PREDNOSTI IN SLABOSTI DVOTIRNEGA ZAPISOVANJA GOVORA V SLOVENSKIH GOVORNIH VIRIH

DARINKA VERDONIK,[1] MITJA TROJAR,[2] ANDREJA BIZJAK[1]
[1] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor, Slovenija
darinka.verdonik@um.si, andreja.bizjak1@um.si
[2] ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana, Slovenija
mitja.trojar@zrc-sazu.si

Zapisovanje govora v govornih korpusih je nedvomno časovno največji vložek v procesu izdelave govornega korpusa in pomemben razlog, da so govorni korpusi neprimerno manjši od pisnih. Zapis govora je prevod iz izvorno večmodalnega kanala komunikacije, v katerem verbalno izraženi pomen sooblikujejo glas in način govora, govorica telesa in situacija, v kateri poteka komunikacija, v eno, pisno modalnost. Zaradi variabilnosti govora na vseh jezikovnih ravneh se zapisovalec pri tem nenehno sooča z vprašanjem, kako naj to, kar sliši, zapiše. Da bi bil zapis čim bolj verodostojen, hkrati pa časovno vseeno izvedljiv za velik obseg gradiv, se je tako poleg standardiziranega zapisa vzpostavil tudi tako imenovani pogovorni zapis, ki sledi načelu zapiši, kakor je izgovorjeno. Toda dvojni zapis zahteva dodaten trud, zato v tem prispevku kritično prevpraševujemo njegovo utemeljenost glede na prakse drugod, zahtevan dodaten trud in njegove prednosti ter kritično analiziramo še druga težavnejša vprašanja zapisovanja.

DOI
https://doi.org/10.18690/um.ff.4.2024.4

ISBN
978-961-286-882-6

Ključne besede:
transkribiranje, standardizirani zapis, ortografska transkripcija, pogovorni zapis, fonetična transkripcija

Univerzitetna založba Univerze v Mariboru

DOI
https://doi.org/
10.18690/um.ff.4.2024.4

ISBN
978-961-286-882-6

Advantages and Disadvantages of Two-Tier Speech Transcription in Slovenian Speech Resources

Darinka Verdonik,[1] Mitja Trojar,[2] Andreja Bizjak[1]
[1] University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
darinka.verdonik@um.si, andreja.bizjak1@um.si
[2] ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Ljubljana, Slovenia
mitja.trojar@zrc-sazu.si

Keywords:
transcribing, standardized transcription, orthographic transcription, literal transcription, phonetic transcription

Transcribing speech in speech corpora is undoubtedly the largest time investment in the process of creating a speech corpus and an important reason that speech corpora are considerably smaller than written ones. Speech transcription is a translation from an originally multimodal channel of communication, in which verbally expressed meaning is shaped by the voice and manner of speaking, body language, etc., and converted into a single, written modality. Due to the variability of speech at all linguistic levels, the transcriber constantly faces the question of how to transcribe what s/he hears. In order to make the transcription as exact as possible, but at the same time feasible when working with large amounts of data, a pronunciation-based transcription was introduced in Slovenian speech corpora along with the standardized transcription. However, two-tier transcription requires additional effort. For this reason, this paper critically assesses its rationale, comparing practices used elsewhere, estimates of the additional effort and its advantages. Additionally, we assess other challenging aspects of speech transcription.

University of Maribor Press

1 Uvod[1]

Časovno in finančno najzahtevnejši korak izdelave govornih korpusov je t. i. transkribiranje posnetkov oziroma natančneje zapisovanje in označevanje posnetkov. Ne gre namreč samo za zapis govora, ampak je treba popisati tudi podatke o govorcih in posnetih govornih dogodkih, segmentirati govor na osnovne enote – segmente, označiti menjavanje govorcev in kdaj kdo govori, označiti akustično ozadje (npr. prisotnost šuma ali glasbe) in akustične dogodke (nenadni zvoki od zunaj ali nastali z govorili, kot so kašljanje, glasni vdihi ipd.) ter osnovne prozodične značilnosti (smeh, premori ipd.). Zaradi časovne in finančne zahtevnosti zapisovanja govora in označevanja posnetkov se ob izdelavi govornih korpusov vedno iščejo načini, kako izvedbo čim bolj ekonomizirati. Predvsem za javno govorjeno rabo, na primer v medijih ali parlamentu, je tako mogoče dobiti arhive posnetkov in včasih tudi zapisov, pri čemer pa se pogosto soočamo s pravnimi in drugimi omejitvami ter posledično njihovo nedostopnostjo (Verdonik 2023: 32). Vse bolj se v proces uvaja uporaba avtomatskega razpoznavanja govora za pripravo zapisa. Toda za posnetke nejavne, zasebne rabe govora lahko še naprej pričakujemo, da bo zaradi zahtevnih akustičnih pogojev in variabilnosti govora na vseh jezikovnih ravneh še dolgo obstajala potreba po ročnem zapisovanju in označevanju.

Na podlagi priporočil EAGLES (Gibbon et al. 1997) in praks v govornih korpusih ločujemo več ravni zapisa govora: bodisi gre za ortografski zapis bodisi za avtomatsko pripravljen fonetični zapis z algoritmi grafemsko-fonemske pretvorbe bodisi za natančen fonemski zapis v fonetični abecedi. Pri ortografskih zapisih lahko nadalje ločujemo:

1. standardni ortografski zapis, to je povsem standardiziran zapis, v katerem se izgovorjene besede in besedne oblike zapiše z ustreznimi standardnimi besedami in besednimi oblikami,
2. razširjeni ortografski zapis (angl. *expanded orthographic transcription*) z navodili, pravili in/ali seznami za dodatno zapisovanje posebnosti izgovorjenih besed in besednih oblik, ki so drugačne od standardnih.

[1] Prispevek je nastal v okviru temeljnega raziskovalnega projekta Temeljne raziskave za razvoj govornih virov in tehnologij za slovenščino (J7-4642) in raziskovalnega programa Slovenski jezik v sinhronem in diahronem razvoju (P6-0038), ki ju financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije.

V slovenskih govornih korpusih je uveljavljena praksa zapisovanja tako v standardnem ortografskem kot v razširjenem ortografskem zapisu. Prvi način je poimenovan standardizirani, drugi pogovorni zapis. To pomeni, da je govor zapisan dvakrat. Ta praksa je bila vzpostavljena s korpusom Gos (Verdonik et al. 2013) ter delno prilagojena in posodobljena ob izdelavi govorne baze in korpusa Artur[2] (Verdonik et al. 2023b). Kot je razvidno iz poglavja 2, najdemo podobno prakso tudi v drugih jezikih. Potem ko je bilo na ta način v korpusu Artur zapisanih več kot 300 ur javnega, nejavnega in parlamentarnega govora, je smiselno vprašanje, ali je takšen dvotirni sistem zapisovanja govora potreben tudi vnaprej: kaj so njegove prednosti in kaj slabosti glede na dodaten zahtevani trud ob že tako ali tako obsežnem delu zapisovanja govora in označevanja posnetkov?

2 Zapisovanje govora v govornih korpusih

Temeljni problem zapisovanja govora za namen izdelave govornega korpusa je izredna variabilnost in inovativnost oblik in besed v govorjeni rabi, ki bistveno presega standardni slovarski nabor besed in njihovih oblik, zato zanje ne obstaja noben uveljavljen način zapisovanja in se je o njem treba šele dogovoriti. V pregled praks smo zajeli nekaj najbolj vplivnih in z vidika zapisovanja govora primerljivih korpusov za različne evropske jezike s posebnim poudarkom na slovanskih jezikih.

2.1 Korpusa Spoken BNC2014 in FOLK

V govorni komponenti enega od bolj vplivnih govornih korpusov, British National Corpus, oz. v njenem najnovejšem dodatku Spoken BNC2014 (Love et al. 2017), avtorji navajajo izbiro med ortografskim in dodatnim fonetičnim zapisom, vendar slednjega zaradi zahtevnosti v Spoken BNC2014 ne vključijo. Toda njihov ortografski zapis vsebuje nabor sprejemljivih oblik za zapisovanje dialektalnih in nestandardnih besed, torej je standardni ortografski zapis razširjen z dodatnimi oblikami. Fonetične zapise običajno najdemo v dialektalnih korpusih, v splošnih govornih korpusih pa redko, in če že, v zelo omejenem obsegu. Pogosto pa se išče kompromis, kako ob čim manjšem vložku vseeno zadovoljivo opisati posebnosti izgovorjave, kot bomo videli v nadaljevanju tega poglavja.

[2] Za razlikovanje med govorno bazo in korpusom gl. Campbell 2005: 114–115.

Najbolj neposredno primerljiv način dvotirnega zapisovanja govora, kot se je uveljavil v slovenskih govornih korpusih, najdemo v nemškem prostoru, med drugim v govornem korpusu FOLK (Schmidt 2016). Ta ima začetke v letu 2009 in je zastavljen kot dolgoročni projekt sistematičnega zbiranja raznolikih govornih interakcij med govorci v Nemčiji. Po številu uporabnikov je eden najbolj uporabljanih govornih korpusov nemščine. Do leta 2019 je obsegal 1,6 mio. pojavnic, letno pa se poveča za okrog 300.000 pojavnic. Sistem zapisovanja in označevanja v korpusu FOLK temelji na smernicah GAT (Selting et al. 2009), ki veljajo po navedbah Schmidta (2016) za enega najbolj uveljavljenih sistemov zapisovanja govora v konverzacijski analizi na Nemškem. Skladno s temi smernicami se uporablja modificiran zapis, poimenovan kot »literarni zapis« (angl. *literary transcription*) ali »očesni dialekt« (angl. *eye dialect*). Besede, ki v izreki odstopajo od standardne, imajo ustrezno prilagojen zapis, npr. *zwo* kot pogovorna različica številnika *zwei*. Tak sistem zapisovanja je tako rekoč identičen kot pogovorni zapis v slovenskih govornih korpusih. Z namenom optimiziranja korpusa za korpusno jezikoslovje in računalniške jezikoslovne metode je v korpusu FOLK nato dodan še zapis v standardni ortografiji, torej podobno kot standardizirani zapis v slovenskih govornih korpusih.

2.2 Korpusi s-hovor-6.0, ORTOFON in HrAL

Primerljivi načini zapisovanja so prisotni tudi v slovanskem prostoru, kjer najdemo večje primerljive govorne korpuse za češki in slovaški jezik. Slovaški govorni korpus v različici s-hovor-6.0 obsega 6,6 mio. pojavnic in vključuje vsakdanje pogovore v najrazličnejših govornih situacijah. Zapis posnetkov je narejen v ortografskem zapisu, ki sledi pravilom standardne slovaške ortografije, s tem da v nekaterih primerih sledijo standardni slovaški izgovorjavi v nasprotju s predpisano uradno – gre za izgovorjavo posameznih glasov, npr. palataliziranega l (Garabík, Rusko 2007: 233). Poleg standardnega zapisa je dodan še delno fonemski zapis oz. zapis izgovorjave, ki pa je narejen s slovaško ortografsko abecedo, ne s fonemsko abecedo – enako kot pogovorni zapis v slovenskih govornih korpusih. Kot navaja Garabík (2023), tak način fonemskega zapisa bistveno pohitri in olajša zapisovanje govora.

Češki korpus ORTOFON (Komrsková et al. 2017) vsebuje spontane pogovore v vsakdanjih situacijah med ljudmi, ki se med seboj poznajo. Vključuje gradivo iz obdobja 2012 do 2017 in obsega 1,2 mio. pojavnic. Tudi v tem korpusu je zapis

govora dvotirni. Ortografski zapis se kljub temu, da je ortografski, v nekaterih elementih razlikuje od zapisa v standardnem pisnem jeziku. Tako na primer vključuje dialektalne značilnosti, kot so različice končnic za vse vrste sklanjatev in spregatev, regionalne različice vokalnih sprememb ali sklanjatev ipd. Seznam vseh izjem skrbno beležijo. Ortografski zapis pa ne označuje na primer različnih dolžin glasov, reduciranih oblik ali soglasniških premen. Po ortografskem zapisu je dodan še poenostavljen fonetični zapis, ki je narejen z ortografsko abecedo, razširjeno z manjšim naborom posebnih simbolov, in ne s fonetično abecedo. Fonetični zapis med drugim beleži različne asimilacijske procese, izpuste v izgovorjavi ipd., nima pa na primer označenega naglašenega zloga.

Neke vrste dvotirni način zapisovanja oz. razširjeni ortografski zapis najdemo tudi v hrvaškem korpusu govorjenega jezika odraslih HrAL (Kuvač Kraljević, Hržica 2016). Ta korpus vsebuje spontane vsakdanje pogovore v obsegu 250.000 pojavnic in je bil posnet v letih od 2012 do 2016. Zapisan in označen je skladno z načeli zapisovanja in označevanja govora v konverzacijski analizi ter je izdan v seriji govornih virov TalkBank,[3] ki sledijo navodilom CHAT (MacWhinney 2000). Ta glede zapisovanja dialektalnih različic v izgovorjavi omogočajo, da se zapiše dialektalni izgovor, v oglatih oklepajih pa se doda standardni zapis. Primer 1 prikazuje vzorec iz korpusa HrAL.

Primer 1: Zapis govora v hrvaškem korpusu HrAL

pa ne zato što se ja njoj nisan [: nisam] niti upucavo [: upucavao] nego san ja nju namješto [: namještao]

Navodila CHAT sicer poleg načina zapisovanja, kot je bil izbran v HrAL, omogočajo še, da izberemo ali fonemski zapis ali pa dialektalne različice ignoriramo, pri čemer je treba to informacijo vključiti v korpus.

2.3 Korpus C-ORAL-ROM

V romanskih jezikih je znan korpus C-ORAL-ROM. Nastajal je od 1999 naprej (Cresti, Moneglia 2005) in vsebuje govorjene vsakdanje pogovore v štirih romanskih jezikih, francoskem, italijanskem, španskem in portugalskem, v skupnem obsegu 121 ur oz. skupno 300.000 besed po jeziku. C-ORAL-ROM je transkribiran podobno

[3] https://talkbank.org

kot hrvaški korpus HrAL, skladno s formatom CHAT. Francoski del korpusa C-ORAL-ROM poleg ortografskega zapisa vključuje tudi fonetični zapis v abecedi SAMPA v primerih, ko izgovorjave odstopajo od predvidene oz. so kakorkoli posebne (Cresti, Moneglia 2005: 114). Tudi španski del korpusa C-ORAL-ROM pri nestandardnih besedah dodatno popisuje izgovorjavo, vendar z ortografsko abecedo (Cresti, Moneglia 2005: 142).

3 Zapisovanje govora v korpusu Artur

V korpusnem delu govorne baze Artur je govor zapisan dvotirno, s pogovornim in standardiziranim zapisom hkrati, skladno z načinom, vzpostavljenim s prvo izdajo govornega korpusa Gos (Verdonik, Zwitter Vitez 2020). Ker pa je korpus Artur prvotno namenjen razvoju avtomatskega razpoznavanja govora za slovenski jezik, so bile vključene nekatere zaželene prilagoditve, nekaj sprememb zlasti v standardiziranem zapisu pa je izhajalo iz izkušenj in analiz zapisov v prvi različici korpusa Gos.

3.1 Priprava pogovornega zapisa

Pogovorni zapis je prva raven zapisovanja govora v slovenskih govornih korpusih. Primer 2 prikazuje izjavo iz korpusa Artur, zapisano v pogovornem in v standardiziranem zapisu.

Primer 2: Pogovorni in standardizirani zapis v korpusu Artur

Pogovorni zapis	*Ja no, s temu mojmu p@rjatlom Jušom midva tud d@rgač velik športava, tud tenis igrava.*
Standardni zapis	*Ja no, s tem mojim prijateljem Jušem midva tudi drugače veliko športava, tudi tenis igrava.*

Cilj pogovornega zapisa je, da »čim bolj olajša avtomatsko fonemsko-grafemsko pretvorbo in silabizacijo. V kombinaciji s standardiziranim zapisom je zasnovan tako, da omogoča čim boljšo ekstrakcijo novih kandidatov za oblikoslovno-fonetični leksikon, ki tako ali drugače odstopajo od normirane rabe.« (Verdonik, Bizjak 2023: 28) Osrednje načelo je, da »/g/ovor zapisujemo v veljavnem slovenskem črkopisu, z dodatnim posebnim znakom za polglasnik (@). Upoštevamo veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pri tem kolikor mogoče natančno

predstavimo glasovno podobo govora.« (Verdonik, Bizjak 2023: 28). Iz pogovornega zapisa so tako še naprej vidne redukcije glasov.

Spremembe v pogovornem zapisu v primerjavi s prvo različico korpusa Gos so predvsem tri. Prvič, uvedena je uporaba ločil in velikih začetnic skladno s pravopisno normo. Drugič, delno je spremenjen način segmentiranja govora, ki zahteva, »da segmenti niso predolge enote in da je tam, kjer naredimo mejo segmenta, dovolj premora v govoru, da lahko določimo mejo segmenta, ne da odrežemo del predhodnega ali del naslednjega fonema. Glavni vodili za meje med segmenti sta zato: (a) kratek premor v govoru in (b) dolžina segmenta, ki naj ne bo predolga, tj. več kot okoli 10 sekund,« (Verdonik, Bizjak, 2023: 9) zaradi česar segmenti ne ustrezajo vedno pojmu izjave. Tretjič, uveden je manjši nabor dodatnih znakov za foneme, po pogostosti izstopa znak @ za polglasnik, pojavljata pa se še $g za zveneči h in $r za mehkonebni r (Trojar, Bizjak 2023: 45).

Premen po zvenečnosti (v nasprotju na primer s češkim sistemom) ne zapisujemo: razlog je, da to zahteva visoko stopnjo pozornosti in se v zapisu, ki ga sicer pripravljajo najeti zunanji izvajalci, pojavlja veliko nedoslednosti. V parlamentarnem delu korpusa Artur, za katerega je bil sistem zapisovanja govora vzpostavljen ločeno že pred uvedbo skupnih smernic za korpus Artur, so premene po zvenečnosti ostale v zapisu, vendar se je potrdilo, da je pri tem doslednost slaba. Po drugi strani so premene po zvenečnosti za slovenščino zelo predvidljive in jih je mogoče precej natančno določiti po pravilih. Podobno velja za zapisovanje dvoustničnega *v*, kjer je navodilo skupnih smernic za korpus Artur podobno, kot je veljalo v prvi različici korpusa Gos:

> *Zvočnik dvoustnični v (ni nosilec zloga) zapisujemo s črko 'v', če se pojavi v besednih oblikah, ki niso knjižne (prov, nav, navm, odpravt, davn, gledavc, pov@n ...). Posebej smo pozorni na primere: lavfati, šlavf, genav, mav (malo), šov (šel), dov (dol), prov (prav), dav (da bo), nov (ne bo), tudi medmet av. Če dvoustnični v nastopa v besedni obliki, ki je knjižna in tudi izgovorjena skladno s standardom, ohranimo knjižni zapis (bil, gledal, siv). Če je glas u samoglasniški, tj. je nosilec zloga, ga pišemo s črko 'u' (pršu, vidu, u tem delu...). Tudi predlog v, izgovorjen kot samoglasniški u, pišemo kot u.* (Verdonik, Bizjak 2023: 32)

Tudi tukaj se je predhodno v parlamentarnem delu korpusa Artur vzpostavila praksa, da so dvoustnični *v* zapisovali kot 'u', pri čemer ga zapisovalci zlasti v zbornih pozicijah niso vedno prepoznali in dosledno zapisali.

Seznam neverbalnih in polverbalnih izrazov (npr. *eee, hm, uh, šš*) se je skozi korpus Artur dopolnjeval in je od prve različice korpusa Gos na podlagi analiz prvotnih zapisov dobil tudi vzpostavljen sistem načel, ki med drugim določajo, da izraze raje zapisujemo z eno kot z več besedami, da ne uporabljamo odvečnih različic za zelo podoben izraz, da jih prednostno zapisujemo s tremi črkami in tako dosežemo razlikovalni zapis ipd. (Verdonik, Bizjak 2023).

Pogovorni zapis v korpusu Artur so izvajali najeti zunanji izvajalci in študenti, nato je sledil pregled naključnih segmentov s strani koordinatorja zapisovanja in označevanja govora. Ob tem smo beležili, kje v pogovornem zapisu se pojavlja največ napak. Pogosto je bila prisotna neustrezna segmentacija govora na osnovne enote, zlasti v primeru zelo strnjenega govora brez premorov. Zapisovalci so občasno pozabljali ustrezno označevati dolge premore ali zvočna ozadja in zvočne dogodke. Pri hkratnem govoru se je dogajalo, da je bilo napačno označeno menjavanje govorcev. Zelo nezaželena napaka je bila izpust besed, ki so bile izgovorjene, zapisane pa ne, kar je izredno zahtevno odkriti. Občasno so imeli zapisovalci težavo razumeti, kaj je bilo izgovorjeno, zlasti če posnetek oz. govor ni bil povsem razločen. Nedoslednosti so se dogajale pri zapisovanju polglasnika, dvoustničnega *v*, neverbalnih in polverbalnih izrazov ter tujih besed, kjer so zapisovalci uporabljali črki *q* in *y*, ki nista del nabora znakov za pogovorni zapis. Precej korekcij je bilo potrebnih tudi pri ločilih in velikih začetnicah, kar je bilo nazadnje izvedeno s prenosom ločil in velikih začetnic iz standardiziranega v pogovorni zapis, da smo dosegli usklajenost med obema zapisoma.

3.2 Priprava standardiziranega zapisa

Standardizirani zapis kot druga raven zapisa govora služi predvsem uspešnejšemu avtomatskemu označevanju besedil ter podpori pri slovarski in slovnični analizi govorjene rabe jezika, tako da razkriva značilno govorjeno besedje in slovnične vzorce. Kot je navedeno v dokumentaciji korpusa Artur, pa so nekatera določila dodana tudi z namenom, »da (1) se podpira avtomatizirana pretvorba pogovornega v standardizirani zapis, (2) omogoča enoznačno ujemanje pojavnic v pogovornem in standardiziranem zapisu, (3) preprečujejo težave s kodiranjem, (4) zagotavlja anonimizacija podatkov o govorcih« (Verdonik et. al. 2023a: 6).

V standardiziranem zapisu so besede, pri katerih ni bistvene razlike v primerjavi s predvideno pravorečno izreko, zapisane standardno. Posebnega znaka za polglasnik ali druge glasove tukaj ni več. Tudi ko so v govoru prisotne bodisi glasovne premene, bodisi površnost, nedoslednost, motnje v govoru ali posebnosti izgovorjave, bodisi lapsusi, so besede zapisane standardno. Izziv pa so oblikoslovne, skladenjske in besedne značilnosti pogovornih in narečnih zvrsti, za katere je v korpusu Artur veljalo navodilo:

> *Kadar prepoznamo oblikoslovne (npr. skladenjski vzorci, ne/določna oblika, pregibanje ipd., npr. fižola, mala namesto majhna, večim), skladenjske (besedni red, vezljivost ipd., npr. ena bolj od okuženih občn; nimam se kaj za pritoževati) ali besedne značilnosti (npr. pasoš, orenk, leder) pogovornega/narečnega jezika, (1) ohranimo izvorno obliko (fora) ali (2) določimo krovno standardizirano obliko te pogovorne/narečne besede oz. njene oblike, če hkrati s slovničnimi ali besednimi značilnostmi govorjenega jezika prepoznamo tudi glasovne premene (npr. zrihtov -> zrihtal).* (Verdonik et al. 2023a: 9)

Vendar prepoznavanje tega ni vedno enostavno in enoznačno, saj

> *se vedno znova odpirajo primeri, za katere ni vnaprej znane najprimernejše standardizirane oblike ali pa se za določene primere pokaže, da niso tako enoznačni, kot se zdi, ko prvič naletimo nanje (npr. poleg besede gučati se naknadno razkrije še varianta gučiti in je treba za nazaj raziskati, katera od teh oblik naj velja kot standardizirana oz. ali naj se vodita dve različni obliki).* (Verdonik et al. 2023a: 9)

Zato se je pri standardiziranem zapisu vodil seznam takih težavnih oblik, ki je objavljen kot del dokumentacije baze.

Enako kot pri pogovornem je tudi pri standardiziranem zapisu novost v primerjavi s prvo različico uvedba ločil in velikih začetnic pri izjavah. Razlog za to je bila predvsem potreba avtomatskega razpoznavanja govora po podatkih za učenje orodja za postavljanje ločil v razpoznani govor (punktuatorja). Vsebinsko pa ocenjujemo, da je v primerjavi s prvo različico korpusa Gos zaznaven premik k večjemu obsegu ohranjanja pogovornih oblik in leksemov ter manjšemu pretvarjanju/prevajanju v vzpostavljene standardne oblike in lekseme. Tak pogost primer je osrednjeslovenski veznik *k*, ki je bil v prvi različici Gos interpretiran v različne veznike (*ker, ko, ki, kot, kjer, kar, kaj*), v korpusu Artur pa je bil uveden unikaten zapis *ke*, ki omogoča enostavnejši avtomatski dostop do primerov teh rab.

Standardizirani zapis v korpusu Artur je bil v prvem koraku avtomatsko pripravljen s pomočjo prevajalnega modela, učenega na bazi Gos Videolectures (Verdonik et al. 2021), in zatem ročno popravljen. Popravki so se izvajali samo v zapisu govora, segmentacija na osnovne enote, označevanje menjavanja govorcev, popisani podatki o govorcih in posnetkih ter označevanje zvočnih ozadij in dogodkov so se pri standardiziranem zapisu popravljali le izjemoma, če je bila opažena očitna napaka oziroma pri označevanju značilnosti izgovorjave. Standardizacija zapisa govora vključuje zaradi zgoraj omenjenih negotovih primerov, ki jih ne moremo vnaprej predvideti, zahtevne odločitve, za katere je zaželeno, da so med seboj čim bolj skladne. Te odločitve imajo hote ali nehote vpliv na rezultate analiz korpusnih podatkov, saj lahko z zapisom govora kakšne značilnosti govorjene rabe nehote zakrijemo. Pri izvajanju standardiziranega zapisa z zunanjimi, laičnimi zapisovalci tega ne moremo zagotoviti, zato je vse standardizirane zapise v korpusu Artur popravljal en sam ustrezno izurjen sodelavec projekta s poglobljenim jezikoslovnim znanjem, ki se je o težavnejših primerih, popisanih v dokumentaciji korpusa (Verdonik et al. 2023a), periodično posvetoval z ožjo skupino sodelavcev. Kljub temu ostaja eden osrednjih problemov standardiziranega zapisa konsistentnost zapisovanja, po eni strani skozi čas in nadgradnje (odločitve so se od prve različice korpusa Gos do Arturja delno prilagodile, kot opisano zgoraj), po drugi strani pa tudi skozi projekt. Usklajevanje odločitev za nazaj ovira predvsem pomanjkanje ustreznega orodja in okolja, v katerem bi se to lahko izvedlo, to bi bil idealno konkordančnik, ki bi omogočal iskanje, poslušanje in hkrati tudi neposredno popravljanje zapisov v jezikovnem viru – zadnji korak pa v obstoječih orodjih ni podprt.

3.3 Naknadni popravki pogovornega zapisa

Hkrati s popravljanjem standardiziranega zapisa se žal vedno znova pokaže, da so določeni popravki potrebni tudi v pogovornem zapisu. V korpusu Artur so bili ti popravki večinoma: (1) ločila in velike začetnice, (2) napačni zapisi skupaj ali narazen ali (3) manjkajoča beseda v zapisu. Napake iz točke 1 so bile pogoste in utrujajoče za usklajevanje med obema zapisoma, zato so bile na koncu avtomatsko prenesene iz standardiziranega v pogovorni zapis. Napake iz točk 2 in 3 pa zahtevajo izredno visoko pozornost zapisovalca, saj so v primeru neusklajenega števila besed v izjavi med pogovornim in standardiziranim zapisom težave pri uporabi korpusa in pretvorbi v formate za javno izdajo, odkrivati in popravljati pa jih je zelo težavno.

Časovno učinkovitost ročnega popravljanja standardiziranega zapisa je ovirala velika količina kratkih datotek v parlamentarnem delu korpusa Artur, saj je pri več tisoč datotekah veliko časa potrebnega samo za rokovanje z datotekami. Paralelna primerjava obeh zapisov, standardiziranega in pogovornega, bi bila bolj učinkovita, če bi bila oba zapisa odprta v enem oknu eden pod drugim. Z uporabljenim orodjem, Transcriber 1.5.1 (Barras et al. 2000), je bilo mogoče odpreti oba zapisa samo v ločenih oknih in ju shranjevati v ločenih datotekah.

4 Kritična analiza zapisovanja govora v korpusu Artur

4.1 Dvotirni način zapisovanja

Dobrushina in Sokur (2022) sta kritična do praks, ko se uporablja dvotirni zapis, češ da je časovno celo bolj zahteven kot fonetični zapis, ker zahteva dve ravni zapisa namesto ene. Časovna zahtevnost zapisovanja in označevanja govora zagotovo zahteva osrednjo pozornost, ko razmišljamo o strategijah za naprej.

Koliko dodatnega časa torej zahteva dvotirni sistem? Ocene po izdelavi korpusa Artur so sledeče: za segmentacijo posnetka, označevanje govorcev in menjavanja govorcev ter ročni zapis povedanega (bodisi v pogovornem bodisi v standardiziranem načinu) potrebujemo okrog 20 ur dela za 1 uro posnetka, pri čemer je treba upoštevati, da je lahko v primeru zelo zahtevnih terenskih posnetkov z veliko hkratnega govora ali zelo narečnim govorom trajanje dela tudi bistveno daljše. Redakcija zapisa in oznak zahteva približno 1 uro dela za 1 uro posnetka, ob predpostavki, da pregledamo samo naključne segmente posnetka, ne celotnega posnetka. Koordiniranje zapisovanja in označevanja govora vključuje organiziranje dela, pripravo navodil, vzpostavitev delotoka, izbor in pripravo orodij in okolij za delo z datotekami na daljavo, iskanje in angažiranje sodelavcev transkriptorjev ter administrativno projektno delo. Običajno zahteva četrtinski do polovični delovni čas za celotno obdobje trajanja, odvisno od intenzivnosti dela.

Izvedba dodatnega nivoja zapisa (v našem primeru standardiziranega) zahteva skupno okvirno 100 ur za razvoj, preverjanje in zaganjanje algoritma za avtomatsko pretvorbo iz pogovornega v standardizirani zapis, pri čemer je prvi pogoj primeren učni korpus, za kar se lahko v prihodnje uporabi korpus Artur ali vsaj del Arturja. Ročno popravljanje avtomatsko predpripravljenega standardiziranega zapisa zahteva okvirno 4 ure dela za 1 uro posnetka. Končna validacija in usklajevanje obeh ravni zapisa besedo na besedo lahko predstavlja vse do dodatne ure dela za uro posnetka,

lahko pa tudi veliko manj, odvisno od obsega napak. V skupnem seštevku delo nikakor ni podvojeno, ampak bi lahko bila groba ocena okrog 25 % dodatnega dela, da izvedemo dvotirni namesto enotirnega zapisa govora.

Če se odločimo samo za enotirni zapis govora, je najverjetnejša izbira standardizirani zapis, saj omogoča primerno nadaljnjo avtomatsko obravnavo podatkov. Izgubimo torej pogovorni zapis, ki pa je v primerjavi s standardiziranim (1) bolje usklajen z govorjeno rabo in bolj natančno odraža dejansko podobo besed in besednih oblik, (2) omogoča bolj podrobno luščenje za govor značilnih besed in besednih oblik iz korpusa, (3) omogoča bolj natančen avtomatsko pripravljen fonemski zapis podatkov, (4) je za zunanje izvajalce transkriptorje lažje usvojljiv in manj zahteven kot standardizirani zapis.

4.2 Težavnejša vprašanja zapisovanja

Kritični pogled na vzpostavljene smernice poleg vprašanja dvotirnega zapisa odpira vsaj še štiri vprašanja.

Prvo se nanaša na segmentiranje govora na osnovne enote in uvedbo ločil (gl. 3.1). Kot že navedeno, v korpusu Artur označeni segmenti ne ustrezajo vedno temu, kar bi lahko interpretirali kot ena izjava ali osnovna enota govora, ampak se bolj opirajo na premore, torej na specifično prozodično lastnost. Tak način segmentiranja odpira tehnične probleme za višje ravni označevanja, zlasti skladenjsko in pragmatično, zaradi prelomov znotraj osnovne enote po eni strani in prehodov prek mej osnovnih enot po drugi. Uvedba ločil lahko pomaga premoščati te probleme, saj omogoča alternativne osnovne enote označevanja, ki izhajajo iz skladenjskih značilnosti. Z vidika čim širše uporabnosti govornih virov za različne vede je obstoječa rešitev v korpusu Artur vseeno lahko dobra, čeprav zahteva nekaj več naknadnega procesiranja korpusnih zapisov pred nadaljnjim označevanjem. Pomembno pa je, da so ločila dodana s premislekom in strokovno, saj dodatno interpretirajo govor. Način, da se dodajo v standardiziranem zapisu, ki ga izvaja strokovno visoko usposobljena oseba, ter nato avtomatsko prenesejo v pogovorni zapis, se zato zdi dobra praksa.

Drugič, vedno znova se odpira problematika zapisovanja opornih signalov, to so običajno besedice *ja, mhm, aha, aja* ipd., ki jih sogovornik izreče, medtem ko drugi govorec govori. V korpusu Artur je natančnost zapisovanja opornih signalov zmanjšana v primerjavi s prvo različico korpusa Gos. Razlog je velika časovna zahtevnost in dodatno segmentiranje govora za povsem natančen zapis. Za potrebe tehnologij je manjša natančnost zapisovanja opornih signalov sprejemljiva, za določene jezikoslovne, sociolingvistične in druge raziskave pa je lahko slabost. Tudi tukaj vidimo rešitev v zamenjavi orodja za zapisovanje in označevanje govora s takim, ki bo omogočalo večtirni način zapisovanja govora v enem oknu.

Tretjič, uvedba posebnega znaka za polglasnik v pogovornem zapisu je po eni strani res omogočila nekoliko bolj enoznačno morebitno avtomatsko pretvorbo v fonetični zapis, a je praksa pokazala, da zapis polglasnika ni dosleden. Kolikor bolj se pogovorni zapis približuje fonemskemu, toliko več nedoslednosti vključuje in več je nejasnih vmesnih primerov, ko se je težko odločiti, kako interpretirati izgovorjeni glas.

Četrtič, potrebna bi bila analiza zapisovanja mejnih primerov standardiziranega zapisa in razširjena sistematična razlaga, po katerih načelih določamo »oblikoslovne, skladenjske ali besedne značilnosti pogovornega/narečnega jezika« (Verdonik et al. 2023a: 7), ki jim ohranimo izvorno obliko, oziroma po kakšnih kriterijih določimo standardizirano obliko, pod katero vodimo tako besedje. Samo za ilustracijo – z dodatno razlago lahko na primer naslovimo primere, kot so: (1) splošnopogovorni ali regionalni pregibni vzorci (npr. *bolan – bolana*) (2) in besedje (npr. *probavati*), (3) poznani jezikovni procesi (npr. maskulinizacija; *mleko – mlek*), (4) prevzete besede (npr. *šoping*, *dugi*), (5) kratice, kot sta *S. P.* ali *D. O. O.*, ipd.

4.3 Orodje za zapisovanje

Za zapisovanje in označevanje govora smo v slovenskih govornih korpusih do zdaj uporabljali orodje Transcriber 1.5.1. Razlogi za njegov izbor so bili vedno znova v preprostosti uporabe za zunanje uporabnike, zanesljivosti, hitri natančni segmentaciji govora in že vzpostavljenih orodjih za uporabo in pretvorbo izhodnega formata datotek Transcriberja v formate za javno uporabo. Čeprav je Transcriber eno od pogosto uporabljanih orodij za zapisovanje in označevanje govora, pa so

pogosto v uporabi vsaj še Praat,[4] ELAN[5] in EXMARaLDA.[6] Za popravljanje dodatnega nivoja zapisa bi bilo idealno, da bi lahko odprli oba zapisa paralelno v enem oknu, kar omogočajo vsa tri navedena orodja, Transcriber pa ne. Razmislek o najučinkovitejšem orodju bo tako potreben tudi v prihodnje in tudi z upoštevanjem zaledne podpore posameznega orodja.

5 Zaključek

V prispevku smo se osredotočili na vprašanje, ali je dvotirni način zapisovanja govora v govornih korpusih, pri katerem se najprej pripravi pogovorni zapis, nato pa še standardizirani zapis, ki podpira nadaljnje avtomatsko označevanje zapisanega govorjenega besedila, v prihodnje še smiseln ali pa je vložek prevelik v primerjavi s koristmi. Zaključek je, da zlasti za jezikoslovne raziskave dvotirni način zapisovanja še vedno prinaša prednosti, zaradi katerih ga je smiselno nadaljevati. Izkušnje kažejo, da potrebno delo ni podvojeno, ampak povečano morda za četrtino, pri čemer vidimo še nekaj možnosti za večjo učinkovitost, ki bi jo lahko dosegli z novim učenjem algoritma za avtomatsko predpripravo standardiziranega zapisa na podatkih korpusa Artur in z uporabo orodja za zapisovanje in označevanje govora, v katerem bi lahko oba zapisa odprli v enem oknu, enega pod drugim.

Izpostavili smo tudi štiri točke v obstoječih standardih zapisovanja govora, na katere je treba biti v prihodnje še posebej pozoren in se nanašajo na navodila za segmentiranje in s tem povezano rabo ločil, na zapisovanje opornih signalov, natančnost pogovornega zapisa z dodatnimi znaki za foneme ter na nadgradnjo smernic za standardizirani zapis.

Ključna težava ostaja neredno financiranje s pretiranimi viški v (pre)kratkem časovnem obdobju, kar zahteva uporabo manj zaželenih bližnjic za zbiranje gradiv in onemogoča temeljite priprave orodij in avtomatskih algoritmov, s katerimi lahko delo pohitrimo, ter dolgimi vmesnimi obdobji brez kakršnegakoli financiranja, ko delo na govornih korpusih popolnoma zamre.

[4] https://www.fon.hum.uva.nl/praat/
[5] https://archive.mpi.nl/tla/elan
[6] https://exmaralda.org/en/

Literatura

Claude BARRAS, Edouard GEOFFROIS, Zhibiao WU, Mark LIBERMAN, 2000: Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication* 33/1–2, 5–22.

Nick CAMPBELL, 2005: Getting to the Heart of the Matter: Speech as the Expression of Affect; Rather than Just Text or Language. *Language Resources and Evaluation* 39, 109–118. Dostop 11. 4. 2024 na https://doi.org/10.1007/s10579-005-2699-y.

Emanuela CRESTI, Massimo MONEGLIA, 2005: *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Nina DOBRUSHINA, Elena SOKUR, 2022: Spoken Corpora of Slavic Languages. *Russian Linguistics* 46, 77–93. Dostop 25. 8. 2023 na https://doi.org/10.1007/s11185-022-09254-9.

Radovan GARABÍK, 2023: Slovenský hovorený korpus. *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: Zbornik povzetkov*. Ur. Mira Krajnc Ivič. Maribor: Univerza v Mariboru, Filozofska fakulteta. Dostop 25. 8. 2023 na https://doi.org/10.18690/um.ff.5.2023.

Radovan GARABÍK, Milan RUSKO, 2007: Corpus of Spoken Slovak Language. *Computer Treatment of Slavic and East European Languages*. Zbornik konference Slovko 2007. Ur. J. Levická, R. Garabík. Brno: Tribun. 222–236.

Dafydd GIBBON, Roger MOORE, Richard WINSKI (ur.), 1997: *Handbook of Standards and Resources for Spoken Language Systems*. Berlin, New York: Walter de Gruyter Publishers. Dostop 25. 8. 2023 na http://wwwhomes.unibielefeld.de/gibbon/Handbooks/gibbon_handbook_1997/index.html

Zuzana KOMRSKOVÁ, Marie KOPŘIVOVÁ, David LUKEŠ, Petra POUKAROVÁ, Hana GOLÁŇOVÁ, 2017: New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný casopis* 68/2, 219–228. Dostop 25. 8. 2023 na https://doi.org/10.1515/jazcas-2017-0031.

Jelena KUVAČ Kraljević, Gordana HRŽICA, 2016: Croatian Adult Spoken Language Corpus (HrAL). *FLUMINENSIA* 28/2, 87–102.

Robbie LOVE, Claire DEMBRY, Andrew HARDIE, Vaclav BREZINA, Tony MCENERY, 2017: The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3, 319–344. Dostop 25. 8. 2023 na https://doi.org/10.1075/ijcl.22.3.02lov

Brian MACWHINNEY, 2000: *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, Ney York: Lawrence Erlbaum Associates.

Thomas SCHMIDT, 2016: Construction and dissemination of a corpus of spoken interaction – tools and workflows in the FOLK project. *Journal for language technology and computational linguistics* 31/1, 127–154.

Margret SELTING, Peter AUER, Dagmar BARTH-WEINGARTEN, Jörg BERGMANN, Pia BERGMANN, Karin BIRKNER, Elizabeth COUPER-KUHLEN, Arnulf DEPPERMANN, Peter GILLES, Susanne GÜNTHNER, Martin HARTUNG, Friederike KERN, Christine MERTZLUFFT, Christian MEYER, Miriam MOREK, Frank OBERZAUCHER, Jörg PETERS, Uta QUASTHOFF, Wilfried SCHÜTTE, Anja STUKENBROCK, Susanne UHMANN, et al., 2009: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion* 10, 353–402.

Mitja TROJAR, Andreja BIZJAK, 2023: Transkribiranje govora pri izdelavi govorne baze Artur: od pogovornih k standardiziranim zapisom. *Razvoj slovenščine v digitalnem okolju*, 39–59. Ljubljana: Založba Univerze v Ljubljani. Dostop 10. 4. 2024 na https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9445.

Darinka VERDONIK, 2023: Zbiranje gradiv za govorne korpuse med Scilo in Karibdo. *Razvoj slovenščine v digitalnem okolju*, 15–37. Ljubljana: Založba Univerze v Ljubljani. Dostop 10. 4. 2024 na https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9447.

Darinka VERDONIK, Iztok KOSEM, Ana ZWITTER VITEZ, Simon KREK, Marko STABEJ, 2013: Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation* 47/4, 1031–1048

Darinka VERDONIK, Ana ZWITTER VITEZ, 2020: *Slovenski govorni korpus Gos*. 1. e-izd. Ljubljana: Znanstvena založba Filozofske fakultete. (Zbirka Sporazumevanje). Dostop 25. 8. 2023 na https://e-knjige.ff.uni-lj.si/, http://www.dlib.si/details/URN:NBN:SI:DOC-X9DAJU5X.

Darinka VERDONIK, Tomaž POTOČNIK, Mirjam SEPESY MAUČEC, Tomaž ERJAVEC, Simona MAJHENIČ, Andrej ŽGANK, 2021: *Spoken corpus Gos VideoLectures 4.2 (transcription)*. CLARIN.SI Data & Tools. Maribor: Faculty of Electrical Engineering and Computer Science, University of Maribor. Dostop 25. 8. 2023 na http://hdl.handle.net/11356/1444.

Darinka VERDONIK, Andreja BIZJAK, 2023: *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru. Dostop 25. 8. 2023 na http://hdl.handle.net/11356/1772.

Darinka VERDONIK, Andreja BIZJAK, Mitja TROJAR, 2023a: *Standardizirani zapis v govorni bazi Artur projekta RSDO*. Maribor: Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru; Inštitut za slovenski jezik Frana Ramovša, ZRC SAZU. Dostop 25. 8. 2023 na http://hdl.handle.net/11356/1772.

Darinka VERDONIK, Andreja BIZJAK, Mirjam SEPESY MAUČEC, Lucija GRIL, Simon DOBRIŠEK, Janez KRIŽAJ, Gregor STRLE, Marko BAJEC, Iztok LEBAR BAJEC, Tjaša ŠOLTES, Jure LOKOVŠEK, Mitja TROJAR, Tomaž ERJAVEC, Mitja BERNJAK, Jerneja ŽGANEC GROS, Peter ČAKŠ, Matevž PUCER, Mitja CVETKO, Jani PAVLIČ, Marijana ZELENIK, Marija IVANOVSKA, Klemen GRM, Jure LONGYKA, Aleš MIHELIČ, Boštjan VESNICER, Naum DRETNIK, 2023b: *ASR database ARTUR 1.0 (transcriptions)*. Maribor: Faculty of Electrical Engineering and Computer Science, University. CLARIN.SI Data & Tools. Dostop 25. 8. 2023 na https://www.clarin.si/repository/xmlui/handle/11356/1772.

Darinka VERDONIK, Andreja BIZJAK, Andrej ŽGANK, Mitja BERNJAK, Špela ANTLOGA, Simona MAJHENIČ, Peter ČAKŠ, Matevž PUCER, Mitja CVETKO, Jani PAVLIČ, Marijana ZELENIK, Simon DOBRIŠEK, Janez KRIŽAJ, Gregor STRLE, Marija IVANOVSKA, Klemen GRM, Marko BAJEC, Iztok LEBAR BAJEC, Tjaša ŠOLTES, Jure LOKOVŠEK, Jure LONGYKA, Mitja TROJAR, Jerneja ŽGANEC GROS, Aleš MIHELIČ, Boštjan VESNICER, Naum DRETNIK, David BORDON, 2023c: *ASR database ARTUR 1.0 (audio)*. CLARIN.SI Data & Tools. Maribor: Faculty of Electrical Engineering and Computer Science, University of Maribor. Dostop 25. 8. 2023 na https://www.clarin.si/repository/xmlui/handle/11356/1776.