Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini

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

Govorni viri ustrezne kakovosti so za razvoj govornih tehnologij in raziskovanje govorjenega jezika ključnega pomena, a jih na področju spontano tvorjene govorjene slovenščine zaradi zahtevnosti zbiranja še vedno primanjkuje. Izgradnja govornih korpusov in podatkovnih baz je stroškovno zahtevna, zato raziskovalci vse pogosteje prepoznavajo potencial občanske znanosti. Ta z uporabo množičenja in drugih metod omogoča učinkovito zbiranje obsežnih govornih podatkov na daljavo. V prispevku obravnavamo ključne dejavnike – tehnične, finančne, pravne, etične in motivacijske – ki jih je treba upoštevati pri načrtovanju trajnostnega in razširljivega sistema za pridobivanje govornih posnetkov. Na podlagi pregleda literature, analiz obstoječih metod in globalnih iniciativ za zbiranje govornih virov podajamo priporočila, primerna za implementacijo v slovenski prostor.

Ključne besede: govorni viri, pridobivanje govora, občanska znanost, množičenje, spontani govor, govorjena slovenščina

Bizjak, A.: Vloga občanske znanosti pri množičnem zbiranju govornih virov v slovenščini. Slovenščina 2.0, 13(1): 58–103

1.02 Pregledni znanstveni članek / Review Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2025.1.58-103
https://creativecommons.org/licenses/by-sa/4.0/

1 Uvod

Pridobivanje virov za govorne korpuse in baze postaja vse pomembnejše, saj po eni strani podpira razvoj z govorom povezanih tehnologij, po drugi strani pa vzpostavlja standard za raziskave govorjenega jezika. Strokovnjaki (Lindén idr., 2022; Eskénazi idr., 2013) ugotavljajo, da bodo obsežne zbirke odprto dostopnih jezikovnih virov pomembno pospešile razvoj in izvajanje različnih na jeziku temelječih aplikacij, povezanih z umetno inteligenco. Sodobna, vse bolj digitalizirana družba od posameznikov zahteva številne digitalne spretnosti. Če uporabnik slabše vidi ali ima težave z gibljivostjo prstov, je prikrajšan za uporabo številnih tehnoloških rešitev, kot so npr. mobilne naprave. Za razvoj rešitev, kot je samodejno glasno branje, s pomočjo katerih povečujemo socialno vključenost invalidov, pa moramo omogočiti rabo govornih podatkov tudi za komercialne namene.

Slovenščina, kakršno jezikovni uporabniki vsakodnevno govorijo po različnih regijah Slovenije, v zamejstvu in drugod po svetu, raziskovalcem nudi izjemno raznolikost izgovorjav, besed, fraz, oblik in struktur, ki pa v primerjavi s pisno slovenščino še vedno ni sistematično dokumentirana in zadostno raziskana. Verdonik (2008; Verdonik in Maučec, 2017) izpostavlja, da govorni podatki, ki jih pridobivamo in objavljamo znotraj govornih korpusov in baz, predstavljajo pomemben empirični vir za jezikoslovje, saj omogočajo natančno analizo fonetičnih, fonoloških, prozodičnih in dialektoloških značilnosti jezika, spoznavanje slovnice govorjenega jezika in raziskovanje pragmatičnih ter diskurznih pojavov v spontano tvorjenem govoru, kot so navezovanje, kohezija, diskurzni označevalci, govorna dejanja, retorična razmerja, namernost ipd. Raba govorjenega in pisnega jezika se razlikujeta tudi glede na leksikalne vzorce in kolokacije, pri čemer pisni korpusi, ne glede na njihovo velikost in strukturo, ne morejo zagotoviti zadostne količine podatkov za opis leksikalnih značilnosti govorjene slovenščine. Pomanjkanje podatkov o tipično govorjenem besedišču pomeni dodatno oviro pri implementaciji jezikovnih tehnologij za slovenski jezik, saj »če določena beseda (npr. *zrihtati, poštimati*) in podatki o njenem izgovoru niso vključeni v zaledne podatkovne baze, na katere se zanaša razpoznava govora, je razpoznavalnik bodisi ne razpozna

ali pa jo razpozna napačno« (Čibej idr., 2024). Prizadevanje za izgradnjo govornih baz pripomore k ohranjanju jezika in kulturne identitete, prenosu znanja in spodbujanju jezikovne raznolikosti – s tem tudi bolj vključujoče družbe –, hkrati pa zmanjšuje obstoječe vrzeli na področju govornih tehnologij, ki dajejo prednost jezikom, ki so v svetu tako ali tako najbolj razširjeni in prepoznavni.[1] Povpraševanje po jezikovnih virih in njihova ponudba nista vedno usklajena, pri čemer ne smemo prezreti, da sta pridobivanje govornih virov in izgradnja govornih korpusov izjemen dosežek sama po sebi (Eskénazi idr., 2013).

V prispevku se bomo dotaknili številnih izzivov, s katerimi se srečamo raziskovalci, ki se ukvarjamo z zbiranjem govornih virov. Eden ključnih je motiviranje govorcev za sodelovanje in soočanje z njihovimi dvomi in strahovi, povezanimi s snemanjem in deljenjem govora (Neale idr., 2017; Rutten idr., 2017). Drugi velik izziv je povezan z vprašanjem pristnosti govora na posnetku. Dialektologi, sociolingvisti in drugi jezikoslovci iščejo načine, kako na posnetku zajeti variabilnost jezika, kar neizogibno pripelje do paradoksa opazovalca, kot ga je formuliral William Labov (Lindén idr., 2022): cilj jezikoslovnih raziskav, osredotočenih na jezik sodobne družbe, mora biti, da ugotovimo, kako ljudje govorijo, kadar jih ne opazujemo sistematično; vendar lahko te podatke pridobimo samo s sistematičnim opazovanjem. Labov paradoks deluje ne glede na to, ali pogovor, v katerem sodelujejo govorci oz. informatorji, snema raziskovalec ali pa ga snema govorec sam – z drugimi besedami, ne glede na to, ali je raziskovalec, ki zbira podatke, prisoten ali ne. Isti paradoks velja tudi za zbiranje podatkov preko interneta, zlasti kadar ne zbiramo branega govora. Pri interakciji, ki poteka z računalnikom namesto s sočlovekom, se samodejno pojavi izziv, da govorci[2] zaradi zavedanja, da so snemani, skrbno nadzorujejo svojo jezikovno rabo.

Z izzivom doseganja spontanosti govora so se soočili tudi raziskovalci projekta Razvoj slovenščine v digitalnem okolju (RSDO),[3] ki so za govorno bazo Artur zbrali 1067 ur govora (od tega 884 ur

1 https://oecd-opsi.org/blog/estonians-donate-a-speech-preserving-language-and-driving-digital-service-delivery/

2 V literaturi se glede na različne pristope pojavljajo številna poimenovanja za posameznike, ki s svojim delom sodelujejo na daljavo – govorci, snemalci, delavci, množica, uporabniki, udeleženci, državljani, občani, člani skupnosti ipd. V tem članku si bomo zaradi osredotočenosti na govorne vire čim bolj dosledno prizadevali uporabljati poimenovanje govorci.

3 https://rsdo.slovenscina.eu/

transkribiranega). Poleg branega, parlamentarnega in javnega govora so zbirali tudi posnetke vsakdanjega, nejavnega govora, vključno s prostimi dialogi in monologi, ki so jih na terenu posneli z 263 govorci. Za razvoj dveh specializiranih razpoznavalnikov govora so zbrali še 10 ur posnetkov opisovanja obrazov, pri čemer je 125 govorcev ob prikazu fotografij različnih oseb prosto opisovalo njihove obrazne poteze. Drugi sklop (7 ur) je bil namenjen področju pametnega doma in je vključeval brane ter spontane ukaze za upravljanje pametnih naprav, v katerih je sodelovalo 148 govorcev (Verdonik idr., 2024).

Čeprav je lahko videti pridobivanje govora preko interneta oz. na daljavo, s svojo zmožnostjo doseči veliko število govorcev, ki sicer morda sploh ne bi sodelovali, hitrejša in gospodarnejša metoda kot tradicionalno terensko delo, si moramo pri načrtovanju zastaviti več vprašanj: kako nadzorovano zbirati govor, ki naj bo čim bolj naraven oz. spontan in ki naj v čim večji meri pokriva aktualno regionalno in socialno raznolikost jezikovnih uporabnikov? Kako zgraditi obsežno podatkovno zbirko, ki tudi funkcionalno predstavlja različne govorne vzorce (Lindén idr., 2022)? Na katere etične, pravne, finančne, motivacijske, organizacijske, tehnične in druge vidike moramo biti ob tem še pozorni?

Pri pridobivanju govornih virov se lahko zaradi pomanjkanja raziskav in literature zgledujemo po konceptih znanstvenoraziskovalnega dela, ki so bili za zbiranje podatkov uporabljeni znotraj drugih znanstvenih disciplin. Eden takšnih konceptov je občanska znanost,[4] ki občanske raziskovalce[5] vključuje v vse faze raziskovalnega procesa in za katero načelno velja, da so podatki in metapodatki, pridobljeni v projektih občanske znanosti, »javno dostopni in, kjer je možno, objavljeni v odprtokodnih formatih«.[6]

Za namen zbiranja govornih posnetkov je v sodobnem digitaliziranem svetu potencialno uporabna metoda mobilnega množičenja[7], ki izkorišča vseprisotnost mobilnih telefonov za množično zbiranje ažurnih

4 Angleški izraz citizen science prevajamo kot občanska znanost v skladu z utemeljitvijo, ki jo podaja Mlinar (2001).

5 Da bi v prispevku zagotovili enotno terminologijo, smo se v skladu z Mlinarjem (2021) odločili za rabo izraza občanski raziskovalec.

6 ECSA (European Citizen Science Association), 2015. Ten Principles of Citizen Science. Berlin. http://doi.org/10.17605/OSF.IO/XPR2N

7 Angleški izraz mobile crowdsourcing (MCS) prevajamo kot mobilno množičenje.

in celovitih podatkov. Takšna metoda ne le zapolnjuje vrzeli, ki jih puščajo konvencionalne metode, temveč s tem, ko skupnost neposredno vključuje v procese spremljanja in upravljanja s podatki, to tudi opolnomoča, kar lahko vodi do natančnejšega zbiranja podatkov in večje ozaveščenosti širše javnosti. Vendar pa je uspeh takšnih pobud močno odvisen od trajne vključenosti občanskih raziskovalcev, kar je odvisno od različnih motivacijskih dejavnikov in koristi, ki jih ti identificirajo (Rutten idr., 2017). Uporaba množičenja je v jezikoslovju na splošno še vedno v povojih, na drugih področjih, kot so računalništvo (vključno z NLP[8]), pa tudi v umetnosti, humanistiki in družboslovju pa je vse bolj razširjena.

S ciljem pripraviti podlage za razmislek, ali in na kak način bi se lahko spletna platforma za pridobivanje posnetkov spontanega, vsakdanjega govora uspešno načrtovala in uporabila za večanje govornih virov za slovenski jezik, bomo v 2. poglavju poleg omembe občanske znanosti, znotraj katere bomo obravnavali tudi prednosti in slabosti množičenja, raziskali nekaj najaktualnejših metod za zbiranje govornih virov na daljavo. V 3. poglavju bomo pregledali izbrane primere dobrih praks in nacionalnih kampanj, na osnovi katerih bomo v zaključnem delu prispevka opozorili na nekatere ključne izzive ter podali priporočila za načrtovanje tovrstne platforme pri nas, upoštevaje čim več vidikov te vse bolj pogoste metode zbiranja govornih podatkov.

2 Različne metode pridobivanja govornih posnetkov

Dialektologi in sociolingvisti tradicionalno pridobivajo podatke z obsežnim empiričnim terenskim delom, medtem ko strokovnjaki, ki se ukvarjajo s fonetiko, preferirajo podatke, pridobljene v laboratorijskih pogojih (Lindén idr., 2022). Vendar pa znanje, ustvarjeno v laboratorijih, pogosto ne more zajeti polne kompleksnosti in raznolikosti realističnega sveta, zato ga je treba dopolnjevati z znanji laične javnosti, kar »naj bi privedlo do novih znanstvenih spoznanj, predvsem na osnovi zbiranja številnih ali skritih podatkov, do katerih profesionalni raziskovalci sami ne bi mogli priti (pravočasno). Hkrati naj bi takšno sodelovanje vplivalo na to, da bi udeleženci povečali svoje znanje in interes za znanstveno delovanje« (Mlinar, 2021).

8 Angleški izraz natural language processing (NLP) prevajamo kot procesiranje naravnega jezika.

Za izgradnjo referenčnega govornega korpusa Gos 2.1[9] in obsežne govorne baze Artur 1.0[10] so bile uporabljene metode snemanja v studiu in na terenu s pomočjo najetih snemalcev, prvi poskusi uporabe množičenja in Games-With-A-Purpose (GWAP) na področju jezikovnih virov in tehnologij pa so bili izvedeni z razvojem mobilne aplikacije Igra besed (Arhar Holdt idr., 2021), med drugim namenjene izboljšanju odzivnega slovarja z avtomatsko pripravo podatkov, Kolokacijskega slovarja sodobne slovenščine. Trenutno je v izvajanju projekt Prilagodljiva obdelava naravnega jezika s pomočjo velikih jezikovnih modelov (PoVeJMo),[11] ki si za namen gradnje velikega jezikovnega modela za slovenščino preko množičenja in drugih tehnik prizadeva zbrati besedila v obsegu 40 milijard besed, med njimi tudi govorjenih.

2.1 Občanska znanost

Z znanostjo so se sprva individualistično ukvarjali predvsem amaterji, ki jih je motivirala želja po učenju in izboljšanju sveta (Soojung-Kim Pang v Mlinar, 2021). Vključevanje ljubiteljskih amaterjev v naravoslovne raziskave sega že v 17. stoletje (Miller-Rushing v Bonney, 2016), čemur je postopoma, s pojavom doktorskega študija ter širitvijo univerz in laboratorijev, ki so v 19. stoletju omogočali znanstveno karierno pot, sledilo razlikovanje med amaterji in profesionalci. Znanost je tako postajala vse manj dostopna, zato je bila praksa zbiranja podatkov s strani laične javnosti v 90. letih prejšnjega stoletja, ko je izraz občanska znanost začel uporabljati Laboratorij za ornitologijo v ZDA (Bonney idr., 2016), razmeroma redka.[12] Koncept občanske znanosti torej ni nov, je pa globalno priljubljenost pridobil v zadnjih dveh desetletjih, kar je povezano z razvojem družbe, temelječe na znanju. Projekti občanske znanosti lahko pozitivno prispevajo k družbeni blaginji, saj med drugim izpostavljajo vprašanja, pomembna za člane lokalne skupnosti, ki so opolnomočeni za iskanje najustreznejših rešitev (Bonney idr., 2016). Razmah občanske

9 https://www.clarin.si/repository/xmlui/handle/11356/1863
10 https://www.clarin.si/repository/xmlui/handle/11356/1776
11 https://povejmo.si/
12 V Združenem kraljestvu so ljubiteljski naravoslovci z zbiranjem in beleženjem podatkov o rastlinskih in živalskih vrstah pričeli stoletje prej (prav tam).

znanosti poleg že omenjenega pojava družbe, temelječe na znanju, omogočata napredek informacijsko-komunikacijskih tehnologij in vseprisotnost mobilnih naprav, ki imajo danes več računske moči kot celotna NASA leta 1969 (Gershenfeld, 2011).

Kar se same definicije tiče, je treba opozoriti na pojav najrazličnejših tipologij in kriterijev, ki so se oblikovali v zadnjih desetletjih in ki pogosto odražajo specifične raziskovalne kontekste projektov občanske znanosti, zlasti v Evropi. Ključno je, da definicija spodbuja odprto in široko razumevanje raznolikih raziskovalnih praks in participativnih dejavnosti, ki nastanejo, ko se v raziskovanje vključijo ljudje, ki raziskovalnega dela ne opravljajo kot del svoje plačane zaposlitve (Vohland, 2021). Po Oxford English Dictionary (2014) je občanska znanost znanstveno delo, ki ga izvajajo člani širše javnosti, pogosto v sodelovanju s ali pod vodstvom profesionalnih znanstvenikov in znanstvenih institucij. Po Mlinarju (2021) je pomemben vidik tega pristopa preseganje tradicionalnega, enosmernega prenosa znanja od strokovnjakov k laikom, ki velikokrat odraža podcenjevanje laičnega, izkustvenega znanja, in spodbujanje interaktivnosti. »Ob graditvi »znanosti od spodaj navzgor« postopoma vse bolj stopata v ospredje pozornost in identifikacija že pojavljajočih se pobud in prizadevanj na ravni posameznika in teritorialnih skupnosti« (prav tam).

Občanski raziskovalci niso reprezentativen prerez družbene realnosti, saj so pogosto višje izobraženi, srednjih let in zaposleni, redkeje pa prihajajo iz socialno-ekonomsko manj privilegiranih okolij. Njihova motivacija je raznolika – nekateri so študenti ali upokojenci, drugi imajo ambicije postati znanstveniki, tretji se ukvarjajo s problematiko varstva okolja ali izboljšanja lokalnih skupnosti (Mlinar, 2021). Pojem občanska znanost označuje obliko odprtega sodelovanja, v kateri posamezniki ali organizacije na različne načine sodelujejo v znanstvenem procesu, vključno z:

- oblikovanjem raziskovalnih vprašanj,
- snovanjem in izpopolnjevanjem raziskovalnih projektov,
- izvajanjem znanstvenih eksperimentov,
- zbiranjem in analiziranjem podatkov,
- interpretacijo rezultatov,
- razvojem tehnologij in aplikacij,

- odkrivanjem novih spoznanj in
- reševanjem problemov.[13]

Glede na deset načel občanske znanosti,[14] projekti s tega področja aktivno vključujejo pripadnike družbe v znanstvena prizadevanja, ki ustvarjajo nova znanja in razumevanja. Posamezniki lahko delujejo kot darovalci, sodelavci ali projektni vodje z vidno in pomembno vlogo pri projektu. Občanski raziskovalci lahko sodelujejo v več fazah znanstvenega procesa, poleg zgoraj naštetih tudi pri objavi in razširjanju rezultatov, pri čemer je pomembno, da so v projektnih rezultatih in publikacijah tudi navedeni. Posamezniki, ki želijo pri nas sodelovati kot občanski raziskovalci, lahko na spletni platformi Citizenscience.si dostopajo do več kot 50 aktivnih projektov.[15]

Vključevanje širšega kroga prostovoljcev in laikov, ne glede na njihovo formalno usposobljenost, v raziskovalni proces, profesionalnim znanstvenikom omogoča, da se posvetijo kompleksnejšim temam. Po Mlinarju (2021) gre za stopnjevanje, kontinuiteto, »ne pa za dihotomno, binarno ločevanje znanosti od neznanosti«. Znanost vse bolj postaja sestavina vsakdanjega življenja, tako lahko »danes izkustveno prepoznavamo aktualnost koncepta podružbljanja (socializacije) znanosti, ki se v enotnosti nasprotij spopada s procesom poznanstvenjenja (profesionalizacije) družbe« (prav tam).

Politika znanosti pogosto negativno sankcionira aplikativno usmerjeno raziskovanje, ki preko objav ne dosega širše mednarodne prepoznavnosti, s čimer se oddaljuje od načel občanske znanosti. Tako akademska skupnost v marsičem ohranja tradicijo zaprtosti, deloma tudi iz strahu, kot navaja Mlinar (2021), da bi intenzivnejše sodelovanje z uporabniki pomenilo »konec univerze«. Kljub prizadevanjem za »odpiranje znanosti« pa ni pričakovati, da bi to vodilo v izenačevanje vlog profesionalnih in občanskih raziskovalcev.

13 US Crowdsourcing and Citizen Science Act (15 USC 3724) (2016): https://www.law.cornell.edu/uscode/text/15/3724
14 https://citizenscience.si/obcanska-znanost/10-nacel-obcanske-znanosti/
15 https://citizenscience.si/aktivnosti/katalog-projektov/

2.2 Množičenje

Pomanjkanje virov na področju procesiranja naravnega jezika (NLP) in izzivi s kakovostjo in/ali pokritostjo obstoječih virov so dolgotrajna ovira, ki je upočasnila raziskave pri vseh jezikih, še zlasti pa tistih, ki že v izhodišču premorejo manj jezikovnih virov. Tradicionalne metode ustvarjanja zbirk podatkov za NLP pogosto zahtevajo precejšnjo količino človeškega dela in strokovnega znanja, saj večine virov ni mogoče pridobiti na povsem avtomatiziran način, kar povzroča visoke stroške in zaradi česar lahko na številnih področjih kot eno od alternativnih rešitev uporabimo množičenje – če le lahko zahtevane naloge opravi množica ljudi z ustreznim in združljivim naborom spretnosti in če upoštevamo, da raba tovrstnih metod terja kompleksno in skrbno upravljanje (Eskénazi idr., 2013; Lyding idr., 2022). V uporabnem jezikoslovju je raba množičenja še razmeroma nov pojav, nekoliko bolj je razširjen zlasti na področju transkribiranja razpoznanega govora in validacije kakovosti prevodov (Neale idr., 2017).

Izraz množičenje je leta 2006 uvedel Jeff Howe, ki ga je opredelil kot zunanje izvajanje,[16] tj. prenos funkcij ali nalog, ki jih je tradicionalno opravljal imenovani zastopnik, na neopredeljeno (običajno številčno) mrežo delavcev. Prenos funkcij izvede podjetje ali institucija v obliki »javnega poziva«,[17] običajno preko interneta. Danes se izraz uporablja za različne pojave, kot so uporabniško generirane vsebine (*user generated content*), soustvarjanje (*co-creation*), družbeno udejstvovanje (*social engagement*), odprte inovacije (*open innovation*), združevanje znanja (*knowledge aggregation*) ali napovedovanje *(prediction)* (Kaufmann idr., 2011).

James Surowiecki v svojem delu Modrost množic (2005) razloži moč modrosti množic, pri čemer je množica definirana kot skupina nestrokovnjakov oz. laikov, ki so se odzvali na javni poziv, da bi izvedli določene naloge.[18] Po Surowieckem lahko skupek mnenj posameznikov,

16 Angleški izraz outsourcing prevajamo kot zunanje izvajanje.

17 Angleški izraz open call prevajamo kot javni poziv.

18 Izrazoslovje na priljubljeni platformi za množičenje, Amazon Mechanical Turk, ki deluje kot nekakšna tržnica za spletno delo, je sledeče: organizacije, ki se ukvarjajo z množičenjem (requesters), objavijo manjše naloge, kot so generiranje vsebin, transkribiranje, označevanje slikovnega gradiva ali sodelovanje v spletnih raziskavah (HIT = human Intelligence tasks), da jih opravljajo predstavniki množice (workers), večinoma za vnaprej dogovorjeno denarno nadomestilo.

ki sprejemajo medsebojno neodvisne odločitve, privede do rezultatov, ki presegajo odločitve strokovnjakov oz. se kakovost rezultatov številnih nestrokovnjakov skupaj približa kakovosti mnenja posameznega strokovnjaka. Raziskovalci s področja jezikovnih tehnologij in govornih virov so v skladu s tem v sodelovanju množic prepoznali možno rešitev za dilemo, povezano s potrebo po obdelavi velikega števila podatkov.[19]

Preden preidemo k prednostim in slabostim množičenja in nekaterim njegovim izpeljavam, si ga na kratko oglejmo še skozi diahrono perspektivo (Eskénazi idr., 2013): sprva so raziskovalci govora posnetke pridobivali v laboratorijih, kjer so lahko natančno nadzirali akustične okoliščine in potek snemanja, izločili vsakršne šume, ustrezno nastavili mikrofon. Temeljna pomanjkljivost te metode je bila, da ni bila razširljiva. Zaradi potrebe po vedno večjih količinah podatkov so prešli na zbiranje govora preko telefona (zbiranje izključno zvočnih posnetkov), kar je zaradi zahtevnih prilagoditev tehnologije povzročilo precejšnje stroške. Temu navkljub je zbiranje govora preko telefona spodbudilo nastanek obsežnih govornih korpusov. Sčasoma so se razvili hibridni sistemi (uporaba tako spleta kot telefona), kar omogoča širšo dostopnost in dokaj preprosto integracijo v platforme za množičenje, kot je npr. globalno prepoznaven Amazon Mechanical Turk (MTurk).[20] Ko je raziskovalna skupnost premagala izziv avtomatizacije zbiranja govornih virov, se je morala soočiti z izzivom cenovno ugodnega pridobivanja govorcev. Kot zelo uspešno se je pri številnih projektih izkazalo sodelovanje z zunanjimi podjetji za tržne raziskave. Za izgradnjo korpusu CALLHOME[21] so govorcem v zameno za posnete pogovore ponudili brezplačne medkrajevne klice, kasneje pa so za izgradnjo korpusov Fisher (Cieri idr., 2004) in Mixer (Cieri idr., 2007) uporabili strategije spletnega marketinga in proaktivnega nagovarjanja govorcev. V novejšem času se postopki avtomatizacije širijo od pridobivanja govornih posnetkov na postopke preverjanja kakovosti, označevanja in transkribiranja. Najsodobnejša orodja za pridobivanje

19 Eden prvih primerov rabe modrosti množic je javni poziv, ki ga je Oxford English Dictionary (OED) v 19. stoletju naslovil na širšo skupnost s prošnjo, da prostovoljno indeksirajo vse besede v angleškem jeziku in zanje poiščejo primere za različne rabe posamezne besede (Eskénazi idr., 2013).

20 https://www.mturk.com/

21 https://doi.org/10.35111/exq3-x930

govora preko spletnih platform in mobilnih naprav omogočajo k uporabniku usmerjene rešitve,[22] kot so enostavnost za uporabo, individualne nastavitve, interaktivnost in igrifikacija.[23] Vendar pa lahko razlike v uporabniških napravah, nezdružljivosti mobilnih operacijskih sistemov in konfiguracijah mikrofonov ter druge zvočne opreme povzročijo razlike v kakovosti snemanja (Eskénazi idr., 2013).

Spletne platforme, ki skušajo izkoristiti modrost množic, se na globalni ravni ves čas spreminjajo. Vedno znova se porajajo nove oblike in koncepti:

- **množično financiranje (*crowdfounding*):** spletni vlagatelji financirajo mala podjetja;
- **državljansko udejstvovanje (*civic engagement*):** kolektivne akcije v zvezi z vprašanji javnega pomena;
- **kolektivno znanje (*collective knowledge*):** deljeno znanje in informacije;
- **odprte inovacije (*open innovation*):** uporaba zunanjih virov za generiranje in izvajanje zamisli.

Občansko znanost lahko zasledimo v najrazličnejših pojavnih oblikah:

- **spletne sodelovalne aktivnosti (*internet-based collaborative activity*):** soustvarjanje in sodelovanje posameznikov ali skupin pri raziskovalnem projektu preko spleta (Eskénazi idr., 2013);
- **človeško računanje (*human computation*):** uporaba človeške procesorske moči za odpravljanje težav, ki jih računalniki še ne morejo rešiti (Quinn in Bederson, 2011);
- **množičenje idej (*idea crowdsourcing*):** »Če imaš problem, vprašaj vse« je nova mantra, ki ji sledijo vodilna podjetja, kot sta IBM in Heineken, da bi izkoristila ustvarjalno moč množic. Množičenje

22 Angleški izraz *user-driven solutions* prevajamo kot k uporabniku usmerjene rešitve.

23 Govorce lahko skušamo spodbuditi na način, da zanje pripravimo zabavne teme in miselne izzive, kot so npr: »Znajdete se v srednjem veku. Vaščanom tistega časa skušajte predstaviti družbo, v kateri živimo danes, pri čemer upoštevajte, da mnogih besed takrat še ne bi razumeli«. Ali pa: »Predstavljajte si, da na sončni gozdni jasi srečate sebe, ko ste bili stari 8 let. Kaj bi si rekli?« Ali tudi: »Opišite svojo najljubšo osebo na način, da je vse, kar o njej povesite, izmišljeno.« Dodatno lahko sodelovanje spodbudimo s točkovanjem: za vsako oddano in odobreno minuto govorec npr. prejme 5 točk. Za vsak zaporedni dan, ko govorec odda nov posnetek, prejme dodatne točke. Ko govorec doseže mejnik (npr. 30 minut govora), prejme dodatnih 5 točk.

idej pomeni, da podjetja z javnim pozivom povabijo širšo skupnost, da predloži svoje ideje o tem, kako inovirati ponudbo izdelkov in storitev tega podjetja. Pojav je v zadnjem desetletju postal široko uporabljena praksa, s katero želijo institucije preseči svoje omejitve (Poesio idr., 2017). Raziskave so pokazale, da lahko sodelujoči predlagajo veliko svežih idej, ki brez težav tekmujejo s tistimi, ki se porodijo znotraj podjetja ali s pomočjo zunanjih strokovnjakov – oz. te celo presežejo (Garaus idr., 2024).

Množičenje se kaže kot priročna rešitev težav, ki pestijo jezikoslovce, ki se ukvarjajo z govornimi viri, vendar je ob tem nujna previdnost, saj lahko zlahka pridobimo zavajajoče ali napačne rezultate. Ključni izzivi rabe množičenja (Eskénazi idr., 2013) so povzeti v Tabeli 1:

Tabela 1: *Ključni izzivi rabe množičenja*

Količina informacij	Zagotoviti je treba zgolj ključne informacije: dovolj, da lahko govorci uspešno rešijo nalogo, in ne preveč, da ne bi z navodili vplivali nanje.[24] To je še zlasti pomembno pri spontanem govoru, saj lahko govorci zaradi prevelike količine informacij svoj govor (nehote) prilagodijo (napačnim) predvidevanjem, kaj se od njih pričakuje, hkrati pa jih lahko preveč podrobna navodila odvrnejo od sodelovanja, saj je razpon pozornosti v svetu digitalne resničnosti izjemno kratek.
Homogenost množic	Rezultati homogene množice običajno niso tako kakovostni, kot če je ta raznolika. Oinas-Kukkonen (2008) je ugotovil, da množica najboljše odločitve sprejme takrat, ko se posamezniki znotraj nje ne strinjajo in med seboj tekmujejo. Prevelika količina informacij homogenost množice le še poveča, prav tako pa tudi preveč komunikacije med naročnikom in govorci oz. med govorci samimi. Zaradi tehnike navzkrižne validacije oz. če imajo govorci dostop do posnetkov drugih, lahko to povzroči efekt posnemanja, kar množico ponovno homogenizira.
Razumna pričakovanja glede delovne obremenitve	Jasno mora biti, koliko časa in napora posamezna naloga zahteva, da preprečimo nezadovoljstvo sodelujočih. Če bodo govorci (nenamerno) zavedeni, da naloga zahteva manj dela, kot ga dejansko (še zlasti, če je delo plačano), bodo nalogo predčasno opustili in sodelovanje (preko različnih komunikacijskih kanalov) odsvetovali tudi drugim.

24 Raziskovalci in snovalci nalog za pridobivanje (dialoških) govornih posnetkov uporabljajo različne metode, da bi v čim večji meri zmanjšali svoj vpliv na govor sodelujočih. Ena od možnosti je podajanje informacij v tabelarični obliki ali s čim krajšimi spremljajočimi opisi. Obstajajo pa tudi pobude, v katerih so navodila in vzorčne primere oblikovali govorci sami, da bi tako zagotovili čim večjo nepristranskost.

Etično-pravni izzivi	Govorci pogosto prejmejo za svoje delo prenizko plačilo, končni stroški pa so lahko celo višji kot pri tradicionalnih metodah, saj poleg plačila dodatne stroške predstavljajo razvoj vmesnikov oz. spletnih platform in drugih rešitev, odkrivanje namernega goljufanja in validacija kakovosti. Izkušnje kažejo, da pravni vidiki, povezani s socialno varnostjo, pokojninami in sindikalnimi pravicami, niso vedno ustrezno upoštevani. Poleg tega je govor tudi biometrični osebni podatek, saj je govorca mogoče identificirati po glasu. Zato zbiranje in obdelava govornih posnetkov zahtevata ustrezno pravno podlago (npr. privolitev, pogodbo ali zakoniti interes) ter ureditev avtorskih in sorodnih pravic (Lindén idr., 2022).
Ustrezna usposobljenost govorcev in nadzor kakovosti	Da bi preprečili oddajo posnetkov, ki so zaradi neustrezne usposobljenosti govorcev prenizke kakovosti ali so rezultat spletnih goljufov, ki iščejo hiter zaslužek (naključno vnašajo odgovore oz. uporabljajo avtomatizirane robote), je smiselno implementirati neke vrste predhodno testiranje (preverimo npr. nekaj vzorčnih posnetkov). Če je časovno in stroškovno to ugodneje, lahko neustrezne posnetke izločimo tudi po tem, ko jih govorci že oddajo.
Motivacija	Zakaj sodelovati? Ali se sodelujoči želi naučiti nekaj novega, ga pritegne igra, želi nekaj doprinesti ali preprosto prejeti plačilo? Motivacijski dejavniki so ključnega pomena, da govorec nalogo izvede do konca in da se k sodelovanju kasneje tudi ponovno vrne.

Množičenje pa prinaša tudi številne prednosti (Eskénazi idr., 2013), kot so potencialno znižanje stroškov pridobivanja govornih virov in posledično povečan in tudi razširljiv obseg podatkov. V nasprotju s tradicionalnimi metodami lahko nagovorimo množico zelo raznolikih govorcev, saj nam omogoča lažji dostop do ljudi, ki bi jih bilo sicer težko doseči. Kar se vprašanja etičnosti tiče, prinaša množičenje nekaj pozitivnih sprememb, med njimi neposredno sodelovanje med delodajalci in delavci, ki ne vključuje posrednikov, v revnejših državah pa lahko pomeni vsaj nek vir prihodkov za posameznike v stiski.

2.3 Collect4NLP in Games-With-A-Purpose

Sodobni trendi rabe množičenja za pridobivanje podatkov na področju NLP med drugim vključujejo tudi metodi Games-With-A-Purpose (GWAP) in Collect4NLP. Tradicionalne platforme za množičenje, kot sta Amazon Mechanical Turk in Clickworker,[25] zagotavljajo finančne spodbude za izvedbo nalog, GWAP pa udeležence pritegne z interakcijo preko privlačnih elementov igrifikacije. Ta pristop, ki podatke za

25 https://www.clickworker.com/

NLP zbira med igranjem, se je pojavil po letu 2010, ciljno občinstvo pa so praktično vsi, ki imajo dostop do računalnika in ki jih zanimajo igre. Po drugi strani nedavna in še ne široko uveljavljena metoda Collect4NLP pridobiva podatke preko spletnega reševanja nalog za učenje tujih jezikov, kar se kaže kot velik potencial za trajnostno zbiranje velikih količin podatkov s pomočjo učečih se množic.

Metoda Collect4NLP predvideva, da če lahko s pomočjo NLP podatkovnih baz samodejno generiramo naloge za učenje tujih jezikov, lahko v naslednji fazi rešene naloge kot »stranski učinek« učne dejavnosti uporabimo za izboljšanje teh istih baz podatkov. Učeče se obravnava kot jezikoslovce, na katere se kot na posameznike ne moremo zanesti, vendar pa velika količina njihovih (potencialno šumnih) odgovorov na isto jezikovno vprašanje skupaj tvori statistično zanesljive rešitve, ki se približajo nivoju strokovnega znanja (podrobneje o modrosti množic gl. v poglavju 2.1). Da zagotovimo kakovost podatkov, potrebujemo robustne in zanesljive mehanizme združevanja odgovorov oz. odločitev učečih se, saj se nivoji njihovega znanja zelo razlikujejo. Mehanizmi združevanja najbolje delujejo, če so vhodni podatki, ki jih pridobimo od množic, čim bolj preprosti. V Tabeli 2 si oglejmo primerjavo med nekaj trenutno najaktualnejšimi metodami zbiranja podatkov na daljavo po različnih kriterijih (Lyding idr., 2022):

Tabela 2: *Primerjava različnih metod zbiranja jezikovnih virov na daljavo (Lyding idr., 2022)*

	Platforme za množičenje	Games-With-A-Purpose	Collect4NLP
Motiviranje množic	Finančne nagrade za vsako opravljeno nalogo (HIT): običajno od 0,01 do 0,20 dolarjev na nalogo, odvisno od zahtevnosti. Višja, kot je finančna spodbuda, višja je motivacija (Poesio idr., 2017).	Zabavne interakcije, ki vključujejo elemente igrifikacije, kot so lestvice najboljših, hitrost, napredovanje po nivojih in dodeljevanje značk. Glede na močno konkurenco klasičnih videoiger je težko ohranjati zanimanje (Tondello idr., 2016).	Gre za izkoriščanje notranje motivacije[26] učečih se za izboljšanje lastnih jezikovnih spretnosti. Bolj kot je izkušnja učenja jezika privlačna in uspešna, višja je motivacija.

26 Angleški izraz intrinsic motivation prevajamo kot notranja motivacija.

	Platforme za množičenje	Games-With-A-Purpose	Collect4NLP
Vključevanje množic[27]	Platforme za množičenje imajo vzpostavljene obsežne uporabniške baze in učinkovite mehanizme za promocijo nalog. Več kot polovica sodelujočih naj bi s platformo sodelovala srednjeročno oz. dolgoročno (od nekaj mesecev do nekaj let).	Množice sodelujejo tako dolgo, dokler jih igre (oz. jezikovne naloge z elementi igričarstva) pritegnejo. V tekmi z videoigrami, ki so prvenstveno namenjene zabavi in imajo višje proračune, pomeni to za raziskovalce precejšen izziv. Tudi zelo (tehnično) izpopolnjene GWAP z obsežno promocijsko kampanjo sčasoma izgubijo uporabnike in propadejo.	Nagovarjanje množic za uporabo celovitih rešitev za učenje jezika z integriranimi pristopi Collect4NLP temelji na promociji stabilnih platform, ki svojim uporabnikom kontinuirano zagotavljajo storitve, koristne zanje.
Kakovost množičenja	Ciljno občinstvo so običajno govorci, ki jim je izbrani jezik J1 ali ga kot J2 tekoče govorijo. Za zagotavljanje kakovosti se pogosto izvajajo predtestiranja ali vmesna testiranja za izključitev goljufov in govorcev s pomanjkljivim znanjem.	Namenjene so veščim govorcem J1 ali J2. Včasih vključujejo usposabljanje za novince, vendar le redko od sodelujočih zahtevajo vključitev v proces predtestiranja.	Namenjeni so govorcem, ki se učijo tujih jezikov ali manj veščim govorcem J1. Ker se po nivoju jezikovnega znanja med seboj močno razlikujejo, je treba njihove dosežke ves čas profilirati, da lahko pri združevanju podatkov odgovore ustrezno utežimo.
Stroški množičenja	Začetni stroški vzpostavitve platforme so precej nizki. Večina platform sodelujočim nudi plačilo, nekatere, kot sta npr. Zooniverse[28] in Distributed Proofreaders[29], pa se po principih občanske znanosti zanašajo na motivacijo sodelujočih, temelječo na altruizmu ali želji po pridobivanju/izgradnji znanja.	Uporabniki sodelujejo prostovoljno, vendar moramo računati na visoke razvojne stroške iger, saj ni veliko obstoječe infrastrukture. Navkljub prosto dostopnim repozitorijem, kot je npr. PythonGameLibraries, ki ponuja implementacijo nekaterih že razvitih igričarskih funkcionalnosti, so začetni stroški precej visoki.	Začetni stroški za razvoj in integracijo metode Collect4NLP so zaradi novosti raziskovalnega trenda še zelo visoki, saj težko najdemo repozitorije z referenčno kodo. Vendar pa naj bi bili v primerjavi s platformami za množičenje in GWAP dolgoročni stroški nižji.

27 Angleški izraz *crowd Involvement* prevajamo kot vključenost množice.

28 Gre za eno največjih in najuspešnejših platform s področja občanske znanosti: https://www.zooniverse.org/

29 Distributed Proofreaders (https://www.pgdp.net/c/) in LibriVox (https://librivox.org/), v sklopu katerega prostovoljci berejo avdio knjige, sta del obsežnega projekta Gutenberg (https://www.gutenberg.org/), ki omogoča brezplačen dostop do več kot 70.000 e-knjig.

Vsaka od navedenih metod ima, če se odločimo zanjo, svoje prednosti. Vendar pa je vzpostavitev spletne platforme rešitev, ki je za raziskovalce, ki z zbiranjem jezikovnih virov na daljavo šele začenjajo, najpreprostejša, saj zaradi že obstoječe infrastrukture ne predstavlja visokih razvojnih stroškov. Pri tem je seveda treba upoštevati, da finančne nagrade pomenijo stalen vir stroškov, v nasprotju z GWAP in Collect4NLP, ki dolgoročno predvidevata bolj trajnostne in etične načine sodelovanja množic.

3 Študije primerov dobrih praks

Opravili smo pregled domačih in tujih iniciativ ter platform za zbiranje govornih podatkov na daljavo, s poudarkom na platformah za množičenje. Kot že omenjeno, se govorni viri za slovenščino trenutno zbirajo znotraj tekočega projekta PoVeJMo[30] (gl. poglavje 2), kjer pa je večina doslej pridobljenih jezikovnih virov s področja pisnega, ne govorjenega jezika. Za zapolnitev te vrzeli je kot del projekta LLM4DH[31] vzpostavljen portal Govorjena slovenščina,[32] ki je namenjen zbiranju posnetkov spontanega, vsakdanjega govora. Trenutno lahko govorci svoje govorne posnetke oddajo tudi na globalni platformi Mozilla Common Voice, ki zbira govorne vire v različnih jezikih, vendar zaenkrat le posnetke branih povedi. V prvem četrtletju leta 2025 Mozilla Common Voice načrtuje uvedbo nove funkcionalnosti[33] – zbiranje spontanega govora, ki bo uporabnikom omogočilo, da se na vprašanja odzovejo z lastnimi besedami.

Iniciative smo identificirali preko različnih virov, vključno s pregledom znanstvene literature, spletnih strani in njihove siceršnje digitalne pojavnosti. Izmed več kot 45 globalnih in nacionalnih, tako komercialnih kot nekomercialnih projektov smo v skladu s kriteriji jezikovne razširjenosti in števila govorcev izbranega jezika, vrste platforme, uporabljenih tehničnih rešitev za pridobivanje posnetkov in namena zbiranja govornih virov, vrste govora, števila in ciljnih skupin govorcev, medijske podpore in načinov motiviranja govorcev ter širše prepoznavnosti projekta tako

30 https://povejmo.si/
31 https://www.cjvt.si/llm4dh/
32 https://govorjena-slovenscina.um.si/
33 https://www.mozillafoundation.org/en/blog/common-voice-navbar-changes-and-spontaneous-speech/?utm_source=chatgpt.com

v smislu nacionalnega pomena kot znanstvene potrebe po rezultatih, identificirali tri uspešne evropske projekte za zbiranje govornih virov, ki jih lahko obravnavamo kot vzorčne primere dobrih praks za morebitno implementacijo v slovensko okolje. Te iniciative so CorCenCC za zbiranje govornih virov za valižanski jezik, Anneta kõnet za estonski jezik in Lahjoita puhetta za zbiranje govornih virov za finski jezik.

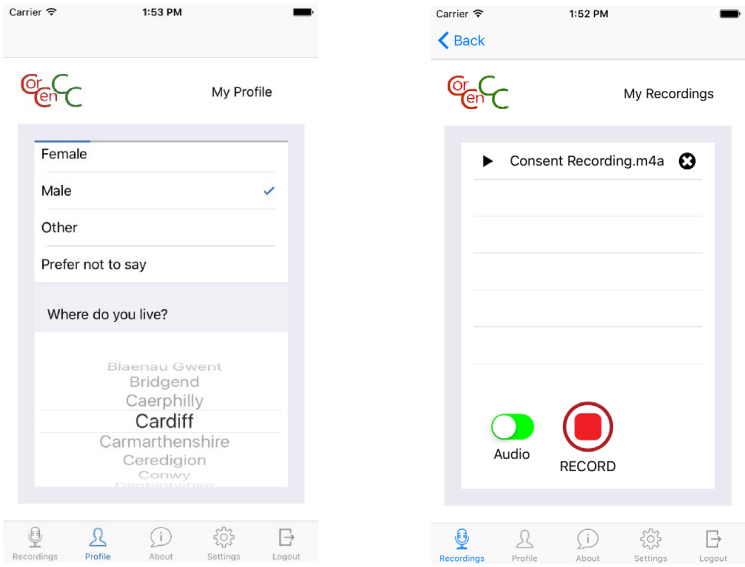
3.1 CorCenCC

CorCenCC (*Corpws Cenedlaethol Cymraeg Cyfoes*), tj. nacionalni korpus sodobne valižanščine (Neale idr., 2017; Knight idr., 2021), je prvi korpus, ki odslikava jezikovno rabo valižanščine v različnih komunikacijskih okoliščinah in kontekstih, žanrih in jezikovnih različicah (regionalnih in družbenih) različnih govorcev. Pri njegovi zasnovi je bil uporabljen pristop »od spodaj navzgor«, ki je razširil obstoječe zasnove dotedanjih korpusov. Cilj projekta, ki je potekal od 2016 do 2020, je bila vzpostavitev obsežnega korpusa z 10 milijoni besed, ki bi odslikaval jezikovno rabo govorcev z vseh območij Walesa, vključno z govorci valižanščine kot drugega ali tujega jezika. Zbrati jim je uspelo več kot 11 milijonov besed, od tega 2,8 milijona besed v sklopu govorjenega jezika, od tega v zasebnem kontekstu okoli 240.000 besed (Knight idr., 2020).

Zbiranje pisnih in e-virov je bilo osredotočeno na podatke, ki so hitro in lahko dostopni. Ker pa govornih virov, tvorjenih spontano, ni bilo v izobilju, je bilo treba poiskati nove in učinkovite načine za pridobivanje posnetkov govorjene valižanščine. Ena od ključnih inovacij projekta CorCenCC je bila zasnova jezikovnega korpusa, v večji meri usklajena z značilnostmi spleta 2.0. Sledila je vzpostavitev sistema, ki generiranje govorne baze omogoča neposredno govorcem samim; »v živo« in brez posrednika. Intuitivna in k uporabniku usmerjena aplikacija, razvita v sklopu projekta, je govorcem omogočila veliko bolj osebno izkušnjo, saj jim je predala lastništvo in nadzor nad lastnimi posnetki (delovanje aplikacije in posamezne funkcionalnosti so predstavljeni na Slikah 1 in 2). Tistim, ki so se prostovoljno odločili, da bodo podarili svoje posnetke, ni bila ponujena nobena finančna nagrada.

Govorne posnetke so govorci lahko oddali preko mobilne aplikacije (na platformah iOS in Android) ali z nalaganjem vnaprej posnetih posnetkov preko interaktivne spletne strani. Razvita mobilna aplikacija je

bila ena izmed prvih tovrstnih aplikacij, ki je v prizadevanjih za uravnotežen jezikovni korpus služila kot dopolnitev bolj tradicionalnih metod zbiranja podatkov. Govorcem je omogočala, da poleg svojih posnetkov govora naložijo tudi soglasja za njihovo uporabo in jih dopolnijo s potrebnimi metapodatki.[34] Da bi v čim večji meri dosegli zastavljene projektne cilje s področja vzorčenja, so morali novi govorci ob registraciji na svojem uporabniškem profilu vnesti precej podrobne metapodatke: regije bivanja, ki so sooblikovale njihovo rabo valižanščine, druge kontekstualne dejavnike, ki so vplivali na njihov jezikovni razvoj, lastno samooceno znanja valižanščine itd. S pomočjo tako izčrpnega profiliranja so snovalci dosegli uravnoteženost in reprezentativnost korpusa.

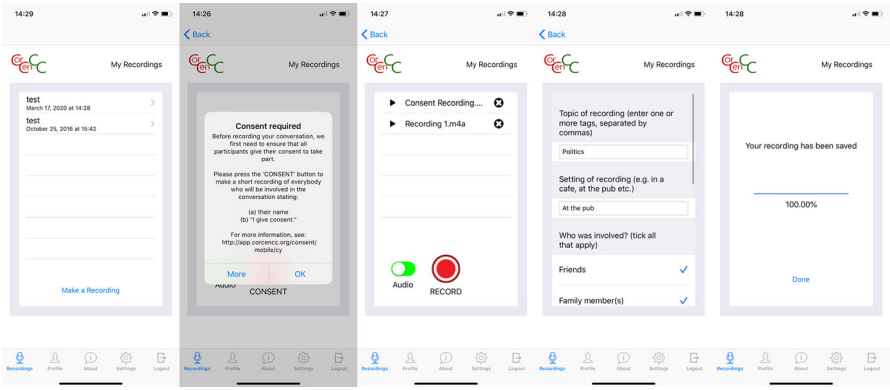


Slika 1: CorCenCC: Vnos metapodatkov na uporabniškem profilu in zaslon za pričetek snemanja (Neale idr., 2017).

V skladu s pravnimi in etičnimi vidiki izdelave korpusa je bilo ključnega pomena zagotoviti, da so uporabniki aplikacije podali vsa potrebna soglasja. Govorce so najprej preko celovitega opisa projekta seznanili z načini uporabe podatkov, nato pa jih pozvali k oddaji soglasja za vključitev posnetkov v CorCenCC in za njihovo rabo

34 Za sistematičen prikaz metapodatkov, ki so se zbirali, vključno s primeri, gl. Knight idr., 2021.

znotraj raziskovalne skupnosti. Govorci so soglasje morali oddati še pred postopkom registracije. Vsi govorci, ki so se pojavili na posameznem posnetku, so bili dodatno spodbujeni, da še preden pričnejo s snemanjem govora, posnamejo ustno privolitev za uporabo govornega posnetka.



Slika 2: Zaslonski posnetki mobilne aplikacije CorCenCC (Knight idr., 2021).

Zbiranje govornih posnetkov je potekalo na dva načina, in sicer preko pridobivanja govorcev, ki so jih posneli člani projektne ekipe, in s pridobivanjem individualnih govorcev, ki so posnetke v bolj zasebnih govornih položajih naložili preko mobilne aplikacije. Ker se je izkazalo, da je pridobivanje slednjih predstavljalo velik izziv, so za motiviranje in splošno ozaveščenost o projektu uporabili različne promocijske strategije, kot so oglaševanje preko televizijskih in radijskih nastopov, prisotnost v družbenih medijih, uporaba promocijskega gradiva[35] in udeležba na lokalnih dogodkih, pomembnih za Valižane in valižanski jezik. Omenjene kampanje so sicer povzročile veliko začetnega navdušenja, vendar pa se to ni odrazilo v velikem številu oddanih posnetkov. Predvidoma so bili govorci zaradi majhnosti jezikovne skupnosti, ki ji pripadajo, nesorazmerno močno zaskrbljeni, da jih bo na posnetkih mogoče identificirati, čeprav so člani projektne ekipe večkrat obveščali o anonimnosti sodelovanja. Govorcem

35 Promocijsko gradivo je vključevalo pisala, podstavke, letake in informativne brošure v velikosti razglednice. Da bi pritegnili pozornost in sodelovanje mlajših od 18 let, so oblikovali neuradno maskoto.

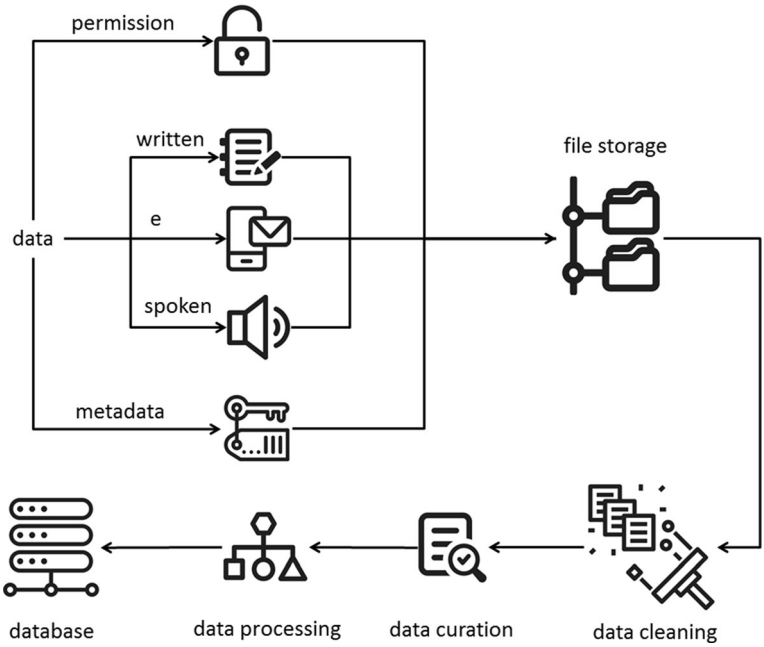
so pojasnjevali, da bodo posnetki pred vključitvijo v korpus ustrezno anonimizirani, in sicer na enak način, kot bi bil anonimiziran prispevek, ki bi ga zbrali jezikoslovci na terenu (Neale idr., 2017), (Knight idr., 2020). Podatki so bili, kolikor je bilo mogoče, anonimizirani s kombinacijo ročnih in avtomatiziranih tehnik. Z orodji za poizvedbo se da poiskati oznake za anonimizirane podatke ali transkripcijske konvencije, ki so bile uporabljene pri govornih podatkih. Za anonimizacijo podatkov se bile uporabljene posebne oznake, npr.:

- osebna imena: <anon> enwg1 </anon> – first male name <anon> enwb1 </anon> – first female name
- telefonske številke: <anon> Rhif ffôn </anon>
- e-poštni naslovi: <anon> cyfeiriad e-bost </anon>
- osebni naslovi: <anon> cyfeiriad </anon>[36]

Kot navajajo Arhar Holdt idr. (2022) je anonimizacijo mogoče izvesti na ravni metapodatkov z izpuščanjem avtorstva ali na vsebinski ravni, kjer se odstranjujejo osebni identifikacijski podatki. Postopki segajo od preprostega nadomeščanja občutljivih podatkov z enoznačnimi kodami do psevdonimizacije z generičnimi imeni. Pri pripravi govorne baze Artur[37] so bili na primer govorci, katerih govor je vključen v bazo, označeni z anonimnimi kodami, ki ne omogočajo njihove identifikacije, njihovi osebni in kontaktni podatki pa niso javno dostopni. Vse morebitne omembe osebnih imen in priimkov v samih posnetkih so bile v postopku postprocesiranja anonimizirane z nadomestitvijo glasu s piskom.

Ključni stebri infrastrukture korpusa CorCenCC so ogrodje, ki podpira zbiranje metapodatkov, interaktivna spletna stran in inovativna mobilna aplikacijo za zbiranje govornih podatkov preko množičenja, zaledna podatkovna zbirka za shrambo pregledanih podatkov in spletni vmesnik, ki uporabnikom omogoča dostop do podatkov preko spleta. Postopek zbiranja podatkov je predstavljen na Sliki 3.

36 https://corcencc.org/outputs/
37 https://www.clarin.si/repository/xmlui/handle/11356/1776



Slika 3: Postopek zbiranja podatkov pri CorCenCC (Knight idr., 2021).

Izvedena je bila študija uporabnosti, s pomočjo katere so bila uporabniška orodja ocenjena ter predlagane usmeritve za izboljšave. Čeprav je bila infrastruktura razvita za zbiranje govornih podatkov v valižanščini, se lahko njena zasnova ponovno uporabi za podporo razvoju korpusov drugih manjšinskih ali tudi večjih jezikov.

3.2 Anneta kõnet

Estonsko ministrstvo za gospodarstvo in komunikacije ter nacionalna agencija za informacijske sisteme[38] sta septembra 2022 lansirala kampanjo Anneta kõnet oz. Podari govor,[39] ki jo je financirala Evropska unija – NextGenerationEU. Kampanja, ki je kot del estonske jezikovne strategije 2021–2035 potekala po vsej državi, je bila namenjena zbiranju spontanega govora, ki bi omogočil razvoj govornih tehnologij

38 Ministry of Economic Affairs and Communications (https://mkm.ee/en), https://www.ria.ee/en

39 https://oecd-opsi.org/blog/estonians-donate-a-speech-preserving-language-and-driving-digital-service-delivery/

v estonščini (npr. razpoznavanje govora[40], glasovno upravljanje naprav ipd.) ter ohranjanju, razvoju in krepitvi jezika. Estonija je želela s kampanjo opolnomočiti podjetja ter javne[41] in raziskovalne institucije za razvoj inovativnih rešitev, s katerimi bi govorcem estonščine, tako domačim kot tujim, tako govorcem narečij kot tistim z okvaro vida ali sluha, omogočili boljši in hitrejši dostop do najsodobnejših jezikovno-tehnoloških storitev ter njihove implementacije v vsakdanje življenje. Samodejno podnaslavljanje se v oddajah estonske nacionalne televizij že uporablja, prizadevali pa so si izboljšati njegovo natančnost in razširjenost. S pomočjo metode množičenja so do aprila 2023 zbrali več kot 100 ur estonskega govorjenega jezika, ki je del odprtokodne podatkovne zbirke.[42]

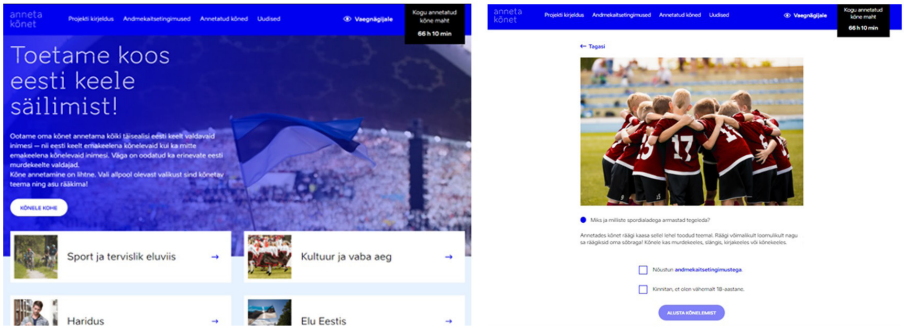
Za zbiranje govornih posnetkov je bila uporabljena metoda množičenja, in sicer so sodelujoči lahko govor podarili preko (tabličnega) računalnika ali pametnega telefona (za prikaz gl. Sliko 4). Donacijo govora so snovalci skušali čim bolj poenostaviti – govorec je moral zgolj obiskati spletno mesto, izbrati eno od vnaprej določenih tem in pričeti govoriti. Na voljo je imel tudi možnost, da govori o temi po lastni izbiri. Pozvani so bili, naj se izogibajo branju ali navajanju besedil drugih avtorjev ter naj uporabljajo čim bolj naraven in spontan govor, ki odraža njihovo vsakdanjo rabo jezika, vključno s slengovskimi izrazi. Prav tako naj ne bi razkrivali osebnih podatkov o sebi ali svojih bližnjih. Omejitve števila posnetkov za posameznega govorca ni bilo.[43]

40 Z namenom podpore razvoju jezikovnih tehnologij za zagotavljanje rešitev, primernih za sodobno estonsko družbo, so želeli z zbiranjem novih jezikovnih virov povečati točnost razpoznavanja spontano govorjene estonščine (vključno z narečji) s tedanjih 85 % na vsaj 91 % (https://oecd-opsi.org/blog/estonians-donate-a-speech-preserving-language-and-driving-digital-service-delivery/).

41 Sistem za razpoznavanje govora naj bi bil integriran v nacionalnega virtualnega pomočnika Bürokratt – orodje, ki temelji na umetni inteligenci in olajšuje komunikacijo z javnimi agencijami (https://www.kratid.ee/en/burokratt).

42 Estonian Open Data (https://avaandmed.eesti.ee/).

43 https://www.kratid.ee/en/keeletehnoloogia-projektid; https://laecwador.ee/en/projekt/donate-a-speech-campaign/



Slika 4: Anneta kõnet: Prikaz nabora razpoložljivih tematik in prikaz posamezne naloge.[44]

Kreativne ideje, uporabljene v promocijski kampanji, so namige črpale iz resničnih življenjskih situacij – zlasti situacij, v katerih skušamo v maternem jeziku posredovati glasovne ukaze pametnim napravam. Marketinška sporočila so izpostavljala dolgoročne cilje projekta v smislu družbenega napredka in ohranjanja narodne identitete. V skladu s tem sta promocijsko kampanjo sestavljala dva vsebinsko-vizualna sklopa: mlajšo populacijo so nagovorili z vidika govornih tehnologij, ki omogočajo hitrejše uvajanje glasovnih ukazov v vsakdanje življenje (gl. zapis na pametnem telefonu na Sliki 5), k starejši ciljni skupini pa so marketinško pristopili z vidika ohranjanja maternega jezika.[45] Zaradi širine ciljnega občinstva, znotraj katerega so posamezniki različno spretni pri upravljanju računalnikov in pametnih naprav, sta bila posneta dva različna TV-oglasa. Prvi se je osredotočal na uporabnost glasovnih tehnologij v vsakdanjem življenju, kot npr. med nabiranjem gob v gozdu, kjer nam lahko glasovne tehnologije rešijo življenje, drugi pa je izpostavil načine, kako je mogoče jezik ohraniti v današnjem času. Zaradi velikega poudarka projekta Anneta kõnet na zbiranju narečnega govora, je radijski posnetek potekal v najbolj razširjenem estonskem narečju, narečju Saare. V objavah v družbenih medijih so uporabili širok nabor estonskih

44 https://medium.com/digiriik/anneta-k%C3%B5net-cdb618d20279

45 Projekt Donate a speech lahko primerjamo z obdobjem estonskega prebujenja v 19. stoletju, ko so sistematično zbirali folkloro in pesmi, kar je prispevalo k vzponu nacionalizma, ki je približno 50 let pozneje utrl pot do neodvisnosti. Danes, skoraj 150 let pozneje, se je Estonija znašla v podobnem položaju – govorno zbirko je mogoče uporabiti in implementirati v tehnologije, ki pomembno prispevajo k sodobni družbi kot celoti (https://laecwador.ee/en/projekt/donate-a-speech-campaign/).

narečij in narečnih zvez, s čimer so jih skušali popularizirati, da ne bi potonili v pozabo.[46]

Slika 5: Anneta kõnet: Primer plakata, na katerem piše: Su hääl on eriline = Vaš glas je poseben!, na zaslonu pametnega telefona pa: »Hej, Siri!« »Žal mi je, ne razumem vas.«[47]

3.3 Lahjoita puhetta

Leta 2020 in v začetku 2021 je na Finskem potekala Lahjoita puhetta (Donate speech), ambiciozna kampanja za zbiranje digitalnih govornih podatkov za akademsko in komercialno rabo, katere namen je bil pospešitev razvoja aplikacij za finski jezik, temelječih na jezikovnih tehnologijah. Kampanjo je izvedla finska nacionalna radiotelevizijska družba Yle v sodelovanju s podjetjem Ilmastorahasto in Univerzo v Helsinkih. Čeprav je bil prvoten cilj, tj. zbrati 10 000 ur govora, precej ambiciozen, je izvajalcem kampanje v nekaj mesecih uspelo zbrati več kot 4000 ur govorjenega pogovornega jezika, tj. 220 000 govornih posnetkov več kot 25 000 govorcev iz celotne Finske (Lindén idr., 2022).

Posamezne jezikovnotehnološke rešitve v finščini so na voljo že več let (npr. govorne funkcije v izdelkih podjetij Apple in Google), a je bila za izboljšanje storitev za končne uporabnike potrebna zanesljivejša

46 https://oecd-opsi.org/blog/estonians-donate-a-speech-preserving-language-and-driving-digital-service-delivery/
47 https://laecwador.ee/en/projekt/donate-a-speech-campaign/

podpora za obdelavo vsakodnevne, spontane finščine. Mozilla Common Voice[48] sicer že od leta 2017 zbira govor na globalni ravni, vendar zbira izključno brani govor. Snovalci kampanje Lahjoita puhetta (Lindén idr., 2022) pa niso želeli usmerjati govorcev k rabi standardiziranega, prilagojenega govora, temveč so želeli zbrati vsakdanji spontani govor čim večjega števila govorcev, ki pripadajo čim bolj raznolikim skupinam jezikovnih uporabnikov.[49]

Ker vnos vseh metapodatkov ni bil obvezen, jih nekateri govorci niso vnesli v celoti.[50] Na osnovi metapodatkov, ki so bili vneseni, je bilo ugotovljeno, da so približno tri četrtine posnetkov prispevale osebe, stare med 20 in 60 let, od katerih jih je 70 % ženskega spola. Skoraj polovica prispevkov je bila donirana v štirih regijah, v katerih so največja mesta na Finskem, vendar so posnetke prejeli tudi iz vseh preostalih regij. Materni jezik večine darovalcev, tj. 95 %, je finščina. Več kot dve tretjini posnetkov so prispevali študenti, upokojenci, učitelji, podjetniki, strokovnjaki in medicinske sestre (v padajočem vrstnem redu glede na število prispevkov), preostanek pa predstavniki več kot 30 drugih poklicev z različnih družbenih področij. Približno 62 % jih je imelo visokošolsko izobrazbo, 28 % pa srednješolsko. Dve tretjini sodelujočih sta za darovanje govora uporabili spletni vmesnik, le 20 % jih je uporabilo aplikacijo za Android, preostali pa aplikacijo za iPhone. Skoraj 90 % od več kot 220.000 posnetkov traja od 10 sekund do 3 minut, v povprečju od 30 do 60 sekund, v skupni dolžini približno 4000 ur (Lindén idr., 2022).

Ključni izzivi pri načrtovanju uporabniškega vmesnika so bili 1. izbrati najustreznejše metode za pridobivanje spontano tvorjenega govora, 2. pridobiti zaupanje govorcev, 3. zagotoviti uporabniku prijazno izkušnjo ob hkratnem upoštevanju pravnih zahtev, ki jih je bilo treba govorcem predstaviti na razumljiv način, 4. podpreti tehnično ustrezne platforme, predstavitvene obrazce in vizualno-slušne povratne informacije o poteku snemanja in kakovosti posnetkov. Govorci so

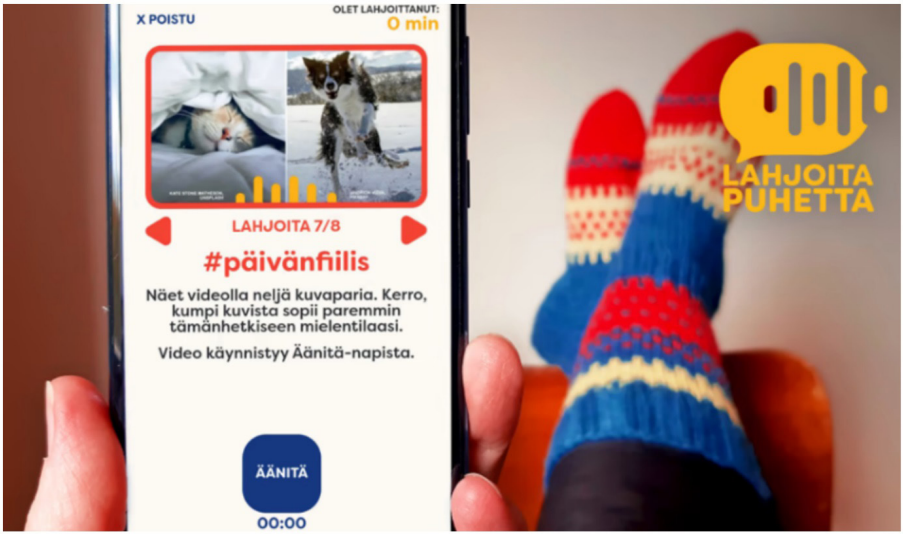
48 https://commonvoice.mozilla.org/

49 Izpostavljeno je bilo, da so dobrodošle vse različice govorjene finščine, tudi govor tujcev, ki se finščino učijo kot drugi ali tuji jezik. Kljub temu je bila za razumevanje obvestila o zasebnosti in navodil potrebna določena raven znanja jezika (Lindén idr., 2022).

50 Vključena so bila npr. vprašanja o dialektološkem ozadju in osnovnih demografskih podatkih, kot so starost, spol, materni jezik, kraj, v katerem oseba živi ali je bila rojena, ter njen poklic in stopnja izobrazbe.

posnetke lahko oddali preko spletišča ali mobilne aplikacije (na platformi iOS ali Android), omogočen pa jim je tudi bil naknaden prenos lokalno shranjenih posnetkov v oblak (v primeru slabe internetne povezave), vendar večji del sodelujočih te funkcionalnosti ni potreboval. Aplikacija za snemanje govora je prikazana na Sliki 6.

Po testiranju več različnih idej za teme so se odločili za preverjene teme,[51] s katerimi so preko krajših videev, slikovnega ali tekstovnega gradiva skušali privabiti govorce, da bi prispevali svoj govor z enim samim, za uporabo enostavnim gumbom za začetek in konec snemanja. Kljub številnim razpravam o tem, kako v projekt uvesti elemente igrifikacije, kot so primerjalne lestvice in timsko delo, je bilo na koncu v projekt vključeno zgolj obveščanje govorcev o skupnem času govora, ki so ga prispevali.

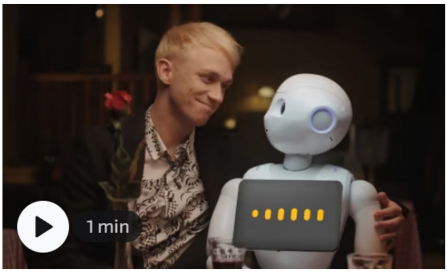


Slika 6: Lahjoita puhetta: Prikaz aplikacije za snemanje govora – naloga za govorca.[52]

51 Podjetje Yle je pripravilo okoli 40 čim enostavnejših tem, katerih namen je bil navdihniti in spodbuditi govorce. Poleg uvodne teme, namenjene vaji in validaciji, navajamo 12 najbolj priljubljenih tem, s katerimi je bila zbrana skoraj polovica vseh posnetkov: moj najljubši hišni ljubljenec; kaj naredi dom prijeten; predmet, ki mi je pomemben; moje najljubše oblačilo; kaj me jezi; moje odvečne stvari; kaj smo se naučili; moji nekdanji hišni ljubljenčki; ko gledam skozi okno; slikovna uganka; kaj sem jedel za zajtrk. Poleg tega so bili govorci pozvani tudi, da lahko v svojem okolju govorijo, o čemer koli želijo.

52 https://www.clarin.eu/impact-stories/donate-speech-digital-speech-database-boost-research-and-innovation

Da bi povečalo ozaveščenost in spodbudilo čim večjo udeležbo, je podjetje YLE pripravilo lahkotno in zabavno marketinško kampanjo. Izkazalo se je, da je bila prav ta, večkrat nagrajena kampanja ključna za uspeh celotnega projekta.[53] Posneli so humoristične oglase, v katerih so splošno javnost pozvali k oddaji govornih posnetkov (gl. Sliko 7). V času epidemije covida-19, poleti in jeseni leta 2020, so bili oglasi predvajani med premori na kanalih nacionalne RTV, nekaj ponovitev predvajanj oglasov je sledilo spomladi 2021.



Slika 7: Lahjoita puhetta: Zaslonska posnetka video oglasov.[54]

Tehnična kakovost govornih posnetkov je odvisna od več parametrov, vključno z izbiro ustreznega formata (priporočljivo je brezizgubno kodiranje, npr. .wav PCM), frekvence vzorčenja (priporočljivo 44,1 kHz) in bitne globine[55] (vsaj 16 bitov), ki vplivajo na razmerje signal/šum (SNR)[56]. Za minimizacijo odmevov in ohranitev optimalnega SNR mora biti mikrofon v neposredni bližini govorca, saj oddaljeno snemanje, npr. s telefonom na mizi, poveča prisotnost odmevov in zmanjša jakost signala. Snemanje naj poteka v akustično nadzorovanem okolju brez govora in šumov v ozadju, pri snemanju v hrupnih razmerah pa je treba izmeriti in dokumentirati SNR (Wieczorkowska, 2025).

Zaradi navedenih izzivov so snovalci iniciative *Lahjoita puhetta* izključili možnost snemanja diskusij z večjo skupino udeležencev. Čeprav

53 Leta 2021 je na festivalu Prix Europa, največjem tekmovanju za televizijske, radijske in spletne produkcije v Evropi, izmed 684 prijavami iz 26 držav prejela nagrado za najboljši evropski digitalni avdio projekt, razvita aplikacija pa je osvojila tudi nagrado za najboljšo mobilno storitev ter častno omembo v kategoriji za najboljšo uporabo podatkov na finski gala prireditvi za digitalne medije GrandOne (Lindén idr., 2022).

54 https://yle.fi/aihe/lahjoita-puhetta

55 Angleški izraz bit deepth prevajamo kot bitna globina.

56 Angleški izraz *Signal-to-Noise Ratio* prevajamo kot razmerje signal/šum.

bi takšen pristop omogočal prosto, dvosmerno komunikacijo, bi hkrati povzročil številne tehnične zaplete, ki bi otežili uporabnost posnetkov. Vsi govorci bi morali biti zbrani v neposredni bližini enega mikrofona ali pa razporejeni dovolj narazen, vsak s svojo napravo, da bi preprečili odmeve in preglase v signalu. Poleg tega bi bilo treba v zalednem sistemu sinhronizirati več signalov ali pa spremljati, kateri telefoni so snemali razpravo z več mikrofoni. Za ohranitev preprostosti so govorni signal zajeli v najvišji razpoložljivi brezizgubni kakovosti ter hkrati beležili metapodatke o sistemu, vrsti telefona in različici. Metapodatki so omogočili določene popravke znotraj postprocesiranja, kot na primer izenačevanje zvoka glede na vrsto mikrofona. Rezultati prvih naključnih vzorčenj in ročnih transkripcij, ki so jih opravili, so spodbudni, saj navkljub občasnim šumom na posnetkih, kot so govor v ozadju ali šum vetra, dokazujejo, da gradiva v povprečju ni težje razpoznati kot predhodnih govornih podatkov, posnetih v bolj nadzorovanih pogojih (Lindén idr., 2022).

Eden pomembnih vidikov kampanje Lahjoita puhetta so bile tudi pravne zahteve v zvezi z varstvom osebnih podatkov v skladu z evropsko in nacionalno zakonodajo o varstvu podatkov, predvsem Uredbo (EU) 2016/679 Evropskega parlamenta in Sveta o varstvu posameznikov pri obdelavi osebnih podatkov (GDPR, 2016). Pravna podlaga za obdelavo osebnih podatkov v okviru projekta temelji na zakonitem interesu. Opravljena je bila tudi ocena učinka v zvezi z varstvom podatkov, ki je pokazala, da obdelava osebnih podatkov v okviru projekta ni povzročila velikega tveganja. V praksi so udeleženci prejeli dva dokumenta: kratko izjavo z navedenimi pogoji sodelovanja, ki so jih morali sprejeti, ter obsežnejši opis politike varstva podatkov, v kateri je bil opisan način obdelave osebnih podatkov v kampanji in možnosti za kasnejšo odstranitev doniranih posnetkov govora, če bi se za to pojavila zahteva. Podatki so shranjeni na portalu Clarin, uporabniki pa so o morebitnih omejitvah dostopa do podatkov obveščeni.

Dostopnost in nadaljnja uporaba govornih virov sta določeni z licenčnimi pogoji. Sprememba obstoječih licenc lahko omogoči širšo rabo podatkov, vključno s komercialno uporabo. To zahteva ponovna pogajanja o licenčnih pogojih, vendar ta pogosto niso izvedljiva, zlasti kadar so govorni podatki že zbrani in licencirani pod strogo določenimi pogoji (Lindén idr., 2022).

4 Priporočila za pridobivanje posnetkov govorjene slovenščine

Kot smo videli doslej, je pri načrtovanju množičenja in drugih metod zbiranja govornih virov na daljavo nujno upoštevati številne dejavnike, med njimi vsaj pravne in tehnične, pa tudi finančne in etične. Še preden se bomo nekoliko podrobneje posvetili vsakemu od njih, si oglejmo nekaj osnovnih informacij o izbranih primerih dobrih praks, povzetih in strnjenih v tabelarični obliki (gl. Tabelo 3):

Tabela 3: *Izbrani primeri dobrih praks: osnovne informacije*

Iniciativa	Št. govorcev izbranega jezika	Obdobje pridobivanja govornih virov	Št. zbranih ur /št. besed	Vrsta govora na posnetku	Število govorcev	Ciljne skupine govorcev
CorCenCC[57]	približno 900.000 govorcev valižanščine[58]	2016–2020	več kot 11 milijonov besed, od tega 2,8 milijona besed v sklopu govorjenega jezika, od tega v zasebnem kontekstu okoli 240.000 besed (Knight idr., 2020)	pisni jezik, e-jezik in govorjeni jezik (preko aplikacije so zbirali spontan, nenapisan govor)	govorec snema sebe ali sodeluje v pogovoru s prijatelji, družinskimi člani	govorci različnih starosti, poklicev in spolov iz različnih geografskih regij Walesa, učenci valižanščine kot tujega jezika
Anneta kõnet[59]	približno 1.100.000 govorcev estonščine[60]	september 2022–april 2023	več kot 100 ur govorjenega jezika	spontani govorjeni jezik	ni omejitve (večinoma en govorec)	vsi odrasli govorci, tudi tisti, ki estonščino govorijo kot tuji jezik
Lahjoita puhetta[61]	približno 4.700.000 govorcev finščine[62]	leta 2020 in v začetku 2021	220.000 govornih posnetkov – 4000 ur govorjenega pogovornega jezika	vsakdanji, spontani govor	en govorec	čim bolj raznolike skupine govorcev, tudi učencev finščine kot drugega ali tujega jezika

57 https://corcencc.org/
58 https://www.gov.wales/
59 https://www.kratid.ee/en/projektid
60 https://www.stat.ee/en/news/population-census-76-estonias-population-speak-foreign-language?utm_source=chatgpt.com
61 https://www.kielipankki.fi/donate-speech/
62 https://stat.fi/index_en.html

Medijske kampanje, ki so vsebinski navdih črpale zlasti iz različnih namenov zbiranja govornih virov, so bile tiste, ki so v največji meri motivirale govorce, da so oddali posnetke svojega govora. Poleg teh v Tabeli 4 navajamo še nekaj drugih izzivov, ki so jih morali izvajalci izbranih kampanj nasloviti.

Tabela 4: *Izbrani primeri dobrih praks: izzivi, podpora in dostopnost*

Iniciativa	Namen zbiranja	Medijska podpora in promocija	Izzivi	Tehnične rešitve zbiranja virov	Dostopnost
CorCenCC	jezikovne rabe v različnih komunikacijskih okoliščinah in kontekstih, žanrih ter regionalnih in družbenih različicah najrazličnejših govorcev	TV- in radijski nastopi, družbeni mediji, promocijsko gradivo, udeležba na dogodkih	zagotavljanje uravnoteženosti in reprezentativnosti korpusa, saj z uporabo metode množičenja pridobivanje posnetkov poteka spontano in sproti; zaskrbljenost govorcev, da bi bili prepoznani	spletna stran in mobilna aplikacija, zaledna podatkovna shramba in spletni vmesnik za končne uporabnike	korpus je licenciran pod Creative Commons Attribution Non Commercial Share Alike 4.0 International, povezana programska orodja pa pod Creative Commons CC-BY-SA v4 (prosto dostopna za uporabo strokovnim skupnostim in posameznikom, ki jih zanima jezik)[63]
Anneta kõnet	razvoj jezikovnih tehnologij in pomoč pri ohranjanju estonskega jezika	TV- in radijski oglasi, objave v družbenih medijih, plakati itd.	motiviranje govorcev, da oddajo svoj govor	oddaja preko (tabličnega) ali računalnika, pametnega telefona	korpus je prosto dostopen in licenciran pod Creative Commons (CC-BY)[64]
Lahjoita puhetta	pospešitev razvoja aplikacij za finski jezik, ki temeljijo na jezikovnih tehnologijah	nagrajena humoristična medijska kampanja je bila ključna za uspeh	izbor metode za pridobivanje spontanega govora; pridobiti zaupanje govorcev; zagotoviti uporabniku prijazno izkušnjo ob hkratnem upoštevanju pravnih zahtev; podpreti tehnično ustrezne platforme; zagotoviti kakovost posnetkov	spletni vmesnik za snemanje preko (tabličnih) računalnikov, mobilna aplikacija za snemanje s pametnimi telefoni	dostopno raziskovalcem, univerzam, inštitutomin podjetjem za namen jezikoslovnih in AI raziskav, razvoj AI rešitev ter visokošolsko izobraževanje; raziskovalcem na voljo brezplačno, za komercialno rabo je predvidena pristojbina za kritje stroškov obdelave[65]

63 https://corcencc.org/outputs/
64 https://metashare.ut.ee/repository/browse/estonian-speech-corpus-donate-speech/ef43652fd7d011ee957893764e77942db3e80cb9a896430ebbcf1877b491e36f/
65 https://www.kielipankki.fi/donate-speech/

4.1 Načrtovanje zbiranja govornih virov

Ko se odločimo, da bomo zbirali govorne vire na daljavo, moramo najprej definirati, kakšne vrste govora želimo pridobiti, kako in kateri govorci lahko v ta namen sodelujejo. Običajno si raziskovalci govorjenega jezika in jezikovnih tehnologij prizadevamo v analizo vključiti čim bolj spontan govor v raznolikih komunikacijskih okoliščinah in kontekstih, pri čemer si prizadevamo zajeti tako regionalno kot socialno obarvane različice govora jezikovne skupnosti, ki jo raziskujemo. K sodelovanju zato vabimo govorce, ki so glede na demografske kriterije med seboj čim bolj različni, pri čemer pa mora biti njihova porazdelitev glede na spol, starost, izobrazbo, kraj bivanja ipd. čim bolj uravnotežena. Snovalci vseh treh analiziranih kampanj so govor zbirali tudi od (nematernih) govorcev po vsem svetu, ki izbrani jezik govorijo kot drugi ali tuji jezik. Govorne posnetke so lahko podarili ali preko (tabličnega) računalnika, s katerim so dostopali do interaktivne spletne strani, ali preko mobilne aplikacije na pametnih telefonih (delujoče na platformah iOS in Android). Pri načrtovanju tehničnih rešitev je treba predvideti še zaledno podatkovno shrambo in varnostno kopiranje.

Eno ključnih vprašanj je, koliko govorcev naj bo vključenih na posnetku. V sklopu projekta CorCenCC so govorci preko aplikacije lahko posneli svoj govor ali pa so sodelovali v pogovoru, za projekt Anneta kõnet nismo našli podatka, da bi snovalci v zvezi s tem navajali kakšne posebne omejitve (ker je projekt že zaključen, so nekateri podatki težje dostopni). Zanimivo pa je, da so pri projektu Lahjoita puhetta identificirali več tehničnih izzivov (gl. poglavje 3.3), na osnovi katerih so sprejeli odločitev, da na posnetku hkrati lahko sodeluje zgolj en govorec. Ker je bila vrsta govora, izbrana v vseh treh vzorčnih primerih, spontani govorjeni jezik (govorci so bili pozvani, naj ne berejo), je odločitev finskih raziskovalcev toliko bolj nenavadna, saj v dialoških govornih situacijah, ki potekajo v domačem okolju in trajajo dlje časa, lahko v večji meri pričakujemo, da bo govorcem uspelo vsaj delno ubežati pastem Labovega paradoksa (gl. 1. Uvod). Da bi kljub temu dosegli, da bi bil govor čim bolj nepripravljen, so govorcem na zaslonih prikazali krajše naloge s slikovnimi in besedilnimi primeri, ki so funkcionirale kot iztočnice za spontane govorne odzive.

Po sprejetju odločitve, kdo vse lahko sodeluje, moramo za izbrane govorce načrtovati zanje ustrezne naloge. To lahko dosežemo na dva načina: z zbirko različnih nalog, med katerimi sodelujoči sami izberejo takšno, ki jim najbolj ustreza, ali pa s členitvijo dane naloge na manjše gradnike, ki jih opolnomoči za izbiro konkretnih podnalog (Rutten idr., 2017). Raziskave (Eskénazi idr., 2013) dokazujejo, da razčlenitev kompleksnejše naloge na več manjših podnalog pri sodelujočih povečuje natančnost. Ugotovitev je v skladna s konceptom kognitivne obremenitve, ki pravi, da več kot je vidikov, ki jih mora sodelujoči pri nalogi upoštevati, slabše bo opravil vsak posamezni sestavni del te naloge. Ko načrtujemo nalogo, skušajmo kot nezainteresirani opazovalec ugotoviti njeno kompleksnost: katere odločitve mora govorec sprejeti, da bi opravil nalogo? Ali je mogoče zahtevnejše naloge razdeliti na več podnalog? Kakšna so navodila za vsako od teh podnalog?

Če želimo, da naš projekt v množici projektov, objavljenih na platformi za množičenje, pozitivno izstopa, upoštevajmo naslednje (Eskénazi idr., 2013):

- Ime naloge naj bo privlačno. Naslov naj čim več pove tudi že o sami vsebini naloge, s čimer dodatno vzbudimo zanimanje (namesto *označevanje govora* lahko nalogo *imenujemo označevanje vsakdanjega govora dveh prijateljev*).
- Naloga naj bo vizualno privlačna (ozadje, pisava, barva in drugi grafični elementi). Gre za subtilen način, kako naredimo nalogo privlačnejšo od drugih.
- Vsaka naloga naj bo na enem zaslonu. Tako dosežemo, da se sodelujočemu za dokončanje naloge ni treba pomikati navzdol ali na rob strani. S tem sporočamo, da naloga ne bo zahtevala dodatnega časa za pomikanje po strani in bo zato videti krajša in prijetnejša.

V procesu načrtovanja naloge moramo iskati ustrezno ravnotežje med zahtevnostjo naloge in točnostjo pridobljenih podatkov. Podatkom, še zlasti tistim, pridobljenim s pomočjo občanske znanosti, končni uporabniki pogosto ne zaupajo, saj dvomijo o njihovi kakovosti, natančnosti in zanesljivosti. Ampak če zaradi tega vztrajamo pri metodah, ki zagotavljajo optimalne in točne podatke, lahko s tem paradoksalno povzročimo, da bo projekt za sodelujoče postal preveč

zahteven in posledično neprivlačen. Dodaten izziv, ki ga moramo nasloviti, je, da končni uporabniki dvomijo, da bodo podatki, pridobljeni v sklopu prostovoljnega sodelovanja, dejansko tudi uporabljeni, zaradi česar moramo biti pozorni, da rezultatov ne diseminiramo zgolj znotraj strokovne javnosti, temveč jih moramo na razumljiv način aktivno predstaviti tudi širši skupnosti (npr. kako, kje in v kolikšnem obsegu so bili rezultati dejansko uporabljeni). Gre za enega od ključnih dejavnikov, ki poleg etičnih vidikov (transparentnost, zaupanje, avtonomija in vzajemno učenje) projektom občanske znanosti zagotavljajo uspeh (Rutten idr., 2017).

Namen, cilji in pričakovanja morajo biti sodelujočim jasno sporočeni na način, da se med akterji vzpostavita zaupanje in legitimnost. Tipično se govorni podatki zbirajo za namen podpore razvoju jezikovnih tehnologij in na njih temelječih rešitev, ki širši jezikovni skupnosti olajšajo vsakodnevne aktivnosti, hkrati pa za namen proučevanja in enakopravne vloge jezika v digitalnem okolju ter njegovi ohranitvi.

4.2 Usposobljenost govorcev

Vsi govorci niso ustrezno usposobljeni za opravljanje vseh tipov nalog. Težave nam kot načrtovalcem in izvajalcem projektov množičenja lahko povzroči neidentificirano sodelovanje:

- robotov (*bots*),
- goljufov,
- posameznikov, ki želijo zaslužiti čim več s čim manj vloženega napora, pri čemer jim kakovost opravljenega dela ne predstavlja vrednote,
- govorcev, ki ne govorijo jezika, za katerega poteka zbiranje govora,
- mladoletnih in drugih govorcev, ki ne podajo soglasja za uporabo posnetkov.

Če platforma to dopušča, lahko neusposobljene ali škodoželjne udeležence izločimo s preverjanem ocen, ki so jih že prejeli v predhodnih projektih. Nivo usposobljenosti dobronamernih udeležencev pa lahko preverimo tako, da predhodno opravijo vzorčni primer naše naloge v manjšem obsegu – predtestiranje (Eskénazi idr., 2013). Crall

idr. (2011) poudarijo pomen analize dejavnikov, ki vplivajo na kakovost podatkov v projektih občanske znanosti. Opozorijo na razlike med udeleženci glede na stopnjo predhodnega usposabljanja in izkušenj, opredeljenosti postopkov raziskovanja in rabe IKT-tehnologije, dokažejo pa tudi pozitivno vlogo uporabe ustreznih protokolov. Mlinar (2021) dodaja pomen avtomatizacije odkrivanja napak in certificiranja usposobljenosti udeležencev, ki dodatno pripomore k dolgoročni angažiranosti prostovoljcev. Vendar pa zagotavljanje kakovosti podatkov »ni nekaj, kar bi laike in širšo javnost apriori diskvalificiralo oziroma izključevalo iz sfere znanstvenega delovanja; prej je to izziv za intenzivnejše uveljavljanje ukrepov in protokolov, ki bi vodili k večji kakovosti podatkov in delovanja« (prav tam).

4.3 Navodila za snemanje

Če govorcem ne podamo dovolj natančnih in jasnih navodil, lahko njihovo napačno razumevanje le-teh občutno zmanjša kakovost rezultatov, dobljenih z množičenjem. Navodila morajo biti objavljena na vidnem mestu, poleg tega je priporočljivo, da uporabimo ali predstavitveni video ali organiziramo (spletne) delavnice. Nekaj dodatnih usmeritev:

- **Začnimo z jasnim opisom naloge**: Čeprav je termin »transkripcija« za večino razumljiv, je pametno uporabiti besede, ki so nedvoumne za čim večje število ljudi. Navedli bi lahko: »Z besedami boste zapisali, kar boste slišali na zvočnih posnetkih.«
- Zagotovimo **zgolj ključne informacije**: Dovolj, da lahko govorci uspešno rešijo nalogo, in ne preveč, da ne bi z navodili vplivali nanje, kar bi posledično lahko vodilo do preveč homogene množice (gl. poglavje 2.2).
- **Razmislimo o mejnih primerih:** Kaj storiti s posnetki, na katerih bo veliko šuma? Kaj če bo na posnetku zgolj govor v tujem jeziku? Premislek o tovrstnih mejnih primerih nam bo pomagal bolje razumeti našo podatkovno bazo.
- **Ponudimo čim več primerov**: Število sodelujočih s predhodnimi izkušnjami na področju pridobivanja govora je zelo omejeno. Izkušnje kažejo, da je eden najučinkovitejših načinov za opis naloge navajanje številnih primerov.

- Za preverjanje, ali so govorci navodila ustrezno razumeli, lahko ti izpolnijo vprašalnik, s pomočjo katerega **ocenimo njihovo razumevanje naloge** (Eskénazi idr., 2013).

Prvotni projekti, ki so za pridobivanje govornih virov uporabili množičenje, so se soočali s povsem banalnimi izzivi, kot so navedeni spodaj:

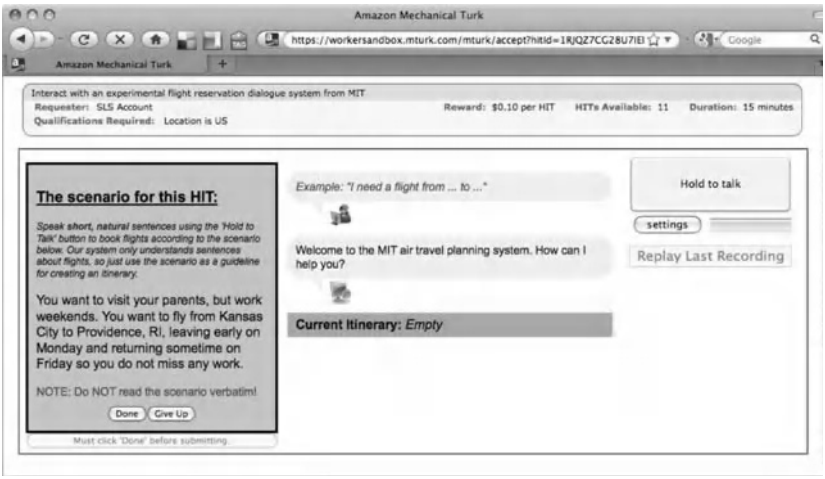
- Posnetek ne vsebuje zvoka.[66]
- Govorec ne uporablja mikrofona/slušalk oz. te niso ustrezno priključene.
- Previsok ali prenizek nivo zvoka.
- Visoka raven hrupa iz okolice.
- Neupoštevanje navodil (Eskénazi idr., 2013).

Zato je priporočljivo, da že takoj na začetku preverimo, ali avdio zares deluje:

- Govorce prosimo, da poslušajo svoj posnetek in ga odobrijo zgolj, če je zvok slišen.
- Validiramo naključne posnetke in govorcem sproti posredujemo povratne informacije.
- Oblikujemo čim bolj razumljiva in nedvoumna tehnična navodila.

Začetniki, ki niso vajeni interakcije s platformo za pridobivanje govornih posnetkov, pomenijo svojevrsten izziv. Če je informacij premalo in so navodila z namenom, da ne bi z njimi vplivali na govor, preskopa, se lahko zgodi, da govorci podani scenarij zgolj glasno preberejo, kot se je to zgodilo v primeru eksperimenta, ki ga je izvedel MIT (Eskénazi idr., 2013). Navodila za nalogo, objavljeno na platformi MTurk, so morali zaradi napačnih odzivov govorcev večkrat dopolniti in popraviti. Končna različica je vidna na Sliki 8:

66 Lane s sodelavci (Eskénazi idr., 2013) je v procesu postprocesiranja odkril, da je približno 10 % izjav, ki so jih pridobili z množičenjem, neuporabnih. Za testiranje zvoka so zato kasneje razvili rešitev, ki je za testiranje zvoka predvidela, da so uporabniki najprej ustno rešili preprost matematični problem.



Slika 8*:* Prikaz vzorčnega HIT, kot ga vidi izvajalec na MTurk (Eskénazi idr., 2013).

4.4 Finančni, pravni in etični vidiki

Govorci so pogosto zaskrbljeni, da bi bili na posnetku prepoznani, zato je nujno, da pridobimo njihovo zaupanje in zagotovimo anonimizacijo (gl. poglavje 3.1). Dimenzija etičnosti je še posebej ključna pri vprašanjih zasebnosti[67] in rabe osebnih podatkov, zaradi česar se morajo prostovoljci od vsega začetka strinjati s predvideno uporabo pridobljenih podatkov, da bi dosegli transparentnost in dolgoročno zaupanje.

Vsak posnetek govora je namreč biometrični osebni podatek, ker je govorca po govoru (npr. intonacija, tempo, barva glasu ...) mogoče identificirati.[68] Iz tega izhaja, da za vsakega govorca na posnetku po GDPR (2016) potrebujemo podlago za obdelavo osebnih podatkov (praviloma gre za osebno privolitev, izpolnjevanje pogodbe ali zakonit

67 Vprašanju zasebnosti se je treba pri mobilni občanski znanosti še intenzivneje posvetiti kot sicer, saj imamo opravka z občutljivejimi podatki. IP-naslovi in gibanje sodelujočih lahko razkrijejo njihov domači ali službeni naslov. Še zlasti previdni moramo biti pri deljenju slikovnega in zvočnega gradiva, ki lahko razkriva osebne podatke (Rutten idr., 2017).

68 »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-osebnih-podatkov/obveznosti-upravljavcev/prijava-biometrijskih-ukrepov/.

interes). Če obdelava govornih posnetkov temelji na privolitvi posameznika, mora biti ta podana na nedvoumen način in jasno ločena od ostalega besedila. Privolitev mora vsebovati informacije o upravljavcu osebnih podatkov, namenu zbiranja, obsegu podatkov, ki se bodo obdelovali, ter o dostopnosti in javni objavi govornih posnetkov. Prav tako mora vsebovati navedbo, komu bodo posnetki dostopni in s kakšnim namenom. Posameznik ima pravico, da privolitev kadar koli prekliče. Ob vzpostavitvi portala za zbiranje govornih posnetkov je treba objaviti tudi izjavo o varstvu osebnih podatkov.

Posnetki vsakdanjega spontanega govora praviloma ne izpolnjujejo pogojev za avtorska dela, kot definira avtorsko delo Zakon o avtorski in sorodnih pravicah (ZASP, 1995, 5. člen), namreč niso »individualne stvaritve s področja književnosti, znanosti in umetnosti, ki so na kakršen koli način izražene«. Avtorsko zaščitena govorna dela so na primer govori, pridige, predavanja, medtem ko vsakdanji govor ni tipičen predstavnik takih del. Iz razlogov previdnosti je sicer mogoče urediti prenos morebitnih avtorskih ali sorodnih pravic, vendar pa moralnih avtorskih pravic ni mogoče prenesti, saj so v skladu z določbami ZASP (1995, 70. člen) neprenosljive. Zaradi navedenega je bolj smiselno, da govorci podajo izjavo, da na posnetkih ni njihovih avtorskih ali sorodnih pravic oziroma sorodnih pravic tretjih oseb ter da posnetki ne predstavljajo avtorskih del.

Kar se plačila tiče, bi nas moralo voditi načelo *»pay enough or not pay at all«* (*plačaj pošteno ali ne plačaj nič*). Ta precej radikalna predpostavka temelji na dejstvu, da se motivi za sodelovanje pri prostovoljnem ali plačanem delu pri večini ljudi zelo razlikujejo.[69] Čeprav se sliši protislovno, lahko naknadno zagotavljanje finančnih spodbud za nalogo, za katero plačilo sprva ni bilo predvideno, uspešnost izvedbe te naloge zgolj zmanjša. Poleg višine plačila je pomembno tudi, da s plačilom ne odlašamo in da posnetek zavrnemo le, če smo prepričani, da je krivda za napako res govorčeva. Za vzpostavitev trajnostnega in

69 Povprečna višina plačila, ki je po nekaterih raziskavah 1,38 dolarja na uro, je za zahodne standarde precej nizka. Ker pa je demografija delavcev zelo raznolika glede na državljanstvo, starost, izobrazbo in dohodek, domnevamo, da na odločitev za sodelovanje vplivajo tudi drugi motivacijski dejavniki, ne le takojšnje plačilo (Garaus idr., 2024). Gre za pogosto vprašanje na področju množičenja, in sicer, ali in v kolikšni meri višina plačila vpliva na kakovost opravljenih nalog. Ker je vprašanje precej kompleksno, bi se mu bilo v prihodnje smiselno podrobneje posvetiti.

stabilnega modela imamo torej na voljo dve možnosti: 1. pošteno plačilo, z možnostjo prilagoditve glede na doseženo kakovost, 2. spodbujanje sodelovanja brez finančnih nagrad, kot je to značilno za sodelovalno znanost[70] ali Games-With-A-Purpose, kjer nalogo načrtujemo nalogo tako, da bo sodelujočim dovolj zanimiva in jo bodo želeli rešiti.

Etične smernice, ki jih je treba upoštevati, če se kot načrtovalci odločimo, da bomo govorcem za njihovo delo nudili plačilo (Eskénazi idr., 2013; Quinn in Bederson, 2011; Gneezy in Rustichini, 2000):

- **transparentnost:** urna postavka in rok izplačila naj bosta znana vnaprej (ne zamujajmo z izplačili);
- **spoštovanje časa**: optimizirajmo naloge tako, da bo čas za opravljeno delo čim bolj učinkovito porabljen;
- **ustrezna delovna obremenitev**: naša pričakovanja glede delovne obremenitve govorcev morajo biti razumna in jasna;
- **jasna pričakovanja in objektivna validacija kakovosti**: sodelujoče pogosto skrbi, da bodo v primeru zavrnitve posnetkov izgubili čas, porabljen za nalogo. Da bi omejili njihovo skrb, jasno definirajmo izločitvene kriterije, ki morajo temeljiti na vnaprej znanih objektivnih merilih. Na primer: Pregledali bomo 10 % oddanih posnetkov. Če bo več kot 50 % vaših validiranih posnetkov neustreznih, bomo zavrnili vse. S tem bomo izločili tudi goljufe.;
- **takojšnje povratne informacije o kakovosti in možnost pritožbe**: Bodimo ažurni ter govorcem zagotovimo vpogled v zavrnjeno delo.
- **zagotovimo kontekst naloge**: pojasnimo kontekst naloge in njen namen;
- **omejimo anonimnost:** naročnikom anonimnost omogoča, da nekaznovano zavrnejo ustrezno opravljeno delo, govorcem pa goljufanje.

Pri obravnavi etičnih vidikov moramo nasloviti še vprašanje izključenosti posameznih skupin, saj lahko preveliko zanašanje na nove tehnologije privede do izključitve tistih udeležencev, ki do teh tehnologij nimajo dostopa (Rutten idr., 2017).

70 Angleški izraz *collaborative science* prevajamo kot sodelovalna znanost.

4.5 Validacija kakovosti

Zagotavljanje kakovosti posnetkov, ki jih pridobimo na daljavo, je v nasprotju s snemanjem v studiu ali na terenu, zahtevnejše, saj samega procesa snemanja ne moremo neposredno nadzorovati. Pozornost moramo zato usmeriti k iskanju načinov, da bi tveganja za nekakovostne posnetke čim bolj zmanjšali. Ker je zagotavljanje ustrezne kakovosti pri zahtevnejših nalogah večje, je priporočljivo, da kompleksnejšo nalogo razdelimo na več podnalog, čeprav to pomeni več administrativnega dela in koordiniranja (gl. poglavje 4.1). Nadzor kakovosti lahko sicer izvajamo pred, med ali po tem, ko udeleženci opravijo naloge, vpeljava validacije pa navkljub dodatnim stroškom zagotavlja, da je delo udeležencev validirano na veljavni podlagi in da v čim večji meri izločimo pomanjkljive rezultate (Eskénazi idr., 2013), (Rutten idr., 2017).

PRED ODDAJO: Udeležencevo usposobljenost za kakovostno opravljanje dela lahko preverimo preko nalog, ki jih je opravil v preteklih projektih, ali njegove uspešnosti pri reševanju vzorčnih primerov za aktualno nalogo. Na ta način preverimo tudi njegovo razumevanje navodil in izločimo goljufe.

MED ODDAJO: Za ocenjevanje kakovosti lahko uporabimo spletno filtriranje. Druga možnost je t. i. zlati standard, kar pomeni, da ustrezne rešitve vnaprej podajo strokovnjaki (slednji običajno opravijo cca. 10 % nalog). Kakovost rezultatov udeležencev nato primerjamo z zlatim standardom.

PO ODDAJI: Tudi v tej fazi lahko uporabimo zlati standard. Naslednja možnost je medsebojna primerjava rezultatov udeležencev. Ta pristop temelji na predpostavki, da večina udeležencev ustrezno reši nalogo, zato lahko npr. preko glasovalne sheme, na kateri večina udeležencev razen nekaj izjem glasuje isto, na osnovi visokega odklona od povprečja identificiramo goljufe ali posameznike, ki niso ustrezno usposobljeni. Tretja tehnika nenadzorovanega pristopa pri nadzoru kakovosti je vpeljava druge skupine, katere naloga je, da pregleda in oceni rezultate dela prve skupine. Ena od možnosti je tudi navzkrižna validacija med govorci samimi, pri čemer moramo paziti, da posledično ne prihaja do posnemanja, ko govorci sami rešujejo nalogo.

4.6 Promocija in motivacija

Da bi spodbudili zanimanje občanov za kontinuirano sodelovanje v projektih, ki predvidevajo prostovoljno sodelovanje pri projektih s področja razvoja znanosti ali osebno angažiranost, je smiselno vključiti lokalne teme. Pri dolgotrajnih projektih si namreč moramo prizadevati, da dlje časa obdržimo iste prostovoljce, saj zaradi več izkušenj zagotovijo zanesljivejše podatke. Nuditi jim moramo organizacijsko podporo, npr. ažurno odzivanje na vprašanja (*helpdesk*) in vzpostavitev spletne skupnosti. Z večanjem zahtevnosti nalog lahko dosežemo, da te dolgoročno ohranijo privlačnost tudi za posameznike, ki so naprednejši in bi sicer lahko prekinili sodelovanje zaradi naveličanosti. Za zagotovitev uspešnosti projekta je treba sistematično raziskati najustreznejše motivacijske dejavnike glede na različne ciljne družbene skupine in v skladu s tem načrtovati komunikacijski načrt, pri promocijskih aktivnostih pa upoštevati dejstvo, da bodo aktivneje sodelovali občani, ki opravljajo poklice, vsebinsko povezane s tematiko projekta (Rutten idr., 2017).

Pri vseh treh izbranih primerih, CorCenCC, Anneta kõnet in Lahjoita puhetta, motivacija za sodelovanje ni izhajala iz finančnih nagrad, temveč so govorci sodelovali prostovoljno, k čemur so jih v največji meri spodbudile uspešne medijsko podprte promocijske kampanje in vključevanje humorja. To potrjuje še ena uspešna evropska kampanja zbiranja govorjene nizozemščine, Sprekend Nederland (Van Leeuwen idr., 2016), kjer so dokazali, da medijski dogodki, v katerih je omenjeno pridobivanje govora, povečujejo količino podatkov, ki jih udeleženci prispevajo. Število gledalcev medijskega dogodka, ki so ga analizirali, je približno 180.000, vključno s spletnimi ogledi po oddaji. Število udeležencev, ki jih je ta medijski dogodek spodbudil k sodelovanju, je bilo ocenjeno na 3000, kar pomeni, da je okoli 1,6 % gledalcev po ogledu TV-oddaje sprejelo odločitev za sodelovanje. Oglasi in medijski nastopi v analiziranih projektih CorCenCC, Anneta kõnet in Lahjoita puhetta so izpostavljali dvoje: po eni strani prizadevanje raziskovalne skupnosti za enakovreden status izbranega jezika v sklopu zgodovinsko prelomnega trenutka razvoja umetne inteligence, na kateri temeljijo rešitve, kot so razpoznavalniki govora, prevajalniki,

pogovorni agenti in klepetalniki, kot tudi skrb za ohranitev narečja ali nacionalnega govorjenega jezika. Pridobili so širšo podporo države in akademskega okolja, brez česar pri pridobivanju govornih virov sicer ne bi bili tako uspešni.

5 Zaključek

V prispevku smo se osredotočili na pridobivanje izključno govornih virov, ki so zaradi varovanja osebnih podatkov in avtorskih pravic, pravice do zasebnosti in dejstva, da gre za biometrični podatek, ki spada med občutljive podatke, posebej zahtevni za pridobivanje in obdelavo. Poleg pravnih vidikov moramo raziskovalci biti pozorni tudi na neizogiben pojav Labovega paradoksa, zaradi katerega je izjemno zahtevno zbirati spontani govor, kakršnega bi govorci uporabljali v zasebnih situacijah, ko niso opazovani, če vemo, da ga lahko analiziramo zgolj s sistematičnim opazovanjem ob upoštevanju demografske uravnoteženosti in drugih komunikacijskih dejavnikov. Glavna pomanjkljivost, ki smo jo opazili pri analiziranih projektih CorCenCC, Anneta kõnet in Lahjoita puhetta, je v tem, da so govorci večinoma oddajali kratke posnetke govora enega samega govorca, kar zmanjšuje verjetnost, da bi bil govor na posnetkih zares spontan in avtentičen.

Pri načrtovanju platforme za pridobivanje govornih virov moramo razmisliti o načinih, kako pri govorcih vzbuditi zaupanje, da bo njihovo sodelovanje anonimno[71], ter kako nalogo oblikovati na način, da bo dovolj privlačna in enostavna, vendar ne preveč dolgočasna, da bi odvrnila zahtevnejše uporabnike. Navodila za snemanje morajo biti jasna in nedvoumna, pred, med ali po oddaji posnetkov pa moramo načrtovati tudi validacijo kakovosti.

Brez tega, da bi bili projekti pridobivanja govornih virov na daljavo na nacionalni ravni prepoznani kot projekti v javnem interesu in zato podprti z obsežnimi medijsko podprtimi kampanjami, ne moremo

71 Če rezultati projekta niso sami po sebi občutljivi biometrični osebni podatki, lahko priznavanje avtorstva zgolj profesionalnim, ne pa tudi občanskim raziskovalcem, pomeni namerno izključevanje. Pri projektih občanske znanosti je zato pomembna skromnost profesionalnih raziskovalcev, ki se izraža v priznavanju prispevka vseh sodelujočih občanskih raziskovalcev z njihovo vključitvijo v seznam avtorjev pri objavi rezultatov (Mlinar, 2021), kar zanje predstavlja dodatno obliko priznanja.

realno pričakovati, da bomo v omejenem časovnem obdobju lahko zbrali veliko količino govornih posnetkov. Celo če projekte spremljajo intenzivne promocijske aktivnosti, trajajo ti zgolj omejen čas, po tem pa usahnejo. Enako zahtevno kot pridobiti govorce za sodelovanje, je dolgoročno ohraniti njihov interes za sodelovanje.

Jezik je namreč živ organizem, ki se spreminja tako, kot se spreminjajo družba in ljudje, ki ga govorijo. To je še toliko bolj očitno pri govorjenem jeziku, ki ga uporabljamo za naše vsakodnevno delovanje, zaradi česar se zbiranje govornih virov nikoli ne konča. Raba govorjenega jezika se od pisnega pomembno razlikuje, in šele kakovostni govorni viri nam omogočajo, da učinkovito analiziramo njegove fonetične, fonološke, prozodične in dialektološke značilnosti. Raziskati in zapisati bi bilo treba slovnico govorjene slovenščine ter sistematično raziskati leksikalne, pragmatične in diskurzne pojave, kot so diskurzni označevalci ter govorna dejanja, v spontano tvorjenem govoru.

V prispevku smo se dotaknili številnih izzivov, ki jih pred raziskovalca postavlja načrtovanje in izgradnja portalov in mobilnih aplikacij za izgradnjo obsežnih govornih baz, a ključni dejavnik za uspeh katere koli tovrstne pobude je motiviranje govorcev, da se sploh odločijo, svoj govor posneti in oddati. Razumevanje različnih motivacijskih dejavnikov, ki pri občanskih raziskovalcih spodbudijo namen za sodelovanje, je zato temelj oblikovanja strategij, prilagojenih za zbiranje govorjenega jezika od različnih demografskih skupin in v različnih funkcionalnih pojavnostih (Rutten idr., 2017), (Lyding idr., 2022). Zaradi pomanjkanja raziskav, ki obravnavajo modele za motiviranje govorcev k oddaji posnetkov govora, tako na domači kot mednarodni ravni, bi bilo smiselno to področje v prihodnje podrobneje raziskati.

Zahvala

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

Literatura

Arhar Holdt, Š., Kosem, I., & Kučuk, M. S. (2022). Metode in orodja za lažjo pripravo korpusov usvajanja jezika. V N. Pirih Svetina in I. Ferbežar (ur.), *Simpozij Obdobja 41: Na stičišču svetov: slovenščina kot drugi in tuji jezik* (str. 23–30). Pridobljeno s https://centerslo.si/wp-content/uploads/2022/11/Arhar-Holdt-et-al_Obdobja-41.pdf

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

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

Cieri, C., Corson, L., Graff, D., & Walker, K. (2007). Resources for new research directions in speaker recognition: the mixer 3, 4 and 5 corpora. V *Interspeech*, *8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27–31* (str. 950–953). doi: 10.21437/Interspeech.2007-340

Cieri, C., Miller, D., & Walker, K. (2004). The Fisher corpus: A resource for the next generations of speech-to-text. V *LREC* (Vol. 4, str. 69–71).

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters, 4*(6), 433–442. doi: 10.1111/j.1755-263X.2011.00196.x

Čibej, J., Robida, N., & Krek, S. (2024). Nadgradnja Digitalne slovarske baze za slovenščino in Slovenskega oblikoslovnega leksikona Sloleks s podatki o govorjeni 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. Pridobljeno s https://press.um.si/index.php/ump/catalog/book/898/chapter/46

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.

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.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don‘t pay at all. *The Quarterly journal of economics*, *115*(3), 791–810. Pridobljeno s https://www.jstor.org/stable/pdf/2586896.pdf

Kaufmann, N., Schulze, T., & Veit, D. (2011). More than fun and money. Worker motivation in crowdsourcing–a study on mechanical turk. Pridobljeno s 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., Loizides, F., Neale, S., Anthony, L., & Spasić, I. (2021). Developing computational infrastructure for the CorCenCC corpus: the national corpus of contemporary Welsh. *Language Resources and Evaluation*, *55*, 789–816. Pridobljeno s https://link.springer.com/content/pdf/10.1007/s10579-020-09501-9.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

Lindén, K., Jauhiainen, 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 Kritko Znanosti, Domisljijo 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

Oinas-Kukkonen, H. (2008). Network analysis and crowds of people as sources of new organisational knowledge. *Knowledge Management: Theoretical Foundation* (str. 173–189).

Oxford English Dictionary. (2014). Citizen science. Oxford: Oxford University Press.

Poesio, M., Chamberlain, J., & Kruschwitz, U. (2017). Crowdsourcing. *Handbook of linguistic annotation* (str. 277–295).

Quinn, A. J., & Bederson, B. B. (2011, May). Human computation: a survey and taxonomy of a growing field. V *Proceedings of the SIGCHI conference on human factors in computing systems* (str. 1403–1412).

Republika Slovenija. (1995). Zakon o avtorski in sorodnih pravicah (ZASP). *Uradni list RS, št. 21/95*. Pridobljeno s https://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO403

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

Surowiecki, J. (2005). The wisdom of crowds/james surowiecki. *NY.: Anchor*.

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

Uredba (EU) 2016/679 Evropskega parlamenta in Sveta z dne 27. aprila 2016 o varstvu posameznikov pri obdelavi osebnih podatkov in o prostem pretoku takih podatkov (GDPR). Pridobljeno s https://eur-lex.europa.eu/legal-content/SL/TXT/?uri=CELEX%3A32016R0679

Van Leeuwen, D. A., Hinskens, 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., Bizjak, A., Žgank, A., Maučec, M. S., Trojar, M., Gros, J. Ž., ..., & Dobrišek, S. (2024). Strategies for managing time and costs in speech corpus creation: insights from the Slovenian ARTUR corpus. *Language*

Resources and Evaluation, 1–26. Pridobljeno s https://link.springer.com/content/pdf/10.1007/s10579-024-09792-2.pdf

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 diskurznih 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.

Wieczorkowska, A. (2025). Methodology for Obtaining High-Quality Speech Corpora. *Applied Sciences, 15*(4), 1848. Pridobljeno s https://www.mdpi.com/2076-3417/15/4/1848

The role of citizen science in crowdsourced collection of speech resources in Slovenian

Speech resources of adequate quality are crucial for the development of speech technologies and spoken language research, but they are still scarce in the field of spontaneously produced spoken Slovenian due to the complexity of collection. The production of speech corpora and databases is costly, which is why researchers are increasingly recognizing the potential of citizen science. By leveraging crowdsourcing and other methods, it enables the efficient remote collection of large amounts of speech data. In this paper, we discuss the key factors - technical, financial, legal, ethical and motivational - that need to be considered when designing a sustainable and scalable system for speech acquisition. Based on a literature review, analysis of existing methods and global initiatives for collecting speech resources, we provide recommendations suitable for implementation in the Slovenian context.

Keywords: speech resources, speech acquisition, citizen science, crowdsourcing, spontaneous speech, spoken Slovenian