Vprašanja zapisovanja govora v govornem korpusu Gos

Darinka Verdonik

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
Smetanova 17, SI-2000 Maribor
darinka.verdonik@um.si

Povzetek

Prispevek obravnava vprašanja, povezana z morebitno nadgradnjo referenčnega govornega korpusa slovenščine Gos, s poudarkom na nekaterih težavnejših vprašanjih zapisovanja govora. Morebitna nadgradnja v smeri skupne platforme z akustično bazo, potrebno za razvoj razpoznavanja tekočega govora, pri vprašanjih zapisovanja govora predvideva nadaljnji zapis po dvotirnem sistemu pogovornega in standardiziranega zapisa, vzpostavljenem v korpusu Gos. Vendar obstoječe gradivo Gosa kaže nekatere nedoslednosti in odprta vprašanja, ki bi jih bilo treba pri nadgradnji odpraviti. Izpostavimo štiri: vprašanje zapisovanja dvoustničnega 'U' in člena 'ta' v pogovornem zapisu, vprašanje zapisovanja neverbalnih in polverbalnih glasov ter vprašanje standardizacije nestandardnih polnopomenskih izrazov.

The questions of speech transcription in the speech corpus GOS

In this paper we discuss issues related with eventual upgrade of the reference speech corpus GOS, with special attention to questions concerning the speech transcription. It is likely that the upgrade of the GOS corpus will be joined with efforts to provide new acoustic speech database for continuous speech recognition. Nevertheless, the transcription of speech should follow the two-level transcription system (pronunciation-based and standardized transcription) specified in the GOS corpus. However, the existing transcriptions of the GOS show some inconsistencies and open questions that need to be discussed before the upgrade. In this paper, we discuss four such issues: the transcription of the sonorant phoneme U and the particle 'ta' in the pronunciation-based transcription, the transcription of non-verbal and semi-verbal sounds in the pronunciation-based and standardized transcription, and the transcription of non-standard lexical items in the standardized transcription.

1. Uvod

Konec leta 2010 je bil v slovenskem prostoru javnosti predstavljen prvi poskusni referenčni korpus govorjene slovenščine – Gos (Verdonik, Zwitter Vitez, 2011; Verdonik idr., 2013). Poskusni pravimo zato, ker pravi referenčni (pisni) korpusi obsegajo po več 100 milijonov besed, govorni le po nekaj milijonov, Gos 1 milijon. Kot s(m)o zapisali avtorji ob njegovi objavi na spletu, zato vsi upamo, da bo v prihodnosti še rasel, in ta prispevek izhaja iz tega upanja.

Toda zdi se, da vsakdanji govorjeni jezik ne požanje prav veliko zanimanja jezikoslovcev; niti toliko, kot ga kažejo sami govorci, ki se ob pomanjkanju aktivnosti na strani stroke v posameznih iniciativah lotijo tudi njegovega opisovanja (gl. npr. http://www.pokazijezik.si/ ali http://razvezanijezik.org/). Na drugi, tehnološki strani jezikovne tehnologije (z vmesnimi padci) vsake toliko »udarijo« z govornimi tehnologijami – na primer s govornim sistemom dialoga, kot je Siri na iPadu, ali s strojnim prevajanjem govora, kot je pred nekaj meseci Microsoft s prevajalnikom govora za Skype. Seveda (se bomo sploh kdaj znebili tega seveda?) pa so ti sistemi narejeni za velike tuje jezike, slovenščine ne pokrivajo, in tako je tudi na tem področju v slovenskem okolju zanimanje za govorjeni jezik prepuščeno le redkim posameznikom, ki se s temi tehnologijami ukvarjajo.

Ob tem, da prav velikega zanimanja za vprašanja vsakdanjega govorjenega jezika v stroki ne zaznamo, pa se hkrati čudimo, da tako pogost in vseprisoten pojav ostaja tako neraziskan in nezanimiv za raziskovalce (in s tem mislimo predvsem jezikoslovce). Morda pa je glavni razlog samo težavnost zbiranja in urejanja gradiva ter majhnost obstoječega govornega korpusa ... in s to mislijo smo se lotili njegove analize in načrtov za prihodnjo rast, optimistično odločeni, da se bo slednja prej ali slej zgodila.

2. Nadaljnja rast v smeri večje podpore razvoju tehnologij

Korpus Gos je bil načrtovan predvsem kot reprezentativni korpus za jezikoslovne raziskave. Kljub temu je vsaj del korpusa, zlasti tisti, ki predstavlja javni diskurz (to je nekaj 10 ur govora), mogoče uporabiti tudi kot akustično bazo za razvoj razpoznavanja tekočega govora,[1] čeprav ni v celoti prilagojen tovrstni uporabi.

Ob rasti akustičnih baz za razpoznavanje tekočega govora v mednarodnem prostoru in primerjavi z razpoložljivimi bazami za slovenščino (med temi predvsem BNSI Broadcast News – Žgank idr. 2004; slovenska govorna baza Broadcast News – Žibert, Mihelič, 2004; SloParl – Žgank idr., 2006) pa postaja več kot očitno, da je velika ovira za nadaljnji razvoj tovrstne tehnologije za slovenščino prav pomanjkanje ustreznega obsega akustičnih baz, ki se meri v tujini v nekaj sto urah, za slovenščino zaenkrat samo v nekaj deset urah. Na drugi strani je za obstoječi reprezentativni govorni korpus Gos prav tako treba načrtovati velik preskok v obsegu, v mednarodnem prostoru postaja referenčni obseg za primerljive vire ca. 10 mio. besed.

Aktualni akcijski načrt za jezikovno opremljenost (Dobrovoljc idr., 2014) predvideva v nadaljnjih načrtih na področju govornih korpusnih in akustičnih virov skupno platformo tako za nadgradnjo referenčnega govornega korpusa Gos kot akustične baze za razpoznavanje tekočega govora. V tej smeri je zastavljena tudi primerjalna analiza korpusa Gos z bazo BNSI Broadcast News kot predstavnikom akustične baze, izdelane za potrebe razpoznavanja govora, objavljena letos na

[1] Možnost uporabe gradiv za razpoznavanje govora velja samo za tisti del gradiv, ki predstavljajo javni diskurz. Posnetke in transkripcije je v ta namen mogoče pridobiti prek konzorcija CLARIN.SI.

konferenci LREC (Žgank idr., 2014). Povzemimo na kratko rezultate te analize.

Razlike med obema viroma so bile ugotovljene na petih ravneh:

(1) (Delno) različna je vsebina enega in drugega vira: reprezentativni govorni korpus zajema vzorce iz vseh najpogostejših tipov govorne interakcije, kar vključuje cel spekter od medijskega diskurza na eni do zasebnih pogovorov na drugi strani, akustična baza pa se osredotoča na zajemanje posnetkov s področij, kjer je najverjetnejša aplikacija uporabe, to so pri BNSI televizijski diskurzi (možnost aplikacije pri avtomatskem podnaslavljanju in prevajanju), sicer pa še drugi bolj kot ne formalni javni govorni nastopi (npr. parlamentarne seje, razna javna predavanja in predstavitve ipd., kjer je možnost aplikacije za avtomatsko izdelovanje dobesednih zapisov), za gluhe in naglušne pa je zanimivo področje aplikacije izobraževanje.

(2) Na akustični ravni je pri korpusu Gos ugotovljen veliko širši spekter različnih akustičnih okolij kot v bazi BNSI, hkrati pa zelo skopa označenost akustičnih okolij in slaba kvaliteta zajema avdio signala. Smernice za skupen jezikovni vir so zato predvidene v smeri večje kvalitete zajema avdio signala in bolj natančnega označevanja akustičnega okolja (zlasti tehnologije zajema signala in akustičnega ozadja, kot je npr. govor, hrup, glasba ...).

(3) Na ravni segmentiranja govora so ugotovljene razlike v načinu določanja segmentov, saj je pri akustični bazi veliko pozornosti usmerjene v ločevanje hkratnega govora in premorov, tudi iskanje mej med segmenti sledi v prvi vrsti ustrezno dolgim premorom v govoru, medtem ko v govornem korpusu označevanje hkratnega govora ni natančno, segmenti pa sledijo v prvi vrsti smiselno zaokroženim izjavam. Skupne smernice so predvidene v smeri, ki jo zastavi korpus Gos, z dodatkom, da se bolj kot v obstoječi praksi transkribiranja v korpusu Gos sledi načelu čim krajših segmentov, zlasti tistih, kjer se pojavlja hkratni govor, in da se bolj podrobno kot doslej označujejo premori v govoru.

(4) Naslednja razlika je opredeljena kot raven označevanja akustičnih dogodkov, to so razni vdihi, izdihi, tleski z jezikom ipd. oz. negovorni zvoki (npr. zvonjenje). Ti dogodki morajo biti v akustični bazi natančneje označeni, v govornem korpusu pa so bili označeni le pragmatično pomembni. Skupne smernice so predvidene v smeri bolj podrobnega označevanja.

(5) Različni praksi sta tudi na področju zapisovanja govora. V akustični bazi BNSI sledi zapis pravopisnemu standardu, če izgovorjava opazno odstopa od predvidene standardne, pa so dodane posebne oznake k takim besedam. V korpusu Gos pa je bil razvit dvotirni sistem zapisovanja govora, kar je bil učinkovit način za obvladovanje številnih izgovornih različic, ki se pojavljajo zlasti pogosto v nejavnem diskurzu. Skupne smernice predvidevajo nadaljevanje dvotirnega zapisovanja, in tega bomo prav zato v tem prispevku nekoliko podrobneje analizirali. Zapis govora je namreč do neke mere vedno interpretacija tistega, kar slišimo. Pri tem se srečujemo z mnogimi vprašanji, kako oblikovati načela zapisovanja, da bomo ohranili vse pomembne jezikovne prvine in hkrati omogočili čim večjo mero avtomatskega prepoznavanja posameznih prvin. Nekaterim od teh vprašanj, ki so se odprla ob uporabi korpusa, se bomo posvetili v nadaljevanju. Zagotovo pa to niso vsa vprašanja zapisovanja govora in zaželeno bi bilo, da se v prihodnosti vedno znova kritično ozremo nazaj.

3. Načela zapisovanja govora v Gosu

Zapisovanje govora v Gosu je bilo zasnovano po dvotirnem sistemu, ki je bolj kot ne unikaten tudi v svetovnem merilu (gl. npr. Verdonik idr., 2013). V specifikacijah korpusa (http://www.korpus-gos.net/Content/Static/Nacela_transkribiranja_in_oznacevanja_posnetkov_v_referencnem_govornem%20korpusu_slovenscine.pdf) je dvotirni zapis utemeljen in opisan takole:

»Pri zapisu govora se je hitro pokazalo, da nekaterih ciljev (hitro in enostavno transkribiranje, dejanska podoba diskurza, avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami) ni mogoče rešiti z eno samo rešitvijo.

Zato smo ustvarili dva nivoja zapisa govora: na prvem nivoju zapisa, ki ga imenujemo 'pogovorni zapis', zapišemo besede sicer ortografsko (ne fonetično!), vendar tako, kot so izgovorjene; na drugem nivoju, ki ga imenujemo 'knjižni zapis' (kasneje spremenjeno v 'standardizirani zapis', op. a.), pa 'poknjižimo' zapis na tak način, da različnim variantam neke besedne oblike (npr. *mam, jemam*) pripišemo krovno knjižno obliko (npr. *imam*).

Tako s prvim nivojem omogočimo dober vpogled v besedje in oblike govorjenega jezika, z drugim nivojem pa razširimo iskalne možnosti ter omogočimo uspešnejše nadaljnje avtomatsko označevanje besedil.«

Za ilustracijo, kako sta oba nivoja zapisa realizirana v praksi, navajamo v nadaljevanju primer iz Gosa:

Pogovorni: *ne sej tak eee tak ko si razložla men mislim veš kak je s temi sanjami ne*

Standardizirani: *ne saj tako eee tako kot si razložila meni mislim veš kako je s temi sanjami ne*

Vseh podrobnosti enega in drugega nivoja zapisa tukaj ne bomo obravnavali, pojasnjene so na spletni strani Gosa (www.korpus-gos.net) v priloženih specifikacijah in v monografiji (Verdonik, Zwitter Vitez, 2011).

4. Zapisi govora v Gosu v številkah

Gos vsebuje 1,035.101 besedo v standardiziranem zapisu. Tabela 1 prikazuje, koliko od teh besed je različnic na nivoju pogovornega in standardiziranega zapisa ter leme.

Tabela 1: Število različnic v Gosu

Nivo	**Št. različnic**
pogovorni zapis	82.648
standardizirani zapis	62.578
lema	31.294

Vsaki besedi v pogovornem zapisu je pripisana ena (izjemoma pa lahko tudi dve ali več) beseda v standardiziranem zapisu. Tabela 2 prikazuje, koliko je vseh tovrstnih parov različnic, koliko je identičnih in koliko neidentičnih ter nakaže strmo padanje frekvenc pojavitve pri neidentičnih parih. Strmo padanje je verjetno delno posledica majhnosti korpusa, je pa tovrstna krivulja frekvenc v jeziku nasploh značilna.

Tabela 2: Pari besed pogovorni – standardizirani zapis

Pari pogovorni – standardizirani zapis	**Število (% vseh parov)**
vseh parov	82.648
identičnih parov	54.822 (66 %)
neidentičnih parov	27.826 (34 %)
neidentičnih parov, ki se pojavijo več kot 5-krat	3.391 (4 %)
neidentičnih parov, ki se pojavijo več kot 100-krat	210 (0,25 %)
neidentičnih parov, ki se pojavijo več kot 1000-krat	18 (0,02 %)

Deset najpogostejših neidentičnih parov je naslednjih, po pričakovanju funkcijskih besed, saj so te v jeziku najpogosteje rabljene:

Po.:	St.:	
tud	*tudi*	*3571*
jz	*jaz*	*3460*
sej	*saj*	*3399*
al	*ali*	*3251*
zdej	*zdaj*	*3036*
tko	*tako*	*2820*
tak	*tako*	*2667*
blo	*bilo*	*2263*
sam	*samo*	*1699*
sn	*sem*	*1620*

5. Nekatera težavnejša vprašanja zapisovanja govora

Tukaj ne bomo obravnavali vseh načel zapisovanja govora v korpusu Gos, ampak samo nekatera težavnejša vprašanja. Prvi dve se nanašata na pogovorni zapis, kjer ponekod opazimo nedoslednosti, tretje na pogovorni in standardizirani zapis ter četrto na standardizirani zapis.

5.1. Dvoustnični 'U'

Načelo pogovornega zapisa številka 3 v specifikacijah transkribiranja za korpus Gos pravi: »Dvoustnični v zapisujemo s črko 'v' (*prov, nav, navm, odpravt, davn*...) oz. tudi z 'l', če tako izhaja iz knjižne norme (*kosil* (v pomenu *kosilo*), *mel* (v pomenu *imel*)). Če je u samoglasniški, ga pišemo s črko 'u' (*pršu, vidu*...).«

Zdi se, da tovrstno načelo govorcem slovenščine vseeno ni popolnoma domače, ko morajo zapisovati besede govorjene slovenščine, ki še nimajo ustaljenega »standarda« zapisovanja, in sicer se marsikje namesto predvidenega zapisa z 'v' ali 'l' vrine zapis z 'u' – npr. *laufati*, *šlauf* ali *genau* se v zapisu z 'u' pojavljajo celo v Besedišču in tudi po korpusu Gigafida močno prevladuje različica z 'u', čeprav bi po zgornjem pravilu pisali *lavfati, šlavf, genav*. Podobno so dvojnice lahko pri medmetih, npr. *au* in *av* (po SSKJ).

V zvezi s tem se pojavlja tudi nekaj več nedoslednosti v pogovornem zapisu korpusa Gos, kjer najdemo po večkrat tudi pogovorne zapise tipa *mau* (*malo*), *biu* (*bil*), *šou* (*šel*), *dou* (*dol*), *prou* (*prav*), *dau* (*da bo*), *nou* (*ne bo*) itd., namesto predvidenega zapisa s črko v/l. Kljub temu pa je večinsko zapis z v/l v tovrstnih vlogah prevladujoč in zdi se, da bi bilo spreminjanje načela v zapis z 'u' še bolj problematično: potem bi namreč besede, ki v glasovni podobi sledijo standardu, še vedno pisali z 'v' ali 'l', npr. *imel*, in kontrast z *meu* namesto *mel* bi verjetno vnesel še več zmede in nedoslednosti. Edina sprejemljiva sprememba tega pravila bi zato bila, da se vodi seznam besed ali oblik, za katere lahko po pisnih korpusih sledimo tendenci po pisanju s črko 'u' v teh položajih, ostale pa se še naprej pišejo z 'v' oz. 'l'. Je pa vprašanje, ali ni tako pravilo še bolj problematično s stališča doslednosti zapisovanja kot obstoječe uniformno vodilo.

5.2. Člen 'ta'

Določila, ki bi posebej omenjalo pisanje člena 'ta' v tipu 'ta rdeči' (kjer je 'ta' nenaglašen in izgovorjen skupaj s sledečim pridevnikom), v specifikacijah transkribiranja ni bilo, iz korpusa pa vidimo, da se je sledilo praksi, da se člen piše kot samostojna beseda. Ob tem pa na nivoju pogovornega zapisa (kot posamezne lapsuse pa posledično tudi na ravni standardiziranega zapisa) vseeno občasno zasledimo stični zapis, zelo pogosto za zvezo *ta mali/ta mala,* npr. *tamal, tamav*, *tamalo, tamali, tamalima, tamavga, tamalga*, poleg te pa bolj kot ne posamično še za zveze *taprav/tapravo, tapravga* (*ta pravi*), *tazaden* (*ta zadnji*), *tamladi* (*ta mladi*), *taprv* (*ta prvi*), *tazadno* (*ta zadnjo*) itd.

Medtem ko je na nivoju standardiziranega zapisa res najbolj praktično in smiselno nestično pisanje, zlasti z vidika kasnejšega oblikoslovnega označevanja, izdelave besednih seznamov in iskanja po besedilu, pa bi veljalo še enkrat razmisliti o možnosti stičnega pisanja v pogovornem zapisu. S tem bi namreč omogočili avtomatsko ločevanje med rabami tipa zaimek + pridevnik (*hvala za ta lep mejl*) in rabami tipa člen + pridevnik (*je bil predračun tak da je šu tist talep lijak ven*), ki jih je mogoče zanesljivo ločevati samo ročno in s pomočjo zvočnega posnetka.

5.3. Neverbalni in polverbalni izrazi

O pisanju neverbalnih in polverbalnih izrazov govori določilo pogovornega zapisa številka pet, ki je (opredeljeno vnaprej, pred začetkom transkribiranja) dokaj skopo: "Podaljšane neleksikalne enote pri iskanju formulacije pišemo s tremi črkami, in sicer: *eee, eem, mmm*... oziroma z nizom črk, ki najbolje ustreza dejanski izgovorjavi."

O pretvorbi teh zapisov v standardizirani zapis je v specifikacijah transkribiranja za korpus Gos določilo: "Onomatopeje, medmete, besedne fragmente in druge glasove, za katere v knjižnem jeziku ni standardnega zapisa, pustimo zapisane tako, kot so bili zapisani v prvotni transkripciji," v monografiji (Verdonik, Zwitter Vitez 2011: 67) pa: "Onomatopeje, medmete, besedne fragmente in druge glasove standardiziramo z enotno krovno obliko, kjer je to mogoče: *jooj, ijoj* > *joj*." Sprememba določila je posledica opažanja, da so v pogovornem zapisu nastajale nedoslednosti pri zapisovanju.

Vseeno pa obstoječa rešitev, da so nekatere glasovno različne realizacije neverbalnih ali poverbalnih glasov vodene pod enotnim krovnim zapisom, ni povsem idealna, saj tukaj večinoma ne moremo govoriti o redukcijah ali glasovnih premenah kot pri bolj verbaliziranih enotah. Za primer: pri *ijoj* ne moremo govoriti o glasovni premeni osnove *joj*.

Neverbalni in polverbalni glasovi so, gledano površinskobesedilno, ena najbolj pogostih in tipičnih

značilnosti govorjenega besedila, ob tem pa povzročajo težave tako lematizaciji kot oblikoslovnemu označevanju, učenemu na pisnih besedilih, in so zato pogosto kar sistematično narobe označeni. Gre torej za vprašanje, ki lahko ima na rezultate iskanja po korpusu precejšen vpliv. Problemu smo se zato podrobneje posvetili: pregledali smo vse tovrstne izraze v Gosu in izdelali predlog natančnejših načel njihovega zapisovanja skupaj s seznamom zapisov za te izraze v obstoječem gradivu korpusa Gos. Čeprav za slovenščino pri ZRC SAZU sicer obstaja slovarček medmetov, ki zajema tudi nekatere polverbalne izraze (http://bos.zrc-sazu.si/cgi_new/medmeti/a01.exe?name=medmeti&expression=*), pa je naš seznam prvi, ki temelji na avtentičnem govorjenem gradivu. Seznam je v prilogi 1 tega prispevka in je (med drugim) pomemben predvsem za uspešnejšo lematizacijo in oblikoslovno označevanje govornega gradiva.

Seznam neverbalnih in polverbalnih izrazov v Gosu smo zbrali tako, da smo ročno pregledali seznam standardiziranih zapisov korpusa Gos in iz njega izločili kandidate črkovnih nizov za tovrstne izraze, nato pa jih preverjali prek Gosovega konkordančnika (www.korpus-gos.net). Po pregledu smo za veliko izrazov predlagali nov, popravljen zapis, ki sledi načelom zapisovanja, kot jih povzemamo spodaj.

Načela zapisovanja izhajajo iz dveh stališč: način zapisa naj bi bil govorcem slovenščine čim bližji, hkrati pa naj bi omogočal največjo možno mero avtomatskega procesiranja teh izrazov v govorjenem besedilu. Načela so:

1. izraze zapišemo raje z eno besedo kot več besedami (npr. *ojoj* namesto *o joj*),
2. kjer ni bistvene razlike v zvočni podobi in funkciji/pomenu, ohranimo enoten zapis za različne rabe (npr. *mhm* bi posamično morda zapisali tudi kot *ehm*, vendar je razmejitev težko objektivno določiti, zato raje ohranjamo vedno *mhm*),
3. izraze zapisujemo prednostno s tremi črkami, tako da se razlikujejo od drugih besed (npr. raje *vaa* kot *va*), razen kjer ni nevarnosti, da bi bil zapis identičen zapisu kakih drugih besed, ali če je drugačen zapis že močno uveljavljen (npr. *eh*),
4. dvoustnični U prednostno pišemo z 'v' (*av, vav*),
5. podaljševanje glasov se ne označuje z več črkami, ampak se ohranja enoten zapis (npr. vedno *jee*, ne *jeee* ali podobno),
6. prednost ima poslovenjen zapis (npr. *jes*, ne *yes*, *okej*, ne *ok* ali *okay*).

Kot izstopajoč neenotni in avtomatsko težko sledljiv zapis v obstoječem gradivu izpostavimo neverbalno glasovno zanikanje. Zasledili smo naslednje različice zapisovanja tega pojava: *n n, m m, a a, e e, nn, aa, mm*. Glasovno le-to dejansko niha od bolj vokalnega, a-jevskega prek polglasniškega do zvočniškega m ali n. Da bi bilo neverbalno glasovno zanikanje avtomatsko sledljivo, bi bil potreben bolj enoten in unikaten zapis. Predlagamo dve različici zapisa: *nn* in *aa*.

Nasproten, sicer redkejši primer je neverbalno glasovno pritrjevanje, za katerega je bil realiziran zapis *mm* – ta je primeren, bi pa bilo dobrodošlo, da z njim niso zapisane še kake druge realizacije, npr. zanikanje (kjer predlagamo *nn*) ali oporni signal (*mmm* oz. *eee*).

Iz zgornjih primerov vidimo, da je lahko transkripcija govora na določenih točkah že močno v vlogi interpretacije funkcij/pomenov izrazov. To se zdi dopustno le izjemoma, ko sicer zelo težko enoumno določimo zapis.

Naš predlog je, da se načelom zapisa, kot so predstavljena zgoraj, sledi že pri pogovornem zapisu. Standardizirani zapis se potem za neverbalne in polverbalne glasove ne bi spreminjal, identičen zapis pa bi dobila tudi lema teh izrazov.

5.4. Nestandardni polnopomenski izrazi

Zadnje vprašanje, ki ga bomo obravnavali, se odpira pri standardizaciji zapisa za nestandardne polnopomenske izraze (s tem mislimo take, ki niso sprejeti v standardni jezik), od katerih imajo mnogi v različnih regijah nekoliko različno glasovno realizacijo. Teh ni toliko, kot bi morda pričakovali, vseeno pa dovolj, da je treba njihov krovni standardizirani zapis bolj natančno določiti. V obstoječih specifikacijah Gosa piše: »Pogovorne besede, ki bi jim težko določili povsem ustrezno knjižno različico, ohranjamo. Pri odločitvah glede zapisa se opiramo na pisne korpuse in druge vire.« (http://www.korpus-gos.net/Content/Static/Navodila_za_standardizacijo_zapisa_govora.pdf) Sledijo primeri, ki dodatno ilustrirajo različne rabe, vendar večinoma razne funkcijske besede, polnopomenske pa le na kratko, in sicer pretežno v naslednjem odstavku: »Ohranimo: a) izposojenke *bek, čuješ, fak, fajrala, ferker, ful, gruntali, hambrt, kafič, kao, kuhla, može, ni mus, ornk, pašeš, plata, pošlihtaš, rajsar, ratati, singl, spedenan, štima, šparati, valjda, ziher, žijaš...*«

Nedoslednosti v zapisu v zvezi s tem problemom smo zasledili bolj kot ne naključno, ob uporabi korpusa in pregledovanju njegovih besednih seznamov. Tako se je na primer v standardiziranem zapisu pojavljalo *fertig* in *fertik*, *frej* in *fraj*, *kafe* in *kofe* ...

Poskus sistematičnega sledenja tem pojavom smo naredili tako, da smo najprej izdelali besedni seznam vseh pojavitev v pogovornem zapisu korpusa Gos, nato pa besedni seznam oblik v korpusu Kres, ki se pojavijo vsaj desetkrat. Nato smo besedni seznam korpusa Gos filtrirali s pomočjo besednega seznama korpusa Kres in ročno pregledali samo tiste pojavitve, ki jih ni bilo na besednem seznamu korpusa Kres. Izločili smo kandidate za podrobnejši pregled ter si zanje izpisali konkordance. Te smo ponovno ročno pregledali. Po pregledu seznamov na črki a in b smo na tak način našli štiri dodatne primere, kar potrjuje, da ne gre za zelo obsežen problem, vseeno pa se ne sme ignorirati. Zadeva namreč leksiko, ki je v pisnih korpusih redko prisotna in lahko na primer pri izdelavi geslovnikov ali analizah ključnih besed izpade iz rezultatov ali je v rezultatih neustrezno rangirana, če ni v različnih realizacijah standardizirana vedno na enoten način. Zato se zdi potrebno, da se pri nadgradnji korpusa Gos na tovrstne primere bolj podrobno opozarja in se jih sistematično vodi.

V glavnem naključno zbrani primeri, ki smo jih sami pregledali, kažejo sledeče:

1. različic ne zasledimo: *kao*, *talam/tala* (za lemo *talati*) ...,
2. različne pogovorne realizacije nestandardnih polnopomenskih izrazov so nedvoumno posledica znanih regionalnih in narečnih glasovnih premen, npr. *šihtu* vs. *šejhti* (za lemo

šiht), *cajt* vs. *cet* (za lemo *cajt*), *pasalo* vs. *pasal* (za lemo *pasati*) ...

Medtem ko v zgornjih primerih neenotnega standardiziranega zapisa ne zasledimo, pa ga v naslednjih primerih:

1. ob redukciji določenega glasu: *luškan/lušno* vs. *luštkan/luštno*, *magari* vs. *magar, glej* vs. *lej* ...; ob tem odločitve niso nujno enostavne, zlasti ko je reducirana oblika zelo pogosta ali dobi nove pomene/funkcije; tako na primer tudi SSKJ obliki *lej* (od *glej*) pripiše posebno geslo, v Gosu pa so v zvezi s tem zanimivi še primeri *kurc* vs. *kurac* ali *dedec* vs. *dec* vs. *ded*; težavnost obravnavanja redukcij se kaže tudi na primeru *čmo* vs. *hočemo* (glej specifikacije standardiziranega zapisa, www.korpus-gos.net) in predstavlja pravi jezikoslovni izziv pri nekaterih funkcijskih besedah, npr. *te* vs. *potem*, *k* v vlogi *ker, ko, ki, kot, kjer, kar, kaj* ...;
2. zaradi premen po zvenečnosti, npr. *fertig* vs. *fertik*, *oreng* vs. *orenk* ...;
3. zaradi premen vokalov, npr. *fraj* vs. *frej*, *kafe* vs. *kofe* ...

Medtem ko pri zgornjih primerih prepoznavamo pomanjkanje enotne standardizirane oblike, pa je treba opozoriti na izredno previdnost, da zapis ne zaide v nasprotno smer, to je v pretirano iskanje skupne standardizirane oblike, ko to ne bi bilo upravičeno, na primer v Gosu *žiher* (v pomenu *lahko*) ni enako kot *ziher* (v pomenu *varno; zagotovo*) ...

Kot ugotavljamo že pri zapisovanju polverbalnih in neverbalnih glasov, transkribiranje izjemoma hote ali nehote zaide na spolzek teren interpretacije funkcij/pomena posameznih izrazov. Tako se v Gosu v standardiziranem zapisu pojavlja trojček izrazov *not* vs. *noter* vs. *notri*, ki pa se v govoru ne rabijo enako kot predvideva standardni jezik, tj. *notri* je lahko v vlogi standardnega *noter*, npr. *morš notri padniti*, in obratno, *noter* je v vlogi *notri,* npr. *pomijejo pa vržejo tam notri*, *not* pa lahko gledamo kot reducirano obliko enega ali drugega ali kot samostojen leksem. V obstoječem gradivu korpusa Gos se pri teh izrazih sledi interpretaciji funkcij/pomenov teh izrazov. Enako kot pri neverbalnih in polverbalnih glasovih tudi tukaj menimo, da naj ostane taka praksa čim bolj izjemna in jo je smiselno tudi za nazaj kritično pretresti od primera do primera.

6. Zaključek

V prispevku smo predpostavili, da bo morebitna nadgradnja korpusa Gos zelo verjetno potekala v obliki enotne platforme in (delno) skupnega vira z akustično bazo za razpoznavanje tekočega govora. V nadaljevanju smo se osredotočili na vprašanja zapisovanja govora, ki bi v tej skupni platformi po našem mnenju potekala na podlagi vzpostavljenega dvotirnega sistema zapisovanja (pogovorni in standardizirani zapis) v korpusu Gos. Opozorili smo na kompleksnost problema zapisovanja, ki je do neke mere vedno tudi interpretacija, ter se nato podrobneje posvetili štirim vprašanjem, ki so se nam odprla skozi uporabo in analize obstoječega Gosovega gradiva.

Že sproti smo opozorili, da je odprtih vprašanj lahko še več. Eno obsežnejših je povezano s funkcijskimi besedami in je prezapleteno, da bi ga lahko obravnavali kot del tega prispevka. Problem lahko ilustriramo s primeroma *k* in *ka*: pogovornemu *k* so v Gosu pripisane standardizirane različice *ker, ko, ki, k, kot, kjer, kar, kaj* ..., pogovornemu *ka* pa *kaj, ka, ker, da, ki, ko, kar* ... Osrednje vprašanje je, kdaj je neki reducirani obliki smiselno iskati interpretacijo v obstoječem (pisnem) standardu in kdaj jo obravnavati kot novo obliko/funkcijsko besedo. V zvezi s tem se kaže potreba po poglobljeni celostni jezikoslovni oz. jezikoslovno-diskurzni analizi rabe funkcijskih besed v govorjenem jeziku, šele potem lahko razmišljamo o prenovljenih ali dopolnjenih navodilih za standardizirani zapis te skupine besed.

V začetku prispevka smo opozorili na majhno zanimanje raziskovalcev, zlasti jezikoslovcev, za vprašanja govorjenega jezika in prepuščenost njegovega opisovanja posameznim iniciativam zunaj stroke. Kot da je to pojav, ki nam je preblizu, da bi se nam zdel neznan in zato zanimiv ter potreben analize in opisa. Toda ravno zato, ker nam je tako blizu, lahko pove veliko o človeku, več, kot se zavedamo ... če se le dovolj poglobimo vanj; tudi (ali pa celo predvsem) s pomočjo transkripcij in posnetkov v obliki govornega korpusa in baze.

7. Literatura

Dobrovoljc, H., Erjavec, T., Krek, S., Snoj, M., Verdonik, D., Vintar, Š., 2014. AKcijski načrt za jezikovno opremljenost. Dostopno na: http://www.mk.gov.si/fileadmin/mk.gov.si/pageuploads/Ministrstvo/slovenski_jezik/Akcijska_nacrta/Akcijski_nacrt_za_jezikovno_opremljenost_javna_razprava.pdf. 1. julij 2014.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., Stabej, M., 2013. Compilation, transcription and usage of a reference speech corpus: The case od the Slovene corpus GOS. *Language Resources and Evaluation*, 47/4, 1031-1048.

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

Žgank, A., Rotovnik, T., Verdonik, D., Kačič, Z., 2004. Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. Zbornik konference IS'04 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/isjt04/zbornik/. 1. julij 2014.

Žgank, A., Rotovnik, T., Grašič, M., Vlaj, D., Kačič, Z., 2006. Slovenska govorna in tekstovna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. Zbornik konference IS'06 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/is-ltc06/proc/22_Zgank_2of2.pdf.

Žgank, A., Zwitter Vitez, A., Verdonik, D., 2014. The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniforme guidelines for future work. Zbornik konference LREC 2014. Dostopno na: http://www.lrec-conf.org/proceedings/lrec2014/index.html. 1. julij 2014.

Žibert, J., Mihelič, F., 2004. Development, evaluation and automatic segmentation of Slovenian Broadcast News Spech Database. Zbornik konference IS'04 – Jezikovne tehnologije. Dostopno na: http://nl.ijs.si/isjt04/zbornik/. 1. julij 2014.

Priloga: Predlog zapisovanja najpogostejših neverbalnih in polverbalnih glasov

Seznam je narejen na podlagi korpusa Gos v obsegu 1 mio. besed, dostopnega na www.korpus-gos.net, marca 2014. Znak # pred zapisom pomeni, da zapis ni enoznačen in je lahko identičen zapisu kake druge besede, npr. veznika, člena ipd. Če pred zapisom ni znaka #, pomeni, da se mu lahko avtomatsko pripiše enaka lema, kot je obstoječi zapis, in oblikoskladenjska oznaka za medmet.

#a
aa (zanikanje)
aaa
aam
aan
ah
aha
ahah
ahaha
ahahaha
ahja
ahjoj
ahm
ahoj
#aj
#aja
ajah
ajaj
aje
ajej
ajo
ajoj
alo
ao
aua
auva
av
#ba
bljeh
brum
bu
bvak
buf
bum
bumč
bvum
bzz
bž
ck
dammm
dh
dum
#e
eee
eem
een
eev
eh
ehe
eheh
ehehe
ej
eje
ejo
ejoj
fuf
fuj
fuu
grr
ha
haha
hahaha
hahahaha
hajaj
#he
heh
hehe
hehehe
hej
hhh
#hi
hihi
hijaj
hijo
hjoj
hjujujujuju
hm
#ho
hoho
hohoho
hohop
hojoj
hopa
hopla
hopsasa
hov
hu
huh
huhu
#i
iii
ija
ijo
ijoj
jah
jaj
jao
jea
jee
#jej
jes
johoho
johoj
joj
jojojojojojo
joo
jov
joz
juhej
juhu
juhuhu
jupi
juu
klink
maa
mahh
mee
mh
mhm
miu
mjav
mm
(pritrjevanje)
mmm
nanananananana
nee
nhn
njam
njm
nn (zanikanje)
nnn
#o
oa
oh
ohja
ohjej
oho
ohoho
ohohoho
#oj
oja
ojej
ojla
ojoj
ojojej
ojojo
ojojoj
ojojojo
ojojojoj
ojojojojoj
ola
ooa
ooo
op
opa
opala
ops
ov
ovh
paf
pavf
pff
pha
plop
pom
puf
ratatatatata
rc
rrr
ssk
sss
tada
tadadada
tadam
tarararata
taratatam
tarararan
tarararararararar
arara
taratataratat
tk
totrolodontodo
tp
tralala
tumbapa
tup
#u
ua
uf
uh
#uhu
uhuhu
ujej
umbapa
#uo
ups
upsala
vaa
vav
vov
zk
šink
šk
ššš
čk
čuf
čuči