SMERNICE ZA ZBIRANJE PODATKOV ZA GOVORNE VIRE

Darinka Verdonik, Januška Gostenčnik

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Marec 2024

VSEBINA

1 UVOD

V projektu Temeljne raziskave za razvoj govornih virov in tehnologij (MEZZANINE, J7-4642), ki ga je financirala Javna agencija za raziskovalno in inovativno dejavnost Republike Slovenije od oktobra 2022 do septembra 2025, je prva aktivnost prvega delovnega sklopa vključevala raziskave potreb po govornih virih v humanističnih in tehničnih znanostih. Na to temo je bila v povezavi s projektom izvedena 6. konferenca Slavistični znanstveni premisleki z naslovom Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah, ki je potekala 18. in 19. maja 2023 v Mariboru (Kranjc Ivič v tisku) ter objavljena prispevka (Verdonik 2023; Trojar, Bizjak 2023). Smernice temeljijo na upoštevanju navedenih objav ter izkušnjah in analizi govornih virov Artur (Verdonik et al. 2023a) in Gos 2.0 (Verdonik et al. 2023b).

Podatki o gradivih, govorcih in geolokaciji ter uravnoteženost gradiv glede na te podatke so odvisni od vsakokratnega namena zbiranja govornih gradiv. V teh smernicah se osredotočamo na namen izdelave osrednjega referenčnega govornega vira, kot je za slovenščino korpus Gos (Verdonik et al. 2023b). Zaradi združljivosti virov in doseganja sinergije pri njihovem razvoju pa je priporočljivo, da so tudi v specializiranih virih podatki kolikor mogoče skladni s temi.

2 ZBIRANJE GRADIV GLEDE NA VRSTO KOMUNIKACIJE

Ločevati je treba kriterije za zajem gradiv, ki pomenijo izhodiščni načrt, katere vrste gradiv bomo zbrali v nekem govornem viru, in metapodatke o zbranih gradivih, ki pomenijo čim bolj natančen popis podatkov o gradivih in govorcih na njih. V teh smernicah oboje združimo v iste kategorije in kot kriterije za zajem gradiv samo izpostavimo določene kategorije.

Pri kriterijih za zajem gradiv priporočamo čim večjo uravnoteženost glede na:

- javnost,
- formalnost,
- interaktivnost,
- kanal,
- področje,
- žanr,
- čas,[1]
- kraj[2] in
- vir.[3]

Tabela 1 podaja podroben seznam kategorij in oznak po kategorijah. Zlasti za specializirane govorne vire, pa tudi za referenčni govorni vir, je treba seznam kritično preverjati in po potrebi prilagoditi in/ali razširiti glede na značilnosti podatkov in namen govornega vira.

Tabela 1: Seznam kategorij in oznak o posnetkih v govornih virih

OPIS	OZNAKE
Snemalec*	Ime, priimek
ID posnetka	
Vir	*Navedba gradivodajalca, npr. RTV Slovenija, Državni zbor RS, Arnes, Videolectures.net, TEDx, lastni posnetek*
URL	*Spletni naslov, če je posnetek dostopen prek spleta.*

[1] Kontinuirano zbiranje gradiv skozi leta.

[2] Ustrezna pokritost vseh regij in narečij, razmerje med urbanim in ruralnim okoljem, slovenščina v zamejstvu in po svetu.

[3] Čim več različnih virov.

Javnost komunikacije	1. javno (za širšo javnost) 2. skupinsko (komunikacija v večji skupini ljudi, znanih poimensko, kot npr. pouk, predavanja na fakultetah, seminarji, delavnice, tečaji, delovni sestanki...) 3. zasebno (komunikacija z enim, dvema, tremi ali štirimi govorci)
Formalnost komunikacije	1. formalno 2. delno formalno (ni povsem formalno niti povsem neformalno) 1. neformalno
Interaktivnost komunikacije	*1.* brez naslovnika *(govor, prebran s pisne predloge, za namene govornega vira)* 2. monolog *(predavanja, javni nastopi, za namene snemanja spodbujen govor enega govorca ob prisotnosti intervjuvarja, ki postavlja iztočnice, ipd.)* 3. dialog *(dva sogovornika, ki se interaktivno izmenjujeta; pogovor med dvema, intervju v medijih ipd.)* 2. multilog *(trije ali več sogovornikov, ki se interaktivno izmenjujejo; pogovor)*
Kanal komunikacije	1. osebna prisotnost *(zasebna komunikacija v živo, skupinska in javna komunikacija, namenjena občinstvu, prisotnemu v živo ob dogodku)* 2. video oddajanje *(video posnetek, namenjen televizijskemu ali internetnemu občinstvu)* *3.* avdio oddajanje *(avdio posnetek, namenjen radijskemu ali internetnemu občinstvu)* 4. video prenos *(nejavna komunikacija med dvema ali več osebami prek telefona ali interneta z video prenosom)* 5. avdio prenos *(nejavna komunikacija med dvema ali več osebami prek telefona ali interneta brez video prenosa)*
Področje družbenega življenja	1. aktualni dogodki *(npr. informativne medijske vsebine, aktualno dogajanje)* *2.* razvedrilo *(npr. razvedrilne medijske vsebine)* 3. izobraževanje *(šolanje, pouk, govorilne ure ipd.)* 4. znanost in tehnologija *(npr. predstavitve in komentarji razvoja v znanosti in tehnologijah)* 5. kultura in umetnost *(kulturni dogodki, oddaje o kulturi ipd.)* 6. zdravje *(npr. javne diskusije na temo zdravja, bolezni, zdravil...)* 7. finance in premoženje *(urejanje finančnih poslov, zavarovanj ipd.)* 8. delo *(govor v delovnem okolju o delu, npr. delovni sestanki)* 9. politika *(govor politikov v parlamentu, javnosti ipd.)* 10. prodaja in storitve *(govor ob prodaji, opravljanju storitev ipd.)* 11. šport *(športni prenosi, športni komentarji ipd.)*

	12. prosti čas *(vsakdanji pogovori, intervjuvani govor o vsakdanjih stvareh ipd.)* 13. splošno *(samo za žanr branja za govorno bazo)* *Ni zaprt seznam.*
Žanr	1. reportaža 2. intervju 3. polemika 4. okrogla miza 5. pogovor za javnost 6. resničnostni šov 7. kuharska oddaja 8. kviz 9. športni prenos 10. moderirani program 11. novinarska konferenca 12. strokovna konferenca 13. seminar/delavnica 14. šolski pouk 15. predavanje 16. delovni sestanek 17. govorilna ura 18. prodajna predstavitev 19. vsakdanji govor 20. seja državnega zbora 21. branje za govorno bazo *Ni zaprt seznam.*
Opis govornega dogodka	*Prosti opis situacije, ki je posneta, v stavku, dveh ali več.*
Čas dogodka	(DD).(MM).LLLL *Kdaj je potekal dogodek; navedeno najmanj z letnico.*
Kraj dogodka	Seznam krajev
Občina dogodka	Seznam občin
Statistična regija dogodka	Seznam regij

Država dogodka	Seznam držav
Vrsta snemalne naprave	1. pametni telefon 2. računalnik z zunanjim mikrofonom 3. računalnik z vgrajenim mikrofonom 4. kamera 5. diktafon 6. prenosni snemalnik, npr. Zoom 7. profesionalna snemalna oprema *Ni zaprt seznam.*
Digitalizacija	1. da *(posnetek je bil digitaliziran)* 2. ne *(posnetek je bil že posnet na digitalni medij)*
Izvorni format posnetka	1. WAV 2. M4A 3. MP3 *Ni zaprt seznam.*
Izvorna frekvenca vzorčenja (kHz)**	1. 32 kHz 2. 44.1 kHz 3. 48 kHz *Ni zaprt seznam.*
Izvorna bitna ločljivost (bit)**	1. 16 bit 2. 32 bit *Ni zaprt seznam.*
Izvorna bitna hitrost (kbps)**	*Vnese se številka, običajno med 110 in 1500.*
Dolžina posnetka	*Na način h:mm:ss.*
Prostor snemanja	1. manjši prostor *(npr. soba v stanovanju, manjša pisarna)* 2. srednje velik prostor *(npr. velik dnevni prostor, velika pisarna, sejna soba, seminarska soba)* 3. velik prostor *(npr. dvorana, predavalnica, avla)* 4. studio *(profesionalno opremljen prostor za snemanje)* 5. odprti prostor *(zunaj zgradb)*
Slušna ocena kvalitete posnetka	• 1 *(nezadostna kvaliteta)* • 2 *(zadostna kvaliteta)* • 3 *(dobra kvaliteta)* • 4 *(zelo dobra kvaliteta)* • 5 *(odlična kvaliteta)*

* Nejavni podatek. ** Neobvezno.

3 ZBIRANJE GRADIV GLEDE NA GOVORCE

Enako kot v poglavju 2 tudi tukaj ločujemo kriterije za zajem in metapodatke o zbranih gradivih združimo v iste kategorije in kot kriterije za zajem gradiv samo izpostavimo določene kategorije.

Pri kriterijih za zajem gradiv priporočamo čim večjo uravnoteženost glede na naslednje lastnosti govorcev:

- spol,
- starost,[4]
- izobrazba,
- regija in
- prvi jezik.[5] [6]

Tabela 2 popisuje seznam priporočljivih kategorij in oznak o govorcih. Zlasti za specializirane govorne vire, pa tudi za referenčni govorni vir, je treba seznam kritično preverjati in po potrebi prilagoditi in/ali razširiti glede na značilnosti podatkov in namen govornega vira.

Tabela 2: Seznam kategorij in oznak o govorcih v govornih virih

OPIS	OZNAKE
ID posnetka	
ID govorca	
Ime in priimek govorca*	
Pojavitev na posnetku	1. prvi zaporedni govorec na posnetku 2. drugi zaporedni govorec na posnetku 3. tretji zaporedni govorec na posnetku

[4] V splošne govorne vire uvrščajo praviloma polnoletni govorci, za mladoletne govorce se običajno izdelajo specializirani viri.

[5] Možnosti so, da je slovenščina prvi jezik, da gre za dvojezičnost slovenščine in še enega tujega jezika od otroštva ali da slovenščina ni prvi jezik, kjer so možne delitve naprej glede na bližino jezikovne skupine prvega jezika s slovenščino.

[6] Po podatkih Statističnega urada Republike Slovenije je leta 2022 v Sloveniji prebivalo 8,5 % tujih državljanov, največ v obalno-kraški regiji in večjih mestnih središčih.

Spol	1. moški 2. ženski 3. neopredeljeno
Starost	*xx let*
Izobrazba	1. osnovna šola ali manj 2. srednja šola 3. višja ali visoka šola 4. fakulteta ali več 5. ni podatka
Prvi jezik	npr. slovenščina, nemščina, hrvaščina itd.; ni podatka
Dvojezičnost**	1. dvojezični govorec; ni podatka
Dodaten prvi jezik**	npr. nemščina, hrvaščina itd.; ni podatka
Bivanje v otroštvu – kraj**	seznam krajev; ni podatka
Bivanje v otroštvu – občina**	seznam občin; ni podatka
Bivanje v otroštvu – država**	seznam držav; ni podatka
Bivanje v sedanjosti – kraj	seznam krajev; ni podatka
Bivanje v sedanjosti – občina	seznam občin; ni podatka
Bivanje v sedanjosti – država	seznam držav; ni podatka
Dodatno daljše bivanje drugod**	1. daljše bivanje drugod
Daljše bivanje drugod – kraj**	seznam krajev; ni podatka
Daljše bivanje drugod – občina**	seznam občin; ni podatka

Daljše bivanje drugod – država**	seznam držav; ni podatka
Zvrstne značilnosti govora	1. standardno (v celoti ali prevladujoče skladno s pravorečno normo) 2. pogovorno ali narečno 3. mešano (pogosto mešanje prvin standardnega in pogovornega/narečnega) 4. tujejezični vplivi (pogosto mešanje prvin slovenskega in katerega tujega jezika)
Narečje	1. ni narečje 2. ni podatka 3. seznam narečij (gl. 4)
Ohranjenost narečja	1. delno 2. srednje 3. dobro 4. ni podatka
Spontanost govora	1. brano (govor je bran z ekrana ali papirja) 2. nespontano (govor je govorjen na pamet, vendar je vidno, da je pripravljen in delno naučen) 3. delno spontano (govor je bil vsebinsko najverjetneje vsaj delno pripravljen vnaprej, npr. z opornimi točkami, vsebinskimi iztočnicami, tvorjen pa je v trenutku govorjenja) 4. spontano (govor ni pripravljen vnaprej, govorec se odloča, kaj povedati, v trenutku govorjenja)
Izgovorne posebnosti**	1. jecljanje 2. izgovor r 3. izgovor l *Ni zaprt seznam.*

* Nejavni podatek. ** Samo če je relevantno za posameznega govorca.

4 GEOLOKACIJSKA URAVNOTEŽENOST GRADIV

Govorjeni jezik geolokacijsko zelo variira, zato je pri govornih podatkih izredno pomembna ustrezna geolokacijska uravnoteženost. Pri tem sta možna pristopa korpusnojezikoslovni in geolingvistični.

4.1 Korpusnojezikoslovni pristop

Za korpusnojezikoslovni pristop dosedanjo prakso povzame citat iz Verdonik et al. (2022):

»Opredelitev regionalnih vplivov na govor govorca ni nujno enoznačna. Tako so se na primer v dodatku h govornemu delu korpusa BNC (British National Corpus) iz leta 2014, v katerem so zajemali samo vsakdanje pogovore, prepustili govorcem, da so sami s svojimi besedami opisali svoj dialekt, in nato te opise preslikali v shemo statističnih teritorijskih enot Velike Britanije (Love et al., 2017). Tudi v slovenskih govornih virih se je uveljavila praksa, da se regija govorcev beleži skozi geopolitične, in ne geolingvistične kategorije. Razlog je bržkone ta, da lahko zanesljive geolingvistične kategorizacije naredi samo stroka, in to naknadno, na podlagi zbranih podatkov. V korpusu GOS so bile tako kategorije za regijo govorcev definirane na podlagi registrskih območij, ki jih je za Slovenijo skupno 11, k temu pa so bile dodane še kategorije za zamejske Slovence (Avstrija, Italija, Madžarska) in govorce, ki jim slovenščina ni prvi jezik (tujina). Taka razdelitev je izredno ohlapna in nenatančna v primerjavi s slovensko dialektalno razpršenostjo. Tudi sam koncept 'regionalna pripadnost', zveden na registrsko označbo na avtomobilu, se zdi neustrezen, čeprav ima za teren zelo koristno lastnost robustnosti. V bazi Artur se je zato iskala bolj natančna, enoznačna, enostavna in manj sporna opredelitev metapodatka, ki bi nosil informacije o regiji govorcev. Ker smo ime kraja, zlasti ko gre za podeželsko okolje, že izpostavili kot problematično zaradi potencialnega razkrivanja identitete govorca, je bila kot osnovna enota izbrana občina. Slovenija je v času zbiranja posnetkov za bazo Artur razdeljena na 212 občin. Prednost te kategorije je tudi ta, da je mogoče občine enostavno enoznačno preslikati na širše geopolitične enote – 12 statističnih regij Slovenije, kot jih v času nastajanja baze definira Statistični urad Republike Slovenije.«

En možni vidik uravnoteženja govornih podatkov glede na geolokacijo je tako po statističnih regijah Slovenije. Leta 2022 so po podatkih Statističnega urada Republike Slovenije (SURS)[7] v Sloveniji veljale statistične regije, predstavljene v tabeli 3.

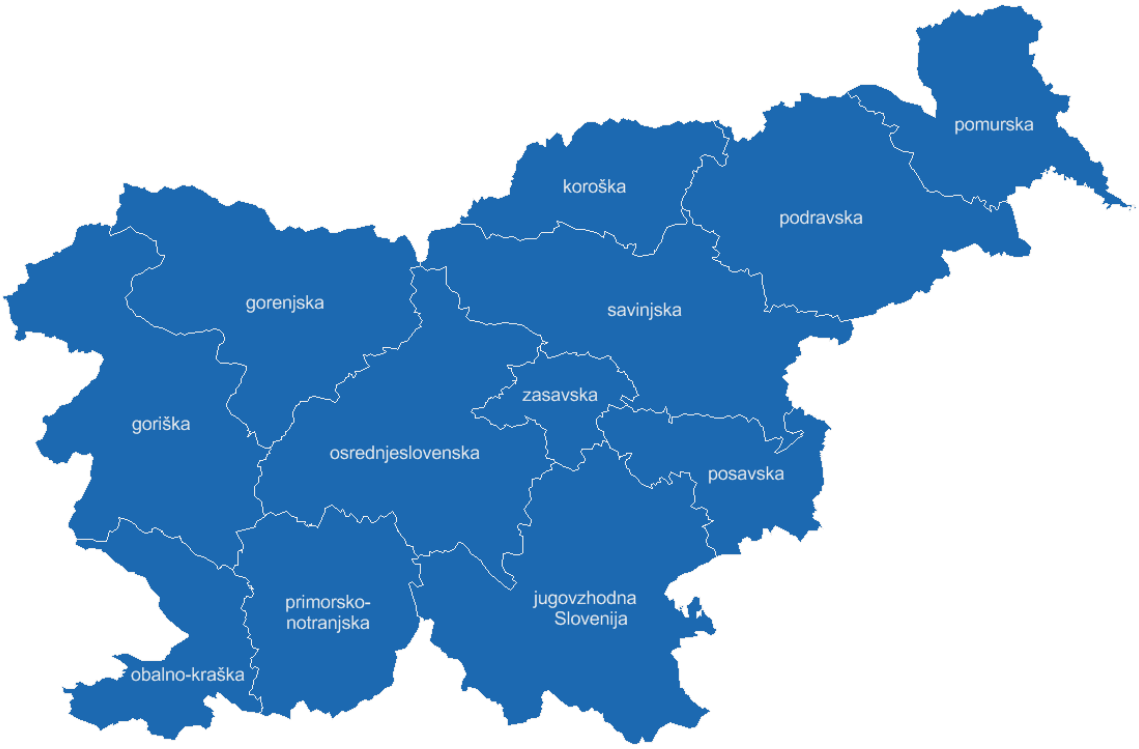
[7] https://www.stat.si/obcine

Tabela 3: Statistične regije Slovenije v letu 2022 in število prebivalcev v njih

Statistična regija	Število prebivalcev	Število prebivalcev v % glede na celotno Slovenijo
Osrednjeslovenska	556.862	26 %
Gorenjska	210.747	10 %
Goriška	118.202	6 %
Obalnokraška	118.426	6 %
Primorsko-notranjska	53.400	3 %
Jugovzhodna Slovenija	146.429	7 %
Zasavska	56.942	3 %
Posavska	75.749	4 %
Savinjska	259.306	12 %
Koroška	70.648	3 %
Podravska	327.858	16 %
Pomurska	114.163	5 %
Skupaj	2,108.732	100 %

Seznam občin po statističnih regijah je na voljo na spletnih straneh SURS.

Slika 1: Statistične regije Slovenije (vir: Wikipedija)



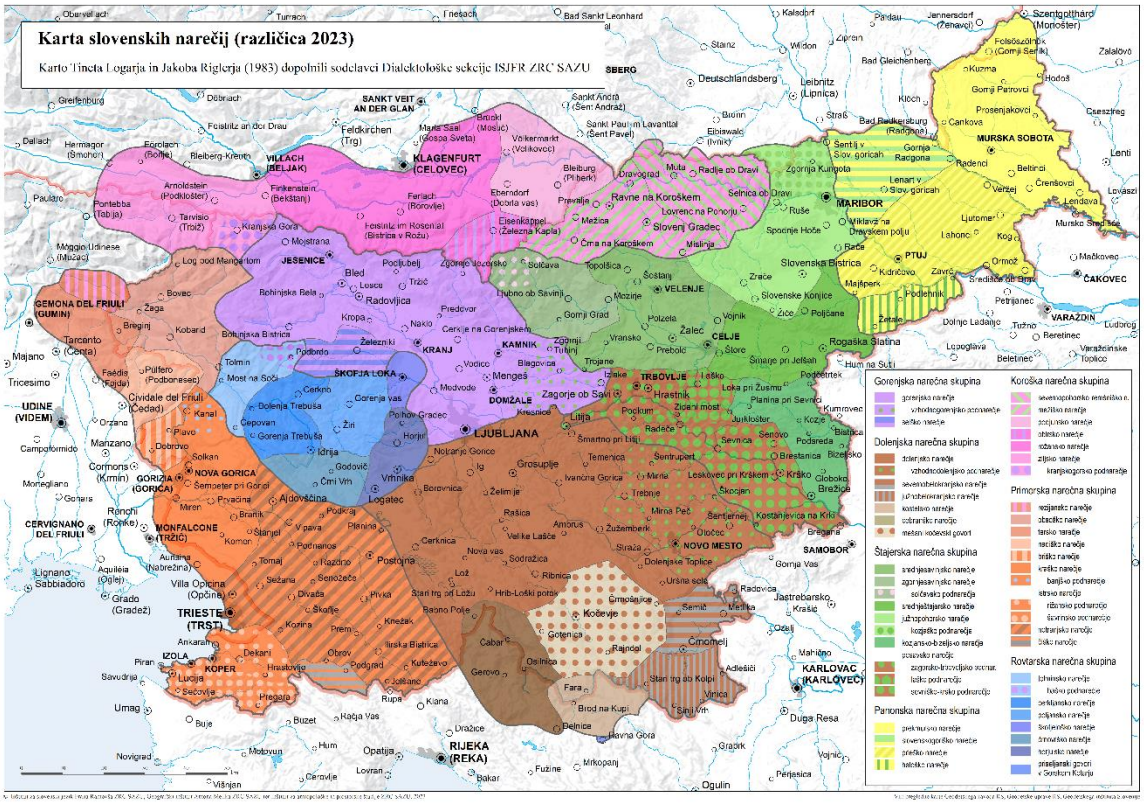
Poleg statističnih regij v Republiki Sloveniji je za celostno pokritost različic govorjene slovenščine glede na geolokacijo smiselno upoštevati tudi:

- govor Slovencev v zamejstvu na Madžarskem, v Avstriji, Italiji in na Hrvaškem,
- govor Slovenskih izseljencev.

4.2 Dialektološki pristop

Dialektološki pristop izhaja iz osnovne dialektološke delitve slovenščine na narečne skupine in narečja. Slovenščina se sinhrono deli na sedem narečnih skupin, znotraj tega na 38 narečij (z mešanimi kočevskimi govori) z 12 podnarečji.

Slika 2: Karta slovenskih narečij (različica 2023) (Vir: SLA 3.1)



Za uravnoteženost govornih podatkov je smiselno obravnavati po vsaj en govor (ali dva govora) iz posameznega narečja in podnarečja. Priporočeni seznam krajev za pridobivanje posnetkov:

Gorenjska narečna skupina: gorenjsko narečje – Zgornje Gorje, Cerklje na Gorenjskem, vzhodnogorenjsko podnarečje – Krašnja, Srednja vas v Tuhinju, selško narečje – Železniki, Dražgoše.

Dolenjska narečna skupina: dolenjsko narečje Ribnica, Borovnica, vzhodnodolenjsko podnarečje – Šentrupert, Mokronog,severnobelokranjsko narečje – Semič, Rožanec; južnobelokranjsko narečje – Stari trg ob Kolpi, Adlešiči, kostelsko narečje – Vas, Delač, Guče Selo (Hrv.), čebranško narečje – Babno Polje, Novi Kot.

Štajerska narečja skupina: posavsko narečje (zagorsko-trboveljsko podnarečje – Podkum, sevniško-krško podnarečje – Brestanica, laško podnarečje - Lahomšek); srednjesavinjsko narečje – Ložnica, Motnik, srednještajersko narečje – Šmarje pri Jelšah, Slivnica pri Celju, kozjansko-bizeljsko narečje – Bistrica ob Sotli, Kapele; zgornjesavinjsko narečje – Spodnje Kraše, Lenart pri Gornjem Gradu, solčavsko podnarečje – Solčava, Podolševa, južnopohorsko narečje – Oplotnica, Zreče, kozjaško podnarečje – Selnica, Spodnje Vrtiče.

Panonska narečna skupina: prekmursko narečje – Gomilica, Gornji Senik (Felsőszölnök), slovenskogoriško narečje – Črešnjevci, Voličina, haloško narečje – Gradišče, Žetale, prleško narečje – Lahonci, Štrigova.

Koroška narečna skupina: ziljsko narečje – Potoče (Potschach), Podklošter (Arnoldstein), kranjskogorsko podnarečje – Kranjska Gora, rožansko narečje – Breznica (Friessnitz), Borovlje (Ferlach), obirsko narečje – Obirsko (Ebriach), Železna Kapla (Eisenkappel), podjunsko narečje – Kneža (Grafenbach), Velikovec (Völkermarkt), mežiško narečje – Kotlje, Črna na Koroškem, severnopohorsko-remšniško narečje – Kapla, Ribnica na Pohorju.

Primorska narečja skupina: rezijansko narečje – Bila (San Giorgio); obsoško narečje – Drežnica, Bovec; tersko narečje – Subid (Subit), nadiško narečje – Špeter (San Pietro), briško narečje – Biljana, Krasna, kraško narečje – Kopriva na Krasu, Renče, banjško podnarečje – Podlešče; notranjsko narečje – Hrušica, Pivka, istrsko narečje (rižansko podnarečje – Dekani, Plavje, šavrinsko podnarečje – Rakitovec, Slum).

Rovtarska narečna skupina: horjulsko narečje – Horjul, Polhov Gradec, škofjeloško narečje – Pungert, Hosta, poljansko narečje – Žiri, Gorenja vas, baško podnarečje – Rut; tolminsko narečje – Grahovo ob Bači, Most na Soči, cerkljansko narečje – Gorenja Trebuša, Cerkno, črnovrško narečje – Črni Vrh, Godovič.

Mešani kočevski govori: Livold, Mozelj.

5 TEHNOLOGIJA IN METODOLOGIJA SNEMANJA LASTNIH POSNETKOV

Pri tehnologiji in metodologiji snemanja izhajamo iz dveh temeljnih zahtev v zvezi z govornimi podatki:

1. Kvaliteta posnetkov: Za potrebe akustičnih analiz in akustičnega procesiranja je pomembno, da je na posnetku čim manj drugih zvokov poleg govora (torej brez glasbe za podlago, hrupa prometa ipd.), da format posnetka in nastavljena glasnost ne odstopata preveč od priporočljivega (gl. Snemalna oprema in tehnične lastnosti posnetkov), pa tudi da je čim manj hkratnega govora oziroma so v primeru hkratnega govora govorci posneti v ločene posnetke.
2. Avtentičnost govora: Zlasti za jezikoslovne in diskurzne raziskave je pomembno, da govorni vir predstavlja dejansko govorno obnašanje govorcev – govor na posnetkih naj ne bi bil prilagojen in izveden izključno za namen izdelave govornega vira, ampak naj bi potekal v neki resnični komunikacijski situaciji.

Oba kriterija je mogoče izpolnjevati pri javni komunikaciji, do neke mere tudi pri skupinski komunikaciji, predvsem pri zasebni komunikaciji pa sta obe zahtevi pogosto nasprotujoči si. Med drugim na govorce vpliva zavedanje, da bo njihov govor posnet; lahko da se morajo delno prilagajati snemalni opremi (npr. paziti na oddaljenost od mikrofona); zahteva po izogibanju hkratnemu govoru vpliva avtentičnost interakcije; mnoge vsebine zaradi varovanja osebnih podatkov in zasebnosti niso primerne za objavo v govornem viru.

V smernicah zato v nadaljevanju navajamo nekaj priporočil predvsem za snemanje zasebne komunikacije, ki izhajajo iz iskanja kompromisa med obema temeljnima zahtevama. Priporočila se lahko prilagajajo glede na to, ali je v neki snemalni kampanji večji poudarek na avtentičnosti gradiv ali na kvaliteti posnetkov.

5.1 Prostor snemanja

Snemanje se izvaja v prostoru, ki je miren in nima premočnih šumov iz okolja, kot so šum promet, gradbena dela, šumenje klimatske naprave, radio ali drug vir glasbe v ozadju, govor drugih govorcev... Za večjo sproščenost govorcev je priporočljivo snemanje v njihovem domačem okolju.

5.2 Teme in način govora

Govorci na posnetkih govorijo kolikor mogoče sproščeno o čim več različnih temah, ki jih izberejo glede na lastne interese, znanja in življenjske izkušnje. Teme so lahko šport, poklicno življenje, kulinarika, avtomobilizem, narava, podnebje, vrtičkarstvo, potovanja, vzgoja, zdravje, zabava, hobiji, glasba, film, gledališče, znanost, tehnologije, digitalizacija, življenjske izkušnje, spomini, navade in običaji, pripovedovanje resničnih ali izmišljenih zgodb ipd.

Oblike govora na lastnih posnetkih so lahko:

- **Intervjuji:** Pretežno monološki govor v obliki intervjuja. Intervjuvar spodbuja intervjuvanca h govorjenju z vprašanji oz. tematskimi iztočnicami, o katerih se pred snemanjem uskladi z govorcem in jih prilagodi vsakemu govorcu posebej. Na posnetku je govor obeh govorcev in oba odstopita svoj govor za govorni vir.
- **Pogovori** med dvema ali več govorci. Vsi govorci odstopijo svoj govor za govorni vir.

5.3 Dolžina in rezanje posnetkov

Posnetki so dolgi okvirno od pol ure do ene ure, lahko tudi več, vendar priporočljivo ne več kot dve uri. Posnetek je lahko obrezan na začetku in na koncu, če vključuje govor, ki ni primeren za korpus, npr. navodila za snemanje, razlaga namena snemanja, nameščanje in upravljanje snemalnih naprav, osebna komunikacija, za katero govorec ne želi, da se objavi, ipd., praviloma pa se ne izreže noben del sredi oddanega posnetka. Če je potreba, da se sredi posnetka kak del izreže, se na tem mestu pusti nekajsekundna tišina. Lahko pa se snemanje z istim govorcem razdeli v dva ali tri posnetke ali posnetke, če so vmes med snemanjem premori.

Posnetek se odda v govorni vir čim bolj tak, kot je bil posnet, brez montaže in obdelave.

5.4 Snemalna oprema in tehnične lastnosti posnetkov

Snemalna oprema mora zagotavljati zadostno kakovost posnetkov. Posnetki slabe kvalitete, z veliko šumi, poki, prenizko ali previsoko glasnostjo ipd. so slabše uporabni za avdio analize in procesiranje.

Za izvajanje lastnih posnetkov je priporočljiv prenosni snemalnik, kot je npr. Zoom h4n Handy recorder, računalnik z zunanjim mikrofonom, pametni telefoni s primerno kvalitetnim zajemom zvoka ali kvalitetnejši diktafoni (npr. Zoom h2n).

Priporočljiv format posnetkov je WAV 44,1 kHz, pcm, 16-bit, mono. Posebej občutljiv parameter nastavitve je nivo glasnosti oz. občutljivosti, saj je lahko ob napačni nastavitvi posnetek neuporaben. Ustrezno občutljivost najnatančneje preverimo s pomočjo grafičnega prikazovalnika na zaslonu snemalne naprave ali programa. Nivo zvoka je optimalen, če je vrednost na prikazovalniku pri govoru na skali med -6 in 0 (merjeno v dB). Če je pretiho (-12 ali manj), povišamo občutljivost, če nivo večkrat sega do 0, pa pomeni, da je zvok preglasen in da moramo občutljivost znižati. Priporoča se, da pred snemanjem prosimo govorca/e, da spregovori/jo nekaj besed za preizkus.

5.5 Izjave govorcev in anonimizacija

Pred začetkom snemanja vsak govorec na posnetku izpolni in podpiše izjavo, s katero dovoli snemanje in uporabo posnetka glasu, obdelavo osebnih podatkov in uporabo avtorskih pravic. Na izjavo se dopiše ID govorca, ki se uporablja v korpusu. Ime in priimek govorca ostaneta vidna samo na izjavi in pri zbiralcu gradiv, ne pa v javni objavi gradiv. Pred kakršno koli javno objavo se vsi posebni podatki v zapisu govora, na posnetkih in med metapodatki anonimizirajo. Primer tovrstne izjave je v prilogi 1.

6 TRANSKRIBIRANJE GRADIV

V dosedanji praksi izdelave govornih virov za slovenščino se je za transkribiranje prevladujoče uporabljalo orodje Transcriber (Barras et al., 2000), verzija 1.5.1, ki je prosto dostopno.[8] Transkripcije, narejene s tem orodjem, so uvozljive v vsa ostala vidnejša orodja za delo z govorom: Praat, EXMARaLDA in ELAN. Prednosti orodja Transcriber so, da je enostavno za uporabo, omogoča podroben prikaz avdio signala ter učinkovito povezavo med zapisom in avdio signalom. Za beleženje metapodatkov o govorcih in posnetkih priporočamo uporabo posebnih, ločenih preglednic. Za morebitne dodatne nivoje zapisa govora in dodajanje oznak v osnovne zapise priporočamo uvoz transkripcij iz Transcriberja bodisi v orodje EXMARaLDA bodisi za fonetične in prozodične analize v Praat bodisi za analize videa v ELAN. V tujini najdemo prakse, da se za transkirbiranje že od začetka uporabljajo tudi EXMARaLDA, Praat ali tudi ELAN.

Priporočila za transkribiranje gradiv izhajajo iz navodil za transkribiranje, ki so bila uporabljena za bazo Artur (Verdonik et al. 2023a), s tem da so ustrezno posplošena. Na podlagi izkušenj so dodatno specificirana glede postavljanja mej med segmenti oz. izjavami.

6.1 Segmentiranje in označevanje govorcev

Pri transkribiranju najprej posnetek segmentiramo na osnovne enote transkribiranja, to so segmenti.

6.1.1 MEJE MED SEGMENTI OZ. IZJAVAMI

Smernice za postavljanje mej med segmenti veljajo po naslednjem prioritetnem vrstnem redu:

1. V govoru je kratek premor, dolg 0,2 sekunde ali več, ki sovpada s koncem stavka ali kratke povedi oz. označuje neko semantično in/ali skladenjsko enoto.
2. Če bi na podlagi 0,2 sekunde ali več dolgih premorov nastali zelo kratki segmenti (počasen govor) po eno ali dve besedi, počakamo do konca stavka, preden naredimo nov segment.

[8] https://sourceforge.net/projects/trans/files/transcriber/1.5.1/

3. Če bi na podlagi 0,2 sekunde ali več dolgih premorov nastali zelo dolgi segmenti (več kot 10 sekund), naredimo nov segment tudi ob nekoliko krajšem premoru ali ob premoru, ki ni hkrati konec stavka.
4. Če je v segmentu hkratni govor, postavimo mejo med segmenti tako, da je v segmentu s hkratnim govorom čim manj nehkratnega govora. Tudi v tem primeru je premor med segmenti lahko krajši od 0,2 sekunde.

Mejo med segmenti postavimo čim bolj na sredino premora, tako da zagotovo ne odrežemo nobenega delčka predhodne ali naslednje besede.

Med segmenti mora biti vedno vsaj malo premora, da je meja med besedami vidna tudi na signalu.

6.1.2 DALJŠI PREMORI, NERAZUMLJIV GOVOR

Za navedene pojave uporabimo oznake:

- {premor} – če je premor v govoru daljši od 1,5 sekunde
- {nerazumljivo} – če gre za več kot eno besedo

Oznako vnesemo kot metaoznako, če uporabljamo Transcriber (Edit/Insert event) ali pa vpišemo v zapis. Pred uvozom v druga orodja je treba metaoznake ustrezno pretvoriti.

Za akustično procesiranje gradiv je priporočljivo, da so daljši premori in nerazumljiv govor označeni kot posebni segmenti brez govorca. Enako se lahko naredi za daljši govor v tujem jeziku, če tega ne želimo zapisati.

6.1.3 OZNAČEVANJE GOVORCA V SEGMENTU

Za vsak del zapisanega govora mora biti določeno, kateri govorec ga govori.

6.1.4 HKRATNI GOVOR

Hkratni govor se lahko pojavlja v začetku ali ob koncu segmenta, ko se prek govora enega govorca sliši hkrati že začetek govora drugega govorca ali pa začneta govoriti oba hkrati, oz. ko govorca dlje časa govorita eden čez drugega. Za akustične analize je priporočljivo, da je tak govor čim bolj natančno ločen v posebne segmente s hkratnim govorom, da se tako loči od gradiva, kjer hkratnega govora ni.

Hkratni govor zapišemo, kolikor je razumljiv, in tudi določimo oba govorca v segmentu.

6.2 Zapisovanje govora

V slovenskih govornih virih je vzpostavljena praksa dvotirnega zapisa govora:

1. pogovorni zapis (primerljiv npr. s prakso t. i. 'literary transcription' (Schmidt 2016) v konverzacijski analizi),
2. standardizirani zapis.

Za zapis gradiva izberemo eno, drugo ali obe različici. Primerjava stroškov in koristi kaže, da je dvotirni zapis smiseln (Verdonik v tisku).

Pogovorni zapis V pogovornem zapisu zapisujemo govor tako, kot ga slišimo, vendar z ortografskim zapisom, ne v fonetični abecedi, in brez naglasov. Po želji se lahko uporabijo posamezni dodatni znaki. V korpusu Gos 2.x in bazi Artur najdemo v nekaterih zapisih dodatne znake:

- @ za polglasnik
- **$g** za zveneči primorski h
- **$r** za mehkonebni koroški r

Standardizirani zapis Zapišemo dobesedno, tako da se besede ujemajo ena na ena s pogovornim zapisom, vendar v standardni obliki.

Kadar prepoznamo posebnosti pogovornega/narečnega jezika:

- oblikoslovje (npr. skladenjski vzorci, ne/določna oblika, pregibanje ipd., npr. *fižola*, *mala* namesto *majhna, večim*),
- skladnja (besedni red, vezljivost ipd., npr. *ena bolj od okuženih občin*; *nimam se kaj za pritoževati)*
- besedje (npr. *pasoš, orenk, leder*),

(1) ohranimo izvorno obliko (fižola, mala, večim, ena bolj od okuženih občin, pasoš, pasoš, orenk, leder) 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).

Priporočljivo je, da za vse dvomljive primere raziščemo predhodno prakso v korpusu Gos in vodimo evidenco lastnih odločitev.

6.2.1 REDUKCIJE

Pogovorni zapis

- Glasov, ki niso izgovorjeni, ne zapisujemo, npr. *tud, neki, tko, mam, čevli*...
- Polglasnik se je do sedaj zapisoval bodisi z znakom @ ali s črko e ali brez nje:
    - pri zvočnikih r, l, m, n: *s@n, p@r, misl@m, hit@r, zlom@l, prjat@lci, fil@m* ... oz. *sn, pr, mislm, hitr, zloml, prjatlci, film*...
    - pri enoglasovnih predlogih, členkih ipd., izgovorjenih vokalizirano, s polglasnikom: *s@, z@, d@* oz. *s, z, d*
    - pri enozložnih besedah: *j@z, n@č*... oz. *jz, nč*
    - v dvo- ali večzložnih besedah: *k@šni (kakšni)* oz. *kešni*
    - v zborni izreki, npr. b@z@g, p@s oz. *bezeg, pes*
- Zapisovanje oblik pomožnega glagola »biti«:
    - redukcije »bi« v »b« zapisujemo kot samostojno besedo, npr. *ne b (ne bi), če b (če bi), pa b mene (pa bi mene)*, *najraj b vidu*...
    - redukcije in premene oblik za prihodnjik (*bom, boš, bo*...) zapisujemo na naslednji način: *čev (če bo), navm (ne bom), nav (ne bo)*...

Standardizirani zapis Zapišemo v standardni slovenščini: *tud -> tudi, neki -> nekaj, tko -> tako, mam -> imam, čevli -> čevlji, s@n -> sem, p@r -> pri, misl@m -> mislim, hit@r -> hitro, zlom@l -> zlomil, prjat@lci -> prijateljici, fil@m -> film; j@z -> jaz, n@č -> nič; k@šni -> kakšni; ne b -> ne bi, če b -> če bi, pa b mene -> pa bi mene ->*, *najraj b vidu -> najraje bi videl*...

6.2.2 PREMENE PO ZVENEČNOSTI

Pogovorni zapis Premen po zvenečnosti v pisavi ne upoštevamo (zapišemo *tud dobr, tud tak*, čeprav se sliši *tut dobr, tut tak*). Izjema so predlogi *s/z* in *k/h* – te pišemo tako, kot so izgovorjeni. Če niso izgovorjeni skupaj z naslednjo besedo, ampak vokalizirano, s polglasnikom, jih lahko zapišemo kot *s@, z@, k@…*, če uporabljamo poseben znak za polglasnik.

Standardizirani zapis Zapišemo v standardni slovenščini: *tud dobr -> tudi dobro, tud tak -> tudi tak* oz. s/z in k/h po pravopisu.

6.2.3 DVOUSTNIČNI V IN SAMOGLASNIK U

Pogovorni zapis 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.

Standardizirani zapis Po pravopisu: *prov -> prav, nav -> ne+bo, navm -> ne+bom, odpravt -> odpraviti, davn -> down, gledavc -> gledalec, pov@n -> poln; lavfati -> lavfati, šlavf -> šlavf, genav -> genav, mav -> malo*, *šov -> šel*, *dov -> dol*, *prov -> prav*, *dav -> da+bo*, *nov -> ne+bo*, *av lavfati, šlavf, genav, mav* (*malo*), *šov* (*šel*), *dov* (*dol*), *prov* (*prav*), *dav* (*da bo*), *nov* (*ne bo*), tudi medmet *av -> av; pršu -> prišel, vidu -> videl, u tem delu -> v tem delu*.

6.2.4 NAREČNO SPECIFIČNI GLASOVI

Diftonge in druge pokrajinsko specifične foneme, ki jih ni v knjižnem jeziku, pišemo z najbližjimi ustreznimi črkami, odvisno tudi od izgovorjave v konkretnih primerih, npr. *»ej«, »ov«, »je«; »u«* za u s preglasom; »h« za zveneči primorski h; »r« za mehkonebni koroški r itd. Lahko se dodajo posebni znaki (npr. znak $ pred črko).

6.3 Tehnikalije

6.3.1 LOČILA

Pogovorni zapis Izjave začenjamo z veliko začetnico. Uporabljamo ločila:

- pika,
- vejica,
- klicaj,
- vprašaj,
- podpičje,
- narekovaj in dvopičje (za dobesedni navedek),
- tri pike (nestično levo in desno) za nedokončane izjave, pri samopopravljanjih, kratkih premorih znotraj stavka …,
- vezaj – samo obojestransko stično pri sklanjanju kratic (npr. *RTV-ja*),

- opuščaj (apostrof) – samo kot del lastnega imena ali za angleške pojme (npr. *O'Rilley*),
- znak za in (&) – samo kot del lastnega imena.

Vezaja v prirednih zloženkah in pomišljaja ne uporabljamo, saj se pojavljajo težave pri analizi podatkov npr. za izdelavo slovarja, zapisa pomišljaja pa orodja običajno ne omogočajo.

Ločila uporabljamo samo v skladenjski rabi, pišemo jih stično in skladno s pravopisom.

Standardizirani zapis Ločila se postavijo identično kot v pogovornem zapisu. V izogib napakam je priporočljivo avtomatizirano usklajevanje ločil v obeh zapisih.

6.3.2 KRATICE

Pogovorni zapis

Kratice pišemo tako, kot so izgovorjene, z malimi črkami in skupaj, če gre za eno kratico. Polglasnik lahko zapisujemo z dodatnim posebnim znakom za polglasnik (@), če ga uporabljamo tudi sicer v zapisu: *erteve, teve, trr oz. t@r@r@.*

Če je kratica lastno ime, jo pišemo z veliko začetnico: *Zrc Sazu oz. Z@r@c@ Sazu, Tevetri, Ajdžiem, Erteve Slovenija.*

Pri sklanjanju kratic, ki bi jih standardizirano zapisali z velikimi črkami, obvezno uporabimo vezaj skladno s pravopisom, da se ob koncu raba vezajev v standardiziranem in pogovornem zapisu ujemala. Primeri: *Amzs-ju oz. Am@z@s@-ju (standardiziran zapis bo AMZS-ja), Sazu-ja (standardiziran zapis bo SAZU-ja).*

Standardizirani zapis

Kratice zapišemo po pravopisu, npr. RTV*, TV, TRR,* ali kot lastno ime, če pravopis to dovoljuje, npr. *Sazuja*. Pri sklanjanju kratice usklajeno s pogovornim zapisom uporabljamo vezaj (npr. *AMZS-ja*).

6.3.3 ŠTEVILKE, OKRAJŠAVE, DATUMI

Številke vedno izpišemo z besedo.

Okrajšav ne uporabljamo, vse, kar je povedano, izpišemo v celoti.

Datume zapisujemo z besedami, kot so izgovorjeni.

6.3.4 ČLEN *TA*

Pogovorni zapis Določni člen *ta* pišemo stično (*je šu tist talep lijak ven; tamali, tamavga, taprav, tazaden*). V različici korpusa Gos 1.0 je bil uporabljen nestični zapis: ta lep, ta mavga.

Standardizirani zapis Zapišemo kot dve besedi (*je šel tisti ta lep lijak ven*).

6.3.5 FRAGMENTI

Pogovorni zapis Besedne fragmente (prekinjene besede ipd.) označimo s praznim oklepajem stično za besedo, npr. *lju()*. Za besedne fragmente štejejo samopopravljanja – ko govorec začne izgovarjati neko besedo, pa jo sredi izgovarjanja prekine in izreče neko drugo besedo.

Standardizirani zapis Ostane identično kot v pogovornem zapisu.

6.3.6 NERAZUMLJIVO

Pogovorni zapis Posamezno nerazumljivo besedo ali kratko frazo označimo z oznako {nerazumljivo} oz. vstavimo metaoznako (Edit/Insert event v Transcriberju).

Če je nerazumljiv daljši del govora (dve ali več besed), je priporočljivo, da ga ločimo v poseben segment brez govorca.

Standardizirani zapis Ostane identično kot v pogovornem zapisu.

6.3.7 ANONIMIZACIJA

Pogovorni zapis Če so v govoru izraženi osebni podatki o govorcih (ime in priimek ipd. – samo ime ne šteje za osebni podatek), ki niso javne osebnosti, jih pripravimo za naknadno anonimizacijo, tako da zapišemo podatke med oglate oklepaje, npr. *[Janez Novak]*. Anonimiziramo tudi osebne podatke oseb, ki so v govoru omenjene, če gre za nejavne osebnosti.

Normalno, brez oglatih oklepajev, zapišemo imena javnih osebnosti, ki so omenjena v govoru, npr. imena politikov, športnikov, novinarjev in voditeljev, umetnikov in drugih kulturnih delavcev, profesorjev, znanstvenikov (ali raziskovalcev), direktorjev ter ostalih medijsko opaznih osebnosti. Če se o teh osebah govori na sovražen ali negativen način, pa anonimiziramo tudi te.

Za vse osebne podatke, ki jih pripravimo za anonimizacijo z zapisom v oglatih oklepajih, dodatno s časovnimi značkami označimo tisti del zvočnega posnetka, v katerem je izgovorjen osebni podatek oz. zaporedoma več osebnih podatkov (v Transcriberju lahko npr. kot Segmentation/Insert background/shh). V kolikor se na posnetku za zelo kratek čas pojavi govor oseb, ki niso podale soglasja za uporabo posnetka, se lahko uporabi isti način označevanja. Tako označeni del posnetka se pred izdajo za javnost prekrije s piskom.

Standardizirani zapis Ostane identično kot v pogovornem zapisu.

6.4 Lapsusi, lastna imena, citatne in tuje besede

6.4.1 LAPSUSI

Pogovorni zapis Zapišemo, kot slišimo: *indidualnih -> individualnih*.

Standardizirani zapis Če so nedvoumni, odpravimo: *indidualnih -> individualnih*.

6.4.2 LASTNA IMENA

Pogovorni zapis Domača lastna imena zapisujemo po pravopisu: *Delo, Brežice, Novo mesto*.

Tuja lastna imena zapisujemo tako, kot slišimo, in z veliko začetnico po pravopisu: *Bler, Hjuston, Nju Jork, Los Endželes, Kanace*.

Standardizirani zapis Po pravopisu: *Blair, Huston, New York, Los Angeles*.

6.4.3 CITATNE BESEDE

Pogovorni zapis Citatne besede, ki niso lastno ime, pišemo tako, kot so izgovorjene, npr. *džakuzi, voš mašina*.

Standardizirani zapis Po pravopisu: *džakuzi* ali *jacuzzi*.

6.4.4 TUJE BESEDE

Pogovorni zapis Če je v tujem jeziku posamezna beseda ali kratka fraza, zapišemo enako kot citatne besede, npr. *jes, let it bi, fani*. Če je v tujem jeziku cela izjava ali več izjav, lahko označimo kot prazen segment brez govora.

Standardizirani zapis Po pravopisu: *yes, let it be, funny*.

6.5 Posebni glasovi in zvoki

6.5.1 NEVERBALNI IN POLVERBALNI GLASOVI

Pogovorni zapis Podaljšan polglasnik ali zvočnik *m* ali *n* in njihove kombinacije, ki pogosto zapolnjujejo premore v govoru, pišemo s tremi črkami in znakom # na začetku, in sicer: *#eee, #eem, #een, #nnn, #mmm* … (in ne kot *#e* ali *#ee*). Druge medmete zapišemo z nizom črk, ki najbolje ustreza dejanski izgovorjavi. Trajanja medmetov ne označujemo posebej. Posebej pozorni smo na zapis neleksikaliziranega glasovnega zanikanja, kjer uporabimo zapis *#nn* ali *#aa*.

Če medmet stoji na začetku povedi, ga zapišemo z veliko začetnico, besedo, ki mu sledi, pa z malo začetnico. Takšno poved zaključimo z ustreznim ločilom. Primer: *#Eee vedno zamudi,*

#eee, brez izjeme. Zlasti medmeta #eee in podobnih ni treba z vejico ločevati od ostalega besedila.

Načela zapisovanja medmetov so:

- izraze zapišemo raje z eno besedo kot več besedami (npr. *#ojoj* namesto *#o #joj*),
- 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*),
- 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*),
- dvoustnični U prednostno pišemo z 'v' (*#av, #vav*),
- podaljševanje glasov se ne označuje z več črkami, ampak se ohranja enoten zapis (npr. vedno *#jee*, ne *#jeee* ali podobno),
- prednost ima poslovenjen zapis (npr. *jes*, ne *yes*, *okej*, ne *ok* ali *okay*).

Standardizirani zapis Ostane identično kot v pogovornem zapisu.

6.5.2 ZVOKI

Pogovorni zapis Zapisujemo:

- **{smeh}**, ki označuje smeh govorca
- **{glas},** ki označuje zvoke, ki nastanejo z govorili, kot so zehanje, vzdih, odkašljanje, pogrkavanje ipd.
- **{dih}**, ki označuje močno slišen vdih ali izdih med govorjenjem
- **{zvok}**, ki označuje zvoke, ki ne nastanejo z govorili, npr. zvonjenje telefone, zapiranje vrat

Lahko jih vnesemo kot metaoznake (v Transcriberju opcija Edit/Insert event). Pred uvozom v druga orodja moramo take metaoznake ustrezno pretvoriti.

Standardizirani zapis Ostane identično kot v pogovornem zapisu.

Zahvala

Delo je nastalo v okviru raziskovalnega projekta ARIS Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik (J7-4642).

7 LITERATURA

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1–2), 5–22.

Krajnc Ivič, M. (ur.) (v tisku). *Stanje in perspektive uporabe govornih virov v raziskavah govora*. Maribor: Univerza v Mariboru, Univerzitetna založba. [Elektronski vir]

Love, R., Dembry, C., Hardie, A., Brezina, V., McEnry, T. (2017). The spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics,* 22(3), 319–344.

Schmidt, T. (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.

SLA 3.1 = Škofic, Jožica, Gostenčnik, Januška, Hazler, Vito, Jakop, Tjaša, Kenda-Jež, Karmen, Kumin Horvat, Mojca, Nartnik, Vlado, Pahor, Nina, Smole, Vera, Šekli, Matej, Zuljan Kumar Danila, (ur.: Škofic, Jožica, Kenda-Jež, Karmen, Kumin Horvat, Mojca). 2023. *Slovenski lingvistični atlas 3: kmetovanje*, 1: atlas. Ljubljana: Založba ZRC, ZRC SAZU (Jezikovni atlasi).

Trojar, M., Bizjak, A. (2023). Transkribiranje govora pri izdelavi govorne baze Artur: od pogovornih k standardiziranim zapisom. V: Arhar Holdt, Š. (ur.), Krek, S. (ur.). *Razvoj slovenščine v digitalnem okolju*. Ljubljana: Založba Univerze, 39–59. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9445.

Verdonik, D., Bizjak, A., Žgank, A., Dobrišek, S. (2022). Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. V: Fišer, D. (ur.), Erjavec, T. (ur.). *Jezikovne tehnologije in digitalna humanistika: zbornik konference.* Ljubljana: Inštitut za novejšo zgodovino, 205–212. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf.

Verdonik, D. (2023). Zbiranje gradiv za govorne korpuse med Scilo in Karibdo. V: Arhar Holdt, Š. (ur.), Krek, S. (ur.). *Razvoj slovenščine v digitalnem okolju*. Ljubljana: Založba Univerze, 15–37. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9447.

Verdonik, D., et al. (2023a). ASR database ARTUR 1.0 (transcriptions). Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1772.

Verdonik, D., et al. (2023b). Spoken corpus Gos 2.1 (transcriptions). Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1863.

Verdonik, D., Tojar, M., Bizjak, A. (v tisku). Prednosti in slabosti dvotirnega zapisovanja govora v slovenskih govornih virih. Krajnc Ivič, M. (ur.). *Stanje in perspektive uporabe govornih virov v raziskavah govora.* Maribor: Univerza v Mariboru, Univerzitetna založba. [Elektronski vir]

PRILOGA 1: PRIMER IZJAVE ZA GOVORCE

Dovoljenje za snemanje in uporabo posnetka glasu

in

privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov

ter dovoljenje za uporabo avtorskih pravic

ID govorca:	
Ime in priimek:	
Spol:	
Starost:	
Izobrazba:	
Prvi jezik:	
Dvojezičnost:	
Dodaten prvi jezik:	
Kraj, občina, država bivanja v otroštvu:	
Kraj, občina, država bivanja v sedanjosti:	
Kraj, občina, država daljšega bivanja drugod:	

1. Spodaj podpisani izjavljam, da sem seznanjen, da **podatki o nosilcu projekta** (v nadaljevanju nosilec projekta) izvaja projekt **podatki o projektu**, ki je bil na javnem razpisu **podatki o razpisu in izboru**. V okviru projekta bo nosilec projekta pripravil govorni korpus v obsegu **xx ur**, ki bo osnova za **namen vira.** Govorni korpus bo javno dostopen pod pogoji prostih licenc (npr. CC BY) in bo na voljo za nekomercialen in komercialen razvoj tehnologij (npr. govorno upravljanje naprav, pogovorni agenti, avtomatsko podnaslavljanje video vsebin ali avtomatsko prevajanje govorjenih vsebin), za jezikoslovne, sociološke in druge raziskave ter za druge raziskovalne namene.

2. Spodaj podpisani soglašam, da se me posname pri naključnem govoru in da se takšen posnetek uporabi za zgoraj navedene namene ter da se takšen posnetek opremi z ustreznim identifikatorjem, ki bo povezan z osebnimi podatki, navedenimi v tej izjavi, ter z metapodatki o snemanju, ki se nanašajo na kraj snemanja, čas snemanja, snemalno opremo, okoliščine snemanja in značilnosti posnetka in govora na posnetku.

3. Zavedam se, da se vsi zgoraj navedeni osebni podatki, vključno s posnetkom glasu, obdelujejo na podlagi moje privolitve in hkrati skladno z zakonitim interesom znanstvenoraziskovalne dejavnosti in tehnološkega razvoja.

4. Spodaj podpisani dajem privolitev za obdelavo podatkov, opisanih v tej izjavi, vključno s posnetkom mojega glasu, v skladu s Splošno Uredbo o varstvu osebnih podatkov (GDPR). Privolitev lahko kadarkoli prekličem tako, da kontaktiram upravljavca na spodaj navedeni e-poštni naslov. Preklic privolitve ne bo vplival na zakonitost obdelave pred preklicem.

5. Spodaj podpisani izjavljam, da sem seznanjen s tem, da se bodo moji zgoraj navedeni osebni podatki hranili za nedoločen čas, dokler traja opisani namen tega projekta.

6. Spodaj podpisani sem seznanjen s tem, da imam v skladu z GDPR pravico, da od nosilca projekta zahtevam dostop do svojih osebnih podatkov; od nosilca projekta zahtevam popravek svojih osebnih podatkov; od nosilca projekta zahtevam izbris svojih osebnih podatkov; od nosilca projekta zahtevam omejitev obdelave svojih osebnih podatkov; ugovarjam obdelavi svojih osebnih podatkov; od nosilca projekta zahtevam prenos svojih osebnih podatkov; ali pri Informacijskemu pooblaščencu RS vložim pritožbo.

7. Spodaj podpisani sem seznanjen, da je upravljavec podatkov: **podatki o upravljalcu, obvezno tudi e-naslov**.

8. Spodaj podpisani s podpisom te izjave nosilcu projekta dovoljujem uporabo omenjenega posnetka v zgoraj navedene namene in zato na nosilca projekta prenašam vse materialne avtorske pravice, druge pravice avtorja v skladu z Zakonom o avtorski in sorodnih pravicah (ZASP) in avtorski sorodne pravice, ki utegnejo nastati pri snemanju in prebiranju besedil, kakor je opisano zgoraj. Materialne avtorske pravice, druge pravice avtorja v skladu z ZASP in avtorski sorodne pravice, ki pri tem nastanejo, se na naročnika prenesejo izključno, geografsko ter časovno neomejeno in neodplačno, brez kakršnihkoli omejitev, vključno z dovoljenjem, da jih nosilec projekta prenese naprej na tretje osebe.

9. Vsebina te izjave ne vpliva na prenos moralnih avtorskih pravic, ki so v skladu z določbami ZASP neprenosljive.

Datum in kraj: _____ Ime in priimek:_____

Podpis: _____