji vera*« izgovarjajo z [*vêra*], vero kot zavest o obstoju boga pa [*véra*]. Dolgi široki *e* izgovarjamo tudi v imenovalniku pri pridevnikih, ki jih premično naglašujemo: torej *vêlik* in *dêbel*. Pri slednjem pa se pojavi kar nekaj odstopanj, saj veliko govorcev izgovarja [*debél*], kar ni pravorečno.

Najmanj odstopanj od pravorečne norme najdemo pri visokih samoglasnikih, torej kratkih in naglašenih *i* in *u*. Najznačilnejše odstopanje se pojavi pri besedi *kùp*, ki se sicer najpogosteje tako izgovarja, zasledimo pa lahko tudi [*kèp*], pogovorno zapisano kot *kep*, in [*kàp*], zapisan kot *kap*.

Za kratek in naglašen samoglasnik *a* je prav tako značilna sprememba v *ʌ*, ki zveni dokaj polglasniško, na primer [*brʌt*] namesto [*bràt*]. Pri omenjeni besedi se pojavita še dve odstopanji: popolnoma polglasniški izgovor, torej [*brət*], ter *e*-jevski izgovor – [*brèt*]. Najdemo pa lahko veliko nedoslednosti pri zapisu; polglasniška izgovorjava namreč ne pogojuje pogovornega zapisa *brt*, beseda se namreč pojavi zapisana tudi kot *brat*. Prav tako se v korpusu pojavijo še izgovorjave [*čʌs*] za *čas/čs*, [*fʌnt*] za *fant/fnt*, [*prʌg*] za *prag*, [*rʌk*] za *rak*, [*zrʌk*] za *zrak/zrk* in [*mrʌz*] za *mraz/mrz*. Pri vseh teh primerih pogovorni zapisi niso dosledni.

Kratko so naglašene tudi nekatere nepregibne besede, kot je prislov *menda*. Slovar slovenskega knjižnega jezika dopušča dve naglasni mesti, torej [*mènda*] in [*mendà*]. Bolj razširjena in priporočljiva oblika naj bi bila prva (Šeruga Prek in Antončič 2003: 45), a kljub temu se v govoru pojavi tudi oblika z ozkim *e*-jem [*ménda*].

Polglasnik je v slovenščini predvidljiv in ga lahko opredelimo z nekaterimi pravili. Poznamo besede s korenskim polglasnikom (naslednji primer ima v imenovalniku ednine celo dva poglasnika – enega v korenu, drugega v zadnjem zlogu), na primer *pekel*, ki se v korpusu pojavi izgovorjen kot [*pəkə̀u̯*], Slovar slovenskega knjižnega jezika pa dovoljuje tudi izgovor [*pə̀kəu̯*]. Tipičen primer korenskega polglasnika je tudi samostalnik *mezda*, kjer pa Slovar slovenskega knjižnega jezika dovoljuje dve možni izgovorjavi, in sicer [*mézda*] in [*mèzda*], slednjo izgovarjamo s polglasnikom, torej [*mə̀zda*]. V korpusu najdemo dve pojavitvi, izgovorjeni z ozkim *e*, torej [*mézda*], ki pa ju artikulira isti govorec. Ostali primeri iz korpusa s polglasnikom v korenu besede so [*də̀ska*], [*də̀ž*] in [*stə̀za*], pojavijo pa se tudi nepravorečne *e*-jevske izgovorjave, kot so [*dèska*], [*dèž*] in [*stèza*]. Pravopis določa tudi polglasniško izgovorjavo besede *megla*, torej [*məglà*], lahko tudi *mègla*, a v izreki se zelo pogosto pojavlja *e*-jevski izgovor [*mègla*]. Tvorjenka *seveda* je nastala iz govorne podstave »*se ve, da*«, kar pa očitno ne pogojuje izgovorjavo besede same – pogosto se v govoru namreč pojavlja izgovorjena s polglasnikom [*sevə̀da*], pravorečno pa bi bilo [*sevéda*].

Pri izgovorjavi samoglasnikov torej prihaja do veliko odstopanj od pravorečne norme, kar pa je med drugim tudi posledica pestrosti narečij slovenskega jezika. Razlike se najpogosteje pojavljajo pri dolgem in širokem *e*, kratkem in naglašenem *a* ter seveda pri polglasniku.

5.3 Naglasno mesto

Naglasno mesto je v slovenščini nepredvidljivo, ni vezano na določen zlog, pač pa se navezuje na vsako besedo posebej in se ga skupaj z njo tudi naučimo. Pri tem si lahko pomagamo s štirimi naglasnimi tipi.

Pri večini govorcev se pojavljajo odstopanja pri premičnem naglaševanju večzložnih deležnikov na *-l*. V slovenščino namreč prodira nepremično naglaševanje, ki pa je tako po slovnici kot po pravopisu napačno (Šeruga Prek in Antončič 2003: 101). Tako najdemo v korpusu GOS nepravorečno izgovorjene nekatere predponske glagole na *-íti -ím* [*zaposlíl*], [*govoríl*], [*zapustíl*], [*pozvoníl*], [*obnovíl*], glagole na *-éti -ím* [*želél*], [*letél*], [*živél*], [*visél*], glagole na *-éti -em* [*hotél*], [*razumél*], glagole na *-íti -im* [*dovolíl*], glagole na *-àti -em* [*peljàl*] ... Omenjene oblike glagolov se pojavljajo tako v javnem, tudi moderiranem, kot nejavnem govoru. Sicer pa je iskanje deležnikov na *-l* v korpusu zelo oteženo, saj predvsem govorci iz osrednjeslovenske in gorenjske regije reducirajo množinsko obliko, izgovarjajo na primer *živel* namesto *živeli*. Tako najdemo pod zadetki vse pojavitve, tako deležnike na *-l* kot glagole v sedanjiku množine, primerne pa lahko določimo le s pomočjo konteksta.

Poznamo nekaj dvojnic, kot je na primer *zavod*. Rodilniško obliko lahko naglasimo [*závoda*] ali [*zavóda*] in tako se pojavlja tudi v korpusu. Enako lahko naglasimo množinsko obliko, v korpusu pa se pojavi le naglas na *o*, torej [*zavódi*]. Šeruga Prek in Antončič (2003: 127) priporočata javnim govorcem naglas na *o*.

Občasno je pomensko razlikovalen naglas pri besedi *zakon*: *zakon* v pomenu zakonska zveza ima premičen naglas, medtem ko pozna *zakon* kot pravni predpis tudi možnost nepremičnega naglasa, sicer jo navaja kot slabšo obliko. V prvem pomenu najdemo v rodilniški obliki tri pojavitve: od tega gre enkrat za javni govor na televiziji in je pravilno naglašen [*zakóna*], dvakrat pa za zasebni pogovor v štajerskem narečju, ki pa naglašuje nepravorečno - [*zákona*]. Po drugi strani pa veliko govorcev *zakon* v rodilniški obliki v pomenu prepisa naglašuje nepravorečno, torej [*zákona*], ne glede na to, ali gre za javni ali nejavni govor.

Pri naglaševanju samostalnikov velja omeniti še nekaj zanimivih primerov, ki jih govorci nepravorečno naglašujejo že v imenovalniku. Pogosto je tako naglaševanje besede *petelin*. V korpusu so tri pojavitve tega samostalnike, od tega vse tri nepravorečne, torej [*petelín*], pravilno bi bilo [*petêlin*]. V to skupino nepravorečno naglašenih lahko dodamo še besedo *Ukrajinka*, v korpusu sta prisotni dve pojavitvi, sicer v neformalnem govoru, obe z naglasom na *u*, torej [*úkrajinka*], medtem ko Govorni pomočnik priporoča naglaševanje na *i*. Besedo *dekle* naglašujemo na koncu besede, a se kljub temu pojavlja izgovarjava [*dêkle*], v korpusu značilna za govorce murskosoboške regije. Pri besedi *megla* SSKJ dovoljuje naglaševanje na poglasniku ali na *a*: v korpusu se res pojavljata obe obliki, vendar v obliki [*məglà*] le pri medijskem govoru govorke iz severovzhodne Slovenije. Podobna dvojnica je beseda *pekel*: v korpusu so tri pojavitve z naglasom na prvem polglasniku, torej [*pə́keu̯*], dve pa na drugem − [*pekə̀u̯*].

Če bi želeli narediti kvalitetno analizo naglasnih mest in samoglasnikov, bi nujno potrebovali prenos posnetkov v računalniški program za fonetično analizo, kot je na primer Praat[16]. Sicer so posnetki zaradi varovanja osebnih podatkov govorcev zavarovani, lahko pa bi do njih dostopali s podpisom individualne pogodbe in jih tako prenesli v program ter govor analizirali. Ponovno pa se pojavi težava, saj so posnetki dostopni v nesegmentirani obliki, ki ne omogočajo poslušanja in analize le ene pojavitve s kontekstom, ampak celotnih posnetkov. Taka vrsta analize bi bila zamudna, še posebej, če bi iskali na primer le vse pojavitve in variante naglaševanja besede *megla*.

5.4 Soglasniki in izgovor fonemov v govorni verigi

O soglasnikih sem pisala v podpoglavjih 2.1.3 in 2.1.4, sicer ločeno o zvočnikih in nezvočnikih. Tako bodo obravnavani tudi v naslednjem podpoglavju, kjer bom izpostavila izgovor glasov v govorni verigi. To je s fonetičnega vidika zanimiva tema, saj opisuje spremembe fonemov zaradi soseščine drugih glasov in tako nastanejo variante fonemov, kar je posebej lepo razvidno pri nepravorečnem govoru.

Pri variantah fonema *n* posebej izstopa zveza *nj* na koncu besede ali pred soglasnikom, ki jo lahko izgovarjamo na dva načina – kot palataliziran glas [*man'*] ali kot malo podaljšan zobni *n* [*maⁿ̄*]. Kljub temu pa se v govoru pojavlja posebna oblika izgovorjave, in sicer zamenjava

[16] Povzeto po elektronski korespondenci z doc. dr. Ano Zwitter Vitez (30. 1. 2014).

omenjenih fonemov, ki do neke mere poenostavi izgovorjavo. Tako se v korpusu pojavita besedi [*majn*] in [*kojn*], ki imata tudi pogovorni zapis temu primeren – *majn, kojn*.

Posebnost sta tudi dva enaka glasova, ki se zlijeta v enega daljšega. V korpusu lahko najdemo veliko takih primerov, na primer *bom mel, z začetkom* in *več časa* (pogovorni zapis), kjer se fonema izgovarjata kot en daljši oziroma se zlijeta, torej [*bomé̯u̯*], [*začétkom*] in [*večása*]. Še posebej je zlivanje enakih fonemov razvidno v primerih s predlogi, kjer pa pogosto pride do pretiranega poudarjanja predloga in tako ne pride do zlitja, na primer [*s sebój*] in [*iz zaljúbljene*].

V primerih, kjer si sledila dva enaka netrajna soglasnika, se pogosto izgovarjata kot en daljši, v Slovenski zborni izreki (2003: 142) pa je dovoljen tudi izgovor vsakega glasu posebej, še posebej takrat, ko želimo biti dovolj razumljivi. V korpusu se pojavi na primer beseda *oddaja*, ki je vedno izgovorjena s podaljšanim glasom *d* [*odája*], v besednih zvezah *pred današnjo, pred dobrima* in *ob bazenu* pa se pojavlja izrazito ločen izgovor: [*pret danášno*], [*pret dôbrima*] in [*op bazénu*].

V besedah, kjer sta kombinirana zobni zapornik in zlitnik (*t, d* + *c, č, dž*) lahko izgovarjamo oba glasova ali pa ustrezni dolgi glas. Tako najdemo v korpusu besedne zveze, kot sta *pod četrto* in *od Džeki*. V prvem primeru sta glasova izgovorjeno ločeno, torej [*potčetŕto*], v drugem pa se zlijeta v en glas [*odžêki*].

Tipičen primer, kjer namesto sičnikov (*s, z, c*) pred šumevci (*š, ž, č, dž*) izgovarjamo šumevce, je beseda *sčasoma*. V korpusu se pojavi dvakrat, obakrat izgovorjena s šumevcem, torej [*ščásoma*]. Zborna izreka (2003: 143) dovoljuje tudi ločen izgovor obeh glasov, ki zveni tako: [*sčásoma*]. Nekateri podobni primeri oziroma besedne zveze z zvezo sičnika in šumevca so še *s številko, s Črne* in *z žogo*, ki so izgovorjeni kot [*številko*], [*ščŕne*] in [*žógo*].

5.5 Diskurzni označevalci

Med pomembnejše prvine spontanega govora uvrščamo diskurzne označevalce. Ti so se v teoriji uveljavili kot izrazi, ki k vsebini in pomenu sporočila ne prispevajo veliko, pač pa imajo komunikacijske, pragmatične oziroma proceduralne funkcije (Verdonik 2008: 25–26). Za iskanje in analizo primernih diskurzivnih označevalcev v korpusu GOS sem uporabila naslednji postopek:

1. s pomočjo člankov Darinke Verdonik (2006 in 2008) sem naredila ožji izbor diskurzivnih označevalcev, ki se pogosto pojavljajo v govorjenih besedilih in so zanimivi za fonetične analize;

2. z iskanjem po standardnem zapisu sem preverila število pojavitev v korpusu;

3. poslušala sem vse pojavitve in si zabeležila odstopanja zapisa ter fonetične podobe.

Najpogostejši diskurzni označevalec v govoru je [*ǝǝǝ*] z različnimi glasovnimi uresničitvami in temu primernimi zapisi, kot so *eee*, *mmm*, *eem*, *emm*, *nnn*, *enn* in *een*. Glede na navodila naj bi jih zapisovali s tremi črkami in pri tem poskušali čim bolj slediti dejanski izgovorjavi, trajanja pa naj ne bi označevali. Zapisovalci so sicer poskušali slediti navodilom, vendar je prav ta vrsta diskurznih označevalcev zaradi specifičnosti izgovorjave (zamolkli zvoki) najbolj nerazumljiva in je zato prihajalo do napak v transkripciji. Verdonik (2008) jih uvršča v vrsto označevalcev procesov tvorjenja, pogosto imenovane tudi mašila (ta izraz bom zaradi pogostosti rabe in razumljivosti uporabljala tudi v nadaljevanju analize) ali zapolnjevalci vrzeli. Tako jih lahko tudi opišemo s pomenskega vidika: najpogosteje služijo za dodajanje časa za dokončanje ali nadaljevanje nedokončane misli oziroma izreka.

Najpogostejše mašilo je [*ǝǝǝ*], v korpusu zastopano več kot dvajsettisočkrat. Tu ne prihaja do odstopanj med govorom in zapisom, saj je mašilo preprosto in razumljivo, večinoma je del izjave in ne samostojna izjava, niti ne nastopa kot prekrivni govor. Pač pa je zanimiv pojav velika razlika med dolžino izgovorjave mašila: od manj kot sekunde pa celo dlje od treh sekund. Za natančnejšo analizo bi morala vsako glasovno uresničitev omenjenega mašila poslušati v izbranem računalniškem programu.

Več odstopanj pa se je pojavilo pri *mmm* (število pojavitev: 1522). Kot sem omenila, naj bi mašila služila zapolnjevanju vrzeli, a v tem primeru pogosto z vsebinskega vidika pomeni pritrditev – torej s rastočo intonacijo v drugem delu mašila. Enak pojav sem zasledila pri zapisu *mhm* (število pojavitev: 4474) – intonacija je rastoča, saj mašilo izraža strinjanje z izrečenim ali razumevanje izrečenega. Verdonik (2008) take primere zato uvršča med interakcijske označevalce. Primerna rešitev teh odstopanj bi bil poenoten zapis vseh mašil z rastočo intonacijo, ki izražajo strinjanje, z *mhm*, saj je njihova glasovna uresničitev praktično enaka.

Pri zapisu *mmm* se pojavi še kar nekaj odstopanj zapisa in glasovne uresničitve. V nekaj primerih gre za izgovor s polglasnikom, torej [*əəə*], kar pomeni napako zapisovalca, saj je standarden zapis te uresničitve *eee*. Zapisovalce je ponekod zavedla tudi dolžina izgovorjave mašila. Zasledila sem na primer podaljšan *mmm*, ki se nadaljuje v *ja* [*məəəja*], zapisano kot *mmm ja*. Primernejša rešitev bi bil zapis skupaj, torej *mmmja* z možnostjo iskanja posameznega dela besede (posamezno mašila *mmm* ali besede *ja*), saj bi s tem nakazali, da gre pri izgovorjavi dejansko za en sam element oziroma se en element brez prekinitve nadaljuje v drugega. Predlagana možnost bi bila zanimiva tudi pri raziskovanju drugih fenomenov nadaljevanja besed brez vmesne prekinitve. Podobno neskladje obravnavanega mašila je ponovno napaka zapisovalcev. Ko namreč poglasnik prehaja v *mmm* [*əəm*], sledi zapis *eee mmm*, kljub temu da med elementoma ni prekinitve in se mašilo nadaljuje v naslednjega. Primernejši bi bil zapis *emm* ali *eem*, odvisno od dolžine izgovorjenega polglasnika.

Do odstopanj je prišlo tudi pri zapisu *nnn*. Ta namreč v nekaterih primerih prehaja v besedo, ki se prične z *n*-jem, na primer [*nnnihče*], zapis pa je *nnn nihče*. Ponovno sem pri primeru, kot sem ga opisala pri analizi prej obravnavanega mašila. Kot rečeno, bi zapis *nnnihče* in možnost ločenega iskanja posameznih elementov ponujala dosti boljše možnosti za analize govora. Drugo odstopanje pa je skrajšan veznik *in*, ki se glasovno uresniči kot *n*. Tako je recimo [*ən pol*] zapisan z *nnn pol*, kljub temu da bi bila primernejša rešitev v pogovorni obliki *n pol*, standardizirani pa *in pol*.

Verdonik (2006: 31) poleg obravnavanega *mhm* k interakcijskim označevalcem uvršča tudi *aha*. V korpusu GOS se pod tem zapisom pojavljata dve glasovni uresničitvi, [*aha*] in [*əhə*], sicer pomensko identični. Glede na to, da smo pri mašilih uporabili za zapis polglasnika črko *e*, na primer *eee* [*əəə*], *eem* [*əəm*], bi lahko pravilo posplošili in določili, naj ga nadomešča še v primeru [*əhə*]. Vendar rešitev *ehe* ne bi bila primerna, saj niti približno ne odraža dejanskega stanja in zavaja uporabnika. Vprašanje primernega zapisa torej ostaja odprto.

5.6 Izgovor črke l v korpusu GOS in t. i. elkanje

Pri izgovoru črke *l* in njenih fonemskih variant si najlažje pomagamo s priročnikoma Slovar slovenskega knjižnega jezika in Slovenski pravopis. Slednji predlaga manj dvojnic kot SSKJ.

Sicer pa obstaja tudi nekaj pravil, ki olajšajo izgovor črke *l* (Šeruga Prek in Antončič 2003: 145).

Zvočniško varianto *u̯*, sicer pisano s črko *l*, izgovarjamo v izglagolskih izpeljankah s priponskimi obrazili *-lc-*, *-lk-*, *-lsk-*, *-lstv-* in v njihovih izpeljankah, če se nanašajo na vršilca dejanja. V korpusu se pojavi kar nekaj primerov nepravorečne rabe, in sicer v primerih vršilke ženskega spola. Med take primere spadata *igralka* in *nosilka*, kjer sta po dve pojavitvi izgovorjeni napačno, večinoma pa je pravilno uporabljen dvoustnični u [*u̯*] – predvsem v medijih.

V nekaterih primerih so dovoljene dvojnice, kjer lahko izgovarjamo [*l*] ali [*u̯*]. Tak tipičen primer je rodilniška oblika *tal*, ki se v korpusu pojavi osemkrat, od tega petkrat izgovorjena [*tál*] in trikrat [*táu̯*].

Dvoglasniški *u* izgovarjamo tudi v zvezah [*ou̯*] pred soglasnikom istega morfema. Zanimiv primer, ki deli mnenja, je *kolk*: v neformalnem govoru se je uveljavil izgovor [*kôlk*], v medijih pa prihaja v ospredje izgovor [*kôu̯k*]. Tako je v korpusu od osmih pojavitev polovica pravorečno in polovica nepravorečno izgovorjenih. Podoben primer je časovni prislov *pol*, na primer v besedni zvezi *pol osmih*, ki se zborno izgovori [*pol ôsmih*]. V pogovornem jeziku pa se večinoma uporablja *u*-jevski izgovor, torej [*pou̯ ôsmih*], medtem ko govorci moderiranega govora in govorci štajerskega narečja večinoma pravorečno izgovarjajo [*pol ôsmih*]. Nasprotje temu je t. i. mernostni prislov *pol*, ki se na primer pojavlja v besedni zvezi *pol kilograma* in se pravorečno izgovarja z [*u̯*], večinoma pa se tega držijo tudi govorci, na primer [*pou̯ métra*]. Je pa pri tem treba omeniti naslednji problem: ko iščemo v konkordančniku besedo *pol* (standardizirani zapis), dobimo skoraj tri tisoč zadetkov, ki pa lahko pomenijo časovni prislov, mernostni prislov in v pomenu *potem*. Zato je treba uporabiti napredno iskanje, a tudi to ni zadovoljivo, saj se pod pojavitvami, ki naj ne bi bile prislov, pojavijo prav te in obratno. Kadar je *pol* sestavni del tvorjenke, ga izgovarjamo z [*u̯*]; izjema pri tem je beseda *polotok*, prisotna tudi v korpusu, ki jo govorci dosledno pravilno izgovarjajo [*pólotok*].

6 ZAKLJUČEK

Za konec diplomskega dela naj izpostavim nekaj ključnih ugotovitev o korpusu govorjenega jezika v povezavi z analizami govora. Kot sem izpostavila že v uvodu, je nujen pogoj za kvalitetne fonetične raziskave obsežen in kvaliteten korpus fonetičnih transkripcij. Mnogi avtorji so poudarili, da je taka vrsta transkripcije zamudna in zahtevna, nekateri so omenili celo njeno nepotrebnost.

V teoretičnem delu naloge so opredeljeni trije standardi transkribiranja in označevanja: TEI, EAGLES in NERC. Priporočila TEI namenjajo veliko pozornosti neverbalnim dogodkom, kot sta kimanje in zvonjenje telefona, sicer pa prepuščajo uporabo fonetičnih oznak transkriptorjem – lahko jih uporabljajo dosledno ali pa jih sploh ne. Kot bolj običajna je navedena ortografska transkripcija z nekaterimi prozodičnimi oznakami, kot sta tonski potek in glasnost. Priporočila NERC prav tako priporočajo ortografsko transkripcijo z osnovnimi podatki o govorcih, prekrivnem govoru in neverbalnih dogodkih, enako velja za priporočila EAGLES, ki veliko pozornosti namenjajo reducirani obliki besed, dialektalnim oblikam, številkam, okrajšavam in medmetom, pa tudi neverbalnim dogodkom. Pri vsem tem pa velja omeniti, da NERC dovoljuje delno uporabo ločil, medtem ko pri EAGLES ta ostaja nedorečena. Sama bi se strinjala s priporočili za bazo izgovorjav SpeechDat, ki popolnoma odsvetujejo rabo ločil, saj transkriptor ta postavlja po občutku. Kot vemo, je v govoru težko določiti stavčne meje, zato je vsakršno predvidevanje ločil nesmiselno, saj ne moremo zagotoviti natančnosti.

TEI in NERC za fonemsko in fonetično transkripcijo priporočata zapis s simboli IPA, Zemljarič Miklavčič pa za zapis na fonetični ravni priporoča zapis MRPA, saj upošteva fonetično abecedo z osmimi samoglasniki, šestimi zvočniki in šestnajstimi nezvočniki, kot pomanjkljivost pa navaja njeno zamudnost in nepotrebnost za namene referenčnega korpusa. Pri transkripcijah MRPA in IPA prihaja le do nekaterih odstopanj: pri samoglasnikih zapis *E* namesto *ɛ*, *O* namesto *ɔ*, *@* namesto *ə*, pri alofonih zapornikov *F* namesto *ɱ*, *N* namesto *ŋ*, *n'* namesto n^j, *l'* namesto l^j in *W* namesto *ʍ*, pri pripornikih *S* namesto *ʃ*, *Z* namesto *ʒ* in *G* namesto *ɣ*. Zemljak (2002) v članku navaja le zapise primerov iz slovnice, ne pa tudi zapisov realnega govora. Zapis MRPA je namenjen preprostejši računalniški uporabi in enotnosti nadaljnih raziskav v slovenskem prostoru.

Korpus GOS je trenutno naš največji korpus govorjenega jezika in ga zlahka postavimo ob bok korpusom govora svetovnih jezikov, kot so LLC, Lancaster/IBM in BNC. Še posebej velja izpostaviti možnost poslušanja vsake od konkordanc v kontekstu v priloženi zvočni datoteki (te možnosti večina svetovnih korpusov še nima) in pa dvostopenjsko transkripcijo: pogovorni in standardizirani zapis. Medtem ko prvi skuša čim bolj slediti dejanskemu govornemu stanju z naborom črk slovenske abecede, pa drugi ponuja boljše možnosti iskanja in avtomatskega oblikoslovnega označevanja. Za diplomsko nalogo je bil ključnega pomena pogovorni zapis, ki pa z omejenim naborom znakov ni dosegel učinka, ki bi ga potrebovali za ustrezno analizo govora oziroma fonetično analizo. Dodatni znaki, ki bi slednje omogočali, bi prinašali podatke o tonemskem in netonemskem naglaševanju, o polglasniku, naglasnem mestu, alofonih in še mnogih drugih segmentih.

Pri tem ne smemo prezreti pomena govornih zbirk, saj te sledijo določenim načelom zbiranja in transkribiranja gradiva. Zbirke SNABI, GOPOLIS, SpeechDat in Onomastica uporabljajo nabor računalniških simbolov SAMPA, ki od IPE odstopa le pri zapisu nekaterih fonemov.

Konkordančnik korpusa GOS le delno omogoča iskanje po posameznih fonemih ali določenih glasovnih zaporednjih, ki se pojavljajo v sredini besed, kar zelo otežuje iskanje primernih zgledov; tak primer je beseda *izvršba*, kjer kombinacija črke *š* in zvenečega nezvočnika povzroči izgovor [*izvržba*]. Ta zgled v korpusu ni prisoten, naključno tudi nisem zasledila nobenega, iskalna zahteva **š* + zveneči nezvočnik v iskalni funkciji seznam in standardizirani zapis pa tudi ni dala ustreznih rezultatov. Pri analizi samoglasnikov sem naletela na nekaj nedoslednosti pogovornega zapisa: kratek in naglašen glas *a* govorci z nekaterih (predvsem osrednjeslovenskih) območij izgovarjajo polglasniško. Transkriptorji pa glede zapisa niso bilo poenoteni, saj je polglasniška izgovorjava ponekod zapisana z *a*, drugje pa brez ustreznega znaka (torej je [*brȁt*] nekje zapisan kot *brat*, drugje kot *brt*). Zanimivo se je pri analizi diskurznih označevalcev najpogosteje pojavljalo odstopanje pri zapisu mašila, ki se nadaljuje v besedo, na primer [*nnnihče*], zapisano kot *nnn nihče*. Sam zapis nam torej poda informacijo o dveh ločenih enotah, ne pa o eni sami enoti, kot kaže dejansko stanje.

Analiza govora iz transkripcij in posnetkov korpusa GOS bi bila zelo zahtevno delo, saj z naborom črk slovenske abecede, ki ga uporablja govorni zapis, težko sledi dejanskemu govornemu stanju. Delo sicer zelo olajšajo priloženi posnetki, ki pa so za resnično kvalitetno analizo govora neuporabni, dokler ne bo možnosti prenosa segmentiranih posnetkov v katerega od programov za analize govora, kot je Praat. Glede na to, da je korpus namenjen

raziskovalcem, zaposlenim v izobraževanju in poklicem, ki se kakor koli ukvarjajo z govorom, bi bilo vredno razmisliti o nadgradnji korpusa z informacijami, ki bi prinašale podatke o fonetičnih značilnostih govora.

Viri in literatura

Atkins, S., Clear, J., Ostler, N. (1991). *Corpus design criteria*. Dostopno prek: http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf, (28. 1. 2014).

Burnard, L. (2002). *Where did we go wrong? A retrospective look at the British National Corpus.* Dostopno prek: http://users.ox.ac.uk/~lou/wip/silfitalk.pdf, (28. 1. 2014).

Dobrišek, S., Gros, J., Ipšić, S., Pepelnjak, K., Mihelič, F., Pavešić, N. (1998). Gopolis: slovenska podatkovna zbirka govorjenih poizvedovanj. *Jezikovne tehnologije za slovenski jezik: zbornik konference*, 105–108.

Gorjanc, V. (2005). *Uvod v korpusno jezikoslovje*. Domžale: Izolit.

Govorni pomočnik. Dostopno prek: http://eizo.rtvslo.si/index.php?option=com_govorni&view=govorni, (13. 2. 2014).

Greenbaum, S., Svartvik, J. (1990). *The London-Lund Corpus of Spoken English*. Dostopno prek: http://icame.uib.no/london-lund/.

Jurgec, P. (2004). Antihiatski pojavi v knjižni slovenščini. *Jezikoslovni zapiski*, 10 (1), 125−144.

Jurgec, P. (2011). Slovenščina ima devet samoglasnikov. *Slavistična revija*, 59 (3), 243−268.

Kačič, Z. (1998). Izgradnja fonetičnega leksikona lastnih imen s pomočjo avtomatske fonetične transkripcije. *Uporabno jezikoslovje: revija Društva za uporabno jezikoslovje Slovenije*, 6, 41–50.

Kačič, Z. (2007). *Govorne tehnologije in jezikovni viri*. FERI.

Kačič, Z., Horvat, B. (1998). Izgradnja infrastrukture potrebne za razvoj govorne tehnologije za slovenski jezik. *Jezikovne tehnologije za slovenski jezik: zbornik konference*, 100–104.

Kaiser, J., Kačič, Z. (1998). Razvoj slovenske baze izgovorjav SpeechDat. *Uporabno jezikoslovje*, 6, 51–57.

Karničar, L. (2008). Fonetično zapisovanje narečnih etnoloških besedil. *Traditiones*: *zbornik Inštituta za slovensko narodopisje in Glasbenonarodnega inštituta ZRC SAZU*, 155–167.

Krek, S. (2012). *Slovenski jezik v digitalni dobi*. Heidelberg: Springer.

Rotovnik, T., Sepesy Maučec, M., Verdonik, D. (2006). Problemi razpoznavanja spontanega govora ob primerih iz govornega korpusa Turdis-1. *SloFon 1, 1. slovenska mednarodna fonetična konferenca, Ljubljana, 20.–22. April 2006*, 66.

Smolej, M. (2006). *Vpliv besedilne vrste na uresničitev skladenjskih struktur (primer narativnih besedil v vsakdanjem spontanem govoru).* Doktorska disertacija. Ljubljana.

Taylor, L. N., Knowles, dr. G. (1988). *Manual of Information to Accompany the SEC Corpus, The Machine-Readable Corpus of Spoken English*. Lancaster. Dostopno prek: http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM.

Tivadar, H. (2004). Fonetično-fonološke lastnosti samoglasnikov v sodobnem knjižnem jeziku. *Jezik in slovstvo*, 52 (1), 31–48.

Tivadar, H. (2005). Slovenski medijski govor v 21. stoletju in pravorečje - RTV Slovenija vs. komercialne RTV-postaje. *Kapitoly z fonetiky a fonologie slovanských jazyků*: *příspěvky z pracovního vědeckého setkání na XVI. zasedání Komise pro fonetiku a fonologii slovanských jazyků při Mezinárodním komitétu slavistů*, 209−226.

Tivadar, H. (2010). Normativni vidik slovenščine v 3. tisočletju – knjižna slovenščina med realnostjo in idealnostjo. *Slavistična revija*, 58 (1), 105−116.

Tivadar, H. (2012a). Nevarna razmerja med pisnim in govorjenim jezikom. *Pravopisna stikanja*, 203−211.

Tivadar, H. (2012b). Nove usmeritve pri raziskavah govora s pogledom v preteklost. Slavistična revija, 60 (4), 587−601.

Toporišič, J. (1992). *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba.

Toporišič, J. (2003). *Slovenski pravopis*. Ljubljana: Znanstvenoraziskovalni center SAZU, Založba ZRC.

Transcriptions of speech. TEI – Text encoding iniciative. Dostopno prek: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html, (28. 1. 2014).

Verdonik, D. (2006). *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora*. Doktorska disertacija. Ljubljana.

Verdonik, D. (2008). Govor in jezikovne tehnologije. *Spisi o govoru*, 231–240.

Verdonik, D. (2008). Označevanje vrste diskurznih označevalcev. *Zbornik šeste konference Jezikovne tehnologije, 16. do 17. oktober 2008*, 25–28.

Verdonik D., Zwitter Vitez, A., Romih, M., Krek, S. (2010). Konkordančnik za govorni korpus GOS. *Zbornik Sedme konference Jezikovne tehnologije*, 14. do 15. oktober 2010, [Ljubljana, Slovenia] : zbornik 13. mednarodne multikonference Informacijska družba - IS 2010, zvezek C, 12–15.

Verdonik, D., Zwitter Vitez, A. (2011). *Slovenski govorni korpus GOS*. Ljubljana: Trojina, zavod za uporabno slovenistiko.

Weiss, P. (2004). ZRCola: vnašalni sistem za jezikoslovno rabo v programu word. *Jezikoslovni zapiski*, 10 (1), 145−152.

Zemljak, M., 1998: Analiza fonemov in alofonov baze izgovorjav SNABI. *Uporabno jezikoslovje*, 6, 68–78.

Zemljak, M., Kačič, Z., Dobrišek, S., Gros, J., Weiss, P. (2002). Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija,* 50 (2), 159–168.

Zemljarič Miklavčič, J. (2008a). *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Zemljarič Miklavčič, J. (2008b). Iskanje odgovorov na vprašanja govorjenega jezika. *Jezik in slovstvo*, 53 (2), 89–106.

Zemljarič Miklavčič, J., Stabej, M., Krek, S., Zwitter Vitez, A. (2009). Kaj in zakaj v referenčni govorni korpus slovenščine. *Infrastruktura slovenščine in slovenistike*, 423−428.

Žganec Gros, J., Mihelič, F., Dobrišek, S. (2003). Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo*, 48 (3−4), 47–59.

IZJAVA O AVTORSTVU

Izjavljam, da sem diplomsko nalogo z naslovom *Korpus govorjenega jezika in njegova uporabnost za analize govora* napisala samostojno, s pomočjo navedene literature in pod vodstvom mentorja.

Ajdovščina, 13. marca 2014 Alenka Laharnar