



Univerzitetna založba
Univerze v Mariboru

Andreja Bizjak

Pridobivanje govornih virov

Prednosti in pomanjkljivosti
različnih pristopov



Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko

Pridobivanje govornih virov

Prednosti in pomanjkljivosti različnih pristopov

Avtorica

Andreja Bizjak

December 2025

Naslov <i>Title</i>	Pridobivanje govornih virov <i>Collection of Speech Resources</i>
Podnaslov <i>Subtitle</i>	Prednosti in pomanjkljivosti različnih pristopov <i>Advantages and Limitations of Different Approaches</i>
Avtorica <i>Author</i>	Andreja Bizjak (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko)
Recenzija <i>Review</i>	Darinka Verdonik (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko)
Lektoriranje <i>Language editing</i>	Andreja Bizjak (Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko)
Tehnični urednik <i>Technical editor</i>	Jan Perša (Univerza v Mariboru, Univerzitetna založba)
Oblikovanje ovitka <i>Cover designers</i>	Jan Perša (Univerza v Mariboru, Univerzitetna založba)
Grafika na ovitku <i>Cover graphics</i>	Three crumpled yellow papers on green surface surrounded by yellow lined papers, foto: Volodymyr Hryshchenko, 2019
Založnik <i>Published by</i>	Univerza v Mariboru Univerzitetna založba Slomškov trg 15, 2000 Maribor, Slovenija https://press.um.si , zalozba@um.si
Izdajatelj <i>Issued by</i>	Univerza v Mariboru Fakulteta za elektrotehniko, računalništvo in informatiko Koroška cesta 46, 2000 Maribor, Slovenija https://feri.um.si , feri@um.si
Izdaja <i>Edition</i>	Prva izdaja
Vrsta publikacije <i>Publication type</i>	E-knjiga
Dostopno na <i>Available at</i>	https://press.um.si/index.php/ump/catalog/book/1061
Izdano <i>Published</i>	Maribor, Slovenija, december 2025

Ime projekta Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik
Project name

Šifra projekta J7-4642
Project number

Financer projekta Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost
Project financier Republike Slovenije



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



© **Univerza v Mariboru, Univerzitetna založba**
/ *University of Maribor, University of Maribor Press*

Besedilo / Text © Bizjak (avtorica), 2025

To delo je objavljeno pod licenco Creative Commons Priznanje avtorstva-Deljenje pod enakimi pogoji 4.0 Mednarodna. / *This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License.*

Licenca dovoli uporabnikom reproduciranje, distribuiranje, dajanje v najem, javno priobčitev in predelavo avtorskega dela, če navedejo avtorja in širijo avtorsko delo/predelavo naprej pod istimi pogoji. Za nova dela, ki bodo nastala s predelavo, bo tako tudi dovoljena komercialna uporaba. Od BY NC SA licence se ta razlikuje samo v tem, da je tu dovoljena tudi komercialna uporaba dela/predelave.

Vsa gradiva tretjih oseb v tej knjigi so objavljena pod licenco Creative Commons, razen če to ni navedeno drugače. Če želite ponovno uporabiti gradivo tretjih oseb, ki ni zajeto v licenci Creative Commons, boste morali pridobiti dovoljenje neposredno od imetnika avtorskih pravic.

<https://creativecommons.org/licenses/by-sa/4.0/>

CIP - Kataložni zapis o publikaciji
Univerzitetna knjižnica Maribor

811.163.6:001.8(0.034.2)

BIZJAK, Andreja, 1981-

Pridobivanje govornih virov [Elektronski vir] : prednosti in pomanjkljivosti različnih pristopov / avtorica Andreja Bizjak. - 1. izd. - E-knjiga. - Maribor : Univerza v Mariboru, Univerzitetna založba, 2025

Način dostopa (URL): <https://press.um.si/index.php/ump/catalog/book/1061>

ISBN 978-961-299-093-0 (PDF)

doi: 10.18690/um.ff.11.2025

COBISS.SI-ID 261103107

ISBN 978-961-299-093-0 (pdf)

DOI <https://doi.org/10.18690/um.feri.11.2025>

Cena
Price Brezplačni izvod

Odgovorna oseba založnika Prof. dr. Zdravko Kačič,
For publisher rektor Univerze v Mariboru

Citiranje Bizjak, A. (2025). *Pridobivanje govornih virov: prednosti in*
Attribution *pomanjkljivosti različnih pristopov*. Univerza v Mariboru,
Univerzitetna založba. doi: 10.18690/um.feri.11.2025

Kazalo

1	Uvod.....	1
1.1	Govorni viri v slovenskem prostoru.....	3
2	Različni načini pridobivanja govornih virov	5
2.1	Snemanje v laboratoriju ali studiu	5
2.2	Snemanje na terenu	7
2.3	Snemanje na daljavo.....	8
2.4	Games-with-a-Purpose.....	9
2.5	Collect4NLP.....	10
2.6	Pridobivanje govornih virov preko množičenja	11
2.7	Pridobivanje govornih virov preko občanske znanosti.....	13
3	Primerjalna analiza različnih pristopov.....	15
4	Tuji primeri dobrih praks.....	17
5	Priporočila za pridobivanje govornih virov	19
	Literatura	23

1 Uvod

Govorni viri, tj. (multimodalni) posnetki govornenega jezika, ustrezno označeni z metapodatki in opremljeni s soglasji govorcev za njihovo uporabo, so temeljnega pomena za razvoj govornih tehnologij, kot sta avtomatsko razpoznavanje govora (ASR) in sinteza govora ter za jezikoslovno preučevanje govornenega jezika. Zaradi zahtevnosti kontinuiranega pridobivanja posnetkov (spontano) govornene slovenščine so tovrstni viri v Sloveniji še vedno v pomanjkanju, kar slovenščino uvršča med manj podprte jezike¹ (Hutin in Allasonnière-Tang, 2022). Izdelava govornih korpusov je časovno in stroškovno zahtevna, zato raziskovalna skupnost preučuje nove možnosti bolj trajnostnega in razširljivega načina pridobivanja govornih virov na daljavo, ob upoštevanju specifičnih tehničnih, finančnih, pravnih, etičnih in motivacijskih vidikov (Bizjak, 2025).

Med ključnimi izzivi sodobnejših načinov izgradnje govornih baz, ki vključuje snemanje na daljavo, so motiviranje govorcev za sodelovanje ter zagotavljanje zvočne kakovosti in pristnosti posnetega govora. Pridobivanje posnetkov vsakdanje, spontano tvorjene slovenščine v avtentičnem okolju z naravnimi govorci je še posebej zahtevno. Govorci se pogosto soočajo z nezaupanjem, saj je govor biometrični osebni podatek, kar pomeni, da je govorca po govoru (npr. intonacija,

¹ Less-resourced (angl.)

tempo, barva glasu ...) mogoče identificirati². Zaradi tega je ključno, da raziskovalci govorce seznani z namenom in načini rabe posnetkov, z obdelavo in hrambo podatkov ter varovanja osebnih podatkov in zaščite avtorskih pravic, predvsem pa v največji možni meri zagotovijo anonimizacijo, kar lahko pozitivno vpliva na pripravljenost govorcev za sodelovanje (Neale idr., 2017; Rutten idr., 2017).

S pripravljenostjo govorcev za sodelovanje se odpre pomembno metodološko vprašanje zagotavljanja avtentičnosti zbranega govora. Dialektologi, sociolingvisti in drugi jezikoslovci opozarjajo na t. i. Labovov paradoks: raziskovalčev cilj je ugotoviti, kako ljudje govorijo, kadar jih nihče ne opazuje, vendar pa tak govor lahko pridobimo le s sistematičnim opazovanjem (Labov, 1973). Čeprav popolnoma nenadzorovan govor ni dosegljiv – in morda tudi ni nujen (Wertheim, 2006) – lahko snemanje brez neposredne prisotnosti raziskovalcev zmanjša samoopazovanje in prispeva k večji naravnosti govora.

Dodaten izziv predstavljajo tehnične omejitve snemanja v neprofesionalnih pogojih: pogosti so šumi iz okolja, nestandardna snemalna oprema (npr. mobilni telefoni) in variabilna kakovost zvoka, kar lahko vpliva na uporabnost posnetkov za nadaljnjo analizo in procesiranje. Nadzor kakovosti je mogoče izvajati pred izvedbo nalog, med njihovim potekom ali po zaključku, pri čemer vpeljava validacije kljub dodatnim stroškom zagotavlja, da je delo udeležencev transparentno validirano in da so neustrezni rezultati v čim večji meri izločeni iz objavljenih jezikovnih baz in korpusov. (Eskénazi idr., 2013)

Za dolgoročnejše prostovoljno sodelovanje govorcev v raziskovalnem projektu je med drugim pomembno vključevanje lokalno relevantnih tem, saj osebni interes povečuje njihovo angažiranost. Sodelujočim morajo raziskovalci v čim večji meri zagotoviti avtonomnost in si prizadevati, da bo njihova izkušnja prijetna, predvsem pa tudi širše prepoznana in cenjena. Ključno je ažurno odzivanje na sprotna vprašanja in dileme govorcev, z višanjem zahtevnosti nalog pa lahko ohranijo motivacijo tudi bolj izkušenih sodelujočih. Uspešnost projekta zahteva ciljno prilagojen komunikacijski načrt, ki temelji na analizi motivacijskih dejavnikov glede

² »Biometrijske ukrepe ureja Zakon o varstvu osebnih podatkov (Uradni list RS, št. 163/22; ZVOP-2) ... ZVOP-2 ureja biometrijske ukrepe v členih od 81 do 84 (4. poglavje II. dela zakona). Obdelava biometričnih osebnih podatkov v nasprotju z določbami tega poglavja ZVOP-2 je prepovedana.« Informacijski pooblaščenec RS. *Prijava biometričnih ukrepov*. Informacijski pooblaščenec Republike Slovenije, www.ip-rs.si/varstvo-osebni-podatkov/obveznosti-upravljavcev/prijava-biometrijskih-ukrepov/.

na različne ciljne skupine govorcev, saj so najbolj aktivni sodelujoči pogosto poklicno povezani s tematiko projekta ali pa se zanjo zanimajo v prostem času (Rutten idr., 2017). Ugotoviti je treba, kaj motivira govorce za oddajo govornih posnetkov, saj je motivacija večplastna in odvisna od vsakokratnega konteksta. Govorce spodbujajo tako notranji dejavniki, kot so zanimanje za jezik, občutek pripadnosti, jezikovna identiteta in želja po zaščiti in ohranjanju jezika, kot tudi zunanje spodbude, kot so javna priznanja in (simbolične) nagrade. Ključni za dolgotrajno sodelovanje so predvsem osebna povezanost s tematiko ter občutek smiselnega in transparentnega prispevka k skupnemu znanju in znanosti (Heinisch, 2022).

Da bi se soočili z nekaterimi zgoraj naštetimi izzivi bomo v elaboratu predstavili različne možnosti sistematičnega zbiranja govornih virov, zlasti tiste, ki omogočajo inovativno in razširjeno zajemanje govornih podatkov tudi izven institucionalnih in raziskovalnih okolij. Cilj tega elaborata je ponuditi praktično uporaben pregled različnih tehnik snemanja govornih virov, pri čemer se osredotočamo na njihove prednosti, omejitve in izvedbeno zahtevnost.

1.1 Govorni viri v slovenskem prostoru

Govorjena slovenščina, kot jo govorci uporabljajo v različnih regijah, v zamejstvu in med izseljenci, je izjemno raznolika na ravni izgovorjave, leksike, oblikoslovja in skladnje. Kljub tej raznolikosti ostaja bistveno manj dokumentirana kot pisna slovenščina. Govorni korpusi omogočajo poglobljeno analizo fonetičnih, fonoloških, pragmatičnih in diskurzivnih pojavov, ki jih iz pisnih virov ni mogoče razbrati. Posebej pomembni so za razumevanje slovnice govornega jezika, diskurzivnih označevalcev, govorne kohezije in spontanosti (Verdonik, 2008; Verdonik in Maučec, 2017). Pomanjkanje podatkov o tipično govornem besedišču (npr. zrihtati, poštimati) pomeni oviro za razvoj tehnologij, kot so razpoznavalniki govora, ki takih besed ne prepoznajo pravilno (Čibej idr., 2024).

V zadnjih letih je bilo v Sloveniji narejenih več korakov v smeri gradnje govornih virov:

- referenčni govorni korpus Gos 2.1³ in govorna baza Artur 1.0⁴ temeljita na studijskih in terenskih snemanjih, ki so jih izvedli najeti snemalci;

³ <https://www.clarin.si/repository/xmlui/handle/11356/1863>

⁴ <https://www.clarin.si/repository/xmlui/handle/11356/1776>

- prvi poskus množičenja z metodo *Games-with-a-Purpose* pri nas predstavlja mobilna igra, imenovana Igra besed (Arhar Holdt idr., 2021);
- v teku je projekt Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov (PoVeJMo)⁵, katerega cilj je zbrati 40 milijard besedil, vključno z govorjenimi;
- portal Govorjena slovenščina⁶ je del projekta LLM4DH⁷, ki se osredotoča na raziskave razvoja velikih jezikovnih modelov za slovenščino in njihovo uporabo v digitalni humanistiki. Portal je namenjen zbiranju posnetkov vsakdanje slovenščine, ki bodo uporabljeni za raziskave govora, razvoj digitalnih orodij za slovenski jezik in kot arhiv jezikovne kulturne dediščine.

Čeprav so ti projekti dragoceni, obstajajo pomembne vrzeli, kot so pomanjkanje sistematične pokritosti geografsko, dialektološko in žanrsko raznolikih govornih variant ter kvantitativno pomanjkanje prosto dostopnih kakovostnih in anotiranih posnetkov spontane govorjene slovenščine. Nadaljnje sistemsko podprto zbiranje in dolgoročno dokumentiranje govora je ključno za jezikoslovne raziskave, razvoj slovenskih govornih tehnologij in ohranjanje jezika v digitalnih okoljih.

⁵ <https://povejmo.si/>

⁶ <https://govorjena-slovenscina.um.si/>

⁷ <https://www.cjvt.si/llm4dh/>

2 Različni načini pridobivanja govornih virov

Zbiranje govornih virov je mogoče izvajati na različne načine, ki se med seboj razlikujejo vsaj po stopnji nadzora nad snemalnimi pogoji, zahtevnosti in načinu izvedbe, kakovosti posnetkov in stopnji spontanosti govora. V nadaljevanju predstavljamo najpogostejše pristope, ki so se uveljavili v domačih in tujih jezikoslovnih projektih, ter nekatere njihove prednosti in omejitve.

2.1 Snemanje v laboratoriju ali studiu

Zbiranje govornih podatkov v nadzorovanih okoljih vključuje dve osnovni izvedbeni možnosti, ki ju bomo za namen tega elaborata opisno poimenovali kot *snemanje v laboratoriju* in *snemanje v studiu*. Poimenovanji nista terminološko uveljavljeni strokovni kategoriji, temveč služita kot opisna označevalca za razlikovanje med zajemom govora za raziskovalne namene na eni strani ter medijsko produkcijo na drugi. Oba pristopa temeljita na tehnični optimizaciji zvoka, omogočata standardizacijo pogojev snemanja in zagotavljata visoko kakovost posnetkov, a se razlikujeta po organizaciji snemanja, vlogi udeležencev ter stopnji spontanosti govora.

Pri govoru, ki se za raziskovalne namene snema v laboratoriju, zbiranje govornih podatkov običajno poteka v akustično izoliranem prostoru, pogosto znotraj znanstvene ustanove. Poudarek je na nadzoru nad vsemi dejavniki snemanja (tip mikrofona in snemalna oprema, pozicija govorca, tematika itd.). Namen je pridobiti ponovljive vzorce govora visoke zvočne kakovosti brez motečih dejavnikov, kar olajša akustično analizo in standardizacijo govornega gradiva. Takšne prakse se uporabljajo predvsem v fonetiki ter pri razvoju in testiranju govornih tehnologij. Zlasti po izbruhu pandemije covid-19, ko snemanje v živo v laboratorijih pogosto ni bilo izvedljivo, so raziskovalci (Awan idr., 2023; Zhang, Jepson in Chuang, 2024) začeli intenzivneje preučevati razlike v kakovosti zvoka med t. i. zlatim standardom posnetkov, narejenih v nadzorovanih laboratorijskih pogojih, in posnetki, zbranimi na daljavo prek pametnih telefonov ali spletnih platform za videokonference.

Tabela 1

Značilnost	Snemanje v laboratoriju	Snemanje v studiu
Akustični pogoji	Zelo dobri (akustično optimizirana, nadzorovana soba, omejen hrup)	Odlični (profesionalna oprema, izolacija)
Tehnična oprema	Snemalna oprema dobre kakovosti	Studijska oprema, večkanalni snemalniki
Nadzor nad spremenljivkami	Visok (natančno določeni in nadzorovani pogoji snemanja)	Srednje visok (vnaprej pripravljena besedila ali scenariji)
Avtentičnost izražanja	Nizka (nadzorovano, nedomače okolje, prisotnost raziskovalca)	Nizka do srednja (odvisno od žanra in izkušenosti govorca)
Pridobivanje soglasij	Tipično interno (znotraj institucij)	Standardizirano (studijska produkcija, pogodbe)
Tipični govorci	Poljubni udeleženci, izbrani glede na demografijo (omejena dostopnost za oddaljene govorce)	Profesionalni govorci, napovedovalci, igralci, gostje
Motivacija govorcev	Denarna nagrada, institucionalna ali študijska obveznost, prispevek k znanosti	Plačilo in profesionalni interes
Ustreznost za spontane govore	Nizka	Nizka do srednja
Stroški izvedbe	Srednji do visoki	Visoki (studio, tehniki, režiserji, koordinacija)

Profesionalni snemalni studii, značilni za medijsko produkcijo, so tehnično optimizirani in opremljeni z akustičnimi oblogami, visokokakovostnimi mikrofoni, mešalnimi mizami in digitalnimi delovnimi postajami. Zagotavljajo jasno, profesionalno reprodukcijo zvoka in javnega govora ter ustvarjanje vsebin, primernih za oddajanje (radio, TV, avdioknjige). Poleg tega vključujejo specializirano

snemalno ekipo, ki skrbi za logistično učinkovitost, kakovost zvoka in konsistentno produkcijo (Nisbett, 2023).

Pri obeh metodah je notranja motivacija govorcev običajno nizka. Govorci sodelujejo bodisi zaradi finančnega nadomestila bodisi zaradi profesionalnih interesov. V laboratorijih pod okriljem raziskovalnih institucij je govor najpogosteje vnaprej pripravljen ali vsaj voden in nadzorovan. V medijskih studiih je sicer mogoče spodbuditi situacije, kot so vodeni dialogi ali intervjuji, ki omogočajo delno spontani govor, vendar kontekst ostaja produkcijsko-formalen, kar vpliva na naravno jezikovno izbiro govorcev.

2.2 Snemanje na terenu

Snemanje na terenu označuje zbiranje govornih podatkov v naravnih okoljih, kjer govor spontano nastaja. Ta okolja obsegajo zasebne domove govorcev, šole, lokale, javne prireditve, pisarne, ulice in druge vsakdanje prostore, kjer se odvijajo avtentične komunikacijske situacije. Tak pristop predstavlja enega ključnih načinov za zajem dialektološkega in sociolingvistično raznoliknega gradiva, saj omogoča zbiranje posnetkov govorcev različnih demografskih značilnosti.

Ena izmed temeljnih prednosti snemanja na terenu je visoka stopnja avtentičnosti, saj govorci delujejo v domačih okoljih, kar zmanjšuje vpliv nadzorovanih pogojev na njihovo govorno obnašanje. Slednje omogoča zajem številnih jezikovnih variacij, vključno z različnimi narečji, sociolekti in registri, kar bistveno prispeva k raznolikosti in reprezentativnosti zbranega gradiva. Pomembna prednost je tudi možnost beleženja spontanе komunikacije z vsemi njenimi značilnostmi, kot so prekinitve, smeh, jezikovna preklapljanja in druge diskurzivne značilnosti, ki se pogosto ne pojavijo v pripravljenem ali nadzorovanem govoru. Z vidika raziskovanja je takšno gradivo neprecenljivo za jezikoslovne in kulturno-antropološke študije, saj omogoča vpogled v govorne prakse, kulturne norme in komunikacijske vzorce določene skupnosti.

Kljub številnim prednostim je snemanje na terenu izjemno zahtevno. Tehnični izzivi so povezani s posnetki, pogosto obremenjeni z motečimi zvoki, kot so prometni hrup, šumenje, hkratni ali nerazumljiv govor in drugi zunanji dejavniki, ki poslabšujejo kakovost posnetka. Potrebna je kakovostna mobilna snemalna oprema

in snemalec, ki ima izkušnje z delom v nepredvidljivih pogojih. Poleg tehničnih ovir je prisotna tudi potreba po ustrezni etični obravnavi – zlasti pri snemanju otrok, ranljivih skupin ali snemanjih v zasebnih prostorih je potrebna visoka stopnja občutljivosti in spoštovanja zasebnosti. Organizacija takšnih snemanj je pogosto zamudna, saj vključuje usklajevanje z govorci, dogovore o terminih, pripravo snemalne opreme in reševanje logističnih izzivov. Motivacija govorcev za sodelovanje v tovrstnem snemanju se lahko bistveno poveča, če so ti vključeni v širši raziskovalni ali dokumentacijski namen, ki ga prepoznajo kot osebno ali družbeno pomembnega. Osebni stik s snemalcem in občutek, da prispevajo k ohranjanju jezika ali podpori skupnosti, pogosto delujeta spodbudno. Nadomestila za sodelovanje so lahko simbolična ali praktična, na primer v obliki darilnih bonov, manjših daril ali možnosti (javno prepoznanega) sodelovanja v projektih.

2.3 Snemanje na daljavo

V razmerah, ko tradicionalne metode zbiranja govornih podatkov naletijo na logistične in finančne omejitve, se kot učinkoviti kažejo pristopi, ki temeljijo na oddaljenem in množičnem sodelovanju posameznikov. Gre za spletno zbiranje govornih prispevkov, ki omogoča neposredno snemanje in nalaganje posnetkov z različnih lokacij brez prisotnosti raziskovalca. Ta pristop se je posebej uveljavil v času pandemije COVID-19, ko so bile tradicionalne oblike zbiranja močno otežene, njegov nadaljnji razmah pa je omogočil pospešen razvoj digitalnih tehnologij.

Snemanje prek spletnih platform, mobilnih aplikacij ali z uporabo snemalnikov na pametnih telefonih se vse pogostejše uporablja tudi v kontekstu množičenja in občanske znanosti. Slednja spodbuja prostovoljno sodelovanje laikov pri zbiranju znanstveno uporabnih podatkov (Robinson idr., 2018), kar povečuje vključevanje javnosti v razvoj jezikovnih virov, zmanjšuje stroške in razbremenjuje raziskovalne ekipe.

Prednosti teh pristopov se kažejo predvsem v geografski, dialektološki in demografski pokritosti ter obsegu zbranih podatkov. Hkrati pa se pojavljajo izzivi glede nadzora kakovosti, etike, varstva osebnih podatkov in motivacije sodelujočih. Kakovost posnetkov je odvisna od opreme in akustike prostora, ki ju uporabnik sam izbere, pri čemer težave, kot so napake brskalnika ali prekinjena internetna povezava, lahko vodijo v izgubo posnetka. Zato je nujna podpora uporabnikom – prek jasnih navodil, možnosti testnega posnetka in tehnične pomoči.

Domače okolje spodbuja naravno komunikacijo, vendar so naloge pogosto vnaprej strukturirane, kar omejuje spontanost. Uporaba odprtih vprašanj in širših tematskih usmeritev lahko poveča avtentičnost govora. Priporoča se daljše trajanje posnetkov, saj se govorci sprostijo šele po določenem času, ko pozabijo na snemalni kontekst (Bizjak, 2025).

Mobilne in spletne aplikacije omogočajo snemanje kjerkoli in kadarkoli, kar povečuje skalabilnost projekta ter omogoča hiter in obsežen zajem podatkov. Igrifikacijski elementi in možnost vpogleda v rezultate povečujejo motivacijo in angažiranost sodelujočih. Nekatere aplikacije samodejno zajemajo tudi metapodatke, kot so geolokacija in vrsta naprave. Kljub temu razvoj in vzdrževanje takšnih aplikacij zahtevata tehnično znanje in sredstva, težave pa lahko povzročata tudi nezdržljivost med različnimi napravami ali sistemi. Uporabniška izkušnja mora biti skrbno testirana, saj v nasprotnem primeru sodelujoči hitro izgubijo motivacijo.

2.4 Games-with-a-Purpose

Games-with-a-Purpose (GWAPs) predstavljajo inovativno metodo zbiranja podatkov, pri kateri uporabniki skozi igrificirane naloge prispevajo govorne vire. V kontekstu jezikovnih tehnologij gre najpogosteje za igre, ki vključujejo naloge, kot so branje, opisovanje, ponavljanje ali sodelovanje v dialogih, pri čemer je osrednjega pomena igriva in tekmovalna uporabniška izkušnja. Ta pristop, ki se je začel uveljavljati po letu 2010, nagovarja predvsem širši krog uporabnikov, ki imajo dostop do računalnika in jih privlači igra (Lyding idr., 2022).

Igrifikacija s komponentami, kot so lestvice najboljših, napredovanje po nivojih in digitalne značke, pomembno prispeva k visoki stopnji vključenosti uporabnikov (Tondello idr., 2016). Ključna prednost tega pristopa je v tem, da se zbiranje podatkov odvija skozi dejavnost, ki je doživeta kot zabavna in ne kot delo, kar zmanjšuje občutek prisile. Udeležba je večinoma prostovoljna in ne zahteva finančnih spodbud, kar olajša organizacijo in izvedbo tovrstnih projektov. Pomembna prednost GWAP-ov je tudi možnost vključevanja zelo raznolikih jezikovnih nalog in različnih stopenj kompleksnosti, s čimer se lahko zajemajo raznoliki tipi govora.

Kljub tem prednostim pa GWAP-i niso brez pomembnih omejitev. Razvoj kakovostne igre zahteva visoke tehnične, oblikovalske in razvojne vire, zaradi česar so začetni stroški visoki, tudi če se uporabljajo obstoječe knjižnice, kot je PythonGameLibraries (Lyding idr., 2022). Dolgoročno ohranjanje zanimanja uporabnikov je še posebej zahtevno, saj tovrstne igre neposredno tekmujejo z uveljavljenimi komercialnimi videoigrami, ki razpolagajo z znatno večjimi proračuni in profesionalnimi ekipami (Tondello idr., 2016). Poleg tega lahko prekomerna osredotočenost na igrivost negativno vpliva na kakovost zbranih podatkov – uporabniki lahko zaradi želje po napredovanju hitijo z reševanjem nalog, kar povečuje tveganje za površne ali nekonsistentne odzive. Demografska struktura udeležencev je še ena pomembna slabost: uporabniki, ki se odločajo za sodelovanje v tovrstnih igrah, običajno niso reprezentativni za splošno populacijo, saj pogosto prevladujejo mlajši in digitalno večji posamezniki.

Z vidika motivacije udeležencev se GWAP-i izkažejo za učinkovite predvsem tedaj, ko so ustrezno zasnovani – ko vključujejo elemente tekmovalnosti, napredovanja ter družbenega priznanja. V takih primerih je motivacija uporabnikov visoka in finančne spodbude niso nujne. Vendar pa dolgoročna vzdržnost projekta zahteva premišljeno načrtovanje, prilagajanje potrebam uporabnikov in kontinuirano optimizacijo uporabniške izkušnje (Lyding idr., 2022).

Za zbiranje spontanega govora so GWAP-i srednje ustrezni. Igre sicer lahko vzpodbudijo rabo avtentičnega govora, vendar so v praksi pogosto zgrajene okoli bolj ali manj strukturiranih nalog, ki omejujejo stopnjo spontane tvorbe. Večjo mero spontanosti omogočajo igre, ki vključujejo elemente interakcije ali ustvarjanja vsebine, denimo v obliki prostega govora, interpretacije slik ali sodelovalnih nalog, kjer uporabniki soustvarjajo zvočne vsebine.

2.5 Collect4NLP

Collect4NLP predstavlja modularni pristop k zbiranju podatkov in temelji na reševanju nalog za učenje tujih jezikov. Ključna zamisel, ki stoji za tem pristopom, je uporaba rešenih nalog, ustvarjenih s pomočjo jezikovnih baz (NLP), kot »stranskega produkta« učenega procesa. Tako pridobljeni podatki posledično izboljšujejo iste baze podatkov, kar pomeni, da učni proces hkrati služi raziskovalnim ciljem. Temeljna zasnova sistema izhaja iz koncepta »modrosti množic«, ki

predpostavlja, da so posamezni odgovori učečih se lahko šumni, a njihova akumulacija omogoča statistično zanesljive izide (Lyding idr., 2022).

Ena izmed glavnih prednosti Collect4NLP je možnost hitre postavitve spletne platforme, saj temelji na že obstoječih rešitvah. Sistem podpira večjezičnost in omogoča razširljivost, kar omogoča uporabo v raznolikih jezikovnih okoljih. Poleg tega platforma nudi podporo za strukturirane naloge, kot so naloge branja, odgovarjanja in interpretacije, kar je posebej pomembno za jezikoslovne raziskave. Njena tehnična zasnova omogoča dobro integracijo z raziskovalnimi postopki, med drugim z zajemom metapodatkov, anonimizacijo podatkov in možnostjo izvoza zbranih podatkov.

Kljub tem prednostim je za učinkovito delovanje Collect4NLP potrebno osnovno tehnično znanje za konfiguracijo komponent, kar lahko predstavlja oviro za uporabnike brez ustreznega predznanja. Možnosti za oblikovno prilagoditev vmesnika so omejene, če uporabnik ne razpolaga z dodatnim programerskim znanjem. Poleg tega je platforma zasnovana predvsem za delo s strukturiranimi nalogami, zaradi česar je manj primerna za zajem spontanega govora. Pomembna omejitev se nanaša tudi na heterogenost jezikovnega znanja med uporabniki, ki se učijo tujih jezikov, zaradi česar je nujna uporaba robustnih mehanizmov za profiliranje in uteževanje njihovih odgovorov (Lyding idr., 2022).

Motivacija udeležencev temelji zlasti na notranji motivaciji za izboljšanje lastnih jezikovnih spretnosti. Ta je izrazitejša, kadar je učna izkušnja hkrati privlačna in učinkovita. Ker platforma sama po sebi ne vključuje elementov igrifikacije, se kot smiselno dopolnilo priporočajo dodatni motivacijski prijemi, kot so pogoste povratne informacije in personalizirana sporočila. Ustreznost platforme za zajem spontane govorjene produkcije je nizka do srednja. Čeprav je sistem prvenstveno zasnovan za strukturiran zajem, je z dodatnim načrtovanjem mogoče zasnovati tudi naloge, ki vključujejo bolj odprte oblike govora (Lyding idr., 2022).

2.6 Pridobivanje govornih virov preko množičenja

Množičenje (angl. *crowdsourcing*) je sodoben pristop k zbiranju govornih virov, ki temelji na vključevanju velikega števila sodelujočih prek spletnih in mobilnih tehnologij. Značilno je, da uporabniki naloge izvajajo samostojno in brez

neposrednega nadzora ali prisotnosti raziskovalca. Tipične naloge vključujejo snemanje govora, transkribiranje ter validiranje ali označevanje govornih posnetkov. Izraz množičenje je leta 2006 uvedel Jeff Howe (Kaufmann idr., 2011) in ga opredelil kot zunanje izvajanje, tj. prenos funkcij ali nalog, ki jih je tradicionalno opravljal imenovani zastopnik, na neopredeljeno mrežo številčnejših delavcev, in sicer preko poziva, običajno objavljenega na internetu. V okviru jezikoslovnih raziskav se je množično zbiranje podatkov uveljavilo kot učinkovit odziv na pomanjkanje govornih virov, visoke stroške tradicionalnih metod zbiranja ter omejitve pri avtomatizaciji procesov (Eskénazi idr., 2013; Lyding idr., 2022).

Zgodovinski razvoj množičenja v jezikoslovju kaže na premik iz nadzorovanih laboratorijskih okolij in telefonskih sistemov v digitalno okolje, kjer spletne platforme omogočajo avtomatiziran in razširljiv zajem zvočnih posnetkov. Pomemben mejnik v tej preobrazbi je predstavljal vzpon platform, kot je Amazon Mechanical Turk, ki so omogočile dostop do širokega nabora uporabnikov in enostavno porazdelitev nalog. Nadaljnji razvoj spletnih in mobilnih rešitev, zlasti tistih, ki vključujejo elemente igrifikacije, je dodatno povečal možnosti prilagoditve nalog različnim ciljnim skupinam in s tem večjo participacijo uporabnikov. Množičenje omogoča hitro in stroškovno učinkovito zbiranje velikih količin govornih podatkov, kar je ključno za podporo manjšim in z viri manj podprtim jezikom. Dodatna prednost je fleksibilna zasnova nalog, ki se lahko prilagodi različnim ravnam zahtevnosti. Z izločitvijo posrednikov in neposrednim vključevanjem posameznikov lahko množično zbiranje podpira tudi bolj etične modele sodelovanja in prispeva k razvoju lokalnih skupnosti, zlasti v državah v razvoju (Eskénazi idr., 2013; Lyding idr., 2022).

Kljub številnim prednostim pa množično zbiranje govornih podatkov odpira tudi vrsto izzivov. Eden najpomembnejših je zagotavljanje kakovosti zbranih podatkov, saj se uporabniki razlikujejo po tehnični opremljenosti, spretnostih in razumevanju nalog, natančnosti in hitrosti izvedbe, nekateri posamezniki pa celo namerno oddajajo neustrezne rešitve. Da bi to preprečili, so potrebni robustni mehanizmi za validacijo, vključno z avtomatiziranim zaznavanjem goljufanja. Nadalje obstaja tveganje, da preveč podrobna navodila zmanjšajo avtentičnost odzivov, medtem ko nejasne usmeritve lahko vodijo do odstopanj ali opustitve sodelovanja (Eskénazi idr., 2013).

Z vidika etike in pravne skladnosti je potrebno skrbno urejanje vseh pogojev zbiranja, hrambe in obdelave podatkov, pri čemer številne platforme sodelujočim za njihovo delo nudijo prenizko plačilo, imajo nerazumna pričakovanja glede delovne obremenitve, pri čemer ne poskrbijo niti za osnovno socialno varstvo (Lindén idr., 2022; Poesio idr., 2017).

Motivacija udeležencev v množičenju je raznolika in vključuje altruizem, željo po učenju, občutek kompetentnosti, tekmovalnost, igrificacijske elemente (npr. točke, značke) ter finančne nagrade. Učinkovitost motivacije je višja, kadar platforme vključujejo povratne informacije, lestvice najboljših ali možnost osebne rasti (Garaus idr., 2024; Poesio idr., 2017). Ustreznost množičenja za zbiranje spontanega govora ocenjujemo kot srednjo, ker četudi se podatki zbirajo v neformalnem okolju, so naloge pogosto strukturirane in vnaprej določene.

2.7 Pridobivanje govornih virov preko občanske znanosti

Občanska znanost (ang. *citizen science*) označuje odprto obliko znanstvenega sodelovanja, v katerem posamezniki ali organizacije sodelujejo v različnih fazah znanstvenega procesa, pri čemer raziskovalnega dela ne opravljajo kot del svoje zaposlitve (Vohland, 2021). Čeprav pojem ni nov – ljubiteljski amaterji so k znanosti prispevali že v 17. stoletju (Miller-Rushing v Bonney idr., 2016) – je večjo globalno odmevnost dobil šele v zadnjih dveh desetletjih, predvsem zaradi razvoja družbe, temelječe na znanju, digitalnih tehnologij in razširjene rabe mobilnih naprav (Gershenfeld, 2011). Vključevanje laične javnosti v znanost presega tradicionalno enosmerno posredovanje znanja in krepi interaktivnost, izmenjavo izkustvenega in znanstvenega vedenja ter poudarja pomen »znanosti od spodaj navzgor« (Mlinar, 2021).

V kontekstu zbiranja govornih virov občanska znanost pomeni, da posamezniki iz različnih okolij prostovoljno prispevajo svoj govor. Njihovo sodelovanje se lahko nanaša na snemanje in oddajo posnetkov lastnega govora ali govora drugih govorcev, validacijo govornih posnetkov, transkribiranje in označevanje govornih posnetkov, promocijo zbiranja v lokalnih skupnostih ali celo soustvarjanje raziskovalnih vprašanj in komentiranje oz. posredovanje rezultatov.

Pristop občanske znanosti k zbiranju govornih virov ima več pomembnih prednosti, ki se nanašajo tako na jezikovno kot na družbeno dimenzijo zbiranja podatkov. Eden ključnih vidikov je vključevanje govorcev iz različnih geografskih, starostnih, socialnih in jezikovnih okolij, kar bistveno prispeva k večji jezikovni in socialni raznolikosti zbranih govornih podatkov. Tak pristop omogoča zajem avtentičnih govornih variant in govorcev, ki jih institucionalni pristopi pogosto spregledajo, zlasti kadar gre za regionalne narečne značilnosti, manjšinske jezikovne skupnosti ali govorce z nizkim družbenoekonomskim statusom. S tem pristopom projekti pridobijo tudi na družbeni legitimnosti, saj vključevanje širše javnosti praviloma povečuje zaupanje lokalnih skupnosti ter zagotavlja podporo pri širjenju informacij in aktivnostih zbiranja podatkov. Dodatno se s tem krepi tudi motivacija sodelujočih govorcev, ki se pogosto zavedajo pomena lastnega prispevka k razvoju jezika, kar posledično vodi do večje angažiranosti in pripravljenosti na sodelovanje. Na dolgi rok se lahko s ponavljajočim sodelovanjem oblikuje stabilna baza uporabnikov, ki podpira trajnostne cilje zbiranja govornih virov in omogoča dolgoročneje nadgradnje zbirke.

Kljub omenjenim prednostim pa pristop občanske znanosti vključuje tudi številne izzive in omejitve. Ena izmed osrednjih težav je potreba po premišljeni in strokovno zasnovani kampanji za mobilizacijo udeležencev, saj je uspešnost zbiranja v veliki meri odvisna od jasne komunikacije namenov in koristi projekta. Vzporedno s tem se pogosto pojavi tudi lokalna odvisnost, kjer učinkovitost zbiranja temelji na posameznih koordinatorjih, lokalnih promotorjih ali mrežah, kar lahko omeji širjenje pobud na druge regije ali jezikovne skupine. Sodelujejo večinoma posamezniki, ki so že digitalno pisмени in vnaprej motivirani za sodelovanje, kar vodi do neuravnotežene zastopanosti v podatkovnih zbirkah. To pomeni, da določene demografske ali jezikovne skupine ostajajo slabo ali sploh nezastopane. Poleg tega je za tak pristop ključno tudi sistematično načrtovanje etičnih protokolov za zaščito sodelujočih ter zagotovitev tehnične podpore, ki omogoča nemoteno sodelovanje tudi manj izkušenim uporabnikom.

Ustreznost takšnega pristopa v kontekstu zbiranja spontane govorne slovenščine je visoka, še posebej če je zbiranje usmerjeno v pristne osebne pripovedi in če udeleženci sami izbirajo tematike in kontekst snemanja. Pomembno je, da imajo govorci možnost snemanja v znanem okolju, s čimer se poveča avtentičnost izraza.

3 Primerjalna analiza različnih pristopov

V nadaljevanju primerjamo opisane tehnike snemanja govornih virov po naslednjih kriterijih:

- kakovost zvoka,
- stroški (snemalna oprema, stroški dela in vzdrževanja, infrastruktura, razvojni stroški),
- razširljivost (število govorcev, geografski doseg) ter
- avtentičnost, spontanost govora.

Med posameznimi pristopi k zbiranju govornih virov ni enoznačno najustrežnejšega. Visoka kakovost zvoka je praviloma dosegljiva le ob višjih stroških, omejenem številu govorcev in nadzorovanih snemalnih pogojih. Nasprotno pa pristopi, ki temeljijo na sodelovanju na daljavo, omogočajo večjo razširljivost in pogosto tudi večjo spontanost govora, a hkrati prinašajo večje tveganje za nižjo tehnično kakovost posnetkov ter nepopolno dokumentacijo metapodatkov. Pri izbiri ustreznega pristopa je zato smiselno upoštevati razpoložljive vire, ciljno strukturo govornih podatkov ter namen projekta ali pobude.

Tabela 2

Tehnika snemanja	Kakovost zvoka	Stroški	Razširljivost	Spontanost govora
Snemanje v laboratoriju ali studiu	Visoka	Visoki	Nizka do srednja	Nizka do srednja
Snemanje na terenu	Srednja	Visoki	Nizka	Visoka
Snemanje na daljavo	Nizka do srednja	Nizki	Visoka	Nizka do srednja
GWAPs	Srednja	Visoki	Visoka	Nizka do srednja
Collect4NLP	Srednja	Srednji	Srednja	Nizka do srednja
Množičenje	Nizka do srednja	Nizki do srednji	Visoka	Nizka do srednja
Občanska znanost	Nizka do srednja	Nizki	Srednja	Visoka

4 Tuji primeri dobrih praks

Občanska humanistika⁸ pomeni razširitev občanske znanosti na področja, kot so jezikoslovje, književnost in arheologija, kjer javnost sodeluje pri raziskavah preko nalog, kot so transkribiranje, označevanje in interpretacija kulturnih podatkov. Ti projekti pogosto potekajo v sodelovanju z javnimi ustanovami ter temeljijo na lokalnem znanju skupnosti. Njihove oblike so raznolike – od digitalnih do terenskih aktivnosti – ter segajo od projektov, ki jih vodijo raziskovalci do projektov »od spodaj navzgor«, katerih pobudniki in koordinatorji so člani lokalne skupnosti. Znotraj občanske humanistike se v zadnjem času kot posebna veja razvija občansko jezikoslovje⁹, kjer prostovoljci prispevajo k razvoju jezikovnih virov in ohranjanju jezikovne raznolikosti kot kulturne dediščine (Heinisch idr., 2021).

Med tovrstne projekte lahko uvrstimo *CorCenCC* (The National Corpus of Contemporary Welsh), nacionalni projekt zbiranja in anotiranja sodobne valižanščine (Knight idr., 2020; Neale idr., 2017), *Lahjoita pubetta* (Donate speech), ambiciozna finska kampanja za zbiranje digitalnih govornih podatkov za podporo razvoju jezikovnih tehnologij (Lindén idr., 2022) in *Teen Speak*, estonski projekt, ki je uporabil participatorni pristop občanskega jezikoslovja za oblikovanje prvega korpusa estonskega govora najstnikov in komuniciranja preko sporočil (Koreinik idr., 2024). Ti in ostali projekti, kot so avstrijski IamDiÖ s področja dokumentiranja

⁸ Citizen humanities (angl.)

⁹ Citizen linguistics (angl.)

in raziskovanja sodobne rabe narečij (Heinisch, 2022), Sprekend Nederland za zbiranje govorne nizozemščine (Van Leeuwen idr., 2016), *What's up, Switzerland?*, večjezični korpus WhatsApp sporočil (Ueberwasser in Stark, 2017), Gschmöis aplikacija za pridobivanje baze švicarskonemškega narečja, projekt Schnëssen za govorno luksemburščino (Entringer idr., 2021) ter platforme, ki eksplicitno podpirajo projekte občanske znanosti, kot sta LanguageARC (Fiumara idr., 2020) in Mozilla Common Voice¹⁰ dokazujejo, da množično zbiranje govornih podatkov na daljavo ni le izvedljivo, temveč tudi metodološko produktivno.¹¹

Skupne ugotovitve omenjenih projektov poudarjajo pomen zasnove intuitivnih in uporabniku prijaznih vmesnikov za oddajo posnetkov (npr. prek mobilnih aplikacij ali spletnih platform), ki omogočajo čim bolj preprosto in minimalno interakcijo z uporabnikom (npr. en sam gumb za začetek in zaključek snemanja). Za uspešno vključevanje uporabnikov so se kot ključne izkazale skrbno načrtovane promocijske kampanje, ki z uporabo humorja in podporo uveljavljenih medijev učinkovito nagovarjajo skupnost ter poudarjajo družbeno vrednost sodelovanja govorcev pri razvoju jezikovnih tehnologij in ohranjanju jezikovne dediščine.

Primeri uveljavljenih tujih praks potrjujejo, da je ob skrbno zasnovani infrastrukturi, zagotovljeni etični in pravni transparentnosti (vključno z ustreznimi soglasji govorcev), jasnih tehničnih navodilih, nalogah za spodbujanje nepripravljenega govora ter ciljno usmerjeni promociji mogoče uspešno vzpostaviti obsežne javno dostopne zbirke spontanega govornega jezika. Kljub raznolikim modelom – od učencev kot jezikovnih ambasadorjev (*Teen Speak*) do množičnih nacionalnih kampanj (*Lahjoita puhetta*) – se vsi projekti srečujejo z izzivi digitalne izključenosti, demografske neuravnoteženosti in trajnostnega sodelovanja. Obenem pa je v večini teh pobud še vedno premalo raziskana ključna komponenta: motivacijski dejavniki, ki vplivajo na pripravljenost posameznikov za sodelovanje, zlasti v primeru spontanega govora.

¹⁰ <https://commonvoice.mozilla.org/en>

¹¹ Nekaj dodatnih iniciativ in projektov: *Lingua Libre* (Hutin in Allasonnière-Tang, 2022), *ABCs of Dialect* (Dorn idr., 2024), zbiranje narečnega gradiva v regiji Palóc (Szabó idr., 2025).

5 Priporočila za pridobivanje govornih virov

Priporočila v nadaljevanju so namenjena raziskovalcem in drugim deležnikom, ki načrtujejo zbiranje govornih posnetkov v slovenskem jeziku, zlasti v primerih snemanja na daljavo. Temeljijo na primerjalni analizi različnih metod pridobivanja govornih virov ter na izbranih primerih tujih praks, ki so se izkazale kot učinkovite in izvedljive. Osredotočajo se na posamezne praktične vidike, kot so izbira govorcev, spodbujanje spontanega govora, tehnična izvedba snemanja ter validacija kakovosti.

1. Načrtovanje in opredelitev ciljev

- Opredelite namen zbiranja govornih virov in njihove uporabnike.
- Določite žanr in vrsto govora (npr. spontani/pripravljeni govor), ki ga želite pridobiti.
- Jasno opredelite različne ciljne skupine govorcev (spol, starost, izobrazba, regija, narečna pripadnost, prvi jezik). Poleg maternih vključite tudi govorce slovenščine kot drugega ali tujega jezika ter govorce iz zamejstva in izseljence.
- Razmislite o možnostih pridobivanja multimodalnih posnetkov (avdio in video).

2. Število govorcev na posnetku

- Odločite se lahko za snemanje monologov, dialogov ali posnetkov z več govorcev.
- Pri dialogih in multilogih je pomembno, da si prizadevate zagotoviti dobro ločljivost zvočnih kanalov (ločeni mikrofoni ali sinhronizacija signalov).

3. Raznolikost in uravnoteženost

- Zagotovite demografsko in geografsko uravnoteženost govorcev (vključite različne statistične regije in narečne skupine).
- Zbirajte gradiva iz različnih komunikacijskih okoliščin (javne, zasebne, formalne, neformalne okoliščine) in vključite širok spekter tem: od vsakdanjih pogovorov do strokovnih diskusij (odvisno od vašega namena in ciljev).

4. Spontanost in avtentičnost

- Spodbujajte rabo spontanega govora z uporabo odprtih nalog, ki vključujejo pripovedovanje, proste opise, sproščene dialoge ipd.
- Spodbujajte snemanje v domačem, zasebnem okolju, ki na govorce vpliva sproščujoče, s čimer povečate možnost bolj avtentičnega govora.
- Spodbujajte oddajo nekoliko daljših posnetkov, saj daljši posnetki (vsaj 20 minut) povečujejo možnost, da se govorci bolj sprostijo.
- Če oblikujete naloge (na primer pri množičenju), naj bodo te kratke, razdeljene na podenote, vizualno privlačne in poimenovane z jasnim, privlačnim naslovom.
- Omogočite sodelujočim možnost izbire med nalogami ali oblikovanje govornih prispevkov po lastni izbiri (avtonomnost je eden od najpomembnejših motivacijskih dejavnikov v sklopu občanske znanosti).

5. Tehnična izvedba snemanja in zbiranje metapodatkov

- Priporočene naprave: pametni telefoni, zunanji mikrofoni, snemalniki (npr. Zoom H2n).
- Priporočene tehnične lastnosti: format WAV, 44.1 kHz, 16 bit, mono.
- Izogibajte se šumom, preglasnemu ozadju, odmevom; snemajte v mirnem okolju.

- Snemalci in/ali govorci naj čim bolj natančno popišejo okoliščine snemanja (snemalno opremo, prostor snemanja itd.) in podatke o govorcih (metapodatke, kot so spol, starost, kraj rojstva, kraj bivanja itd.)¹²

6. Usposabljanje in podpora sodelujočim

- Ponudite jasna, enostavna in vizualno podprta navodila (video, delavnice, *FAQ*).
- Vključite vzorčne naloge in možnost testnega posnetka.
- Vzpostavite pomoč za uporabnike (*helpdesk*, e-naslov, spletna skupnost).
- Dvosmerna komunikacijska podpora sodelujočim naj bo čim bolj ažurna, spoštljiva in transparentna.

7. Validacija in nadzor kakovosti

- Odločite se lahko za eno izmed naslednjih možnosti oz. te medsebojno kombinirate:
 - predhodno preverjanje govorcev (na osnovi že oddanih posnetkov ali uspešnosti pri reševanju vzorčnih primerov za aktualno nalogo),
 - medprocesni nadzor (npr. primerjava rezultatov z zlatim standardom, ki ga oblikujete vnaprej s pomočjo strokovnjakov) ter
 - validacijo po zaključenem snemanju (npr. z medsebojno primerjavo sodelujočih ali navzkrižno validacijo med govorci samimi) (Eskénazi idr., 2013).
- Omogočite govorcem vpogled v rezultate validacije in možnost izboljšav.

8. Pravna in etična podlaga

- Natančno opredelite namen zbiranja govornih virov ter predvidene končne uporabnike. Jasno določite tudi obseg podatkov, ki bodo zbrani, načine njihove uporabe in pogoje za njihovo morebitno javno objavo.
- Od govorcev pridobite osebno privolitev za snemanje, obdelavo osebnih podatkov in njihovo nadaljnjo rabo. Spoštujte načela GDPR (pravica do pozabe, transparentnost, anonimizacija) in uredite avtorske pravice.

¹² Za natančna priporočila gl. *Smernice za zbiranje podatkov za govorne vire* (Verdonik in Gostenčnik, 2024).

9. Finančno nadomestilo

- Če sodelujoče plačate, jih plačajte pošteno (etičnost) in jasno komunicirajte merila za sprejem/zavrnitev.
- Če plačila ni, poskrbite za privlačno motivacijo (npr. igrifikacija, priznanja, dostop do rezultatov in njihova javna objava, prispevek k znanosti, pridobivanje novega znanja in spretnosti, simbolične nagrade, vključenost v skupnost itd.).

10. Promocija in motivacija

- Uporabite ciljno usmerjene medijske kampanje (TV, radio, družbena omrežja, plakati).
- Vključite humor, lokalne teme in poudarite družbeni pomen sodelovanja.
- Raziščite, kateri motivacijski dejavniki v največji meri pozitivno vplivajo na pripravljenost za snemanje in oddajo govornih posnetkov glede na izbrano ciljno skupino govorcev.

11. Odprta dostopnost in trajnostna uporaba

- Korpusi naj bodo dostopni v odprti obliki (npr. licenca CC BY), opremljeni z metapodatki, soglasji in tehničnimi opisi. Priporočljiva je objava v repozitoriju Clarin, raziskovalni infrastrukturi za jezikovne vire in tehnologije.¹³
- Uporaba naj bo dovoljena za raziskovalne, pedagoške in komercialne namene (če to dovoljuje licenca).
- Predvidite dolgoročno shranjevanje, varnostne kopije in nadgradljivost sistema.

12. Povezovanje z nacionalnimi cilji

- Če je mogoče, projekt umeščajte v širši kontekst nacionalnega razvoja jezikovnih tehnologij, digitalizacije in kulturne dediščine.

¹³ <https://www.clarin.si/info/o-projektu/>

Literatura

- Arhar Holdt, Š., Logar, N., Pori, E., & Kosem, I. (2021). "Game of Words": Play the Game, Clean the Database. *EURALEX XIX*. Pridobljeno s XIX-Euralex-Proceedings-Lexicography-for-Inclusion.pdf
- Awan, S. N., Shaikh, M. A., Awan, J. A., Abdalla, I., Lim, K. O., & Misono, S. (2023). Smartphone recordings are comparable to "gold standard" recordings for acoustic measurements of voice. *Journal of voice*.
- Bizjak, A. (2025). Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 13(1), 58-103.
- Bonney, R., Phillips, T. B., Ballard, H. L., & Enck, J. W. (2016). Can citizen science enhance public understanding of science?. *Public understanding of science*, 25(1), 2–16. doi: 10.1177/0963662515607406
- Čibej, J., Robida, N., & Krek, S. (2024). Nadgradnja Digitalne slovarske baze za slovenščino in Slovenskega oblikoslovnega leksikona Sloleks s podatki o govorni slovenščini: načrti in cilji. V M. Krajnc Ivič (ur.), *Stanje in perspektive uporabe govornih virov v raziskavah govora* (str. 27–39). Univerza v Mariboru, Univerzitetna založba. <https://press.um.si/index.php/ump/catalog/book/898/chapter/46>
- Dorn, A., Stocker, R., & Stöckle, P. (2024). Old dialect words through the ages—the ABCs of dialect project. *ARPHA Proceedings*, 6, 213-216.
- Entringer, N., Gilles, P., Martin, S., & Purschke, C. (2021). Schnëssen. Surveying language dynamics in Luxembourgish with a mobile research app. *Linguistics Vanguard*, 7(s1), 20190031.
- Eskenazi, M., Levow, G. A., Meng, H., Parent, G., & Suendermann, D. (2013). *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. John Wiley & Sons.
- Fiumara, J., Cieri, C., Wright, J., & Liberman, M. (2020). LanguageARC: Developing language resources through citizen linguistics. In *Proceedings of the LREC 2020 Workshop Citizen Linguistics in Language Resource Development (CLLRD 2020)*.
- Garaus, C., Garaus, M., & Wagner, U. M. (2024). Getting users involved in idea crowdsourcing initiatives: An experimental approach to stimulate intrinsic motivation and intention to submit. *IEEE Transactions on Engineering Management*, 71, 3700–3711. Pridobljeno s <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10387737>
- Gershenfeld, N. (2011). Physics of the future: how science will shape human destiny and our daily lives by the year 2100. *Physics Today*, 64(10), 56–56.
- Heinisch, B., Oswald, K., Weißpflug, M., Shuttleworth, S., Belknap, G. (2021). Citizen Humanities. In: Vohland, K., et al. *The Science of Citizen Science*. Springer, Cham. https://doi.org/10.1007/978-3-030-58278-4_6
- Heinisch, B. (2022). The Influence of Intrinsic and Extrinsic Motivation on the Creation of Language Resources in a Citizen Linguistics Project about Lexicography. In *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022* (pp. 58-63).
- Hutin, M., & Allassonnière-Tang, M. (2022). Crowd-sourcing for less-resourced languages: Lingua libre for Polish. In *1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)*.

- Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. worker motivation in crowdsourcing—a study on mechanical turk. https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/45694/file/More_than_fun_and_money_Worker_Motivation_in_Crowd.pdf
- Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., & Thomas, E. M. (2020). The national corpus of contemporary Welsh: Project report| Y corpws cenedlaethol Cymraeg cyfoes: adroddiad y prosiect. Pridobljeno s <https://arxiv.org/abs/2010.05542>
- Koreinik, K., Mandel, A., Pilvik, M. L., Praakli, K., & Vihman, V. A. (2024). Outsourcing teenage language: A participatory approach for exploring speech and text messaging. *Linguistics Vanguard*, 9(s4), 389–398.
- Labov, W. (1973). *Sociolinguistic patterns* (No. 4). University of Pennsylvania press.
- Lindén, K., Jauhainen, T., Lennes, M., Kurimo, M., Rossi, A., Kurki, T., & Pitkänen, O. (2022). Donate Speech: Collecting and Sharing a Large-Scale Speech Database for Social Sciences, Humanities and Artificial Intelligence Research and Innovation. V CLARIN: the infrastructure for language resources (Digital Linguistics; Vol. 1). De Gruyter. doi: 10.1515/9783110767377-019
- Lyding, V., Nicolas, L., & König, A. (2022). About the applicability of combining implicit crowdsourcing and language learning for the collection of NLP datasets. V *Proceedings of the 2nd Workshop on Novel Incentives in Data Collection from People: models, implementations, challenges and results within LREC 2022* (str. 46–57). Pridobljeno s <https://aclanthology.org/2022.nidcp-1.8.pdf>
- Mlinar, Z. (2021). Kaj nam prinašata koncept in gibanje občanska znanost/Citizen Science? Uveljavljanje raziskovanja kot sestavine vsakdanjega življenja. *Casopis za Kritiko Znanosti, Domislijo in Novo Antropologijo (Journal for the Critique of Science, Imagination & New Anthropology)*, 49(282).
- Neale, S., Spasic, I., Needs, J., Watkins, G., Morris, S., Fitzpatrick, T., ... & Knight, D. (2017). The CorCenCC crowdsourcing app: A bespoke tool for the user-driven creation of the national corpus of contemporary Welsh. V *Corpus Linguistics Conference, Birmingham*. Pridobljeno s <https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper273.pdf>
- Nisbett, A. (2003). *Sound Studio: Audio Techniques for Radio, Television, Film and Recording*. Routledge.
- Poesio, M., Chamberlain, J., & Kruschwitz, U. (2017). Crowdsourcing. *Handbook of linguistic annotation* (str. 277–295).
- Robinson, L. D., Cawthray, J. L., West, S. E., Bonn, A., & Ansine, J. (2018). Ten principles of citizen science. In *Citizen science: Innovation in open science, society and policy* (pp. 27–40). UCL Press.
- Rutten, M., Minkman, E., & van der Sanden, M. (2017). How to get and keep citizens involved in mobile crowd sensing for water management? A review of key success factors and motivational aspects. *Wiley Interdisciplinary Reviews: Water*, 4(4), e1218. Pridobljeno s <https://wires.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/wat2.1218>
- Szabó, G., Fazakas, N., Kocsis, Z., Krizsai, F., & Vargha, F. S. (2025). Researching dialects with high school students: a citizen science approach. *Linguistics Vanguard*, (0).
- Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016). The gamification user types hexad scale. V *Proceedings of the 2016 annual symposium on computer-human interaction in play* (str. 229–243). Pridobljeno s <https://dl.acm.org/doi/pdf/10.1145/2967934.2968082>
- Ueberwasser, S., & Stark, E. (2017). What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik online*, 84(5).
- Van Leeuwen, D. A., Hinsken, F., Martinovic, B., Van Hessen, A., Grondelaers, S., & Orr, R. (2016). Sprekend Nederland: A heterogeneous speech data collection. *Computational Linguistics in the Netherlands Journal*, 6, 21–38. Pridobljeno s <https://www.clinjournal.org/clinj/article/view/62/55>
- Verdonik, D., & Gostenčnik, J. (2024). *Smernice za zbiranje podatkov za govorne vire*. Univerza, Fakulteta za elektrotehniko, računalništvo in informatiko. <https://mezzanine.um.si/rezultati/#tehn%C4%8Dne-smernice>

- Verdonik, D., & Maučec, M. S. (2017). A speech corpus as a source of lexical information. *International Journal of Lexicography*, 30(2), 143–166.
- Verdonik, D. (2008). Označevanje vrste diskurzivnih označevalcev. V T. Erjavec in J. Žganec Gros (ur.), *Zbornik šeste konference Jezikovne tehnologije, 16.–17. oktober 2008, Ljubljana* (Vol. 12, str. 25). Pridobljeno s <https://nl.ijs.si/isjt08/IS-LTC08-Proceedings.pdf#page=33>
- Vohland, K. (2021). The Science of Citizen Science.
- Wertheim, S. (2006). Cleaning up for company: Using participant roles to understand fieldworker effect. *Language in Society*, 35(5), 707–727.
- Zhang, C., Jepson, K., & Chuang, Y. Y. (2024). Investigating differences in lab-quality and remote recording methods with dynamic acoustic measures. *arXiv preprint arXiv:2404.17022*.

PRIDOBIVANJE GOVORNIH VIROV: PREDNOSTI IN POMANJKLJIVOSTI RAZLIČNIH PRISTOPOV

ANDREJA BIZJAK

Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko, Maribor,
Slovenija
andreja.bizjak1@um.si

Elaborat sistematično obravnava različne pristope k pridobivanju govornih virov za slovenščino – od laboratorijskih in terenskih snemanj do sodobnejših spletnih oblik, kot so GWAPs, Collect4NLP, množičenje in občanska znanost. S primerjalno analizo, ki vključuje kriterije kakovosti zvoka, stroškov, razširljivosti in spontanosti govora, razkriva prednosti, omejitve in izvedbeno zahtevnost posameznih pristopov. Vključuje praktična priporočila za raziskovalce in razvijalce govornih tehnologij, ki se soočajo z izzivi snemanja na daljavo. Z združevanjem globalnih trendov in tujih praks z razmerami v slovenskem prostoru prispeva k razvoju digitalnih govornih virov ter krepitvi jezikovnih tehnologij za manj podprte jezike.

DOI

[https://doi.org/
10.18690/um.feri.11.2025](https://doi.org/10.18690/um.feri.11.2025)

ISBN

978-961-299-093-0

Ključne besede:

pridobivanje govornih
virov,
snemanje na daljavo,
spontani govor,
govorjena slovenščina,
digitalna humanistika



Univerzitetna založba
Univerze v Mariboru

DOI
[https://doi.org/
10.18690/um.feri.11.2025](https://doi.org/10.18690/um.feri.11.2025)

ISBN
978-961-299-093-0

Keywords:

speech resource collection,
remote speech recording,
spontaneous speech,
spoken Slovenian,
digital humanities

COLLECTION OF SPEECH RESOURCES: ADVANTAGES AND LIMITATIONS OF DIFFERENT APPROACHES

ANDREJA BIZJAK

University of Maribor, Faculty of Electrical Engineering and Computer Science,
Maribor, Slovenia
andreja.bizjak1@um.si

This study systematically examines various approaches to collecting speech resources for Slovenian – from laboratory and field recordings to more contemporary online methods such as GWAPs, Collect4NLP, crowdsourcing, and citizen science. Through a comparative analysis that considers sound quality, cost, scalability, and speech spontaneity, it reveals the advantages, limitations, and implementation challenges of each approach. The work provides practical recommendations for researchers and developers of speech technologies who face the challenges of remote recording. By integrating global trends and international practices with the Slovenian context, it contributes to the development of digital speech resources and the advancement of language technologies for under-resourced languages.



University of Maribor Press



Univerza v Mariboru

Fakulteta za elektrotehniko,
računalništvo in informatiko