Transkribiranje govora pri izdelavi govorne baze Artur: od pogovornih k standardiziranim zapisom

Mitja TROJAR
ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša

Andreja BIZJAK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Povzetek

Prispevek predstavlja načela za zapis govora pri izdelavi govorne baze Artur in opis izvedbe transkribiranja govora v projektu RSDO. Opisana so načela za zapis govora za pripravo pogovornih zapisov in praktični vidiki njihove priprave. Sledi opis priprave standardiziranih zapisov, ki so bili pripravljeni z ročnim popravljanjem avtomatskih pretvorb pogovornih zapisov. Prispevek zaokrožuje opis izzivov pri izdelavi pogovornih in standardiziranih zapisov ter priporočila za podobne projekte v prihodnosti.

Ključne besede: govorjeni jezik, transkripcije, pogovorni zapisi, standardizirani zapisi, govorna baza Artur

Abstract

This chapter presents principles of transcribing speech in the making of the Artur speech database and a description of speech transcription in the project Development of Slovene in a Digital Environment. It includes a description of principles used in the creation of orthographic transcriptions as well as its practical aspects, which is followed by an account of the making of standardised transcriptions, which were created by making manual corrections to automatic conversions of orthographic transcriptions. The chapter concludes with a presentation of challenges encountered in

the making of orthographic and standardised transcriptions and with recommendations for similar future projects.

Keywords: spoken language, transcriptions, orthographic transcriptions, standardised transcriptions, Artur speech database

1 Uvod

Delovni sklop 2 projekta RSDO[1] je bil namenjen razvoju govornih tehnologij za slovenščino, pri čemer je bil osnovni cilj projekta izdelava razpoznavalnika za slovenščino.[2] Za razvoj strojnega razpoznavanja govora je bilo treba zagotoviti dovolj veliko bazo transkribiranih posnetkov avtentičnega govora v raznolikih komunikacijskih okoliščinah. V ta namen je bila ustvarjena govorna baza Artur (**A**vtomatsko **r**azpoznavanje govora **R**azvoj slovenščine v digi**t**alnem okolj**u**), ki skupaj vsebuje 1094 ur posnetega govora. Visoko kakovost transkripcij smo v projektu RSDO poskušali zagotoviti z dvotirnim načinom transkribiranja: najprej so bili izdelani t. i. pogovorni zapisi, ki so bili nato pretvorjeni v t. i. standardizirane zapise. Odločitev za dvotirni način transkribiranja govora je bila sprejeta iz dveh razlogov. Prvi je ta, da je bil dvotirni način uporabljen že pri izdelavi govornega korpusa Gos (Verdonik in Zwitter Vitez, 2011). Primerljiva metodologija transkribiranja je omogočila razširitev korpusa Gos z izborom posnetkov in transkripcij, ki so nastali v projektu RSDO (gl. Gos 2.0).[3] Drugi razlog je ocena, da bi izdelava samo standardiziranih zapisov preveč obremenila transkriptorje, kar bi po predvidevanjih precej

1 Projekt Razvoj slovenščine v digitalnem okolju sta med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020).

2 Aktivnosti izdelave govorne baze je koordinirala Darinka Verdonik (Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru), v projektnem sklopu pa so sodelovale še 3 partnerske ustanove iz znanstvenoraziskovalnega okolja (Univerza v Ljubljani, Institut »Jožef Stefan« in ZRC SAZU), 2 partnerja iz gospodarstva (Slovenska tiskovna agencija, d. o. o., Alpineon, d. o. o.), zunanji izvajalci (Fixmedia, d. o. o., Kreativist, d. o. o., Akademija INT, d. o. o., TAIA, d. o. o.) ter študentke in študenti Univerze v Ljubljani in Univerze v Mariboru.

3 https://viri.cjvt.si/gos/

povečalo število napak v transkripcijah (gl. Verdonik, Trojar in Bizjak, 2023a).

V prispevku je najprej predstavljena struktura baze Artur, opis delotoka, opis načel, po katerih so bili pripravljeni pogovorni in standardizirani zapisi, nakazane pa so tudi težave, s katerimi smo se srečali pri gradnji baze, ter priporočila za gradnjo primerljivih baz v prihodnosti.

2 Struktura govorne baze Artur in opis delotoka njene izgradnje

Govorna baza Artur je sestavljena iz 4 sklopov:[4]

1. **Brani govor** (573 ur posnetkov): posnetki prebranih povedi, zajetih iz dela korpusa Gigafida 2.0, ki je dostopen pod ustrezno licenco (CC BY); povedi je bralo pribl. 1000 govorcev (pribl. 30 minut govora na posameznega govorca), ki so ustrezali vnaprej določenim demografskim kriterijem (uravnoteženost govorcev po spolu, starosti, statistični regiji stalnega bivališča in prvem jeziku);[5] za ta sklop baze Artur transkripcije niso bile izdelane, ker so funkcijo transkripcij (standardiziranega zapisa) opravljale povedi iz pisnega korpusa.[6]
2. **Javni govor** (208 ur posnetkov): posnetki javnih dogodkov, ki vključujejo novinarske konference, okrogle mize, intervjuje, nagovore, predavanja, seminarje, posvete in moderirane pogovore, ki so v času pandemije covida-19 večinoma potekali prek spleta. Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 62 ur posnetkov.

4 Spodaj navedeni podatki o bazi Artur in opis njene strukture so povzeti po Verdonik, Bizjak in Dobrišek (2023).

5 Poleg opisanega branja povedi sklop branega govora vsebuje še posnetke govora enega samega (izšolanega profesionalnega) govorca v obsegu 50 ur (za razvoj avtomatske sinteze govora) in posnetke črkovanj (v obsegu 10 ur).

6 Povedi so bile izbrane tako, da izbor povedi odraža dejansko distribucijo trifonov v slovenskih povedih, dobljeni nabor povedi pa je bil v nadaljevanju avtomatsko in ročno prefiltriran tako, da so bile izločene povedi, ki so vsebovale besede, katerih zapis se bistveno razlikuje od zapisa po slovenskem črkopisu (zlasti citatne besede, npr. *pole position*), kratice, jezikovne napake, ali pa so bile povedi kako drugače neustrezne (žaljiv govor, nepopolne povedi ipd.). Za podrobnejši opis izbora povedi gl. Žganec Gros in Vesnicer (2021) in Žganec Gros idr. (2023).

3. **Nejavni govor** (112 ur posnetkov): posnetki govora v nejavnih govornih položajih, in sicer gre za tri tipe govornih dogodkov: proste dialoge med sogovornikoma, proste monologe ter razlaganje in opisovanje. Govorci so bili izbrani po istih demografskih kriterijih kot v sklopu branega govora. V nejavni govor so vključeni tudi posnetki, namenjeni razvoju dveh specializiranih razpoznavalnikov govora, ki vključujejo naslednje govorne dogodke: opis obraza ter brane in spontane ukaze za upravljanje pametnega doma. Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 74 ur posnetkov.
4. **Parlamentarni govor** (201 ura posnetkov): posnetki javnih sej Državnega zbora Republike Slovenije iz dveh sklicev med letoma 2010 in 2018. Vsaka datoteka vsebuje govor enega samega govorca. Govor posameznega govorca je lahko zajet na več posnetkih, vendar v celoti ne presega 3,5 ure. Za pripravo pogovornih zapisov so bili uporabljeni zapisi sej, ki jih pripravljajo v Državnem zboru. Študentke in študenti so zapise sej uredili in popravili tako, da so ustrezali standardom za pogovorni zapis (gl. spodaj). Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za 201 uro posnetkov.

Transkripcije (pogovorni in standardizirani zapisi) so bile pripravljene za javni, nejavni in parlamentarni govor oziroma za skupno 337 ur posnetega govora. S pogovornimi in standardiziranimi zapisi je torej opremljena pribl. tretjina posnetkov v govorni bazi Artur, s čimer so bili projektni cilji doseženi in celo preseženi.

Časovno najzahtevnejši fazi delotoka sta bili priprava pogovornih in standardiziranih zapisov. Delotok je bil razdeljen v več faz, ki so se pri javnem, nejavnem in parlamentarnem govoru zvrstile v podobnem zaporedju: oddaja izvornih avdio posnetkov, pridobljenih od različnih virov ali posnetih na terenu, validacija posnetkov in odobritev ustreznih posnetkov glede na tehnično kakovost in vsebinsko ustreznost (npr. odsotnost sovražnega govora), oddaja soglasij govorcev in dokumentacije z metapodatki o posnetkih in govorcih, ročna priprava pogovornih zapisov, validacija pogovornih zapisov, avtomatska

pretvorba standardiziranih zapisov iz pogovornih zapisov, pregled in ročno popravljanje avtomatsko tvorjenih standardiziranih zapisov. Ročnemu pregledu so sledili še avtomatski pregledi, napake, ki so bile z njimi odkrite, pa so bile popravljene ročno ali avtomatsko (z izdelavo ustreznih skriptov). Na osnovi tako pridobljenih standardiziranih zapisov in avdio posnetkov je bil razvit splošni razpoznavalnik govora za slovenščino ter dva domensko specifična razpoznavalnika.

Pri načrtovanju delovnega procesa je pomembno zagotoviti transparentnost in sledljivost posameznih korakov, kar sodelujočim omogoča, da so seznanjeni, kateri posnetki ali zapisi posnetkov so že validirani, kateri so odobreni, zavrnjeni ali trenutno še v obdelavi. Delotok je bil zasnovan v vseh fazah priprave dvonivojsko (vsaka mapa je imela svojo kopijo kot varnostni arhiv z omejenimi pooblastili za spreminjanje njene vsebine). Tovrstna sledljivost je namreč ključnega pomena pri morebitnem kasnejšem iskanju izvora ali tipa napak, hkrati pa prepreči izgubo datotek. Zasnova delotoka mora biti tudi dovolj fleksibilna, da omogoča naknadno dodajanje novih faz, če se zanje kadar koli med pripravo baze pojavi potreba, kot se je izkazalo pri Arturju (npr. naknadno dodana faza prenosa ločil in velikih začetnic iz standardiziranih v pogovorne zapise).

Pri izdelavi govorne baze Artur je bila za obdelavo in shranjevanje datotek uporabljena oblačna platforma Nextcloud.

3 Načela, uporabljena pri izdelavi pogovornih in standardiziranih zapisov

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.«[7] (Verdonik in Bizjak, 2023)

7 Opozoriti velja, da je v projektu RSDO pogovorni zapis služil le kot sredstvo za doseganje vmesnega cilja, tj. izdelave standardiziranega zapisa. Pri izdelavi razpoznavalnika, ki je bil končni cilj projekta, je bil namreč uporabljen samo standardizirani zapis. Pogovorni zapis je torej imel izrazito pomožno vlogo in ni nadomestek za (znanstveno/jezikoslovno) fonetično transkripcijo govora.

Govor je zapisan v slovenskem črkopisu v skladu z veljavnimi načeli, po katerih se glasovi zapisujejo s črkami. Pri tem se upoštevajo omejitve, ki izhajajo predvsem iz omejenega nabora črk, da bi karseda natančno predstavili glasovno podobo govora (Verdonik in Bizjak, 2023). Novost v pogovornem zapisu je npr. poseben znak za polglasnik (@), ki ga do Arturja v govornih korpusih za slovenščino ni bilo.

Pogovorni zapis poleg zapisa govora vključuje še segmentacijo govora, označevanje menjavanja govorcev, označevanje akustičnega ozadja (npr. prisotnost šuma ali glasbe), akustičnih dogodkov (npr. nenadni krajši zvoki, kašljanje, glasni vdihi) ter osnovnih neverbalnih značilnosti (npr. smeh ali premor). Pri izdelavi baze Artur so ga v orodju Transcriber 1.5.1. ročno pripravili zunanji izvajalci in študenti, koordinator transkribiranja pa je naključne dele pogovornih zapisov validiral in po potrebi popravil. Glede na izkušnje z Arturjem se je izkazalo, da je za 1 uro posnetka govora potrebnih okrog 20 ur dela za zapis, segmentiranje in označevanje govora (prim. Verdonik, Trojar in Bizjak, 2023a). Povedano še drugače, za pripravo pogovornega zapisa 2 oz. 3 minut govora je v povprečju potrebna ena ura dela.

Eden od prvih korakov pri izdelavi pogovornega zapisa je segmentiranje govora, ki je bilo pri pripravi govorne baze Artur delno prilagojeno za potrebe razvoja splošnega razpoznavalnika govora. Glede na to naj segmenti ne bi bili (pre)dolgi, tj. trajajoči več kot 10 sekund. Poleg tega smo upoštevali, da lahko mejo segmenta določimo le tam, kjer je v govoru dovolj premora, tj. vsaj 0,2 sekunde, ne da bi odrezali del predhodnega ali del naslednjega fonema. Glavni usmeritvi pri postavljanju meja med segmenti sta bila (1) kratek premor v govoru in (2) dolžina segmenta, ki ne sme biti predolga (Verdonik in Bizjak, 2023). Pri tako prilagojenem načinu segmentiranja označeni segmenti ne sovpadajo vedno s stavki oz. izjavami kot semantično in skladenjsko zaključenimi enotami, kar je bilo identificirano kot težava na višjih ravneh označevanja.

Navodilo za daljše premore, trajajoče več kot 1,5 sekunde, govor v tujem jeziku in nerazumljiv govor je, da se jih označi kot prazen segment ali izjavo brez govorca. Če je nerazumljiva zgolj posamezna beseda ali kratka fraza, se vstavi oznako *neraz.* Hkratni govor, ki se

pojavi v začetku ali ob koncu segmenta, ko govorca govorita hkrati, se ustrezno označi in se ga, če je razumljiv, tudi zapiše. Ob vsaki menjavi govorcev je treba paziti, da se menjava ustrezno označi.

Akustično ozadje se označi, kadar se v ozadju govora nenadoma pojavijo dalj časa trajajoči zvoki (najmanj 3 sekunde), ter določi, ali je šum v ozadju govor, glasba ali kaj drugega (npr. aplavz, zvonjenje telefona, prometni hrup). Kadar pa se med govorom pojavijo krajši zvoki (pribl. do ene besede), se vstavijo kot akustični dogodek (npr. zehanje, kihanje, vdih, izdih).

Besedni fragmenti (prekinjene besede, samopopravki) so označeni s praznim oklepajem stično za besedo, npr. *dru()*. Če se v govoru pojavijo osebni podatki o govorcih, ki niso javne osebnosti (npr. ime in priimek), se jih s piskom anonimizira. Številke (tudi vrstilni števniki) se izpišejo z besedo znotraj oglatih oklepajev, npr. *[tretje]*. Datumi se zapisujejo znotraj zavitih oklepajev, npr. *{peti osmi dva tisoč devet}*.

Novost v jezikovni bazi Artur v primerjavi s preteklimi govornimi korpusi pri nas je uvedba nekaterih dodatnih znakov za foneme, od katerih po pogostosti izstopa @ za polglasnik, omenimo pa še *$g* za zveneči *h* in *$r* za mehkonebni *r*. Nova je tudi vpeljava ločil in velikih začetnic v pogovorni in standardizirani zapis.

Redukcije glasov so v pogovornih zapisih upoštevane, saj se neizgovorjeni glasovi ne zapisujejo, npr. *tud* (Verdonik in Bizjak, 2023), premene po zvenečnosti pa se niso zapisovale, saj smo predvidevali, da bi bili najeti zunanji izvajalci ali študenti pri njihovem zapisu preveč nedosledni. Drugače je pri parlamentarnem govoru, saj je bil ta zapisan še pred uvedbo skupnih smernic za govorno bazo Artur. Premene po zvenečnosti so v parlamentarnem govoru zapisane, čeprav ne dosledno, npr. *gdo*, različen pa je tudi zapis dvoustničnega *u̯*, ki je praviloma zapisan s črko *u*, npr. *obraunavau*, in zapis kratic, ki so mestoma zapisane z veliki črkami, npr. *ZOFI*.

S ciljem čim bolj poenotenega zapisa neverbalnih in polverbalnih izrazov (npr. *eee, hm, uh, šššš*) so bile dopolnjene smernice za njihov zapis in razširjen seznam identificiranih neverbalnih in polverbalnih izrazov. Določili smo, da jih prednostno zapisujemo z največ

eno besedo, in če le gre, s tremi črkami, zelo podoben izraz pa zapišemo vedno na isti način, brez variacij (Verdonik, Bizjak 2023). Na začetku vedno dodamo znak #, npr. *#eem.*

V spodnji tabeli je navedenih nekaj smernic za izdelavo pogovornega in standardiziranega zapisa, pripravljenih za izgradnjo govorne baze Artur (Verdonik in Bizjak, 2023), skupaj s konkretnimi primeri iz iste baze.

Tabela 1: Smernice za pripravo pogovornega in standardiziranega zapisa po posameznih problemskih sklopih.

Sklop	Pogovorni zapis	Primer	Standardizirani zapis	Primer
Redukcije	Glasov, ki niso izgovorjeni, ne zapisujemo.	mamo, tko	Uporabljamo nereducirane oblike, skladno s pravopisno normo.	imamo, tako
	Redukcijo pomožnega glagola *bi* v *b* zapisujemo kot samostojno besedo, redukcije in premene oblik za prihodnjik pa kot: *čev* (*če bo*), *navm* (*ne bom*), *nav* (*ne bo*).	ne b navm	Standardizirane oblike zanikanega pomožnega glagola, ki so v pogovornem zapisu zapisane kot ena beseda, npr. *navm* pišemo z znakom + in stično: *ne+bom*.	ne bi ne+bom
	Polglasnik vedno zapisujemo z znakom @.	misl@m, fil@m, z@, @ldje, j@t, p@r	Posebnega znaka za polglasnik ne uporabljamo. Polglasnik se zapisuje skladno s pravopisno normo.	mislim, film, z, ljudje, iti, pri
Premene po zvenečnosti	Premen po zvenečnosti, razen pri predlogih *s/z* in *k/h*, v pisavi ne upoštevamo.	tud, fizka, j@z	Premene po zvenečnosti se načeloma ne upoštevajo oz. zapis besed sledi pravopisni normi.	tudi, fizika, jaz
Dvoustnični *u̯* in samoglasnik *u*	Dvoustnični *u̯*, ki ni nosilec zloga, v neknjižnih oblikah zapisujemo s črko *v*.	šov, prov	Dvoustnični *u̯* se zapisuje skladno s pravopisno normo, tj. s črkama *v* in *l*. Zapis se ravna po pravopisni normi, reducirane oblike deležnikov na *-il*, *-al*, *-el* se pišejo s samoglasnikom.	šel, prav
	Če dvoustnični *u̯* nastopi v knjižni besedni obliki, izgovorjeni skladno s standardom, ohranimo knjižni zapis.	bil, gledal		bil, gledal
	Če je glas *u* samoglasniški, tj. je nosilec zloga, ga pišemo s črko *u*.	odloču, padu, izpelu		odločil, padel, izpeljal
	Enako velja za predlog *v*, izgovorjen kot samoglasniški *u*.	u sobi	Predlog *v* se zapisuje skladno s pravopisno normo, tj. vedno s črko *v*.	v sobi

Sklop	Pogovorni zapis	Primer	Standardizirani zapis	Primer
Narečno specifični glasovi	Diftonge in druge pokrajinsko specifične foneme, ki jih v knjižnem jeziku ni, pišemo z najbližjimi ustreznimi črkami.	guvurim, tku, gučali, fseh, fsekakor	Oblike besed s pokrajinsko specifičnimi variantami fonemov oz. fonemi se nadomeščajo z ustreznimi knjižnimi oblikami besed.	govorim, tako, gučali, vseh, vsekakor
	Zveneči primorski *h* lahko zapišemo tudi z znakom *$g*, mehkonebni koroški *r* pa z znakom *$r*.	knji$g	Narečnim oblikam, ki nimajo ustreznih oblik v knjižnem jeziku, se priredi standardizirana oblika, ki sledi pravilom slovenskega črkopisa. Pri standardizaciji se prednostno uporablja standardizirane oblike iz narečnih slovarjev.	knjig
Lastna imena, citatne in tuje besede	Domača lastna imena zapisujemo skladno s pravopisom, tuja lastna imena pa tako, kot so izgovorjena.	Avstro-Ogrske, R@dovlci Mark Kjub@n, Zum, Heri Poter	Domača lastna imena zapisujemo skladno s pravopisom. Tuja lastna imena se prav tako zapisujejo v skladu s pravopisom, tj. bodisi podomačeno bodisi citatno (če konkretno lastno ime še ni podomačeno v pravopisnih priročnikih, se prednostno uporabi citatno obliko tujega lastnega imena).	Avstro-Ogrske, Radovljici Mark Cuban, Zoom, Harry Potter
	Citatne besede in besede oz. kratke fraze v tujem jeziku se pišejo, tako kot so izgovorjene.	fen, pojnt, trejler komon sens, riz@ning, Vourld of Vorkreft	Citatne besede in besede oz. kratke fraze v tujem jeziku se praviloma pišejo citatno, lahko pa tudi podomačeno, če je podomačeni zapis že uveljavljen oz. registriran v korpusih in slovarjih slovenskega jezika.	fan, point, trailer common sense, reasoning, World of Warcraft
Ločila Pisanje skupaj, narazen ali z vezajem	Ločila uporabljamo v njihovi skladenjski rabi in skladno s pravopisom. Tako zapisujemo tudi besede skupaj, narazen ali z vezajem.	pisiar-testi	Ločila uporabljamo v njihovi skladenjski rabi in skladno s pravopisom. Tako zapisujemo tudi besede skupaj, narazen ali z vezajem.	PCR-testi
Člen *ta* Kratice	Izjema so določni člen *ta*, ki ga pišemo stično, in kratice, ki jih pišemo tako, kot so izgovorjene, z malimi črkami in skupaj. Če je kratica lastno ime, jo pišemo z veliko začetnico. Okrajšav ne uporabljamo.	tamali tapravi A@g@r@f@t@c@p@p@-ja ajti-podjetjem Estea-jem	Določni člen ta pišemo z znakom + in stično. Kratice se pišejo skladno s pravopisom, tj. z vezajem med osnovo in končnico. Tvorjenke s kraticami se pišejo skladno s pravopisom.	ta+mali ta+pravi AGRFT CPP-ja IT-podjetjem STA-jem

Tabela 1 nakazuje razmerje med pogovornimi in standardiziranimi zapisi: standardizirani zapis je zapis govora, pri katerem se govorjeni jezik zapiše tako, kot bi bil zapisan v pisnem knjižnem jeziku. Standardizirani zapis lahko nastane na podlagi predhodnega zapisa govora,[8] ki se ga prilagodi (spremeni) tako, da nastali zapis (z možnimi predvidenimi odstopanji) ustreza pravilom slovenskega pravopisa, ki veljajo za pisni knjižni jezik. V standardiziranem zapisu oblike besed, značilne za govorjeni jezik, nastopajo v obliki, določeni za pisni knjižni jezik, besedilo je smiselno členjeno na povedi, stavke in besede (npr. besede v naslonskem nizu so ločene s presledki kot v knjižnem jeziku), uporabljena so ustrezna ločila.[9] Besedam, ki (še) niso registrirane v jezikovnih priročnikih za pisni knjižni jezik (zlasti Slovenski pravopis, SSKJ2, eSSKJ)[10] oz. niso zastopane v pisnih korpusih slovenskega jezika (predvsem Gigafida 2.0),[11] se priredi oblika, ki bi jo besede pričakovano imele, če bi se uporabljale v pisnem knjižnem jeziku.[12]

Končni validaciji pogovornih zapisov je v projektu RSDO sledila faza avtomatske pretvorbe pogovornih zapisov v standardizirane. Ročno preverjanje in popravljanje tako tvorjenih standardiziranih zapisov je v primeru zahtevnejših in nejasnih delov zahtevalo poslušanje posnetka in primerjavo z ustreznim segmentom v pogovornem zapisu.[13]

8 V projektu RSDO je standardizirani zapis nastal na podlagi pogovornega zapisa, načeloma pa bi lahko nastal tudi na podlagi npr. (dialektološke) fonetične transkripcije govora. Možno je seveda tudi, da bi standardizirani zapis nastal kot prvi zapis govora (tj. brez predhodnega pogovornega zapisa ali fonetične transkripcije).

9 Največja odstopanja od pravil in konvencij knjižnega jezika se pojavljajo na ravni besednega reda (ta se ni popravljal, ker se za razvoj razpoznavalnika zahteva, da so besede v transkripciji sinhronizirane z ustreznimi signali na posnetku) in zgradbe besed in povedi (ponavljanja besed, nedokončane besede in povedi, očitni lapsusi in drugi pojavi, značilni za govorjeni jezik (npr. nedoločni členi), niso bili izločeni), predvsem pa zaradi doslednega beleženja neverbalnih in polverbalnih glasov ter akustičnega ozadja in akustičnih dogodkov (smeh, vzdihi, hrup itd.). Ena od maloštevilnih izjem od pravopisnih pravil pri zapisovanju leksemov je zapis reducirane oblike veznikov *k@* (standardizirano v: *ke*), ki ji v knjižnem jeziku ustreza več veznikov, npr. *ker*, *ko*, *ki*. Za takšna odstopanja smo se odločili, ker smo poskušali zagotoviti, da bi bila pretvorba pogovornih zapisov v standardizirane čim manj odvisna od subjektivne interpretacije pripravljavca.

10 https://fran.si/

11 https://viri.cjvt.si/gigafida/

12 Oziroma se uporabi že standardizirano obliko, če gre npr. za narečne besede, ki so že registrirane v narečnih in/ali zgodovinskih slovarjih.

13 V idealnem primeru bi pregledovalec avtomatskih pretvorb v standardizirani zapis ob validaciji vsake transkripcije poslušal celotni avdio posnetek govora. V praksi bi to še precej bolj obremenilo pregledovalca in znatno podaljšalo pregledovanje standardiziranih zapisov.

Avtomatska pretvorba je bila sestavljena iz petih korakov: tokenizacije, pretvorbe v slovarske začetnice (ang. *truecasing*), prevoda, pretvorbe v besedilne začetnice (ang. *detruecasing*) in detokenizacije. Prevod je bil izveden z uporabo prevajalskega in jezikovnega modela, naučenega na bazi Gos VideoLectures 4.2 (Verdonik et al., 2021).

Avtomatske pretvorbe v standardizirane zapise so vsebovale napake, ki so bile ob pregledu odstranjene. V Tabeli 2 spodaj so vzporedno prikazani pogovorni zapisi, avtomatske pretvorbe v standardizirane zapise in ročno popravljeni standardizirani zapisi.

Tabela 2: Primerjava pogovornih zapisov, avtomatskih pretvorb v standardizirane zapise in ročno popravljenih standardiziranih zapisov.

	Pogovorni zapis	Avtomatska pretvorba v standardizirani zapis	Ročno popravljeni standardizirani zapis
1.	Čeprav jo bol uporablajo v@ ljudskem zdravilstvu, se danes spet več uporablja kot fčasih.	Čeprav jo bolj uporabljajo v ljudskem zdravilstvu, se danes spet več uporablja kot včasih.	Čeprav jo bolj uporabljajo v ljudskem zdravilstvu, se danes spet več uporablja kot včasih.
2.	Uporablamo tuji za astmo, plučni katar ...	Uporabljamo tudi za astmo, plučni katar	Uporabljamo tudi za astmo, pljučni katar ...
3.	Kakš@n nevljud@n?	Kakšen nevljudn?	Kakšen nevljuden?
4.	Če misl@n, da mi je tu nekaj stoplo v glavo, nekšne, nekšna ... črna energija.	Če mislim, da mi je to nekaj stoplo v glavo, nekakšne, neka ...črna energija	Če mislim, da mi je tu nekaj stopilo v glavo, nekšne, nekšna ... črna energija.
5.	To jaz bi rejs reko, to ne vem, kak so to bli vzgojeni, doma, f šoli, kje drugje, to so ... #puf.	To jaz bi res rekel, to ne vem kako so to bili vzgojeni, doma, v šoli, kje drugje, to so ...puf	To jaz bi res rekel, to ne vem, kako so to bili vzgojeni, doma, v šoli, kje drugje, to so ... #puf.

Zgledi v zgornji tabeli kažejo, da je bila avtomatska pretvorba v splošnem koristna in je pregledovalcu olajšala delo: pogosto nadaljnji popravki povedi niso bili potrebni (gl. zgled 1). Zgleda 2 in 3 kažeta, da so bile avtomatske pretvorbe v povedih včasih samo deloma ustrezne (zlasti znak @ za polglasnik je bil v pretvorbah pogosto samo izpuščen, ne pa tudi nadomeščen z ustrezno črko, tj. *e*). Zgled 4 kaže, da je bil leksem *nekšen* (po SSKJ2 narečno ‘nekak,

nekakšen') pretvorjen v leksem *nekakšen*, ki je bil nato ročno popravljen nazaj v *nekšen*. Pogosto je avtomatska pretvorba povzročala napake tako, da so bila določena ločila izbrisana (zlasti tri pike, gl. zgleda 2 in 5) in/ali premeščena na drugo mesto. Tovrstne napake je bilo treba odpraviti ročno. Pomembno je poudariti tudi to, da avtomatska pretvorba načeloma ni prizadela stave ločil in velikih začetnic (razen v zgoraj opisanih primerih, ko so bila ločila izbrisana), zato je bilo treba napačno rabljena ločila in velike začetnice (tj. napake transkriptorjev, ki so pripravljali pogovorne zapise) popravljati vzporedno v pogovornih in standardiziranih zapisih.

Osnovno načelo pri popravljanju avtomatskih pretvorb je bilo, da se segmentov govora v transkripcijah ne spreminja, segmenti v pogovornih in standardiziranih zapisih se morajo torej natančno ujemati. Prav tako pripravljavec standardiziranih zapisov načeloma ni spreminjal akustičnega ozadja in akustičnih dogodkov (smeh, odkašljanje, vdih, izdih, drugi zvoki), je pa lahko opozoril na napake pri njihovem označevanju v pogovornih zapisih. V pogovornih in standardiziranih je bilo dovoljeno uporabljati omejen nabor ločil: piko, vejico, klicaj, vprašaj, podpičje, opuščaj (kot del besede, pisan stično z besedo ali sredi besede), znak za *in* (*&*), narekovaje, dvopičje, tri pike, vezaj.[14] Ločila, uporabljena v določenem segmentu pogovornega zapisa, so morala biti uporabljena tudi v ustreznem segmentu standardiziranega zapisa. Oznake besednih fragmentov so v standardiziranem zapisu ohranjene takšne, kot se pojavijo v pogovornem zapisu (se jih torej ne spreminja). Prav tako se pri pripravi standardiziranih zapisov ne spreminjajo oznake anonimiziranih osebnih podatkov in oznaka za nerazumljiv govor (*neraz*). Tudi številke in datumi so načeloma ohranjeni tako, kot so bili zapisani v pogovornem zapisu, torej številke znotraj oglatih oklepajev in datumi znotraj zavitih oklepajev.

V Tabeli 3 so prikazani zgledi pogovornih in ustreznih standardiziranih zapisov iz vseh treh sklopov baze Artur, za katere so bili pripravljeni tako pogovorni kot standardizirani zapisi.

14 Pri vezaju je prišlo do odstopanja od pravopisne norme, ker program Transcriber 1.5.1 ne omogoča razlikovanja med vezajem in pomišljajem, zato je bil vezaj uporabljen tudi namesto pomišljaja.

Tabela 3: Primerjava pogovornih zapisov in ročno popravljenih standardiziranih zapisov (primeri iz javnega, nejavnega in parlamentarnega govora).

Sklop baze Artur	Pogovorni zapis	Ročno popravljeni stand. zapis
Javni govor	Nekoč je bil človk navaj@n na trpljenje in je z večjo lahkoto šel skozi življenje. Sodob@n človk veliko hitreje podleže težavam, ker nanje ni pripravlj@n.	Nekoč je bil človek navajen na trpljenje in je z večjo lahkoto šel skozi življenje. Sodoben človek veliko hitreje podleže težavam, ker nanje ni pripravljen.
Javni govor	Če gledamo starostno st@nd@rdizirano stopnjo, s k@tero mi primerjamo bremena med državami, je to okol [dvejs@t] na [sto tisoč]. S tako stopnjo smo mi začel v@ začetku [še@s@tih] let prejšnj@ga stoletja. In zakaj so p@ršli h n@m?	Če gledamo starostno standardizirano stopnjo, s katero mi primerjamo bremena med državami, je to okoli [dvajset] na [sto tisoč]. S tako stopnjo smo mi začeli v začetku [šestdesetih] let prejšnjega stoletja. In zakaj so prišli k nam?
Nejavni govor	#Eee, mogoče na konci pri@m na isti nivo sposobnosti oziroma še na večji nivo, zato k@r bon se mogo san navučiti uravnavate moč, #e, k@r mi ne dobena elektronika pomagala, al pa kupiš novejši motor,	#Eee, mogoče na koncu pridem na isti nivo sposobnosti oziroma še na večji nivo, zato ker bom se mogel sam naučiti uravnavati moč, #e, ker mi ne nobena elektronika pomagala, ali pa kupiš novejši motor,
Nejavni govor	Ge sem se včakala, ja, jaz s@m šla f penzijo. #Eee, s@m tudi v Muri napredovala potem, sledi s@m šla f kontorlo, s@m f kontroli delala, takrat se mi je to, tudi lepo bilo, ne. #Eee v@ ... Pri plači se mi je poznalo pa še ovačik, s@m z veseljem to delo delala pa opravljala ga, s@m kontrolirala izdelke.	Ge sem se včakala, ja, jaz sem šla v penzijo. #Eee, sem tudi v Muri napredovala potem, sledi sem šla v kontrolo, sem v kontroli delala, takrat se mi je to, tudi lepo bilo, ne. #Eee v ... Pri plači se mi je poznalo pa še ovačik, sem z veseljem to delo delala pa opravljala ga, sem kontrolirala izdelke.
Parlamentarni govor	torej u bistvu fsa kritika ki je bla danes usmerjena u predlok poslanske skupine SDS je na nek način usmerjena v DESUS @k tuki jz z DESUSA opozarjam da so u tej koaliciji torej to kr vi zagovarjate to kr vi zagovarjate kje() kar je eden vaših temelnih točk programa kot stranke u bistvu rezon detre DESUSA	torej v bistvu vsa kritika, ki je bila danes usmerjena v predlog poslanske skupine SDS, je na neki način usmerjena v Desus, ker tukaj jaz iz Desusa opozarjam, da so v tej koaliciji, torej to, kar vi zagovarjate, to, kar vi zagovarjate, kje() kar je ena vaših temeljnih točk programa kot stranke, v bistvu raison d'être Desusa,
Parlamentarni govor	Js bi reku seveda potem ko je bilo potrebno dobiti soglasje za poroštvo tm se je pa pol krepko upela politika pa da ne rečem konkretno tudi nas držauni zbor. Ampak poglejte tudi u takrat naprej ko je biu dejansko #eee poroštveni zakon	Jaz bi rekel, seveda, potem ko je bilo potrebno dobiti soglasje za poroštvo, tam se je pa pol krepko vpela politika, pa da ne rečem konkretno tudi nas, državni zbor. Ampak poglejte, tudi od takrat naprej, ko je bil dejansko, #eee, poroštveni zakon

V splošnem je mogoče reči, da je bil z vidika priprave standardiziranih zapisov pričakovano najmanj problematičen oz. zahteven javni govor (gl. prva dva primera v Tabeli 3). V njem se je namreč pojavljalo zelo malo narečnih in pogovornih besed (oz. besed, ki niso zajete v splošnih slovarjih slovenskega jezika). Večji izziv je predstavljal nejavni govor, v katerem smo identificirali večjo pojavnost narečnih besed, ki se praviloma uvrščajo med težavnejše primere standardizacije.[15] Parlamentarni govor ni bil problematičen v smislu zahtevnosti standardizacije, saj gre za govor v javnem formalnem govornem položaju, za katerega izrazito narečna leksika ni značilna. Parlamentarni govor je bil najzahtevnejši v smislu vloženega časa zaradi nihajoče kakovosti pogovornih zapisov (gl. razdelek 4).

Težavnejše primere standardizacije besed je pripravljavec standardiziranega zapisa sproti beležil in oblikoval predloge za standardizirani zapis (včasih po posvetu s kolegi dialektologi). Te je pregledala in potrdila skupina treh strokovno usposobljenih projektnih sodelavcev.

V spodnji Tabeli 4 so navedeni izbrani primeri s Seznama težavnejših primerov standardiziranega zapisa v bazi Artur (Verdonik, Trojar in Bizjak, 2023b: 14–24).

Tabela 4: Izbor primerov s Seznama težavnejših primerov standardiziranega zapisa v bazi Artur.

Primer	Predlog za standardizirani zapis
ajnfah, ajnfoh	ajnfah
bohlonaj, boglonaj, bohloni	boglonaj
cajt, cet	cajt
dugi (‘dolg’)	dugi
fancy, fensi	fensi
gniliti (‘gniti’, narečno)	gniliti
jajčka (‘jajce’), ž. sp.	jajčka

15 Pri težavnejših primerih se je pripravljavec standardiziranega zapisa posvetoval s kolegi dialektologi. Prim. npr. besedo ***ovačik*** v 2. primeru nejavnega govora v Tabeli 3, ki se ne pojavi v nobenem slovarju na portalu Fran (se pa v Pleteršnikovem slovarju pojavita besedi ***ovače*** in ***ovači***, slednja pa je zabeležena tudi v Slovarju stare knjižne prekmurščine).

Primer	Predlog za standardizirani zapis
kejpop	K-pop
mezmes (‘vmes’)	mezmes
obično	obično
parajt, berajt (‘pripravljen’)	berajt
ušeta (‘ušesa’)	ušeta
vjutro	vjutro

V grobem je mogoče težavnejše primere standardiziranega zapisa razdeliti v tri skupine, in sicer na narečne besede (npr. *mezmes*), pogovorne besede, ki niso vezane na eno narečje ali manjšo skupino narečij (npr. pokrajinskopogovorne in tudi splošnoslovenske pogovorne besede, npr. *ajnfah*, *cajt*), in prevzete besede (npr. *K-pop*).

4 Težave pri izdelavi pogovornih in standardiziranih zapisov, rešitve zanje in priporočila za prihodnje projekte

Najprej velja opozoriti, da je govorna baza Artur nastajala v času pandemije covida-19, ko so bili medosebni stiki in javni dogodki močno omejeni ali celo prepovedani. Omejitev gibanja znotraj občinskih meja je pomenila dodatno oviro pri snemanju in vključevanju govorcev iz različnih regij, na samo kakovost govora pa je vplivalo tudi nošenje obraznih mask. Količino takšnih posnetkov smo zato močno omejili, metapodatek o nošenju mask pa sproti beležili.

Pri izdelavi tako obsežne govorne baze je ključno, da koordiniranje aktivnosti poteka ažurno, da je komuniciranje med deležniki periodično in da se sproti iščejo rešitve za morebitne probleme. S tem se optimizira čas v zaključnih fazah priprave govorne baze, ko se ponovno izvedejo avtomatske validacije podatkov in na njihovi osnovi časovno zelo potratni ročni popravki identificiranih napak.

Pri pripravi pogovornih zapisov za govorno bazo Artur je bilo zelo problematično pogosto menjavanje transkriptorjev in njihovo časovno zamudno uvajanje. Vsak izmed njih je namreč tvoril drug tip napak, kar je bilo treba vsakič znova identificirati. Iskanje in uvajanje

zanesljivega transkriptorja tako ostaja ena od zahtevnejših nalog, ki jo je zaradi neželene fluktuacije sodelavcev pri zapisovanju govora treba večkrat ponoviti.

V nadaljevanju so navedene in opisane nekatere najpogostejše napake, identificirane pri pripravi pogovornih zapisov. Zaradi hitrega in močno strnjenega govora posameznih moderatorjev so meje segmentov transkriptorji postavili preblizu predhodnih fonemov ali tistih, ki so jim sledili. Ko je moderator govoril neprekinjeno več kot 10 sekund, je bilo treba mejo postaviti v tistem delu signala, ko je zajel sapo. Posamezne posnetke radijskega govora smo prejeli že obrezane in v njih ob menjavi govorcev ni bilo vidnih premorov, kar je dodatno otežilo proces segmentiranja. Zaradi prilagoditve segmentacije tehničnim zahtevam za razvoj razpoznavalnika transkriptorji meja segmentov pogosto niso postavili glede na semantično-skladenjski vidik, temveč glede na premore kot prozodično značilnost.

Segmenti pri hkratnem govoru so bili včasih predolgi, zapisi pa nenatančni. V tem smislu smo zaznali nedosleden zapis opornih signalov, kot sta *ja* in *mhm,* saj je njihov natančen zapis pri hkratnem govoru časovno zelo zamuden. Precej popravkov je bilo potrebnih tudi zaradi napačnih označevanj menjav govorcev.

Mestoma zvočna ozadja in zvočni dogodki niso bili označeni ali pa so bili neustrezno označeni zgolj kot deli segmenta in ne segmenti kot celota. Daljši premori so bili pogosto izpuščeni.

Posamezne besede, ki so bile izgovorjene, niso bile zapisane ali pa so bile zapisane napačno. Gre za tip napake, ki ga je izjemno težko odkriti z avtomatskimi preverjanji in ki zahteva veliko dodatnega časa za ročno preverjanje. Če želimo v prihodnje doseči čim višjo kakovost zapisa izrazito narečnega govora, kar zahteva dodatno natančnost pri poslušanju, ga mora zapisati ustrezno usposobljen dialektolog.

Ugotovili smo, da je v tako obsežni bazi, kot je Artur, dosledno zapisovanje polglasnikov (z znakom @) skoraj nemogoče doseči, poleg tega je precej primerov, pri katerih različni transkriptorji različno interpretirajo slišani glas. Podobno je pri fonemu *v*, katerega zapis

je lahko nedosleden, npr. *vprašal* namesto *fprašal*. Dodatno smo v pogovornih zapisih zaznali rabo neustreznih črk, ki niso del slovenskega črkopisa, npr. *q* in *y*.

Izkazalo se je, da je težko ohranjati doslednost pri zapisu neverbalnih in polverbalnih glasov. Pojavili so se neenotni zapisi z eno, dvema ali tremi črkami; včasih so transkriptorji pred ali za njimi vstavili vejico, drugič ne; pogosto na začetku segmenta niso zapisali velike začetnice ali pa so pri označevanju izpustili znak #.

Zaradi prevelikega števila napak pri vstavljanju oklepajev so bili ti pred javno objavo govorne baze iz nje odstranjeni. V oglatih oklepajih so se poleg številnikov pojavljali tudi samostalniki (npr. *tretjina*), pridevniki (npr. *drugi ljudje*) in nedoločni členi (npr. *en lep dan*). Mestoma so bili datumi namesto v zavitih oklepajih zapisani v oglatih.

Zaradi pomanjkljive jezikoslovne izobrazbe transkriptorjev in njihove pogoste fluktuacije so se v zapisih pojavljale različne pravopisne napake, zlasti pri veliki začetnici, zapisu skupaj ali narazen in ločilih, najpogosteje pri vejici in vezaju. Kot posebno problematičen se je tu izkazal parlamentarni govor: na osnovi transkripcij, ki jih izdelujejo v Državnem zboru RS, so pogovorne zapise namreč pripravljali (popravljali) študentje nejezikoslovnih smeri. To je praviloma vodilo do zelo velikega števila napak v pogovornih zapisih (gl. zgleda iz parlamentarnega govora v Tabeli 3, v katerih manjka večina ločil). Slednje pomeni izjemno povečano obremenitev za pripravljavca standardiziranega zapisa, ker mora večino popravkov vnašati dvakrat, tj. hkrati v standardizirani in pogovorni zapis. Težava je bila razrešena tako, da so bila ločila in velike začetnice pri parlamentarnem govoru naknadno avtomatsko (s posebnim skriptom) prenesena iz standardiziranih v pogovorne zapise. Ključno je spoznanje, da je nadzor nad kakovostjo pogovornih zapisov bistvenega pomena; če so namreč pogovorni zapisi kakovostni, pomeni to precej manj oz. hitrejše delo za pripravljavca standardiziranega zapisa.[16] Ker se z vstavljanjem ločil govor dodatno interpretira, je neizbežno, da ločila

16 Pogovorni zapisi, ki so jih pripravljali zunanji izvajalci (podjetja), so se praviloma izkazali za precej kakovostnejše od tistih, ki so jih pripravljali študentje. Za prihodnje projekte se zato priporoča najemanje zunanjih izvajalcev.

skladno s pravopisno normo vstavljajo ustrezno usposobljeni strokovnjaki z jezikoslovno izobrazbo.

Med težavami pri pripravi pogovornih zapisov za bazo Artur je bil tudi zapis dialogov, ki so bili pri nejavnem govoru z namenom zagotavljanja višje kakovosti zvoka posneti 2-kanalno preko dveh mikrofonov. Isti dialog je bilo tako treba zapisati dvakrat, vsakič za drugega govorca, kar pomeni višje finančne stroške. Nastopi pa lahko še dodatna težava, ko se pri hkratnem govoru v ozadju sliši govor drugega govorca.

Dodatni časovno zahteven izziv je bil velik obseg zelo kratkih posnetkov, trajajočih tudi manj kot eno minuto, in zapisov zanje, ki jih je bilo treba vsakič znova prenesti, preimenovati, obdelati in shraniti.

Po pripravi standardiziranih zapisov je bilo sprva načrtovano preverjanje konsistentnosti popravkov (npr. konsistentnosti uporabe izbranih rešitev na Seznamu težavnejših primerov standardiziranega zapisa v bazi Artur, gl. Tabelo 4). Za tovrstno preverjanje in popravljanje napak je zmanjkalo časa, bi se pa mu bilo smiselno posvetiti v prihodnje, saj v standardiziranih zapisih prihaja do nedoslednosti.

Pri izdelavi baze je bil za transkribiranje uporabljen program Transcriber 1.5.1. Njegova odlika je izjemna stabilnost, je pa rokovanje z njim časovno zelo zamudno. Gre za to, da je treba pogovorni zapis in ustrezni standardizirani zapis odpreti vsakega v svojem oknu, nato pa še posebej odpreti zvočno datoteko s posnetkom govora ter ročno vsakič posebej nastaviti kodiranje (na UTF-8) in morebitne dodatne nastavitve.[17] To se je v projektu RSDO izkazalo za večjo pomanjkljivost. Pri pripravi standardiziranega zapisa je bilo namreč pregledanih 2871 parov datotek parlamentarnega govora,

17 Pri programu Transcriber 1.5.1 je zelo moteče in zamudno tudi to, da program pri uporabi smernih tipk na tipkovnici ne omogoča prehajanja kurzorja s sredine izbranega segmenta na sredino segmenta, ki leži tik nad ali tik pod njim. Pri uporabi smernih tipk ↑ in ↓ je prehod med segmenti namreč mogoč le na konec višje ali nižje ležečega segmenta, pri uporabi smernih tipk ← in → pa mora kurzor prepotovati celotno pot do začetka ali konca trenutno izbranega segmenta in šele nato do želenega mesta sredi višje/nižje ležečega segmenta. V praksi to največkrat pomeni veliko zamudnega klikanja z miško za postavljanje kurzorja na ustrezno mesto oz. za premikanje med segmenti. Delo s Transcriberjem je zamudno tudi zato, ker morata biti pogovorni in ustrezni standardizirani zapis odprta vsak v svojem oknu (ker gre za dve ločeni datoteki), pri čemer je treba vsakemu segmentu v pogovornem zapisu ročno poiskati ustrezni vzporedni segment v standardiziranem zapisu.

100 parov datotek javnega govora in 375 parov nejavnega govora (skupno 3346 parov datotek oz. 6692 datotek s standardiziranimi in pogovornimi zapisi). Ob predpostavki, da priprava delovnega okolja za 1 par datotek vzame 2 minuti (tj. odpiranje Transcriberja, odpiranje pogovornega in standardiziranega zapisa, odpiranje avdio datoteke, nastavljanje ustreznih nastavitev in vpisovanje podatkov v evidenco, ker se evidenca o opravljenem delu ni vodila avtomatsko), je bilo zgolj za pripravo na delo (!) potrebnih 111,53 ure ali 14,87 delovnega dne (po 7,5 ure). Ta čas ne vključuje dejanskega popravljanja transkripcij, poleg tega sem ni vključeno še naknadno popravljanje datotek (npr. po pregledu ujemanja v številu pojavnic).

5 Zaključek

Na osnovi izkušenj pri pripravi govorne baze Artur za prihodnje projekte priporočamo prehod na orodje OrthoNormal[18] ali kako drugo primerljivo orodje za transkribiranje, ki omogoča več fleksibilnosti pri delu (podpora za transkribiranje velikega števila kratkih posnetkov, hitro in avtomatsko prehajanje med njimi, avtomatski vzporedni prikaz segmentov v pogovornem in standardiziranem zapisu itd.). Idealno bi bilo vzpostaviti lastno spletno okolje z vgrajenimi rešitvami za obdelavo zvočnih posnetkov, ročno transkribiranje, avtomatsko razpoznavanje govora (avtomatsko generirane transkripcije), projektno vodenje in s črkovalnikom (ki bi preverjal konsistentnost popravkov) ipd. Takšno spletno okolje bi na enem mestu podpiralo izdelavo govornih baz, vanj pa bi integrirali tudi rezultate projekta RSDO: avtomatski razpoznavalnik bi transkriptorju predpripravil transkripcije (npr. pogovorne in/ali standardizirane zapise), ki bi bile že ustrezno segmentirane, transkriptor pa bi jih po potrebi le popravil v skladu s strokovnimi oz. projektnimi zahtevami. Okolje bi samodejno beležilo spremembe in vodilo evidence o opravljenem delu, kar bi bistveno povečalo informiranost vodje projekta in sodelavcem olajšalo vodenje evidenc v projektu. Zasnova in realizacija tovrstnega spletnega okolja bi seveda

18 https://exmaralda.org/de/orthonormal-de/

zahtevali stabilno financiranje, ki bi segalo onkraj sporadičnih triletnih projektov. Šele tako široko zastavljeno dolgoročno zbiranje posnetkov govora bi omogočilo tudi kakovostnejše slovnične in leksikalne opise govorjenega slovenskega jezika, ki je bil v dosedanjih raziskavah izrazito prešibko zastopan.

Projekt RSDO je kot primer dobre prakse omogočil spoznanje in utemeljil zavedanje o tem, da je v slovenskem prostoru na področju procesiranja govora in jezikovnih tehnologij nujno tudi interdisciplinarno povezovanje čim več institucij tako iz akademsko-raziskovalnega okolja kot tudi iz gospodarstva.

Literatura

Verdonik, D., & Bizjak, A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. https://dk.um.si/Dokument.php?lang=slv&id=170009&dn

Verdonik, D., Trojar, M., & Bizjak, A. (2023a). Prednosti in slabosti dvotirnega zapisovanja govora v slovenskih govornih virih = Advantages and Disadvantages of Two-level Speech Transcription in the Slovenian Speech Resources. *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov*, 111-114. https://press.um.si/index.php/ump/catalog/book/774

Verdonik, D., Trojar, M., & Bizjak, A. (2023b). *Standardizirani zapis v govorni bazi Artur projekta RSDO*. Univerza v Mariboru. https://dk.um.si/Dokument.php?id=170007&lang=slv

Verdonik, D., Bizjak, A., & Dobrišek, S. (2023). *Opis govorne baze Artur projekta RSDO*. Univerza v Mariboru. https://dk.um.si/IzpisGradiva.php?id=85199

Verdonik, D., Potočnik, T., Sepesy Maučec, M., Erjavec, T., Majhenič, S., & Žgank, A. (2021). *Spoken corpus Gos VideoLectures 4.2 (transcription).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1444

Verdonik, D., & Zwitter Vitez, A. (2011). *Slovenski govorni korpus Gos*. Trojina, zavod za uporabno slovenistiko.

Žganec Gros, J., & Vesnicer, B., (2021). Izbor fonetično uravnoteženih besedilnih predlog za bazo branega govora. V T. Mirtič in M. Snoj (ur.), *1. slovenski pravorečni posvet* (pp. 111–119). Slovenska akademija

znanosti in umetnosti. https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Žganec Gros, J., Vesnicer, B., Mihelič, A., Trojar, M., Dobrišek, S., Bizjak, A., & Verdonik, D. (2023). Izbor povedi za govorno bazo Artur v projektu Razvoj slovenščine v digitalnem okolju. Projektno poročilo DS2-2.1.1. https://dk.um.si/IzpisGradiva.php?id=85200