Zbiranje gradiv za govorne korpuse med Scilo in Karibdo

Darinka VERDONIK
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

Povzetek

Govorni korpusi niso pomembni samo za tehnološki razvoj, ampak tudi za sodobno jezikoslovje. Ker zahtevajo velik časovni vložek, mora biti njihovo načrtovanje toliko bolj premišljeno. V prispevku se osredotočamo na prihodnji razvoj govornih korpusov in iščemo odgovor na vprašanji: Kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih? Katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom? Med bolj aktivnimi uporabniki govornih korpusov so mnoge jezikoslovne discipline kot tudi govorne in semantične tehnologije. V obstoječih slovenskih govornih korpusih že obstaja večja količina gradiv za medijski, parlamentarni in akademski govor, manjka pa avtentičnih vsakdanjih govornih interakcij, kjer bi bila potrebna bolj podrobna regionalna pokritost, visoka kvaliteta posnetkov in zajem videa, kjer je mogoče. V Sloveniji je problem nekontinuirano snemanje v izredno kratkih časovnih obdobjih, pri čemer se veliko sredstev izgublja za koordiniranje množice sodelavcev ter ni časa za podrobno načrtovanje in pripravo orodij za bolj učinkovito delo.

Ključne besede: govorni viri, razpoznavanje govora, snemanje, uporabniki

Abstract

Speech corpora are important for technological development and for modern linguistics. They require a large investment of time, therefore their planning must be all the more thoughtful. In this paper, we focus on the future development of Slovenian speech corpora and seek answers to the following questions: Who are the users of speech corpora and what are their needs for data? What are the practices of collecting data for speech

corpora and how can we synergistically address as many different needs as possible with a single source? Among the more active users of speech corpora are many linguistic disciplines as well as speech and semantic technologies. In the existing Slovenian speech corpora, there is already a large amount of media, parliamentary and academic speech. There is a lack of authentic everyday speech interactions, which would require more detailed regional coverage, high quality recordings and video capture where possible. In Slovenia, the problem is non-continuous recording in extremely short periods of time where a lot of resources are wasted on coordinating a multitude of collaborators while there is no time for detailed planning and preparation of tools for more efficient work.

Keywords: speech resources, speech recognition, recording, users

1 Uvod

Slovenščina se po jezikovnotehnološki podprtosti uvršča na rep držav s fragmentarno tehnološko podporo. Z vidika pripravljenosti na digitalno prihodnost je primerljiva z bolgarskim, slovaškim, hrvaškim, baskovskim, velškim, galicijskim in islandskim jezikom (Giagkou idr., 2023: 81). Na tehnološko podprtost jezikov seveda vplivajo razni socioekonomski in politični dejavniki in razumljivo je, da se po podprtosti tehnologij slovenski jezik nikoli ne bo mogel primerjati z nemškim, francoskim ali španskim jezikom, da angleškega ne omenjamo; vsekakor pa je treba ohranjati prizadevanja, da postane naš jezik digitalno podprt vsaj primerljivo zahodnoslovanskim in skandinavskim jezikom. Tehnološka oziroma digitalna podprtost jezika vključuje širok spekter jezikovnih tehnologij in virov, od katerih so bili mnogi podprti v projektu Razvoj slovenščine v digitalnem okolju.[1] V tem prispevku se osredotočamo samo na področje govornih virov, natančneje govornih korpusov oz. govornih baz. Za slovenščino sta na tem področju potekali do zdaj dve večji kampanji, obe zelo kratkoročni. V okviru projekta Sporazumevanje v slovenskem jeziku, ki ga je v obdobju 2008–2013 omogočilo Ministrstvo

1 https://slovenscina.eu

za izobraževanje, znanost in šport ob podpori sredstev iz Evropskega socialnega sklada, je v letih 2009–2010 nastal referenčni govorni korpus Gos (Verdonik idr., 2013) v obsegu 112 ur/1 mio. besed. Sledilo je desetletno zatišno obdobje z minimalnimi vlaganji v govorno infrastrukturo do leta 2020, ko je Ministrstvo za kulturo s pomočjo sredstev iz Evropskega sklada za regionalni razvoj spodbudilo projekt Razvoj slovenščine v digitalnem okolju, v katerem je v dveh letih in pol nastala govorna baza in korpus Artur v obsegu 1000 ur, kjer pa ne gre več samo za korpusne podatke, ampak je polovica gradiva po pisni predlogi govorjen in posnet govor, slabih dvesto ur pa ostaja brez transkripcij, samo s posnetki. S pomočjo gradiv iz Arturja je v okviru projekta Razvoj slovenščine v digitalnem okolju tudi referenčni govorni korpus Gos zrasel za več kot dvakrat, na 300 ur oz. 2,4 mio. besed.

Govorni viri niso pomembni samo za tehnološko podprtost jezika (predvsem avtomatsko razpoznavanje govora), čeprav je ta danes v središču pozornosti in nas upravičeno skrbi. Enako pomembni so za sodobno jezikoslovno znanost. Spoznavanje svojega jezika, instrumenta, prek katerega komuniciramo in se povezujemo v skupnost(i), je eden temeljnih humanističnih postulatov. Čeprav ne prinaša neposrednih ekonomskih učinkov, pomeni opazovanje jezika, človeške komunikacije in interakcije preusmeritev pozornosti nazaj k človeku in k temu, kar nas povezuje. Pomeni vrnitev znanosti nazaj k njenim primarnim izhodiščem, stran od prevladujoče kapitalistične ideologije, v kateri tudi znanost vse bolj pristaja v vlogo orodja za dodatno gospodarsko rast, ki dolgoročno izčrpava tako planet kot človeka. Najlažje, najbolj zanesljivo in najbolj široko dostopno lahko jezik in komunikacijo opazujemo prav v govornih virih, skozi posnetke govora v številnih vsakdanjih situacijah, ki smo jim izpostavljeni ali v njih aktivno sodelujemo. Korpusno jezikoslovje že več desetletij aktivno uporablja korpusne podatke v slovaropisju in slovnici (Adolphs in Carter, 2013). Tudi dialektologija v svojih raziskavah vse pogosteje posega po korpusnih podatkih (Goláňová idr., 2013; Šumenjak, 2012). Jezikoslovne discipline, ki so bolj povezane s sociološkimi (sociolingvistika, etnografija komunikacije, konverzacijska analiza) ali kognitivnimi

disciplinami (pragmatično jezikoslovje), prav tako temeljijo vedno več svojih raziskav na korpusnih podatkih (Aijmer in Rühlemann, 2015; govorni viri v okviru TalkBanka[2]), enako številne discipline, povezane z različnimi zdravstvenimi stanji ali razvojem jezika (CHILDES[3], PhonBank[4]) (MacWhinney, 2018). Podobno velja za uporabno jezikoslovje oz. bolj specifično za učenje jezika (CLARIN L2 Learner Corpora[5]). Korpusni podatki so lahko v pomoč tudi fonetičnim/fonološkim disciplinam, vključno s pravorečjem (Verdonik, 2021).

Nadaljnji razvoj govornih korpusov za slovenščino je torej ključen tako za razvoj njene tehnološke podprtosti kot tudi za razvoj slovenskega jezikoslovja. Prvi naslednji mejniki so 5 in 10 mio. besed v referenčnem govornem korpusu slovenščine ter dodatni področno specializirani in na višjih jezikovnih ravneh označeni (manjši) govorni korpusi. Ker pa govorimo o virih, ki zahtevajo velik časovni vložek, je toliko večja potreba po natančnem premisleku o njihovih potencialnih uporabnikih, njihovih potrebah, najbolj učinkovitih načinih zbiranja gradiv in ovirah pri tem, da lahko poteka razvoj v smeri, ki je najbolj smiselna in združuje čim več zaželenih učinkov. Zato sta vprašanji, ki ju naslavljamo v tem prispevku: Kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih? Katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom?

2 Vzorčni tuji modeli in obstoječi govorni korpusi za slovenščino

Govorni korpusi obstajajo za večino evropskih jezikov. Vse bolj intenzivno se govorni viri (ne samo korpusni, ampak tudi kot baze govora s posnetki po pisnih predlogah) razvijajo tudi za druge jezike s premalo jezikovnimi viri, t. i. »under-resourced languages« (npr. centralni kurdski jezik – Veisi idr., 2022; lugandski jezik – Mukiibi idr., 2022; švicarska nemška narečja – Plüss idr., 2022). Po drugi strani

2 https://www.talkbank.org
3 https://childes.talkbank.org
4 https://phon.talkbank.org
5 https://www.clarin.eu/resource-families/L2-corpora

so jeziki velikih jezikovnih skupnosti, kot so v Evropi angleška, nemška, francoska ali španska, tisti, ki pogosto služijo kot zgled za ostale jezike. V tem razdelku bomo podrobneje pogledali angleški govorni korpus, kjer je govorna komponenta British National Corpusa že od začetkov korpusnega jezikoslovja pogost referenčni vir za ostale jezike. Tehnološko dokaj zadovoljivo je med evropskimi jeziki podprta še nemščina, kjer je med govornimi korpusi najbolj prepoznaven korpus FOLK. Kot tretji primer bomo izbrali korpus, ki je slovenščini primerljiv po socio-ekonomskem statusu države, po številu govorcev in je prav tako slovanski jezik, to je slovaški govorni korpus.

2.1 Vzorčni tuji korpusi

British National Corpus (BNC) je bil pionir ne samo kot pisni, ampak tudi kot govorni korpus, saj je že tri desetletja nazaj, 1994, izdal govorno komponento v obsegu 4,2 milijona besed – t. i. BNC1994. Takrat je bil to eden prvih javno dostopnih korpusov svoje vrste. BNC1994 vključuje demografsko uravnotežen in besedilnovrstno uravnotežen del ter skuša biti reprezentativen za govorjeno britansko angleščino. Predstavljal je pomemben vir za raziskave v različnih jezikoslovnih disciplinah, od slovnice (Rühlemann, 2006; Smith, 2014) do sociolingvistike (McEnery, 2005; Säily, 2011; Xiao in Tao, 2007), konverzacijske analize (Rühlemann in Gries, 2015) in pragmatike (Wang, 2005; Capelle idr., 2015; Hatice, 2015), pa tudi za raziskave učenja jezika (Alderson, 2007; Flowerdew, 2009) in drugo. Love idr. (2017) navajajo kot razloge za njegovo popularnost, da obsega ortografsko zapisane podatke v velikem obsegu, da gre za splošen, reprezentativen vzorec govorjenih besedil in predvsem da je javno dostopen. Skozi čas pa je postajalo vedno bolj problematično, da se za raziskave današnje govorjene angleščine uporablja več kot dve desetletji staro gradivo. V obdobju od 2012 do 2016 je bil zato govorni del BNC nadgrajen s Spoken BNC2014, ki pa vsebuje samo demografsko uravnotežen del s posnetki v neformalnih kontekstih, ne pa tudi besedilnovrstno uravnoteženega dela. Kot razlog za to navajajo avtorji (Love idr., 2017), da po njihovem

opažanju obstaja večja potreba in zahteva po gradivu iz konverzacije in da imajo raziskovalci, ki želijo raziskovati britansko angleščino v specifičnih kontekstih, svoje lastne, specializirane korpuse oz. so takšni korpusi javno izdani (npr. BASE – korpus britanske govorjene akademske angleščine). Spoken BNC2014 obsega 11,5 milijona besed, 1251 posnetkov in sodelujočih 668 govorcev. Gre za vsakdanji neformalni govor, govorci pa so uravnoteženi glede na spol, starost, socio-ekonomski status in regijo. Za uporabnike je na voljo prek konkordančnika Sketch Engine.

Nemški govorni korpus FOLK (Schmidt, 2014) podobno kot angleški BNC izhaja iz potrebe po odprto dostopnih virih, ki so bili v času zasnove korpusa za nemščino redki in omejeni na specifične situacije, ne pa reprezentativni. Namenjen je tako za raziskovalne potrebe kot tudi za uporabo v šolskem okolju (Schmidt, 2016). Sledi ciljem, da pokrije širok nabor govornih interakcij v zasebnih, institucionalnih (predvsem interakcije v izobraževanju ter v delovnem okolju) in javnih situacijah (mediji). Kontrolirati skušajo tudi demografske kriterije, kot so regija, spol in starost govorcev. Da dokumentirajo komunikacijske prakse, vedno posnamejo in vključijo celotno interakcijo, ne samo izbranih segmentov. Ker so vidne oblike komunikacije pogosto enako pomembne kot slišne, skušajo v zadnjem času vedno, kjer je mogoče, zajeti tudi video posnetek, ne samo avdio. Projekt se je začel leta 2008 (Schmidt, 2016). V prvi izdaji je korpus obsegal 1 mio. besed (Schmidt, 2014), ker pa gre za dolgoročni načrt, se korpus ves čas dograjuje. Julija 2022 (verzija 2.18) je korpus FOLK obsegal 3,2 milijona pojavnic oz. 336 ur posnetkov, od tega 151 ur z videom (Schmidt, 2023). Korpus FOLK je za uporabnike dostopen prek konkordančnikov DGD (Datenbank für Gesprochenes Deutsch).[6]

Tudi korpus govorjene slovaščine s-hovor sodi v sklop t. i. velikih reprezentativnih govornih korpusov (Garabík, 2023). Prvič je bil izdan decembra 2008 in se od takrat ves čas nadgrajuje. Trenutna različica s-hovor-7.0 obsega 851 ur posnetkov oziroma 7,8 mio.

6 https://dgd.ids-mannheim.de/DGD2Web/jsp/Welcome.jsp

pojavnic.[7] Približno tretjino posnetkov za korpus je prispeval slovaški Nacionalni institut spomina (Nation's Memory Institute – UPN). 4,2 mio. pojavnic so posnetki iz drugih virov, poleg medijev in parlamenta je zelo veliko tudi terenskih posnetkov, pri čemer upoštevajo osrednje demografske kriterije (spol, starost, izobrazbo, regijo izvora in skladnost s standardnim jezikom), pa tudi vrsto diskurza (Garabík in Rusko, 2007). Korpus je osredotočen na splošni govorjeni jezik in ne vključuje dialektalnega govora. Za uporabnike je dostopen prek konkordančnika Sketch Engine.

2.2 Slovenski govorni korpusi

Slovenci smo potrebe po reprezentativnih govornih korpusih hitro zaznali (Stabej in Vitez, 2000), do prve izvedbe pa je prišlo desetletje kasneje (Verdonik idr., 2013). V letu 2023 je bil izdan še en pomemben govorni vir: govorna baza in korpus Artur (Verdonik idr., 2023a; Verdonik idr., 2023b), katerega cilj je bil zagotoviti gradiva za razvoj avtomatskega razpoznavanja govora za slovenščino. Gradiva iz Arturja so bila uporabljena tudi za nadgradnjo referenčnega govornega korpusa Gos v različico 2.x (Verdonik idr., 2023c), kjer so bili združeni obstoječi viri z namenom zagotavljanja nadgradnje reprezentativnega korpusa za jezikoslovne raziskave. Izdelan je bil tudi prenovljen uporabniško prijazen konkordančnik,[8] ki omogoča uporabo korpusa tudi v šoli oz. nasploh zunaj raziskovalne sfere. V Tabeli 1 so predstavljene osnovne informacije o govorni bazi in korpusu Artur. Kot vidimo iz nje, približno polovica baze vključuje posnetke branja povedi po pisnih predlogah. Čeprav gre za demografsko uravnotežen nabor velikega števila govorcev, pa besedila izhajajo iz pisnih virov (Žganec Gros idr., 2022). Od preostale polovice precejšen delež nima transkripcij, ampak samo posnetke. Največji del na novo zbranih posnetkov z ročno narejenimi kvalitetnimi zapisi govora tako obsega gradivo iz Državnega zbora Republike Slovenije, torej parlamentarni govor. Preostanek se deli dokaj enakomerno na

7 https://korpus.sk/en/corpora-and-databases/snc-corpora/publicly-available-snc-corpora/corpus-of-spoken-slovak/

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

javni govor in nejavni govor, pri čemer nejavni govor v veliko primerih ni interakcija, ampak razlaganje ali opisovanje po vnaprej določenih vsebinskih iztočnicah. Poleg teh večjih sklopov vključuje Artur še nekaj manjših, prilagojenih potrebam razvoja tehnologij. Na novo pridobljeno gradivo je torej z vidika potreb jezikoslovja zelo omejeno, njegova bistvena prednost v primerjavi s posnetki iz prvega vala snemanja v letu 2010 pa so kvalitetni avdio posnetki, ki omogočajo raziskave in razvoj na podlagi analize ali procesiranja avdio signala.

Tabela 1: Osnovni podatki o govorni bazi in korpusu Artur.

	Št. govorcev	Št. posnetkov	Trajanje v urah
Brane povedi	884	257.942	485
Črkovanje	345	676	10,5
Studijski posnetki za sintezo	1	10.109	27
Pogovori/opisovanje	263 (181 trans.)	301 (210 trans.)	94 (61 trans.)
Pametni dom (za avtomatsko razpoznavanje govora)	148 (148 trans.)	195 (189 trans.)	7,5 (7 trans.)
Opis obraza (za avtomatsko razpoznavanje govora)	125 (86 trans.)	125 (86 trans.)	10 (6 trans.)
Mediji, javni dogodki	811 (240 trans.)	400 (100 trans.)	207 (62 trans.)
Parlament	158	2799	201 (vse trans.)
Skupaj	**2222 (1586 trans.)**	**286.064**	**1067 (884 trans.)**

V letu 2023 je bila izdana tudi nadgrajena različica korpusa Gos, ki vključuje zbir vsega, kar je bilo od njegove prve izdaje na voljo pod ustrezno licenco in je bilo mogoče smiselno vključiti v reprezentativni govorni korpus, ne da se pretirano poruši uravnoteženost gradiv. Korpus Gos 2.x tako obsega sledeče vire in vsebine:

- Gos 1.1: 1 mio. besed/112 ur; avtentični posnetki, izogibanje branemu govoru, vsebuje besedilnovrstno in demografsko uravnotežen del po vzoru BNC; posnetki so pogosto slabše kvalitete, v zasebnem delu je izredno veliko segmentov s prekrivanjem govora dveh ali več govorcev;
- GosVL: 180.000 besed/22 ur; 55 predavanj ali delov predavanj, izbranih s portala Videolectures.net z upoštevanjem

uravnoteženosti po vedah in demografskih značilnosti govorcev, kolikor je bilo o njih mogoče sklepati iz posnetkov in na spletu dostopnih podatkov;

- Artur (1,2 mio. besed/185 ur):
  - javni govor, 422.000 besed/62 ur,
  - nejavni govor, 324.000 besed/61 ur,
  - parlamentarni govor, 450.000 besed/62 ur.

3 Uporabniki govornih korpusov in njihove potrebe po gradivih

V tem razdelku skušamo odgovoriti na vprašanje, kdo so uporabniki govornih korpusov in kakšne so njihove potrebe po gradivih. S pomočjo pregleda literature v mednarodnem prostoru in posebej tudi v slovenskem prostoru bomo ugotavljali, v katerih disciplinah pogosto posegajo po korpusnih podatkih, analizirali gradiva obstoječega referenčnega govornega korpusa Gos 2.x in ugotavljali, kje so pomanjkljivosti, ki jih je treba nasloviti ob prihodnjih nadgradnjah korpusa.

3.1 Uporabniki

Love idr. (2017) navajajo kot pomembne uporabnike govornega korpusa BNC slovnico, sociolingvistiko, konverzacijsko analizo, pragmatiko in učenje jezika kot drugega jezika. Schmidt (2016) posveti posebno pozornost šolskemu okolju in izobraževanju kot sicer neraziskovalnemu, a enako zainteresiranemu uporabniku govornih korpusov. Večinoma specializirani govorni korpusi v okviru projekta TalkBank opozorijo na uporabnike iz psihologije (razvoj govora) in medicine (npr. raziskave govora pri osebah z demenco, poškodbami desne hemisfere, travmatološkimi poškodbami možganov, afazijo, logopedskimi težavami). Tudi za potrebe dialektologije se večinoma razvijajo specializirani korpusi (Goláňová idr., 2013; Šumenjak, 2012). Fonetika in fonologija sta precej specifičen uporabnik, ki bolj kot same korpuse potrebujeta določene jezikovnotehnološke servise za avtomatsko predpripravo korpusnih podatkov za analizo, kot

jih ponuja na primer WebMAUS.[9] Na drugi strani so velik in zelo aktiven uporabnik govornih korpusov tehnologije: avtomatsko razpoznavanje govora (Gril idr., 2021), klasifikacija (Vlaj in Žgank, 2023) in prepoznavanje govorcev (Ljubešić in Rupnik, 2022), procesiranje govorjenega jezika (Lee idr., 2021), govorjeni sistemi dialoga (Chen idr., 2021) itd.

Med razpoložljivimi govornimi korpusi je poleg referenčnih kar nekaj korpusov specializiranih, pri čemer prevladujejo korpusi parlamentarnega govora (Ogrodniczuk idr., 2020) in govor v akademskem okolju (Verdonik, 2018; korpus MICASE[10]), verjetno predvsem zaradi lahke dostopnosti tovrstnih podatkov v primerjavi z drugimi področji. Mednarodno eden najbolj pogosto procesiranih govornih korpusov je Switchboard (Godfrey in Hollimann, 1993)[11], ki vsebuje nekoliko specifično izzvane interakcije med dvema neznancema na eno od tem, ki so bile pripravljene vnaprej, torej delno simulirano, in ne avtentično govorno situacijo.

V slovenskem okolju je tradicija raziskovanja govorne interakcije v jezikoslovju šibka in raziskave v primerjavi s pisnim jezikom redke. Tehnološki uporabniki so v slovenskem prostoru morda nekoliko bolj aktivni uporabniki govornih korpusov in baz kot jezikoslovci. V letu 2023 je bila konferenca Slavistični znanstveni premisleki, ki jo organizira Oddelek za slovanske jezike in književnosti Univerze v Mariboru, posvečena tematiki infrastrukture za raziskave govora. Raziskovalci iz slovenskega prostora, ki so se odzvali, so naslavljali vprašanja govorne infrastrukture z vidika sociolingvistike, leksike, skladnje, jezikovnih tehnologij, dialektologije, pragmatike in učenja drugega jezika, med specializiranimi področji pa so med drugim izstopali parlamentarni govor, govorjeni jezik v literaturi, v gledališču, na radiu in televiziji (Krajnc Ivič, 2023). V primerjavi z mednarodnim prostorom v slovenskem ni zaznati uporabnikov specializiranih korpusov s področja razvoja govora, čeprav je to raziskovalno področje aktivno (Marjanovič Umek idr., 2006), in ne iz logopedije in drugih

9 https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic
10 https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase
11 https://catalog.ldc.upenn.edu/LDC97S62

disciplin, povezanih z medicino. O uporabi korpusov v šolstvu in splošni javnosti je po drugi strani kar nekaj razmislekov (Logar idr., 2023), kar se kaže tudi skozi tovrstni javnosti prilagojene konkordančnike, tudi za korpus Gos.[12]

3.2 Potrebe uporabnikov

Potrebe uporabnikov so tukaj obravnavane z vidika, da skušamo z enim osrednjim referenčnim korpusom zadovoljiti potrebe čim več različnih disciplin. Čeprav so potrebe včasih kontradiktorne, je v primeru manjših skupnosti to edini način, da se zagotovi gradivo za različne uporabnike, saj je razpoložljivih finančnih sredstev malo, potreben finančni in časovni vložek pa velik.

Potrebe uporabnikov v zvezi z govornim korpusom lahko razdelimo v več ravni. Prva se nanaša na vrste situacij, ki so zajete v govorni korpus. V ta namen lahko ločujemo specializirane in referenčne korpuse. Specializirani govorni korpusi za slovenščino zelo dobro pokrivajo parlamentarni govor (Pančur idr., 2020), deloma akademski govor (Verdonik, 2018), ostalih vrst govora pa tako rekoč ne, z izjemo majhnega, 1 uro trajajočega korpusa govora Koprive na Krasu (Šumenjak, 2012) kot do zdaj edinega primera dialektološkega korpusa za slovenščino. Znotraj referenčnega korpusa Gos je sicer še dokaj obsežno zastopan tudi akademski govor, vendar samo v javnih situacijah. Tudi medijski govor je široko zastopan v referenčnem korpusu Gos, pa tudi s specializiranimi viri (npr. BNSI Broadcast News, Žgank idr., 2005). Pomemben potencialni uporabnik govornih virov v slovenskem prostoru je dialektologija. V referenčnem govornem korpusu Gos ločevanje med dialektološkim in nedialektološkim gradivom ni vzpostavljeno. Ne v prvem ne v drugem snemalnem valu na terenu snemanje ni bilo osredotočeno samo na urbana središča, ampak je potekalo mešano po vaseh in mestih v vseh slovenskih regijah. Tudi sicer ni jasnih podatkov, kakšne so v manjših urbanih središčih razlike v govoru med mestom in okoliškimi vasmi. V referenčnem korpusu Gos najdemo tako posamezne primere zelo

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

narečnega govora (tudi iz zamejstva v vseh treh sosednih državah), vendar so to samo posamični naključno vključeni narečni govori. Kriteriji za zajem gradiv so se namreč v prvem snemalnem valu ravnali po registrskih enotah, v drugem pa po statističnih regijah. Čeprav je podobna praksa običajna (Love idr., 2017), bi jo bilo v prihodnje smiselno ponovno premisliti tudi z dialektološkega vidika.

Kot vidimo na primeru Spoken BNC2014, je ključen segment govornih korpusov za mnoge jezikoslovne discipline vsakdanja govorna interakcija v zasebnih situacijah. Za slovenščino je lahko ta segment še dodatno pomemben zaradi velike dialektalne razpršenosti. Te vsebine so bile v slovenskem govornem korpusu Gos kvalitetno pokrite v prvem valu snemanja, v drugem pa veliko manj zaradi zahtev tehnologij po visoko kakovostnih posnetkih brez prekrivanja govora. V drugem valu snemanja je tako veliko vsebin celo kar monoloških in niso primerne za raziskave interakcije. Po drugem valu snemanj torej beležimo v referenčnem govornem korpusu slovenščine pomanjkanje posnetkov avtentičnih vsakdanjih nejavnih in institucionalnih govornih interakcij. Razmisliti je treba tudi o morebitni vključitvi govornih situacij, ki do zdaj niso bile zajete v korpus Gos, najdemo pa zainteresirane raziskovalce (npr. gledališki govor, dramatika), ter posebno pozornost posvetiti vprašanjem otroškega in mladostniškega govora ter govora neprvih govorcev slovenščine.

Naslednje vprašanje zajemanja gradiv je odločitev o tem, kje se vključeni posnetek začne in konča. Kot vidimo pri Schmidtu (2023), obstajajo argumenti, da se vključijo posnetki celotne interakcije od začetka do konca. V prvi izdaji korpusa Gos ta praksa ni bila dosledno upoštevana, veliko bolj v drugem valu zbiranja gradiv. To je mogoče v primeru javne in institucionalne komunikacije, kjer imajo dogodki jasne začetke in konce. V vsakdanji zasebni komunikaciji pa začetki in konci niso nujno jasni, predvsem pa so ob snemanju začetki lahko obremenjeni z razlaganjem namena snemanja, nameščanjem naprav, podpisovanjem strinjanja in uvodno nervozo govorcev zaradi snemanja. Družabni dogodki lahko potekajo tudi več ur, z vmesnimi premori, spremembami prisotnih govorcev ipd. V primeru nejavnih

terenskih posnetkov odgovor, kaj šteje kot celotna interakcija, tako ni vedno enoumen.

Tretji vidik potreb uporabnikov glede gradiv se nanaša na modalnosti zajema in tehnično kvaliteto gradiv. V prvem valu snemanja za korpus Gos je bila tehnična kvaliteta v terenskih posnetkih pogosto nizka, prednost se je dajalo avtentičnosti situacije, čim manjši invazivnosti snemalnih naprav in preprostosti njihove uporabe. Takšni posnetki ne zadoščajo potrebam disciplin, kjer se procesira ali analizira avdio signal, zato je zahteva po višji kvaliteti posnetkov tako rekoč nujna. Posebej težavno je vprašanje hkratnega govora, ki je v vsakdanji interakciji široko prisoten, po drugi strani pa so segmenti s hkratnim govorom neprimerni za avdio procesiranje. V drugem valu snemanja je bil tako isti terenski pogovor ločen v dva posnetka, za vsakega govorca en. Slabosti takega načina sta pogosto prisoten presluh in fizična ločenost transkripcij. Na ta način gradiva niso primerno pripravljena za pragmatične, diskurzne in sociolingvistične analize. Pri snemalni tehnologiji je nadalje treba nasloviti tudi vprašanje video zajema. Mnogi govornokorpusni centri (prim. Schmidt, 2023) že nekaj časa zajemajo tudi video, ne samo avdia. Govorna komunikacija ni samo slušna, ampak v večini primerov tudi vidna in ob odsotnosti vizualnega dela ne moremo celostno analizirati govora, hkrati pa je za vizualne podatke vedno bolj zainteresirana tudi tehnologija (npr. za razvoj pogovornih agentov z vizualnim vmesnikom).

Četrti vidik se nanaša na vprašanja metapodatkov o govorcih in posnetih situacijah. Ta vprašanja so že bila podrobno obravnavana v Verdonik (2022). Vsekakor je treba upoštevati, da imajo discipline, kot so sociolingvistika, pragmatika, analiza diskurza, tudi dialektologija, potrebo po čim bolj natančnem opisu konteksta, zato je lahko možnost, da se kontekst vsakega posnetka opiše v nekaj stavkih, zelo dobrodošel dodaten podatek, čeprav ni strukturiran. Podobni opisi bi se lahko dodajali tudi za govorca, saj imajo discipline, kot je dialektologija ali sociolingvistika, potrebo po čim bolj natančni predstavitvi govorca in njegove podvrženosti različnim jezikovnim vplivom. Metapodatke o govorcih in posnetkih, ki jih popišemo, moramo pri tem ločevati od kategorij, ki jih postavimo kot kriterije za

zajem. Medtem ko so popisani podatki idealno čim bolj podrobni, so kriteriji za zajem v demografsko uravnoteženem delu korpusa praviloma: spol, starost, izobrazba, regija, prvi jezik. V angleškem Spoken BNC2014 najdemo kot kriterij še socio-ekonomski status, kar v slovenskem okolju že ob začetkih govornega korpusnega jezikoslovja ni bilo prepoznano kot primeren kriterij za naše okolje. Pač pa bi bilo glede na široko narečno razpršenost smiselno razmisliti o tem, ali je treba pri kriterijih za zajem določiti regijo govorcev bolj podrobno kot samo na ravni statističnih regij.

Nazadnje je vprašanje tudi, kakšne so potrebe glede zapisa govora. V slovenskih govornih korpusih je vzpostavljena praksa dvojnega ortografskega zapisa, pogovornega in standardiziranega, primerljivo kot v nemškem (Schmidt, 2023) ali slovaškem korpusu (Garabík, 2023). Vprašanja smotrnosti takšnega zapisovanja so naslovljena v Verdonik (v tisku) in končno priporočilo je, da se s tem nadaljuje. Nekatere discipline imajo potrebo po bolj natančnih, fonemskih ali fonetičnih zapisih, zlasti dialektologija, slovaropisje, fonetika in fonologija. Na tak način je seveda mogoče zapisati le manjši del gradiv, na primer učni korpus, kar bi bil smiseln korak v prihodnosti zlasti z namenom, da se vzpostavi servis za avtomatsko pretvorbo ortografskega v fonetični zapis, kot jo omogoča na primer WebMAUS. V zapisih se ves čas ohranjajo tudi zabeležke o nekaterih osnovnih neverbalnih dogodkih med govorom, kot so daljši premori, smeh, nerazumljiv govor, govor v tujem jeziku ipd.

Iz zgornjega pregleda je vidno, da so potrebe nekaterih uporabnikov govornih korpusov do določene mere nasprotne ena drugi. Največja težava je po eni strani potreba po kvalitetnem zvoku in videu z malo ali nič hkratnega govora, po drugi pa potreba po posnetkih avtentičnih govornih dogodkov. Z namestitvijo govorcev v studio, snemanjem na ločene kanale, govorjenjem »na ukaz« zagotovimo visoko kvaliteto posnetkov, a postavimo govorce v stresno in nenaravno situacijo, za katero ne moremo trditi, da so se govorci obnašali v njej enako, kot bi se v avtentičnem okolju. Avtentično okolje pa je lahko polno hrupov in šumov, govorci se vmes premikajo po prostoru, namestitev snemalne opreme v domače okolje govorcev je hkrati

tudi veliko večji vdor v zasebno življenje govorcev kot snemanje v studiu. Za doseganje kompromisa je zato potreben zelo premišljen izbor vsake snemalne situacije in govorcev, to pa ob intenzivnih snemalnih kampanjah, omejenih s kratkimi časovnimi roki, ni mogoče.

4 Prakse zbiranja gradiv za govorne korpuse

Na podlagi pregleda tuje literature o govornih korpusih ter informacij in izkušenj pri projektih, v katerih so se snemala gradiva za slovenske govorne vire, v tem razdelku predstavljamo možne načine pridobivanja gradiv in probleme, ki jih imamo pri tem v Sloveniji.

Gradiva za govorne korpuse prihajajo iz dveh bistveno različnih virov: prvi so že obstoječi, večinoma javno tako ali drugače predvajani ali dostopni posnetki, kot so posnetki medijskih hiš, na internetu, v parlamentu, v okviru različnih javnih dogodkov ipd. Za te posnetke je treba doseči dogovore z nosilci avtorskih pravic, ki so lahko institucije (npr. državna RTV, državni zbor, Arnes), podjetja (komercialne radijske in TV-postaje), posamezniki (medijski posnetki, če govorci niso prenesli pravic na medij, predvsem pa različni javni dogodki) ali tudi mednarodne korporacije (npr. Youtubova licenca). Če posamezen akter ne vidi jasnega lastnega interesa za sodelovanje, je velika možnost, da dogovarjanje ni uspešno. Za gradiva, ki jih uspemo pridobiti in skleniti dogovor z nosilci avtorskih pravic, je treba za vsak posamezen vir izvesti test zakonitega interesa. Govor sam po sebi je namreč bibliometrični podatek (Data Protection Working Party, 2003), in tudi če govorci v njem ne navajajo osebnih podatkov, kot sta ime in priimek, je treba zagotoviti ravnanje z gradivi skladno z zakonodajo.

Drugi sklop posnetkov so terenski posnetki vsakdanjih govornih interakcij. Kot smo videli v Razdelku 3.2, je za te posnetke interes zelo velik. Tradicionalno to poteka tako, da se angažirajo študenti, ki vsak prek svoje lastne socialne mreže nagovorijo govorce, da jih smejo posneti, in vsak govorec podpiše v ta namen pripravljeno izjavo, s katero se zagotovijo vse potrebne pravice za nadaljnje deljenje posnetkov (to vključuje dovoljenje za snemanje in uporabo

posnetka, privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov ter dovoljenje za uporabo avtorskih pravic). V drugem snemalnem valu v Sloveniji sta večji del terenskega snemanja prevzela najeta zunanja izvajalca, saj samo s pomočjo študentov v takšnem obsegu in kratkem časovnem roku ni bilo izvedljivo. Zunanji izvajalci so na primer podjetja, ki se ukvarjajo z avdio in/ali video produkcijo in pogosto tudi že imajo lastne sezname kontaktov ljudi, ki jih angažirajo kot govorce. Tretji način snemanja je, da povabimo govorce v studio in se tam pogovarjajo. Takim načinom se izogibamo, saj težko pridobimo ljudi iz oddaljenih krajev in bi bila potrebna večja finančna nagrada govorcem, kot jo običajno omogočajo razpoložljiva sredstva. Zanimiva alternativna možnost je mobilni studio, na primer ustrezno preurejen in opremljen kombi – tak način se je med drugim uporabljal pri snemanju branega govora za bazo Artur. V prihodnje bi bilo smiselno raziskati še tehnološko podprte pristope prek spletnih platform. Za množičenje v Sloveniji sicer ne obstaja že vzpostavljena skupnost, na katero bi se lahko obrnili, in dosedanji poskusi množičenja niso dali spodbudnih rezultatov (npr. na platformi Mozzilla CommonVoice so do julija 2023 zbrali samo 14 ur slovenskega govora, čeprav je bila platforma postavljena že leta 2017). Je pa vsekakor dandanes možnost, da govorci uporabijo lastne snemalne naprave (npr. pametne telefone), veliko bolj dostopna kot včasih in je lahko pridobivanje posnetkov na način, da jih govorci sami oddajo na neko spletno mesto, zanimiva rešitev. Tako v primeru množičenja kot v primeru snemanja v studiu je ključna težava motiviranje govorcev. Plačevanje honorarja govorcem za vsak oddan posnetek pomeni velik finančni vložek, ki lahko prestavlja tudi izredno obsežno administrativno delo in velike stroške oglaševanja, če gre za kratkoročno, intenzivno snemalno kampanjo.

Primerjava z vzorčnimi tujimi govornimi korpusi (gl. Razdelek 2.1) pokaže bistveno razliko s slovensko prakso: pri vseh navedenih tujih govornih korpusih gre za dolgoročne projekte, medtem ko smo v Sloveniji vse dosedanje gradivo v govornih korpusih posneli skupaj v dobrih treh letih, v dveh izredno intenzivnih snemalnih valih z dolgim vmesnim obdobjem brez financiranja in brez kakršnega koli

signala, kdaj se bo financiranje ponovno nadaljevalo. Na tak način v delo ni mogoče učinkovito vključevati študentov, kar je škoda med drugim za študijski proces na vsebinsko povezanih študijskih smereh. Finančni stroški so znatno višji zaradi izredno zahtevnega koordiniranja. Časa za podrobno načrtovanje snemanja in transkribiranja ter preučitev in pripravo podpornih orodij in okolij, ki bi lahko pohitrili delo in zmanjšali potreben čas za izvedbo posameznih korakov izdelave govornega korpusa, pa je premalo.

5 Diskusija in zaključek

V prispevku smo izhajali iz stališča, da je nadaljnji razvoj govornih korpusov za slovenščino ključen tako za razvoj njene tehnološke podprtosti kot tudi za razvoj slovenskega jezikoslovja. Zastavili smo si vprašanji, kdo so uporabniki govornih korpusov in kakšne so njihove potrebe glede gradiv ter katere so prakse zbiranja gradiv za govorne korpuse in kako lahko sinergično naslovimo čim več različnih potreb z enotnim virom. Kot bolj aktivne uporabnike govornih korpusov smo prepoznali jezikoslovne discipline leksikologijo, slovnico, sociolingvistiko, dialektologijo, konverzacijsko analizo, pragmatiko, uporabno jezikoslovje (učenje jezika kot tujega jezika); med tehnološkimi vedami predvsem govorne in semantične tehnologije; posamično so uporabniki tudi nekatere družboslovne (npr. razvojna psihologija) in naravoslovne discipline (logopedija ipd.); in šolstvo ter drugi neakademski uporabniki. Nekatere od navedenih disciplin potrebujejo specializirane korpusne vire, kljub temu pa smo skozi prispevek iskali možnosti pokrivanja potreb čim več disciplin skozi enoten, skupen govorni korpus, ki se lahko po potrebi deli na specializirane podenote. Ugotavljali smo, da v obstoječih slovenskih govornih korpusih že obstaja večja količina gradiv za medijski, parlamentarni in akademski govor. V prihodnje smo priporočali preusmeritev v terenski zajem avtentičnih vsakdanjih govorjenih situacij, kjer bi bilo smiselno zagotavljati bolj podrobno zastavljeno regionalno pokritost posnetkov, višjo kvaliteto posnetkov, kot je bila v prvem snemalnem valu, in vključen zajem videa, kjer koli bo mogoče. Pri

praksah zbiranja gradiv smo ugotavljali kot ključni problem nekontinuirano delo oz. snemanje v izredno kratkih časovnih obdobjih ter opozorili, da se tako veliko sredstev izgublja za koordiniranje množice sodelavcev in da ni zadosti časa za podrobno načrtovanje in pripravo orodij, s katerimi bi lahko bilo delo bolj učinkovito in bolj kvalitetno.

V naslovu smo nakazali, da vidimo snemanje gradiv za govorne korpuse v Sloveniji kot plutje med Scilo in Karibdo. Med obema grozečima skalama barka slovenskih govornih korpusov pluje večkrat: prvič, ko je ujeta v kratke roke in omejena finančna sredstva, za katere so pričakovanja po končnem obsegu gradiv izredno visoka; drugič, ko skuša ustreči včasih tudi precej nasprotujočim si željam različnih uporabnikov; tretjič, ko se sooča z nedostopnostjo in omejitvami že obstoječih posnetkov v različnih institucijah. Upajmo, da je barka svoje Scile in Karibde srečno preplula in da jo v prihodnje čakajo mirnejše vode v obliki dolgoročnega, stabilnega načrtovanja, kjer bo mogoče kontinuirano in premišljeno nadgrajevati govorne korpuse ter tako za ista ali celo manjša sredstva doseči več in boljše.

Zahvala

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

Literatura

Adolphs, S., Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Routledge.

Aijmer, K., Rühlemann, C. (ur.) (2015). *Corpus Pragmatics: A Handbook*. Cambridge University Press. https://doi.org/10.1017/CBO9781139057493

Alderson, C. J. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383–409. https://doi.org/10.1093/applin/amm024

Cappelle, B., Dugas, E., Tobin, V. (2015). An afterthought on let alone. *Journal of Pragmatics*, 80, 70–85. https://doi.org/10.1016/j.pragma.2015.02.005

Chen, N., You, C., Zou, Y. (2021). Self-Supervised Dialogue Learning for Spoken Conversational Question Answering. *Proceedings of the Interspeech 2021*, 231–235. https://doi.org/10.21437/Interspeech.2021-120

Data Protection Working Party. (2003). *Working document on biometrics*. Article 29 of Directive 95/46/EC. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjn_Km6kcCAAxUhS_EDHTOCAggQFnoECB0QAQ&url=https%3A%2F%2Fec.europa.eu%2Fjustice%2Farticle-29%2Fdocumentation%2Fopinion-recommendation%2Ffiles%2F2003%2Fwp80_en.pdf&usg=AOvVaw0NtFl7DWh5OLKSW3ZrVQik&opi=89978449

Flowerdew, J. (2009). Corpora in language teaching. V M. H. Long, C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 327–350). Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch19

Garabík, R. (2023). Corpus of Spoken Slovak. V M. Krajnc Ivič (ur.), *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov* (pp. 5–6). 6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki, Maribor, Slovenija. Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.5.2023

Garabík, R., Rusko, M. (2007). Corpus of Spoken Slovak Language. V J. Levická, R. Garabík (ur.), *Computer Treatment of Slavic and East European Languages,* Zbornik konference Slovko 2007 (pp. 222–236). Brno: Tribun.

Giagkou, M., Lynn, T., Dunne, J., Piperidis, S., Rehm, G. (2023). European Language Technology in 2022/2023. V G. Rehm, A. Way (ur.), *European language Equality: A Strategic Agenda for Digital Language Equality*. Springer. https://doi.org/10.1007/978-3-031-28819-7

Godfrey, J. J., Holliman, E. (1993). *Switchboard-1 Release 2 LDC97S62*. Web Download. Linguistic Data Consortium. https://doi.org/10.35111/sw3h-rw02

Goláňová, H., Waclawičová, M., Komrsková, Z., Lukeš, D., Kopřivová, M., Poukarová, P. (2017). DIALEKT: nářeční korpus, verze 1 z 2. 6. 2017. Praha: ÚČNK FF UK. http://www.korpus.cz

Gril, L., Sepesy Maučec, M., Donaj, G., Žgank, A. (2021). Avtomatsko razpoznavanje slovenskega govora za dnevnoinformativne oddaje. Slovenščina 2.0, 9(1), 60–89. https://revije.ff.uni-lj.si/slovenscina2/article/view/9899/9554

Hatice, C. (2015). *Impoliteness in Corpora: A Comparative Analysis of British English and Spoken Turkish*. Sheffield: Equinox.

Krajnc Ivič, M. (ur.). (2023, 18. in 19. maj). *Infrastruktura za raziskave govora v humanistiki in jezikovnih tehnologijah: zbornik povzetkov.* 6. mednarodna znanstvena konferenca Slavistični znanstveni premisleki, Maribor, Slovenija. Univerza v Mariboru, Univerzitetna založba. https://doi.org/10.18690/um.ff.5.2023

Lee, H., Yun, J., Choi, H., Joe, S., Gwon, Y.L. (2021). Enhancing Semantic Understanding with Self-Supervised Methods for Abstractive Dialogue Summarization. *Proceedings of the Interspeech 2021.* 796–800, doi: 10.21437/Interspeech.2021-1270

Ljubešić, N., Rupnik, P. (2022). The ParlaSpeech-HR benchmark for speaker profiling in Croatian. V D. Fišer, T. Erjavec (ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference,* 117–123. Inštitut za novejšo zgodovino. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf

Logar, N., Gorjanc, V., Arhar Holdt, Š. (2023). Korpus Gigafida 2.0: mnenje uporabnikov. *Jezik in slovstvo* 68(2), 75–91. https://doi.org/10.4312/jis.68.2.75-91

Love, R., Dembry, C., Hardie, A., Brezina, V., McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics,* 22(3), 319-344. https://doi.org/10.1075/ijcl.22.3.02lov

MacWhinney, B. (2019). Understanding spoken language through TalkBank. *Behavior Research Methods,* 51, 1919–1927. https://doi.org/10.3758/s13428-018-1174-9

Marjanovič Umek, L., Kranjc, S., Fekonja, U., Saksida, I. (ur.). (2006). *Otroški govor: razvoj in učenje*. Izolit.

McEnery, T. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. New York, NY: Routledge.

Mukiibi, J., Katumba, A., Nakatumba-Nabende, J., Hussein, A., Meyer, J. (2022). The Makerere Radio Speech Corpus: A Luganda Radio Corpus for Automatic Speech Recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 1945–1954). Marseille, France: European Language Resources Association.

Ogrodniczuk, M., Osenova, P., Erjavec, T., Fišer, D., Ljubešić, N., Çöltekin, Ç., Kopp, M., Meden, K. (2022). ParlaMint II: the show must go on. V

D. Fišer idr. (ur.), *Proceedings of the ParlaCLARIN III,* 1–6. http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/pdf/2022.parlaclariniii-1.1.pdf

Pančur, A. Erjavec, T., Ojsteršek, M., Šorn, M., Blaj Hribar, N. (2020). *Slovenian parliamentary corpus (1990-2018) siParl 2.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1300

Plüss, M., Hürlimann, M., Cuny, M., Stöckli, A., Kapotis, N., Hartmann, J., Ulasik, M. A., Scheller, C., Schraner, Y., Jain, A., Deriu, J., Cieliebak, M., Vogel, M. (2022). SDS-200: A Swiss German Speech to Standard German Text Corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, (pp. 3250–3256). Marseille, France: European Language Resources Association.

Rühlemann, C. (2006). Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, 11(4), 385–409. https://doi.org/10.1075/ijcl.11.4.03ruh

Rühlemann, C., Gries, S. (2015). Turn order and turn distribution in multiparty storytelling. *Journal of Pragmatics*, 87, 171–191. https://doi.org/10.1016/j.pragma.2015.08.003

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141. https://doi.org/10.1515/cllt.2011.006

Schmidt, T. (2014). The Research and Teaching Corpus of Spoken German — FOLK. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 383–387, Reykjavik, Iceland. European Language Resources Association (ELRA).

Schmidt, T. (2016). Construction and Dissemination of a Corpus of Spoken Interaction – Tools and Workflows in the FOLK project. *Journal for Language Technology and Computational Linguistics*, 31(1), 127–154.

Schmidt, T. (2023). FOLK – Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch. *Korpora Deutsch als Fremdsprache,* 3(1), 166–169. https://doi.org/10.48694/kordaf.3737

Smith, A. (2014). Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics*, 34(1), 118–138. https://doi.org/10.1080/07268602.2014.875458

Stabej, M., Vitez, P. (2000). KGB (korpus govorjenih besedil) v slovenščini. V T. Erjavec, J. Gros (ur.), *Informacijska družba IS'2000, Jezikovne tehnologije* (pp. 79–81). Inštitut Jožef Stefan.

Šumenjak, K. (2012). Zasnova dialektološkega korpusa na primeru govora Koprive na Krasu. V B. Krakar Vogel (ur.), *Slavistika v regijah – Koper* (pp. 73–78). Zbornik 23. Slovenskega slavističnega kongresa, Zveza društev Slavistično društvo Slovenije.

Verdonik, D. (2018). Korpus in baza Gos Videolectures. V D. Fišer, A. Pančur (ur.), *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, 265–268. Znanstvena založba Filozofske fakultete. http://www.sdjt.si/wp/wp-content/uploads/2018/09/JTDH-2018_Verdonik_Korpus-in-baza-Gos-Videolectures.pdf

Verdonik, D. (2021). Govorni viri za pravorečje. V T. Mirtič, M. Snoj (ur.), *1. slovenski pravorečni posvet* (pp. 120–132). Slovenska akademija znanosti in umetnosti. https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S., Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. Language Resources and Evaluation, 47(4), 1031–1048.

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., Pucer, M., Cvetko, M., Zelenik, M., Pavlič, J., Dobrišek, S., Križaj, J., Strle, G., Ivanovska, M., Grm, K., Bajec, M., Lebar Bajec, I., Jelovšek, T., Lokovšek, J., Longyka, J., Trojar, M., Žganec Gros, J., Mihelič, A., Vesnicer, B., Dretnik, N., Bordon, D. (2023a). ASR database ARTUR 1.0 (audio). Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1776

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., Pucer, M., Cvetko, M., Zelenik, M., Pavlič, J., Dobrišek, S., Križaj, J., Strle, G., Ivanovska, M., Grm, K., Bajec, M., Lebar Bajec, I., Jelovšek, T., Lokovšek, J., Longyka, J., Trojar, M., Žganec Gros, J., Mihelič, A., Vesnicer, B., Dretnik, N., Bordon, D. (2023b). *ASR database ARTUR 1.0 (transcriptions)*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1772

Verdonik, D., Zwitter Vitez, A., Zemljarič Miklavčič, J., Krek, S., Stabej, M., Erjavec, T., Verdonik, D., Potočnik, T., Sepesy Maučec, M., Majhenič, S., Žgank, A., Bizjak, A., Gril, L., Dobrišek, S., Križaj, J., Bajec, M., Lebar Bajec, I., Jelovšek, T., Trojar, M., Bernjak, M., Dretnik, N., Strle, G., Dobrovoljc, K., Ljubešić, N., Rupnik, P. (2023c). *Spoken corpus Gos 2.0 (transcriptions).* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1771

Vlaj, D., Žgank, A. (2023). Acoustic Gender and Age Classification as an Aid to Human–Computer Interaction in a Smart Home Environment. *Mathematics*, 11(1). https://doi.org/10.3390/math11010169

Žganec Gros, J., Vesnicer, B., Dobrišek, S. (2022). A method for selection of phonetically balanced sentences in read speech corpus design. *Proceedings of the 30th European Signal Processing Conference (EUSIPCO 2022)* (pp. 1136-1139). Belgrade, Serbia: EURASIP. https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0001136.pdf

Žgank, A., Verdonik, D., Zögling Markuš, A., Kačič, Z. (2005). BNSI Slovenian broadcast news database - speech and text corpus. *Interspeech Lisboa 2005: proceedings of the 9th European conference on speech communication and technology* (pp. 1537-1540). Bonn: Universität, Institut für Kommunikationsforschung und Phonetik.

Wang, S. (2005). Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *Journal of Pragmatics*, 34(4), 505–540. https://doi.org/10.1016/j.pragma.2004.08.002

Xiao, R., Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies*, 1(2), 231–273. https://doi.org/10.1558/sols.v1i2.241