IZBOR IN UREJANJE GRADIV ZA UČNI KORPUS GOVORJENE SLOVENŠČINE ROG

Darinka VERDONIK,[1] Nikola LJUBEŠIĆ,[2] Peter RUPNIK,[2] Kaja DOBROVOLJC,[3] Jaka ČIBEJ[3]

[1] Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko
[2] Institut Jožef Stefan
[3] Univerza v Ljubljani, Filozofska fakulteta

Učni korpusi vsebujejo premišljen nabor gradiv in zanesljive, praviloma ročno pripisane oznake na različnih jezikoslovnih ravneh. Služijo tako za učenje avtomatskih označevalnikov kot za temeljne jezikoslovne raziskave. Učni korpus za slovenščino SUK vsebuje samo pisna besedila, medtem ko je bilo stanje za govorjeni jezik zelo fragmentarno. Prispevek predstavlja izbor in urejanje gradiv za učni govorni korpus slovenščine ROG. Korpus obsega približno 75.000 besed oziroma preračunano okrog 8 do 9 ur govora. Gradiva za ROG so bila izbrana iz aktualne različice korpusa Gos 2.1 in ga delimo v tri podenote. V prispevku so podrobno opisane sestavne enote korpusa ROG skupaj s številčnimi podatki o gradivih. Korpus bo predvidoma do konca leta 2024 objavljen v repizitoriju CLARIN.SI pod licenco Creative Commons.

Ključne besede: govorni viri, govor, učni korpus

1 UVOD

Z razvojem tehnologije in umetne inteligence je poudarek na razumevanju in obdelavi govorjenega jezika vse večji. Eden ključnih korakov za razvoj tovrstnih tehnologij je učni korpus, tj. »premišljeno grajene besedilne množice z zanesljivimi (tipično ročno pripisanimi ali pregledanimi) dodatnimi informacijami, ki se uporabljajo pri nadzorovanem strojnem učenju postopkov za obdelavo naravnega jezika« (Arhar Holdt in sod., 2023, str. 121). Za slovenski jezik se učni korpus, ki vsebuje pisna besedila, razvija že več kot desetletje (Krek in sod., 2020). Najnovejšo izdajo pod imenom SUK je doživel v 2022 (Arhar Holdt in sod., 2022) in obsega 1 mio. pojavnic. Vključuje ročno pregledane jezikoslovne oznake na naslednjih ravneh: tokenizacija, stavčna segmentacija, lematizacija, oblikoskladnja MULTEXT-East, oblikoslovje ter

skladnja Universal Dependencies, skladnja JOS-SYN, udeleženske vloge, imenske entitete in koreference.

Korpus SUK vključuje besedila iz referenčnega pisnega korpusa Gigafida, iz slovenskih novičarskih portalov, iz Wikipedijinih člankov, skratka, iz pisnih virov. Številne raziskave so potrdile pomembne razlike med pisno in govorjeno rabo jezika tako na leksikalni kot slovnični in drugih jezikoslovnih ravneh (Akinnaso, 1982; Henrichsen in Allwood, 2005; Adolphs in Carter, 2003; Biber, 2012; Dobrovoljc in Nivre, 2016). Jezikovni viri, ki vključujejo samo pisno rabo jezika, zato ne morejo biti zadostni za obravnavo govorjenega jezika (Siepmann, 2015; Verdonik in Sepesy Maučec, 2017), ampak so za uspešno procesiranje kot tudi za celostno razumevanje jezikovne rabe potrebni tudi viri, ki vključujejo avtentične primere govorjene rabe. S tem namenom ne samo za slovenščino (Verdonik in sod., 2024), ampak za mnoge jezike nastajajo t. i. govorni korpusi, ki vključujejo posnetke in zapise govora (npr. Komrsková in sod., 2023; Kuvač in Hržica, 2016; Schmidt, 2016; Love in sod., 2017). Nezapolnjena vrzel pa ostajajo govorni korpusi, ki bi imeli ročno pripisane oz. popravljene jezikoslovne oznake. Za slovenščino je bila edini tovrstni vir doslej drevesnica govorjene slovenščine Spoken Slovenian Treebank (SST) v obsegu 30.000 pojavnic, ki ima ročno pripisane leme, oblikoskladenjske oznake MULTEXT-East ter oblikoslovne in odvisnostne skladenjske oznake po sistemu Universal Dependencies (Dobrovoljc in Nivre, 2016).[1] V primerjavi s pisnimi viri tako beležimo precejšnjo vrzel.

S ciljem zapolniti to vrzel v projektu Temeljne raziskave za razvoj govornih virov in tehnologij – MEZZANINE poteka aktivnost izdelave učnega korpusa govorjene slovenščine, poimenovanega ROG (Ročno označeni govorni korpus), ki bo imel ročno pripisane oznake na naslednjih ravneh: tokenizacija, lematizacija, oblikoskladnja po sistemu MULTEXT-East v6,[2] skladnja Universal Dependencies, prozodične enote, netekočnosti in dialoška dejanja. Ta prispevek opisuje prvi korak izdelave tega korpusa, to je izbor gradiv in pripravo gradiv za označevanje. Označevanje je v času pisanja članka še potekalo in v tem prispevku ni obravnavano.

[1] Korpus Spoken Slovenian Treebank je bil v času pisanja tega prispevka objavljen z razširjenim gradivom, ki je opisano v tem prispevku.

[2] Oblikoskladenjske oznake po sistemu MULTEXT-East v6: https://nl.ijs.si/ME/V6/msd/html/msd-sl.html.

2 IZBOR GRADIV

Za učni korpus je pomembno, da zajema raznovrstne vzorce in primere jezikovne rabe in s tem omogoči dovolj raznoliko množico učnih podatkov. Hkrati so ročno označeni učni podatki omejeni z razpoložljivimi sredstvi in časom ter posledično majhni, zato je potrebna premišljena sestava gradiv, ki jih vključimo v korpus. V govornih korpusih se vse od govorne komponente korpusa British National Corpus (Crowdy, 1993) praviloma upoštevata dve vrsti kriterijev za zajem gradiv: besedilnovrstni in demografski. Obe vrsti kriterijev smo upoštevali tudi ob sestavi učnega korpusa ROG, pri čemer pa smo se omejili z obstoječimi gradivi, saj je bil cilj osredotočiti se na označevanje.

Gradiva smo zajeli iz najnovejše izdaje referenčnega govornega korpusa Gos 2.1 (Verdonik in sod., 2024), ki že vključuje smiseln nabor ob času nastanka razpoložljivih govornih virov. Pri tem smo upoštevali že izdelani in skladenjsko ročno označeni nabor gradiv iz korpusa Gos 1.1 (Verdonik in sod., 2013). Zaradi potreb procesiranja in analiziranja na akustični ravni je bil cilj, da je vsaj polovica učnega korpusa javno dostopna tudi s kvalitetnimi avdio posnetki. Glede na navedene cilje korpus ROG vključuje tri podenote:

1. Izbor iz tistega dela Gos 2.1, ki izhaja iz korpusa Artur, kjer so posnetki kvalitetni in javno dostopni.

2. Izbor iz stare različice korpusa Gos 1.1 z manj kvalitetnimi in za dostop omejenimi posnetki, med katerimi pa mnogi vključujejo bolj interaktivne, bolj spontane in bolj raznolike nejavne govorne situacije kot korpus Artur. Ta del vključuje že obstoječo podenoto 30.000 pojavnic (Dobrovoljc in Nivre, 2016) in dodatno enoto na novo izbranih 10.000 pojavnic.

V nadaljevanju poglavja so vse tri podenote opisane po kriterijih za izbor gradiv in s številčnimi podatki.

2.1 Izbor iz korpusa Artur

Iz korpusa Artur je bil cilj izbrati gradiva v obsegu 40.000 pojavnic. Izbor gradiv iz korpusa Artur je potekal v dveh korakih: (1) ročni izbor posnetkov, (2) avtomatski izbor odsekov posnetkov. Kriteriji za ročni izbor posnetkov iz korpusa Artur:

1. V posnetku ni presluha in prisotnosti šuma.
2. Uravnotežen nabor različnih tipov glede na opis govornega dogodka v korpusu.
3. Uravnoteženost po spolu govorcev.
4. Uravnoteženost po regiji stalnega bivališča govorcev.

Avtomatski izbor segmentov v posnetkih je sledil kriterijem:

1. Izbrani odsek posnetka vključuje okvirno med 700 in 800 besed.
2. Pri posnetkih javnega govora se izbere odsek, v katerem govori predhodno ročno izbrani govorec.
3. Pri nejavnem govoru, kjer je isti pogovor posnet v dveh ločenih posnetkih, vsak za enega govorca, se izbereta oba posnetka pogovora in v obeh posnetkih isti odsek pogovora.
4. Začetek izbranega odseka posnetka je pri menjavi vloge.
5. Konec izbranega odseka je konec segmenta, ki si konča s piko.

Na podlagi tako definiranih kriterijev so bila izbrana gradiva tako v formatu TEI kot v formatu TRS,[3] prav tako so bili pripravljeni tudi avdio posnetki na način, da je bil celoten neizbrani del posnetka utišan. Posnetki so tako ostali enako dolgi kot izvorno, slišen pa je samo odsek, izbran za učni korpus, kar je pomembno za ohranitev obstoječe segmentacije na časovne enote glede na izvorni posnetek. Tabeli 1 in 2 predstavljata podatke o izbranih gradivih.

Tabela 1: Izbor iz korpusa Artur – posnetki.

Tip diskurza	*Govorni dogodek*	*Število posnetkov*	*Število besed*[4]
Javni		23	19.165
	Spletni dogodek	7	5792
	Okrogla miza	6	5039
	Intervju	6	5123
	Novinarska konferenca	3	2479
	Nagovor na dogodku	1	732

[3] Format TRS je izhodni format programa Transcriber, s katerim so bile narejene transkripcije za korpus Gos. Gre za XML-format, ki je enostaven za razčlenjevanje ter uvozljiv v pomembnejša orodja za označevanje in analizo govora (Praat, ELAN, EXMARaLDA).

[4] V izvornih datotekah korpusa Artur.

Nejavni		28	15.533
	Prosti dialog med dvema sogovornikoma	14	5771
	Prosti monološki govor	7	4799
	Razlaganje in opisovanje	7	4963
Parlamentarni	Seja državnega zbora	6	4241
SKUPAJ		57	38.939

Kot vidimo iz tabele 1, obsega izbor 57 približno enako dolgih posnetkov. Povprečen obseg enega pogovora je 683 besed. Razmerje med javnim, nejavnim in parlamentarnim je 50 % javni diskurz, 40 % nejavni diskurz in 10 % parlamentarni diskurz.

Tabela 2: Izbor iz korpusa Artur – govorci.

Značilnosti govorca	*Vrsta*	*Število govorcev*	*Število besed*
Spol	Moški	38	20.073
	Ženski	34	18.866
Starost	18 do 34	11	6785
	30 do 59	36	18.381
	nad 60	13	6561
	Nedoločeno	12	7212
Statistična regija*	Osrednjeslovenska	18	10.319
	Podravska	10	4566
	Savinjska	6	3149
	Pomurska	3	1584
	Goriška	5	2591
	Gorenjska	2	1459
	Jugovzhodna	2	870
	Koroška	2	676
	Primorsko-notranjska	3	995
	Posavska	4	1847
	Nedoločeno	17	10.883
SKUPAJ		72	38.939

* Oznaka pomeni statistično regijo stalnega bivališča.

Tabela 2 predstavlja podatke za celoten izbor. Odstotek nedoločenih podatkov je dokaj visok, ker za javni govor pogosto ni vseh podatkov o govorcih. Povprečno število besed na enega govorca je 540. Število oseb moškega in ženskega spola je približno enakomerno, zastopane so vse starostne skupine in skoraj vse statistične regije.

2.2 Izbor iz korpusa Gos 1.1

Izbor iz korpusa Gos 1.1 vsebuje že obstoječi korpus SST, razširjen z dodatnimi 10.000 pojavnicami prav tako iz korpusa Gos 1.1. Obe podenoti ostajata v ločenih datotekah in ju opisujemo v ločenih podpoglavjih.

2.2.1 ŽE OBSTOJEČI IZBOR V OBSEGU 30.000 BESED

V raziskavi (Dobrovoljc in Nivre, 2016) je predstavljen nabor gradiv iz korpusa Gos 1.1 v skupnem obsegu 30.000 pojavnic, tj. korpus Spoken Slovenian Treebank – SST. Gradiva so bila izbrana tako, da se je iz vsakega posnetka od 287 posnetkov v korpusu Gos 1.1 izbral proporcionalen del pojavnic. Vsak zajeti odsek posnetkov vključuje eno ali več zaporednih vlog govorcev. Ta nabor je bil uporabljen za dve ročni označevalni kampanji: označevanje večbesednih diskurznofunkcijskih stalnih besednih zvez (Dobrovoljc, 2018) in slovnično označevanje v okviru izdelave prve drevesnice govorjene slovenščine (Dobrovoljc in Nivre, 2016), v kateri so bili besedilom ročno pripisani podatki o lemah in oblikoskladenjskih oznakah po sistemu JOS/MULTEXT-East ter oblikoslovne in skladenjske oznake po medjezikovno usklajeni označevalni shemi Universal Dependencies. Drevesnica SST je bila prva govorna drevesnica, označena s shemo UD, v slovenskem prostoru pa je bila uporabljena za raziskave in razvoj prilagojenih slovničnih označevalnikov za govor (Dobrovoljc in Martinc, 2018; Verdonik in sod., 2024).

Tabela 3: SST izbor iz korpusa Gos 1.1 – posnetki.

Tip diskurza	*Kanal*	*Število posnetkov*	*Število besed*[5]
Javni informativni/ izobraževalni	Televizija	61	3068
	Radio	27	2310
	Osebni stik	33	3555

[5] V izvornih datotekah XML TEI korpusa Gos 1.1.

Javni razvedrilni	Televizija	14	2792
	Radio	27	3439
Nejavni nezasebni	Osebni stik	26	3007
	Telefon	17	1018
Zasebni	Osebni stik	49	5580
	Telefon	18	1562
SKUPAJ		272	26.331

Ker vključuje korpus SST del vsakega posnetka v izvornem korpusu Gos 1.1, je število datotek dokaj veliko, 272. Povprečen obseg enega (po)govora je 97 besed, tako da je korpus precej fragmentiran.

2.2.2 DODATEN IZBOR V OBSEGU 10.000 BESED

S ciljema, da se tudi iz korpusa Gos 1.1 zajamejo gradiva v obsegu 40.000 pojavnic in da se učni korpus razširi s posnetki pogovorov oz. medosebne interakcije, je bil korpus SST dopolnjem z dodatnimi 10.000 pojavnicami. Gradiva so bila izbrana tako, da vključujejo posnetke interaktivnih neformalnih govornih situacij, ki v korpusu Artur večinoma niso zajeta, v korpusu Gos 1.1 pa so. Izbor je potekal podobno kot pri korpusu Artur z ročnim izborom posnetkov in avtomatskim izborom odsekov iz posnetkov. Za izbor posnetkov so bili upoštevani kriteriji:

1. javni diskurz:
    a. samo posnetki neformalne interakcije
    b. enakomerna zastopanost posnetkov s televizije in radia
    c. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
2. nejavni nezasebni diskurz:
    a. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
    b. raznolikost tematik
    c. 85 % posnetkov v osebnem stiku in 15 % posnetkov po telefonu
3. nejavni zasebni diskurz:

a. vključijo se vrste govornih dogodkov, ki še niso bile zajete v gradivih iz korpusa Artur
b. raznolikost tematik
c. 85 % posnetkov v osebnem stiku in 15 % posnetkov po telefonu
d. zastopanost različnih regij

Hkrati smo v celotnem izboru sledili cilju, da bi bili ustrezno zastopani oba spola in različne starostne skupine.

Izbor odsekov iz posnetkov je bil izveden tako, da se je nadaljeval od tam naprej, kjer se je obstoječi izbor iz iste datoteke nehal. Začetna ocena je bila, da naj bi bili izbrani odseki v vseh datotekah dolgi od 350 do 360 besed, vendar smo za dosego cilja 10.000 pojavnic ta obseg nekoliko povečali.

Tabela 4: Dodaten izbor iz korpusa Gos 1.1 – posnetki.

Tip diskurza	*Kanal*	*Število posnetkov*	*Število besed*[6]
Javni razvedrilni	Televizija	2	817
	Radio	2	874
Nejavni nezasebni	Osebni stik	5	2347
	Telefon	1	425
Zasebni	Osebni stik	10	4301
	Telefon	2	855
SKUPAJ		22	9.619

Dodaten nabor obsega 22 posnetkov govora, kot vidimo iz tabele 4. Vsi posnetki so približno enako dolgi, povprečen obseg govora na enem posnetku pa je 437 besed.

Tabela 5: Dodaten izbor iz korpusa Gos 1.1 – govorci.

Značilnosti govorca	*Vrsta*	*Število govorcev*	*Število besed*
Spol	Moški	24	3624
	Ženski	37	5995
Starost	10 do 18	2	60

[6] V izvornih datotekah XMLTEI korpusa Gos 1.1.

	18 do 34	20	3988
	30 do 59	21	3019
	Nad 60	5	861
	Nedoločeno	13	1691
Statistična regija*	Osrednjeslovenska	14	2546
	Podravska	4	735
	Obalno-kraška	4	589
	Jugovzhodna Slovenija	4	792
	Goriška	3	297
	Pomurska	3	440
	Posavska	2	421
	Gorenjska	1	115
	Savinjska	1	199
	Mešano	15	2229
	Nedoločeno	10	1256
SKUPAJ		61	9619

* Oznaka pomeni vse statistične regije, v katerih je oseba bivala dalj časa.

Tabela 5 predstavlja podatke za celoten izbor. Odstotek nedoločenih podatkov je dokaj visok, ker za javni govor pogosto ni vseh podatkov o govorcih. Povprečno število besed na enega govorca je 157. Več oseb je ženskega spola. Zastopane so vse starostne skupine. Geografsko so pokrite vse večje slovenske regije. Velik delež govorcev ima označeno več kot eno regijo daljšega bivanja in so v tabeli uvrščeni pod kategorijo mešano.

Celoten učni korpus ROG skupaj obsega približno 75.000 besed. Preračunano v čas trajanja posnetkov znaša to skupaj med 8 in 9 ur govora.

3 UREJANJE GRADIV

3.1 Usklajevanje segmentacije

Ker korpus Gos 2.1 združuje posnetke iz različnih virov (Gos 1.1, Gos Videolectures in Artur), naletimo na nekaj razlik tako na ravni metapodatkov o posnetkih in govorcih (glej prispevek Verdonik in sod., 2022) kot na ravni zapisovanja in segmentiranja govora.

Za učni korpus je predstavljala poseben izziv razlika med načeli segmentacije

na enote govora v gradivu, ki je vključeno v GOS 1.1, in tistim gradivom, ki je bilo v korpus dodano v različici 2.1. Transkripcije posnetkov so bile namreč v različici 1.1 razdeljene na izjave in segmente, ki so bile obravnavane kot osnovna enota govora, ki približno ustreza pojmu povedi v pisnem jeziku in je določena tako, da je prozodično, semantično in skladenjsko zaokrožena (Verdonik in sod., 2013). Del, ki izhaja iz zbirke Artur (Verdonik in Bizjak, 2023), pa je bil predvsem zaradi potreb razvoja razpoznavalnika govora segmentiran z večjim upoštevanjem prozodičnih kriterijev, tj. s strožjim ločevanjem izjav na segmente glede na premore. Premor, ki je bil dolg vsaj 0,2 sekunde, je bil obravnavan kot mejnik med segmentoma, semantična in skladenjska zaključenost pa je bila šele drugotnega pomena. Razlika v segmentaciji je razvidna iz primerov v tabeli 6. Ker je predstavljala težavo pri označevanju v uporabljenih orodjih, ki so prilagojena za označevanje pisnih besedil, smo se odločili za dodatno usklajevanje segmentacije. Preučitev orodij, prilagojenih označevanju govora, za izvajanje vseh ravni označevanja, tudi skladnje, z vzporednim označevanjem osnovnih enot, ki so lahko različne glede na prozodijo, skladnjo, pragmatiko ali tehnične zahteve in podobno, ostaja eden od izzivov za naprej.

Tabela 6: Razlike v segmentaciji v gradivih iz Gos 1.1 in Artur.

Različica korpusa	*Primer segmentirane transkripcije*
GOS 1.1	<seg>kot vedno naši eem sodni mlini delujejo zelo zelo zelo počasi</seg>
Artur	<seg>Drage prijateljice, dragi prijatelji</seg> <seg>govorjene slovenščine.</seg> <seg>razmišljal sem, kako naj začnem ta</seg> <seg>svoj nastop.</seg>

Segmenti, razdeljeni na podlagi premorov, so pogosto kratki in pri označevanju skladenjskih povezav le-te potekajo preko mej segmentov, kar povzroča težave na nivoju tehnične implementacije pri označevalnih platformah in programski opremi, kot je Q-CAT (Brank, 2021). Pred označevanjem smo tako segmente strojno preporazdelili glede na ločila, ki so bila postavljena med transkripcijo posnetkov. Uporabili smo končna ločila pika (.), vprašaj (?) in

tropičje (...) in algoritem resegmentacije je vsakič, ko je naletel na eno od teh ločil, zaključil segment in začel novega. Pri tem so bile ohranjene vse informacije o prvotni delitvi na segmente (tj. identifikacijske kode prvotnih segmentov za vsako pojavnico), kar je omogočilo popolno sledljivost in skladnost z izvornimi podatki, obenem pa je poenotilo reprezentacijo segmentov med tistimi deli, ki so bili vzorčeni iz Gos 1.1, in tistimi iz korpusa Artur. Primer resegmentirane transkripcije prikazuje tabela 7.

Tabela 7: Primer strojno resegmentirane transkripcije iz korpusa Artur.

Izvorna segmentacija	*Resegmentirana transkripcija*
<seg>in tu se navezujem na</seg> <seg>misli eee direktorja</seg> <seg>ZRC Sazu,</seg> <seg>čez mejo</seg> <seg>in še čez eno mejo,</seg> <seg>v Prago.</seg>	<seg>in tu se navezujem na misli eee direktorja ZRC Sazu, čez mejo in še čez eno mejo, v Prago.</seg>

Strojna resegmentacija je število segmentov v vzorcu iz korpusa Artur zmanjšala s 5.587 na 1.968. V povprečju vsak nov segment vsebuje po tri stare segmente (najmanj enega in največ 70), polovica pa vsebuje manj kot dva (s standardnim odklonom 3,1 segmenta). Iz tabele 8 je razvidno, da gradivo iz korpusa Artur vsebuje nekoliko daljše segmente – v povprečju skoraj 25 pojavnic na segment, kar je več kot dvakrat več od gradiva iz Gos 1.1, kjer segment v povprečju zajema okrog 9 pojavnic. To je do določene mere pričakovana razlika, saj je bila v gradivu iz Gos 1.1 ob transkribiranju preferirana krajša dolžina segmenta, prav tako je v korpusu Artur manj interaktivnih in neformalnih govornih situacij, za katere so značilni krajši segmenti kot pri govornih dogodkih, kakršna so predavanja, okrogle mize ipd.

Tabela 8: Statistika resegmentacije vzorca iz korpusa Artur in primerjava z vzorcem iz Gos 1.1.

Statistika	*Vzorec GOS 1.1*	*Vzorec Artur*	*Vzorec Artur z resegmentacijo*
Število segmentov	4.912	5.587	1.968
Povprečno število pojavnic na segment	9,23	8,68	24,71

Mediana števila pojavnic na segment	6,00	7,00	18,00
Največje število pojavnic na segment	120	51	452

3.2 Sledljivost pojavnic

Izbrani segmenti za korpus ROG so bili izločeni iz korpusa Gos 2.1 (Verdonik in sod., 2023) v formatih TRS, TEI in WAV. Sledilo je sočasno označevanje na več nivojih: prva skupina raziskovalcev je izvajala ročno popravljanje oblikosladenjskih oznak in lem, druga skupina skladenjsko označevanje, tretja skupina označevanje netekočnosti in četrta označevanje prozodičnih enot. V uporabi so bili različni programi za označevanje: leme in oblikoskladenjske oznake so se pregledovale v razpredelnicah Google Sheets; odvisnostna skladnja se je označevala v orodju Q-Cat; za označevanje netekočnosti je bilo izbrano orodje EXMARaLDA,[7] ki je namenjeno podpori pri razvoju in označevanju govornih virov in omogoča označevanje na podlagi časovnih značk, ki ohranjajo povezavo s signalom; za označevanje prozodičnih enot je bilo uporabljeno orodje Praat, ki je prilagojeno analizam na akustični ravni.

Cilj je vse različne nivoje ročno pregledanih in dodanih oznak združiti v skupno XML-datoteko z več nivoji oznak. Da bi zagotovili združljivost gradiv, so bile v izbrana gradiva, uvožena v različna orodja in posledično orodjem prilagojene formate, dodane identifikacijske oznake pojavnic iz korpusa Gos2.1 (t. i. xml:id, na primer xml:id="Artur-P-G7036-P701111.tok603"). To omogoča naknadno preverjanje sprememb (denimo popravki ali naključne napake) in olajša prehajanje med različnimi formati.

4 ZAKLJUČEK

V prispevku smo predstavili izbor in urejanje gradiv za nov učni korpus govorjene slovenščine ROG. Korpus obsega skupaj okvirno 75.000 besed oz. po oceni približno 8 do 9 ur govora. Polovica korpusa izhaja iz korpusa Artur, kjer so prednost kvalitetni avdio posnetki, dostopni brez omejitev, polovica pa iz korpusa Gos 1.1, kjer je prednost večja avtentičnost in interaktivnost zlasti

[7] https://exmaralda.org/en/

nejavnega, delno pa tudi nekaterih posnetkov javnega govora. Celoten korpus bo ročno pregledan in označen na ravni lematizacije, oblikosladnje in skladnje, tisti del, ki izhaja iz korpusa Artur in ima na voljo prosto dostopne in kvalitetne avdio posnetke, pa tudi z ročnimi oznakami prozodičnih enot, netekočnosti in dialoških dejanj. Učni korpus ROG bo objavljen v repozitoriju CLARIN.SI, potem ko bodo dodani vsi nivoji oznak in združeni v eno datoteko, pod eno od licenc Creative Commons predvidoma do konca leta 2024.

Načrti za nadaljnje delo vključujejo razširitev korpusa s kvalitetnimi in odprto dostopnimi avdio posnetki pogovornega gradiva in razvijanje dodatnih nivojev oznak, ki omogočajo raziskave in razvoj na področju razumevanja govorne komunikacije.

ZAHVALA

Prispevek je nastal v okviru raziskovalnega projekta *Temeljne raziskave za razvoj govornih virov in tehnologij za slovenski jezik* (MEZZANINE, J7-4642), raziskovalnega projekta *Na drevesnici temelječ pristop k raziskavam govorjene slovenščine* (SPOT, Z6-4617) in raziskovalnega programa *Jezikovni viri in tehnologije za slovenski jezik* (P6-0411), ki jih financira Javna agencija za znanstvenoraziskovalno in inovacijsko dejavnost Republike Slovenije (ARIS).

ITERATURA

Arhar Holdt, Š. in sod. (2022). *Training corpus SUK 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1747.

Arhar Holdt, Š., Čibej, J., Dobrovoljc, K., Erjavec, T., Gantar, P., Krek, S., Munda, T., Robida, N., Terčon L. in Žitnik, S. (2023). Nadgradnja učnega korpusa ssj550k v SUK 1.0. V Š. Arhar Holdt, S. Krek (ur.), *Razvoj slovenščine v digitalnem okolju* (str. 119-156). Ljubljana: Založba Univerze. https://ebooks.uni-lj.si/ZalozbaUL/catalog/view/522/852/9441.

Adolphs, S. in Carter, R. (2003). And she's like it's terrible, like: Spoken discourse, grammar and corpus analysis. *International Journal of English Studies*, 3(1), 45–66.

Akinnaso, F. N. (1982). On the differences between spoken and written language.

Language and Speech, 25(2), 97–125.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, 8(1), 9–37.

Brank, J. (2021). Q-CAT Corpus Annotation Tool 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, 8(4), 259–265. doi: 10.1093/llc/8.4.259

Dobrovoljc, K. in Martinc, M. (2018). Er … well, it matters, right? On the role of data representations in spoken language dependency parsing. V *Proceedings of the workshop*. Second Workshop on Universal Dependencies (UDW 2018), November 1, 2018, Brussels. Strasbourg: Association for Computational Linguistics, 2018. Str. 37-46. https://aclanthology.info/papers/W18-6005/w18-6005.

Dobrovoljc, K. in Joakim, N. (2016). The Universal Dependencies Treebank of Spoken Slovenian. V *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16), Portorož, Slovenia.

Dobrovoljc, K. (2018). Raba tipično govorjenih diskurznih označevalcev na spletu. *Slavistična revija: časopis za jezikoslovje in literarne vede* 66(4), str. 497-513. https://srl.si/ojs/srl/article/view/2018-4-1-6.

Henrichsen, P. J. in Allwood, J. (2005). Swedish and Danish, spoken and written language: A statistical comparison. *International Journal of Corpus Linguistics* 10(3), 367–399.

Komrsková, Z., Kopřivová, M., Lukeš, D., Poukarová, P. in Goláňová, H. (2017). New Spoken Corpora of Czech: ORTOFON and DIALEKT. *Journal of Linguistics/Jazykovedný casopis* 68(2), 219–228. https://doi.org/10.1515/jazcas-2017-0031.

Krek, S., Erjavec, T., Dobrovoljc, K., Gantar, P., Arhar Holdt, Š., Čibej, J. in Brank, J. (2020). The ssj500k training corpus for Slovene language processing. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 23–33). Ljubljana, Slovenija. Inštitut za novejšo zgodovino.

Kuvač Kraljević, J. in Hržica, G. (2016). Croatian Adult Spoken Language Corpus (HrAL). *FLUMINENSIA* 28(2), 87–102.

Love, R., Dembry, C., Hardie, A., Brezina, V. in 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

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.

Siepmann, D. (2015). Dictionaries and spoken language: A corpus-based review of French dictionaries. *International Journal of Lexicography*, 28(2), 139–168.

Verdonik, D. in Bizjak A. (2023). *Pogovorni zapis in označevanje govora v govorni bazi Artur projekta RSDO*. Maribor: Univerzitetna založba.

Verdonik, D. in Sepesy Maučec, M. (2017). A speech corpus as a source of lexical information. *International Journal of Lexicography* 30(2), 143–166. DOI: 10.1093/ijl/ecw004.

Verdonik, D., Kosem, I., Zwitter Vitez, A., Krek, S. in 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., Dobrišek, S. Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. V D. Fišer in T. Erjavec (Ur.), *Jezikovne tehnologije in digitalna humanistika: zbornik konference* (str. 206-2012), Ljubljana, Slovenija. Inštitut za novejšo zgodovino.

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

Verdonik, D., Dobrovoljc, K., Erjavec, T. in Ljubešić, N. (2024). Gos 2: A New Reference Corpus of Spoken Slovenian. V *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (str. 7825–7830). Torino, Italia. ELRA and ICCL.

SELECTION AND PREPARATION OF DATA FOR THE ROG 1.0 TRAINING CORPUS OF SPOKEN SLOVENIAN

The article presents the selection and preparation of data for ROG, the training corpus of spoken Slovenian. The corpus comprises approximately 75,000 words, equivalent to around 8 to 9 hours of speech. Materials for the corpus were selected from the current version of the Gos 2.1 corpus and are divided into three subunits. The first subunit consists of data derived from the Artur corpus and encompasses 40,000 tokens. The advantage of these data is the availability of high-quality and freely accessible audio recordings. The other two subunits are selected from the Gos 1.1 corpus: the existing SST spoken language learning corpus, comprising 30,000 tokens, and an additional selection to fill the gap in interactive, conversational data, comprising 10,000 tokens. The data were exported in TEI, TRS, and WAV formats. In the WAV format, unselected parts of the recordings were muted to preserve the alignment with original segment time codes, while in the TEI and TRS formats, unselected parts were removed. The learning corpus will contain manually corrected and added annotations for lemmas, morphosyntax, syntax, prosodic units, disfluencies, and dialog acts, and is expected to be published by the end of 2024 in the CLARIN.SI repository under a Creative Commons license.

Keywords: spoken resources, speech, training corpus