REPUBLIKA SLOVENIJA
MINISTRSTVO ZA KULTURO

EVROPSKA UNIJA
EVROPSKI SKLAD ZA REGIONALNI RAZVOJ
NALOŽBA V VAŠO PRIHODNOST

Gos 2.0

Poročilo projekta Razvoj slovenščine v digitalnem okolju

Aktivnost DS1.5

Avtorji: Nejc Robida[○,□], Kaja Dobrovoljc[○,□], Luka Terčon[○], Darinka Verdonik[✧]

[○] Fakulteta za računalništvo in informatiko Univerze v Ljubljani
[□] Filozofska fakulteta Univerze v Ljubljani
[✧] Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru

Ljubljana: Center za jezikovne vire in tehnologije, Univerza v Ljubljani, 2023

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Vsebina

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

1 Nejc Robida: Projektni okvir

Poročilo oz. kazalnik *Gos 2.0* je nastalo pod okriljem projekta Razvoj slovenščine v digitalnem okolju, ki sta ga med leti 2020 in 2023 sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj (Operacija se je izvajala v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014–2020). Cilj projekta je bil zadovoljiti potrebe po računalniških izdelkih in storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, podjetja in širšo javnost. Spletna stran projekta z dostopom do rezultatov: https://slovenscina.eu/.

Kazalnik se umešča v prvi projektni delovni sklop z naslovom *Jezikovni viri*. Namen delovnega sklopa je bil nadgraditi slovenske besedilne korpuse in leksikon besednih oblik. Prenovili smo učne množice in postopke za strojno označevanje sodobne slovenščine. Rezultat so osveženi in povečani jezikovni viri, ki so na voljo tako uporabniški skupnosti kot za strojno rabo. Z razvitimi postopki in orodji bo posodabljanje slovenskih korpusov v prihodnosti hitrejše in preprostejše.

Med besedilne korpuse, ki smo se jim na projektu posebej posvetili, sodi tudi referenčni korpus govorjene slovenščine Gos 2.0. Korpus ponuja pomemben uvid v govorjeno slovenščino, kakršna se pojavlja v različnih govornih situacijah.

2 Nejc Robida: Opis projektnih aktivnosti

Cilj aktivnosti 1.5 Referenčni korpus govorjene slovenščine Gos je bil pripraviti načrt za kontinuirano nadgrajevanje korpusa govorjene slovenščine Gos, korpus povečati z novimi gradivi ter dati Gos 2.0 na voljo v specializiranem korpusnem konkordančniku.

Aktivnost se je posvečala referenčnemu korpusu govorjene slovenščine Gos, ki zajema posnetke govora v različnih slovenskih regijah, situacijah in govornih položajih. Korpus Gos 1.0, ki je na voljo v prosto dostopnem konkordančniku za širšo rabo, vsebuje transkripcije okrog 120 ur posnetkov. Kot podatkovna baza na repozitoriju CLARIN.SI je na voljo tudi dopolnilni del, ki zajema transkripcije 22 ur javnih strokovnih predavanj s portala VideoLectures. V sklopu projekta smo podatke obeh baz združili

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

in korpus Gos 1.0 še dodatno povečali z gradivi, ki so bila pripravljena v projektni aktivnosti 2.1 Izdelava govorne baze.

Korpus Gos je uravnotežena in reprezentativna besedilna zbirka, zato so bili vanj vključeni samo tisti posnetki govorne baze, ki so najbolje ustrezali jezikoslovnim kriterijem za povečavo korpusa. V povezavi z delom na aktivnosti 2.1 so v specifikacijah natančneje opredeljene tudi želene vsebine za nadaljnji razvoj korpusa ter povezana pravna vprašanja in metodologija transkribiranja ter označevanja. Na tak način bo poskrbljeno, da bo tudi po koncu projekta korpus mogoče nadalje razvijati na metodološko skladen način. V aktivnosti smo sodelovali z besedilodajalci, ki jih zastopa STA (Slovenska tiskovna agencija). Tako smo lahko v specifikacije vključili tudi identificirane težave in primere dobrih praks glede kontinuiranega zbiranja gradiva. Na drugi strani je bilo treba poskrbeti za vključitev novonastalega korpusa Gos 2.0 v specializirani konkordančnik (kot del obstoječe infrastrukture viri.cjvt.si), ki prikazuje transkripcijo na dveh nivojih ter vsebuje posnetke govora. V nekoliko prilagojeni obliki se predvideva tudi vključitev korpusa Gos 2.0 v konkordančnike CLARIN.SI.
Na projektni aktivnosti je sodelovalo več raziskovalcev in zunanjih sodelavcev, kot so študenti. Soavtorstvo je navedeno pri vsakem posameznem projektnem rezultatu.

Delo je potekalo prek celotnega časa trajanja projekta po naslednjih korakih:

- V sodelovanju z aktivnostjo 2.1 Izdelava govorne baze smo spremljali izkušnjo zbiralcev in dajalcev besedil in evalvirali preizkušene delotoke ter metodologijo. Identificirali in naslovili smo tudi pravna vprašanja, ki so se pojavila v zvezi z zbiranjem oz. snemanjem gradiva. Spoznanja in rešitve smo vključili v pričujoče specifikacije za nadgradnjo korpusa Gos, kjer so poleg želene vsebine za nadaljnje širitve opredeljena pravna vprašanja, metodologija in delotoki transkribiranja ter označevanja.
- Pripravili smo korpus Gos 2.0, v katerega smo združili obstoječe korpusno gradivo, gradivo Gos VideoLectures 4.0 in novopripravljene posnetke, ki jih je izbrala in posredovala aktivnost 2.1 Izdelava govorne baze. Posnetki, ki so iz govorne baze vključeni tudi v Gos, so bili izbrani glede na jezikoslovne kriterije tako, da je ohranjena reprezentativnost in uravnoteženost korpusa kot celote. V projektu je bil zagotovljen novi korpus Gos 2.0 v obsegu vsaj 200 ur posnetkov.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- Novopripravljeni korpus Gos 2.0 smo vključili v obstoječe konkordančnike, s čimer je gradivo dostopno za raziskave celotni jezikovni skupnosti. Korpus je vključen na portal cjvt.si, v nekoliko prilagojeni obliki pa tudi v konkordančnike CLARIN.SI.

REZULTATI AKTIVNOSTI

- Korpus Gos 2.0, povečan z gradivi, zbranimi v aktivnosti DS2, kot baza na repozitoriju CLARIN.SI: http://hdl.handle.net/11356/1771, na portalu viri.cjvt.si: https://viri.cjvt.si/gos/ in vključen v konkordančnika na CLARIN.SI:
    - NoSketch Engine Crystal: https://www.clarin.si/ske/#dashboard?corpname=gos20
    - KonText: https://www.clarin.si/kontext/query?corpname=gos20
- (Pričujoče) specifikacije za nadgradnjo korpusa Gos z opredelitvijo želene vsebine, pravnih vprašanj, metodologije in delotokov transkribiranja ter označevanja.

3 Nejc Robida: Izhodišče: Gos 1.0 in Gos VL

Korpus Gos 1.0 je referenčni govorni korpus, ki naj bi sprva predvidoma predstavljal govorni podkorpus referenčnega korpusa slovenskega jezika v okviru projekta Sporazumevanje v slovenskem jeziku[1]. Ta je bil sofinanciran s strani Evropskega socialnega sklada ter Republike Slovenije, Ministrstva za šolstvo in šport. Projekt pa je izvajal konzorcij v sestavi Institut Jožef Stefan, Univerza v Ljubljani, Znanstvenoraziskovalni center SAZU, Trojina, zavod za uporabno slovenistiko, in Amebis, d. o. o, trajal je dobrih pet let. Izdelava samega korpusa je potekala od septembra 2008 do decembra 2010. Leta 2011 pa je bil Gos 1.0 končan kot prvi sklop, ki se je potem nadaljeval leta 2016 s projektom izdelave dodatnega korpusa avdiobaze h korpusu Gos, poimenovali so ga Gos Videolectues[2] (skrajšano Gos VL). Zadnjega je izvajala fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, zaključili pa so ga leta 2018.

Korpus Gos 1.0 je namenjen raziskavam govorjene podobe slovenščine v različnih govornih položajih, natančneje je bil prvi sklop korpusa predviden predvsem za naslednje uporabnike:

[1] http://ssj.slovenscina.eu/
[2] http://hdl.handle.net/11356/1222

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- »znotraj projekta Sporazumevanje v slovenskem jeziku kot vir avtentičnih govorjenih diskurzov za leksikalno bazo, pedagoško korpusno slovnico in slogovni priročnik, za poučevanje slovenščine kot maternega ali tujega jezika,
- za raziskave govorjenega jezika in diskurza (jezikoslovne in dialektološke, sociolingvistične, pragmatične, jezikovnotehnološke ...),
- za poklicne govorce in pisce (govorci na radiju in televiziji, igralci, scenaristi, pisatelji, lektorji, prevajalci, tolmači ...),
- za slehernega maternega ali nematernega govorca slovenščine, ki mu brskanje po prijaznem spletnem vmesniku prinaša nova spoznanja o različnih regionalnih, starostnih, žanrskih in drugih značilnostih današnje govorjene slovenščine« (Verdonik in Zwitter Vitez 2020: 16).

Bil je tudi eden izmed virov za izdelavo leksikalne podatkovne baze s podatki o frekvenci, pomenski strukturi in z zgledi rabe in še za izdelavo pedagoške korpusne slovnice. V njem pa so zajeti:

- vzorčni primeri različnih govornih situacij in različnih tipov govora,
- govor demografsko reprezentativnega vzorca govorcev slovenskega jezika,
- predvsem tiste govorne situacije, v katerih so uporabniki jezika najbolj pogosto produktivno-receptivno udeleženi (Verdonik in Zwitter Vitez 2020).

Z Gos VL so ga razširili še z izbranimi posnetki javnih predavanj s portala Videolectures.net zaradi aktualnosti področja za jezikoslovne raziskave in avtomatsko razpoznavanje govora. Uporabniki lahko s tem raziskujejo tudi akademski jezik, javni diskurz in šolski jezik, ki močno vpliva na vsakdanji jezik. Korpus pa je tako postal primeren tudi za raziskave terminologije različnih strok.

3.1 Gradiva v korpusih Gos in Gos VL

3.1.1 Gos 1.0

Govorni korpus Gos 1.0 je v splošnem namenjen korpusnim raziskavam govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah. Naravnan je v kar se da referenčni zajem govorjenih diskurzov in obsega milijon besed (Verdonik in Zwitter 2020: 17).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Pri zajemanju gradiva za korpus Gos 1.0 so avtorji poskušali obdržati avtentično dolžino govorjenih diskurzov, predvsem pri zasebnih pogovorih in predavanjih pa so morali dolžine posnetkov zaradi potrebe po razpršenosti gradiva skrajšati na približno pol ure. Pri gradivu so pazili, da so dobili avtentične posnetke, to je bilo nemogoče v primerih, ko so morali paziti na varovanje pravic govorcev. Upoštevali so tudi pravno-etični vidik, nekaj govorcev je snemanje tudi zavrnilo. Pri snemanju so se izogibali hrupnega okolja in poskusili pridobiti predvsem spontani diskurz. Manjša spontanost pogovora se je pojavila le pri novinarskih prispevkih in pri nekaterih predavanjih.

V korpusu je tudi govor mladostnikov in otrok, ta se je pojavljal predvsem v šolskem okolju in pri družinskih posnetkih. Govor tujih govorcev (večinoma južnoslovanski) slovenščine obsega dva odstotka korpusa (približno 130 minut), en odstotek korpusa pa je govora zamejskih Slovencev.

Pomembna kriterija pri gradnji korpusa sta bila predvsem javnost diskurza in prenosnik. Natančna razporeditev je prikazana na Sliki 1.

tip diskurza*	št. besed	%	kanal	št. besed	%
javni informativno-izobraževalni	359.549	35%	televizija	102.263	10%
			radio	94.536	9%
			osebni stik	162.750	16%
javni razvedrilni	228.765	22%	televizija	105.613	10%
			radio	123.152	12%
nejavni nezasebni	153.471	15%	osebni stik	119.987	12%
			telefon	33.484	3%
nejavni zasebni	290.990	28%	osebni stik	222.907	22%
			telefon	68.083	7%
skupaj	1.032.775	100%		1.032.775	100%

Slika 1: Razporeditev gradiva glede na osnovne besedilnovrstne kriterije (Vir: Verdonik in Zwitter 2020: 48)

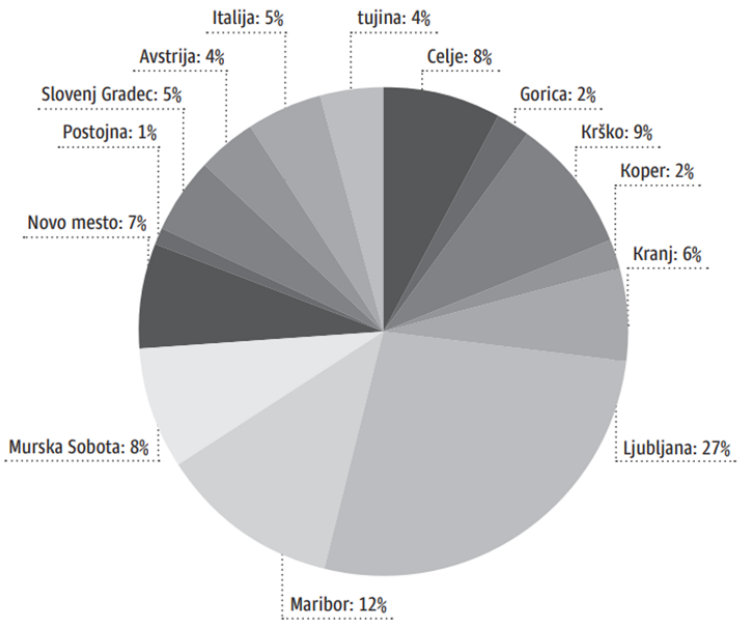
Če podatke podrobneje razčlenimo: Javni diskurz obsega 56 % korpusa Gos 1.0, nejavni pa 44 %. Za analizo govorjenega jezika je pomemben predvsem drugi. Še največ težav zaradi dovoljenj in avtorskih pravic so imeli pri snemanju šolskega diskurza. Skoraj polovico gradiva so zajeli z osebnim stikom (predvsem nejavni in šolski diskurz). 20 odstotkov korpusa je posnetkov s televizije, zajeti so bili

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

programi, ki oddajajo po celi Sloveniji. Tretji prenosnik je radio, ta obsega 21 odstotkov korpusa, poudarek je bil tudi na lokalnih radijih. Pri izbiranju gradiva z radia in televizije so upoštevali še podatke o poslušanosti oziroma gledanosti televizije. Zadnji prenosnik, prek katerega so zajemali govorne diskurze, je telefon.

Večino šolskega diskurza obsegajo fakultetna predavanja (19 %), osnovnošolske učne ure (37 %) in srednješolske učne ure (34 %), 8 odstotkov je javnih predavanj in 2 odstotka tečajev. Korpus Gos je uravnotežen tudi po spolu, in sicer so 51 % govora v korpusu opravile govorke, 49 % pa govorci. Povsem uravnotežena shema pri starosti govorcev je bila težje dosegljiva, saj so bili snemalci večinoma študentje, takšne starosti pa je tudi njihova socialna mreža. Kar 87 odstotkov govorcev je bilo mlajših od 24 let, 11 odstotkov je govorcev do 34 let, starejših od 34 let pa je za 2 odstotka. Govorci so bili večinoma osnovnošolsko izobraženi, po deset odstotkov je bilo tistih s srednjo in višjo šolo, 20 odstotkov govorcev ima višjo ali visoko izobrazbo.

Regijsko pripadnost govorcev navajamo na Sliki 2.



Slika 2: Regijska pripadnost govorcev (Vir: Verdonik in Zwitter 2020: 53)

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

3.1.2 Gos VL

Gos Videolectures skuša ustrezno upoštevati tako potrebe jezikoslovja kot potrebe govornih tehnologij po jezikovnih virih. Zajema 55 predavanj, to je več kot 22 ur (178.960 besed transkripcij) izbranih posnetkov javnih predavanj s portala Videolectures.net. Posnetki so izbrani tako, da zastopajo različna strokovna področja in različne skupine govorce (predavateljev) glede na spol, starost in regijo. Korpus so sproti nadgrajevali, delo pa je bilo financirano prek različnih organizacij (npr. Ministrstvo za kulturo Republike Slovenije, Clarin …).

V korpusu Gos VL so izbrana predavanja razdelili na pet strokovnih področij, in sicer na: humanistiko, družboslovje, medicino, naravoslovje oziroma matematiko in tehniko. Da bi čim bolj uravnotežili gradivo, so se (kot že pri Gos 1.0) odločili, da omejijo trajanja predavanj, ta potem niso bila daljša od petinštirideset minut, skoraj polovico posnetkov pa je bilo krajših od 20 minut. Na tabeli 1 navajamo natančnejšo razporeditev gradiva, glede na področje, poleg pa je tudi nekaj kvantitavnih podatkov o posameznih področjih (dolžina vseh posnetkov, število besed in število posnetkov).

Tabela 1: Vsebina korpusa Gos VL

Področje	**Število posnetkov**	**Trajanje posnetkov**	**Število besed**
Humanistika	9	4:31:38	34100
Družboslovje	14	6:16:40	52667
Medicina	8	2:13:59	17721
Naravoslovje/matematika	13	4:33:50	37747
Tehnika	11	4:25:24	36725
Skupaj	55	22:01:31	178960

Tako kot že korpus Gos 1.0 so poskušali čim bolj demografsko uravnotežiti tudi Gos VL. A so imeli tokrat drugačne cilje in želje, ter povsem drugačne specifike gradiva. Tokrat reprezentativnost tujih govorcev in zamejskih Slovencev ni igrala posebne vloge, prav tako ni bil smiseln kriterij izobrazbe, govorci so imeli večinoma visoko izobrazbo, zato so se osredotočili na spol, starost in regijsko pripadnost. Spol pa je edini zanesljivi demografski podatek, saj so se lahko o drugih odločali le na podlagi videoposnetkov (izvor posnetka, regionalni govor …). Moški so tako prispevali 57 odstotkov dolžine posnetkov predavanj, ženske pa 43 (Verdonik 2018: 266).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4 Nejc Robida: Pravna vprašanja

Korpusi Gos 1.0, Gos VL 4.2 in Artur 1.0 obsegajo posnetke in njihove transkripcije iz medijev, posnetke šolskega pouka (osnovna šola, srednja šola in fakulteta), posnetke iz parlamenta in tudi posnetke nejavnega govora. Zato so pri projektih morali slediti zahtevam po varovanju gradiva, osebnih podatkov in spoštovanju avtorskih pravic.

»S pravnega in etičnega vidika je zajemanje gradiva za govorni korpus omejeno predvsem zato, ker je treba zagotoviti predhodno seznanitev in soglasje vseh govorcev, katerih govor bo posnet za namene govornega korpusa. Izkušnje pri tem kažejo, da vsi govorci niso vedno pripravljeni za takšno sodelovanje, še pomembneje pa je, da se govorci praviloma vsaj v začetku vedejo drugače kot sicer, če vedo, da se njihov govor snema« (Sporazumevanje v slovenskem jeziku – govorni korpus[3]).

»V Republiki Sloveniji je področje zbiranja, hranjenja in obdelave osebnih podatkov zakonsko urejeno z *Zakonom o varstvu osebnih podatkov* (uradno prečiščeno besedilo *ZVOP-1-UPB1*)[4] ter z *Zakonom o varstvu dokumentarnega in arhivskega gradiva ter arhivih* (*ZVDAGA*[5]). *Zakon o varstvu osebnih podatkov* določa pravice, obveznosti, načela in ukrepe, s katerimi se preprečujejo neustavni, nezakoniti in neupravičeni posegi v zasebnost in dostojanstvo posameznika oziroma posameznice pri obdelavi podatkov. Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih ureja način, organizacijo, infrastrukturo in izvedbo zajema ter hrambe dokumentarnega gradiva v fizični in elektronski obliki, veljavnost oziroma dokazno vrednost takega gradiva in pogoje za njegovo uporabo, naloge arhivov in javne arhivske službe ter s tem povezane storitve in nadzor nad izvajanjem. Področje avtorskih pravic je v Republiki Sloveniji urejeno z *Zakonom o avtorskih in sorodnih pravicah* (*ZASP*)[6] in podzakonskimi akti« (Verdonik in Zwitter 2020: 32).

[3] http://www.korpus-gos.net/Content/Static/Splosne_specifikacije_zbiranja_gradiva.pdf
[4] 15. 12. 2022 je bil sprejet ZVOP-2: http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO7959.
[5] http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO4284
[6] 29. 9. 2022 je bil sprejet ZASP-I: http://www.pisrs.si/Pis.web/pregledPredpisa?id=ZAKO8300.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4.1 Korpus Gos 1.0

V okviru konzorcija projekta Sporazumevanje v slovenskem jeziku so bile avtorske pravice za gradiva, ki so jih za Gos prejeli od medijev, urejene s pogodbo o odstopu avdio in video posnetkov radijskih in televizijskih oddaj ali drugih programskih vsebin. Prenos pravic je urejal 4. člen, ki se glasi:

»Lastnik gradiv s to pogodbo neizključno, neodplačno in brez časovnih omejitev na nosilca projekta prenaša pravico reprodukcije, distribucije, dajanja v najem, priobčitve javnosti in predelave gradiv in njegovih predelav, na način kot to določa licenca Creative Commons: “priznanje avtorstva” + “nekomercialno” + “deljenje pod istimi pogoji”. Ta licenca dovoli uporabnikom avtorsko delo in njegove predelave reproducirati, distribuirati, dajati v najem, priobčiti javnosti in predelovati samo pod pogojem, da navedejo avtorja, da ne gre za komercialno uporabo in da tudi oni naprej širijo izvirna dela/predelave pod istimi pogoji. Uporaba te licence za podatkovno zbirko referenčni besedilni korpus z govornim podkorpusom je določena v 19. členu Pogodbe o sofinanciranju izvedbe projekta št. 3311-08-986003 v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013 “Sporazumevanje v slovenskem jeziku”, sklenjene med Ministrstvom za šolstvo in šport Republike Slovenije in podjetjem Amebis, d.o.o., Kamnik« (Verdonik in Zwitter 2020: 33).

»Avtorske pravice za terenske posnetke so bile v okviru konzorcija projekta Sporazumevanje v slovenskem jeziku urejene z izjavo o odstopu pravic, v kateri govorec izjavlja:

S to pogodbo brezplačno in brez časovnih omejitev prenašam pravico reprodukcije, distribucije, dajanja v najem, priobčitve javnosti in predelave na posnetkih, na katerih sodelujem s svojim govorjenjem, na transkripcijah teh posnetkov in njegovih predelavah, na v tej izjavi naveden projektni konzorcij in dajem dovoljenje, da se »posnetki, na katerih sodelujem s svojim govorjenjem, in transkripcije teh posnetkov (v nadaljevanju: »gradivo«) uporabijo za izgradnjo govornega korpusa.

Konzorcij projekta Sporazumevanje v slovenskem jeziku pa se zavezuje:

Projektni konzorcij in tretje osebe, ki bi izkazale interes, bodo govorni korpus uporabljaje v skladu z licenco Creative Commons: “priznanje avtorstva” + “nekomercialno” + “deljenje pod istimi pogoji”. Ta licenca dovoli uporabnikom avtorsko delo in njegove predelave reproducirati, distribuirati, dajati v

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

najem, priobčiti javnosti in predelovati samo pod pogojem, da navedejo avtorja, da ne gre za komercialno uporabo in da tudi oni naprej širijo izvirna dela/predelave pod istimi pogoji. Uporaba te licence za podatkovno zbirko referenčni besedilni korpus z govornim podkorpusom je določena v 19. členu Pogodbe o sofinanciranju izvedbe projekta št. 3311-08-986003 v okviru Operativnega programa razvoja človeških virov za obdobje 2007-2013 "Sporazumevanje v slovenskem jeziku", sklenjene med Ministrstvom za šolstvo in šport Republike Slovenije in podjetjem Amebis, d. o. o., Kamnik« (Verdonik in Zwitter 2020: 33).

V nekaterih zvočnih posnetkih in tudi zbranih obrazcih s podatki o posnetkih in govorcih se pojavljajo tudi osebni podatki. Po *Zakonu o varstvu osebnih podatkov* (*ZVOP-1*[7], UL RS 86/4) je osebni podatek katerikoli podatek, ki se nanaša na posameznika, ne glede na obliko, v kateri je izražen. Za gradivo Gos so ob zajemanju na posebnem obrazcu zapisali tudi podatke o spolu, starosti, izobrazbi, regionalni pripadnosti in prvem jeziku govorcev. Vsi podatki so v korpus vključeni tako, da so anominizirani in jih ni mogoče povezati s posameznikom. Konzorcij Sporazumevanje v slovenskem jeziku je zagotovil varovanje osebnih podatkov in gradiva tako pri transkripcijah kot pri posnetkih. Govorci so v transkripcijah in metapodatkih poimenovani s šiframi. Če se posamezniki pojavijo v posnetkih pa so v zapisu govora anonimizirani, označena je samo vrsta podatka, npr.: [ime], [priimek], [podjetje] … Zakrita so tudi imena manjših podjetij, prek katerih bi lahko posredno ugotovili identiteto posameznika. Anonimizirani pa so tudi podatki oseb v primeru žaljivega konteksta v govoru.

Na posnetkih so osebni podatki anonimizirani s pomočjo piska, spremenjena pa je tudi frekvenca posnetkov, da v nemedijskih in zasebnih diskurzih ni mogoče prepoznati osebe na posnetku.

Vsi, ki so obdelovali gradovi, so podpisali tudi izjavo »o varstvu osebnih podatkov skladno z *ZVOP-1* in izjavo o uničenju gradiv, s katerimi so delali, po opravljenem delu, upravljavec teh podatkov pa ravna v skladu z določili o zavarovanju osebnih podatkov po *ZVOP-1*« (Verdonik in Zwitter 2020: 34).

[7] http://pisrs.si/Pis.web/pregledPredpisa?id=ZAKO3906

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

4.2 Korpus Gos VL

Manj težav z anonimizacijo je bilo pri korpusu Gos VL, saj so bila predavanja javna. So pa v metapodatkih, ki so jih zbrali ob posnetkih predavateljev, podatki o samem dogodku, kot so: ime dogodka/datoteke, dolžina posnetka, tip diskurza[8], vrsta situacije[9], opis diskurza[10], regija (snemanja), vir (snemanja), kraj (dogodka), čas, število aktivnih udeležencev, opis govornega dogodka in zvočna kakovost posnetka. Pri informacijah o govorcih je zabeležena ID-KODA govorca v obliki: Ym-vlxxx; Yf-vlxxx, kjer m ozačuje moški spol govorca in f ženski spol govorke. Podatek o starosti je bil ocenjen na videz (do 35 let in nad 35 let), regionalna pripadnost pa se je ocenila na podlagi biografskih podatkov s spleta in narečnih vplivov pri govoru govorca.

»Korpus in avdio posnetki Gos Videlectures so na voljo za raziskave prek konkordančnika NoSketchEngine pri IJS. Prav tako so dostopne izvorne datoteke, ki jih lahko uporabniki snamejo v repozitoriju CLARIN.SI. Tam so dostopni avdio posnetki v formatu wav, katerih uporaba je vezana na izvorne licence pri Videolectures.net in niso na voljo za komercialno rabo (licenca CC BY-NC-ND 4.0). Transkripcije so na voljo v treh formatih: kot TEI xml, kot vertikalna tabela Sketch Engina in kot izvorne transkripcijske datoteke, ki se lahko odprejo s programom Transcriber 1.5.1 ali drugim podobnim, ki podpira format .trs, oz. v tekstovnem urejevalniku. Dostopne so pod licenco CC-BY 4.0« (Verdonik 2018: 267).

4.3 Govorna zbirka Artur

Tako za pisna kot govorjena besedila velja, “da so njihovi tvorci hkrati tudi avtorji z avtorskimi pravicami nad besedili in pogosto obstajajo pogodbene zaveze, da bo ta podatek v jezikovnem viru ustrezno naveden. Pri posnetkih govora se v zvezi z avtorskimi pravicami in navajanjem vira srečujemo s štirimi vrstami situacij: (1) Če gre za posnetek na terenu, ki je bil narejen za namene govornega vira in zajema avtentični govor v vsakdanjih situacijah, govorci prenesejo avtorske pravice praviloma na nosilca projekta, v katerem nastaja govorni vir. Praksa je, da je v takih primerih kot vir označeno terenski/lastni

[8] V Gos Videolectures so vsi posnetki opredeljeni kot javni informativno-izobraževalni tip diskurza.
[9] V Gos Videolectures je vrsta situacije vedno opredeljena kot predavanje.
[10] V Gos Videolectures je vrsta opis diskurza vedno opredeljen kot javno predavanje.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

posnetek. (2) Če gre za posnetek, ki je bil predvajan prek radia ali televizije, so pogosto nosilci avtorskih pravic medijske hiše in so posledično te navedene kot vir. Tudi pri spletnih virih (npr. posnetki na Youtubu) je treba pogosto urediti avtorske pravice z njihovim/-i nosilcem/-i in v metapodatkih ustrezno navesti vir. Če gre za spletne dogodke, ki jih sicer organizira in objavi neka institucija (npr. spletne konference, delavnice, seminarji), je pogosto treba urejati avtorske pravice z neposrednimi tvorci teh besedil. Pri tem se pojavi vprašanje, kako je najbolj smiselno definirati metapodatek o viru: kot posameznika/e, ki je/so pravice odstopil/-i in nastopa/-jo na posnetku, ali kot institucijo, ki je organizirala in objavila spletni dogodek. V bazi Artur je bila pri tovrstnih posnetkih izbrana druga možnost. (3) Določeni internetni viri že imajo urejene avtorske pravice na način, ki omogoča nadaljnjo uporabo, in sicer pod pogoji katere od licenc Creative Commons. Taka večja vira posnetkov v slovenščini sta portala Videolectures.net in Arnes Video. V takih primerih se v obstoječih bazah za slovenščino kot vir navaja kar ime portala. (4) Določena govorjena besedila niso avtorsko varovana. V skladu z 9. členom ZASP so taka »uradna besedila z zakonodajnega, upravnega in sodnega področja«. Čeprav še ni tovrstne sodne prakse ali doktrine, se lahko kot tovrstna med drugim štejejo govorjena besedila, ki nastajajo v Državnem zboru RS v okviru zakonodajnih postopkov. V tem primeru se kot vir v bazi Artur, kjer se pojavljajo tovrstni posnetki, navaja kar Državni zbor Republike Slovenije" (Verdonik et al. 2022a: 210).

Pri ustvarjanju govorne zbirke Artur se je opravil test zakonitega interesa, v katerem smo se opredelili do razlogov za obdelovanje podatkov, do koristi obdelave podatkov in same narave osebnih podatkov, ki so v zbirki (parlamentarni govor, predavanja, posnetki izobraževalnega procesa …).

Z zbirko smo želeli zadovoljiti potrebe po računalniških storitvah s področja jezikovnih tehnologij za slovenski jezik za raziskovalne organizacije, za podjetja in za širšo javnost. Predvsem pa je bil cilj izdelava govorne baze, ki bo osnova za izdelavo splošnega razpoznavalnika govora ter podpornih tehnologij, vključno z domeno izobraževanja. Koristi od obdelave podatkov pa ima tudi javnost, saj je zbirka namenjena raziskovalnim organizacijam, podjetjem in širši javnosti. Zelo velike koristi pa ima predvsem slovenščina kot jezik saj si slovenski javni in zasebni sektor ne moreta privoščiti velikih investicij.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Za čim manjšo količino osebnih podatkov smo se na primer pri pripravi Baze Arnes Video odločili za različne korake:

- Izključitev posnetkov, kjer je večja verjetnost pojavljanja otrok
- Odstranitev posnetka v končni verziji baze za strojno učenje ter možnost brisanja posameznega posnetka na zahtevo prizadetega posameznika iz baze prehodno shranjenih posnetkov za namen priprave baze
- Dodatno ročno preverjanje, če se na katerem od posnetkov še vedno pojavlja osebno razpoznaven otroški govor, ter umik teh posnetkov pred njihovo vključitvijo v bazo za strojno učenje.
- Kjer je mogoče, izključitev govora odraslih oseb, ki niso predavatelji ali uradni govorci na dogodkih (npr. študenti), iz baze za strojno učenje.

Najbolj občutljivi podatki v bazi so bili posnetki govora otrok, saj se pojavljajo na nekaterih posnetkih med izobraževalnim procesom v šoli, vendar njihovi posnetki niso namenjeni strojnemu učenju. Posnetki, ki smo jih pridobili od tretjih oseb, so bili javni, govorci so zato vedeli, da bodo objavljeni (v skladu s CC licencami), niso pa se zavedali, da bodo uporabljeni za potrebe strojnega učenja. Sicer smo predvideli, da bo vpliv na posameznika zanemarljiv. Lahko bi prišlo do primera, ko bi se komu iz načelnih razlogov zdelo neprimerno, da se njegov/njen govor uporablja za navedeni namen. Posamezniki bi lahko do določene mere tudi izgubili nadzor nad svojimi podatki, saj bo tisti, ki bo za namene strojnega učenja prenesel bazo s posnetkom, postal tudi upravljavec te baze. Seveda bo moral tak upravljavec izpolniti zahteve GDPR. Zavedali smo se, da bi lahko po snemanju posamezniki ugovarjali obdelavi podatkov zlasti starši, ki bi izvedeli, da se obdelujejo posnetki z njihovimi otroki. Zato je bil otroški govor iz baze izključen.

Za odprtost in transparentnost projekta bo poskrbljeno na spletni strani, na kateri bo razložen tudi postopek obdelave posnetkov in druga pojasnila. Na že prenesenih bazah ne bo mogoče izvršiti izbrisa posameznikovega govora, saj bi izbris iz izvirne baze preprečil ponovljivost raziskav oziroma

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

eksperimentov. Tudi pravica do ugovora ne bi bila mogoča, saj gre za »nujne legitimne razloge za obdelavo«, ki prevladajo nad interesi, pravicami in svoboščinami posameznika.[11]

Pred izdelavo baze se je izvedla tudi ocena učinka v zvezi z varstvom osebnih podatkov (DPIA) za posnetke z Arnes Videa in za posnetke parlamentarnega govora. Pri obdelavi podatkov so bila predvidena naslednja dejanja:

»Pridobivanje podatkov iz baze Arnes Video. Določene kategorije, zlasti Vzgoja in izobraževanje ter učenje jezikov, bodo zaradi redne prisotnosti otrok na posnetkih izvzete;

- Odstranitev video zapisa za potrebe trajne hrambe v bazi za strojno učenje;
- Čiščenje vira, vključno z odstranitvijo otroškega govora;
- Urejanje transkripcij in koordiniranje študentov, vključenih v projekt;
- Ročno popravljanje in kontrola standardiziranega zapisa; in
- Dajanje na voljo baze uporabnikom za potrebe strojnega učenja prek repozitorija raziskovalne infrastrukture CLARIN.SI. Za prenos baze, ki je javno dostopna, ni potrebna registracija ali plačilo. Za te potrebe se baza trajno hrani« (*Ocena učinka v zvezi z varstvom osebnih podatkov* (DPIA) 2022).

V *Oceni učinka v zvezi z varstvom podatkov* so identificirali naslednja tveganja:

1. »Prenos posnetka s strani uporabnika, ki ima drug namen, kot je predvideni namen strojnega učenja. Tveganje ni nepomembno, a se po drugi strani z obdelavo v obliki vključitve v bazo za potrebe strojnega učenja realno ne povečuje. Namreč, isto tveganje obstaja, kadar bi se posnetek pridobil iz baze Arnes Video, ki je javnosti dostopna prek spleta. Razlika v smislu urejenosti oziroma prilagojenosti posnetkov v bazi za potrebe strojnega učenja po naši oceni ne povečuje možnosti zlorabe posnetkov, npr. za namene posegov v čast in dobro ime govorca.
2. Izguba nadzora nad posnetkom/podatki, ko je baza prenesena v sistem uporabnika kot tretje osebe oziroma novega upravljavca. Sprememb v že preneseni bazi ne bi bilo mogoče zanesljivo doseči, saj bi to zahtevalo sodelovanje vseh, ki so bazo prenesli, hkrati pa bi taka sprememba

[11] Ta vrsta govora je v celoti javna, dostopna prek drugih baz in že izvirno namenjena javnosti.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

tudi pomenila, da raziskav oz. eksperimentov na stari bazi ne bi bilo mogoče replicirati, kar je v nasprotju z načeli odprte znanosti. V tem pogledu bo omejena tako pravica do izbrisa po Členu 17 GDPR, saj bodo podatki trajno »potrebni v namene, za katere so bili zbrani ali kako drugače obdelani«, kakor tudi pravica do ugovora po Členu 21 GDPR, saj zahteve odprte znanosti tu predstavljajo »nujne legitimne razloge za obdelavo, ki prevladajo nad interesi, pravicami in svoboščinami posameznika«.

3. Občutek, da se podatki uporabljajo za namen, ki v času snemanja ni bil predviden. To tveganje obstaja, vendar je po naši oceni majhno, saj je strojno učenje za potrebe razvoja slovenščine v digitalnem okolju manjši poseg kot predhodna javna objava posnetka. Zaradi narave baze za potrebe strojnega učenja zelo verjetno ne bo mogoče upoštevati ugovora zoper obdelavo po Členu 21 GDPR, vendar pa predstavlja priprava orodij za rabo slovenščine v digitalnem okolju »nujne legitimne razloge za obdelavo, ki prevladajo nad interesi, pravicami in svoboščinami posameznika, na katerega se nanašajo osebni podatki.« Enaka omejitev bo veljala za pravico do izbrisa po Členu 17 GDPR, saj bodo po točki 17(1)(a) osebni podatki še vedno potrebni za namen, za katerega so bili obdelani, to je strojno učenje« (*Ocena učinka v zvezi z varstvom osebnih podatkov* (DPIA) 2022).

Sicer pa glede na ugotovitve ni bilo razvidno, da bi obdelava povzročila velika tveganja, če ne bi sprejeli ukrepov za blažitev tveganja. V *Oceni* so navedeni tudi nekateri ukrepi za obvladovanje navedenih tveganj, in sicer:

»Objava informacij po Členu 14 GDPR v javno objavljeni Politiki zasebnosti na spletnih straneh upravljavca, s čimer se zagotovi ustrezna informiranost o projektu;

- Pazljivost, da se uporablja izključno govor iz javno dostopne baze Arnes Video;
- Poleg vnaprejšnje izključitve določenih Kategorij, kjer se utegnejo pojavljati otroci, dodatno ročno preverjanje, če se na katerem od posnetkov še vedno pojavlja osebno razpoznaven otroški govor, ter umik teh posnetkov pred njihovo vključitvijo v bazo za strojno učenje;
- Kjer je mogoče, izključitev govora odraslih oseb, ki niso predavatelji ali uradni govorci na dogodkih (npr. študenti), iz baze za strojno učenje.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Varnostni ukrepi: repozitorij CLARIN.SI, kjer bodo podatki deponirani, je registriran kot Center tipa B pri mednarodni raziskovalni infrastrukturi CLARIN ERIC in je certificiran kot zaupanja vreden repozitorij pri mednarodni organizaciji za certificiranje Core Trust Seal (pravilniki repozitorija CLARIN.SI so dostopni na spletni strani[12]). Varnostni ukrepi vključujejo uporabo protokola HTTPS, kar pomeni, da podatkov med prenosom z repozitorija do uporabnika ni mogoče prestreči; da repozitorij za licence, ki zahtevajo identifikacijo uporabnika, uporablja sistem avtentikacije in avtorizacije AAI, kar pomeni, da se uporabnik prijavi prek svoje matične institucije, ki so ji dostopni osebni podatki uporabnika; ter da se integriteta podatkov redno preverja prek njihove MD5 kode, kar preprečuje, da bi se podatki na strežnikih pokvarili« (*Ocena učinka v zvezi z varstvom osebnih podatkov (DPIA)* 2022).

Za posnetke, kjer je bilo to potrebno, so se pridobila tudi soglasja. V Prilogi 1 dodajamo obrazec *Dovoljenje za snemanje in uporabo posnetka glasu in privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov ter dovoljenje za uporabo avtorskih pravic* kot primer. Tudi za javne nastope, konference, okrogle mize idr. smo pridobili dovoljenja od vsakega govorca na posnetku, kdor ga ni podal, je bil iz posnetkov izrezan. Vse to pa smo vnesli na posebne sezname vseh posnetkov, ki so bili namenjeni za govorni korpus. Na seznamu so bili poleg imena dogodka navedeni tudi govorci in pri vsakem od njih še določeno, ali je oseba soglasje podala ali ne.

5 Nejc Robida: Delotoki transkripcije in označevanja (Arturja)

5.1 O bazi Artur

»Baza ARTUR (Avtomatsko Razpoznavanje govora Razvoj slovenščine v digiTalnem okoljU) je nastala v delovnem sklopu 2 projekta Razvoj slovenščine v digitalnem okolju, ki je potekal od 4. 5. 2020 do 28. 2. 2023. Projekt je izvajal konzorcij 6 partnerjev iz akademskega okolja (Univerza v Ljubljani, Univerza v Mariboru, Univerza v Novi Gorici, Institut »Jožef Stefan«, Znanstvenoraziskovalni center SAZU, Inštitut za novejšo zgodovino) in 6 partnerjev iz gospodarskega okolja (Alpineon, d. o. o., Amebis, d. o.

[12] https://www.clarin.si/repository/xmlui/page/about

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

o., Aikwit, d. o. o., Vitasis, d. o. o., Slovenska tiskovna agencija, d. o. o. in Pošta Slovenije, d. o. o.) pod vodstvom Univerze v Ljubljani.

Delovni sklop 2, namenjen govornim tehnologijam, je bil eden od šestih vsebinskih sklopov. Poleg baze ARTUR so se v njem razvijale tudi tehnologije za razpoznavanje slovenskega govora. Projekt sta sofinancirali Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj.

Pri izdelavi govorne baze ARTUR so sodelovali:

- Laboratorij za digitalno procesiranje signalov, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru: koordiniranje celotnega procesa izdelave baze, koordiniranje izdelave pogovornega zapisa za nejavni in javni govor, koordiniranje pridobivanja posnetkov javnega govora
- Inštitut za medijske komunikacije, Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru: koordiniranje in snemanje posnetkov branega in nejavnega govora v SV Sloveniji, sodelovanje pri pridobivanju posnetkov javnega govora
- Laboratorij za umetno zaznavanje, sisteme in kibernetiko, Fakulteta za elektrotehniko, Univerza v Ljubljani: koordiniranje in snemanje posnetkov branega in nejavnega govora v JZ Sloveniji, koordiniranje in izvajanje končne validacije in pretvorb v končni format
- Laboratorij za podatkovne tehnologije, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani: pridobivanje in priprava posnetkov parlamentarnega govora; koordiniranje pogovornega transkribiranja parlamentarnega govora; koordiniranje podsklopa baze za sintezo govora
- Inštitut za slovenski jezik Frana Ramovša, Znanstvenoraziskovalni center Slovenske akademije znanosti in umetnosti: izvajanje standardiziranega zapisa za parlamentarni, nejavni in javni govor
- Slovenska tiskovna agencija: pridobivanje posnetkov javnega govora
- Alpineon, d. o. o.: priprava pisnih predlog za sklop branega govora
- Fixmedia, d. o. o.: zunanji izvajalec, snemanje posnetkov branega govora v SV Sloveniji
- Kreativist, d. o. o.: zunanji izvajalec, snemanje posnetkov branega govora v JZ Sloveniji

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- Akademija INT, d. o. o., in TAIA, d. o. o: zunanja izvajalca, transkribiranje pogovornega zapisa javnega in nejavnega govora
- študenti Univerze v Ljubljani in Univerze v Mariboru: snemanje branega in nejavnega govora na terenu (Verdonik et al. 2023a)

Načrt za bazo govorjenih besedil Artur je bil, da bo obsegala 1000 ur in bo namenjena razvoju razpoznavanja govora. Ciljna zasnova baze je bila 500 ur branega govora (branje vnaprej pripravljenih povedi), 100 ur prosto govorjenega, nejavnega govora (dialogi, prosto opisovanje in narekovanje), 200 ur javnega (intervjuji, tiskovne konference, okrogle mize, predavanja, strokovni dogodki …) ter 200 ur parlamentarnega govora (posnetki sej Državnega zbora RS). Predvidevali so, da bo vključenih skupno okrog 1800 govorcev (Žganec Gros et al. 2021).

Skupaj s še štirimi podpornimi orodji za izdelavo splošnih ali domensko specializiranih razpoznavalnikov, tj. akustičnim normalizatorjem, sintaktičnim normalizatorjem, punktuatorjem in fonemizatorjem, bo govorna baza Artur pripomogla k boljšemu položaju slovenščine pri razvoju govorne komunikacije človeka s strojem. Sama baza bo omogočala tudi raznovrstne druge jezikoslovne raziskave, saj bo podala prvi obsežnejši vpogled v dejansko rabo sodobnega jezika z vidika povprečnega jezikovnega uporabnika oz. bralca slovenskih besedil (Žganec Gros et al. 2021).

Pri izbiranju povedi za brani del govorne baze Artur je bil osrednji cilj, da povedi odslikavajo naravno oziroma dejansko porazdelitev trifonov v povedih, zato so morali izbrane povedi razporediti med 1000 govorcev. Za izvedbo so uporabili aproksimativni algoritem in modificiran pristop izbire parov, kot množico vseh povedi za analizo dejanske porazdelitve trifonov v povedih pa so uporabili korpus Gigafida 2.0 (Krek et al. 2020), tj. 60 milijonov povedi (Žganec Gros et al. 2021).

Izbrane povedi so predstavljale osnovni izbor besedila za branje oziroma brani govor. Izmed njih so bile izločene vse povedi, pri katerih bi lahko prišlo do razlik med zapisanim in izgovorjenim govorom (povedi s številkami, okrajšavami, akronimi, slovničnimi in pravopisnimi napakami …). Nato so te povedi še enkrat pregledali ročno, posebej so preverili, ali vsebujejo vulgarno izražanje, sovražni govor, težko izgovorljive besede, tuje besede, če je bilo besedilo nesmiselno … Niso pa bile izločene povedi, ki so

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

vsebovale napake, kot so: napačna raba sklonov, napačno stopnjevanje pridevnikov, napačna raba veznikov ki/kateri …

5.2 Obseg in sestava baze Artur

V korpusu Gos 2.0 bo ostal le del baze Artur, zato bomo nekaj podatkov o sestavi celotne baze navedli v tem poglavju, ki se osredotoča samo na Arturja. V tabeli 2 navajamo podatke o posnetkih, ki so bili končno odobreni in transkribirani, ne vseh, ki so bili zajeti.

Tabela 2: Podatki o posnetkih v celotni bazi Artur

Baza Artur	**Število posnetkov**	**Število govorcev**	**Trajanje posnetkov**
Brani govor	282.244	885	546 ur 32 min 36 sek
Nejavni govor	485	303	74 ur 4 min 20 sek
Javni govor	100	240	62 ur 3 min
Parlamentarni govor	2.799	158	201 ura 23 min 21 sek

Tabela 3: Število govorcev glede na spol (Vir: Verdonik et al. 2023b)

Spol	**Brani govor**	**Nejavni govor**	**Javni govor**	**Parlament**	**Artur 1.0**
Ženski	494	209	379	46	1128
Moški	398	159	442	112	1111
Neopredeljeno					

Tabela 4: Število govorcev glede na starost (Vir: Verdonik et al. 2023b)

Starost	**Brani govor**	**Nejavni govor**	**Javni govor**	**Parlament**	**Artur 1.0**
12-17 let	40	13	0		53
18-29 let	359	158	22		539
30-49 let	298	109	238		645
50-59 let	140	51	142		333
60+ let	55	37	113		205
Neopredeljeno			306	158	464

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 5: Število govorcev glede na izobrazbo (Vir: Verdonik et al. 2023b)

Izobrazba	Brani govor	Nejavni govor	Javni govor	Parlament	Artur 1.0
Osnovna šola – nedokončana	12	3			15
Osnovna šola – dokončana	83	12			95
Nižje poklicno izobraževanje	5	2			7
Srednje poklicno izobraževanje	113	68	6		187
Gimnazije, SSI in PTI	242	76	14		332
Višješolski programi	67	25	9		101
VS in UNI programi (1. bolonjska stopnja)	276	110	115		501
Magisterij stroke (2. bolonjska stopnja)	69	51	74		194
Magisterij znanosti (pred bolonjsko reformo)	14	7	36		57
Doktorat znanosti	11		251		271
Neopredeljeno		5	316	158	479

Tabela 6: Število govorcev glede na statistično regijo stalnega bivališča (Vir: Verdonik et al. 2023b)

Stat. regija	Brani govor	Nejavni govor	Javni govor	Parlament	Artur 1.0
Osrednjeslovenska regija	255	133	269		657
Podravska regija	200	105	44		349
Savinjska regija	139	14	43		196
Gorenjska regija	84	51	32		167
Koroška regija	44	8	5		57
Pomurska regija	36	11	13		60
Jugovzhodna Slovenija	30	13	11		54
Posavska regija	27	1	7		35
Obalno-kraška regija	23	7	9		39
Goriška regija	20	12	12		44
Zasavska regija	17		6		27
Primorsko-notranjska regija	12		7		25

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tujina	4	3			7
Neopredeljeno	1		363	158	522

Tabela 7: Število govorcev glede na prvi jezik

Prvi jezik	Brani govor	Nejavni govor	Javni govor	Parlament	Artur 1.0
Slovenski	872	364	731	156	2123
Hrvaški	5		1		6
Srbski	4	2	5		11
Bosanski	2		1		3
Makedonski	3	1	1		5
Madžarski	4		2	1	7
Italijanski			1	1	2
Ruski			2		2
Ukrajinski		1			1
Japonski			1		1
Drugo	1				1
Neopredeljeno	1		76		77

5.3 Metapodatki o posnetkih in govorcih v govorni bazi Artur

»Z vidika razvoja govornih tehnologij oziroma razpoznavalnikov govora je glavni razlog za zbiranje podatkov o govorcih in posnetkih predvsem ta, da se v govorni bazi zagotovi čim bolj ustrezna reprezentativna zastopanost vseh izrazitih govornih značilnosti, ki se spreminjajo med različnimi govorci in različnimi govornimi okoliščinami« (Verdonik 2022a: 2). Pomembni so tudi vsi podatki, ki bi lahko imeli kakršen koli vpliv na akustične in jezikovne posebnosti govora, saj lahko z obdelavo govora pridemo do informacij, ki so sicer del različnih disciplin, npr. fonetike ali celo sociolingvistike, semantike itd. Na podlagi metapodatkov lahko tvorci govornih zbirk poskrbijo za ustrezno zastopanost vseh kategorij govorcev in govora. Za razvoj splošnega samodejnega razpoznavalnika govora moramo upoštevati predvsem metapodatke o enoznačni oznaki govorca, o prvem jeziku govorca, o narečni skupini (pri spontanem nejavnem govoru), o snemalnih zvočnih okoliščinah in o spolu ter starosti govorca.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Baza Artur vključuje 4 večje sklope različnih vrst govora: brani govor po pisnih predlogah (546 ur), javni govor (javni dogodki, mediji ipd. – 62 ur), parlamentarni govor (Državni zbor RS – 201 ura) in nejavni govor (terenski posnetki prosto govorjenih monologov in dialogov – 74 ur).

»Podatki o posnetkih in govorcih so v bazi Artur organizirani kot TSV-datoteka in v obliki XML-zapisa po standardu TEI. V primerjavi s predhodnimi govornimi viri za slovenščino vključujejo predvsem zelo podroben popis tehničnih lastnosti posnetkov (npr. podatke o lastnostih izvornih posnetkov in tehnični opremi, uporabljeni za snemanje) ter vseh okoliščin, ki bi lahko na te lastnosti vplivale (od velikosti prostora snemanja, prisotnosti hkratnega govora vse do uporabe maske pri govorcih, ki je bila pogosta v času epidemije COVIDA-19)« (Verdonik 2022a: 3).

V Tabeli 8 in 9 navajamo vse oznake, ki smo jih uporabljali v okviru metapodatkov v govorni zbirki Artur.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 8: Seznam metapodatkov o posnetkih (Verdonik 2022a: 3–4)

Identifikacijski podatki in kategorizacija posnetkov	ID-posnetka	Sestavljen je iz imena baze (Artur), podatka o tipu govora (brani – B, javni – J, nejavni – N in parlamentarni govor – P), štirimestne identifikacijske številke govorca (Gxxxx), šestmestne identifikacijske številke posnetka (Pxxxxxx) ter podatka o vrsti datoteke (-avd). Pri posnetkih javnega govora, na katerih se običajno pojavlja večje število govorcev, je namesto štirimestne identifikacijske številke govorca navedba Gvecg (s pomenom več govorcev). Primer ID-posnetka: Artur-N-G5134-P600134-avd. Dodan je lahko tudi ID snemalca, ID potrjevalca in ID govorca.
	vrsta govornega dogodka	javni, nejavni, parlamentarni ali brani govor
	opisi govornih dogodkov oz. topiki	seja državnega zbora, okrogla miza, intervju, novinarska konferenca …
		Za potrebe razvoja specializiranih razpoznavalnikov v projektu Razvoj slovenščine v digitalnem okolju so v bazi Artur opredeljeni še govorni dogodki, kjer je snemanje potekalo po vnaprej pripravljenih scenarijih z dveh področij: opisovanje obrazov in upravljanje pametnega doma.
Podatki o okoliščinah snemanja	datum snemanja	mesec in leto
	podatek o občini snemanja	
	prostor snemanja	stanovanje, pisarna, studio …
	velikost prostora	20 m^2, 20–80 m^2, 80+ m^2
	prisotnost šuma	npr. šelestenje, šumenje, prometni hrup, zvok ventilatorja
	presluh	Presluh se občasno pojavi pri 2-kanalnem snemanju nejavnega govora, ko je spontani pogovor dveh sogovornikov posnet z dvema ločenima mikrofonoma. Prisotnost presluha je označena, če se pri

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

		takem snemanju pogosto in jasno sliši govor govorca z drugega kanala.
	pogost hkratni govor	
	maska	Podatek o tem, ali govorec nosi masko, je bil aktualen v času epidemije COVIDA-19, ko je veliko javnih dogodkov potekalo ob uporabi obrazne maske.
Podatki o formatu izvornih posnetkov	Vsi posnetki v bazi Artur so v formatu WAV; 44,1 kHz; pcm; 16-bit; mono.	Pridobljeni posnetki so bili tudi v formatih WAV, MP3, M4A …
Podatki o opremi, uporabljeni za snemanje	snemalna naprava	prenosni računalnik, prenosni snemalnik, pametni telefon, kamera …
	tehnične lastnosti snemalne naprave	opis naprave, vrsta in model mikrofona, snemalni program …
Podatki o viru posnetkov	vir posnetka	Vir posnetka je lahko lastni posnetek, ki ga je naredila ekipa govorne baze Artur namensko za to bazo – to so vsi posnetki branega in nejavnega govora. V primeru parlamentarnega in javnega govora pa gre za arhivsko ali drugo gradivo, pridobljeno od različnih gradivodajalcev: Državni zbor RS, STA, Arnes, ZRC SAZU, Univerza …
	spletna povezava	Pri javnem govoru je včasih na voljo tudi spletna povezava do samega posnetka.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 9: Seznam metapodatkov o govorcih (Verdonik 2022a: 4–5)

Identifikacijski in sociodemografski metapodatki	ID-govorca	Zajema ime baze (Artur), oznako vrste govornega dogodka (B, J, N in P) ter vnaprej določeno štirimestno identifikacijsko številko govorca (Gxxxx). Primer ID-govorca: Artur-N-G5097.
	spol	moški, ženski, drugo
	izobrazba	9 kategorij (osnovna šola – nedokončana; osnovna šola – dokončana; nižje poklicno izobraževanje …)
	starost	12–17 let, 18–29 let, 30–49 let, 50–59 let, 60+ let
Metapodatki o regiji govorca	občina stalnega bivališča	v RS in v tujini
	statistična regija	
	občina bivanja v otroštvu	morebitni narečni vplivi na govor
	prvi jezik govorca	Poleg govorcev, katerih prvi jezik je slovenščina, so v bazo Artur v manjši meri vključeni tudi govorci, katerih prvi jezik je hrvaščina, srbščina, makedonščina, bosanščina, ruščina, madžarščina idr.
	značilnosti govora	standardni jezik, pogovorni jezik in narečje
	izgovorne težave	težave z izgovorom glasov r, l, jecljanje ali podobno

5.4 Zapisovanje govora

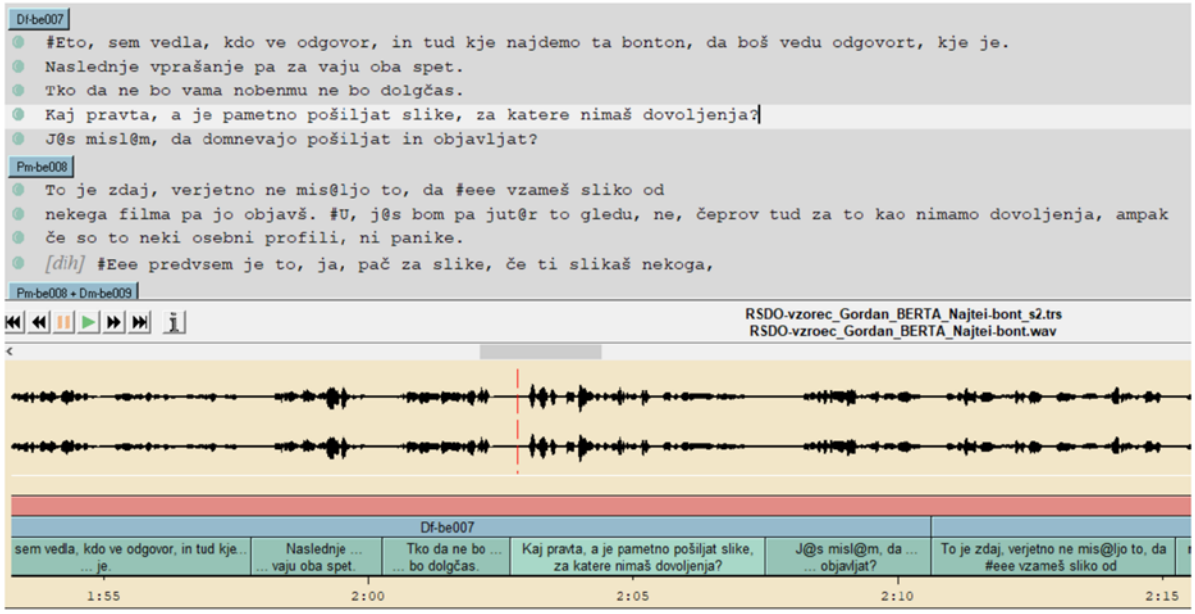
5.4.1 Segmenti in vloge

Transkribiranje je potekalo v programu Transcriber. Zapisovalci pa so morali v programu ustrezno nastaviti kodiranje (UTF-8). Vsak posnetek je bil razdeljen na ustrezno dolge enote, imenovane segmenti oziroma izjave. Segmenti niso smeli biti predolgi in niso smeli biti določeni na sredini fonema. Premor, ki je bil ustrezen za razrez, je moral biti načelom dolg vsaj 0,2 sekunde. Pri krajših premorih, so morali pri odločitvi o razrezu transkriptorji upoštevati naslednje točke:

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- če bi bili sicer segmenti predolgi,
- če se menjajo govorci ali
- če bi v segment s hkratnim govorom tako zajeli tudi veliko nehkratnega govora.

Zapisovalci so morali paziti tudi, da izjav ne režejo sredi vzdiha ali izdiha. Slediti so morali tudi navodilom, da je najbolje, če segmenti sledijo tudi semantično in skladenjsko zaključenim enotam. Primer na sliki 3 prikazuje segmente v transkripciji – rdeča črta na signalu označuje mesto, na katerem so posnetek začasno ustavili. V zelenem segmentacijskem traku pod njim se sproti izpisuje transkribirano besedilo, ki so ga vpisovali neposredno v siv urejevalnik besedil.



Slika 3: Segmentiranje (Vir: *Navodila za zapisovanje in označevanje govora v govorni bazi Artur projekta RSDO*)

Pojave, kot so premori, govor v tujem jeziku, nerazumljiv govor so označevalci označevali kot posebne segmente oziroma kot izjave brez govorca. Dele posnetke brez govora, ki so bili daljši od sekunde in pol pa kot premor. Če je bil govor nerazumljiv, pretih, hkraten in prekriven in ga niso mogli

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

transkribirati, so naredili poseben segment in vnesli oznako *neraz*, ki pomeni nerazumljivo. Tudi daljše fraze, odstavke v tujem jeziku so označili kot poseben segment, in sicer z oznako *Language*.[13]

Omenjali smo že izjave oziroma segmente, ki so najmanjša strukturna enota zapisa. Če je mogoče, so tudi prozodično, semantično in skladenjsko približno zaključena enota (meje so podobne tistim pri stavku ali povedi). Pri diskurzivnem govoru pa so morali zapisovalci označevati tudi vloge. Vlogo lahko definiramo kot govor enega govorca, dokler ga ne prekine drug govorec. Lahko je sestavljena iz ene ali več izjav oziroma segmentov. Vsako menjavo govorca so ustrezno označili in mu dodali tudi novo kodo govorca, če je bilo to potrebno.

Če je bila transkribirana datoteka z imenom *Artur-N-G5003-P600003-avd.wav*, je bila izpeljana koda govorca *Artur-N-5003*. Vse kode govorcev so bile shranjene v posebnih evidencah govorcev, njihov izvor pa je bil odvisen od koordinatorjev snemanja, tudi evidence so morali transkriptorji sproti preverjati in jih dopolnjevati. Pri vsaki menjavi so lahko pri posameznem govorcu določili še kanal signala (npr. studio ali telefon).

Paziti so morali na oporne signale, to so besede, ki jih izreče sogovornik med govorom drugega (npr. besede *ja*, *mhm*, *aha*, *aja* …). Pri takšnih dogodkih glasov v ozadju niso transkribirali, ampak so pri aktivnem govorcu le označili del govora, kjer se pojavlja *govor v ozadju.* Podobno se lahko pojavi tudi hkratni govor, po navadi na začetku ali na koncu segmenta. Tak segment se je označil kot *overlapping speech* oz. *hkratni govor*. Če je bil govor obeh govorcev razumljiv, so ga zapisali. Paziti so morali le, da je bil posamezni govor določen prvemu ali drugemu govorcu.

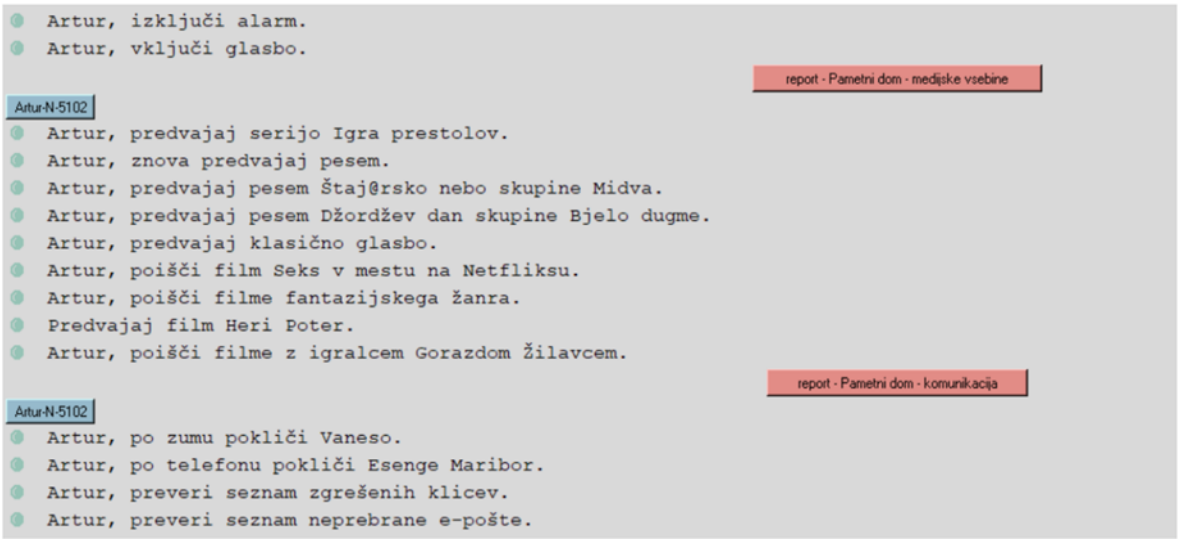
5.4.2 Označevanje tem pri posnetkih *Pametni dom*

Del nejavnega govora je bil posnet tudi po scenariju, in sicer posebej za razvoj specializiranega razpoznavalnika za pametni dom. V pogovornem zapisu pa so morali zapisovalci označiti tudi vsebinske sklope kot *teme* oziroma kot *topic*. Vnaprej je bilo določenih pet vsebinskih sklopov. Ti so:

[13] Tuja imena, ki jih ni bilo mogoče zapisati s slovenskimi fonemi, so označili z isto oznako, dodatno pa so transkriptorji izbrali še izvorni jezik besede.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- Pametni dom – bivanjsko okolje
- Pametni dom – medijske vsebine
- Pametni dom – komunikacija
- Pametni dom – upravljanje naprav
- Pametni dom – poljubno



Slika 4: Označevanje tem pri posnetkih *Pametni dom* (Vir: Verdonik et al. 2022b)

5.4.3 Pogovorni zapis

Pogovorni zapis je zastavljen na način, da bi čim bolj olajšal grafemsko-fonemsko pretvorbo in omogočal preprosto izločanje nestandardnega besedišča, namenjenega oblikoslovno-fonetičnemu leksikonu, kot je Sloleks. Govor so zapisovali v veljavnem slovenskem črkopisu, dodali pa so posebni znak za polglasnik.[14]

Vsaka izjava se je začela z veliko začetnico. Pri transkribiranju pa so uporabljali naslednja ločila: piko, vejico, klicaj, vprašaj, podpičje, opuščaj (apostrof), znak za in (&), narekovaje in dvopičje (za dobesedni navedek ipd.), vezaj in tri pike za nedokončane misli, pri samopopravljanju ali kratkih premorih znotraj

[14] Najpomembnejša razlika glede nabora glasov od transkribiranja gradiva za korpus Gos 1.0.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

stavka. Vsa ločila so morali zapisovati po trenutno še aktualnih pravopisnih pravilih (SP 2001), uporabljala pa so se samo v skladenjski rabi. Z oklepajem so se označevali tudi besedne fragmente pri samopopravljanju ali pri prekinitvah, na primer *lju()*. Posebej pa so pazili na anonimizacijo osebnih podatkov (poglavje 4.1), uporabili so oznako *shh*.

Glavne in vrstilne številnike so označevali z besedo in v oglatih oklepajih, na primer [petinštirideset] in [petsto dvaindevedeset]. Imeli so navodila, da govor zapisujejo v knjižni normi, če izgovor ni bistveno odstopal od zbornega, izjema je bil le polglasnik. Pri zapisu t. i. neknjižnih oz. nestandardnih besed so sledili načelom, ki so bili zastavljeni posebej za projekt v *Navodilih za zapisovanje in označevanje govora v govorni bazi Artur projekta RSDO*.

Za polglasnik so transkriptorji uporabljali znak @. Zapisovali so ga tudi pri oziroma med in pred zvočniki, npr. *r@*, *misl@m*, *hit@r*, pri enoglasovnih členkih in predlogih (*s@*, *z@* …), enozložnih besedah (*n@č*), v večzložnih besedah (*k@šni*) in zborni izreki (*p@s*). Kadar je prišlo do redukcij pri izgovoru oblik pomožnega glagola *biti*, polglasnika niso zapisovali, na primer zapisi *ne b* (ne bi), *pa b mene* (pa bi mene). Premen po zvenečnosti niso upoštevali (tud tak). Druače je bilo pri predlogih s/z in k/h, te so zapisali, kot so jih govorci izgovorili (nezveneče ali zveneče) skupaj z besedo ali pa kot samostojni izgovor s polglasnikom.

Kadar dvoustnična varianta fonema *v* ni bila nosilec zloga, so jo zapisali s črko *v*, na primer *prov*, *nav*, *navm*, *odpravt*, *davn*, *gledavc*, *pov@n* ... Če je bila dvoustnična varianta fonema *v* izgovorjena skladno s standardom (oziroma knjižno), so ohranili tudi knjižni zapis s črko *v* ali *l*, na primer *bil*, *gledal, siv*. Kadar pa je bil *v* izgovorjen kot samoglasniški *u*, torej kot nosilec zloga, so ga zapisali s črko *u*, na primer *pršu*, *vidu*, *u tem delu* … S črko *u* je bil zapisan tudi predlog *v*, kadar je bil izgovorjen samoglasniško oziroma »u-jevsko«.

Diftonge in druge pokrajinsko specifične foneme, ki jih ni v knjižnem jeziku, so zapisovali z najbližje ustreznimi črkami, odvisno tudi od izgovora v konkretnih primerih, npr. *ej, ov, je; u* za ü; »h« za zveneči primorski h in r za mehkonebni koroški r. Za zadnja glasova so uporabljali tudi oznaki *$g* in *$r*.

Mašila, ki pogosto zapolnijo premore v govoru, so zapisovali s tremi črkami in znakom #, na primer *#eee*, *#eem*, *#een*, *#nnn*, *#mmm*. Ali pa so uporabili nize črk, ki so izgovoru najbolje ustrezali. Imeli pa

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

so tudi poseben seznam zapisovanja medmetov (tabela 10). Še posebej pozorni pa so bili na zapise neleksikaliziranega glasovnega zanikanja, takrat so uporabljali zapisa, kot sta #*nn* ali #*aa*. Pri vsaki izjemi so se morali zapisovalci dogovoriti s koordinatorjem transkribiranja.

Tabela 10: Del seznama dogovorjenega zapisa medmetov oziroma mašil v govoru (Vir: Verdonik et al. 2022b)

#a	#ahoj	#brum	#dh
#aa (zanikanje)	#aj	#bu	#dum
#aaa	#aja	#bvak	#e
#aaaa (a z vprašalno intonacijo v vlogi vprašanja)	#ajah	#buf	#eee
	#ajaj	#bum	#eem
	#aje	#bumč	#een
#aam	#ajej	#bvum	#eev
#aan	#ajo	#bzz	#eh
#ah	#ajoj	#bž	#ehe
#aha	#alo	#ccc (tleskajoči zvok z jezikom, ki se trikrat ponovi in izraža, da nečesa ne odobravamo)	#eheh
#ahah	#ao		#ehehe
#ahaha	#aua		#ej
#ahahaha	#auva		#eje
#ahja	#av		#ejo
#ahjoj	#ba	#ck	#ejoj
#ahm	#bljeh	#dammm	#fuf

Načela zapisovanja medmetov so bila:

- »izraze zapišemo raje z eno besedo kot več besedami (npr. #ojoj namesto #o #joj),
- kjer ni bistvene razlike v zvočni podobi in funkciji/pomenu, ohranimo enoten zapis za različne rabe (npr. #mhm bi posamično morda zapisali tudi kot #ehm, vendar je razmejitev težko objektivno določiti, zato raje ohranjamo vedno #mhm),
- izraze zapisujemo prednostno s tremi črkami, tako da se razlikujejo od drugih besed (npr. raje #vaa kot #va), razen kjer ni nevarnosti, da bi bil zapis identičen zapisu kakih drugih besed, ali če je drugačen zapis že močno uveljavljen (npr. #eh),
- dvoustnični U[15] prednostno pišemo z 'v' (#av, #vav),
- podaljševanje glasov se ne označuje z več črkami, ampak se ohranja enoten zapis (npr. vedno #jee, ne #jeee ali podobno),

[15] Gre za varianto fonema /v/, izgovorjenega za samoglasnikom in pred soglasnikom ali na koncu besede.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- prednost ima poslovenjen zapis (npr. jes, ne yes, okej, ne ok ali okay)« (*Navodila za zapisovanje in označevanje govora v govorni bazi Artur projekta RSDO)*.

Členek ta so v pogovornem zapisu transkribirali skupaj (*talep*, *tamal* …), v standardnem pa po trenutnih pravopisnih pravilih, torej narazen. Lastna imena so zapisovali skladno s pravopisom (*Delo*, *Novo mesto*), tuja pa po izgovoru in z veliko začetnico (*Bler*, *Hjuston*, *Nju Jork* …), prav tako so pisali citatne in tuje besede, ki niso lastno ime (*džakuzi*, *let it bi*, *fani* …).

Med posnetki je tudi črkovanje povedi in kratic. Pri črkovanju so upoštevali trenutno kodifikacijo in dva mogoča načina navajanja črk, in sicer z dodajanjem samoglasnikov /a/ ali /e/ ali pa izgovor s polglasniki ([ˈbeː], [ˈhaː], [ˈbə] …). Črkovani pogovorni zapisi so bili tvorjeni samodejno in nato popravljeni po potrebi, če je govorec črkoval po svoje. Pri tujih črkah so *Q* glede na izgovor zapisali kot *ku* ali *kju*, *Y* kot ipsilon, *X* kot *iks*, dvojni v (W) pa kot dvojni-v@[16].

5.4.3.1 Zapis parlamentarnega govora

V pogovornem zapisu so lahko kratice zapisovali tudi z velikimi črkami (S@M@C@ju, ZOFI), premene po zvenečnosti so lahko zapisali, kot so jih slišali, torej (*knjigovotski*, *potpremo*, *gdo*). Dvoustnično varianto fonema /v/[17] so praviloma zapisovali z u (biu, instrumentou, nedauna, držaunega). Polglasnik zlasti ob zlogotvornih zvočnikih ni bil nujno zapisan z znakom @, ampak brez (*kr*, *mogl*, *pršu*, *prpelal*, *d je šou*, *prpelal*). Transkriptorji so lahko podaljšanje izgovora samoglasnikov označili z dvema istima zaporednima črkama (*daa*).

5.4.4 Standardni zapis

Tako kot pogovorni zapis so tudi standardni zapis obdelovali v programu Transcriber. Standardni zapis je bil prvotno samodejno tvorjen iz pogovornega zapisa, vanj pa so zapisovalci vnašali popravke, saj je lahko samodejna pretvorba poiskala napačno standardno ustreznico pogovorni obliki besede.

Vsak posnetek je imel v evidenci naslednje podatke:

[16] Zapis z vezajem kot ena beseda.
[17] Gre za varianto fonema /v/, izgovorjenega za samoglasnikom in pred soglasnikom ali na koncu besede.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- podatke o diskurzih/posnetkih: ID posnetka, leto, kraj, vir, kanal, ročna ocena kvalitete posnetka, tip diskurza, tip govornega dogodka, opis okoliščin;
- podatke o govorcih: ID, spol, starost, regija, prvi jezik, izgovorne težave (da/ne, opis).

Transkriptorji so ob zapisovanju vodili tudi dodatno evidenco standardiziranega zapisa, kjer so lahko težavnejše primere tudi predebatirali z drugimi koordinatorji, tako so dosegli, kolikor je mogoče enoten standardni zapis v zbirki. Segmentiranje ni bilo potrebno, saj so ga opravili že na ravni zapisovanja pogovornega zapisa.

Standardni zapis, poleg pogovornega, je nujen, ker omogoča:

- uspešno samodejno oblikoslovno, skladenjsko in drugo označevanje,
- samodejno pretvorbo pogovornega zapisa v standardizirani zapis,
- dopolnjevanje podatkov o obstoječih in novih besedah v slovarju.

Število besed v pogovornem in standardnem zapisu se je moralo vedno ujemati, tudi pri kraticah in zloženkah, le ločila so bila lahko postavljena drugače kot v pogovornem zapisu. Zapisi fragmentov so v standardnem zapisu ohranili isti zapis kot v pogovornem, enako je bilo z anonimizacijo podatkov. Tudi številke so zapisovali v oglatih oklepajih, vendar so morali upoštevati knjižni zapis (*petsto dvaindevetdeset*). Zapis datumov so ohranili v oglatih oklepajih (*{peti osmi dva tisoč devet}*). Nerazumljivi odstavki so ohranili oznako *neraz*, napake pri izgovoru pa so se pri nedvoumnosti popravile (na primer *indidualnih* v *individualnih*).

Če je izgovor besede odstopal od izgovora, kot je predviden s pravorečjem, in je bodisi prišlo do glasovnih premen bodisi do površnosti, nedoslednosti, motenj v govoru, posebnosti, so besede zapisali tako, kot bi jih zapisali po knjižni normi. Primeri so:

- lej -> glej
- luškan(o) -> luštkan(o)
- pomožni glagol sma -> sva
- zele -> zdajle
- predloga s/z -> s pred nezvenečimi soglasniki, z pred ostalimi glasovi

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- predloga k/h -> h pred g in k, pred ostalimi k
- pogovorni nedoločnik -> dolgi nedoločnik
- dons -> danes
- p@rčakvala -> pričakovala (Verdonik et al. 2022c).

Kadar so imele besede posebne narečne ali pogovorne značilnosti tudi na primer na oblikoslovni ravni, so ohranili izvorno obliko ali poiskali novo enotno obliko. Na primer:

- kao -> kao
- fora -> fora
- probavu -> probaval
- trofu -> trofil
- zrihtov -> zrihtal
- naštimaš -> naštimaš
- mahnjen -> mahnjen
- pol (v pomenu potem) -> pol
- žiher (v pomenu lahko) -> žiher
- ziher (v pomenu varno, zagotovo) -> ziher
- (po)tagati -> (po)tagati
- trendati -> trendati
- snapnati -> snapniti (Verdonik et al. 2022c*)*

Pri pisanju skupaj, narazen ali z vezajem so upoštevali pravopisno normo, oblika pa je morala biti enaka tisti v pogovornem zapisu. Kratice so zapisovali običajno (*RTV*, *TRR*, *Sazuja* …), člen ta pa s pomočjo znaka za plus (*ta+lep*, *ta+mali* …). Enako so zapisovali standardizirane oblike zanikanega pomožnega glagola, ki so bile v pogovornem zapisu zapisane kot ena beseda, na primer *navm*, *navš* kot *ne+bom*, *ne+boš*. V standardiziranem zapisu so bila domača in tuja lastna imena so bila zapisana po pravopisnih pravilih (*Novo mesto*, *Blair*, *New York* …), enako tudi citatne besede. Pri tujih besedah pa so upoštevali zapis v tujem jeziku (*yes*, *let it be*, *funny* …).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

6 Nejc Robida: Združevanje virov za novi standard

6.1 Razhajanja v metapodatkih

»/V/ projektu Razvoj slovenščine v digitalnem okolju je bil iz velikega obsega posnetkov za govorno bazo Artur izbran primeren del za nadgradnjo govornega korpusa Gos. Ob tem pa se je v veliki meri ravno v zvezi z metapodatki o govorcih in posnetkih zgodilo precej razhajanj, ki so večinoma posledica bolj natančnega popisovanja podatkov, specifik ali pa namena baze, povzročajo pa težave ob združevanju gradiv« (Verdonik et al. 2022a: 209).

Osnovne razlike se pokažejo že v tipih diskurza. Pri bazi Artur na primer manjka kategorija javnega razvedrilnega diskurza, ki ga v Arturju na primer ni, saj bi lahko vsa besedila uvrstili v kategorijo javni informativno-izobraževalni diskurz. Skupen v obeh korpusih je na primer brani govor. Tako baza Artur kot korpus Gos imata še 20 podrobnejših vrst govornih dogodkov, pri Arturju na primer črkovanje, področno specifični razpoznavalniki za pametni dom in opisovanje obrazov … Vrsta govornega dogodka je zelo pomemben metapodatek, saj omogoča naknadno prekategorizacijo ob združevanju različnih virov.

Podatka o času in lokaciji snemanja sta obvezna v obeh korpusih. V korpusu Gos lahko najdemo podatka o kraju snemanja (anonimiziran) in regiji, v Gosu so se oprli na registrska območja. V Arturju pa so zabeležili občino snemanja, tako so dosegli pravo mero med natančnostjo in ohranitvijo anonimnosti govorcev.

Razlike pri metapodatkih o govorcih se pojavljajo skoraj pri vseh kategorijah. Težave so bile predvsem pri opredelitvi o regionalnem vplivu na govor. V korpusu Gos so uporabljali 15 kategorij, in sicer je bilo 11 registrskih območij, tri kategorije so imeli za zamejske Slovence in eno posebej za tiste, ki jim slovenščina ni prvi jezik. Ta razdelitev je bila zaradi razvejanosti slovenskih narečij preohlapna. V času zbiranja gradiva za bazo Artur je bilo 212 občin, te pa so lahko preprosto razdelili na 12 statističnih regij. Ker se marsikdo seli po Sloveniji in imajo posamezniki različna ozadja v družini, se lahko regionalni vplivi tudi mešajo. Zato so lahko govorci v korpusu Gos zase izbrali do pet regionalnih pripadnosti. »V bazi Artur je bila opredelitev geografske mobilnosti skozi čas poenostavljena na dve vrsti metapodatkov, prva se nanaša na občino bivanja v otroštvu, druga na občino trenutnega stalnega

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

bivališča. S tem se izgubi precej informacij o morebitni dodatni mobilnosti posamezne osebe, ki bi sicer bile pomembne za podrobno analizo govora posameznega govorca, vprašljivo pa je, koliko so relevantne za (kvantitativno) korpusno analizo ali za morebitno prilagajanje razpoznavalnika govorcem po regijah« (Verdonik et al. 2022a: 211). V obeh korpusih so tudi govorci, ki jim slovenščina ni prvi jezik, zato je to v obeh korpusih pomemben metapodatek. Velikokrat metapodatka o starosti in izobrazbi v korpusih Gos in Artur velikokrat manjkata, pri posnetkih, kjer jih lahko najdemo, so informacije med korpusoma različno podrobne. Tako je tudi pri podatki o izobrazbi, v Gos-u je 4-stopenjska, v Arturju pa ima 9 stopenj.

Seveda se je videlo, da je veliko lažje število kategorij zmanjšati in jih prilagajati glede na različne skupne lastnosti različnih korpusov, zato je bolje imeti podrobnejše metapodatke, ki jih lahko naknadno poenostavimo ali prekategoriziramo.

7 Nejc Robida: Gos 2.0 kot baza

Gos je referenčni korpus govorjene slovenščine. Obsega transkripcije okrog 310 ur posnetkov (po)govora v najrazličnejših situacijah, ki smo jim izpostavljeni vsak dan: od radijskih in televizijskih oddaj prek šolskih ur in predavanj do zasebnih pogovorov med prijatelji ali v krogu družine ter raznih delovnih sestankov, svetovanj, pogovora ob prodaji, storitvah ipd. Zapis govora na posnetkih je narejen v dveh različicah, standardizirani in pogovorni, in obsega skoraj 2,4 milijona besed.

Korpus Gos 2.0 je najnovejša različica korpusa, ki je nastala v okviru projekta Razvoj slovenščine v digitalnem okolju z združitvijo korpusov Gos 1.1, Gos VideoLectures 4.2 in dela govorne baze Artur. V primerjavi s prvotno različico korpus Gos 2.0 vsebuje več kot podvojeno količino posnetkov in transkripcij, zaradi poenotenja vseh treh virov pa so bila nekoliko spremenjena tudi načela zapisovanja govora. Pri zasnovi korpusa Gos 2.0 je bila posebna pozornost namenjena njegovi aktualnosti (posnetki zajemajo obdobje od leta 2007 do 2022) in uravnoteženosti glede na različne tipe govornih dogodkov.

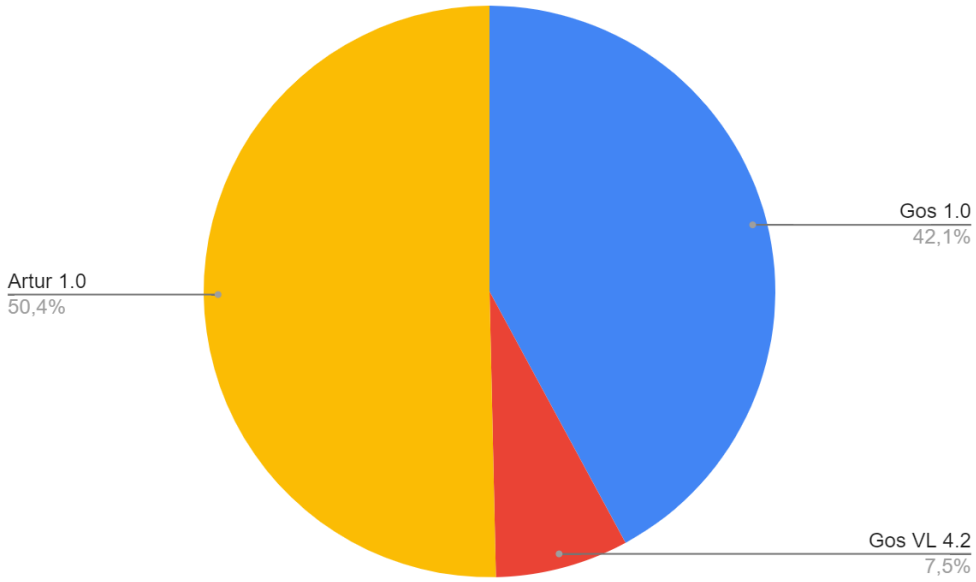
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Po korpusu lahko iščemo prek spletnega vmesnika na spletni strani http://viri.cjvt.si/gos, ki poleg različnih možnosti iskanja po obeh ravneh zapisa omogoča tudi poslušanje pripadajočih posnetkov in filtriranje zadetkov po bogatem naboru metapodatkov, kot so vrsta dogodka in komunikacijski kanal ter spol, starost, izobrazba in regionalna pripadnost govorca. Za jezikoslovno rabo je korpus prosto dostopen tudi v konkordančnikih NoSketch Engine[18] in KonText[19], ki ju vzdržuje raziskovalna infrastruktura CLARIN.SI.

7.1 Vizualna predstavitev korpusa Gos 2.0

Tabela 11: Sestava korpusa Gos 2.0 glede na število besed

Korpus	Število besed	Odstotek
Gos 1.0	1.000.000	42,1 %
Gos VL 4.2	179.000	7,5 %
Artur 1.0	1.196.000	50,4 %
Gos 2.0	**2.379.000**	**100 %**



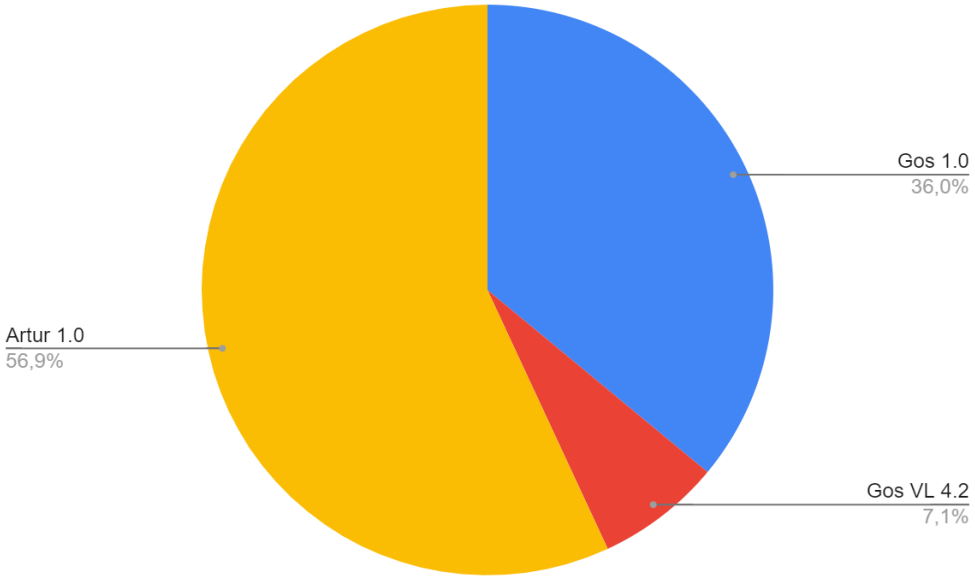
Graf 1: Sestava korpusa Gos 2.0 glede na število besed

[18] https://www.clarin.si/ske/#dashboard?corpname=gos20
[19] https://www.clarin.si/kontext/query?corpname=gos20

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Tabela 12: Sestava korpusa Gos 2.0 glede na število ur posnetkov

Korpus	Število ur posnetkov	Odstotek
Gos 1.0	112	36,0 %
Gos VL 4.2	22	7,1 %
Artur 1.0	177	56,9 %
Gos 2.0	**310**	**100 %**



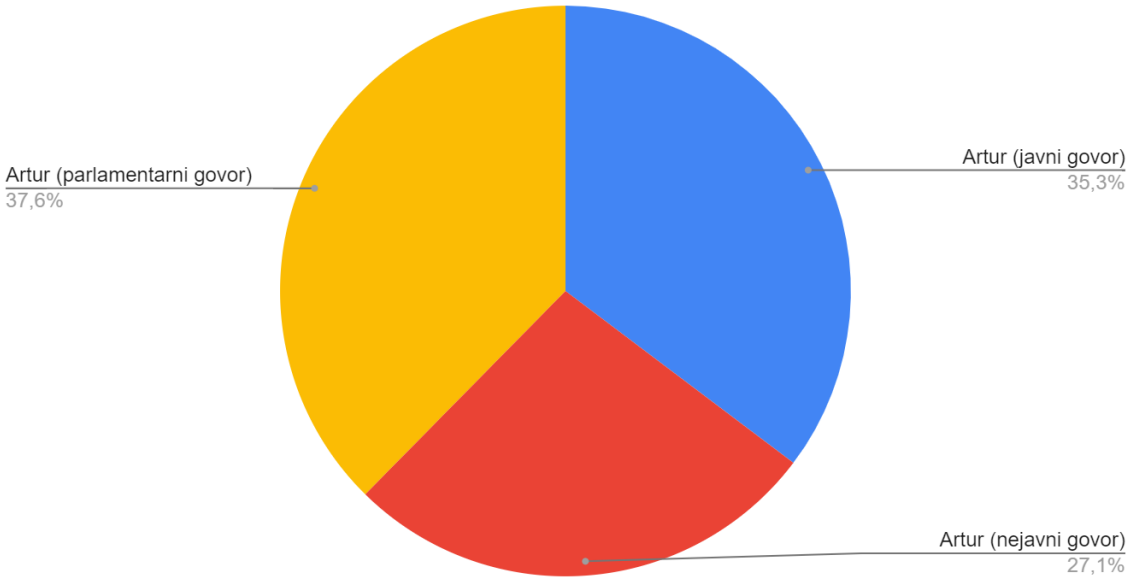
Graf 2: Sestava korpusa Gos 2.0 glede na število ur posnetkov

7.2 Vizualna predstavitev korpusa Artur 1.0

Tabela 13: Sestava korpusa Artur 1.0 (del v Gosu 2.0) glede na število besed

	Število besed	Odstotek
Artur (javni govor)	422.000	35,3 %
Artur (nejavni govor)	324.000	27,1 %
Artur (parlamentarni govor)	450.000	37,6 %
Skupaj	1.196.000	100 %

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

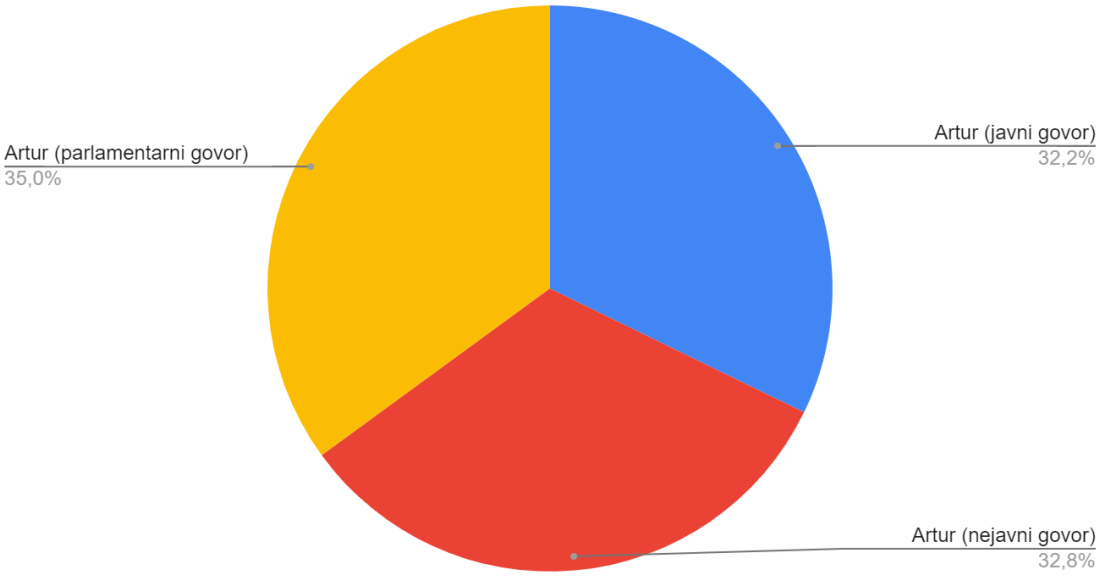


Graf 3: Sestava korpusa Artur 1.0 glede na število besed

Tabela 14: Sestava korpusa Artur 1.0 glede na število ur posnetkov (del v Gosu 2.0)

Artur 1.0	**Število ur posnetkov**	**Število posnetkov**	**Število govorcev**	**Odstotek ur posnetkov**
Artur (javni govor)	57	92	218	32,2 %
Artur (nejavni govor)	58	187	158	32,8 %
Artur (parlamentarni govor)	62	913	158	35,0 %
Skupaj	177	1.192	534	100 %

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Graf 4: Sestava korpusa Artur 1.0 glede na število ur posnetkov

8 Kaja Dobrovoljc, Luka Terčon: Konkordančniki korpusa Gos 2.0

Da bi vsebino korpusa Gos 2.0 približali raziskovalcem govorjenega jezika in drugim uporabnikom, smo v okviru projekta razvili tudi novi specializirani spletni vmesnik (konkordančnik) v okviru spletnega portala Centra za jezikovne vire in tehnologije Univerze v Ljubljani, dodatno pa je korpus za brskanje na voljo tudi na standardnih konkordančnikih infrastrukture CLARIN.SI noSketchEngine in KonText. Njihove funkcionalnosti podrobneje predstavimo v nadaljevanju, poglavje pa sklenemo s pregledom sorodnih konkordančnikov po svetu in iz tega izpeljanimi priporočili za nadaljnji razvoj.

8.1 Novi konkordančnik na spletnem portalu CJVT

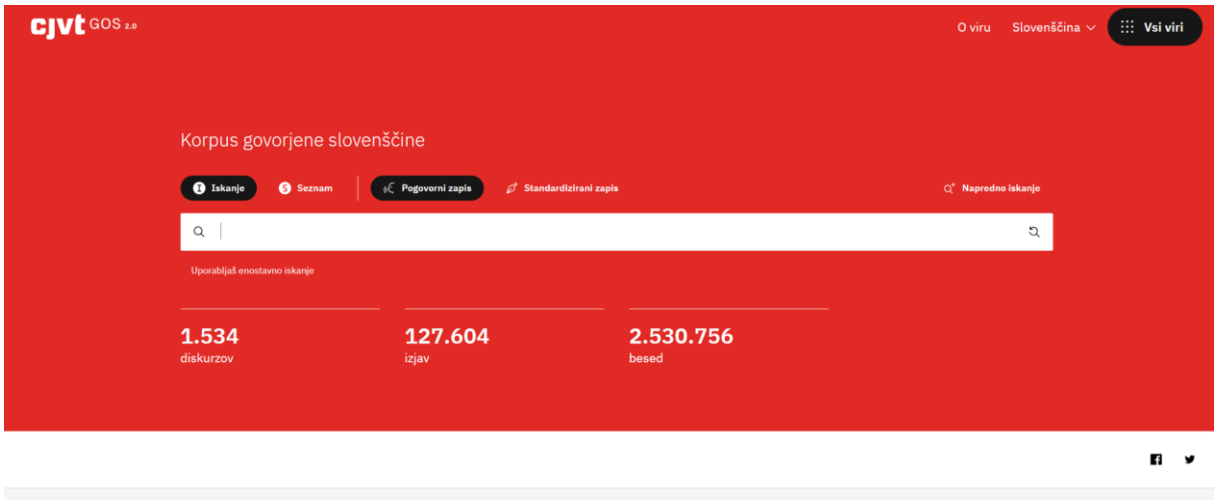
Ob objavi korpusa Gos 1.0 v okviru projekta Sporazumevanje v slovenskem jeziku (2008-2013) je bil zanj izdelan tudi specializirani konkordančnik,[20] ki danes ni več vzdrževan. Za bazo Gos 2.0 je bil zato v okviru projekta izdelan povsem nov specializirani konkordančnik za brskanje po novem referenčnem

[20] https://www.korpus-gos.net/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

korpusu govorjene slovenščine na spletnem portalu CJVT UL[21]. Programska koda konkordančnika je odprto dostopna na repozitoriju CLARIN.SI[22].

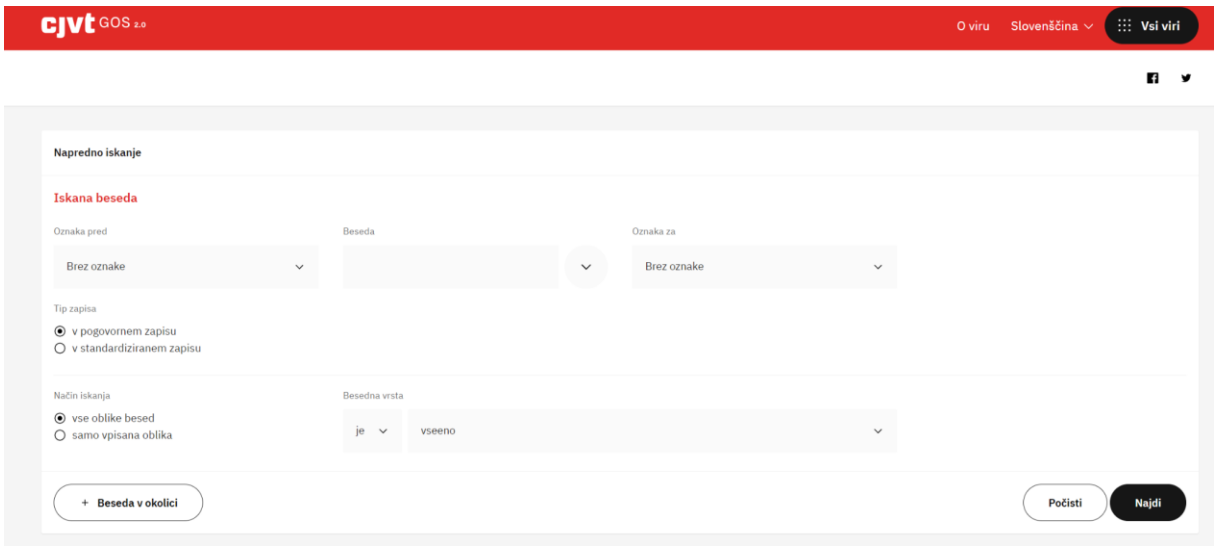
Ob prihodu na vstopno stran konkordančnika (Slika 5) uporabniki vpišejo iskano besedo ali besedno zvezo in pri tem izberejo način iskanja (privzeto iskanje po konkordancah ali iskanje po seznamih oblik) in raven transkripcij, po katerih naj se iskanje izvede (pogovorni ali standardizirani zapis). Možen je tudi preklop na vmesniško okno za napredno iskanje, v katerem lahko uporabniki opredelijo podrobnejše slovnične lastnosti iskane besede in besed v njeni okolici (Slika 6).



Slika 5: Vstopna stran konkordančnika z okencem za enostavno iskanje

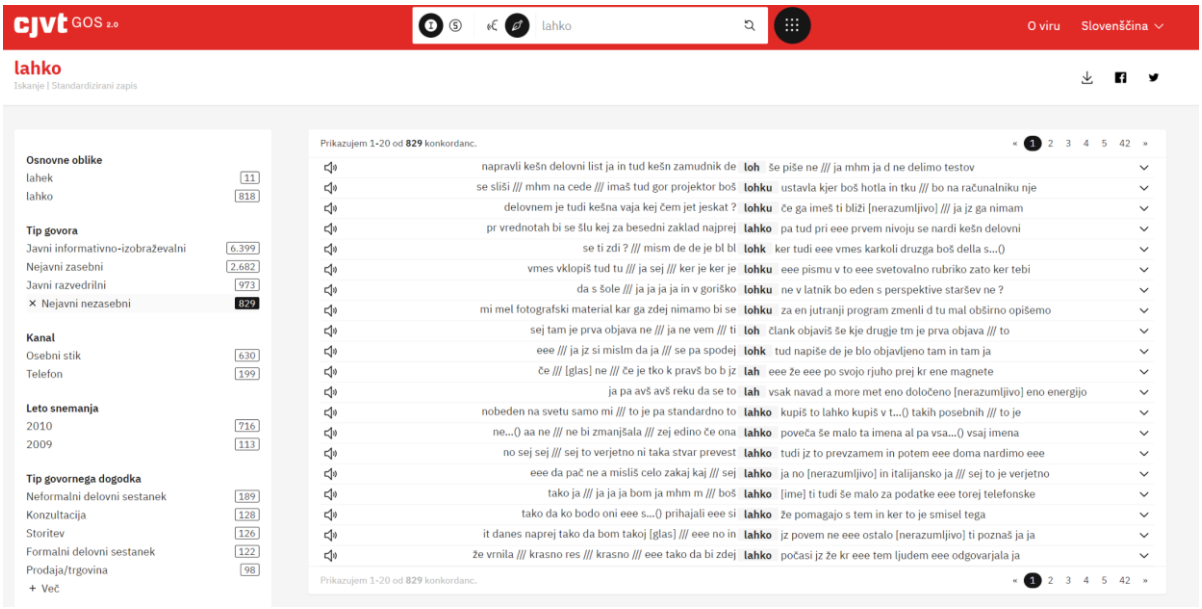
[21] https://viri.cjvt.si/gos/
[22] https://github.com/clarinsi/rsdo_gos

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 6: Vmesniško okno za napredno iskanje

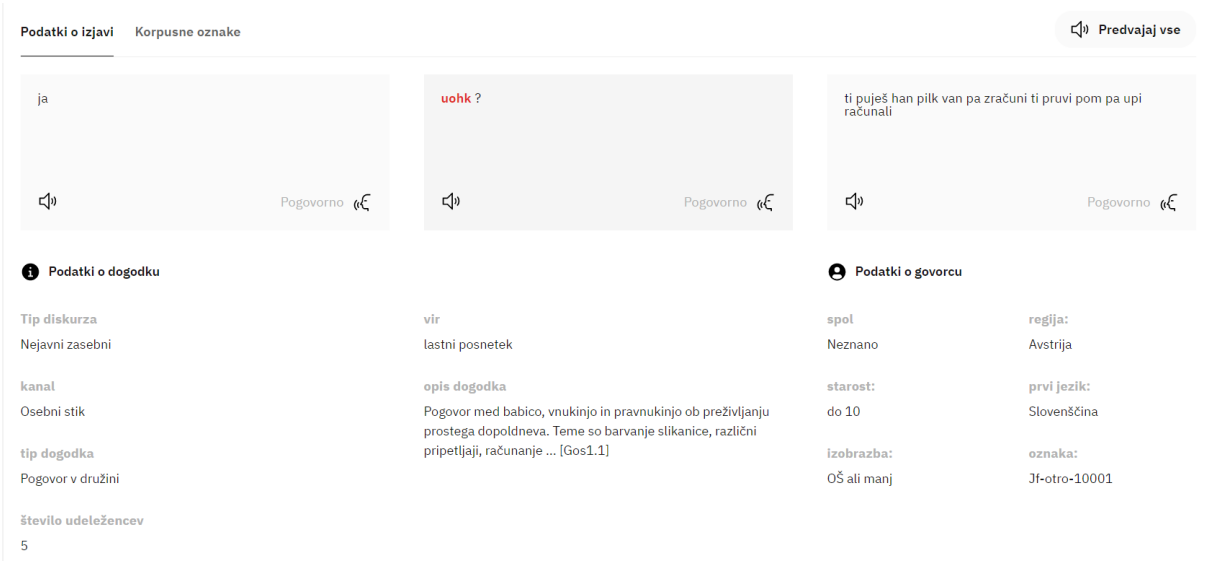
Po izvedbi (običajnega) iskanja se rezultati prikažejo v obliki konkordančnega niza, ob katerem so na levi strani navedene različne vrste metapodatkov o posnetkih, okoliščinah in govorcih, ki jih je mogoče uporabiti tudi kot filtre za nadaljnje ožanje nabora rezultatov (Slika 7). Konkordance lahko v obliki tabelarične besedilne datoteke uporabniki prenesejo tudi na svoj računalnik ali povezavo do njih delijo preko različnih družbenih omrežij.



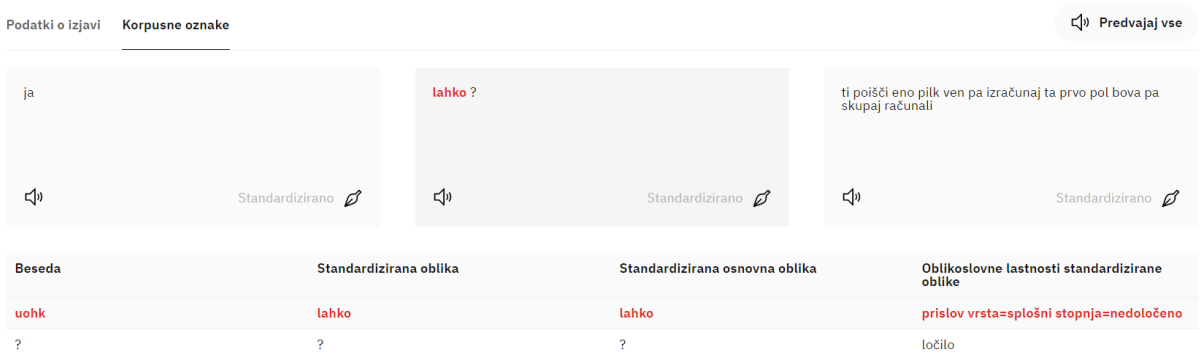
Slika 7: Stran s prikazom rezultatov iskanja po konkordancah.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Primere rabe v konkordančnem nizu je mogoče tudi poslušati s klikom na ikono zvočnika na levi, klik na puščico na desni strani konkordance pa uporabnika vodi do podrobnejših informacij o primeru. Na prvi ravni (Slika 8) te vključujejo izpis predhodne oz. naslednje izjave ter podrobnejše podatke o dogodku in govorcu. Na drugi ravni pa si s klikom na zavihek *Korpusne oznake* uporabnik lahko ogleda tudi pripisano osnovno obliko (lemo) in slovnične lastnosti posameznih pojavnic v izjavi (Slika 9).



Slika 8: Prikaz podrobnejših podatkov o izjavi

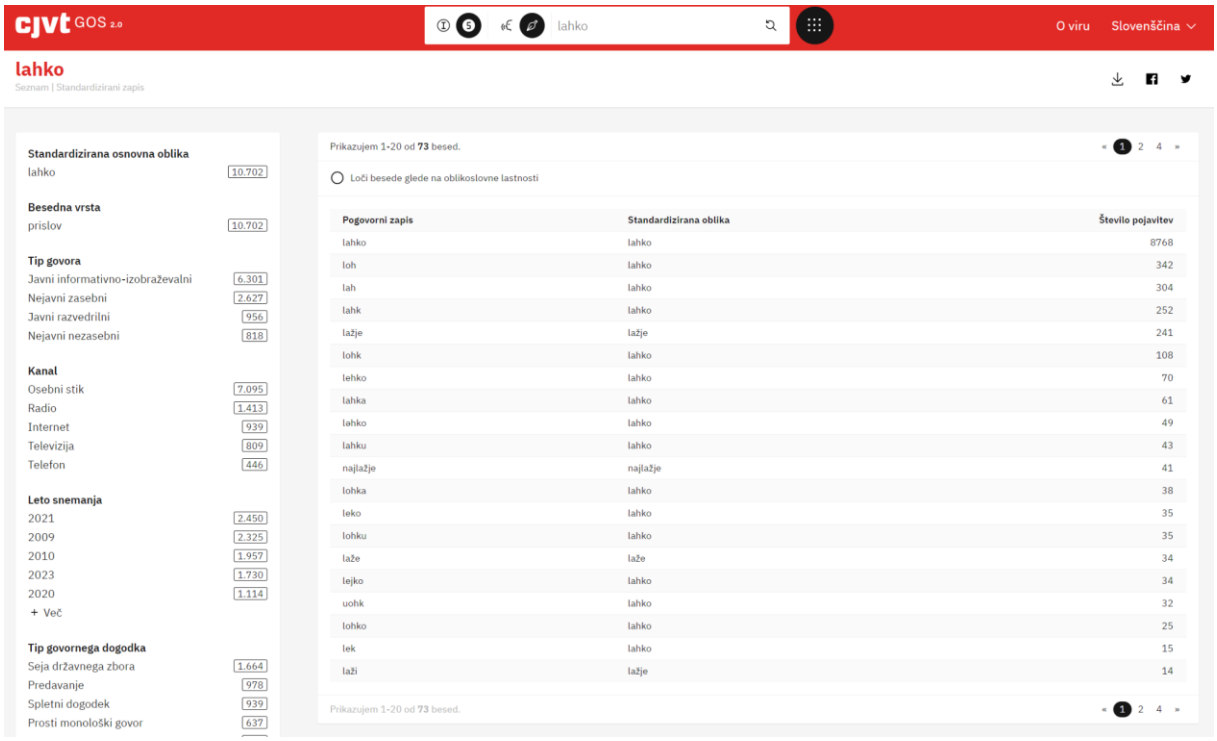


Slika 9: Prikaz podrobnejših podatkov o slovničnih oznakah izjave

Zgornji del zaslona z rezultati (Slika 7) uporabniku omogoča tudi spremembo iskanja oz. izvedbo novega, ne da bi se moral vračati na vstopno spletno stran. Konkretno lahko uporabniki za obstoječi iskalni pogoj s klikom na ikono s črko *I* oz. *S* preklopijo z običajnega iskanja na iskanje po seznamih in

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

obratno (Slika 10), s klikom na ikono govorca oz. pisala pa se lahko na enak način preklaplja med iskanjem po pogovornem oz. standardiziranem zapisu. V iskalno okence lahko uporabniki vnesejo tudi povsem nov iskalni pogoj, pri čemer je s klikom na lupi podobno ikono mogoč tudi ogled zgodovine vseh izvedenih iskanj. Poleg iskanja po samem viru pa konkordančnik omogoča tudi izvedbo enakega poizvedovanja (npr. iste besede) po drugih virih na portalu CJVT, tj. Slovenskem oblikoslovnem leksikonu Sloleks, Slovarju sopomenk sodobne slovenščine, Kolokacijskem slovarju sodobne slovenščine ali Korpusu pisne standardne slovenščine Gigafida.



Slika 10: Stran s prikazom rezultatov po preklopu na iskanje po seznamih.

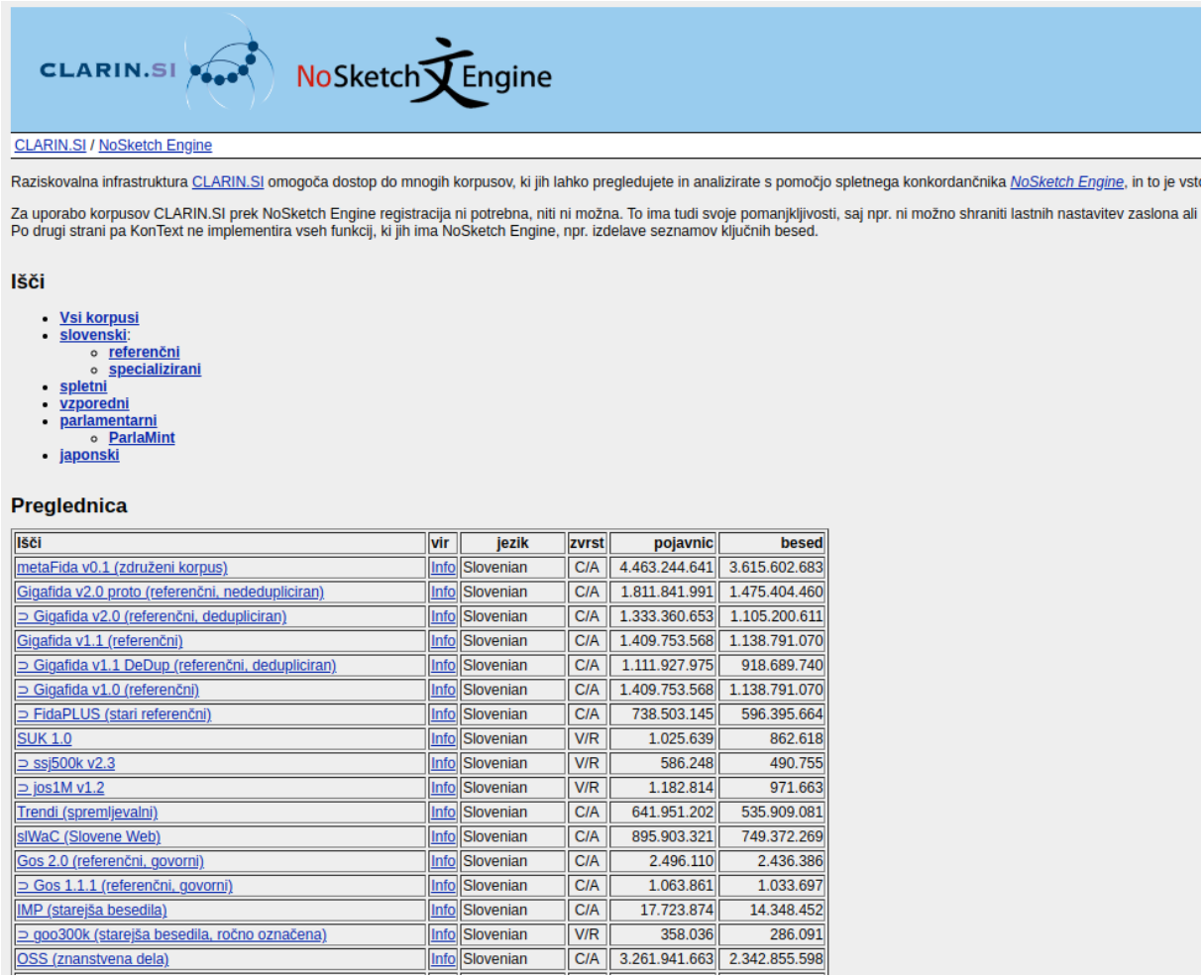
8.2 Konkordančnika noSketchEngine in KonText

Do korpusa Gos 2.0 lahko dostopamo tudi prek konkordančnikov noSketchEngine in KonText. V tem poglavju predstavimo uporabo obeh vmesnikov za brskanje po novi različici korpusa Gos. Podobno kot v prejšnjem poglavju so opisom za lažjo predstavo dodane slike, ki prikazujejo uporabo.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

8.2.1 Konkordančnik noSketchEngine

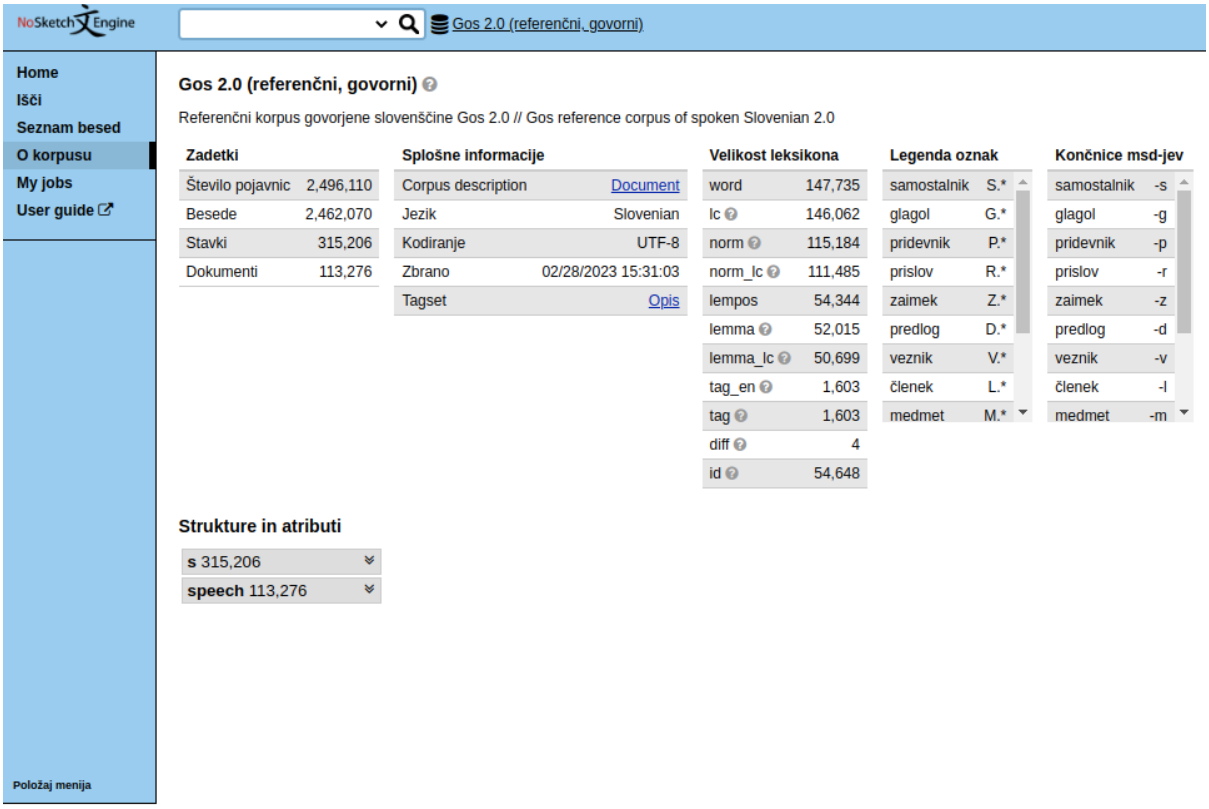
Konkordančnik noSketchEngine (Richly 2007; Kilgarriff et al. 2014) je dostopen na spletnem naslovu https://www.clarin.si/noske/. V preglednici s korpusi, ki so na voljo (Slika 11), najprej izberemo Gos 2.0, nato pa se prikaže začetni zaslon s pregledom osnovnih značilnosti in statistik korpusa (Slika 12). Na levi lahko izbiramo med šestimi različnimi zasloni: *Home* nas povrne nazaj na vstopno stran orodja, *Išči* nam prikaže iskalni zaslon, *Seznam besed* nam omogoča sestavljanje seznamov besed, lem ali drugih korpusnih oznak, urejenih glede na pogostost pojavitev v korpusu, zaslon *O korpusu* nam prikaže osnovne statistike in informacije o korpusu, *My jobs* nam ponuja pregled nad procesi, ki se izvajajo v ozadju, *User guide* pa nas preusmeri na navodila za uporabo orodja.



Išči	vir	jezik	zvrst	pojavnic	besed
metaFida v0.1 (združeni korpus)	Info	Slovenian	C/A	4.463.244.641	3.615.602.683
Gigafida v2.0 proto (referenčni, nededupliciran)	Info	Slovenian	C/A	1.811.841.991	1.475.404.460
⊃ Gigafida v2.0 (referenčni, dedupliciran)	Info	Slovenian	C/A	1.333.360.653	1.105.200.611
Gigafida v1.1 (referenčni)	Info	Slovenian	C/A	1.409.753.568	1.138.791.070
⊃ Gigafida v1.1 DeDup (referenčni, dedupliciran)	Info	Slovenian	C/A	1.111.927.975	918.689.740
⊃ Gigafida v1.0 (referenčni)	Info	Slovenian	C/A	1.409.753.568	1.138.791.070
⊃ FidaPLUS (stari referenčni)	Info	Slovenian	C/A	738.503.145	596.395.664
SUK 1.0	Info	Slovenian	V/R	1.025.639	862.618
⊃ ssj500k v2.3	Info	Slovenian	V/R	586.248	490.755
⊃ jos1M v1.2	Info	Slovenian	V/R	1.182.814	971.663
Trendi (spremljevalni)	Info	Slovenian	C/A	641.951.202	535.909.081
slWaC (Slovene Web)	Info	Slovenian	C/A	895.903.321	749.372.269
Gos 2.0 (referenčni, govorni)	Info	Slovenian	C/A	2.496.110	2.436.386
⊃ Gos 1.1.1 (referenčni, govorni)	Info	Slovenian	C/A	1.063.861	1.033.697
IMP (starejša besedila)	Info	Slovenian	C/A	17.723.874	14.348.452
⊃ goo300k (starejša besedila, ročno označena)	Info	Slovenian	V/R	358.036	286.091
OSS (znanstvena dela)	Info	Slovenian	C/A	3.261.941.663	2.342.855.598

Slika 11: Začetna stran z naborom korpusov

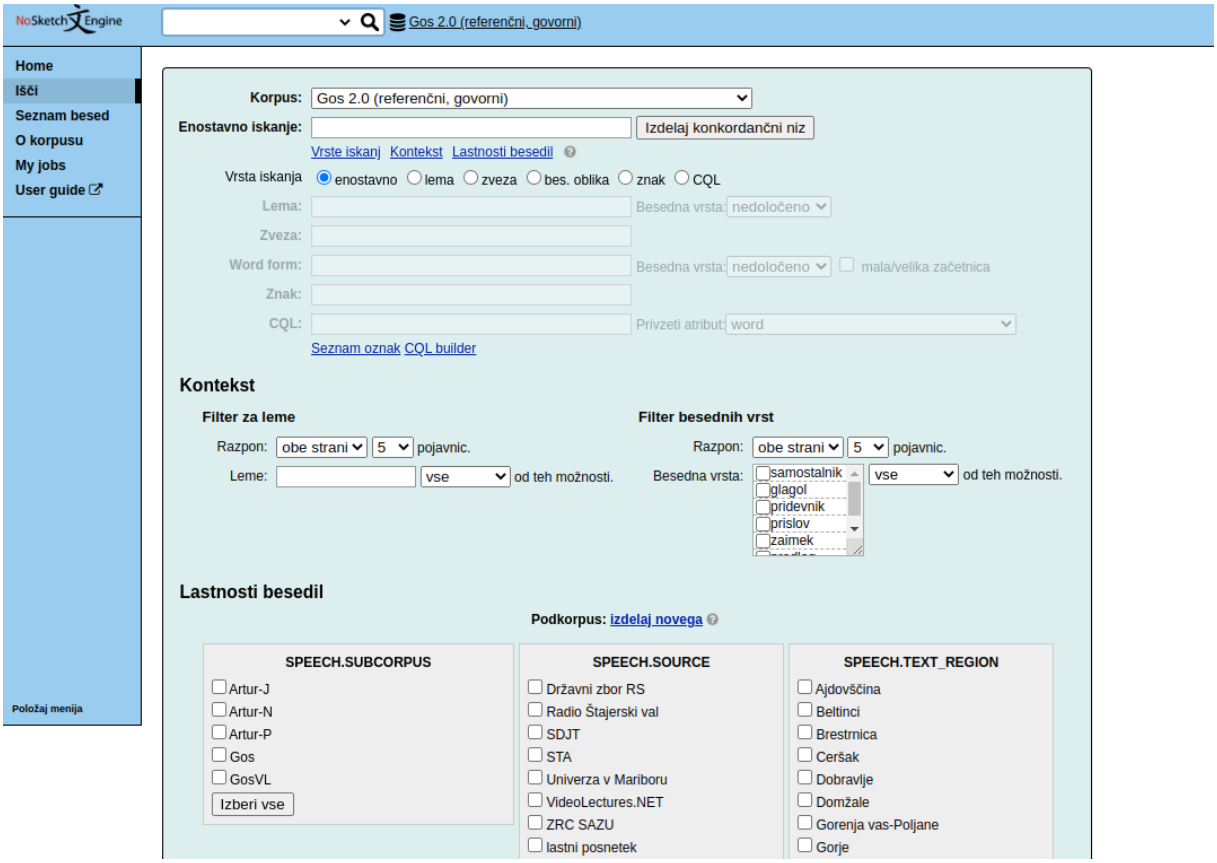
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 12: Zaslon s pregledom osnovnih značilnosti in statistik korpusa

S klikom na *Išči* se nam izriše zaslon z različnimi možnostmi iskanja (Slika 13). Sprva je prikazan le način *Enostavno iskanje*, ki omogoča enostavno iskanje po iskalnem nizu. S klikom na *Vrste iskanj, Kontekst* in *Lastnosti besedil* pa se nam prikaže še več opcij za bolj podrobno iskanje po bazi. *Vrste iskanj* nam prikaže več različnih nastavitev: iskanje po lemah, besednih zvezah, specifičnih oblikah neke besede, zaporedju znakov in iskanje s pomočjo iskalnega jezika CQL (angl. *Corpus Query Language*), ki podpira tudi regularne izraze. S klikom na *Kontekst* se nam prikažejo možnosti za prilagajanje prikazanega konteksta okrog iskane besede. *Lastnosti besedil* pa nam omogoča omejevanje rezultatov glede na tip besedila, spol, starost, regijo govorca in ostale metapodatke.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 13: Možnosti iskanja po korpusu

Konkordančnik nam rezultate iskanja prikaže v obliki ključnih besed v kontekstu (angl. *KeyWords In Context - KWIC*) (Slika 14). Ob vsakem zadetku nam na levi strani prikaže še metapodatke za poved, ki vsebuje iskano besedo, ob sami besedi pa oznake, ki ji pripadajo. Na vrhu strani, nad prikazom rezultatov, se izpiše še število zadetkov in pogostost na milijon pojavnic. S klikom na metapodatke, izpisane na levi strani v modrem, se nam na dnu strani odpre rumeno okence, ki nam prikaže še več metapodatkov, ki se nanašajo na poved. Ob kliku na povezavo ob podatku *s.speech* nam bo vmesnik predvajal tudi zvočni posnetek izbrane povedi (Slika 15).

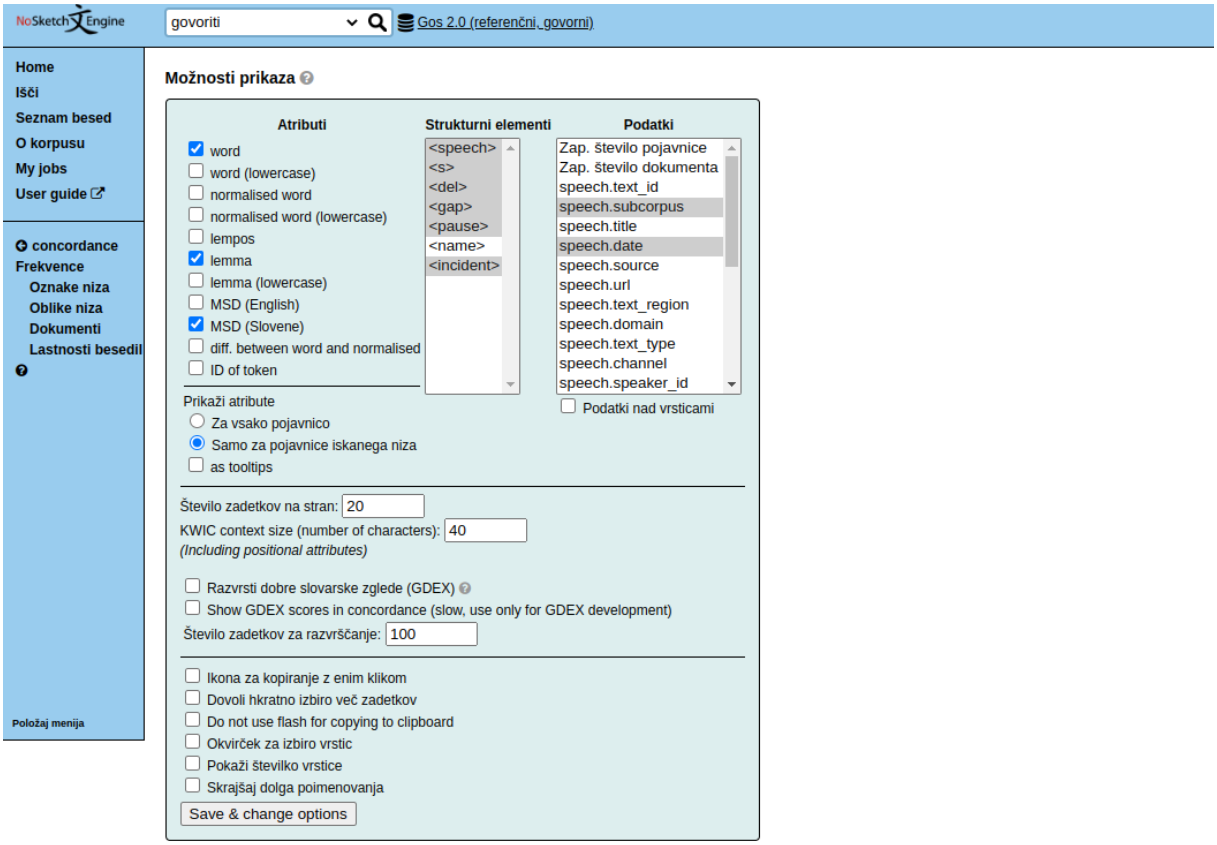
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Slika 14: Rezultat iskanja

Slika 15: Rumeno okno z metapodatki

V modrem okencu na levi se nam pod črto prikažejo še dodatne opcije za prilagajanje prikaza rezultatov. Posebno pomemben je zaslon *Možnosti prikaza* (Slika 16), ki nam omogoča prilagajanje prikazanih metapodatkov, prikazanih korpusnih oznak za vsako besedo v kontekstu, število vrstic na stran in še veliko drugih uporabnih nastavitev.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 16: Prilagajanje prikaza rezultatov

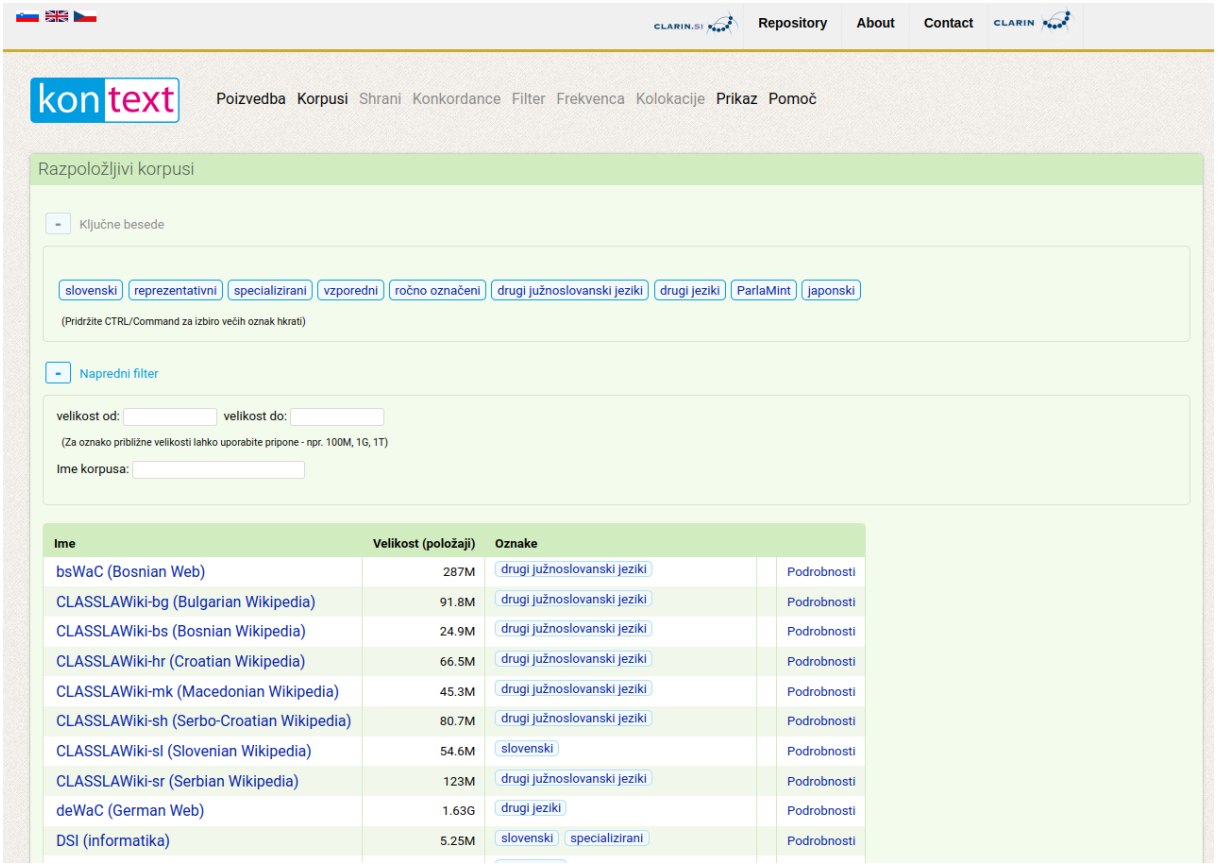
8.2.2 Konkordančnik KonText

Konkordančnik KonText (Machálek 2014, Machálek 2020) je dostopen na spletnem naslovu https://www.clarin.si/kontext. Po naboru funkcij in možnosti iskanja je zelo podoben orodju noSketchEngine.

Na začetnem zaslonu spet najdemo seznam korpusov, ki so na voljo za uporabo (Slika 17). Do korpusa Gos 2.0 pridemo z nekajkratnim klikom na gumb *Naloži več.* Ob izbiri korpusa nas vmesnik pripelje do zaslona za vnašanje iskalnega ukaza (Slika 18). Tu lahko preklapljamo med osnovnim načinom iskanja in pa naprednejšim, do katerega pridemo s klikom na napis *Napredno iskanje*. Osnovni način je podoben osnovnemu iskanju v orodju noSketchEngine, s tem da je tu podprto tudi iskanje s pomočjo

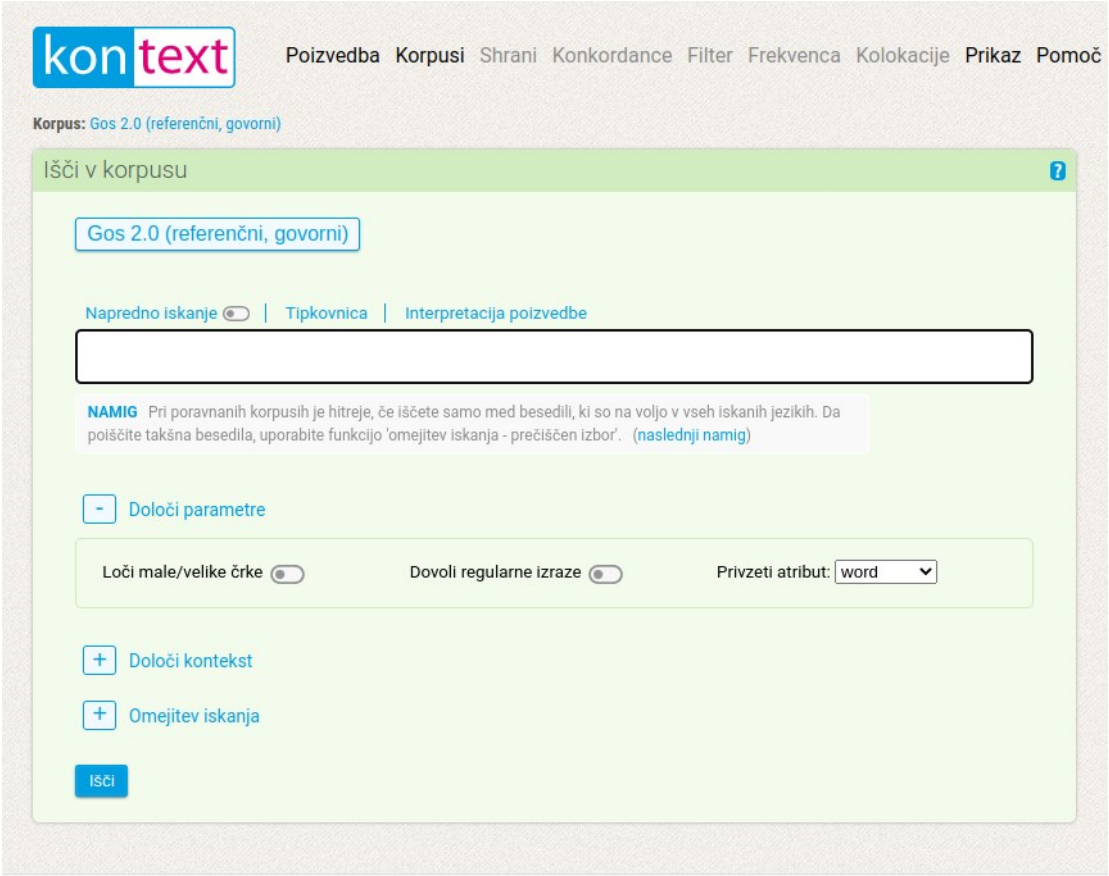
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

regularnih izrazov, naprednejši način pa ustreza iskanju s pomočjo jezika CQL. Poleg tega imamo na tem zaslonu še nabore nastavitev *Določi kontekst,* s katerim lahko urejamo prikazani kontekst, *Omejitev iskanja*, s katerim lahko omejimo prikazani rezultat le na tiste zadetke, ki vsebujejo določene metapodatke in pa *Določi parametre*, s katerim lahko določamo, ali bo iskanje podpiralo regularne izraze in ali bo razlikovalo med veliko in malo začetnico.



Slika 17: Začetni zaslon

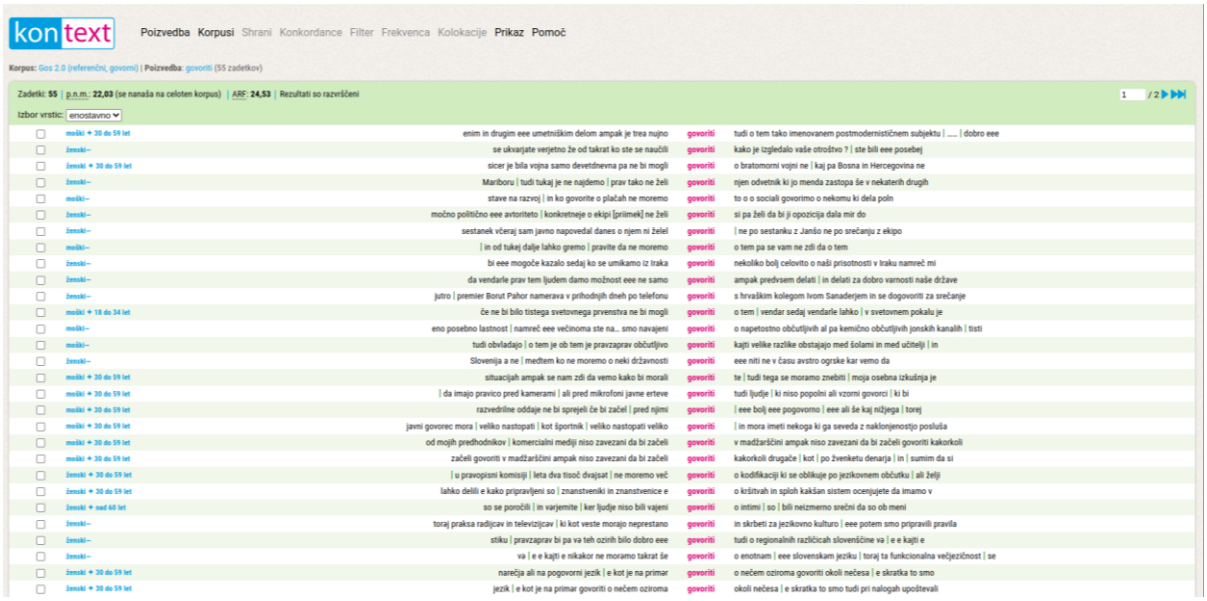
Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



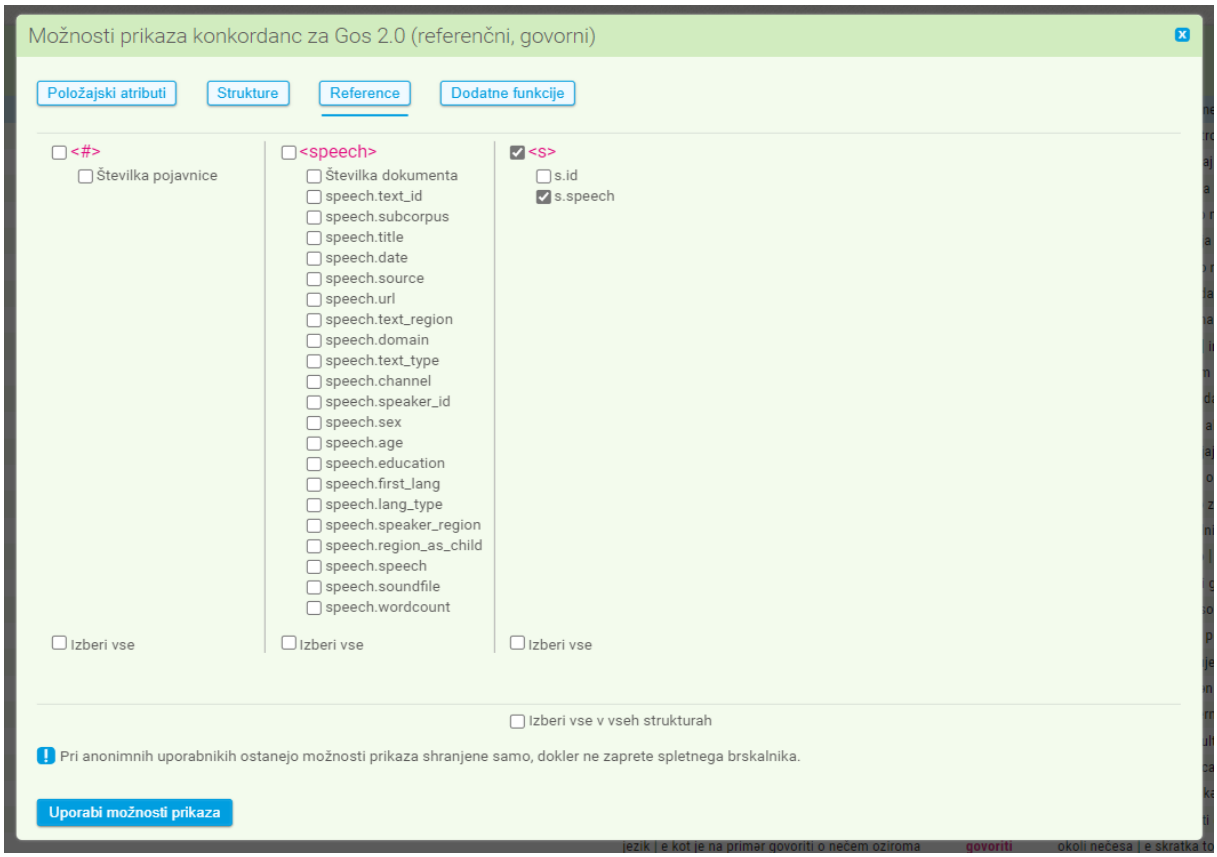
Slika 18: Zaslon z iskalnim poljem

Prikaz rezultatov v konkordančniku KonText je podoben kot pri vmesniku noSketchEngine in omogoča podobne možnosti prilagajanja prikaza (Slika 19). Do nastavitev prikaza rezultatov pridemo s klikom na gumb *Prikaz* na vrhu strani. Če želimo dostopati do zvočnih posnetkov vsake povedi, moramo najprej spremeniti nastavitve s klikom na *Prikaz > Specifične nastavitve za korpus > Reference,* nato pa obkljukamo opcijo *s.speech* v stolpcu *<s>* (Slika 20). Do posnetka lahko nato dostopamo s klikom na prikazane metapodatke levo od konteksta in klikom na povezavo ob podatku *s.speech* v okencu na dnu strani (Slika 21). Konkordančnik KonText je podrobneje opisan v poglavju 8.3.3.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

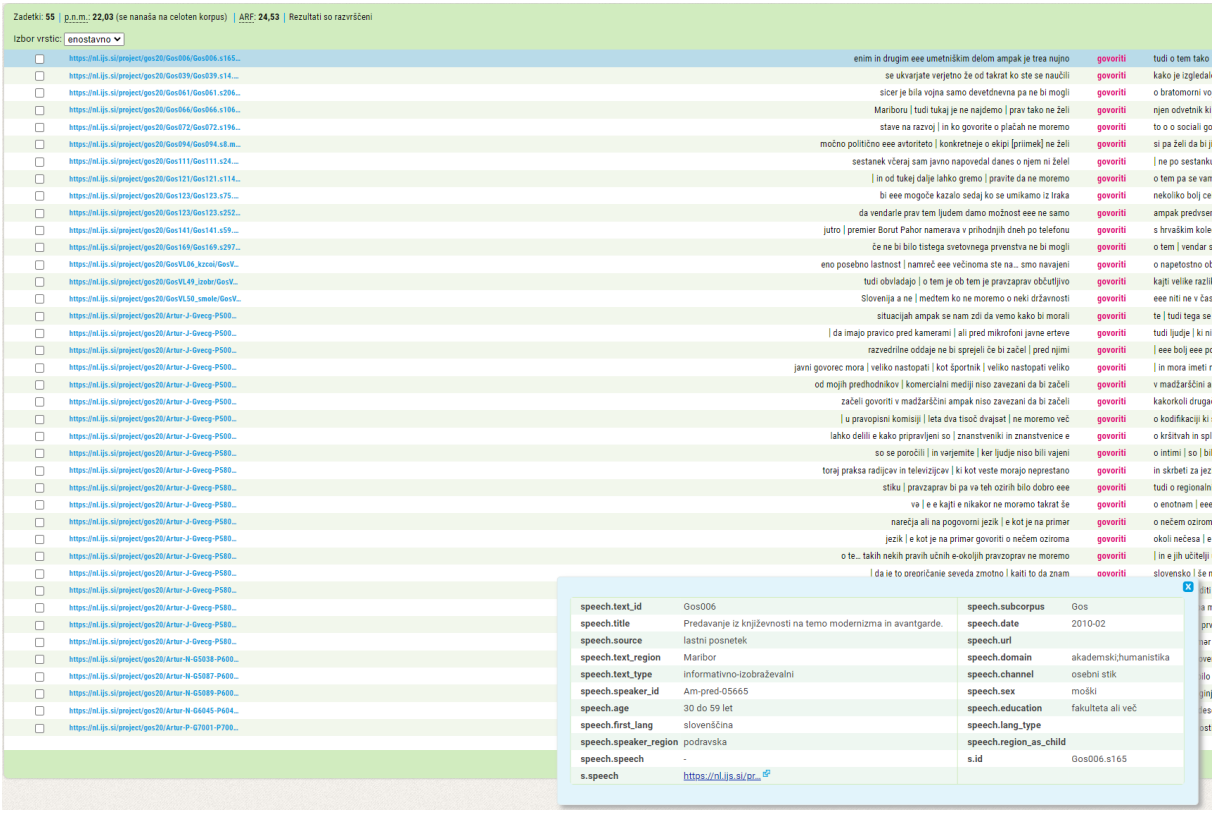


Slika 19: Prikaz rezultatov iskanja



Slika 20: *Prikaz > Specifične nastavitve za korpus > Reference*

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.



Slika 21: Izris okenca z metapodatki na dnu strani

8.3 Pregled sorodnih konkordančnikov in smernice nadaljnjega razvoja

V okviru projekta RSDO smo pripravili tudi pregled podobnih konkordančnikov za iskanje po govornih korpusih v drugih jezikih. Tovrsten pregled sorodnih virov bo v prihodnosti služil kot popis dobrih praks in vodilo za nadaljnji razvoj trenutnega vmesnika za Gos 2.0. V sledečih podpoglavjih je predstavljenih osem različnih spletnih orodij za iskanje po govornih korpusih, nato pa sledijo priporočila za nadaljnji razvoj, ki temeljijo na najboljših praksah in najbolj uporabnih funkcijah konkordančnikov. Pregled orodij smo izvedli v februarju 2023.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

8.3.1 BNClab

BNClab[23] je spletno orodje za iskanje po korpusih British National Corpus in Spoken BNC2014, ki v ospredje postavlja predvsem sociolingvistično komponento korpusne analize (Brezina et al. 2018). Orodje omogoča vizualni prikaz statističnih podatkov o iskanih besedah.

Glavna prednost orodja BNClab so vizualni prikazi, ki jih vmesnik sam generira na podlagi rezultatov iskalnega ukaza (ki podpira tudi regularne izraze). Izbiramo lahko med 7 načini prikaza rezultatov iskanja: *Usage, Change, Gender, Age, Social Class, Region* in *Summary*. Prikaz *Usage* sproducira običajno shemo primerov uporabe s kontekstom (KWIC). Tu lahko omejujemo rezultate glede na levi in desni kontekst in obliko iskane besede, omogočeno pa je tudi brskanje glede na atribute govorca (ime, spol, starost, družbeni sloj). Ostali prikazi (*Change, Gender, Age, Social Class* in *Region*) nam izrišejo grafikone, ki prikazujejo pogostost besede glede na časovno obdobje, spol, starost, družbeni sloj in regijo govorca. Prikaz *Summary* nam izpiše rezultate numerične analize glede na vse statistične dejavnike. Orodje BNClab sicer ne omogoča predvajanja zvočnih posnetkov govorjenega dela korpusa.

Zaradi grafičnih prikazov (grafikoni, zemljevidi, histogrami) je to orodje zelo uporabno tako za strokovnjake kot tudi za nestrokovnjake. Meniji so urejeni po zelo intuitivnem načelu, ki bazira na vizualnem prikazu funkcij in gumbov, zato je uporabniku lahko takoj jasno kako priti do iskanih informacij. Orodje je sicer zelo usmerjeno v potrebe uporabnikov, ki jih zanima sociolingvistični aspekt korpusne analize, zato ne vsebuje številnih funkcij, ki so pogosto prisotne pri drugih konkordančnikih.

8.3.2 CQPWeb

CQPWeb[24] (Hardie 2012) je platforma za več govorjenih korpusov: Spoken BNC2014, Longman Spoken American Corpus idr.
Ne podpira možnosti predvajanja zvočnih posnetkov, vsebuje samo transkripcije govora.

Konkordančnik ima šest načinov iskanja: *Standard query, Restricted query, word lookup, Frequency lists, Keywords* in *Analyse corpus*. *Standard query* ima še tri različne načine vnašanja oznak: *CQP syntax,*

[23] http://corpora.lancs.ac.uk/bnclab/search
[24] https://cqpweb.lancs.ac.uk/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Simple query (ignore case) in *Simple query (case-sensitive)*. Izpis rezultatov prikaže število pojavitev, število besedil v korpusu, število besed v korpusu in pogostost na milijon pojavnic (angl. *parts per million*).

Spoken BNC2014 nudi nekaj več podatkov o govorcu, do katerih se pride s klikom na oznako govorca ob povedi. Med prikazanimi podatki so tudi število besed in število povedi, pripisanih temu govorcu.

Funkcij, ponujenih informacij in možnosti za omejevanje iskalnega pogoja je veliko, vendar meniji in prikazi niso najbolj intuitivno urejeni, zato vmesnik ni najbolj prijazen za neprofesionalnega uporabnika. Uporabljenih je tudi veliko strokovnih izrazov, ki se izven področja jezikoslovja redko pojavijo. Vizualna ureditev vmesnika tudi ni najbolj optimalna, saj so meniji razširjeni prek celotnega zaslona (od enega roba do drugega), posledično je tudi besedilo preveč raztegnjeno in težko berljivo. Uporabljena barvna kombinacija rdeče in modre deluje nekoliko neprijetno in posledično slabo vpliva na preglednost.

8.3.3 KonText

Konkordančnik KonText[25] (Machálek 2014; Machálek 2020) je platforma za številne govorjene korpuse.

Ima pet načinov iskanja: *Basic* (enostavno iskanje), *Phrase* (zaporedje besed, ne razlikuje med malimi in velikimi črkami, podpira regex), *Word form* (samo ena oblika besede, podpira regex), *Word part* (zaporedje znakov v besedi, podpira regex), *CQL* (iskanje s pomočjo jezika Corpus Query Language).

Iskanje po kontekstu je možno na podoben način kot s prvo različico konkordančnika Gos. Območje iskanja se lahko omeji glede na specifične podkorpuse, dokumente, lastnosti govorca itd.

Pri prikazu rezultatov se poleg števila zadetkov izpišeta tudi pogostost na milijon pojavnic in povprečna reducirana frekvenca. Možnosti predvajanja zvočnega posnetka so podobne kot pri prvi različici Gos konkordančnika, vendar je gumb za predvajanje vstavljen v tekst na mesto, kjer je v govoru narejen premor. Uporabnik lahko tudi naredi poljubno selekcijo zadetkov in jih iz prikaza izbriše ali pa obdrži le izbrane in ostale skrije. Uporabnik lahko tudi do neke mere prilagaja prikaz rezultatov (npr. prikaz po

[25] https://lindat.mff.cuni.cz/services/kontext/corpora/corplist

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

povedih/KWIC, št. zadetkov na stran itd.). Poleg teh funkcij so uporabniku na voljo tudi druge, vendar le, če se vpiše z uporabniškim računom. Te vključujejo ustvarjanje podkorpusov, shranjevanje rezultatov, prikaz kolokacij, dodajanje filtrov, druge informacije o frekvenci itd.

Konkordančnik je zelo priročen za uporabnike, ki so že seznanjeni z jezikovnimi tehnologijami in strokovnimi termini, ki se na tem področju uporabljajo. Za uporabnike, ki pa se s tem še niso srečevali, utegne biti konkordančnik nekoliko nedostopen.

8.3.4 Sketch Engine

Sketch Engine[26] je platforma, ki ponuja širok nabor orodij za korpusno analizo (Kilgarriff et al. 2014). Ponuja dostop do številnih korpusov v različnih jezikih, med katerimi je tudi družina korpusov TenTen. Ni posebej adaptiran za govorne korpuse, zato ne ponuja možnosti predvajanja zvočnih posnetkov oz. lahko uporabniki do teh dostopajo samo preko povezav na zunanje vire.

Funkcije Sketch Engina so podobne kot pri odprtokodnem konkordančniku noSketch engine (Richly 2007, Kilgarriff et al. 2014), s tem da Sketch Engine ponuja nekaj novih načinov iskanja po jezikovnih bazah, kot so *Word Sketch* (prikaže kolokacije oz. besede, ki se pogosto pojavijo v kombinaciji z iskano), *Thesaurus* (neke vrste tezaver, iskalnik besed, ki so pomensko blizu), *N-grams* (iskanje po n-gramih), *Parallel Concordance* (iskanje po prevedenih ustreznicah v paralelnih korpusih) in druge. Funkcij, filtrov, možnosti iskanja je zelo veliko, vendar Sketch Engine vedno najprej prikaže le osnovne načine iskanja, naprednejši pa so skriti za spustnimi seznami oz. za gumbom *advanced*. Na ta način se uporabnika ne zasuje z informacijami.

Sketch Engine je enostaven za uporabo, saj nudi osnovne načine iskanja po korpusih, hkrati pa ne zanemari naprednejših funkcij, ki so sprva sicer skrite, vendar vedno dostopne. Na ta način se omogoči dostopnost osnovnim uporabnikom in hkrati ohrani možnost bolj poglobljenega dela za naprednejše uporabnike.

[26] https://www.sketchengine.eu/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

8.3.5 Glossa

Orodje Glossa[27] so razvili na Univerzi v Oslu in omogoča brskanje po korpusih v številnih svetovnih jezikih (Nøklestad et al. 2017, Kosek et al. 2015). Med drugimi vključuje tudi korpus Big Brother v norveščini, ki vsebuje zvočne posnetke, video posnetke in transkripcije pogovorov iz norveške televizijske oddaje Big Brother.

Korpus je dobro prilagojen govornim korpusom, saj omogoča predvajanje zvočnih posnetkov, video posnetkov (če so na voljo) in način *waveform* (prikaz oblike zvočnega valovanja) za pogovore pridobljene iz televizijske serije. Po transkripcijah lahko iščemo s pomočjo treh iskalnih načinov: *Simple*, kjer lahko iščemo po poljubnem številu besednih nizov, *Extended*, kjer lahko podamo eno ali več iskanih besed in iskanje omejimo glede na besedno vrsto, mesto pojavitve v posnetku, način izgovorjave (nejasno, smeje, šepetaje…) ipd. in *CQP Query*, kjer lahko iščemo po sistemu CQP. Glossa omogoča tudi razvrščanje podatkov glede na sociolingvistične dejavnike (spol, ime, regija govorca itd.).

Vmesnik je dokaj intuitivno urejen, zato je precej enostaven za uporabo. Možnosti iskanja je veliko, precej poudarka je namenjenega zvočnim posnetkom, zato je vmesnik dobro prilagojen uporabnikom, ki jih ne zanimajo le transkripcije govora, ampak tudi zvočna in video vsebina. Barvna shema je nekoliko pusta.

8.3.6 Korp

Konkordančnik Korp[28] (Borin et al. 2012) omogoča brskanje po raznih korpusih nordijskih jezikov (švedščina, finščina, laponski jeziki idr.). Po načinu uporabe je precej podoben orodju Glossa, vendar ni posebej prilagojen govornim korpusom, zato vsebuje le pisana besedila in transkripcije, ne podpira pa predvajanja zvočnih posnetkov.

Korp omogoča tri iskalne načine: *Simple, Extended* in *CQP query* (podobno kot pri Glossi), s tem da vključuje tudi možnost prikaza *word picture*, ki prikaže vizualni prikaz najpogostejših kolokatorjev v različnih skladenjskih vlogah. Ponuja nam iskanje in prikazovanje rezultatov glede na številne

[27] https://tekstlab.uio.no/glossa2/
[28] https://spraakbanken.gu.se/korp/#?lang=en

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

jezikoslovne lastnosti, kot so besedne vrste, leme, semantični pomen besede, vloga v odvisnostni drevesnici in druge.

Vmesnik orodja Korp je dokaj zapleten. Ureditev menijev je nekoliko neintuitivna, saj so vsi gumbi zbrani na enem delu zaslona, drugi del pa ni zapolnjen. Poleg tega nam vmesnik že takoj ponudi veliko informacij, kar utegne biti nekoliko moteče. Orodje je v prvi vrsti prilagojeno le potrebam strokovnjakov, saj nudi veliko uporabnih možnosti za jezikoslovno analizo, a je zaradi tega nekoliko manj dostopno osnovnemu uporabniku.

8.3.7 Voice

Spletni konkordančnik za korpus VOICE[29] – Vienna-Oxford International Corpus of English – omogoča iskanje po transkripcijah in poslušanje zvočnih posnetkov pogovorov med L2 govorci angleščine (The Vienna-Oxford International Corpus of English 2021). Projekt je usmerjen v analizo govorjene angleščine v ELF kontekstih (angleščina kot lingua franca - angl. *English as a Lingua Franca*).

Zaslon je razdeljen na tri okna, v prvem se lahko izbira med različnimi dokumenti, dodaja razne filtre iskalnemu ukazu in dodaja zaznamke. Drugo okno vsebuje iskalno vrstico in prikaže rezultate iskanja, v tretjem oknu pa lahko pregledujemo celotni dokument, v katerem se izbrani rezultat nahaja, in predvajamo zvočni posnetek. Drugo in tretje okno nam ponujata štiri različne tipe prikazovanja podatkov: prikaz polnih transkripcij posnetka, prikaz samega besedila brez dodatnih oznak, prikaz s POS oznakami in prikaz XML dokumenta z rezultatom.

Vmesnik je precej zapleten in namenjen strokovnim uporabnikom, na kar nakazuje prisotnost opcij, kot je npr. prikaz dokumenta v XML obliki. Sama razporeditev informacij in funkcij v tri različna okna je nekoliko neintuitivna, zato lahko traja kar nekaj časa preden se postavitve privadimo.

8.3.8 Scottish Corpus of Texts & Speech

Korpus SCOTS[30] (Scottish Corpus of Texts & Speech) sestavljajo pisna in govorjena besedila, ki so nastala na Škotskem – ne le v angleščini, ampak tudi v drugih jezikih, na primer v jeziku Scots

[29] https://voice3.acdh.oeaw.ac.at/
[30] https://www.scottishcorpus.ac.uk/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

(germanski jezik, ki se govori na Škotskem) in škotski gelščini (Anderson et al. 2007). Vsebuje tekste, ki so izvorno pisni, transkripcije govorjenih pogovorov in besedil, omogoča nam pa tudi ogled videoposnetkov in poslušanje zvočnih posnetkov, kjer so ti na voljo.

Za iskanje po bazi imamo na voljo štiri glavne načine: *search, advanced search, browse* in *tools*. Način *search* nam ponuja osnovno iskanje glede na iskani niz, ki je lahko ena sama beseda ali pa zaporedje večih. Poleg iskanega niza lahko vnesemo še druge pogoje iskanja, kot so npr. spol, priimek, rojstna regija govorca, ime, leto in tip dokumenta ter možnost predvajanja zvočnega posnetka dokumenta. Rezultate nam iskalnik prikaže v obliki konkordance z dodatnimi podatki o posameznem dokumentu, frekvenci besede in normalizirani frekvenci. Ko izberemo enega od rezultatov, se nam izkaže celoten dokument, v katerem se rezultat pojavi. Tu lahko predvajamo tudi zvočni posnetek besedila, če je ta na voljo. Poleg osnovnega načina *search* imamo možnost iskati tudi z načinom *advanced search*, kjer lahko dodajamo bolj podrobne iskalne pogoje in omejujemo nabor dokumentov, po katerih hočemo iskati. Način *advanced search* nam ponuja tudi prikaz zemljevida britanskega otočja, ki nam grafično ponazori razporeditev tekstov glede na regijo nastanka. Način *browse* nam omogoča brskanje po vseh tekstih, ki so vključeni v bazo, način *tools* pa nudi razna dodatna orodja za iskanje po korpusu (iskanje po kolokacijah ipd.). Iskanje glede na slovnične lastnosti besed ni podprto, niti v naprednejšem načinu iskanja.

Sama ureditev spletne strani je dokaj intuitivna in pregledna, vendar pa so nekatere iskalne opcije skrite, njihova uporaba pa ni zelo natančno pojasnjena. Uporabnika še posebej lahko zmede način *advanced search*, ki je nekoliko manj pregleden kot ostali. Zelo dobrodošel je prikaz zemljevida, ki pritegne pozornost uporabnika in nekoliko popestri iskanje. Barvna shema ni preveč moteča.

8.3.9 Priporočila za prihodnjo izboljšavo vmesnika

Ker je Gos namenjen zelo širokemu krogu uporabnikov (učitelji, tolmači, lektorji govora, raziskovalci), je ključno vključevanje elementov, ki bodo zanimivi in dostopni za vse, hkrati pa bodo omogočali naprednejšim uporabnikom čim več naprednejših možnosti iskanja po podatkih.

Zelo primerna za korpus Gos bi bila možnost izrisa vizualizacije rezultatov v obliki grafikonov, zemljevidov ipd., kot to ponuja orodje BNClab. Korpus Gos je namreč v primerjavi z drugimi

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

referenčnimi korpusi opremljen s številnimi kategorijami metapodatkov, ki so kot nalašč za tovrstne sociolingvistične analize. Pri BNClabu je še posebej zanimiv prikaz *Region*, ki izriše zemljevid Britanskega otočja in prikaže, kje se iskani izraz pogosto uporablja. Podobno bi se lahko storilo za korpus Gos, saj vsak posnetek vključuje informacijo o regiji govorca in regiji snemanja. Delno se uporabniku prijaznejši vizualizaciji podatkov korpusa Gos že posveča projekt SLOKIT[31], v okviru katerega bo razvit specializirani spletni portal za prikazovanje najzanimivejših podatkov različnih referenčnih korpusov slovenskega jezika, vendar bi veljalo tovrstne infografike vključiti tudi v temeljni konkordančnik korpusa Gos.

Večina ostalih konkordančnikov nudi možnost iskanja s pomočjo širšega nabora regularnih izrazov. Trenutno iskanje po Gosu podpira le znaka * in ?. Za naprednejše uporabnike bi bilo morda koristno vključiti še nekoliko večji nabor regularnih izrazov.

Ker korpus Gos vsebuje tudi dodatek Gos VideoLectures, bi se lahko izkoristilo tudi možnost predvajanja videoposnetka govorjene vsebine. Pri tem se je vredno zgledovati po konkordančniku Glossa, ki za vsak rezultat iskanja omogoča tudi predvajanje videoposnetka, če je ta na voljo za izbrani korpus. Kot naprednejša možnost dela z zvočnimi posnetki je za korpus Gos zanimiv tudi način *waveform*, ki ga ponuja vmesnik Glossa. Ta je koristen za naprednejše fonetične raziskave.

9 Darinka Verdonik: Prioritete za nadaljnji razvoj

Spoznanja zbiranja in urejanja gradiv v bazi ARTUR, katere del je vključen tudi v korpus Gos 2.0, so sledeča:

- Razlike v standardih beleženja metapodatkov o govorih in posnetkih ter v načinih transkribiranja pogosto rezultirajo v manj natančne informacije ob združevanju virov in izgubo podatkov. Določene razlike pa je treba vedno znova pričakovati zaradi razlik v potrebah različnih ved po govornih podatkih.

[31] https://slokit.ijs.si/

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- V prihodnje je mogoče sčasoma pričakovati prehod na novo orodje za transkribiranje, saj je Transcriber 1.5.1 precej staro in po funkcionalnih zelo omejeno orodje. Kot zanimiva alternativa se kaže Transana.
- Snemanje na terenu je finančno zahteven zalogaj. Pokazalo se je, da je v velikem obsegu pravočasno izvedbo lažje zagotoviti z najemom zunanjega izvajalca, čeprav je to cenovno manj ugodno od študentskega dela.
- Za ročno narejene transkripcije ustrezne kvalitete je potreben resen zunanji izvajalec in veliko ur redno zaposlenih oseb v projektu, ki kontrolirajo narejene transkripcije, ali pa se transkribiranje v celoti izvede z zaposlenimi osebami. Sodelovanje študentov pri transkribiranju je izredno težko koordinirati in rezultat so transkripcije slabše kvalitete.
- Zahteve po velikem obsegu gradiv v kratkem časovnem obdobju vodijo v iskanje neoptimalnih načinov zbiranja posnetkov, kot je branje pisnih povedi. Brani govor, posnet izključno v namene govornega vira, lahko zadosti določenim potrebam po fonetično-fonemskih raziskavah ali pri razvoju na akustiki temelječih strojnih modelov, vendar v takem viru manjkajo jezikovno-kontekstne dimenzije, ki so značilne za govorjeno rabo. Izdelava vsebinsko optimalnih govornih virov v ustreznem obsegu je dolgotrajen projekt, ki zahteva stabilno financiranje, in ne more biti izvedena v kratkem času.

Strategija razvoja v nadaljevanju temelji na projektnih izkušnjah z izdelavo govornih virov ARTUR in Gos 2.0 kot tudi predhodnih, v repozitoriju CLARIN.SI deponiranih govornih virov, na pregledu in poznavanju stanja vzorčnih govornih virov ter govornih virov za primerljive jezike. Za vzorčne govorne vire štejemo predvsem govorno komponento korpusa BNC, korpus romanskih jezikov C-ORAL-ROM in češke govorne vire, za primerljive jezike pa štejemo jezike z do 10 mio. govorcev iz držav Evropske unije, ki so se uniji pridružile po letu 2004. To so hrvaški, estonski, irski, latvijski, litovski in slovaški jezik. Primerjamo obseg virov in zajete tipe govora. Na podlagi tega priporočamo sledeče usmeritve v nadaljnji strategiji razvoja:

- Naslednji realni mejnik **obsega** govornega korpusa slovenščine je 5 mio. besed. Po trenutnih standardih idealni ciljni obseg bi bil 10 mio. besed. Pričakovati je, da se bo s časom ta mejnik višal.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- Pomemben segment med **tipi govora** v referenčnem govornem korpusu mora ostati neformalni vsakdanji pogovor, ki pokriva vse demografske in regionalne segmente govorcev ter je posnet v avtentičnih situacijah na terenu. Izkušnje namreč kažejo, da je v tem segmentu zelo veliko raziskovalnega zanimanja in potenciala tako za razvoj posameznih ved kot aplikacij z vključenimi govornimi tehnologijami. Snemanje tega segmenta se najlažje izvaja prek široke mreže terenskih snemalcev in ga ni mogoče izvesti v večjem obsegu v kratkem časovnem obdobju. V tem segmentu se lahko ustrezno naslovijo tudi potrebe po dialektoloških korpusih. Določene tipe govora je zaradi varovanja osebnih in/ali zaupnih podatkov izredno težko zajeti, to je predvsem t. i. nejavni nezasebni govor, kamor sodijo različne vsakdanje komunikacijske situacije v uradih, prodaji, storitvah, informacijah, klicnih centrih, zdravstvu, sodstvu, na delovnem mestu ipd., zato so v obstoječih virih slabo zastopani, kar negativno vpliva na reprezentativnost virov. Pozornost je treba nameniti tudi možnostim za zajem tovrstnih tipov govora.
- Na področju **transkribiranja** bi bilo smiselno raziskati, kako se lahko najnovejša tehnologija strojnega razpoznavanja govora najučinkoviteje uporabi za izdelavo transkripcij ter ali lahko pohitri proces izdelave ročnih transkripcij. V bazi ARTUR je ostalo pribl. 200 ur netranskribiranih posnetkov in prvi naslednji smiselni korak bi bila zagotovitev transkripcij za teh 200 ur.
- Pri **načinih zbiranja posnetkov** bi bilo smiselno raziskati, ali lahko množičenje in platforme za komuniciranje na daljavo koristno uporabimo za zajem posnetkov. Kot del strategij zbiranja ostajajo vsi obstoječi viri posnetkov, ki jih ne omejujejo avtorske pravice in varstvo osebnih podatkov, ter široka mreža terenskih snemalcev.
- Zbiranje posnetkov močno omejuje **pravne omejitve**. Z novostmi na tem področju je treba raziskati, kako bi se lahko v prihodnje dosegala večja zastopanost gradiv iz različnih medijskih hiš in drugih virov javnega govora. Posebno pozornost bi veljalo posvetiti gradivom pri nacionalni radioteleviziji, za katera so delno že zagotovljene transkripcije za potrebe podnaslavljanja, četudi so te transkrpicije prilagojene in bi jih bilo treba za potrebe govornih virov dodatno urejati.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

- V čim večjem delu podatkov bi bil koristna nadgradnja z **video posnetki**, saj se samo z avdio posnetki določen del kontekstualnih in pomenskih informacij izgubi. Prek neverbalnega telesnega izražanja se najbolj tipično izražajo odnos do teme in sogovornikov, emocije, metadiskurzni procesi tvorjenja, odzivanja in upravljanja poteka pogovora, družabni stiki kot tudi deiktični signali in simbolne geste ipd.
- Na področju **označevanja** govornih podatkov je treba prilagoditi orodja za oblikoslovno in skladenjsko označevanje ter semantično analizo na značilnosti govorjenega jezika. Velik raziskovalni in razvojni potencial je na nivojih označevanja, ki so specifični za govorjeno modalnost: prozodične lastnosti govora, govorica telesa (mimika obraza, gibanje glave, oči, rok, telesa...), razkrivanje procesov tvorjenja (npr. skozi t. i. netekočnosti), specifike govorjene rabe, ki izhajajo iz visoke interaktivnosti in neposrednega medosebnega stika (npr. diskurzni označevalci, dialoška dejanja) ipd.

10 Priloge

Priloga 1: Dovoljenje za snemanje in uporabo posnetka glasu in privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov ter dovoljenje za uporabo avtorskih pravic

11 Literatura

Anderson et al. 2007 = Anderson, J., David Beavan in Christian J. Kay. "SCOTS: Scottish Corpus of Texts and Speech." (2007).

Borin et al. 2012 = Lars Borin, Markus Forsberg in Johan Roxendal. 2012. Korp – the corpus infrastructure of Språkbanken. V: *Proceedings of LREC 2012*. Istanbul: ELRA, str. 474–478.

VOICE. 2021. The Vienna-Oxford International Corpus of English (version VOICE 3.0 Online). https://voice3.acdh.oeaw.ac.at (dostop: 23. 2. 2023).

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Brezina et al. 2018 = Brezina, V., Gablasova, D. in Reichelt, S. (2018). BNClab. http://corpora.lancs.ac.uk/bnclab [electronic resource], Lancaster University.

Hardie 2012 = Hardie, Andrew (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17(3), 380–409.

Kilgarriff et al. 2014 = Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, Vít Suchomel. The Sketch Engine: ten years on. Lexicography, 1: 7-36, 2014.

Nøklestad et al. 2017 = Nøklestad, Anders, Hagen, Kristin, Johannessen, Janne Bondi, Kosek, Michal and Joel Priestley. 2017. A modernised version of the Glossa corpus search system. V: Jörg Tiedemann (ed.): Proceedings of the 21st Nordic Conference on Computational Linguistics (NoDaLiDa). 2017, 251-254.

Kosek et al. 2015 = Kosek, Michal , Anders Nøklestad, Joel Priestley, Kristin Hagen, and Janne Bondi Johannessen. 2015. V: Gintarė Grigonytė, Simon Clematide, Andrius Utka and Martin Volk (eds.): *Visualisation in speech corpora: maps and waves in the Glossa system, Proceedings of the Workshop on Innovative Corpus Query and Visualization Tools at NODALIDA 2015*, May 11-13, 2015, Vilnius, Lithuania, NEALT Proceedings Series 25, 23–31.

Machálek 2014 = Tomáš Machálek (2014): KonText – Corpus Query Interface. FF UK, Praha. http://kontext.korpus.cz/.

Machálek 2020 = Tomáš Machálek (2020): KonText: Advanced and Flexible Corpus Query Interface. V: *Proceedings of LREC 2020*, str. 7005–7010.

Ocena učinka v zvezi z varstvom osebnih podatkov (DPIA) 2022 = *Ocena učinka v zvezi z varstvom osebnih podatkov (DPIA): Izdelava govorne baze za strojno učenje za rabo slovenščine v digitalnem okolju: Arnes Video,* marec 2022.

Rychlý 2007 = Rychlý, Pavel. Manatee/Bonito-A Modular Corpus Manager. V: RASLAN. 2007. str. 65-70.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Verdonik et al. 2018 = Verdonik, Darinka. Korpus in baza Gos Videolectures. V *Zbornik konference Jezikovne tehnologije in digitalna humanistika, 20. September—21. September 2018*. Ljubljana, Znanstvena založba Filozofske fakultete, 2018. http://nl.ijs.si/jtdh18/JTDH-2018-Proceedings.pdf.

Verdonik in Zwitter Vitez 2020 = Verdonik, Darinka, Zwitter Vitez Ana. *Slovenski govorni korpus Gos*. Ljubljana: Znanstvena založba Filozofske fakultete, 2020. http://www.dlib.si/stream/URN:NBN:SI:DOC-X9DAJU5X/0ac91ae2-56e1-459e-baf8-e28e0c566f08/PDF.

Verdonik et al. 2022a = Verdonik, Darinka, Bizjak, Andreja, Žgank, Andrej, Dobrišek, Simon. Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. V Zbornik konference Jezikovne tehnologije in digitalna humanistika 15.–16. september 2022. Ljubljana, Inštitut za novejšo zgodovino, 2022. https://nl.ijs.si/jtdh22/pdf/JTDH2022_Proceedings.pdf.

Verdonik et al. 2022b = Verdonik, Darinka, Bizjak, Andreja. *Navodila za zapisovanje in označevanje govora v govorni bazi Artur projekta RSDO.* Maribor, Fakulteta za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, 28. 11. 2022.

Verdonik et al. 2022c = Verdonik, Darinka, Bizjak, Andreja, Trojar, Mitja. *Navodila za standardizirani zapis v govorni bazi Artur projekta RSDO.* Maribor, 22. 11. 2022.

Verdonik et al. 2023a = Verdonik Darinka, Bizjak, Andreja, Dobrišek Simon. *Specifikacije posnetkov v govorni bazi Artur*, januar 2023.

Verdonik et al. 2023b = Verdonik Darinka, Bizjak, Andreja, Dobrišek Simon. *Opis govorne baze Artur projekta RSDO*, februar 2023.

Žganec Gros et al. 2021 = Žganec Gros, Jerneja, Vesnicer, Boštjan, Mihelič, Aleš, Trojar, Mitja, Dobrišek, Simon, Bizjak, Andreja, Verdonik, Darinka. *Izbor povedi za govorno bazo Artur v projektu Razvoj slovenščine v digitalnem okolju: Projektno poročilo DS2-2.1.1*, junij 2021.

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

Priloga 1: Dovoljenje za snemanje in uporabo posnetka glasu in privolitev v obdelavo osebnih podatkov z informacijami o obdelavi osebnih podatkov ter dovoljenje za uporabo avtorskih pravic

DOVOLJENJE ZA SNEMANJE IN UPORABO POSNETKA GLASU

IN

PRIVOLITEV V OBDELAVO OSEBNIH PODATKOV Z INFORMACIJAMI O OBDELAVI OSEBNIH PODATKOV

TER

DOVOLJENJE ZA UPORABO AVTORSKIH PRAVIC

(BRANI GOVOR)

ID govorca:	
Ime in priimek:	
Spol:	
Starost:	
Izobrazba:	
Kraj in občina stalnega bivališča:	
Kraj in občina bivanja v otroštvu:	
Prvi jezik:	

1. Spodaj podpisani izjavljam, da sem seznanjen, da Univerza v Ljubljani, Kongresni trg 12, 1000 Ljubljana, matična številka: 5085063000, davčna številka: 54162513 (v nadaljevanju nosilec projekta), izvaja projekt **Razvoj slovenščine v digitalnem okolju – RSDO** (v nadaljevanju projekt RSDO), ki je bil na javnem razpisu Razvoj slovenščine v digitalnem okolju – jezikovni viri in tehnologije (JR-ESRR-Razvoj slovenščine v digitalnem okolju), objavljenem v Uradnem listu RS št. 70/19 dne 29. 11. 2019, sprejet v sofinanciranje in katerega vsebina je razvidna s spletnih strani https://slovenscina.eu in https://www.gov.si/zbirke/javne-objave/javni-razpis-razvoj-slovenscine-v-digitalnem-okolju-jezikovni-viri-in-tehnologije/ in v okviru katerega bo nosilec projekta pripravil govorno bazo v obsegu 1000 ur, ki bo osnova za izdelavo splošnega avtomatskega razpoznavalnika govora slovenskega jezika in ki bo javno dostopna pod pogoji prostih licenc (npr. CC BY) in bo na voljo za nekomercialen in komercialen razvoj tehnologij (npr. govorno upravljanje naprav, pogovorni agenti, avtomatsko

podnaslavljanje video vsebin ali avtomatsko prevajanje govorjenih vsebin), za raziskave in za druge raziskovalne namene. Seznanjen sem, da bodo poleg omenjene baze javno dostopne pod pogoji prostih licenc tudi vsa programska koda in zbirke podatkov, ki bodo nastale med projektom, in vse aplikacije, ki bodo razvite s pomočjo te baze in ki bodo na voljo tudi na javnem portalu RSDO, kjer jih bodo lahko preizkusili in uporabljali posamezniki, raziskovalne in izobraževalne institucije, neprofitne organizacije, državni organi, organizacije z javnimi pooblastili in gospodarske družbe v Sloveniji in tujini.

2. Spodaj podpisani soglašam, da se me posname pri prebiranju vnaprej pripravljenih besedil, ki so bila izbrana s strani nosilca projekta, in da se takšen posnetek uporabi za zgoraj navedene namene ter da se takšen posnetek opremi z ustreznim identifikatorjem, ki bo povezan z osebnimi podatki, navedenimi v tej izjavi, ter z metapodatki o snemanju, ki se nanašajo na kraj snemanja, čas snemanja, snemalno opremo in okoliščine snemanja.

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

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

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

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

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.

projekta zahtevam prenos svojih osebnih podatkov; ali pri Informacijskemu pooblaščencu RS vložim pritožbo.

7. Spodaj podpisani sem seznanjen, da je upravljavec podatkov: Univerza v Ljubljani, Kongresni trg 12, 1000 Ljubljana, e-pošta: dpo@uni-lj.si (Oddelek za upravljanje s tveganji in varstvo osebnih podatkov), pooblaščena oseba za varstvo podatkov: Klemen Kraigher Mišič.

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

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

10. Spodaj podpisani ne jamčim in ne odgovarjam za kakršnekoli morebitne posege v avtorske ali druge pravice tretjih oseb, do katerih lahko pride zaradi uporabe vnaprej pripravljenih besedil iz točke 2 te izjave.

Datum in kraj: _____ Ime in priimek:_____

Podpis: _____

Operacijo Razvoj slovenščine v digitalnem okolju sofinancirata Republika Slovenija in Evropska unija iz Evropskega sklada za regionalni razvoj. Operacija se izvaja v okviru Operativnega programa za izvajanje evropske kohezijske politike v obdobju 2014-2020.