www.slovenščina.eu
sporazumevanje

Darinka Verdonik, Ana Zwitter Vitez
Slovenski govorni korpus Gos

Zbirka *Sporazumevanje*
Urednik zbirke *Simon Krek*

Recenzenta *Tomaž Erjavec, Hotimir Tivadar*
Urednici *Ana Zwitter Vitez, Darinka Verdonik*
Oblikovanje in prelom *Tomato Košir*
Avtor črkovne vrste »BadNews« *Samo Ačko*

Založila *Znanstvena založba Filozofske fakultete Univerze v Ljubljani*
Izdal *Center za jezikovne vire in tehnologije Univerze v Ljubljani*
Za založnika *Roman Kuhar, dekan Filozofske fakultete Univerze v Ljubljani*

Ljubljana, 2020

Prva e-izdaja.
Publikacija je v digitalni obliki prosto dostopna na
https://e-knjige.ff.uni-lj.si/
DOI: 10.4312/9789610603528

Izid knjige je podprla Javna agencija za knjigo

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
E-knjiga
COBISS.SI-ID=21000195
ISBN 978-961-06-0352-8 (pdf)

Darinka Verdonik,
Ana Zwitter Vitez
Slovenski
govorni korpus Gos

Kazalo vsebine

Kazalo tabel

Kazalo slik

Kazalo grafov

Zahvala

Pričujoča monografija predstavlja korpus govorjene slovenščine Gos. Da lahko avtorici o njem piševa, je bilo poleg najinega potrebnega veliko dela in dobre volje še številnih drugih ljudi.

Hvala vsem, ki so tako ali drugače pomagali, da je korpus nastal in lahko po njem brskamo …

… koordinator in vodja projekta Sporazumevanje v slovenskem jeziku, v okviru katerega je nastajal korpus, Simon Krek in Miro Romih

… sodelavci, ki so sodelovali pri izdelavi specifikacij za korpus: Simon Krek, Marko Stabej, Jana Zemljarič Miklavčič, Tina Verovnik, Andrej Žgank

… sodelavci pri snemanju korpusa: Brigita Bec, Mojca Bizjak, Rebeka Dragič, Aja Barbo Gruden, Jernej Golobič, Andreja Gregorič, Pija Kapitanovič, Ana Kočevar, Katja Krapež, Jaruška Majovski, Iztok Mikulan, Alenka Mirkac, Dusán Mukics, Barbara Omahen, Neža Pahovnik, Tomaž Potočnik, Lucija Ramovš, Lucija Rap, Erika M. Roblek, Mateja Strmšek, Ivana Šlaus, Maja Štefančič, Jure Tompa, Andrej Tomše, Slavka Vesenjak, Pija Vrezner

… sodelavci pri urejanju in transkribiranju korpusa: Aja Barbo Gruden, Mariša Bizjak, Mojca Bizjak, Rebeka Dragič, Jernej Golobič, Ana Gorinšek, Marko Kos, Katja Krapež, Jaruška Majovski, Iztok Mikulan, Alenka Mirkac, Barbara Omahen, Neža Pahovnik, Tomaž Potočnik, Erika M. Roblek, Mateja Strmšek, Maja Štefančič, Maja Šučur, Andrej Tomše, Bojana Zevnik

… sodelavci, ki so poskrbeli za ta ali oni tehnični vidik korpusa: Tomaž Erjavec za izdelavo XML-sheme, Simon Rozman za skrb z FTP-strežnikom, Damjan Vlaj za procesiranje zvočnih datotek

… sodelavci pri snovanju in izdelavi konkordančnika: Simon Krek, Iztok Kosem, Simon Rigač, Rok Rejc

Prav posebna hvala vsem govorcem, ki so prijazno dovolili, da posnamemo njihov pogovor za korpus.

Prav tako hvala vsem, ki so posredovali pri pridobivanju posnetkov za korpus iz medijskih hiš.

Hvala tudi vsem ravnateljem, ki so dovolili snemanje v okviru svojih šol, in drugim odgovornim, ki so dovolili snemanje v okviru podjetij.

V spomin …

… dragi kolegici Jani Zemljarič Miklavčič, ki se je od nas tik ob koncu nastajanja korpusa za vedno poslovila.

Preučevanje jezika prehaja v novo dobo, v kateri bo gonilo razvoja uporaba računalnikov. Z njimi lahko preverjamo svoje hipoteze, pokažejo nam lahko stvari, ki jih morda še nismo vedeli, in tudi stvari, ki precej zamajejo naše zaupanje v obstoječe modele in nas spodbudijo, da temeljito premislimo svoje predstave. Moje stališče pri vsem tem je, da zaupajmo besedilu.
(The study of language is moving into a new era in which the exploitation of modern computers will be at the centre of progress. The machines can be harnessed in order to test our hypotheses, they can show us things that we may not already know and even things which shake our faith quite a bit in established models, and which may cause us to revise our ideas very substantially. In all of this my plea is to trust the text.)
John Sinclair

1 Uvod

Zadnje desetletje lahko v jezikoslovju označimo kot ero korpusov. Izdelava različnih korpusnih virov za jezike se je široko razmahnila in hkrati s tem tudi njihova uporaba v jezikoslovnih raziskavah. Med najbolj prepoznavnimi tovrstnimi viri so t. i. nacionalni, po zasnovi referenčni korpusi, ki običajno pomenijo več deset- ali zadnji čas tudi več sto milijonov, ponekod pa tudi že nekaj milijard besed veliko uravnoteženo elektronsko bazo avtentičnih besedil (npr. nemški Dereko[1]), ki naj bi v osnovi predstavljala referenčni vir podatkov o jeziku na različnih ravneh (tako besedišče kot slovnica) in z različnih vidikov, za jezikoslovje in za druge vede, katerih predmet raziskovanja je jezik (zlasti npr. jezikovne tehnologije). Tudi za slovenščino že imamo tovrstne baze: sprva komercialni referenčni korpus Fida je kasneje narasel v Fido+[2], prostodostopno za zainteresirane uporabnike, širok in velik spekter besedil, čeprav manj uravnoteženih, pokriva tudi Nova beseda[3], pred vrati pa je že nova nadgradnja Fide+, korpus Gigafida[4]. Posebej zadovoljivo ob tem je dejstvo, da se omenjene baze ves čas razvijajo, saj je med drugim ideja referenčnega korpusa, da gre v korak z razvojem jezika in zagotavlja vedno sveže podatke, zato njegova izdelava ni nikoli dokončna.

Potem ko so bila zbrana in v elektronske baze urejena pisna besedila, je veliko jezikov(slovcev) po svetu začelo dopolnjevati tovrstne baze s (pod)korpusom govorjenih besedil[5]. Kot pravi Stabej v spremni misli k prvi govornokorpusni monografiji na Slovenskem (Zemljarič Miklavčič, 2008): »Zato se raziskovalci povsod po svetu lotevajo gradnje govornih korpusov, saj si v sodobnem komunikacijskem in informacijskem okviru jezikoslovje ne more več privoščiti avtoritarne nevednosti, ne more si privoščiti pojmovanja govorjenja kot odstopanja od idealizirane slovnice, izluščene na podlagi prestižne pisne besedilne produkcije.« Prav ta misel je ves čas spremljala ideje, priprave in izdelavo govornega korpusa Gos, ki ga tukaj predstavljamo.

Kot jezik, s katerim se ukvarja največ raziskovalcev po svetu, je bila seveda prva opremljena z govornim korpusom angleščina (govorni del korpusa British National Corpus – BNC[6], ki je še danes najpogosteje citiran in zgledovan), danes pa imajo referenčne govorne korpuse kot del nacionalnih korpusov ali kot samostojen korpus tudi vsi drugi večji (francoski – Corpus de la parole[7], nemški – Datenbank Gesprochenes Deutsch[8], italijanski – CLIPS[9], španski – El Corpus de referencia del español actual – CREA[10], Corpus del Espanol[11]) in nekoliko manj veliki (švedski – Goeteborg Spoken Language Corpus – GSLC[12], nizozemski – Spoken Dutch Corpus/Corpus Gesproken Nederlands – CGN[13], idr.) evropski jeziki, zelo aktivno gredo v korak s časom tudi (predvsem večji) slovanski jeziki (ruski – Russian National Corpus[14],

1 http://www.ids-mannheim.de/kl/projekte/dereko/
2 http://www.fidaplus.net/
3 http://bos.zrc-sazu.si/s_beseda.html
4 http://www.gigafida.net
5 V tej knjigi bomo raje govorili o govorjenem diskurzu ali govoru. Termin besedilo uporabljamo predvsem v zvezi s pisnimi besedili, v zvezi z govorjenimi besedili pa, ko povzemamo ali se sklicujemo na druge avtorje, sicer pa le v tehničnem smislu, če imamo v mislih samo zapis jezikovne podobe govorjenega diskurza.
6 http://www.natcorp.ox.ac.uk/
7 http://corpusdelaparole.in2p3.fr/
8 http://dsav-oeff.ids-mannheim.de/DSAv/SUCHMASK.HTM
9 http://www.clips.unina.it/it/
10 http://corpus.rae.es/creanet.html
11 http://www.corpusdelespanol.org/
12 http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3
13 http://lands.let.kun.nl/cgn/ehome.htm
14 http://www.ruscorpora.ru/en/index.html

češki – Česky narodni korpus[15], poljski – PELCRA[16], idr.). Temu seznamu se z referenčnim govornim korpusom Gos[17] (ne povsem zadnja) pridružuje tudi slovenščina.

Ideje o tem, da bi potrebovali govorni korpus slovenskega jezika, so ob začetku izdelave korpusa stare že najmanj dobro desetletje (Stabej, 1998; Stabej, Vitez, 2000). Njihovo uresničitev so spočetka zavirali pomanjkanje znanja in izkušenj z govornimi viri, delno negativni prvi odzivi na na videz neuresničljivo idejo, kasneje, ko je bilo glasov o potrebi po govornem korpusu vse več (Weiss, 2001: 422; Gorjanc, 2005: 53; Kenda Jež, 2004: 271; Verdonik, 2006; idr.) in so bila pripravljena tudi že teoretična izhodišča njegove izdelave, preizkušena na manjšem učnem korpusu (Zemljarič Miklavčič, 2007), pa začetek ni bil mogoč brez vsebinsko primernega razpisa razpoložljivih sredstev in ekipe ljudi, ki bi projekt pripravila, prijavila in izvedla. Po desetletju razmišljanj o referenčnem govornem korpusu in priprav nanj pa je situacija leta 2008 vendarle dozorela do realizacije in korpus je našel svoje mesto v okviru projekta Sporazumevanje v slovenskem jeziku.

Korpus Gos je namenjen naslednjim uporabnikom:

- znotraj projekta Sporazumevanje v slovenskem jeziku kot vir avtentičnih govorjenih diskurzov[18] 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.

1.1 Predstavitev poglavij

V tej knjigi bo govora o tem, kako je korpus Gos nastajal (poglavje 2), opisana bodo gradiva, ki so rezultat tega dela in dostopna za zainteresirane uporabnike (poglavje 3), predstavljen bo spletni vmesnik (konkordančnik), ki omogoča dostopno, enostavno in učinkovito brskanje po korpusnih gradivih prek spleta (poglavje 4), uporaba korpusa v praksi pa bo tudi ponazorjena z nekaj zanimivejšimi jezikovnimi korpusnimi analizami (poglavje 5). Cilj te monografije je torej podati popoln pregled vseh informacij, povezanih s korpusom Gos, ki bi zanimale tako njegove uporabnike kot prihodnje graditelje podobnih jezikovnih virov.

15 http://ucnk.ff.cuni.cz/
16 http://korpus.ia.uni.lodz.pl/conversational/
17 www.korpus-gos.net
18 V monografiji (enako kot skozi ves proces izdelave korpusa) uporabljamo termin diskurz, ki sicer ni značilen za korpusno jezikoslovje, ampak izhaja iz jezikovne pragmatike in analize diskurza. Taka odločitev temelji predvsem na želji, da ohranimo zavest, da gre za realizacijo jezika v govoru in da za njegovo spoznavanje ni dovolj, da gledamo samo zapis govora (govorjeno besedilo), ampak da ga tudi poslušamo in po potrebi upoštevamo kontekst rabe – in vse to korpus Gos tudi omogoča. V strokovnem jeziku termin diskurz najbolj ustreza temu namenu, saj se nanaša tako na jezikovno kot družbeno in kognitivno razsežnost govornega dogodka.

Poglavje 2: Govorni korpus Gos je v splošnem namenjen korpusnim raziskavam govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah. Naravnan je v kar se da referenčni zajem govorjenih diskurzov, vendar je pri tem omejen z obsegom milijon besed. Ta omejitev izhaja iz obsega odobrenega projekta za korpus in obsega dela, potrebnega za njegovo izdelavo. Definirani so bili trije osnovni cilji, ki so pomagali določiti njegovo zasnovo:

- zajeti vzorčne primere različnih govornih situacij in različnih govorjenih diskurzov,
- zajeti govorjeni diskurz demografsko reprezentativnega vzorca govorcev slovenskega jezika,
- zajeti predvsem tiste govorne situacije, v katerih so uporabniki jezika najbolj pogosto produktivno-receptivno udeleženi.

Pri določanju kriterijev za zajem posnetkov v korpus Gos smo kombinirali demografske in besedilnovrstne kriterije. Demografski kriteriji so definirani za zasebni diskurz v naslednjih izhodiščnih razmerjih:

- prvi jezik: slovenščina 98%, drugi jeziki 2%,
- država bivanja: Slovenija 97%, Avstrija 1%, Italija 1%, Madžarska 1%,
- spol: moški 50%, ženski 50%,
- starost: do 34 let 40%, nad 35 let 60%,
- dosežena izobrazba: nižja (osnovna in srednja šola) 70%, višja (več kot srednja šola) 30%,
- regijska pripadnost (glede na večja regionalna mestna središča):
- zasebni diskurz: govor JZ Slovenije brez ljubljanske regije (NM, KK, KR, GO, PO, KP) 35%, govor ljubljanske regije 25%, govor SV Slovenije brez mariborske regije (MS, SG, CE) 25%, govor mariborske regije 15%,
- nezasebni diskurz: JZ Slovenija 60%, SV Slovenija 40%.

Besedilnovrstni kriteriji so določeni v naslednjih razmerjih:

- javnost: javni diskurz 60% (razvedrilni 20%, nerazvedrilni 40%), nejavni diskurz 40% (nezasebni 15%, zasebni 25%),
- prenosnik: osebni stik 50%, telefon 10%, radio 20%, televizija 20%.

Posebej so znotraj javnega diskruza v osebnem stiku definirani kriteriji za šolski diskurz:

- stopnja šolanja: osnovna šola 55% (drugo triletje 50%, tretje triletje 50%), srednja šola 45% (gimnazije 40%, nižje in srednje poklicno, srednje strokovno in poklicno-tehniško izobraževanje 60%),
- regija: JZ 60%, SV 40%,
- učni predmet: naravoslovni in tehnični predmeti 50%, družboslovni in humanistični predmeti 50%.

Glede uresničevanja zastavljene sheme je bila sprejeta odločitev, da Gos izpolnjuje zastavljene kriterije, če so ti uresničeni v okviru 30-odstotnega odstopanja relativno, ter zelo dobro izpolnjuje zastavljene kriterije, če so uresničeni v okviru 10-odstotnega odstopanja relativno. Prav tako štejemo, da Gos izpolnjuje zastavljene kriterije, če samo določen del gradiva vključuje podatke, na podlagi katerih je mogoče preveriti pokritost definiranih kriterijev, in ta del Gosa izkazuje ustrezno zastopanost.

Korpus Gos vključuje posnetke in transkripcije posnetkov iz medijev (radio, televizija), posnetke šolskega pouka in predavanj ter posnetke nejavnih pogovorov. V zvezi s tem je potrebno slediti veljavnim pravnim zahtevam po varovanju gradiva in morebitnih osebnih podatkov ter spoštovanju avtorskih pravic. Avtorske pravice za gradiva, prejeta od medijev, so bile v okviru konzorcija projekta Sporazumevanje v slovenskem jeziku urejene s pogodbo o odstopu avdio in video posnetkov radijskih in televizijskih oddaj ali drugih programskih vsebin. Avtorske pravice za terenske posnetke so bile urejene z vsakim posnetim govorcem oz. zakonitim zastopnikom pri mladoletnih osebah posebej z izjavo o odstopu pravic. Uporaba korpusa je predvidena v skladu z licenco Creative Commons: "priznanje avtorstva" + "nekomercialno" + "deljenje pod istimi pogoji". Vsi podatki so v Gos vključeni tako, da so anonimizirani.

Zajemanje avtentičnih posnetkov pogovorov na terenu v obsegu, predvidenem za korpus Gos, zahteva ustrezno organizacijo in koordinacijo dela. Za korpus Gos so bile izbrane naslednje rešitve:

- snemalna oprema: Samson Zoom H4,
- velika ekipa snemalcev (ob koncu snemanja jih je bilo skupaj okrog 30), ki izhajajo iz različnih slovenskih regij,
- delo z datotekami na daljavo preko FTP-strežnika,
- pridobivanje posnetkov in soglasij za snemanje na institucionalni ravni za javni diskurz in prek posameznih snemalcev za zasebni diskurz.

Pri pripravah na transkribiranje so bila upoštevana načela EAGLES in TEI, v slovenskem prostoru pa dialektološka praksa, praksa besedilo-slovnih in pragmatičnih raziskav govorjene slovenščine in jezikovnotehnološka praksa. Definirana sta bila dva nivoja transkribiranja, pogovorni in standardizirani zapis. V pogovornem zapisu zapisujemo govor na način, zapiši, kot slišiš. V standardiziranem zapisu zapisujemo govor na način, zapiši, kot pišemo. Od orodij za transkribiranje so bili preizkušeni ELAN, Exmaralda, Praat in Transcriber, izbran pa slednji. Posebna pozornost pri izdelavi osnovne transkripcije s pogovornim zapisom je bila ob velikem številu transkriptorjev namenjena zagotavljanju kvalitete. V ta namen je transkribiranju sledilo dvostopenjsko pregledovanje. Standardizacijo zapisa je izvajala ena sama oseba.

Poglavje 3: Gradiva, zajeta v korpus Gos, vključujejo naslednje tipe govornih dogodkov: vsebinsko zaključene odseke iz moderiranih vsebin in pogovorov na radiu in televiziji, novinarske prispevke, odseke iz moderiranih oddaj, resničnostnih šovov in športnih prenosov; posnetke osnovnošolskih in srednješolskih učnih ur, predavanj na fakulteti in javnih predavanj; posnetke govora na formalnih in neformalnih delovnih sestankih, konzultacijah na fakulteti, ob raznih storitvah, v formalnih razgovorih, ob prodaji in v trgovini, ob svetovanju, posredovanju informacij itd.; posnetke pogovorov v družini ter med prijatelji in znanci.

Dejanska razmerja med besedilnovrstnimi in demografskimi kriteriji so naslednja:

Besedilnovrstno:

- javnost: javni diskurz 56% (razvedrilni 22%, nerazvedrilni 34%), nejavni diskurz 44% (nezasebni 15%, zasebni 29%),
- prenosnik: osebni stik 49%, telefon 10%, radio 21%, televizija 20%.

Demografsko – samo zasebni diskurz:

- prvi jezik: slovenščina 95%, drugi jeziki 5%,
- spol: moški 49%, ženski 51%,
- starost: do 34 let 63%, nad 35 let 37%,
- dosežena izobrazba: nižja (osnovna in srednja šola) 70%, višja (več kot srednja šola) 30%,
- regijska pripadnost: govor JZ Slovenije brez ljubljanske regije 31%, govor ljubljanske regije 27%, govor SV Slovenije brez mariborske regije 21%, govor mariborske regije 12%, Avstrija 4%, Italija 5%.

Gradiva korpusa Gos vključujejo tudi natančne podatke o diskurzih in govorcih, in sicer:

- diskurzi: identifikacijska koda, dolžina posnetka, tip diskurza, vrsta institucije/situacije, tip govornega dogodka, vir posnetka, regija, kraj in čas odvijanja/predvajanja diskurza, število aktivnih udeležencev in kratek prosti opis govornega dogodka v stavku ali dveh,
- govorci: spol, starost, regionalna pripadnost (lahko jih je več, če je govorec dalj časa bival v različnih regijah), izobrazba, prvi jezik.

Pri transkribiranju so diskurzi segmentirani na vloge (govor enega govorca, dokler ga ne prekine drug govorec) in izjave (najmanjša strukturna enota zapisa, ki je prozodično, semantično in skladenjsko približno zaključena enota). Če hkrati govorita dva ali trije govorci, je za začetek hkratnega govora označen začetek izjave, v kateri se vključi drug govorec, za konec hkratnega govora pa konec zadnje izjave, v kateri se pojavlja hkratni govor.

Govor je zapisan v pogovornem zapisu in standardiziranem zapisu. V pogovornem zapisu je cilj karseda zvesta predstavitev glasovne

podobe govora v karseda berljivi obliki. Zapisan je v veljavnem slovenskem črkopisu, brez dodatnih znakov, vendar ponazarja redukcije, pokrajinsko specifične glasove in druge značilnosti govorjenega jezika. Namen standardiziranega zapisa je, da odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Ciljna oblika je standardna različica istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru.

Tekstovni del korpusa je bil v zadnjem koraku zapisan v standardu XML in sloni na priporočilih TEI (Text Encoding Iniciative), različice TEI P5. Dodane oblikoslovne oznake temeljijo na naboru oznak, ki je bil definiran v priporočilih JOS. Korpus je dostopen tudi v tej različici in se lahko sname s spletne strani.

Poglavje 4*:* Konkordančnik korpusa Gos je bil izdelan v posebnem projektu in uporabnikom omogoča napredne metode iskanja po transkripcijah in spremljevalnih metapodatkih ter poslušanje pripadajočih segmentov govora v izvornih posnetkih za vsako izjavo med iskalnimi zadetki. Pri tem smo upoštevali potrebe vsakdanjih uporabnikov korpusa govorjene slovenščine, potrebe v izobraževanju in potrebe znanstvenoraziskovalne skupnosti, ki bo pri svojem delu lahko uporabljala tovrstno infrastrukturo. Konkordančnik je postavljen na spletnem mestu www.korpus-gos.net, ki omogoča tudi dostop do spremljevalnih podatkov, povezav, objav ipd.

Zasnova konkordančnika za korpus Gos temelji na izčiščenosti in intuitivnem ravnanju uporabnika, hkrati pa omogoča izkoriščanje kompleksnih podatkov o govorcih in diskurzih, primerno za zahtevnejše uporabnike.

Osnovni način iskanja po korpusu je enostavno iskanje. Ker smo zajeti govor zapisali na pogovornem in standardiziranem nivoju, je omogočeno tudi iskanje po obeh nivojih.

Napredno iskanje je namenjeno zahtevnejšim uporabnikom, ki iščejo po slovničnih lastnostih neke besede, sopojavljanje različnih besed ali kombinirajo iskanje določene besede s posebnimi dogodki v govoru (smeh, nejezikovni zvoki…).

Zavihek Seznam nudi hiter vpogled v število pojavitev vseh realiziranih oblik določene besede. Pri tem lahko specificiramo iskano besedo tudi glede na besedno vrsto in druge oblikoslovne lastnosti.

Če želimo več podatkov o posamezni konkordanci, kliknemo nanjo. Prikaže se razširjeni kontekst, podatki o diskurzu, podatki o govorcu in standardizirani zapis razširjenega konteksta. Rezultate iskanja lahko tudi filtriramo po tipih diskurzov in govorcev. Prav tako jih lahko izvozimo v tekstovni format ter natisnemo ali shranimo.

Uporabniki lahko tekstovni del korpusa Gos snamejo s spletne strani www.korpus-gos.net. Tekstovno gradivo je dostopno pod

pogoji, ki jih določa licenca Creative Commons: »nekomercialno« + »priznanje avtorstva« + »deljenje pod istimi pogoji«.

Poglavje 5: Na koncu predstavimo primere uporabe korpusa. Najprej primerjamo korpusne podatke o rabi diskurznih označevalcev *aha* in *aja*. Rezultati potrjujejo, da sta v javnih informativno-izobraževalnih televizijskih in radijskih oddajah *aha*, še bolj pa *aja* redko rabljena. Pogostejša v javnem izobraževalnem diskurzu je njuna raba v osebnem stiku. Bolj ko se odmikamo od formalnih situacij, bolj njuna raba narašča, tako je pri javnem razvedrilnem diskurzu že višja, v splošnem pa najvišja v nejavnem diskurzu. Vendar nas v nejavnem diskurzu preseneti *aha*, ki je veliko pogosteje kot v zasebnih rabljen v nezasebnih situacijah. *Aja* je nasprotno pogosteje rabljen v zasebnih kot v nezasebnih nejavnih diskurzih, pri njem je torej verjeten močan vpliv formalnosti situacije na rabo.

Aha in *aja* se pogosto pojavljata v začetku izjave skupaj z drugimi diskurznimi označevalci v neke vrste gručah diskurznih označevalcev. Z naprednim iskanjem po okolici besed vidimo, da prevladujejo gruče, v katerih se sopojavljata samo dva različna diskurzna označevalca, vendar se pri teh pogosto ali eden ali drugi ali oba lahko dva- ali večkrat ponovita. Najpogostejše so gruče *aha* ali *aja* in *ja*, gruče *aha* in katerega od diskurznih označevalcev *dobro/v redu/okej* ter gruče *aha* ali *aja* in *no*. V teh gručah je običajneje, da je na prvem mestu ali *aha* ali *aja*. *Aha* in *aja* v isti gruči se pojavljata redko. Trije ali štirje diskurzni označevalci v gruči se pojavljajo redkeje, pri tem ni posameznega prevladujočega vzorca, ampak so gruče zelo različne, spet pa večinoma z *ja, dobro/v redu/okej* ter *no*.

V drugem primeru z iskanjem po Seznamu iščemo pogovorne različice besede *lahko*. Število njenih različnih pogovornih zapisov nas preseneti, vendar številke kažejo, da jih je večina zelo redkih. Daleč najpogostejša je vendarle standardna izgovorna različica *lahko*, z desetkrat manjšo pogostnostjo ji sledijo nekatere reducirane različice (*loh, lah, lahk, lohk*). Reducirane oblike so redke predvsem v mariborski in murskosoboški regiji, najpogostejše pa v novogoriški regiji.

V zadnjem primeru se posvetimo izrazu *in tako naprej* in njegovim različicam. Raba tovrstnih izrazov je v javnem diskurzu precej večja kot v nejavnem. Različica z veznikom *pa* (*pa tako naprej*) je veliko redkejša kot različica z veznikom *in* ter se – nasprotno – pojavlja prevladujoče v nejavnem diskurzu. Reducirane oblike (s *tko, tk, tak*) so presenetljivo v skupnem seštevku skoraj enako pogoste kot nereducirane (s *tako*). Razvrstitev zadetkov glede na regije kaže, da so različice s *tko* in *tk* značilne za osrednjo, južno in zahodno Slovenijo, s *tak* pa za severno in vzhodno Slovenijo. Izraz *in/pa tako naprej* je v splošnem veliko bolj pogost kot *in/pa tako dalje*.

2 Snovanje in izdelava korpusa Gos

Referenčni govorni korpus slovenskega jezika Gos naj bi sprva predvidoma predstavljal govorni podkorpus referenčnega korpusa slovenskega jezika v okviru projekta Sporazumevanje v slovenskem jeziku[19] (SSJ – www.slovenscina.eu), vendar je z odobrenim projektom za lasten konkordančnik[20] dobil bolj samostojno vlogo, ki je omogočala večjo izraznost posebnosti govorjenega jezika (brskanje po dveh nivojih zapisa govora, povezava zapisa z zvokom...).

V tem poglavju je opisano, kako so bili postavljeni kriteriji za zajemanje gradiva za korpus Gos, kako je potekalo snemanje, kako so bila zastavljena načela za transkribiranje gradiva in kako je potekalo transkribiranje. Podrobna pravila transkribiranja pa so opisana v poglavju 3, pri pregledu korpusnih gradiv.

2.1 Specificiranje kriterijev za zajem gradiv

2.1.1 Pregled tujih primerljivih in domačih predhodnih projektov

V svetu se v zadnjih desetletjih gradijo številni govorni korpusi. V nadaljevanju na kratko navajamo nekatere po obsegu in namenu najbolj primerljive, predvsem evropske projekte gradnje govornih korpusov ter nekatere najbolj relevantne slovenske korpuse in publikacije.

Korpus britanske angleščine **(British National Corpus – BNC[21])** vključuje govorni podkorpus v obsegu deset milijonov besed. Ta je sestavljen iz demografskega dela, ki je uravnotežen po demografskih kriterijih: družbeni status, spol, regionalna razpršenost (govorci z 38 različnih lokacij) in starost. Po spolu, družbenem statusu in starosti so bile različne skupine enakomerno zastopane. Drugi del govornega podkorpusa je kontekstno usmerjen, in sicer skuša v enakomernih deležih zajeti štiri kategorije družbenega konteksta: izobraževanje in informiranje, poslovne dogodke, institucionalne in javne dogodke (kot so maše, politični govori, parlamentarna zasedanja) ter zabavo (športni komentarji, klubska srečanja, nagovori ob večerji...). Referenčni korpus angleščine je tudi **Collins Corpus z Bank of English**[22], ki pa je zelo skopo dokumentiran. Osnovno vodilo avtorjev tega korpusa je kvantiteta in ažurnost, tako da mu mesečno dodajajo novo gradivo.

Češki govorni korpus[23] je sestavljen iz treh enot: praškega govornega korpusa (PMK, zajema predvsem govor Prage in bližnje okolice, obseg je

19 Projekt Sporazumevanje v slovenskem jeziku je bil sofinanciran s strani Evropskega socialnega sklada ter Republike Slovenije, Ministrstva za šolstvo in šport. Projekt je izvajal konzorcij v sestavi Institut Jožef Stefan, Univerza v Ljubljani, Znanstvenoraziskovalni center SAZU, Trojina, zavod za uporabno slovenistiko, in Amebis, d. o. o., Kamnik, kot nosilec, trajal pa je od junija 2008 do decembra 2013. Znotraj projekta je izdelava govornega korpusa potekala od septembra 2008 do decembra 2010.

20 Konkordančnik za iskanje po govornem korpusu Gos je bil izdelan v okviru projekta Spletni konkordančnik za nacionalni govorni korpus slovenskega jezika. Trajal je od septembra 2009 do oktobra 2010, izvajal pa ga je konzorcij v sestavi Univerza v Mariboru (Fakulteta za elektrotehniko, računalništvo in informatiko), Trojina, zavod za uporabno slovenistiko, in Univerza v Ljubljani (Filozofska fakulteta). Kot podizvajalec je pri izdelavi konkordančnika sodelovalo podjetje Amebis, d. o. o., Kamnik.

21 http://www.natcorp.ox.ac.uk/corpus/creating.xml

22 http://mycobuild.com/about-collins-corpus.aspx

23 http://ucnk.ff.cuni.cz/english/index.html

675.000 besed), brnskega govornega korpusa (BMK, zajema predvsem govor Brna in bližnje oklice, obseg je 490.000 besed) ter ORAL2006 (zajema govor čeških narečnih področij in obsega milijon besed). Vsi trije podkorpusi se opirajo na sociolingvistične kriterije: starost (od 20 do 35, nad 35), spol, izobrazba (osnovna ali višja) in formalnost govora (formalni – predvsem monolog, neformalni – predvsem dialog).

Enega največjih in pravkar začetih projektov gradnje govornega korpusa predstavlja nacionalni **poljski** korpus (National Corpus of Polish – NKJP; Przepiorkowski et al., 2008), ki bo vključeval govorni podkorpus v obsegu 30 milijonov besed (javni govori, parlamentarne debate, televizijske oddaje, pogovorne oddaje, radijski intervjuji, dnevnoinformativne oddaje, tri milijone besed pa naj bi obsegal podkorpus vsakdanjih pogovorov).

Švedski govorni korpus (Goeteborg Spoken Language Corpus – GSLC[24]) obsega 1,4 milijona besed. Sestavljajo ga samo posnetki spontanega govora. Bolj kot demografska uravnoteženost jih je zanimala pokritost čim večjega spektra različnih govornih dogodkov, tako da vključuje več kot 25 družbenih aktivnosti.

Nizozemski govorni korpus (Spoken Dutch Corpus/Corpus Gesproken Nederlands – CGN[25]) obsega skoraj devet milijonov besed. V prepleteni strukturi ločijo 14 kategorij: monolog – multilog/dialog, javno – zasebno, javno nadalje delijo na radio in televizijo, brano – spontano, formalno – neformalno in na koncu besedilne tipe: pogovor, intervju, telefonski pogovor itd. Pri tistih kategorijah, kjer je bilo smiselno, so upoštevali tudi demografske značilnosti (spol, starost, regijski izvor, socialno-ekonomski status) kot kriterij za vzorčenje.

Korpus govorjene **estonščine** (Hennoste et al., 2008) obsega okoli milijon besed. Kriteriji za zajem gradiva so družbena in narečna pripadnost govorcev, dialog – monolog, stopnja spontanosti, prenosnik (osebni stik, telefon, množični medij) in stopnja formalnosti s štirimi podkategorijami: razmerje med sogovorniki (poznani, nepoznani), vloga sogovornikov (privatna oseba, predstavnik institucije), scena (zasebni prostor, uradni prostor), namen interakcije (udeležba, informiranje). Vsebina korpusa so tako predvsem institucionalni ali delno institucionalni pogovori, manj pa zasebni pogovori.

Korpus govorjene **italijanščine** (CLIPS[26]) obsega sto ur govora, kar bi znašalo okoli 900.000 besed. Posnetki so narejeni v 15 mestnih regijskih središčih po Italiji. Za vsako od teh središč so zajeti mediji (radio in televizija), dialogi, brani govor profesionalnih in neprofesionalnih govorcev in telefonski pogovori s simulirano hotelsko recepcijo. Eden večjih korpusov za italijanščino je tudi LABLITA[27], ki je razdeljen v dva podkorpusa: govor odraslih govorcev v obsegu 640.000 besed vključuje pogovore v osebnem stiku, govor prek masovnih medijev in govor prek telefona. Pogovori v osebnem stiku so nadalje uravnoteženi glede na družbeni kontekst (družina, privatno, javno), vrsto interakcije

24 http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3
25 http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm
26 http://www.alphabit.net/Glottophilia/2007/02/clips-corpus-of-spoken-italian.html
27 http://lablita.dit.unifi.it/corpora/descriptions/lablita/

28 http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php

(vodena, prosta) in strukturo komunikacijskega dogodka (monolog, dialog, pogovor). Drugi podkoprus LABLITA vključuje govor zgodnjega učenja italijanščine (pri otrocih v starosti od 15 do 36 mesecev).

Referenčni korpus sodobne **portugalščine** (Reference Corpus of Contemporary Portugese – CRPC[28]) vključuje govorni podkorpus v obsegu 2,5 milijona besed. Pokriva portugalščino vseh portugalsko govorečih dežel po svetu.

Za slovenski jezik je pred korpusom Gos že obstajalo nekaj srednje velikih in manjših specializiranih govornih korpusov, ki zajemajo:

- televizijske dnevnoinformativne in pogovorne oddaje:
  - BNSI Broadcast News, 72 ur (Žgank et al., 2004),
  - Broadcast News Speech Database, 34 ur (Žibert, Mihelič, 2004),
- parlamentarni govor:
  - določen delež korpusa Fida+ predstavljajo transkripcije razpra iz državnega zbora, ki pa so prilagojene drugim, neraziskov namenom in niso vedno dobesedne; obsegajo okrog dva milijona besed,
  - baza Sloparl vključuje prav tako transkripcije razpra državnega zbora, ki pa se trenutno urejajo v raziskov namene; obsega okrog sto ur govora (Žgank et al., 2006),
- telefonske pogovore s turističnimi agencijami, hotelsko recepcijo in turistično pisarno – Turdis, obseg okrog 30.000 besed (Verdonik, Rojc, 2006),
- govorjene diskurze različnih tipov, osnutek referenčnega korpusa, obsega okrog 15.000 besed (Zemljarič Miklavčič, Stabej, 2005; Zemljarič Miklavčič, 2006).

Poleg navedenih obstajajo še druge govorne baze, ki pa ne vključujejo avtentičnih govorjenih diskurzov, in nekaj transkripcij sicer avtentičnih govorjenih diskurzov, nastalih predvsem v okviru različnih diplomskih, magistrskih in doktorskih projektov, ki pa niso urejene do te mere, da bi omogočale avtomatsko iskanje brez dodatnega urejanja.

Od publikacij je za definiranje specifikacij za gradnjo referenčnega govornega korpusa slovenskega jezika relevantna predvsem disertacija, katere cilj je bil definirati načela za gradnjo govornega korpusa slovenščine (Zemljarič Miklavčič, 2007), in iz nje izhajajoče publikacije (Zemljarič Miklavčič, 2006; Zemljarič Miklavčič, Stabej, 2005), prvi grob osnutek korpusa govorjenih besedil pa je bil predstavljen že v Stabej, Vitez (2000).

Iz zgornjega pregleda vidimo, da se avtorji odločajo za različne pristope k definiranju kriterijev za zajem gradiv. Nekateri poudarjajo predvsem lastnosti govorcev, kot so spol, starost, izobrazba, kar imenujemo s skupnim terminom demografski kriteriji (npr. češki korpus). Drugi poudarjajo zlasti različnost zajetih besedil in pri tem upoštevajo

spontanost, število udeležencev, formalnost, tipologijo govorjenih besedil itd., kar bomo imenovali s skupnim terminom besedilnovrstni kriteriji (enako Zemljarič Miklavčič, 2007; 2008). Tak pristop je uporabljen na primer pri švedskem korpusu. Večina projektov pa kombinira oba pristopa na različne načine: vsak sklop kriterijev upoštevajo v svojem podkorpusu (BNC), izhajajo iz besedilnovrstnih kriterijev in vključujejo demografske, kjer je to smiselno (nizozemski korpus), ali pa izhajajo iz regijske razdeljenosti in nato vključujejo besedilnovrstne kriterije (italijanski korpus).

2.1.2 Cilji izdelave govornega korpusa Gos

Govorni korpus Gos je v splošnem namenjen korpusnim raziskavam govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah. Naravnan je v kar se da referenčni zajem govorjenih diskurzov, vendar je pri tem omejen z obsegom milijon besed. Ta omejitev izhaja iz obsega odobrenega projekta za korpus in obsega dela, potrebnega za njegovo izdelavo.

V projektu Sporazumevanje v slovenskem jeziku je bil korpus Gos eden od virov za:

- izdelavo leksikalne podatkovne baze s podatki o frekvenci, pomenski strukturi in z zgledi rabe,
- izdelavo pedagoške korpusne slovnice.

Še bolj pomembna kot vloga znotraj projekta, v katerem je nastajal, pa je razpoložljivost korpusa Gos za zainteresirane zunanje uporabnike, ki lahko do njega prosto dostopajo prek spletnega naslova www.korpus-gos.net. Cilj korpusa Gos je tako tudi, da omogoča razne korpusne raziskave[29] ter da je na voljo zainteresiranim uporabnikom v šolstvu in raznih poklicih, povezanih z govorom. Gos naj bi predstavljal nujni začetni korak k boljšemu poznavanju spontane govorjene slovenščine v vsakdanjih okoliščinah.

Ob upoštevanju zgoraj navedenega so bili definirani trije osnovni cilji, ki so pomagali določiti zasnovo Gosa. Želeli smo zajeti:

- vzorčne primere 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.

29 Zlasti v leksikografiji, skladnji, analizi diskurza/konverzacijski analizi in govornih tehnologijah, potencialno pa tudi v drugih vejah jezikoslovja in raznih vejah psihologije, sociologije, antropologije, kognitivnih in informacijskih znanosti, ki se dotikajo človeškega govora in govorne komunikacije.

2.1.3 Opredelitev glede posameznih vidikov zajemanja gradiva

Pri zajemanju gradiva za govorni korpus je treba upoštevati tudi razne okoliščine snemanja gradiv na terenu. Poleg zgoraj navedenih ciljev korpusa so na definiranje kriterijev za zajem gradiv zato vplivala še naslednja stališča (o tem, kako so bila dejansko uresničena, glej 3.1):

- Dolžina posnetkov: kolikor mogoče ohraniti avtentično dolžino govorjenih diskurzov.
- Avtentičnost posnetkov: zajeti predvsem avtentične govorjene diskurze, tj. diskurze, ki potekajo v naravnem okolju in niso zrežirani ali umetno sproženi. Ta kriterij je pomembnejši kot popolna demografska in besedilnovrstna uravnoteženost.
- Pravno-etični vidiki snemanja: zagotoviti predhodno seznanitev in soglasje vseh govorcev, katerih govor bo posnet za namene Gosa. 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.
- Tehnični vidiki snemanja: ustrezno namestiti in aktivirati snemalno tehnično opremo (mikrofone, nosilce podatkov ipd.). Te aktivnosti lahko zmotijo običajen potek diskurza. Nadalje tudi s tem, ko namestimo mikrofon, govorce spomnimo, da bo njihov govor posnet, kar lahko vpliva na njihovo vedenje. Ker je za natančno transkripcijo potreben kvaliteten zvok, pa je s tehničnih vidikov zelo težavno in vprašljivo snemanje v zelo šumnih okoljih ali v skupinah, kjer je prisotnih zelo veliko govorcev.
- Spontanost govora: izločiti govorne situacije, v katerih je govorjeni jezik prevladujoče le govorna predstavitev pisnega besedila (brano ali naučeno besedilo). Čeprav lahko pričakujemo, da je tudi brani diskurz različen od pisnega diskurza, pa je ta razlika bistveno manjša kot pri spontanem diskurzu, poleg tega že obstaja nekaj korpusov, ki beležijo predvsem brani ali delno spontani govorjeni diskurz v slovenskem jeziku[30]. Ker je Gos po obsegu zelo omejen, doslejšnji korpusni viri za slovenščino pa so najskromnejši ravno na področju spontanega govorjenega diskurza, je kriterij za zajem v Gos prevladujoča vsaj delna ali popolna spontanost, tj. da govorec vsaj delno sproti tvori besedilo (npr. ima vnaprej pripravljene samo začetne izjave) oz. da v celoti sproti tvori besedilo (kot je značilno za zasebni diskurz).
- Govor mladoletnikov: ne povsem izključiti govorcev, mlajših od 18 let. Govor otrok in mladostnikov se pogosto zbira ločeno od govora odraslih govorcev. Toda pomemben del govorne komunikacije odraslih je tudi pogovor z otroki in mladostniki (zlasti v okviru družine). Ker bo Gos tudi jezikovni vir za izdelavo

30 Npr. diskurz po televiziji: BNSI Broadcast News (Žgank et al., 2004), Broadcast News Speech Database (Žibert, Mihelič, 2004), diskurz v parlamentu: Sloparl (Žgank et al., 2006).

pedagoške slovnice, pa bo pomemben del korpusa zajemal tudi šolski diskurz za otroke od desetega leta starosti naprej.
- Prvi jezik govorcev: zajeti tudi določen delež govorcev, za katere slovenščina ni prvi oz. materni jezik. Pri tem bomo skušali upoštevati realno demografsko sestavo govorcev slovenščine.
- Slovenci v zamejstvu in po svetu: zajeti tudi določen delež govorcev, ki živijo v zamejstvu. Govor Slovencev, ki živijo po svetu, pa v tej fazi ni vključen v Gos.

31 Spol, starost, izobrazba, regijska pripadnost, socialni status, višina prihodkov ipd.
32 Javnost/nejavnost govora, formalnost/neformalnost govora, število sogovornikov v diskurzu, prenosnik, namen diskurza, tematika ipd.

2.1.4 Kriteriji za zajem gradiva

Glede na zastavljene cilje Gosa se zdi najustreznejša metoda tista, pri kateri kombiniramo demografske[31] in besedilnovrstne[32] kriterije. Pri tem upoštevamo tudi zastavljene cilje Gosa, tuje izkušnje na tem področju, specifike slovenskega okolja, tehnične in pravne vidike zajemanja gradiva, potrebe korpusnih raziskav in statistično reprezentativnost posameznih skupin glede na predvideni obseg korpusa.

2.1.4.1 Demografski kriteriji

Demografski kriteriji so: spol, starost, dosežena izobrazba in regijski izvor. Gre za kriterije, ki glede na obstoječe raziskave v slovenskem jezikoslovju (prim. Zemljarič Miklavčič, 2007; 2008) in glede na hipotetična predvidevanja najbolj vplivajo na razlike v govoru. Pri tem ni upoštevan kriterij socialnega izvora, ker je (z današnjega stališča) za slovenske razmere težko določljiv in vprašljiv. Vsekakor bi demografske kriterije lahko še bistveno bolj natančno specificirali, vendar korpusno zbiranje gradiva zahteva čim bolj robustno in enostavno shemo, zato smo se omejili le na najbolj bistvene dejavnike.

Demografski kriteriji so bili uravnoteženi glede na najnovejše podatke SURS (dostopne l. 2008) in so usklajeni v skladu s cilji Gosa. Za večje skupine so zaokroženi na deset odstotkov. Tabela 1 predstavlja izhodiščna razmerja demografskih kriterijev v odstotkih.

Tabela 1: Izhodiščna razmerja demografskih kriterijev v odstotkih

tip diskurza	kategorija	podkategorija	%
zasebni	prvi jezik	slovenščina	98%
		drugi jeziki	2%
zasebni	država bivanja	Slovenija	97%
		Avstrija	1%
		Italija	1%
		Madžarska	1%
zasebni	spol	moški	50%
		ženski	50%

tip diskurza	kategorija	podkategorija	%
zasebni	starost	do 34 let	40%
		nad 35 let	60%
zasebni	dosežena izobrazba	nižja (osnovna in srednja šola)	70%
		višja (več kot srednja šola)	30%
zasebni	regijska pripadnost*	govor JZ Slovenije brez ljubljanske regije (NM, KK, KR, GO, PO, KP)	35%
		govor ljubljanske regije	25%
		govor SV Slovenije brez mariborske regije (MS, SG, CE)	25%
		govor mariborske regije	15%
javni in nezasebni	regijska pripadnost*	JZ Slovenija	60%
		SV Slovenija	40%

* Regijsko pripadnost označujemo glede na večja regionalna mestna središča, h katerim gravitira posamezno področje in ki sovpadajo z registrskimi območji v Sloveniji.

2.1.4.2 Besedilnovrstni kriteriji

Kriteriji, na podlagi katerih razvrščamo govorjene diskurze, so lahko zelo različni (npr. klasifikacija na podlagi namernosti, prenosnika, funkcije, strukture diskurza, tematike, socialne zvrstnosti itd.). Večinoma se avtorji pri definiranju besedilnovrstnih kriterijev za zajem gradiva v reprezentativni govorni korpus odločajo za kombinacijo več kriterijev. Zemljarič Miklavčič (2007: 96) za slovenščino predlaga naslednje besedilnovrstne kriterije: struktura besedila (monolog, dialog/multilog), okoliščine (javna besedila, zasebna besedila), govorni položaj (formalni, neformalni), prenosnik (osebni stik, telefon, avdio, video).

Definiranje besedilnovrstnih kriterijev je težavnejše kot definiranje demografskih kriterijev, saj govorjenim diskurzom pogosto ne moremo nedvoumno določiti neke lastnosti (npr. spontanost ali nespontanost, formalnost ali neformalnost) oziroma se lahko v istem diskurzu meša več lastnosti (npr. prevladovanje monologa, ki pa občasno preide v dialog), tako da je lahko smiselnost posameznih faktorjev kot kriterijev za zajem celo vprašljiva. Zastavljeni cilji Gosa, tuje izkušnje, specifike slovenskega okolja, potrebe korpusnih raziskav in statistična reprezentativnost posameznih skupin glede na predvideni obseg korpusa so pri zasnovi Gosa vodili do naslednjih odločitev:

- Struktura besedila (monolog/dialog) za Gos ni relevanten kriterij. To utemeljujemo s ciljem, da zajamemo predvsem spontane ali delno spontane govorne situacije, kar pomembno vpliva na strukturo zajetih gradiv: monologi so namreč praviloma vnaprej pripravljena, brana besedila v javnem diskurzu. Večina diskurzov v Gosu bo torej dialoških ali multiloških.
- Stopnje spontanosti ni smiselno posebej izpostavljati kot kriterij za zajem. Cilj Gosa je namreč, da zajame predvsem spontani

govor, le v manjši meri vnaprej pripravljenega, branega pa sploh ne. Poleg tega je lahko stopnja spontanosti zelo težko določljiva, zlasti v javnem diskurzu.
- Formalnost ni kredibilen besedilnovrstni kriterij, ker se nanaša predvsem na izbiro jezikovnih sredstev, in za posamezne tipe govornih situacij je zelo težko vnaprej predvidevati te izbire, sploh ker na tem področju še ni bilo veliko raziskav v slovenskem jeziku.
- Namen in tematika besedil nista enostavno prenosljiva na govorjeni diskurz v slovenskem jeziku, saj sta bila pretežno klasificirana ali za tuje kulturno okolje ali za pisni diskurz.

Kot bistvena besedilnovrstna kriterija sta bila tako izbrana javnost diskurza in prenosnik. Njuna podrobnejša klasifikacija in izhodiščna razmerja v odstotkih so predstavljena v tabeli 2.

Tabela 2: Izhodiščna razmerja besedilnovrstnih kriterijev v odstotkih

kategorija	podkategorija	%
javnost	javni razvedrilni	20%
	javni informativno-izobraževalni	40%
	nejavni nezasebni	15%
	nejavni zasebni	25%
prenosnik	osebni stik	50%
	telefon	10%
	radio	20%
	televizija	20%

Za javni diskurz smo šteli tisti diskurz, ki je odprt za širšo javnost ali naslavlja veliko skupino ljudi, vsak drugi diskurz smo šteli za nejavnega.

Nejavni diskurz smo nadalje ločili na zasebni diskurz, tj. diskurz v zasebnem življenju posameznikov (v okviru družine, prijateljev, znancev). Nejavni nezasebni diskurz vključuje različne uradne in poluradne diskurze (v uradih, trgovinah, ob storitvah, v profesionalnem življenju ipd.).

V javnem diskurzu smo ločili medijske vsebine razvedrilnega programa in drug javni diskurz, katerega namen je predvsem razvedrilen (razvedrilni javni diskurz), ter medijske vsebine informativnega/izobraževalnega/kulturnega programa in drug javni diskurz, katerega namen je predvsem informativen/izobraževalen/socialen (nerazvedrilni oz. informativno-izobraževalni javni diskurz – sem sodi tudi večina diskurza v okviru različnih stopenj šolanja, javna predavanja ipd.).

2.1.4.3 Dodatni kriteriji za zajemanje šolskega diskurza

V skladu s cilji Gosa znaten del korpusa zajema šolski diskurz pri pouku v osnovni in srednji šoli. Ta del je zato dodatno specificiran po kriterijih, predstavljenih v tabeli 3. V besedilnovrstni klasifikaciji zgoraj je zajet v kategorijah javni nerazvedrilni oz. informativno-izobraževalni diskurz, prenosnik je osebni stik. Kriteriji so uravnani glede na podatke MŠŠ in SURS.

Tabela 3: Izhodiščna razmerja za šolski diskurz v odstotkih

kategorija	podkategorija	dodatna podkategorija	%
stopnja šolanja	osnovna šola	2. triletje	27,5%
		3. triletje	27,5%
	srednja šola	gimnazija	18%
		nižje in srednje poklicno, srednje strokovno in poklicno-tehniško izobraževanje	27%
regija	JZ		60%
	SV		40%
učni predmet	naravoslovni in tehnični predmeti		50%
	družboslovni in humanistični predmeti		50%

2.1.4.4 Skupna shema kriterijev za zajemanje gradiva

Tabela 4 povzema predstavljene kriterije v skupni preglednici.

Tabela 4: Shema predvidene sestave Gosa – celoten korpus

tip diskurza	%	podtip diskurza	%	prenosnik	%	regija	%
javni diskurz	60%	informativno-izobraževalni	40%	tv*	10%	SV	4%
						JZ	6%
				radio**	10%	SV	4%
						JZ	6%
				osebni stik	20%	SV	8%
						JZ	12%
		razvedrilni	20%	tv*	10%	SV	4%
						JZ	6%
				radio**	10%	SV	4%
						JZ	6%
							60,00%

tip diskurza	%	podtip diskurza	%	prenosnik	%	regija	%
nejavni diskurz	40%	nezasebni	15%	telefon	5%	SV	2,00%
						JZ	3,00%
				osebni stik	10%	SV	4,00%
						JZ	6,00%
		zasebni	25%	telefon	5%	SV	1,25%
						MB	0,75%
						JZ	1,75%
						LJ	1,25%
				osebni stik	20%	Italija	1,00%
						Avstrija	1,00%
						Madžarska	1,00%
						neslovenci	2,00%
						SV	3,75%
						MB	2,25%
						JZ	5,25%
						LJ	3,75%
							40,00%

* Najbolj gledani televizijski programi in oddaje.

** Najbolj poslušane radijske postaje in vsebine po posameznih slovenskih regijah.

Iz zasnove vidimo, da je predvidena delitev na JZ in SV regijo tudi pri medijskem diskurzu. Zajem gradiva je pokazal, da v resnici nekateri najbolj gledani in poslušani medijski programi pokrivajo celotno Slovenijo, zato je bila shema, predstavljena v tabeli 4, nekoliko popravljena, kot prikazuje tabela 5 – v televizijske vsebine so bili zajeti samo programi, ki so v času zajemanja gradiva oddajali po celotni Sloveniji, pri radijskih vsebinah pa smo tovrstnim programom namenili skupno 4% korpusa, po 8% pa radijskim postajam s področja SV in JZ Slovenije.

Tabela 5: Popravljena shema predvidene sestave Gosa – javni diskurz

tip diskurza	%	podtip diskurza	%	prenosnik	%	regija	%
javni diskurz	60%	informativno-izobraževalni	40%	tv	10%	celotna Slovenija	10%
				radio	10%	celotna Slovenija	2%
						SV	4%
						JZ	4%
				osebni stik	20%	SV	8%
						JZ	12%
		razvedrilni	20%	tv	10%	celotna Slovenija	10%
				radio	10%	celotna Slovenija	2%
						SV	4%
						JZ	4%
skupaj							60,00%

2.1.4.5 Toleranca odstopanj

Zgoraj navedena razmerja med kriteriji so »idealna« in v končnem izdelku so seveda nekoliko drugačna, saj teh kriterijev pri snemanju avtentičnih diskurzov (kar je eden osrednjih ciljev Gosa in pomembnejša zahteva kot popolna demografska in besedilnovrstna uravnotežitev) ni mogoče popolnoma nadzorovati. Zato je bila sprejeta odločitev, da korpus izpolnjuje zastavljene kriterije, če so ti uresničeni v okviru 30-odstotnega odstopanja relativno, ter zelo dobro izpolnjuje zastavljene kriterije, če so uresničeni v okviru 10-odstotnega odstopanja relativno. Prav tako štejemo, da korpus izpolnjuje zastavljene kriterije, če samo določen del gradiva vključuje podatke, na podlagi katerih je mogoče preveriti pokritost definiranih kriterijev, in ta del korpusa izkazuje ustrezno zastopanost.

Kako so bili zastavljeni kriteriji v korpusu Gos dejansko realizirani, je razvidno iz statistik v poglavju 3 in v prilogi.

2.1.5 Avtorske pravice in varovanje osebnih podatkov

Korpus Gos vključuje posnetke in transkripcije posnetkov iz medijev (radio, televizija), posnetke šolskega pouka in predavanj ter posnetke nejavnih pogovorov. V zvezi s tem je bilo treba slediti veljavnim pravnim zahtevam po varovanju gradiva in morebitnih osebnih podatkov ter spoštovanju avtorskih pravic.

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) ter z Zakonom o varstvu dokumentarnega in arhivskega gradiva ter arhivih (ZVDAGA). 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 avtorski in sorodnih pravicah (ZASP) in podzakonskimi akti.

Avtorske pravice za gradiva, prejeta od medijev, so bile v okviru konzorcija projekta Sporazumevanje v slovenskem jeziku urejene s pogodbo o odstopu avdio in video posnetkov radijskih in televizijskih

oddaj ali drugih programskih vsebin. Prenos pravic je urejen v 4. členu, ki določa:

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.

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 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.

Vprašanje varovanja osebnih podatkov je aktualno v zvezi z gradivom, saj se ti pojavljajo tako v zvočnih posnetkih kot v zbranih obrazcih s podatki o posnetkih in govorcih.

Ob zajemanju gradiva za Gos so bili s posebnim obrazcem zbrani podatki o spolu, starosti, izobrazbi, regionalni(h) pripadnosti(h) in prvem jeziku govorcev. Od tega je podatek o prvem jeziku po ZVOP-1 občutljivi osebni podatek. Vsi podatki so v Gos vključeni tako, da so anonimizirani, kar pomeni, da jih ni več mogoče povezati s posameznikom.

Nadalje je konzorcij Sporazumevanje v slovenskem jeziku v skladu z veljavno zakonodajo zagotovil varovanje osebnih podatkov in gradiva v transkripcijah posnetkov in samih posnetkih. Identiteta govorcev iz transkripcij in spremljajočih metapodatkov gradiva namreč ni več razvidna, saj so govorci poimenovani s šiframi, morebitni podatki o samih govorcih (tudi pri javnem diskurzu), ki se lahko pojavljajo v posnetkih in transkripcijah, pa so v zapisu govora anonimizirani. Tako je označena samo vrsta podatka: [ime], [priimek] itd., tudi [podjetje], če je mogoče prek imena (manjšega) podjetja iz konteksta razbrati identiteto oseb. Imena oseb, podjetij, institucij ali druga imena, prek katerih bi lahko razbrali identiteto oseb, o katerih je govora, so zakrita tudi v primerih žaljive konotacije govora.

Varovanje osebnih podatkov v avdio posnetkih je zagotovljeno tako, da so vsa mesta v govoru, kjer se pojavljajo podatki o govorcih, ki so v transkripciji anonimizirani, prekrita s piskom. Identiteto govorcev lahko razkrije tudi govorčev glas, zato je v nemedijskih diskurzih nekoliko spremenjena tudi frekvenca posnetkov.

Vsi posamezniki, ki so imeli dostop do katerihkoli datotek ali dokumentov, iz katerih je še bila razpoznavna identiteta govorcev, so podpisali 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.

2.2 Zajemanje gradiva na terenu

Zajemanje avtentičnih posnetkov pogovorov na terenu v obsegu, predvidenem za korpus Gos, zahteva ustrezno organizacijo in koordinacijo dela. Najti je bilo treba rešitve za naslednje osrednje probleme zajemanja gradiva:

- izbira snemalne opreme in karakteristike posnetkov,
- sestava ekipe za snemanje in prenos snemalne opreme med kraji snemanja,
- delo z datotekami na daljavo,
- pridobivanje posnetkov in soglasij za snemanje.

2.2.1 Izbira ustrezne snemalne opreme in karakteristike posnetkov

Za snemanje terenskih posnetkov pogovorov v osebnem stiku je bil izbran snemalnik Samson Zoom H4 z naslednjimi karakteristikami:

- 2-sledno snemanje, stereo,
- 4-sledno snemanje ob priključitvi dodatnih mikrofonov,
- X/Y vzorec stereo mikrofonov,
- mp3-snemanje do 320 kbps,
- wav-snemanje do 24 bitov/96 kHz,
- avdio in podatkovni vmesnik USB,
- XLR-priključka s fantomskim napajanjem,
- razširitvena reža za spominske kartice SD.

Pri snemanju je bil dodatno uporabljen trinožni podstavek za snemalnik, zato da ta nikoli ni bil v neposrednem stiku s podlago. Izbrani snemalnik je zadostoval za snemanje vseh vrst diskurzov v osebnem stiku (tudi javnih: šolske ure, predavanja). Snemanje je potekalo v glavnem v formatu mp3, 256 kbps, 44 kHz, stereo.

Za snemanje telefonskih pogovorov je bil izbran mobilni telefon Nokia N78 z dodatno nameščeno aplikacijo za snemanje pogovorov. Telefon je omogočal snemanje v formatu wav, 128 kbps, 8 kHz, mono.

Vsi posnetki so bili ob urejanju pretvorjeni v enoten format wav (Windows PCM), 256 kbps, 16 kHz, mono.

2.2.2 Sestava ekipe za snemanje in prenos snemalne opreme med kraji snemanja

Pri snemanju korpusa Gos je bilo treba zajeti veliko različnih tipov diskurzov in različne govorce, razpršene po celotnem geografskem območju, kjer se govori slovenski jezik (vključno z zamejstvom). Za izvedbo tega je bila potrebna velika ekipa snemalcev (ob koncu snemanja jih je bilo skupaj okrog 30). Zelo pomemben faktor pri izbiri snemalcev je, da izhajajo iz različnih slovenskih regij, saj najlažje vsak v svojem domačem okolju posnamejo potrebno gradivo. Ob praktični izvedbi se je še pokazalo, da lahko posamezen snemalec priskrbi povprečno dva do tri posnetke. Redki, zlasti taki s široko socialno mrežo in komunikacijskimi sposobnostmi, pa so uspeli zagotoviti tudi do pet, osem ali deset posnetkov.

Snemalci so lahko prevzeli snemalno opremo na točki Gosa v Mariboru ali Ljubljani, jo imeli pri sebi teden, dva ali tri (odvisno od potreb in razpoložljivega časa) in v tem času zagotovili posnetke iz regije, iz katere izhajajo. Za ekonomično izvedbo snemanja je zato zaželeno, da so snemalci osebe, ki pogosto potujejo med tema središčema

in domačim krajem – to je tipično študentska populacija. Druga možnost je še uporaba poštnih storitev za migracijo snemalne opreme, vendar v tem primeru ni vedno možnosti, da dobi snemalec pred snemanjem osebne inštrukcije o namenu snemanja, okoliščinah snemanja in uporabi snemalne opreme, če je to njegovo prvo snemanje.

2.2.3 Delo z datotekami na daljavo

Posnete zvočne datoteke so bile običajno velike nekaj deset MB, zato prenos prek e-pošte ni mogoč. Kljub temu pa je zaradi geografske razpršenosti sodelavcev pri korpusu ekonomično, da se datoteke prenašajo na daljavo (zlasti ker so prenosi med različnimi osebami potrebni tudi za urejanje, transkribiranje posnetkov ipd.). V ta namen je bil vzpostavljen FTP-strežnik, ki je omogočal izvajanje vseh del, povezanih z izdelavo korpusa, na daljavo. Delo z FTP-strežnikom je bilo organizirano v natančno določeno strukturo sodelavcev, njihovih nalog in pravic, in sicer so bile potrebne naloge za celotno izvedbo del, povezanih s korpusom, naslednje: snemalec, urejevalec posnetkov, ocenjevalec posnetkov, transkriptor 1 (pogovorni zapis), kontrolor transkripcij 1, validator transkripcij 1, transkriptor 2 (standardizacija zapisa), administrator. FTP-strežnik je tudi omogočal, da so bile vse korpusne datoteke vedno na enem mestu, v jasni strukturi in da so se sproti izdelovale potrebne varnostne kopije.

2.2.4 Pridobivanje posnetkov in soglasij za snemanje

Pridobivanje posnetkov in soglasij je potekalo na različne načine, odvisno od tipa diskurza.

Za pridobitev posnetkov javnega medijskega diskurza je potekala komunikacija na institucionalni ravni, med vodjo projekta Sporazumevanje v slovenskem jeziku in izbranimi televizijskimi in radijskimi postajami. Posnetki so bili vzeti iz arhivov teh postaj.

Pri šolskem diskurzu je tekla komunikacija na institucionalni ravni, občasno pa je bila predhodno vzpostavljena tudi komunikacija s posameznimi učitelji. Če so bila zagotovljena potrebna soglasja s strani učiteljev in šole, je nato s posredovanjem šole steklo še obveščanje staršev o snemanju in pridobivanje njihovih soglasij, šele nato je lahko bilo izvedeno snemanje.

Za pridobitev posnetkov javnih predavanj, tečajev, delavnic in podobnega je najprej stekel kontakt med vodjo projekta Sporazumevanje v slovenskem jeziku in dotičnim predavateljem. Po predhodnem soglasju predavatelja je moral snemalec na predavanju o snemanju ob-

vestiti še publiko in zagotoviti pisna soglasja tistih oseb iz publike, ki so slišne s svojim govorom na posnetku.

Pri snemanju zasebnih posnetkov v osebnem stiku in zasebnih telefonskih posnetkov je moral snemalec sam osebno zagotoviti vnaprejšnje pisno soglasje oseb, ki so bile udeležene v posnetem pogovoru.

Kot najtežavnejše se je pokazalo snemanje nejavnih nezasebnih diskurzov (sestanki, konzultacije, pogovori ob raznih storitvah, kupovanju, posredovanje informacij, svetovanja itd.). Pri snemanju v osebnem stiku je bil najuspešnejši način, da je snemalec sam sodeloval v takem pogovoru in si osebno zagotovil vnaprejšnje soglasje sogovornika za snemanje. Poskusi pridobivanja soglasij na institucionalni ravni pri tem tipu diskurza niso prinesli uspeha, je pa bila včasih potrebna podpora komunikacije na institucionalni ravni po predhodnem dogovoru z osebami, ki bi bile udeležene v pogovoru. Še bolj zapleteno je bilo pridobivanje soglasij za nejavni nezasebni telefonski diskurz: tudi tam je bil na eni strani snemalec hkrati eden od sogovornikov v diskurzu, za sogovornika na drugi strani pa je bilo treba najprej pridobiti soglasje na institucionalni ravni (z organizacijo, v okviru katere je potekal pogovor: trgovina, svetovalna pisarna, turistična pisarna, trženje, klicni center...), nato pa s pomočjo organizacije še soglasje govorca, ki je na posnetku. Uporaba obstoječih posnetkov klicnih centrov in drugih organizacij, ki snemajo pogovore s svojimi strankami, ni bila mogoča, saj je po obstoječi zakonodaji potrebna predhodna informirana privolitev posameznika za takšno uporabo njegovega posnetka.

2.3 Specificiranje načel transkribiranja gradiva

V tej sekciji opisujemo priprave na transkribiranje gradiva.

2.3.1 Obstoječi standardi in prakse transkribiranja

Od mednarodnih standardov sta bila upoštevana standarda EAGLES in TEI.

Evropska iniciativa EAGLES[33] je nastala leta 1993 na pobudo Evropske komisije z namenom, da pospeši oblikovanje skupnih standardov za izdelavo obsežnih jezikovnih virov. Pri izdelavi priporočil za govorjena besedila (EAGLES, 1996) upoštevajo obstoječe prakse (priporočila TEI, NERC, SpeechDat, tradicijo konverzacijske analize idr.) in skušajo najti skupne elemente različnih tradicij transkribiranja.

33 Expert Advisory Group on Language Engineering Standards; http://www.ilc.cnr.it/EAGLES/home.html.

V povzetku priporočajo označevanje naslednjih elementov: glasovnih polleksikalnih (*eee*, *mhm*, *aha* itd.) in neleksikalnih (smeh, cmokanje, kašljanje, kihanje, jok, zehanje itd.) enot, negovornih nekomunikacijskih dogodkov, identitete govorca, menjavanja govorcev, hkratnega govora, izpustitev pri branih besedilih, samopopravljanj, besednih fragmentov in nerazumljivih fragmentov.

Pri zapisu govora ločujejo tri ravni:
- S1 – ortografska predstavitev besedila,
- S2 – fonemska predstavitev besed v citatni obliki (tj. v obliki, kot so besede izgovorjene v izolaciji),
- S3 – fonetična transkripcija, ki predstavlja dejansko glasovno podobo izjave.

Za ortografski zapis priporočajo, da je narejen v standardni (tj. knjižni) normi in da so vse enote (tudi okrajšave, številke, črkovanja ipd.) polno izpisane. To priporočilo temelji na predpostavki, da obstaja možnost avtomatske povezave med ortografskim (S1) in fonemskim (S2) zapisom. Za fonemski in fonetični zapis priporočajo uporabo fonetične abecede SAMPA oz. X-SAMPA[34]. Na prozodični ravni priporočajo vsaj označevanje izjav in premorov.

TEI[35] je organizacija, katere namen je definiranje standardov za kodiranje besedil. Njihova priporočila[36] v posebnem (osmem) poglavju obsežno obravnavajo tudi transkribiranje govora. Priporočila EAGLES-a že upoštevajo bistvena priporočila TEI.

Glede samega kodiranja transkripcij pa je pomembno izpostaviti, da so standardi TEI izraženi v danes široko uporabljanem kodifikacijskem jeziku XML, čeprav sicer kodna shema TEI ni odvisna od tega jezika.

V slovenskem prostoru zasledimo tri usmeritve bistveno različnih praks transkribiranja daljših besedil:

a) V slovenski dialektologiji se običajno uporablja t. i. tradicionalna slovenska fonetična transkripcija oz. fonetična transkripcija OLA (Slovanski lingvistični atlas) z dodanimi različnimi diakritičnimi znaki, npr. krožec pod zvočnikom za silabem, pika pod e ali o za ozki izgovor vokala, strešica pod e ali o za nevtralni izgovor, e in o brez diakritičnega znamenja za široka vokala, dvopičje za črko za dolgi vokal itd. (Zorko, 1995).
b) V besediloslovnih in pragmatičnih raziskavah govorjene slovenščine se večinoma uporablja ortografski zapis govorjenega jezika. Skupnega standarda pri tem sicer še ni, na podlagi zapisov govora in besedilnih fragmentov nekaterih avtorjev transkripcij (npr. Kranjc, 1999; Krajnc, 2005; Smolej, 2006) pa lahko ugotavljamo naslednji temeljni skupni značilnosti:

34 http://www.phon.ucl.ac.uk/home/sampa/
35 Text Encoding Iniciative; http://www.tei-c.org/index.xml
36 TEI, 2007; http://www.tei-c.org/Guidelines/index.xml

- Uporabljen je slovenski knjižni črkopis brez dodatnih posebnih znakov za različne glasove (npr. za polglasnik, dvoustnični u, diftonge ipd.). Izjemoma najdemo dodan poseben znak za nekatere rabe polglasnika (npr. *sevȍda*; Krajnc, 2005).
- Zapis ponazarja pojave moderne vokalne redukcije in drugih pogovornih in narečnih prvin govorjene slovenščine (npr. *maš čevle, to je zloml, jes mam, notr držim, hitr kuple, beu avto, neki, k bi lohk, je poneso, ceno svojih živlen, tko je blo, maš polhen kufer, mislm*). Seveda se pri različnih avtorjih pojavljajo tudi nekoliko različne rešitve pri zapisu nekaterih pogovornih oblik (npr. *jes* proti *js* za *jaz*, *neč* proti *nč* za *nič*, različni zapisi polglasnika v vlogi diskurznega označevalca: *ee, {eee}, ə* itd.).
  Podoben zapis se zadnji čas vse širše uporablja v spletnih forumih, klepetalnicah, blogih in drugih zapisih, ki jih uporabniki objavljajo v spletu, kot tudi v nekaterih literarnih delih. Tako npr. pišejo *čeprov, niti najmnj, se nej, pusti pr mer, nč ne rečm, tut js, kšni prjatlci, s kero kol, pejt spat, js sm zlo, odpravt* (http://www.cveka.com, 22. 12. 2008) oz. *najraj b vidu, če bse mi, čeb šu sam pa bmene pustu, tud jest, čeu pa ta, mi nau hotu povedat, ker neb mel* (Welsh, 1997); *razmete, rejsan, nej pred sebo, pred bougon, san ges živa, tou ka znam, pejneze, fkraj, z menof* (Šarotar, 2007).

c) V jezikovnotehnološki praksi (Žgank et al., 2004; 2006; Zemljarič Miklavčič, 2007; 2008) se uporablja v glavnem poknjižen zapis govorjenega jezika, iz katerega niso več vidni pogovorni in narečni pojavi, kot je npr. moderna vokalna redukcija. Dopuščeno je le omejeno število določenih najpogostejših odstopanj, npr. kratki nedoločnik, *k* za veznike *ki, ko, ker, pol* v pomenu *potem* itd. V tej smeri je nastal tudi poskus definiranja pravil za transkribiranje govornih korpusov (Zemljarič Miklavčič, 2007). Verdonik (2006) pa poleg poknjiženega zapisa dodaja fonetični zapis v abecedi SAMPA, vendar samo v primerih, ko določena besedna oblika ne sovpada s knjižno normo.

O tem, katere vrste podatkov in kako podrobne podatke o samem diskurzu in govoru vključujejo transkripcije, so se avtorji odločali različno, glede na cilje in potrebe posameznih transkripcij.

2.3.2 Izhodišča za definiranje pravil transkribiranja

Zaradi obsega dela (milijon besed) in zaradi zagotavljanja homogenosti transkripcij je cilj, da so transkripcije kar se da enostavne in skrčene na zapisovanje za namene Gosa najnujnejših podatkov, zato:

- vključujejo najpomembnejše kontekstne informacije, zlasti vse tiste, ki so pridobljene ob samem zbiranju gradiva in so pomembne kot potencialni iskalni pogoj,
- vključujejo besedilnovrstno razvrstitev diskurzov, ki bo omogočala omejevanje iskanja na različne tipe govora,
- vsebujejo samo pragmatično[37] najpomembnejše informacije o strukturi diskurza,
- omogočajo čim hitrejše transkribiranje,
- kar najbolj nazorno predstavljajo dejansko govorjeno podobo diskurza,
- omogočajo avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami,
- vsebujejo samo pragmatično najpomembnejše informacije o nebesednih in nejezikovnih zvokih, ki so pomembni za uporabnikovo boljše razumevanje poteka diskurza.

37 Za pragmatično pomembne štejemo tiste dogodke, ki imajo vidnejši učinek na diskurz, pri čemer mislimo tako jezikovno rabo kot socialna razmerja in kognitivno dimenzijo.

Gos je zasnovan predvsem kot jezikovni govorni vir, zato transkripcije ne vključujejo informacij o kretnjah, mimiki, gestah in drugih nezvočnih spremljevalnih kanalih govornega sporazumevanja.

Pravila transkribiranja so bila v skladu z zgoraj opisanimi cilji definirana na naslednjih ravneh:

- podatki o govorcih,
- podatki o diskurzu,
- struktura diskurzov,
- zapis govora,
- nebesedni in nejezikovni zvoki ter prozodija.

Ta pravila so podrobneje predstavljena v poglavju 3 pri opisu gradiva in podatkov. Dodatna pojasnila so potrebna samo v zvezi z odločitvami o zapisu govora, in sicer:

Za namene jezikovnotehnološke uporabe korpusa je zelo priporočljivo, da je govor zapisan v knjižni normi, kot izhaja tudi iz jezikovnotehnološke prakse transkribiranja govora v slovenščini. Vendar na ta način izredno popačimo resnično jezikovno podobo, v kateri je veliko redukcij glasov ter neknjižnih besednih oblik in besed, tako da iz samega zapisa uporabnik dobi zelo nenatančen vtis o diskurzu ter nepopoln oz. napačen vtis o jezikovni podobi. Zato smo se odločili za dva nivoja zapisa: pogovornega in standardiziranega.

Pogovorni zapis sledi smernicam transkribiranja govora, ki se oblikujejo v besediloslovnih in pragmatičnih raziskavah, ter praksi pogovornega pisanja v nekaterih spletnih in leposlovnih besedilih. Tak zapis poteka veliko hitreje, kot bi potekal fonetični zapis z abecedo SAMPA, in uporabnikom korpusa na enostaven način in v poznanem črkopisu predstavi govor. Tak način zapisovanja tudi omogoči za cilje

Gosa zelo pomembne raziskave najbolj tipičnih neknjižnih besednih oblik in njihovih oblikoskladenjskih vlog. Osrednje vodilo za zapisovanje pogovornega zapisa je: govor zapisujemo v veljavnem slovenskem črkopisu in upoštevamo veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pa pri tem kolikor mogoče zvesto predstavimo glasovno podobo govora.

Zatem je transkripciji dodan standardizirani zapis, katerega osrednji namen je izboljšati in razširiti korpusne iskalne možnosti. Osrednje vodilo standardizacije zapisa je: pri pretvorbi pogovornega zapisa v standardizirani zapis odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki. Izhodišče je knjižna različica istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Za ločevanje, kdaj gre za glasovno premeno in kdaj ne, se ob primerih oblikujejo načela dobre prakse. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru.

2.3.3 Izbira orodja za transkribiranje

Orodje za transkribiranje mora biti uporabniško prijazno za transkribiranje, hkrati pa mora podpirati:

- dolge zvočne datoteke,
- šumnike,
- vnos metaoznak za opis različnih pragmatičnih, akustičnih in drugih dogodkov,
- vnos podatkov o govorcih, zajetih s popisnim obrazcem,
- vnos besedilnovrstnih podatkov o diskurzu.

Obenem mora orodje za transkribiranje omogočati:

- segmentacijo zvočnega signala na poljubne enote na enostaven način in
- enostavno vnašanje informacij o povezavi med posameznimi segmenti in pripadajočimi deli transkripcije.

Obstaja več orodij, veliko tudi prosto dostopnih, ki jih lahko uporabimo za transkribiranje zvočnih posnetkov, vendar nobeno ni bilo narejeno specifično za namene, kot jih imamo pri gradnji Gosa. V literaturi najdemo tudi nekaj primerjav in vrednotenj različnih orodij.

Garg et al. (2004) med drugim primerjajo naslednja orodja za transkribiranje: Praat, Transcriber, TASX in Anvil. Zaključijo, da je za splošno transkripcijo najprimernejši Transcriber.

Rohlfing et al. (2006) primerjajo prednosti in slabosti orodij za multimodalno označevanje avdio in video posnetkov: Anvil, ELAN, Exmaralda, TASX in MacVista. V zaključku ugotavljajo, da je bilo vsako od naštetih orodij razvito v različne namene, zato je tudi izbor orodja odvisen predvsem od potreb. Od naštetih so za naše cilje potencialno

primerni ELAN, Anvil in Exmaralda. Orodij, ki so izdelana samo za transkribiranje, ne vključijo v primerjavo.

Zemljarič Miklavčič (2007) podrobneje primerja Transcriber, Praat in WinPitch. Opozarja na pomanjkljivost Transcriberja, da ne omogoča hkratnega zapisa govora več kot dveh govorcev – če torej hkrati govorijo trije ali več govorcev, lahko zapišemo samo govor dveh. Za Praat ugotavlja, da te pomanjkljivosti nima, hkrati pa navaja, da je po njenih izkušnjah »v najboljšem primeru še mogoče transkribirati besedilo, ki ga hkrati izrečejo trije govorci; če govorijo več kot trije naenkrat, je običajno mogoče transkribirati samo posamezne fragmente iz posameznih izjav« (Zemljarič Miklavčič, 2007: 137). Tudi verzija WinPitchPro programa WinPitch omogoča transkribiranje, vendar v primerjavi s Transcriberjem in Praatom ni prosto dostopen. Avtorica sklene, da se za ortografsko transkribiranje zdita najprimernejša programa Transcriber in Praat.

V slovenskem prostoru že obstaja nekaj specializiranih govornih korpusov, za izdelavo katerih je bilo uporabljeno transkripcijsko orodje: v glavnem je bil to Transcriber (za baze BNSI Broadcast News (Žgank et al., 2004), Broadcast News Speech Database (Žibert, Mihelič, 2004), Turdis (Verdonik, Rojc, 2006), Sloparl (Žgank et al., 2006)), v enem primeru pa Transcriber in Praat (Zemljarič Miklavčič, Stabej, 2005; Zemljarič Miklavčič, 2006).

Pred izborom smo še sami testno preizkusili najbolj obetavna orodja za transkribiranje: Transcriber, Praat, Exmaralda in ELAN.

ELAN je namenjen predvsem označevanju zvočnih in video posnetkov, za uporabo pri zapisovanju govora pa ni preveč uporabniško prijazen. Prav tako ni posebej uporabniško prijazen za določanje in spreminjanje mej segmentov/izjav. Vnašanje podatkov o govorcih in diskurzih je mogoče samo v obliki sledi. Po naši oceni je orodje bolj kot za samo transkribiranje govora primerno za označevanje pri gradnji multimodalnih govornih baz, kjer so označene tudi geste, mimika, kretnje ipd. Posebnost ELAN-a je, da omogoča med drugim uvoz datotek, zapisanih s programom Transcriber.

Exmaralda je edino od navedenih orodij, ki podpira vnos poljubnih podatkov o govorcih in diskurzih, kar je gotovo njegova pozitivna lastnost za naše namene. Osrednja pomanjkljivost pa je slaba povezava med transkripcijo in zvočnim posnetkom: funkcije za pomikanje po zvočnem signalu so premalo natančne, kar lahko bistveno upočasni transkribiranje, vzpostavljanje povezave med posameznimi segmenti in pripadajočimi deli transkripcije pa je s tem orodjem zamudno. Zaradi tega Exmaralda ni posebej primerna za naše namene.

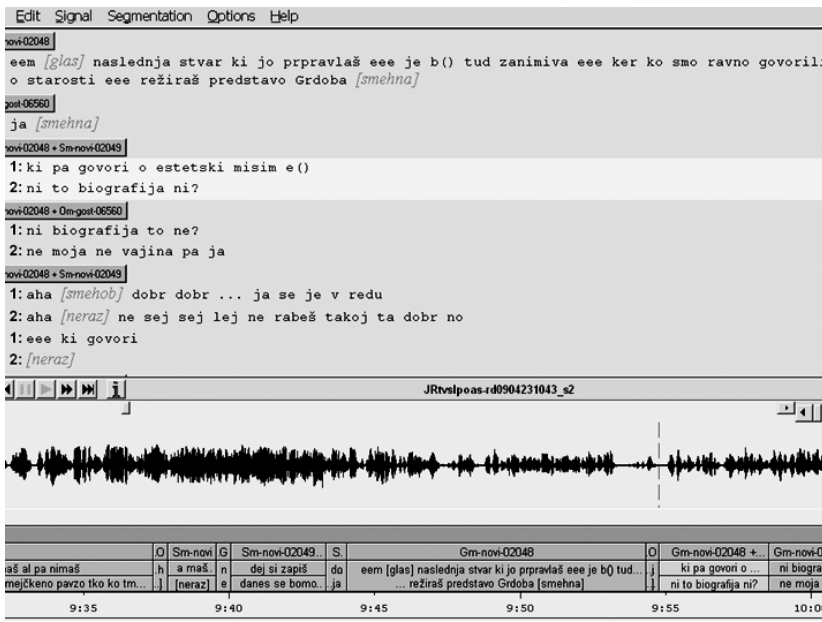
Praat je v osnovi namenjen akustičnim analizam, vendar omogoča tudi transkribiranje daljših zvočnih datotek. Njegova prednost v primerjavi s Transcriberjem je, da omogoča zapis hkratnega govora več kot dveh govorcev. Tudi predvajanje zvočnega posnetka

in segmentiranje signala na manjše enote je enostavno. Glavna pomanjkljivost je, da zaradi številnih funkcij akustične analize ni najbolj uporabniško prijazen, če ga želimo uporabljati samo za transkribiranje. Prav tako ne podpira vnosa podatkov o govorcih in diskurzih ter vnosa metaoznak. V nasprotju z ostalimi programi Praatova izhodna datoteka ni v XML-formatu, ampak v posebni Praatovi skripti (ki pa je sicer široko podprta v drugih transkripcijskih orodjih).

Transcriber je za naše namene najbolj uporabniško prijazen in tudi po svoji zasnovi najbližji: narejen je bil namreč za namene transkribiranja televizijskih informativnih oddaj. Dodatna prednost je, da v slovenskem prostoru že obstaja praksa njegove uporabe. Povezava med zvočnim posnetkom in transkripcijo je izredno dobra, po zvočnih posnetkih se enostavno premikamo in posnetek enostavno segmentiramo. Orodje tudi podpira vnos metaoznak v transkripcijo, prav tako vnos nekaterih podatkov o govorcih in diskurzu, čeprav za naše namene nekoliko preveč omejeno. Njegova največja pomanjkljivost pa je, da ne omogoča zapisa hkratnega govora več kot dveh govorcev.

Potem ko smo pretehtali prednosti in slabosti testiranih orodij, smo se odločili, da naše cilje najbolje izpolnjuje Transcriber. Slika 1 prikazuje okno Transcriberja s transkripcijo. V spodnjem delu okna je prikazan zvočni signal, nad njim so ikone za navigacijo po zvoku. V zgornjem delu okna transkriptor zapisuje govor. Polja pravokotne oblike označujejo menjavo govorcev, pike pa izjave/segmente.

Slika 1: Transcriber



Za vnos podatkov o govorcih in diskurzu smo uporabili posebne obrazce, pripravljene v Excelu, ki smo jih nato shranili kot tekstovno datoteko. Za zapis govora tretjega hkratnega govorca smo določili poseben niz znakov, ki omogoča naknadno avtomatsko razširitev izhodne XML-kode Transcriberja z vrstico za tretjega hkratnega govorca. Govora četrtega in ostalih hkratnih govorcev nismo zapisovali. Tudi za standardizacijo zapisa govora smo uporabili Transcriber in naredili ta zapis v posebno trs-datoteko (tj. izhodna Transcriberjeva datoteka), ločeno od pogovornega zapisa. Delovni kodni format je bil CP1250.

Po končanem postopku transkribiranja smo tako dobili več datotek za isti posnetek, kar je predstavljalo delovno končno verzijo korpusa. Iz te verzije smo pripravili posebno datoteko, v kateri so vse transkripcije korpusa z vsemi spremljajočimi podatki zapisane po XML-shemi, ki sledi standardu TEI, za končne uporabnike. Za potrebe delovanja spletnega konkordančnika korpusa Gos so bile delovne datoteke uporabljene za izdelavo SQL-baze. Gradiva korpusa Gos so podrobneje predstavljena v poglavju 3.

2.4 Izvedba transkribiranja in zagotavljanje kvalitete

Za izdelavo korpusa Gos je bilo treba transkribirati okoli 7000 minut (približno 115 ur) govora, in to na dva načina: v pogovornem in standardiziranem zapisu. To pomeni veliko količino enoličnega ročnega dela. Na izvedbo transkribiranja vpliva tudi dejstvo, da se zlasti v zasebnih pogovorih pojavljajo tudi posnetki s posameznih geografskih območij ali določenega tipa govorcev (zlasti starejših), ki jih lahko v prvem koraku kvalitetno zapiše samo oseba, ki dobro pozna govor tega območja, ali še bolje kar oseba, ki je pogovor posnela. Zaradi navedenega je bilo skupno število transkriptorjev na osnovnem nivoju kar veliko, 20, s tem da so mnogi od teh transkribirali predvsem posnetke, ki so jih sami posneli. Povprečno je transkribiranje dveh minut posnetka trajalo eno uro, torej je bilo samo za osnovni pogovorni zapis potrebnih okrog 3500 ur dela.

Posebna pozornost pri izdelavi osnovne transkripcije s pogovornim zapisom je bila glede na veliko število transkriptorjev namenjena zagotavljanju kvalitete. V ta namen je transkribiranju sledilo dvostopenjsko pregledovanje: najprej kontrola transkripcij, ki so jo izvajali štirje najbolje usposobljeni transkriptorji, pri čemer so natančno pregledali celotno transkripcijo in popravili napake ali pa transkripcijo zavrnili, če ni ustrezala zahtevani kvaliteti, da jo je transkriptor popravil, in jo nato ponovno prekontrolirali. Povprečno je kontroliranje sedmih minut posnetka potekalo eno uro.

Prekontrolirane transkripcije je v tretjem koraku pregledal validator, ki je vnesel morebitne manjše popravke oz. po potrebi opozoril na transkripcije preslabe kvalitete, da jih je kontrolor transkripcij ponovno popravil. Validator je bila samo ena, visoko usposobljena oseba.

Tudi standardizacijo zapisa je izvajala ena sama, visoko usposobljena oseba, ki je hkrati sodelovala tudi pri izdelavi načel za standardizacijo zapisa. S tem smo skušali doseči kar najbolj usklajen način standardizacije zapisa in njegovo visoko kvaliteto. Standardizacija je potekala tako, da je bila transkripcija v pogovornem zapisu najprej delno avtomatsko predprocesirana, nato pa je bil samo zapis (in noben drug element transkripcije) ročno popravljen tako, da je ustrezal načelom standardizacije. Ročni del standardizacije zapisa je izvajala ista oseba, ki je validirala transkripcije. Pri tem se je občasno pokazala potreba po manjših popravkih pogovornega zapisa, tako da je del validacijskih popravkov transkripcij nastajal kar hkrati s standardizacijo zapisa.

Tako ob transkribiranju v pogovornem zapisu kot ob standardizaciji zapisa so se občasno odprla vprašanja, o katerih je ob koordinaciji administratorja diskutirala skupina, odgovorna za specifikacije korpusa.

Posamezne napake predvsem tehnične narave so se odkrile in odpravile še ob končnem procesiranju in uvozu datotek v uporabniške različice korpusa.

Enako kot pri snemanju je tudi v vseh fazah transkribiranja delo potekalo na daljavo, prenos datotek pa prek FTP-strežnika.

3 Gradiva korpusa Gos

V tem poglavju bomo predstavili končno sliko gradiva, ki smo ga vključili v referenčni korpus Gos. Pri tem se bomo osredotočili na ključne komponente gradiva:

- uresničitev posebnih kriterijev za zajem gradiv,
- besedilnovrstno in demografsko sestavo zajetih posnetkov,
- sistem označevanja podatkov o diskurzih in udeležencih,
- načela zapisa govora na dveh nivojih,
- zapis korpusa v XML-datoteki, namenjeni za zahtevnejše uporabnike.

3.1 Uresničitev posebnih kriterijev za zajem gradiv

Glede na zastavljene cilje korpusa Gos (gl. poglavje 2.1.2) smo opredelili posebne kriterije v zvezi s posameznimi vidiki zajemanja gradiv (gl. poglavje 2.1.2), ki smo jih uresničili v naslednjih točkah:

- Avtentična dolžina govorjenih diskurzov: ta kriterij je bil upoštevan pri medijskem diskurzu, telefonskih pogovorih, šolskem diskurzu, v veliki meri tudi pri nejavnem nezasebnem diskurzu. Pri zasebnih pogovorih in predavanjih, pa tudi nekaterih nejavnih nezasebnih diskurzih pa se je pokazalo, da je običajno naravna dolžina teh diskurzov predolga, da bi hkrati dosegli ustrezno razpršenost gradiva, zato so bili praviloma skrajšani na pol ure do največ tri četrt ure.
- Avtentični govorjeni diskurz: kriterij je bil precej zvesto upoštevan, izjemoma so bili pogovori izvršeni samo v namene snemanja le pri nekaterih nejavnih nezasebnih telefonskih posnetkih, kjer zaradi varovanja pravic govorcev pogosto ni bilo mogoče priti do povsem avtentičnih posnetkov.
- Pravno-etični vidiki: veliko oseb, ki smo jih prosili za snemanje, je prošnjo zavrnilo zaradi občutka nelagodja pred snemanjem. V nekaterih posnetkih je mogoče čutiti določeno stopnjo nelagodja govorcev zaradi snemanja, v drugih pa manj ali sploh ne.
- Tehnični vidiki: snemanju v šumnih okoljih smo se izogibali. Za diskurze, ki imajo formalen začetek (nejavni nezasebni pogovori, javni diskurz) oz. tehnično določen začetek (telefonski pogovori), je bilo treba vključiti snemalno napravo nekoliko pred začetkom diskurza, pri zasebnih pogovorih pa smo običajno začetni del posnetka, v katerem je pogosto še razlaga o namenu snemanja, izločili.

- Spontanost govora: zajeti diskurzi so prevladujoče spontani, noben ni v celoti osnovan na vnaprej pripravljeni pisni predlogi. Manjša spontanost govora se lahko pojavlja le v novinarskih prispevkih in nekaterih predavanjih, vendar tudi tovrstni diskurz nikoli ni v celoti vnaprej napisan.
- Govor otrok in mladoletnikov: govor mladostnikov se, pričakovano, pojavlja v šolskem diskurzu ter v posameznih družinskih posnetkih.
- Prvi jezik govorcev: govor tujih govorcev slovenščine je zajet v različnih situacijah (doma v okviru družine, doma s prijatelji, tečaj v jezikovni šoli, delovni sestanek itd.). Prvi jezik tujih govorcev je pogosto južnoslovanski. Celotnega spektra tujih govorcev slovenščine pri omejenem obsegu (2% korpusa oz. 130 minut govora) nismo mogli zajeti.
- Govor Slovencev v zamejstvu in po svetu: zajeli smo vzorce govora Slovencev, ki živijo v Italiji, v Avstriji in v Madžarski. Ker je bilo opredeljeno, da bo tega govora za vsako od teh dežel okrog 1% (tj. okrog 70 minut), to ni reprezentativni vzorec govorjene slovenščine zunaj meja Republike Slovenije, saj je v vsaki od teh dežel slovenščina tako raznolika in razpršena, da bi morali v tak namen zajeti veliko več gradiva in v več različnih situacijah. Zajem pogovorov zunaj Slovenije je izredno zahteven tako z vidika snemanja (veliko težje je najti ustrezne sodelavce snemalce kot za ostale slovenske regije) kot transkribiranja (transkriptor mora biti govorec ali dober poznavalec zamejskega narečja). V korpus smo tako zajeli dva posnetka zasebnega pogovora med Slovenci v Italiji, dva posnetka zasebnega pogovora med Slovenci v Avstriji ter dva posnetka intervjujev v slovenščini s tamkaj živečimi Slovenci na radiu, ki oddaja na Madžarskem. Govor Slovencev po svetu ni zajet v korpus.

3.2 Sestava zajetih posnetkov

3.2.1 Besedilnovrstna sestava korpusa Gos

Pri gradnji korpusa Gos sta bila kot bistvena besedilnovrstna kriterija izbrana javnost diskurza in prenosnik. V javni informativno-izobraževalni diskurz smo zajeli vsebinsko zaključene izseke, kot so novinarski prispevki, intervjuji, okrogle mize in šolske učne ure oz. predavanja[38], v razvedrilnega pa na primer kontaktne oddaje, resničnostne šove in športne komentarje. V nejavni nezasebni diskurz smo uvrstili posnetke delovnih sestankov, konzultacij, telefonskih storitev in svetovanj, znotraj zasebnega diskurza pa smo posneli telefonske ali osebne pogovore med prijatelji ali v družinskem krogu. Tabela 6 prikazuje končno sliko gradiva korpusa Gos.

38 Če je bil tak posamezen dogodek predolg, npr. več kot 35 minut, je bil izbran samo vsebinsko zaključen izsek, saj bi sicer posamezen diskurz preveč vplival na uravnoteženost gradiva. Pri tem je bilo upoštevano tudi število govorcev – več kot je bilo govorcev, večja je bila toleranca do dolžine posnetka.

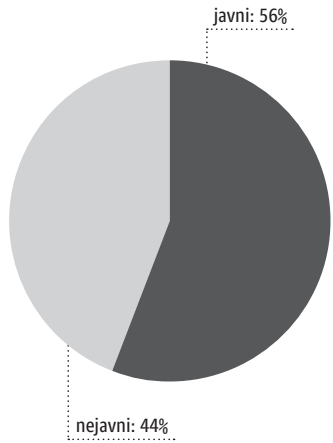
Tabela 6: Razporeditev gradiva glede na osnovne besedilnovrstne kriterije

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%

* Kategorijo Tip diskurza smo zaradi lažje razumljivosti v konkordančniku preimenovali v kategorijo Tip govora (glej poglavje 4).

Iz tabele je razvidno, da je končni razrez zajetih diskurzov dobro usklajen z zastavljeno shemo v specifikacijah (gl. poglavje 2.2), morebitni odstopi pa so večinoma posledica kombinacije različnih dejavnikov (prilagodljivost in socialna mreža snemalcev, pridobivanje avtorskih pravic in soglasij za snemanje ipd.). Podrobna razmerja med različnimi kriteriji so razložena s spodnjimi grafičnimi prikazi.

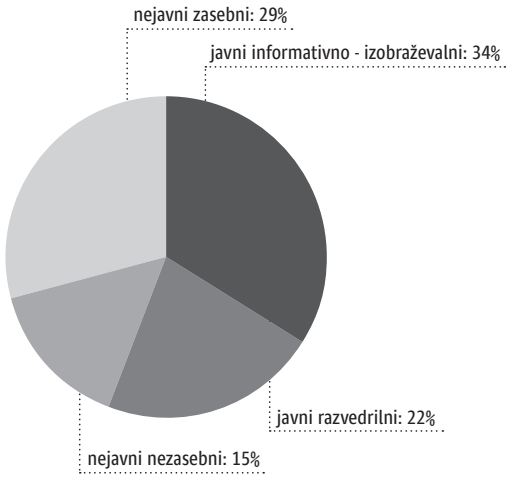
Graf 1: Javnost diskurza



Glede na zastavljeno razmerje (60% javnega in 40% nejavnega diskurza) lahko vidimo, da je bilo zajemanje nejavnega, zlasti zasebnega diskurza lažje, kot smo pričakovali, saj se je mreža snemalcev zelo dobro razvila po vsej Sloveniji. Zato smo namenoma nekoliko prevesili tehtnico v prid zasebnega dela, saj prav tam pričakujemo največ variabilnosti jezika in novih informacij o govorjeni

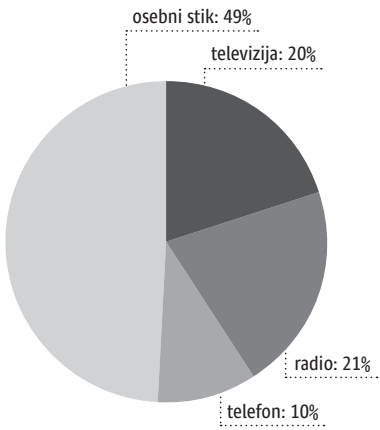
podobi slovenščine. Pri javnem diskurzu se je za trši oreh izkazalo zgolj zajemanje šolskega diskurza, povezano z dovoljenji za snemanje in urejanjem avtorskih pravic.

Graf 2: Tip diskurza



Iz grafa je razvidno, da smo glede na zastavljeno shemo zajeli nekoliko manj informativno-izobraževalnih vsebin (34,2% od predvidenih 40%). Shema se je v prid razvedrilnih prevesila zlasti na račun radijskih vsebin, kjer se je pokazalo, da je ločitev na informativne in razvedrilne zelo težka, sploh potem, ko izločimo vse brane novice. Večina prostega moderiranega radijskega programa namreč skuša podajati informativne vsebine na razvedrilen način, zato jih je bolj primerno razvrstiti v razvedrilni kot v informativni sklop.

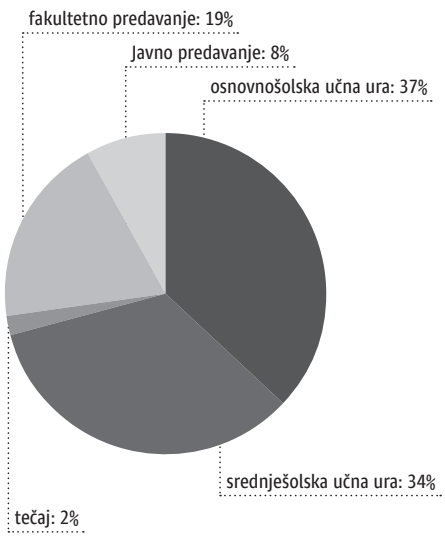
Graf 3: Prenosnik



Skoraj polovica gradiva je zajetega prek osebnega stika. V to kategorijo spada velik del kategorije nejavnega diskurza in celotna kategorija šolskega diskurza (torej posnetkov pouka v osnovnih, srednjih šolah in na fakultetah).

V televizijske vsebine so bili zajeti samo programi, ki so v času zajemanja gradiva oddajali po celotni Sloveniji, saj imajo ti na področju televizije prevladujoč vpliv. Pri radijskih vsebinah pa smo tovrstnim programom namenili skupno 4% korpusa, po 8% pa radijskim postajam s področja SV in JZ Slovenije, saj imajo nasprotno s televizijo lokalni radijski programi enak ali morda celo bolj pomemben položaj kot vseslovenski radijski programi. Radijske in televizijske vsebine za korpus Gos smo zajeli glede na interne podatke Fakultete za družbene vede o poslušanosti oziroma gledanosti slovenskih medijev.

Graf 4: Šolski diskurz

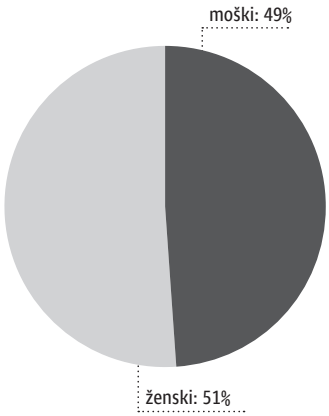


Znotraj šolskega diskurza smo zajeli približno enak odstotek gradiv v osnovnih in srednjih šolah, ostalo gradivo javnega informativno-izobraževalnega diskurza v osebnem stiku smo zajeli s posnetki tečajev ter visokošolskih in javnih predavanj.

3.2.2 Demografska sestava korpusa Gos

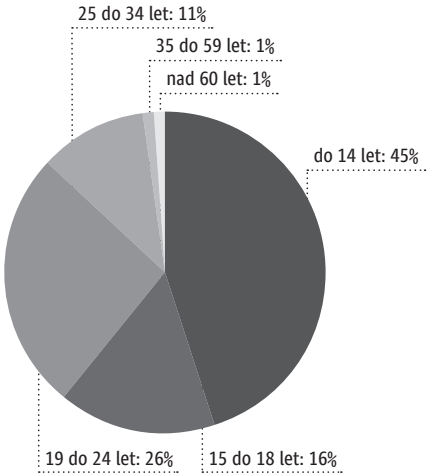
Podatki o govorcih so v celoti na voljo samo v zasebnem diskurzu, zato se predstavljena demografska sestava govorcev nanaša na ta segment korpusa Gos. V zasebni diskurz je zajet govor 186 govorcev, podatki pa so predstavljeni glede na število izgovorjenih besed.

Graf 5: Spol govorcev



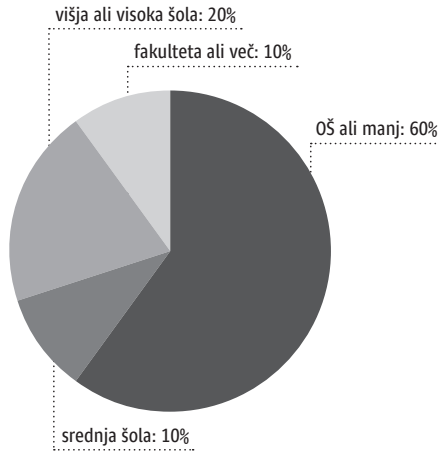
Zasebni diskurz smo želeli uravnotežiti glede na spol govorca. Končni zajem govorcev je usklajen z zastavljenim razmerjem 50:50, kar je zelo velik uspeh, saj je med zbiranjem gradiva precej težko nadzorovati demografske kriterije že glede na število govorcev, števila besed, ki jih izrečejo eni ali drugi govorci, pa v avtentičnih posnetkih nikakor ni mogoče nadzorovati.

Graf 6: Starost govorcev



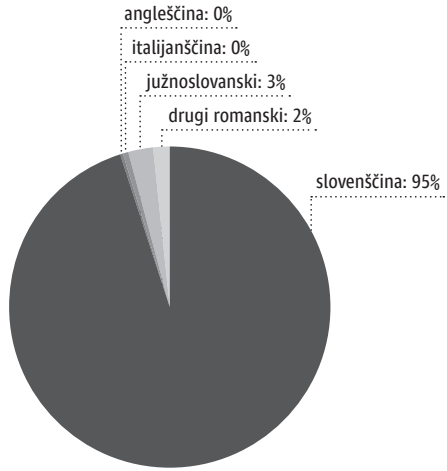
Pri starostnem razrezu se zastavljena in dejansko zajeta shema gradiva nekoliko razlikujeta. Namesto razmerja 40% proti 60% v prid govorcem, starim nad 35 let, se je razmerje obrnilo. V veliki meri gre za posledico dejstva, da je bila ekipa snemalcev sestavljena iz študentov, zato so tudi v njihovi socialni mreži prevladovale mlajše osebe, teže pa so našli govorce, stare nad 35 let, ki bi privolili v snemanje.

Graf 7: Izobrazba govorcev



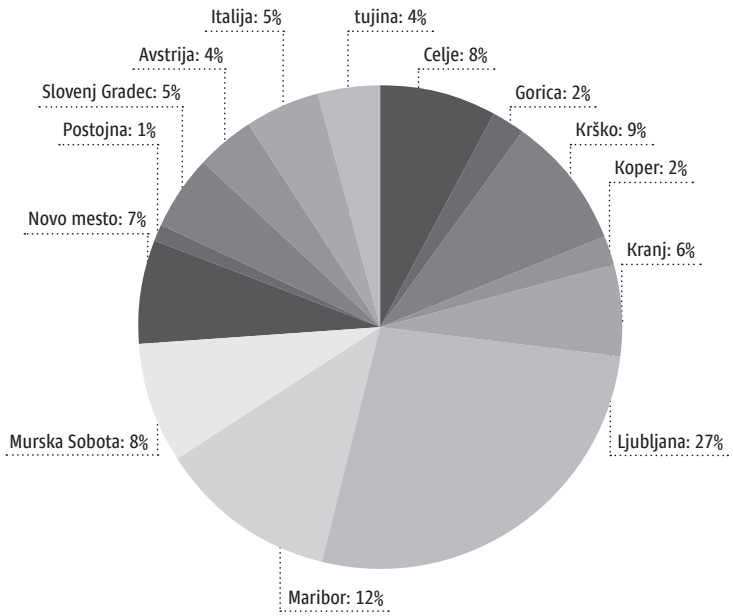
Pri kategoriji izobrazbe govorcev se je razmerje med prebivalstvom z nižjo stopnjo izobrazbe (osnovna in srednja šola) in višjo stopnjo izobrazbe (fakulteta ali več) prekrilo z zastavljeno shemo, najvišji odstotek govorcev (61%) pa ima končano srednješolsko izobrazbo.

Graf 8: Prvi jezik govorcev



V zasebni diskurz smo zajeli tudi 5% tujcev, ki živijo in se jezikovno udejstvujejo v Sloveniji.

Graf 9: Regija govorcev



Regijsko pripadnost smo pri pridobivanju podatkov o govorcih uskladili z registrskimi območji[39]. Ta odločitev temelji na metodološki preglednosti, zaradi katere odločanja o dialektalni pripadnosti nismo želeli prepuščati govorcem samim. Posledica te odločitve pa je med drugim dejstvo, da pod določena registrska območja pogosto spada tudi govor druge dialektalne skupine (npr. Grosuplje pod registrskim območjem LJ), kar moramo upoštevati pri analizi rezultatov.

V končni rešitvi spletnega vmesnika pa smo se odločili za uporabnikom prijaznejša pokrajinska poimenovanja (ljubljanska, celjska, posavska regija). Govorec ima lahko označenih več regij, če je živel dlje časa (več kot eno leto) v različnih regijah.

V korpus smo zajeli tudi govor slovenskih manjšin v Italiji, Avstriji in na Madžarskem. Govor manjšin v Avstriji in Italiji je bil posnet na terenu, kot je bilo predvideno, govor slovenske manjšine v Madžarski pa smo lahko zajeli samo iz manjšinske pogovorne radijske oddaje, zato je v korpusni strukturi formalno razvrščen v javni diskurz.

V pojasnilo dodajmo še to, da je podatek o regiji zabeležen dvakrat: kot regija govorca in kot regija snemanja. V praksi je namreč regija snemanja lahko drugačna od regijske pripadnosti govorca, poleg tega pa je regija snemanja lahko dvojna pri telefonskem pogovoru, če se sogovornika nahajata v različnih krajih.

39 Pravilnik o registrskih tablicah motornih in priklopnih vozil v tretjem členu določa: »(1) V Republiki Sloveniji so naslednja registracijska območja:
1. Celje: za območje upravnih enot Celje, Laško, Mozirje, Slovenske Konjice, Šentjur pri Celju, Šmarje pri Jelšah, Velenje in Žalec;
2. Koper: za območje upravnih enot Izola, Koper, Piran, Sežana in Ilirska Bistrica;
3. Kranj: za območje upravnih enot Jesenice, Kranj, Radovljica, Škofja Loka in Tržič;
4. Krško: za območje upravnih enot Brežice, Krško in Sevnica;
5. Ljubljana: za območje upravnih enot Domžale, Grosuplje, Hrastnik, Kamnik, Kočevje, Litija, Ljubljana, Logatec, Ribnica, Trbovlje, Vrhnika, Zagorje ob Savi in Cerknica;
6. Maribor: za območje upravnih enot Lenart, Maribor, Ormož, Pesnica, Ptuj, Ruše in Slovenska Bistrica;
7. Murska Sobota: za območje upravnih enot Gornja Radgona, Lendava, Ljutomer in Murska Sobota;
8. Nova Gorica: za območje upravnih enot Ajdovščina, Idrija, Nova Gorica in Tolmin;
9. Novo mesto: za območje upravnih enot Črnomelj, Metlika, Novo mesto in Trebnje;
10. Postojna: za območje upravne enote Postojna;
11. Slovenj Gradec: za območje upravnih enot Dravograd, Radlje ob Dravi, Ravne na Koroškem in Slovenj Gradec.«

40 Za sestavo imena datoteke je bil izdelan poseben program.

3.3 Podatki o diskurzih in govorcih

Vsak posnetek je opremljen z dodatnimi dokumenti, ki zajemajo podatke o diskurzu in udeležencih. Za vnos podatkov o govorcih in diskurzu smo uporabili posebne obrazce. Vsakemu posnetku torej pripada ena datoteka s podatki o diskurzu in toliko datotek s podatki o govorcih, kolikor govorcev je v diskurzu udeleženih.

3.3.1 Podatki o diskurzih

Tabela 7 predstavlja kategorije, v katere so snemalci in transkriptorji vnašali podatke o posameznem diskurzu. Ti podatki so izredno pomembni za različne možnosti iskanja po govornem korpusu.

Tabela 7: Podatki o diskurzu

kategorija	**podatek**
snemalec	[ime in priimek]
identifikacijska koda diskurza[40]	npr. Jirasvvaak_ab1003181413_s2
dolžina posnetka	[število minut in število sekund]
tip diskurza	javni informativno-izobraževalni
	javni razvedrilni
	nejavni nezasebni
	nejavni zasebni
vrsta institucije/situacije, v okviru katere je potekal diskurz	
	radio
	televizija
	osnovna šola
	srednja šola
	jezikovna šola
	fakulteta
	telefon
	osebni stik
opis govornega dogodka	TVSlo, Dnevnik
	PopTV, As ti tud
	Val202, Aktualna
	Maribor1, jutranji
	2. triletje, družboslovje
	gimnazija, naravoslovje
	akademski, družboslovje
	formalni delovni sestanek
	neformalni delovni sestanek
	doma, družina
	delovno mesto, sodelavci
	itd.

kategorija	podatek
regija snemanja	SV Slo, JZ Slo, celotna Slo
	MS, MB, SG, CE, LJ itd.,
	Italija, Avstrija, Madžarska, nedoločno...
vir posnetka	terenski posnetek
	RTV Slovenija
	ProPlus
	Radio Center
	Itd.
kraj poteka diskurza	[ime kraja]
čas snemanja/predvajanja	[datum in ura začetka diskurza]
št. aktivnih udeležencev	[število različnih oseb, ki govorijo v diskurzu]
prosti opis govornega dogodka	[ročni opis situacije, teme in oseb v pogovoru]

Podatki iz zgornje tabele so bili skrbno beleženi, saj predstavljajo izredno pomembne kontekstualne informacije o posnetih govornih položajih. Končni spletni vmesnik sicer ne omogoča iskanja po vseh zgornjih podatkih, je pa v veliki meri možen vpogled vanje s klikom na kontekst (gl. poglavje 4). Nekateri podatki, na primer kraj poteka diskurza, so bili zaradi varovanja identitete govorcev pri izdelavi konkordančnika anonimizirani.

3.3.2 Podatki o govorcih

Za vsakega govorca, ki je udeležen v diskurzu, so transkriptorji tvorili tekstovno datoteko, v katero so vnesli podatke v kategorije, predstavljene v tabeli 8.

Tabela 8: Podatki o govorcih

kategorija	podatek
ID koda govorca[41]	npr. lf-posl-02149
spol	m, f
starost	do 10, 10 do 14, 15 do 18, 19 do 24,
	25 do 34, 35 do 59, nad 60
regionalna pripadnost	
(lahko tudi več regij,	
v katerih je govorec	
bival več kot eno leto)	MB, MS, SG, CE, LJ, KR itd.
	Italija, Avstrija, Madžarska
	tujina
izobrazba	OŠ ali manj
	srednja šola
	višja ali visoka šola
	fakulteta ali več

41 Za generiranje kode govorca je bil izdelan poseben program.

kategorija	podatek
prvi jezik	slovenščina
	angleščina
	nemščina
	italijanščina
	itd.

3.3.3 Podatki v zapisu govora

Poleg transkripcije izrečenih besed je v datotekah z zapisom govora še veliko informacij, ki se nanašajo na strukturo diskurza ter nebesedne zvoke in prozodijo. Te informacije so bile vnesene v okviru transkripcije z orodjem Transcriber. Večino med njimi je mogoče poiskati tudi prek spletnega vmesnika.

3.3.3.1 Metapodatki o transkripciji

Na začetku Transcriberjeve datoteke se nahajajo naslednji metapodatki:

- uporabljena verzija Transcriberja,
- kodni format (CP1250),
- tip datoteke,
- transkriptor (ime in priimek prvega transkriptorja, ki je tvoril transkripcijo),
- ime pripadajoče avdio datoteke in
- datum zadnje spremembe.

3.3.3.2 Struktura diskurzov

Diskurzi so segmentirani na izjave in vloge.

Izjava/segment

Najmanjša strukturna enota zapisa, ki je prozodično, semantično in skladenjsko približno zaključena enota (stavek ali poved), npr.:

Izjava 1
ja recimo tud ne
Izjava 2
no pa je tak blo da od pribora se ful čuje

Vloga/turn

Govor enega govorca, dokler ga ne prekine drug govorec. Vloga je lahko sestavljena iz ene ali več izjav/segmentov, npr:

Vloga 1 (Govorec 1)
ja

Vloga 2 (Govorec 2)
eee mejhn te sune s pa dol al pa de se t mal zavrti s pa tud uhka dol
Vloga 3 (Govorec 1)
ja ja

Vloga lahko vključuje tudi govor dveh ali treh govorcev hkrati (hkratni govor), npr.:

Vloga 1
(Govorec 1)
ka sn te zbudla ... kak to?
(Govorec 2)
aja dans mnda nao ha? dans mnda nau piknika ja zele sn malo zadremala eem dans nau uniga mmm piknika pa tete degustacije ko ne vem či lahko vsi al ka ne vem

Pri hkratnem govoru dveh ali več govorcev ni natančno označeno, kateri deli govora so izgovorjeni hkrati. Za začetek hkratnega govora štejemo začetek izjave, v kateri se vključi drug govorec, za konec hkratnega govora štejemo konec zadnje izjave, v kateri se pojavlja hkratni govor. Zapisan je govor največ treh hkratnih govorcev. Če govorijo hkrati več kot trije, to ni posebej označeno.

Za uporabo v spletnem iskalniku so bile zvočne datoteke razsekane po izjavah/segmentih glede na to, kako je transkriptor posamezen posnetek v programu Transcriber segmentiral na manjše enote. Procesiranje zvočnih datotek so izvedli na Fakulteti za elektrotehniko in računalništvo in informatiko v Mariboru.

3.4 Zapis govora

Referenčni korpus Gos predstavlja zbirko zapisov govora, pri kateri lahko predvidevamo zelo različne iskalce z zelo različnimi predznanji in interesi. Zato smo pri označevanju posnetkov skušali ohraniti tiste kontekstne informacije, ki so pomembne kot potencialni iskalni pogoj za uporabnika (podatke o diskurzu, podatke o govorcih, dejansko govorjeno podobo diskurza). Hkrati pa smo želeli, da bi uporabnik glede na pravila zapisa govora čim hitreje našel želeno obliko (gl. tudi poglavje 2.3.2).

Pri zapisu govora se je hitro pokazalo, da vseh različnih ciljev (hitro in enostavno transkribiranje, dejanska podoba diskurza, avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami) ni mogoče doseči z eno samo rešitvijo.

Zato smo ustvarili dva nivoja zapisa govora: na prvem nivoju zapisa, ki ga imenujemo pogovorni zapis, zapišemo besede sicer ortografsko (ne fonetično!), vendar tako, kot so izgovorjene; na drugem nivoju,

ki ga imenujemo standardizirani zapis, pa zapis standardiziramo na tak način, da različnim variantam neke besedne oblike (npr. *mam, jemam*) pripišemo krovno standardno obliko (npr. *imam*). Tako s prvim nivojem omogočimo dober vpogled v besedje in oblike govorjenega jezika, z drugim nivojem pa razširimo iskalne možnosti ter omogočimo uspešnejše nadaljnje avtomatsko označevanje.

3.4.1 Pogovorni zapis

Osnovni cilj pogovornega zapisa je karseda zvesta predstavitev glasovne podobe govora v karseda berljivi obliki. Zato smo govor zapisali v veljavnem slovenskem črkopisu, od pravopisne norme pa zapis odstopa na mestih, kjer določena izgovorjena beseda odstopa od standardne izreke.

Zapis specifičnih pojavov govorjenega jezika

1 Redukcije

Glasov, ki niso izgovorjeni, ne zapisujemo, npr. *tud, neki, tko, mam, čevli*...

Polglasnika ne zapisujemo posebej pri:

- zvočnikih r, l, m, n: *sn, pr, mislm, hitr, zloml, prjatlci*...,
- enoglasovnih predlogih, členkih ipd.*: s, z, d*... (tudi če so izgovorjeni zložno, s polglasnikom),
- enozložnih besedah: *nč, jz* (za *jaz;* lahko tudi *jst*, če je izgovorjeno s *t* na koncu; ali tudi *jez*, vendar samo, če je *e* izgovorjen široko, ne kot polglasnik)...

Polglasnik lahko zapisujemo z »e« v dvo- ali večzložnih besedah, npr. *kešni (kakšni),* razen pred zvočniki m, n, r, l (*zloml, mislm, hitr*...).

Zapisovanje oblik pomožnega glagola »biti« in zaimka »se/si«:

- redukcije »bi« v »b« in »se/si« v »s« zapisujemo kot samostojno besedo, npr. *ne b (ne bi), če b (če bi), pa b mene (pa bi mene), najraj b vidu, da s*...,
- redukcije in premene oblik za prihodnjik (*bom, boš, bo*...) zapisujemo na naslednji način: *čev (če bo), čem (če bom), dam (da bom), navm/nam/nemo/nevm (ne bom), nav (ne bo),* vendar kot dve besedi pri zvezi polnopomenske besede in redukcije prihodnje oblike pomožnika: *rekla m (rekla bom)*...

2 Premene po zvenečnosti

V pisavi jih ne upoštevamo (*tud dobr, tud tak, grandž scena*...).

3 Dvoustnični v

Zvočnik dvoustnični v (ni nosilec zloga) zapisujemo:

- s črko »v«, npr. *prov, nav, navm, odpravt, davn, tip pršov, delov, gledov* (deležnik stanja na –l z glasovno premeno),
- oz. tudi z »l«, če tako izhaja iz knjižne norme, npr. *kosil* (samostalnik *kosilo*), *mel* (*imeli*) – redukcije končnega vokala v deležniku na –l.
- Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog v, izgovorjen samoglasniško), ga pišemo s črko »u« (*pršu, vidu, u tem delu*...).

4 Pokrajinsko specifični glasovi

Diftonge in druge pokrajinsko specifične foneme, ki jih ni v knjižnem jeziku, pišemo z najbližjimi ustreznimi črkami, odvisno tudi od izgovorjave v konkretnih primerih.

Seznam diftongov:

- ***ej*** *(rejs = res)*
- ***aj*** *(lajtus = letos)*
- ***uj*** *(tujdi = tudi)*
- ***ov*** *(sovh = suh, povvati = puvati; prov = prav)*
- ***av*** *(tav = to)*
- ***ev*** *(kjevk = koliko)*
- ***uo*** *(puole = pol)*
- ***je*** *(rjekla = rekla)*
- ***ju*** *(kjuk = koliko, tjuk = toliko)*
- ***ue*** *(zmuetlu = zmotilo)*
- ***ea*** *(nea = ne)*
- ***ua*** *(uaknu = okno)*
- ***uo*** *(duobru = dobro) itd.*

Drugi pokrajinsko specifični fonemi:

- *»u«* ali tudi *»i«* za u s preglasom,
- *»h«* ali tudi *»g«* za zveneči primorski h,
- *»r«* za mehkonebni koroški r itd.

5 »Mašila«

Podaljšani polglasnik ali zvočnik m ali n in njihove kombinacije, ki pogosto zapolnjujejo premore v govoru, pišemo s tremi črkami, in sicer:

eee, eem, een, nnn, mmm...

Druge medmete zapišemo z nizom črk, ki najbolje ustreza dejanski izgovorjavi. Trajanja medmetov ne označujemo posebej.

6 Besedni fragmenti

So označeni z oklepaji '()', npr.: *jz sn ej k() pač*.

7 Prozodija

Nekatere izstopajoče prozodične pojave označujejo naslednja ločila:

- »?« (vprašaj) za vprašanja,
- »!« (klicaj) za izrazito ukazujoč govor oz. ob zavpitju ali vzkliku,
- »...« (tri pikice) za mejo med izjavami istega govorca v hkratnem govoru, ko tam, kjer je meja med izjavama, zaradi govora drugega govorca ne moremo narediti novega segmenta.

8 Petje

Če je npr. zapet le kratek verz ali par besed, zapeto besedilo vključimo v zapis govora in ga ne označujemo posebej. Če je zapet daljši odsek (več verzov, kitica, cel refren, cela pesem...), ne transkribiramo, ampak označimo kot premor (prazna izjava).

9 Nerazumljiv govor

Nerazumljiv govor je označen z oznako [neraz], npr.: *nehaj se zaj [neraz] zaradi tega.*

10 Posebni govorni zvoki

Zvoki, ki nastanejo z govorili in prispevajo k vsebini diskurza, pa jih nikakor ne moremo ustrezno zapisati s črkami, kot tudi pragmatično pomembni zvoki, ki nastanejo z govorili, kot so npr. jok, hlipanje oz. hlipajoč govor, zehanje, vzdih, odkašljanje ipd., so označeni z oznako [glas] npr.: *[glas] sej ne raita.*

Vdihi in izdihi, kašljanje, tleski z jezikom in drugi zvoki, ki ne nosijo sporočila in niso pragmatično pomembni za diskurz, niso označeni.

11 Smeh

Označeno je samo mesto, kjer se smeh začne, in sicer z oznako:

- [smehgo], če se smeji samo govorec,
- [smehna], če se smejijo samo naslovniki oz. občinstvo, in
- [smehob], če se smejijo vsi.

Ni označeno, kakšen je smeh, kako glasen je, kako dolgo traja, prav tako ni zapisan z medmeti, ki običajno označujejo smeh (npr. haha). Primer:

- *nehaj se zaj zaradi tega*
- *[smehna]*

12 Nejezikovni zvoki

Gre za zvoke, ki ne nastanejo z govorili, ampak so zunanjega izvora. Če je tak zvok pragmatično pomemben, npr. če zvonjenje telefona prekine potek diskurza in se eden od govorcev odzove na klic, je označen z oznako [zvok] na mestu v transkripciji, kjer se tak zvok začne. Če zvok ne vpliva na diskurz, ni označen. Primer:

- *[zvok] mhm*

13 Začetek izjav

Izjave začenjamo z malo, ne z veliko začetnico.

Zapis ostalih jezikovnih struktur

1 Zloženke

Če gre za eno besedno enoto, jih pišemo skupaj in brez vezaja (ne glede na to, ali gre za podredno ali priredno zloženko).

Če ne predstavljajo ene besedne enote oz. gre za zvezo prislova in pridevnika, ju zapišemo s presledkom kot dve besedi.

2 Lastna imena

Domača lastna imena zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico skladno s pravopisom, npr. *Delo, Brežice*. Večbesedna stvarna in zemljepisna lastna imena dodatno označimo z zavitimi oklepaji: *{Novo mesto}, {Lenart v Slovenskih goricah}, {Ministrstvo za kulturo Republike Slovenije}, {Občina Starše}, {Osnovna šola Ivana Cankarja}* itd.

Tuja lastna imena zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico, npr.:

- *Bler, Hjuston.* Če so večbesedna, jih označimo z zavitimi oklepaji, npr.: *{Nju Jork}, {Los Endželes}*.

3 Citatne besede

Citatne besede, ki niso lastno ime, pišemo tako, kot so izgovorjene.

4 Govor v tujem jeziku

Če je del stavka, cel stavek ali več stavkov povedanih v tujem jeziku, je tak govor označen z oznako [tujjez], npr.:

- *[tujjez]*
- *blank durh najlon ajn tip štet in dize majsten prufen šteln an der tages ornung de najlon štrumf cum štifn polirn*
- *[tujjez]*

Če je v tujem jeziku samo beseda ali fraza, to ni posebej označeno.

5 Osebni podatki

Osebne podatke o samih govorcih (tudi pri javnem diskurzu) anonimiziramo, tako da označimo samo vrsto podatka, *[ime], [priimek], [podjetje], [ulica],* npr.: *zadnjič sn pri [ime] snemala.*

Prav tako anonimiziramo osebne podatke o osebah, ki sicer niso prisotne v diskurzu, so pa vseeno omenjene, če gre za nejavne osebnosti. Anonimiziramo tudi imena manjših in zasebnih podjetij, zlasti če so omenjena v kontekstu kakšne afere. Imena večjih in javnih podjetij lahko izpišemo.

Normalno, kot lastno ime, pa zapišemo imena javnih osebnosti, ki so omenjena v diskurzu, npr. imena politikov, športnikov, novinarjev

in voditeljev, umetnikov in drugih kulturnih delavcev ter ostalih medijsko opaznih osebnosti. Kraj bivališča in poštna številka ne štejeta za osebni podatek in ju ne anonimiziramo.

6 Kratice
Pišemo jih tako, kot so izgovorjene, vendar skupaj, če gre za eno kratico, npr.: *erteve, teve, trr, Sds.* Če je kratica lastno ime, jo pišemo z veliko začetnico, npr.: *Sazuja, Tevetri*. Spletni naslovi se zapišejo na način: *veveve pika {Radio Capris} pika si*.

7 Številke
Vedno jih izpišemo z besedo.

3.4.2 Standardizirani zapis

Zaradi boljših iskalnih možnosti in avtomatskega oblikoslovnega označevanja smo zajeto gradivo transkribirali tudi s standardiziranim zapisom, ki vsaki besedi pripisuje najbližjo standardno besedno obliko.

Osrednje vodilo pri pretvorbi pogovornega v standardizirani zapis je, da odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Ciljna oblika je standardna različica istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru (gl. tudi poglavje 2.3.2).

3.4.2.1 Načela standardizacije

Standardizirani zapis govora poenoti različne, predvsem glasoslovne variacije določene besedne oblike tako, da jih lahko v nadaljevanju vodimo pod eno enoto, ki naj sovpada s knjižno enoto, če taka enota obstaja v knjižnem jeziku. To pomeni, da standardizirani zapis ohranja vse jezikovne lastnosti gradiva, ki so pomembne za nadaljnje stopnje označevanja:

- tokenizacijo: število besednih enot je enako kot pri pogovornem zapisu ali pa je sprememba ustrezno označena,
- lematizacijo: upoštevanje potencialne leme vsake posamezne besedne oblike, s čimer se olajša avtomatska lematizacija korpusa,
- oblikoslovno označevanje: ohranjanje formalnih oblikoslovnih lastnosti besede (besedna vrsta, spol, sklon, število, oseba),
- skladenjsko označevanje: ohranjanje skladenjskih lastnosti govora (stavčni členi, struktura besednih zvez).

3.4.2.2 Primeri dobre prakse

Zgoraj opisana načela standardizacije najbolje prikazujejo primeri uresničenih pretvorb pogovornega zapisa v standardizirani zapis.

I) GLASOSLOVNA RAVEN

Spremenimo:

redukcije

- *bl ⇢ bolj, kešnmu ⇢ kakšnemu, kolk ⇢ koliko, pršl ⇢ prišlo, tolk ⇢ toliko, tle > tule...*

premene

- *a uš ⇢ a boš, bliži ⇢ bliže, druzga ⇢ drugega, fčeraj ⇢ včeraj, gnar ⇢ denar, de na bi ⇢ da ne bi, dej ⇢ daj, jejli ⇢ jedli, jeskat ⇢ iskat/i, kukr ⇢ kakor, lohka/lah ⇢ lahko...*

Nepolnopomenske besede, kot so k, ku, ka, ko...

Ker se v različnih regijah pojavljajo različne besede s podobno problematiko standardizacije zapisa, jih obravnavamo vse na enak način, čeprav ne gre za povsem analogne primere. Primeri so navedeni v nadaljevanju in se nenehno dopolnjujejo.

- ***k ⇢ ker, ki, ko, kot, kjer...*** (*kašn si k posušen jurček ⇢ kakšen si kot posušen jurček; tiste k sem mela za valeto ⇢ tiste ki sem imela za valeto; zato k smo se umaknli ⇢ zato ker smo se umaknili*)
- ***ka ⇢ ki, kaj, ker...*** (*tisti ka so bolj zaspane sorte ⇢ tisti ki so bolj zaspane sorte; jah veš ka on zaj še šče ⇢ ja veš kaj on zdaj še hoče; zaj bi še to trbelo malo polejvati ka je tak fejst sujo ⇢ zdaj bi še to trebalo malo polivati ker je tako fejst suho*)[42]
- ***ko ⇢ kot, ki...*** (*tak ko solarij ziher ne škodi ⇢ tako kot solarij ziher ne škodi; to je tisti ko je tak eee večkotn ⇢ to je tisti ki je tak eee večkoten*)
- ***ku ⇢ kot, ko, ki, kako, ker...*** (*tko ku prej ⇢ tako kot prej; tlele no ku sem ⇢ ja tlele no ko sem; ma in kej tiste ku so lani hodile ⇢ ma in kaj tiste ki so lani hodile; ku si bla lejpa ⇢ kako si bila lepa; zato ku vzamejo ⇢ zato ker vzamejo*)

II) LEKSEMSKA RAVEN

Pogovorne besede, ki bi jim težko določili povsem ustrezno knjižno različico, ohranjamo. Pri odločitvah glede zapisa je ključni kriterij najpogostejši zapis v pisnih korpusih in rabi na spletu.

42 Razen ko beseda ka oblikoskladenjsko in pragmatično ustreza vezniku da.

Ohranimo:

pogovorne izposojenke

- *bek, čuješ, fak, fajrala, ferker, ful, gruntali, hambrt, kafič, kao, kuhla, može, ni mus, oreng, pašeš, plata, pošlihtaš, rajsar, ratati, singl, spedenan, štima, šparati, valjda, ziher, žiher, žijaš...*

pogovorne besedilne aktualizatorje, zaimke, veznike, pridevnike, členke

- ***a*** *(a se čva midva umakniti [ime]?)*
- ***anche*** *(zdaj bojo oni ki so kriminalci imeli anche svoj urad ne)*
- ***bem*** *(bem vsake toliko zadržijo kaj zase ne)*
- ***en, ena, eno*** *(*nosilec besedilne nedoločnosti: *eno štorijo ti bom povedal o enem profesorju fizike)*
- ***hal*** *(*avstrijska Koroška: *tisto pa to sem lepo naredila hal?)*
- ***jel*** *(ko že zdaj tudi nisi tako grozno mlad jel)*
- ***ma*** *(saj pol vidim ma če jih nimam gor pa nankar gospoda h oltarju ne vidim)*
- ***nankar*** *(je pa rekla da je bilo tako fajn da ne ve kako bo upala vsem sošolcem nankar povedati)*
- ***ta*** (ki ustreza funkciji določnega člena pred pridevniško besedo in je za vse tri spole enak, prav tako je enak v vseh sklonih*: ta mlada, ta stara, ta rdeč avto)*
- ***te*** (MB: *kaj si te tam delala?; kako jo te to ožemaš; kaj sem te hotela zdaj reči; koliko sta te dala za to?; kaj te vem)*
- ***ka*** *(*ki oblikoskladenjsko in pragmatično ustreza vezniku da: *mislim ka ne, ti si pravil ka sem čudna)*

Spremenimo:

Ko ima pogovorna oblika enako oblikoskladenjsko in pragmatično vlogo kot knjižna, jo pretvorimo:

- ***te*** ⇢ ***potem*** *(ja v to v to sploh ne dvomim ne ker potem bi se po mojem skoz kregali; kaj pa še potem ima [ime] sploh?)*

pogovorne besede, ki jim knjižni izvor sicer lahko določimo, vendar knjižna različica tako rekoč ni v rabi

- ***tlele***

Spremenimo:

če je raba neenotna in oblika besede variira po regijah ali govorcih, določimo enotno krovno obliko

- ***jest, jz, jst*** ⇢ ***jaz***
- ***kva, kej, ka, kogà*** ⇢ ***kaj*** *(in koga si kupu?* ⇢ *in kaj si kupil?)*
- ***pol, pole, pouli, puole*** ⇢ ***pol***
- ***lej, glej*** ⇢ ***glej***

- ***ta, toti, teti ↛ ta*** *(sicer so nori ti kolegi ker pol so šli na Damjana Murkota; ti si zdaj ekspert za te Čehinje pa te Poljakinje pa to)*
- prislov ***ene*** *(ob ene sedmih,* prisotna tudi pregibna oblika: *imam v enih treh vrečkah ↛ ima v ene treh vrečkah)*
- ***un, gun, uen, oni ↛ oni*** *(*kazalec zunajbesedilne predmetnosti*: oni je šel pa kar za njo; pol smo pa kupili ono ta drugo kljuko)*
- ***anke, anka ↛ anche***
- nenka, nenkar, nankar ↛ nankar
- ***al, ali ↛ ali*** *(kaj bi pa ti izbral to ali to)*
- ***ta, tada ↛ tedaj*** *(*avstrijska Koroška*: tedaj pa ta luftbalon poštrihajva)*
- ***oba, aba ↛ aber*** *(*avstrijska Koroška*: ja aber pol pa morava spet po travi)*

III) OBLIKOSLOVNA RAVEN

Na oblikoslovni ravni pogovorne oblike besed spremenimo, pri čemer pazimo, da ohranimo prvotno oblikoskladenjsko in pragmatično vlogo besed.

Spremenimo:

pogovorne oblike glagolov

- *narest, naret ↛ naredit(i)*
- *rečt, pečt, vržt ↛ vreč(i)*
- *najdit, najdt ↛ najt(i)*
- *najdli ↛ našli*

kratki nedoločnik

- *se niso pripravljeni pogovarjat ↛ pogovarjati*
- *ne bi želel reč nič drugega kot to ↛ reči*
- *bi se pa upal trdet in mogoče malo korigirat eee je pa definitivno mimo doba špekulacij in pač potegnt določene poteze določene ukrepe ↛ trditi, korigirati, potegniti*

regionalne različice besednih oblik

- boma ↛ bova (*pa saj bo mislim boš videla da bomo nekaj si bova našle midve bova itak milijonarke pol ko bova velike*)

Ohranimo:

če knjižna oblika obstaja, vendar z drugačno oblikoskladenjsko vlogo

- ***čem, češ, čmo ↛ čem*** *(kako čmo temu reči a je to šlamparija; a čmo iti v kino)*

Spremenimo:
TODA
- ***čem/-š...*** spremenimo, če je oblikoskladenjska vloga enaka tisti, ki jo ima normativna različica: *češ ↛ hočeš (a češ čokolado ↛ a hočeš čokolado)*
  daljšanje osnove (s 't')
- *Markota*

IV) SKLADENJSKA RAVEN

Ohranimo:
Na skladenjski ravni pogovorne prvine ohranjamo, ker ne vplivajo na lematizacijo. Primeri:
- *sva šle; ste šla; ste izjavil; bova počasi morale zaključiti...*
- *te dva problema; vso drevje se suši; ta dvigalo je pokvarjen...*
- *s čem vse; s svojimi otroci; od kje vse so prišli; moram živeti v temu delu...*
- *Enrique Iglesias nam je polepšal ponedeljkov dopoldne...*
- *to si mi že zadnjič načel pa nisi nič dokončal to debato; potem pa so zavodi ugotovili da to ne morejo financirati ne...*
- *bi mogli it; on more biti pripravljen in bo mogla hoditi na ta drug oddelek; moreš preživeti kaj češ...*
- *to so vse male naklade; zdaj imam pa čisto mali avto...*
- *sem bil skoz večji; mater kakšno gužvo imata skozi...*
- *kaj ji je že ime; noben si ne bo dovolil; gre za eden interes farmacevtske industrije po razvoju in profitu; to se nisem jaz zmislil...*

3.4.2.3 Tehnične oznake za standardizacijo

Vsak diskurz je zapisan v pogovornem in standardiziranem zapisu. Tema zapisoma pripadata po strukturi identični Transcriberjevi datoteki. Načela standardizacije zapisa so opisana v sekciji 3.4.2.1, v nadaljevanju pa so navedena dodatna, tehnična navodila za standardizacijo.

1. Če eno besedo v prvotni transkripciji pretvorimo v dve ali več besed v standardiziranem zapisu, zapišemo te besede v standardiziranem zapisu stično z znakom »+« (plus):
- *navm ↛ ne+bom*
- *jevm ↛ jaz+bom*

2. Če po dve ali več besed v prvotni transkripciji pretvorimo v eno besedo v standardiziranem zapisu, zapišemo to besedo v standardiziranem zapisu z znakom »_« (podčrtaj) stično:
- *v pričo ↛ v_pričo*
- *Mak Donalc ↛ Mc_Donald's*

3. Onomatopeje, medmete, besedne fragmente in druge glasove standardiziramo z enotno krovno obliko, kjer je to mogoče:
   - *jooj, ijoj* ↛ *joj*

4. Lapsuse v izgovorjavi, če so nedvoumni, odpravimo:
   - *indidualnih* ↛ *individualnih*

5. Zloženke pišemo enako kot v prvotni transkripciji, samo skupaj ali narazen, brez vezajev.

6. Kratice zapišemo z velikimi črkami:
   - *veveve* ↛ *WWW*
   - *es i* ↛ *S_I*
   - *ha pe* ↛ *H_P*
   - *s o n c h e k* ↛ *S_ O_ N_ C_ H_E_K*
   - *ertees* ↛ *RTS*

   Če so že izpričane v korpusih in drugod v besedilih, imajo prednost izpisane besede:
   - *piar* proti *PR*

   Spletne naslove pišemo po načelu:
   - *WWW pika Maribor T_O_U_R_I_S_M pika S_I ali*
   - *WWW pika Kompas pika SI (ali S_I, če je na prvem nivoju es i).*

   Spletne naslove, ki vsebujejo večdelna ali osebna imena, pišemo po načelu:
   - *WWW {Radio Aktual} pika SI*
   - *WWW [ime] [priimek] pika com*

7. Tuja lastna imena zapišemo po pravopisni normi:
   - *{Kos Porta}* ↛ *{Cost Porta}*
   - *{Šarm el Šejk}* ↛ *{Sharm El Sheikh}*

8. Citatne občne besede zapišemo po pravopisni normi, ali citatno ali v poslovenjenem zapisu, pri čemer se odločimo za tisto obliko, ki je v virih (korpusi, internet...) bolj pogosta. Vezajev ne pišemo, ampak se odločimo za zapis ali skupaj ali narazen:
   - *granč scena* ↛ *grunge scena*
   - *ofišl suporterja* ↛ *official supporterja*
   - *rentakar* ↛ *rentacar*
   - *pab* ↛ *pub*

9. Ločil ne spreminjamo in ne dodajamo (tudi ne pik in vejic).

10. Začetki izjav ostanejo z malo začetnico.

3.4.3 Tipične težave transkribiranja

Pri izdelavi velikih zbirk zapisov govora, katerih končni cilj je pregledno in učinkovito iskanje po zbirki, je nujno postaviti natančna pravila transkribiranja. Po drugi strani pa veliko pravil pomeni veliko priložnosti, da gre kaj narobe. Ekipa transkriptorjev je bila po končanem usposabljanju in preizkusni dobi zaradi zagotavljanja kakovosti transkripcij pod neprestanim drobnogledom kontrolorja in validatorja (gl. poglavje 2.8). Za nekatere pojave pa se je izkazalo, da se, globoko vtkani v jezikovno podzavest, pojavljajo pogosteje kot drugi in se radi izmuznejo še tako natančni kontroli.

Zdi se, da je bilo za transkriptorje najteže usvojiti osnovni princip pogovornega zapisa: *»besede zapisujemo v veljavnem slovenskem črkopisu, od pravopisne norme pa zapis odstopa* **samo** *na mestih, kjer izgovorjena beseda odstopa od standardne izreke«*. To pomeni, da bo beseda *prišel*, če je izgovorjena v skladu s standardno izreko, tudi zapisana *prišel* (in ne *prišeu*), po drugi strani pa bo njena nestandardna različica zapisana kot *pršu*. Transkriptorji, ki so doumeli to ključno razliko, so praviloma hitro usvojili tudi podrobnejša pravila zapisa govora.

Največ težav pa je predstavljalo pravilo, ki se nanaša na zvočnik *dvoustnični v*, ko ta ni nosilec zloga. Ker smo želeli ohraniti karseda berljivo podobo zapisanih besed, smo po analogiji s pisnim in zbornim standardom (*prav, rokovnik, mivka* ipd.) izdelali navodilo, da ta glas tudi v primeru odstopa od standardne izreke zapisujemo:

- *s črko »v«, npr. prov, povšter, navm, odpravt, pa tudi pršov, delov, gledov (deležnik stanja na –l z glasovno premeno),*
- *oz. tudi z »l«, če tako izhaja z knjižne norme, npr. smo šli na kosil, je mel velik problemov (deležnik stanja na –l brez glasovne premene).*

Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog v, izgovorjen samoglasniško), ga pišemo s črko »u« (pršu, vidu, u tem delu...).
(gl. poglavje 3.4.1)

Pri procesu transkribiranja pa so transkriptorji ta glas večkrat zapisovali z *»u«* (*je pršou, smo mel kosiu)*, čeprav naj bi se odločali med *»v«* in *»l«* in kljub pogostim opozorilom, naj bodo na to pozorni. Pravil transkribiranja sicer v tem primeru nismo spreminjali, ker bi taka odločitev pomenila ogromno ročnega popravljanja že potrjenih transkripcij. V primeru nadgradnje korpusa Gos pa bi bilo smiselno razmisliti o prilagojeni različici tega pravila, ki bi bila bolj usklajena z že zasidranimi koncepti zapisovanja.

3.5 XML-sheme

Tekstovni del korpusa je bil v zadnjem koraku zapisan v standardu XML in sloni na priporočilih TEI (Text Encoding Iniciative), različice TEI P5. Dodane oblikoslovne oznake temeljijo na naboru oznak, ki je bil definiran v priporočilih JOS (Erjavec idr., 2010). V tej obliki je tekstovni del korpusa na voljo zahtevnejšim uporabnikom in si ga lahko snamejo s spletnih strani *www.korpus-gos.net*.

XML-shema je skladna z ostalimi shemami korpusnih virov projekta Sporazumevanje v slovenskem jeziku[43]. Določa formalno sintakso in pomen elementov in atributov, dovoljenih v korpusu Gos. Skladnost s shemo se preverja preko validatorjev XML, bodisi preko sheme DTD ali preko RelaxNG. Shema je izdelana s pomočjo spletnega servisa Roma[44].

V XML-shemo so zajeti vsi podatki v zvezi s korpusom Gos, tj.:

- metapodatki o korpusu,
- podatki o diskurzih oz. posnetkih,
- podatki o govorcih,
- zapis govora v pogovornem zapisu,
- zapis govora v standardiziranem zapisu,
- avtomatsko dodana lema in oblikoslovne oznake.

Podroben opis sheme XML je vključen v paket za prenos korpusa oziroma je dostopen na spletni strani *http://nl.ijs.si/ssj/gos/schema/tei_gos_doc.pdf*.

43 Prim. http://nl.ijs.si/ssj/.
44 http://www.tei-c.org/Roma/

4 Spletni konkordančnik za korpus Gos

Spletni iskalni vmesnik – konkordančnik za iskanje po govornem korpusu Gos je bil izdelan v okviru projekta Spletni konkordančnik za nacionalni govorni korpus slovenskega jezika (Operativni program krepitve regionalnih razvojnih potencialov 2007–2013)[45].

Konkordančnik uporabnikom omogoča napredne metode iskanja po transkripcijah in spremljevalnih metapodatkih korpusa Gos ter poslušanje pripadajočih segmentov govora v izvornih posnetkih za vsako izjavo med iskalnimi zadetki. Pri tem smo upoštevali potrebe vsakdanjih uporabnikov korpusa govorjene slovenščine, potrebe v izobraževanju in potrebe znanstvenoraziskovalne skupnosti, ki bo pri svojem delu lahko uporabljala tovrstno infrastrukturo (to so zlasti vse veje jezikoslovja ter tiste veje sociologije, psihologije, kognitivnih in informacijskih znanosti, ki v raziskave vključujejo govor). Konkordančnik je postavljen na spletnem mestu www.korpus-gos.net, ki omogoča tudi dostop do spremljevalnih podatkov, povezav, objav ipd.

Glede na cilje spletnega vmesnika bomo v tem poglavju predstavili:

- pregled obstoječih sorodnih aplikacij v primerljivih tujih projektih,
- analizo potreb različnih skupin uporabnikov konkordančnika,
- ključne funkcije konkordančnika.

4.1 Pregled nekaterih tujih sorodnih konkordančnikov

Eden prvih in še vedno najboj znanih govornih korpusov je govorna podsekcija Britanskega nacionalnega korpusa – BNC[46]. Korpus je prosto dostopen na spletu. Govorna podsekcija obsega 10 milijonov besed, vendar zajema samo transkripcije brez izvornih zvočnih posnetkov. Iskanje po korpusu je omogočeno z vmesnikom XAIRA, ki ni spletni iskalnik, ampak si ga je treba namestiti na svoj računalnik. Z njegovo pomočjo lahko iščemo po celotnem ali omejenem gradivu in dobimo primere konkordanc ali informacije o frekvencah besed in besednih vrst. Za BNC obstajajo sicer še drugi spletni strežniki, omeniti velja predvsem iskalni strežnik Marka Daviesa[47], ki omogoča veliko funkcij, vendar je po zasnovi sklepati, da je namenjen

45 Projekt je izvedel konzorcij v sestavi Univerza v Mariboru (Fakulteta za elektrotehniko, računalništvo in informatiko), Trojina, zavod za uporabno slovenistiko, in Univerza v Ljubljani (Filozofska fakulteta). Kot podizvajalec je pri izdelavi konkordančnika sodelovalo podjetje Amebis, d. o. o., Kamnik.
46 http://www.natcorp.ox.ac.uk/
47 http://corpus.byu.edu/bnc/

predvsem za raziskovalno sfero. Korpus akademske gopvorjene angleščine MICASE[48] v nasprotju z BNC vključuje samo govorno gradivo. Spletni iskalnik po korpusu omogoča prosto iskanje konkretnih besed in besednih zvez ali pa brskanje po transkripcijah korpusa in njegovih podsekcijah, vendar brez možnosti poslušanja izvornih zvočnih posnetkov.

Med slovanskimi jeziki omenimo samo največjega, govorno podsekcijo ruskega nacionalnega korpusa[49]. Podobno kot pri BNC so dostopni samo zapisi govora, brez zvoka. Spletni vmesnik je sicer mogoče hitro osvojiti in omogoča iskanje po obliki, kanalu, slovničnih ali semantičnih značilnostih oz. po metapodatkih o govorcih in diskurzu.

Francoski korpus Corpus de la parole[50] je, kot ime pove, samo korpus govorjenega jezika. Spletni iskalnik po korpusu odlikuje grafično privlačen vmesnik, ki omogoča iskanje po jeziku (zastopani so namreč vsi govorjeni jeziki v Franciji), po besedah ali frazah ali po metapodatkih o diskurzu in govorcih.

Pregled ostalih tujih korpusov kaže, da tako kot navedeni večinoma ne omogočajo dostopa do zvočnih posnetkov[51], pri iskanju pa izkoriščajo možnosti, ki jih ponuja korpusno gradivo (poleg osnovnega iskanja po besedah ali frazah še iskanje po jezikovnih oznakah na različnih ravneh ter metapodatkih o govorcih in diskurzih).

4.2 Potrebe ciljnih skupin uporabnikov

Kot ključne ciljne skupine uporabnikov korpusa Gos smo opredelili raziskovalce govora, učitelje slovenščine kot maternega ali tujega jezika in skupino drugih poklicev, tesno povezanih z govorom, med katere spadajo lektorji, razni pisci, prevajalci in poklicni govorci.

4.2.1 Raziskovalci govora

Med raziskovalna področja, ki se dotikajo človeškega govora in govorne komunikacije, spadajo:

- vse veje jezikoslovja, tudi dialektološke, fonetične, pragmatične, stilistične, sociolingvistične in forenzične,
- razne veje psihologije, sociologije, antropologije, kognitivnih in informacijskih znanosti,
- jezikovne tehnologije (npr. strojna sinteza in razpoznavanje govora).

Raziskovalci govora potrebujejo čim več informacij o gradivu in čim več iskalnih možnosti:

- dostop do pogovornega in standardiziranega zapisa, kar omogoča bolj kvalitetne rezultate iskanja in vpogled v korpusno zasnovo, potreben za ustrezno interpretacijo zadetkov,

48 http://quod.lib.umich.edu/m/micase/
49 http://www.ruscorpora.ru/en/index.html
50 http://www.corpusdelaparole.culture.fr/
51 Poslušanje zadetkov omogoča npr. nemški vmesnik (http://dsav-oeff.ids-mannheim.de/DSAv/SUCHMASK.H TM), mnogi drugi pa ne, npr. češki – http://ucnk.ff.cuni.cz/english/hledat_v_cnk.php, italijanski – http://badip.uni-graz.at/, poljski – http://korpus.ia.uni.lodz.pl/conversational/...

- čim več in čim natančnejše kontekstne informacije o posameznih zadetkih (tip diskurza, kanal, regija, leto snemanja, tip govornega dogodka, spol, starost, regionalna pripadnost, izobrazba in prvi jezik udeležencev),
- omejevanje gradiva po posameznih žanrih, kanalih in situacijah, ki omogoči medžanrsko primerjavo,
- omejevanja gradiva glede na značilnosti govorcev, ki omogoči primerjavo jezikovne rabe po različnih demografskih skupinah,
- predvajanje zvoka za bolj natančno interpretacijo zadetkov in akustične analize,
- shranjevanje rezultatov za njihovo nadaljnjo obdelavo,
- frekvenčne sezname za statistično analizo.

4.2.2 Izobraževanje

Odzivi na dosedanje objave in predstavitve[52] kažejo, da predstavlja korpus Gos zelo zanimiv vir za učitelje slovenščine kot maternega in tujega jezika:

- kot vir številnih avtentičnih primerov različnih govorjenih žanrov in govora različnih slovenskih regij,
- za spoznavanje glasoslovne, oblikoslovne in skladenjske podobe govorjenega jezika,
- za spoznavanje pragmatične narave govorjenega diskurza.

Podobno je korpus Gos pomemben pripomoček tudi za študente in predavatelje visokošolskih in fakultetnih programov, ki so ali jezikoslovni ali pa se jezikoslovja dotikajo.

V izobraževanju je v nasprotju s potrebami v raziskovanju bolj kot pestrost iskalnih možnosti pomembna enostavnost iskanja in grafična privlačnost konkordančnika. Zato so pomembne naslednje značilnosti konkordančnika:

- enostavnost uporabe, ki temelji na intuitivnem pristopu in uporabniku omogoča, da skozi samo uporabo odkriva potenciale konkordančnika in različne načine njegove uporabe,
- kratka, jasna in priročna navodila in pomoč brez uporabe strokovnih izrazov ter s primeri uporabe konkordančnika,
- predvajanje zvoka za bolj avtentičen stik z gradivom,
- jasna in privlačna grafična oblikovanost,
- enostaven dostop do konkordančnika.

4.2.3 Drugi poklici

Korpus Gos je kot vir avtentičnih primerov govorjene slovenščine relevanten tudi za naslednje poklice[53]:

- lektorji govorjenega jezika,

52 Konferenca Sirikt 2011 (Zwitter Vitez, Krapš Vodopivec, 2011). Izobraževanje za učitelje slovenščine na Centru za slovenščino kot drugi/tuji jezik 2011 (Zwitter Vitez, 2011).
53 Predstavitev korpusa Gos na AGRFT UL 2010 (Zwitter Vitez, 2010) in lektorjem govorjenega jezika na Slavističnem kongresu 2011 (Verdonik, 2011).

- razni pisci (scenaristi, pisatelji, novinarji),
- tolmači in prevajalci,
- poklicni govorci (na radiu in televiziji).

Pri omenjenih poklicih so ključne predvsem naslednje možnosti konkordančnika:

- omejevanje gradiva po posameznih žanrih, kanalih in situacijah ter omejevanje gradiva glede na določene značilnosti govorcev, kar omogoči iskanje določenih pogovornih oblik (npr. slengovski izrazi med mladimi),
- predvajanje zvoka (zlasti za poklicne govorce), da slišijo, kako so iskane besede ali fraze izgovorjene, in za bolj avtentičen stik z gradivom.

54 Razen tistih, ki nam jih nalagajo obveznosti do financerja projekta.

4.3 Predstavitev konkordančnika

Glede na predstavljene ugotovitve o potrebah uporabnikov smo zasnovali osnovni cilj spletnega konkordančnika za korpus Gos: uporabnika (pa naj bo motiviran, izkušen ali povsem naključen) pripraviti do tega, da sproži vsaj eno iskanje, v nadaljnjih fazah pa mu omogočati hitro in elegantno pot do rezultata. Na podlagi tega cilja smo predvideli zasnovo konkordančnika, ki temelji na izčiščenosti in intuitivnem ravnanju uporabnika, hkrati pa omogoča možnost izkoriščanja kompleksnih podatkov o govorcih in diskurzih, primerno za zahtevnejše uporabnike.

Tako smo vstopno stran očistili vseh elementov, ki bi lahko preusmerili pozornost uporabnika od iskalnega okenca: edini grafični element je logotip[54], pri funkcijah in spremljevalnih besedilih pa smo uporabili poimenovanja, ki vsebujejo zgolj splošno besedišče (odpovedali smo se strokovnim korpusnojezikoslovnim izrazom konkordanca, kolokator, korpusni zadetki...). To pomeni, da ima navigacijska funkcija prednost pred opisno, pa čeprav ponekod na račun terminološke jasnosti poimenovanj.

Vse te odločitve smo sprejeli z namenom, da obiskovalec vstopi v svet korpusnega jezikoslovja lahkotno in brez truda, da mu prvi rezultati zbudijo radovednost, v nadaljevanju pa željo k bolj poglobljenemu raziskovanju vsega, kar mu vmesnik in baza avtentičnih posnetkov lahko ponudita.

4.3.1 Enostavno iskanje

Osnovni način iskanja po korpusu Gos je enostavno iskanje. Ker smo zajeti govor zapisali na pogovornem in standardiziranem nivoju, je omogočeno tudi iskanje po obeh nivojih.

Slika 2: Enostavno iskanje

Ko odpremo spletno stran www.korpus-gos.net, je privzeto iskanje po pogovornem zapisu. Ta način je primeren, ko iščemo točno določeno realizacijo neke besede (*pršov*), in tudi rezultati so identični iskani obliki (*pršov*):

tip iskanja	iskana beseda	rezultat
Iskanje po pogovornem zapisu	*pršov*	*pršov*

Iskanje po standardiziranem zapisu je ustrezno, ko želimo najti vse realizacije neke besede, ki so bile standardizirane z določeno standardno obliko. Za razliko od iskanja po pogovornem zapisu pa iskanje po standardiziranem zapisu ponuja dva načina vpisovanja iskane oblike:

- če vpišemo samo standardizirano obliko (*prišel*), bo iskanje avtomatsko lematizirano, kar pomeni, da bodo rezultati zajemali vse oblike in realizacije osnovne oblike[55] (*prišel, pršli, prhajajo, prideš, pršu*),
- če vpišemo standardizirano obliko v narekovajih *(»prišel«)*, bodo rezultati vsebovali vse realizacije iskane besede (*prišel, pršov, pršu, prišo*).

tip iskanja	iskana beseda	rezultat
iskanje po standardiziranem zapisu	*prišel*	*prišel, pršli, prhajajo, prideš, pršu...*
	»prišel«	*prišel, pršov, pršu, prišo*

55 Pri iskanju je smiselno upoštevati dvoumnost nekaterih izrazov, ki lahko pripadajo različnim lemam (npr. je za lemo *biti* ali *jesti* ali *on*; lahko kot prislov ali pridevnik itd.). Pri tem si lahko pomagamo z naprednim iskanjem (gl. sekcijo 4.3.2) ali zavihkom Seznam (gl. sekcijo 4.3.3).

4.3.2 Napredno iskanje

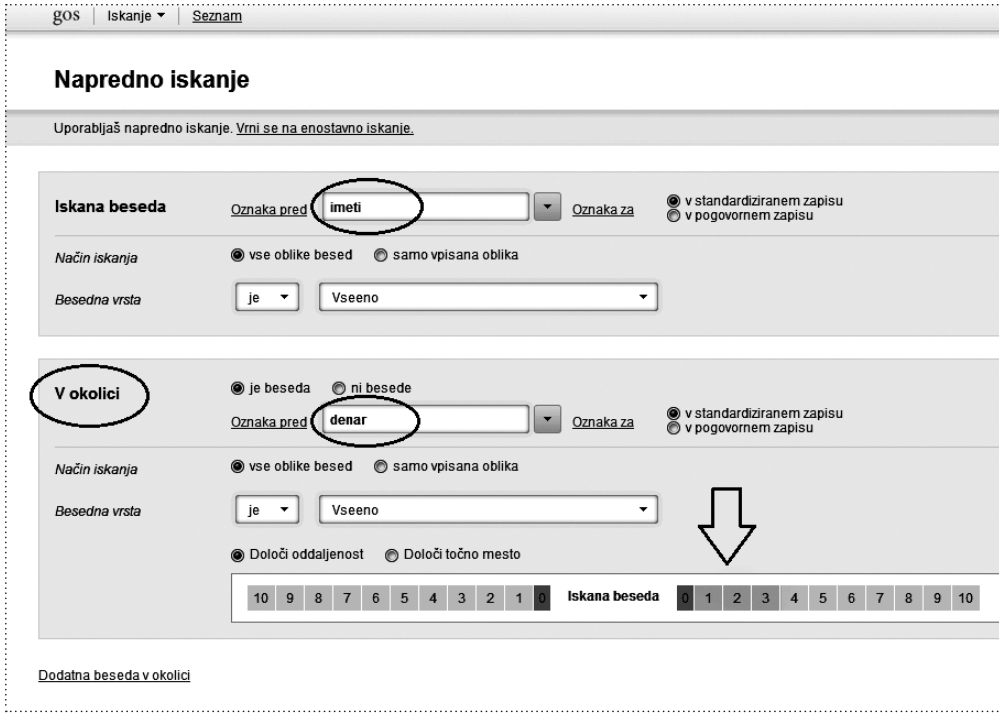
Napredno iskanje je namenjeno zahtevnejšim uporabnikom: ti izvajajo iskanje besed v specifični okolici, po slovničnih lastnostih neke besede ali pa iskanje določene besede kombinirajo s posebnimi dogodki v govoru (smeh, nejezikovni zvoki, anonimizirani podatki o govorcu ali diskurzu...).

Specificiranje okolice besede

Ta način iskanja je smiselno uporabiti, ko nas zanima raba neke besede v specifičnem kontekstu oziroma iskanje več besed, ki ne stojijo nujno ena za drugo.

Zanima nas na primer raba besede *imeti*, če je v njeni okolici beseda *denar*. V iskalno polje vpišemo *imeti*, kliknemo polje Beseda v okolici, da se odpre novo iskalno okence. Vpišemo besedo *denar* in pustimo odkljukano opcijo Vse oblike te besede. Na koncu še grafično označimo položaj in oddaljenost obeh iskanih besed: druga iskana beseda naj bo na primer nič besed pred in tri besede za prvo iskano besedo.

Slika 3: Napredno iskanje: okolica besede

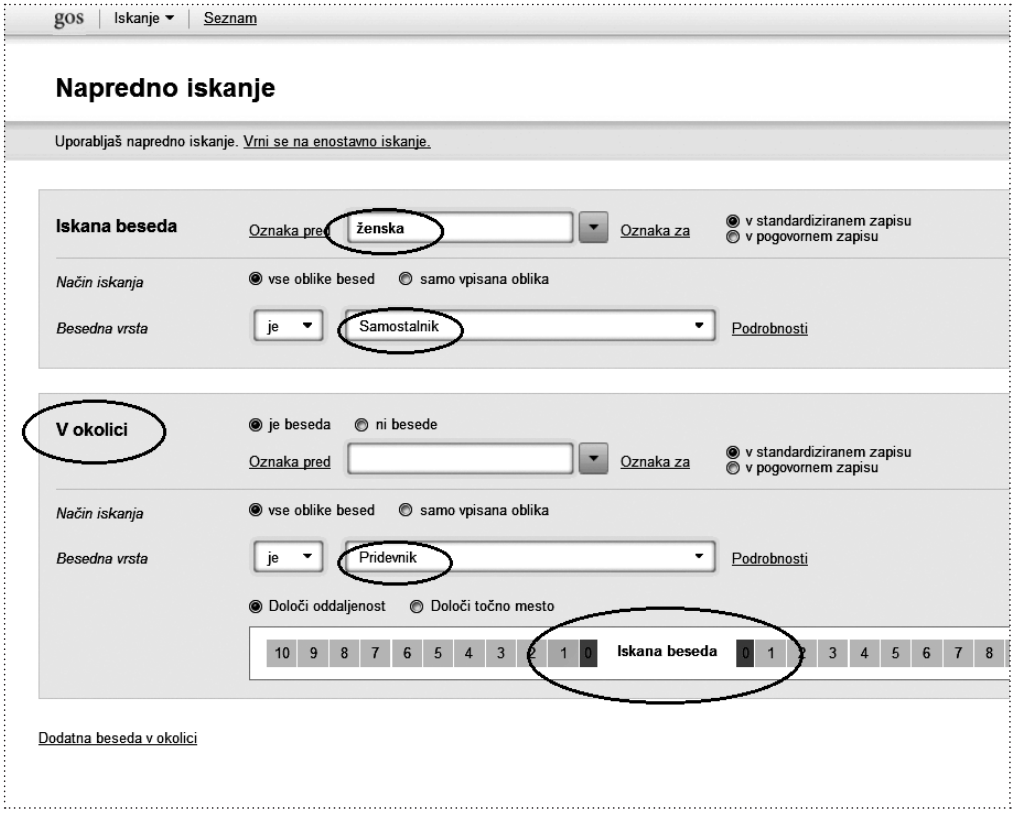


V rezultatih bomo torej dobili vse primere, ko govorec uporabi katero koli obliko besed *imeti* in *denar*, če sta med njima dve besedi ali manj.

Iskanje po slovničnih lastnostih besede

Ta način iskanja je primeren takrat, ko nas zanima, katere besede določene slovnične kategorije se pojavljajo v določenem kontekstu.

Slika 4: Napredno iskanje: slovnične lastnosti



Recimo, da želimo vedeti, kateri pridevniki se pojavljajo pred besedo *ženska*. V polje Iskana beseda vpišemo *ženska*, označimo, da gre za samostalnik (da se izognemo klasifikaciji besede *ženska* kot pridevnik), kliknemo polje Beseda v okolici in pri besedni vrsti označimo Pridevnik. Nato na grafični lestvici označimo, naj se pridevnik nahaja eno besedo *pred* in nič besed *za* iskano besedo. Med rezultati bomo tako dobili zveze *znana ženska, samska ženska, simpatična ženska* ipd.

Iskanje po posebnih dogodkih

V govoru so pogosto prisotni številni elementi spontanosti (smeh, nerazumljiv govor, hkratni govor, premor) ali posebni elementi, ki smo jih označili (anonimizirani tipi podatkov o govorcih). Zato lahko pri naprednem iskanju iščemo tudi po tovrstnih posebnih dogodkih v govoru.

Slika 5: Napredno iskanje: posebni dogodki

Predpostavimo, da nas zanimajo besede, za katerimi se poslušalci zasmejejo. V iskalno okence vpišemo kakšno pogosto besedo, na primer *ne*, pri polju *Oznaka za* pa izberemo možnost *Smeh poslušalcev*. Med rezultati bo prikazana beseda *ne*, ko ji sledi smeh naslovnika.

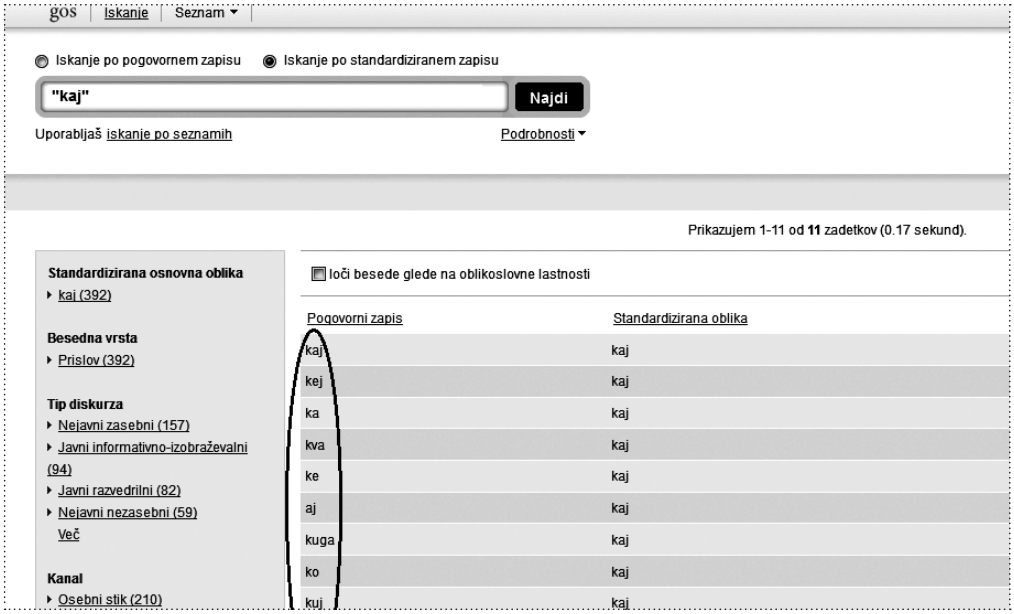
4.3.3 Iskanje po zavihku Seznam

Zavihek Seznam nudi hiter vpogled v število pojavitev vseh realiziranih oblik določene besede.

Slika 6: Zavihek Seznam

Če nas na primer zanima število dejanskih uresničitev besede *kaj*, v zavihku Seznam izberemo iskanje po standardiziranem zapisu in vpišemo besedo *»kaj«*. Rezultati bodo predstavljeni v obliki seznama dejanskih uresničitev standardne besede *kaj* (*kaj, kej, ka, kva, ke, kuga*...).

Slika 7: Zavihek Seznam: rezultati*



* V posodobljeni verziji konkornadčnika je zaradi lažje razumljivosti kategorija Tip diskurza preimenovana v Tip govora.

V primeru, da iskana beseda lahko pripada dvema različnima lemama (na primer *lahko* kot prislov in *lahko* kot pridevnik), sta možna dva scenarija:

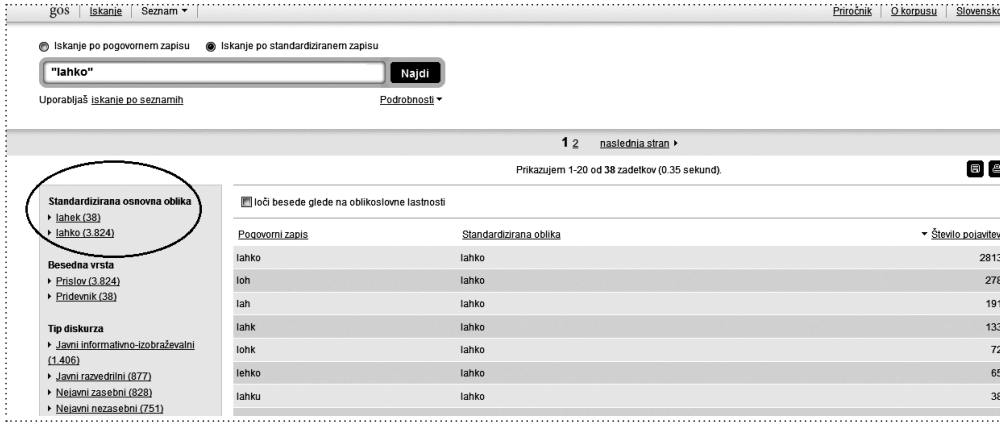
Če dvoumnost predvidimo, lahko, preden sprožimo postopek iskanja, pri opciji Podrobnosti določimo besedno vrsto in lastnosti besede.

Slika 8: Zavihek Seznam: podrobnosti iskanja

Če dvoumnosti ne predvidimo, nas bo na to spomnila dodatna vmesniška funkcija Standardizirana osnovna oblika, ki se prikaže na vrhu seznama filtrov pri rezultatih.

56 Seveda lahko znaka ? in * uporabljamo na poljubnem mestu iskane besede: miz?, miz*, bed? ipd.

Slika 9: Zavihek Seznam: razdvoumljanje rezultatov



Zavihek Seznam omogoča tudi zanimiv način iskanja z nadomestnimi znaki:

- Znak * nadomešča poljubno število znakov. Če nas zanima, katere besede se končajo na *–uk*, v pogovornem zapisu vpišemo iskanje **uk*. Dobimo oblike *kuk (koliko), tuk (toliko), tuk (tako), fejsbuk (Facebook)* ipd.
- Znak ? nadomešča en znak. Če želimo vedeti, katere oblike s tremi črkami se končajo na *–eš*, v pogovornem zapisu vpišemo iskanje *?eš*. Dobimo oblike *veš, češ, peš, ješ* ipd.[56]

Slika 10: Iskanje z nadomestnimi znaki

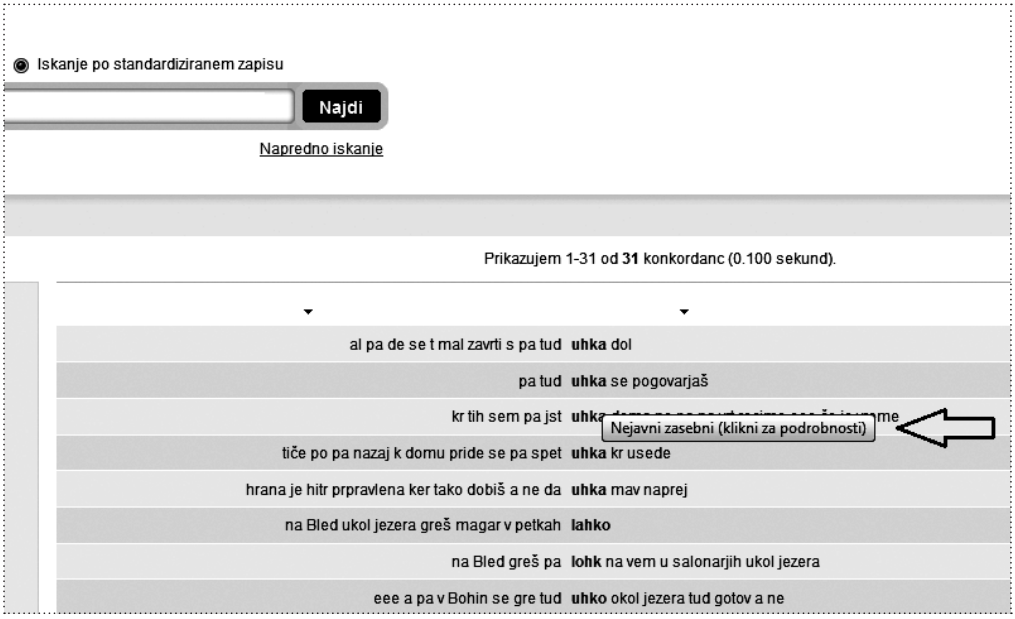
4.3.4 Prikaz rezultatov

Glede na osnovne načine iskanja se rezultati prikažejo v obliki:

- seznama konkordanc pri zavihku Iskanje,
- seznama besed s podatki o frekvenci in standardizirani obliki pri zavihku Seznam.

Rezultati iskanja pri zavihku Iskanje so prikazani v pogovornem zapisu. Če se z miško sprehodimo čez posamezno izjavo, se prikaže oznaka za tip diskurza (JI – javni informativno-izobraževalni, JR – javni razvedrilni, NN – nejavni nezasebni, NZ – nejavni zasebni). S klikom na grafično ikono zvočnika desno od konkordanc lahko poslušamo eno ali več izjav, v kateri(h) je bil izgovorjen iskani niz.

Slika 11: Prikaz rezultatov: poslušanje izjave

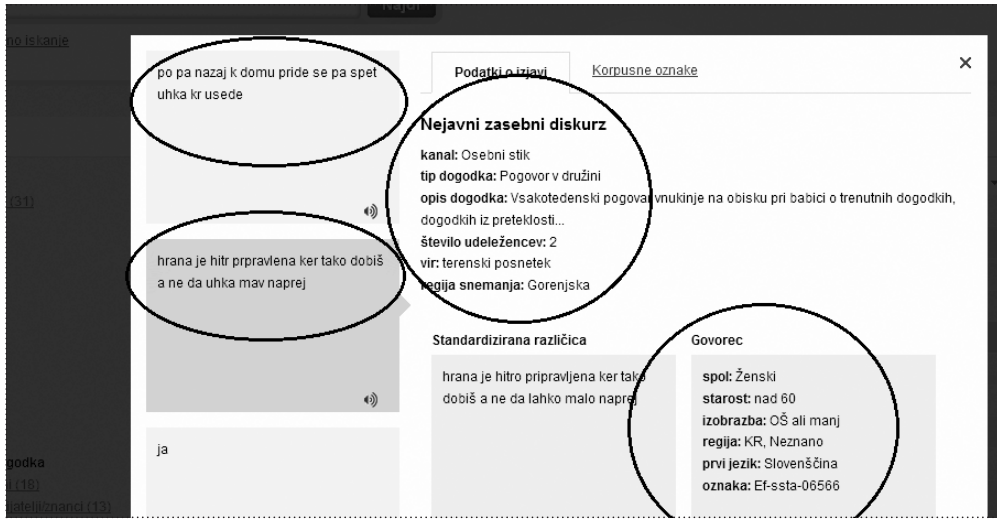


Če želimo več podatkov o posamezni konkordanci, kliknemo nanjo. Prikažejo se naslednji podatki:

- razširjeni kontekst (izjava, v kateri je bil izgovorjen iskani niz, ena predhodna in ena naslednja izjava[57]),
- podatki o diskurzu (tip diskurza, kanal, opis govornega dogodka, regija, kjer je potekal diskurz, datum in čas poteka diskurza, vir posnetka, opis diskurza),
- podatki o govorcu (spol, starost, izobrazba, regionalna pripadnost, prvi jezik),
- standardizirani zapis razširjenega konteksta.

57 Če kliknemo na predhodno ali naslednjo izjavo, se podatki o diskurzu in govorcih prilagodijo izbrani izjavi.

Slika 12: Podrobnosti izjave



Pri opazovanju razširjenega konteksta dodajmo še pojasnilo glede hkratnega govora: v spontanem govoru se pogosto zgodi, da govori več govorcev hkrati. Ko sočasno govorijo (največ) trije govorci, smo njihove izjave zabeležili, ni pa natančno označeno, kateri deli govora so izgovorjeni hkrati[58]. Zato je v podrobnostih izjave hkratni govor dveh govorcev prikazan kot dve ločeni izjavi, ena kot predhodna, druga kot sledeča, čeprav sta dejansko izgovorjeni deloma istočasno.

58 Za začetek hkratnega govora štejemo začetek izjave, v kateri se vključi drug govorec, za konec hkratnega govora štejemo konec zadnje izjave, v kateri se pojavlja hkratni govor.

Slika 13: Podrobnosti izjave: hkratni govor

Kontekst izjave lahko tudi poslušamo in pregledamo korpusne oznake zapisane izjave (lema, oblikoslovne oznake).

Pri zavihku Seznam so rezultati prikazani v obliki padajočega frekvenčnega seznama oblik določene besede ali besedne družine. Če kliknemo na število pojavitev določene oblike na desni strani vrstice z rezultati, se za izbrano obliko prestavimo v rezultate zavihka Iskanje, kar pomeni, da si lahko to obliko ogledamo tudi v obliki konkordančnega niza (slika 14). Rezultate lahko, tako kot pri rezultatih zavihka Iskanje, filtriramo glede na okoliščine diskurza in značilnosti udeleženih govorcev (slika 14).

Slika 14: Prikaz rezultatov: zavihek Seznam

4.3.5 Filtriranje rezultatov

Rezultate iskanja lahko filtriramo po značilnostih diskurzov in govorcev. Ta funkcija omogoča iskanje znotraj podatkov o diskurzih in govorcih, obenem pa predstavlja koristno informacijo o številu zadetkov znotraj posamezne kategorije (tip govora, spol in starost govorca ipd.).

Slika 15: Filtriranje rezultatov*

* V posodobljeni različici konkordančnika smo kategorijo Tip diskurza preimenovali v Tip govora.

Podatki o govorcih vključujejo naslednje izbire:

- spol,
- starost,
- izobrazba,
- regionalna pripadnost in
- prvi jezik.

Pri regionalni pripadnosti lahko izbiramo med imeni pokrajin, vendar je bil dejanski podatek o regiji govorca pridobljen glede na registrsko območje, ki mu pripada (glej tudi 2.2.1). Izberemo lahko tudi več regij za posameznega govorca.

Podatki o diskurzih zajemajo izbire po:

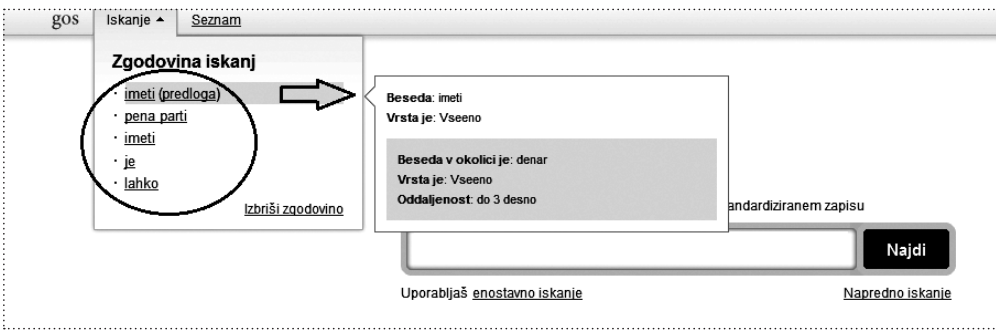
- regiji snemanja,
- letu snemanja in
- klasifikaciji diskurza.

Diskurz je grobo segmentiran na javni informativno-izobraževalni in razvedrilni ter nejavni nezasebni in zasebni diskurz. Glede na kanal poteka oziroma prenosa diskurza lahko izbiramo med televizijo, radiem, osebnim stikom in telefonom. V kategoriji Govorni dogodek pa je izbira možna med kategorijami novinarski prispevek, moderirani pogovor, osnovnošolska ali srednješolska učna ura, tečaj, javno predavanje, formalni delovni sestanek, storitev, pogovor v družini ipd.

4.3.6 Ostale funkcije konkordančnika

Poleg zgoraj predstavljenih funkcij konkordančnika si lahko nadaljnja iskanja poenostavimo tako, da uporabimo Zgodovino iskanj, ki se odpre s klikom na puščico poleg zavihkov Iskanje in Seznam. Pri tem lahko uporabimo tudi bližnjico za hitro ponastavljanje naprednega iskanja, ki jo znotraj zgodovine iskanja omogoča oznaka (Predloga).

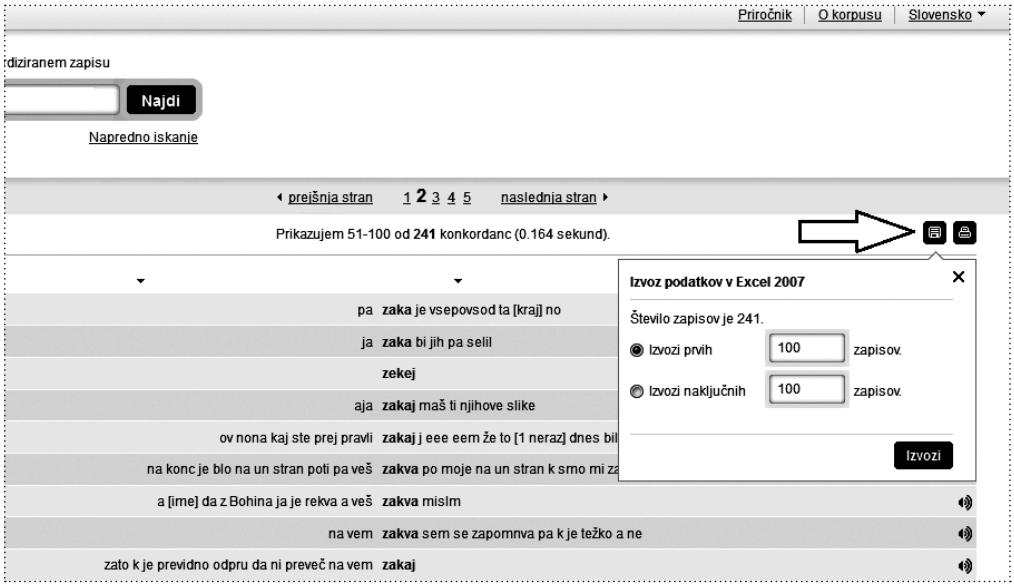
Slika 16: Zgodovina iskanj



Rezultate lahko izvozimo v tekstovni format, ki je pripravljen za tiskanje in opremljen z naslednjimi podatki:

- datum in ura iskanja,
- število izpisanih konkordanc,
- število vseh zadetkov iskanja,
- tip iskanja in
- iskani niz.

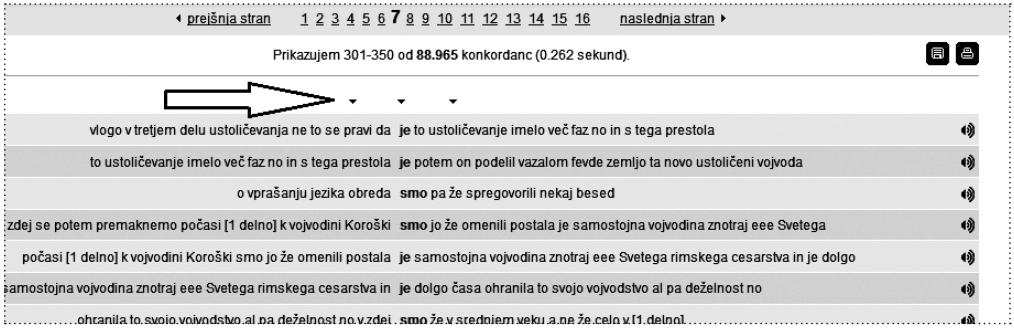
Slika 17: Izvoz podatkov



Rezultate iskanja lahko razvrstimo po abecedi glede na:

- jedrno besedo/besede,
- levo sobesedilo,
- desno sobesedilo.

Slika 18: Razvrščanje rezultatov



Ključne lastnosti konkordančnika korpusa Gos lahko strnemo v naslednje točke:

- omogoča povezavo med zvokom in zapisom: vsako konkordanco lahko poslušamo v kontekstu,
- omogoča intuitivno iskanje za manj zahtevne uporabnike in vso potrebno podporo v primeru morebitnih težav (priročnik, sprotna navodila, video o uporabi),
- izkorišča podatke o korpusnem gradivu za poglobljeno iskanje zahtevnih uporabnikov (tudi v angleškem jeziku).

V prihodnosti pa bi si želeli tudi dodatno opremljenost gradiva, npr. fonetični zapis ter skladenjske in semantične oznake.

4.4 Dostopnost gradiv

Lastnik korpusa Gos je Ministrstvo za šolstvo in šport Republike Slovenije na podlagi pogodbe »Pogodba o sofinanciranju izvedbe projekta Sporazumevanje v slovenskem jeziku v okviru Operativnega programa razvoja človeških virov za obdobje 2007–2013«, št. pogodbe 3311-08-986003, sklenjene med Republiko Slovenijo, Ministrstvom za šolstvo in šport, ter podjetjem Amebis, d. o. o., Kamnik.

Uporabniki lahko tekstovni del korpusa snamejo s spletne strani www.korpus-gos.net. Tekstovno gradivo je dostopno pod pogoji, ki jih določa licenca Creative Commons: »nekomercialno« + »priznanje avtorstva« + »deljenje pod istimi pogoji«[59].

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 Operativne-ga 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.

59 http://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode

60 Vir: http://www.korpus-gos.net/Support.aspx/About.

5 Zapodimo se v Gos

Potem ko smo korpus Gos predstavili, se ustavimo še pri vprašanju njegove uporabe. Kaj z njim početi? Kaj nam lahko pove o slovenščini in njeni rabi? Kako se lotiti iskanja po njem skozi predstavljeni spletni konkordančnik? Korpusi imajo pogosto žalostno usodo, da je veliko hrupa okrog njihovega nastanka in obstoja, njihova uporaba pa se le stežka in počasi širi. Da prebijemo zid, navajamo tri primere uporabe Gosa, izbrane tako, da predstavijo hkrati tudi različne načine iskanja z Gosovim konkordančnikom. Pri tem si bomo za ustrezno interpretacijo večkrat pomagali tudi s statističnimi podatki za Gos, predstavljenimi v tabeli 9.

Tabela 9: Nekaj osnovnih statističnih podatkov o Gosu[60]

tip diskurza	št. besed	kanal	št. besed
javni informativno-izobraževalni	359.549	televizija	102.263
		radio	94.536
		osebni stik	162.750
javni razvedrilni	228.765	televizija	105.613
		radio	123.152
nejavni nezasebni	153.471	osebni stik	119.987
		telefon	33.484
nejavni zasebni	290.990	osebni stik	222.907
		telefon	68.083
skupaj	1.032.775		1.032.775

5.1 Korpusna analiza posameznih izrazov: '*aha*' in '*aja*'

Aha in *aja* sodita med tako imenovane diskurzne označevalce. O njih je bilo v zadnjih letih že precej napisanega (glej npr. Fraser, 1999; Schouroup, 1999; Blakemore, 2005; Verdonik, 2006; 2007). Gre za izraze, ki funkcionirajo predvsem na pragmatični ravni in ne prispevajo k vsebini diskurza, to so npr. *ja, mhm, aha, aja, no, dobro, v redu, okej, veste, glejte, a ne*... V delu Verdonik (2010) je s korpusnim pristopom analiziran vpliv žanrov na rabo diskurznih označevalcev. Ugotovitve te razprave, ki nas bodo zanimale kot izhodišče za brskanje po korpusu Gos, so naslednje:

a) *aha*:
- v formalnem žanru (televizijski intervju) ni rabljen,
- veliko pogosteje je rabljen v nezasebnih telefonskih pogovorih (spraševanje po turističnih informacijah) kot v zasebnih pogovorih v osebnem stiku;

b) *aja*:

- v zasebnih pogovorih v osebnem stiku je pogosteje rabljen kot v nezasebnih telefonskih pogovorih (spraševanje po turističnih informacijah),
- v formalnem žanru (televizijski intervju) ni rabljen.

Poglejmo, ali rezultati iz Gosa potrjujejo te ugotovitve in kaj še povedo v zvezi z njimi.

Način iskanja: uporabimo enostavno iskanje po standardiziranem zapisu, iskano besedo vpišemo med narekovaji (tako iščemo samo vpisano obliko).

Rezultate za *aja* prikazuje tabela 10, za *aha* pa tabela 11. Podatki so preračunani tudi v število pojavitev na 100.000 besed glede na podatke o številu besed v celotnem korpusu (prim. tabelo 9). Število pojavitev je s pomočjo konkordančnega filtra na levi strani konkordanc prikazano ločeno glede na tip diskurza in kanal.

Tabela 10: Rezultati korpusnega iskanja za *'aja'*

tip diskurza	št. pojavitev	na 100.000 besed	kanal	št. pojavitev	na 100.000 besed
javni informativno-izobraževalni	110	31	televizija	2	2
			radio	3	3
			osebni stik	105	65
javni razvedrilni	119	52	televizija	74	70
			radio	45	37
nejavni nezasebni	148	96	osebni stik	130	108
			telefon	18	54
nejavni zasebni	799	275	osebni stik	511	229
			telefon	288	417
skupaj				1176	114

Tabela 11: Rezultati korpusnega iskanja za *'aha'*

tip diskurza	št. pojavitev	na 100.000 besed	kanal	št. pojavitev	na 100.000 besed
javni informativno-izobraževalni	192	53	televizija	8	8
			radio	18	19
			osebni stik	166	102
javni razvedrilni	306	134	televizija	132	125
			radio	174	141
nejavni nezasebni	692	451	osebni stik	366	305
			telefon	326	974
nejavni zasebni	527	181	osebni stik	321	144
			telefon	206	298
skupaj				1717	166

61 www.korpus-gos.net

Rezultati potrjujejo zgoraj navedene ugotovitve. V javnih informativno-izobraževalnih televizijskih in radijskih oddajah sta tako *aha* (8 in 19 pojavitev na 100.000 besed), še bolj pa *aja* (2 in 3 pojavitve na 100.000 besed) resnično redko rabljena. Pogostejša v javnem izobraževalnem diskurzu je njuna raba v osebnem stiku (*aha* 102 in *aja* 65 pojavitev na 100.000 besed), kjer se za korpusnimi številkami skrivajo predvsem šolske učne ure in predavanja. Bolj ko se odmikamo od formalnih situacij, bolj njuna raba narašča, tako je pri javnem razvedrilnem diskurzu (to so predvsem razvedrilne televizijske oddaje in razvedrilne radijske vsebine) že višja (*aha* 134 in *aja* 52 pojavitev na 100.000 besed), v splošnem pa najvišja v nejavnem diskurzu.

Vendar nas v nejavnem diskurzu ponovno preseneti *aha*, ki je veliko pogosteje kot v zasebnih (181 pojavitev na 100.000 besed) rabljen v nezasebnih situacijah (451). Če to konkretiziramo s pomočjo opisa gradiv, ki je dostopen v zavihku O korpusu[61], vidimo, da se pod nezasebnim diskurzom skrivajo delovni sestanki in konzultacije, pogovori ob storitvah, prodaji, svetovanju, posredovanju informacij... *Aja* je nasprotno pogosteje rabljen v zasebnih (275) kot v nezasebnih nejavnih diskurzih (96 pojavitev na 100.000 besed), kar pomeni, da je pri njem verjetno prisoten močan vpliv formalnosti situacije na rabo. V zasebnih diskurzih je *aja* resnično precej pogostejši kot *aha* (*aja* 275, *aha* 181 pojavitev na 100.000 besed).

Na podlagi zgornjih podatkov lahko primerjamo tudi, ali kanal (telefon vs. osebni stik) vpliva na pogostost rabe *aha* in *aja*. Za *aha* lahko potrdimo, da je v telefonskem pogovoru nasploh (v zasebnem – 974 pojavitev na 100.000 besed – kot tudi nezasebnem – 298) dva- do trikrat pogosteje rabljen kot v osebnem stiku (zasebni 305 in nezasebni 144 pojavitev na 100.000 besed). S tem lahko potrdimo že večkrat domnevano trditev/ugotovitev (gl. npr. Verdonik, 2007), da odsotnost vidnega stika med sogovornikoma pri telefonskem pogovoru zviša rabo signalov, kot je *aha*, s katerimi sogovornika signalizirata svojo pozornost in razumevanje. Raba *aja* v primerjavi med osebnim stikom in telefonom je bolj zapletena: v nezasebnih nejavnih diskurzih je pogostejša v osebnem stiku (108), v zasebnih pa po telefonu (419 pojavitev na 100.000 besed). Sam kanal torej nima tolikšnega vpliva na njegovo rabo.

V zvezi z *aha* in *aja* je zanimiva tudi ugotovitev, da se pogosto pojavljata v začetku izjave skupaj z drugimi diskurznimi označevalci v neke vrste gručah diskurznih označevalcev. V Verdonik (2006) je kot najznačilnejše tako zaporedje označeno naslednje (znak # pomeni, da se lahko ta diskurzni označevalec ponovi dva- ali večkrat):

aha#/mhm#/aja	*ja#*	*no#*	*dobro/okej/v redu/prav#*	*glejte*	*zdaj*

O tem želimo izvedeti več. Poglejmo, v katerih gručah in katerem zaporedju se v korpusu Gos pojavljata *aha* in *aja*.

Način iskanja: izberemo napredno iskanje, kot prvo besedo vpisujemo *aha*, ohranimo privzete nastavitve (standardizirani zapis, vse oblike besed), kot drugo besedo vpisujemo zapovrstjo: *ja, no, dobro, v* in *redu, okej, glejte, zdaj, aja*, ohranimo prav tako privzete nastavitve, pri določanju okolice besede pa nastavimo –3 in +6 (tj. iščemo okolico predhodne tri besede in naslednjih šest besed). Rezultate si po vsakem izpisu shranimo na svoj računalnik za nadaljnjo obdelavo. Postopek ponovimo z *aja*.

Nato podatke obdelamo na svojem računalniku: shranjene zadetke (samo osrednji stolpec brez levega in desnega konteksta) združimo v eno datoteko, jih shranimo kot tekstovno datoteko in sortiramo po abecedi. Ročno jih pregledamo in naredimo seznam gruč diskurznih označevalcev v začetku izjave, v katerih se pojavljata *aha* in/ali *aja*.

Podatke preverimo s ponovnim iskanjem po korpusu (enostavno iskanje, standardizirani zapis).

V tabeli 12 predstavljamo seznam gruč diskurznih označevalcev na začetku izjav, v katerih se pojavljata *aha* in *aja*, skupaj s podatki o pogostosti njihove rabe.

Tabela 12: Seznam gruč diskurznih označevalcev z '*aha*' in '*aja*'

1	2	3	4	št. rab
aja	ja			64
aha	ja			56
aha	okej			32
aha	dobro			25
aja	no			21
aha	no			18
aha	zdaj			13
ja	aja			10
aha	v redu			8
ja	aha			7
aha	aja			6
aja	zdaj			5
aja	aha			4
no	aja			3
dobro	aha			1
aha	glej			1
aha	ja	veste		2
aha	no	dobro		2
aha	okej	zdaj		2
aha	aja	ja		2
aha	no	okej		1

1	2	3	4	št. rab
aha	no	zdaj		1
aha	okej	v redu		1
aha	okej	glej		1
aha	v redu	glejte		1
aha	v redu	ja		1
aha	dobro	okej		1
aja	dobro	ja		1
aha	poglejte	zdaj		1
aja	ma	ja		1
aja	no	zdaj		1
aja	ja	dobro		2
aja	okej	v redu		1
ja	aha	okej		1
aha	ja	no		1
no	aha	aja		1
aja	mislim	ja		1
aja	dobro	ja ja ja ja	no	1
aha	ja	veš kaj	lej	1
aha	dobro	v redu	no	1
aha	okej	no	zdaj	1
ja	no	ja	aja	1
			skupaj	305

Kot vidimo, prevladujejo gruče, v katerih se sopojavljata samo dva različna diskurzna označevalca (takih je skupno 274 od 305), vendar se pri teh pogosto ali eden ali drugi ali oba lahko dva- ali večkrat ponovita (npr. *aha aha aha aha ja ja ja*). Najpogostejše so gruče *aha* ali *aja* in *ja*, gruče *aha* in katerega od diskurznih označevalcev *dobro/v redu/okej* ter gruče *aha* ali *aja* in *no*. V teh gručah je res običajneje, da je na prvem mestu ali *aha* ali *aja*. *Aha* in *aja* v isti gruči se pojavljata redko (13-krat). Trije ali štirje diskurzni označevalci v gruči se pojavljajo redkeje (31 primerov), pri tem ni posameznega prevladujočega vzorca, ampak so gruče zelo različne, spet pa večinoma z *ja, dobro/v redu/okej* ter *no*.

5.2 Pogovorne različice besed: '*lahko*'

Za govorjeno slovenščino je značilen močan vpliv vokalne redukcije. V korpusu Gos so njeni pojavi do določene mere zabeleženi v pogovornem zapisu. Za naslednje iskanje bomo izbrali besedo *lahko*, ki je pogosto rabljena in zanjo iz izkušenj vemo, da ima več pogovornih različic. S pomočjo Gosa bomo ugotavljali, katere so te različice, ter se osredotočili zlasti na tiste, ki so pogosto rabljene in so pod vplivom vokalne redukcije.

Način iskanja: uporabimo iskanje pod zavihkom Seznam, izberemo standardizirani zapis, določimo, da je besedna vrsta prislov, pod podrobnosti pa še vrsto splošni in stopnjo nedoločeno.

Izpišejo se podatki, kot jih prikazuje tabela 13.

Tabela 13: Rezultati iskanja po seznamu za besedo '*lahko*'

pogovorni zapis	št. pojavitev	standardizirana oblika	standardizirana osnovna oblika
lahko	2780	lahko	lahko
loh	277	lahko	lahko
lah	191	lahko	lahko
lahk	133	lahko	lahko
lohk	72	lahko	lahko
lehko	65	lahko	lahko
lahku	37	lahko	lahko
lohku	33	lahko	lahko
leko	32	lahko	lahko
lejko	32	lahko	lahko
uohk	27	lahko	lahko
lohka	25	lahko	lahko
lahka	24	lahko	lahko
lehku	13	lahko	lahko
lohko	12	lahko	lahko
lek	12	lahko	lahko
uoh	11	lahko	lahko
lh	7	lahko	lahko
leh	7	lahko	lahko
uhka	6	lahko	lahko
lako	6	lahko	lahko
lahke	3	lahko	lahko
leka	2	lahko	lahko
lehk	2	lahko	lahko
laho	2	lahko	lahko
uhko	1	lahko	lahko
lejk	1	lahko	lahko
lekof	1	lahko	lahko
ljehku	1	lahko	lahko
lahkko	1	lahko	lahko
loho	1	lahko	lahko
loko	1	lahko	lahko
lhk	1	lahko	lahko
jahka	1	lahko	lahko
lejkof	1	lahko	lahko
uhke	1	lahko	lahko
leku	1	lahko	lahko
ko	1	lahko	lahko

Čeprav smo predvidevali, da bo imela beseda *lahko* več različic, nas število različnih pogovornih zapisov vseeno preseneti. Vendar številke kažejo, da jih je večina zelo redkih. Daleč najpogostejša je vendarle standardna izgovorna različica *lahko*, z desetkrat manjšo pogostnostjo ji sledijo nekatere reducirane različice (*loh, lah, lahk, lohk*).

Primerjajmo najprej rabo izgovorne različice *lahko* z reduciranimi različicami *loh, lah, lahk, lohk* glede na tip diskurza. Predpostavljamo namreč, da bodo reducirane različice omejene predvsem na rabo v nejavnem diskurzu. Rezultati za izgovorno različico *lahko* so prikazani v tabeli 14 in za reducirane različice v tabeli 15. Dodani so tudi podatki o številu pojavitev na 100.000 besed.

Tabela 14: Pogostost rabe izgovorne različice '*lahko*' glede na tip diskurza

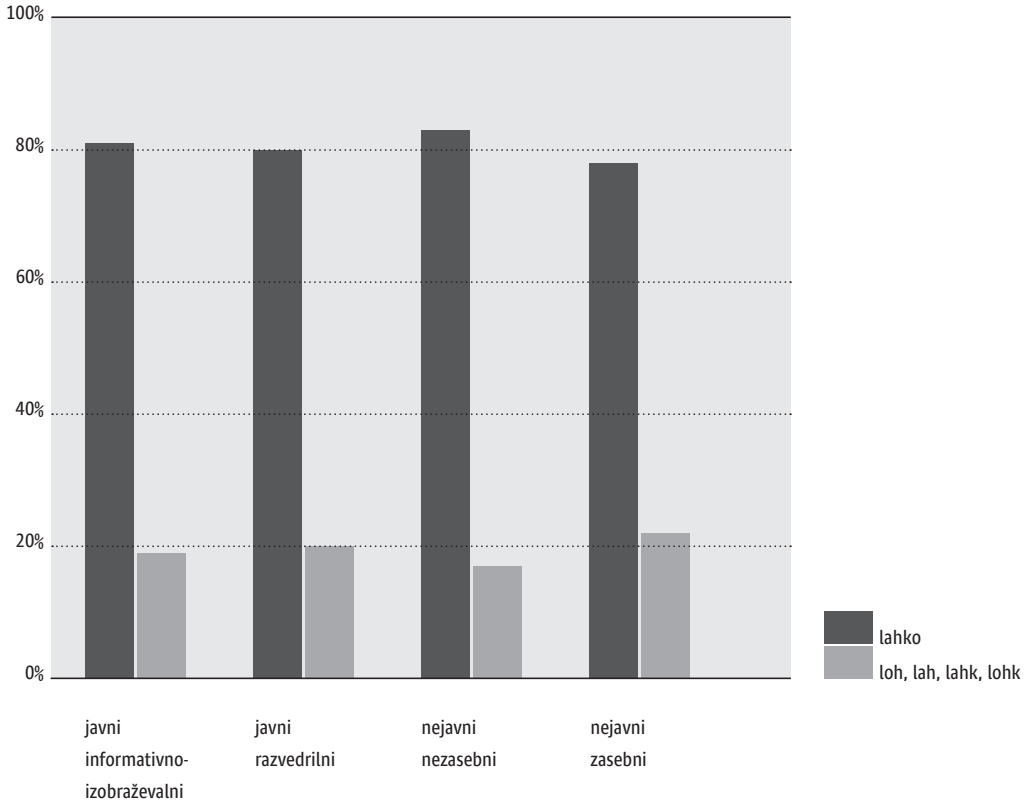
standardna oblika	**lahko**	
tip diskurza	**št. pojavitev**	**na 100.000 besed**
javni informativno-izobraževalni	1242	345
javni razvedrilni	666	291
nejavni nezasebni	586	382
nejavni zasebni	319	110
skupaj	**2813**	**272**

Tabela 15: Pogostost rabe reduciranih izgovornih različic '*loh*', '*lah*', '*lahk*', '*lohk*' glede na tip diskurza

reducirane oblike	**loh**	**lah**	**lahk**	**lohk**		
tip diskurza	**št. rab**	**št. rab**	**št. rab**	**št. rab**	**skupaj**	**skupaj na 100.000 besed**
javni informativno-izobraževalni	116	89	57	36	**298**	**83**
javni razvedrilni	69	41	38	21	**169**	**74**
nejavni nezasebni	47	36	24	9	**116**	**76**
nejavni zasebni	46	25	14	6	**91**	**31**
skupaj	**278**	**191**	**133**	**72**	**674**	**65**

Primerjava podatkov v tabelah 14 in 15 kaže, da v vseh tipih diskurza (tudi v zasebnem!) močno prevladuje raba izgovorne različice *lahko*. Reducirane različice so nekajkrat redkeje rabljene. Primerjajmo te številke še preračunano v odstotke, kar prikazuje graf 10.

Graf 10: Razmerje pogostosti rabe različice *'lahko'* in reduciranih različic v različnih tipih diskurza



Proti pričakovanjem ugotavljamo, da razmerje med odstotkom rab različice *lahko* in reduciranih različic *loh, lah, lahk, lohk* ne variira bistveno med različnimi tipi diskurza. To si lahko razlagamo s tem, da ima različica *lahko* dve naglasni varianti: poleg standardne *lahkó* še pogovorno *láhko*, ki je razširjena v koroški, štajerski in prekmurski regiji. O tem se lahko prepričamo, če zadetke za *lahko* v zasebnem diskurzu filtriramo glede na regijo govorca in poslušamo pripadajoče posnetke. Podatki v filtru na levi nam ob tem povedo, da je več kot polovica zadetkov v zasebnem diskurzu prav za govorce iz navedenih severovzhodnih regij.

Reducirane oblike *loh, lah, lahk, lohk* so v severovzhodnih regijah seveda zelo redko rabljene, kot vidimo iz podatkov v tabeli 16, ki prikazuje, kolikokrat so posamezne reducirane različice v zasebnem diskurzu rabljene pri govorcih iz različnih slovenskih regij. Podatki so omejeni na zasebni diskurz, tudi zato, ker so v korpusu Gos samo v tem segmentu ti podatki dosledno zbrani in uravnoteženi.

Tabela 16: Število rab reduciranih različic *'lahko'* v zasebnem diskurzu glede na regijo govorca

regija govorca	loh	lah	lahk	lohk	skupaj	na 100.000 besed
novogoriška	17	0	0	0	**17**	**249**
ljubljanska	75	21	13	34	**143**	**188**
krška	1	25	16	0	**42**	**172**
slovengraška	0	20	5	0	**25**	**168**
celjska	0	12	15	0	**27**	**115**
kranjska	12	3	0	1	**16**	**102**
novomeška	4	7	4	0	**15**	**72**
murskosoboška	1	1	2	0	**4**	**17**
mariborska	0	0	2	0	**2**	**6**
koprska	0	0	0	1	**1**	-
neznano	6	0	0	0	**6**	-
skupaj	116	89	57	36	**298**	**102**

Kot vidimo, so reducirane oblike redke predvsem v mariborski in murskosoboški regiji, najpogostejše pa v novogoriški regiji. Za koprsko regijo je premalo podatkov (preračun na podlagi ene pojavitve ni smiseln).

5.3 Večbesedni izrazi: *'in tako naprej'*

V Verdonik (2008) je bila izpostavljena posebna skupina izrazov, poimenovana interpretacijski označevalci. Gre predvsem za izraze, ki so bili v angleščini raziskovani pod termini general extenders, co-ordination tags, set markers, discourse extenders idr. in se rabijo praviloma na koncu izjave, začenjajo pa se ali z besedico *in* oz. *pa* ali z *ali* (npr. *in tako naprej, pa tako, ali pa kaj*). Overstreet (2005) definira naslednje osnovne funkcije teh izrazov: signalizirajo predpostavko, da naslovnik ve, kaj ima tvorec v mislih, in da zato nadaljnje tvorjenje v nakazani smeri ni potrebno; spodbujajo naslovnika k solidarnosti, naj se vživi situacijo, ki jo tvorec opisuje; nakazujejo, da bi lahko rekli še veliko več o predmetu pogovora oz. da bi lahko rekli še več, ampak je tisto nepomembno; opozarjajo, da to, kar je bilo rečeno, ni povsem natančno; ublažijo izjave, ki bi lahko prizadele naslovnika; poudarjajo povedano in spodbujajo odgovor.

V nadaljevanju navedenega vira (Verdonik, 2008) je naveden seznam tovrstnih izrazov v gradivu. Najpogostejši je *in tako naprej*. Poglejmo, v kolikšni meri in v katerih različicah se izraz pojavlja v korpusu Gos.

Način iskanja: zadostuje, da izberemo enostavno iskanje. Vpišemo izraz *in tako naprej* ter izberemo standardizirani zapis, vse oblike.

Pregled zadetkov pokaže, da se pojavljajo izgovorne različice: *in tako naprej, in tko naprej, in tk naprej, in tak naprej*. Ker nas zanima še pogostost posameznih različic, ponovimo iskanje, tako da namesto

po standardiziranem zapisu iščemo po pogovornem zapisu. Rezultate prikazuje tabela 17, ločeno za posamezen tip diskurza. Z enakim postopkom iskanja zatem primerjamo še rabe *pa tako naprej* (tabela 18) ter *in tako dalje* (tabela 19) in *pa tako dalje* (tabela 20).

Tabela 17: Rezultati iskanja za izraz *'in tako naprej'*

tip diskurza	in tako naprej	in tko naprej	in tk naprej	in tak naprej	skupaj	na 100.000 besed
javni informativno-izobraževalni	96	44	0	5	145	40
javni razvedrilni	22	16	0	1	39	17
nejavni nezasebni	5	23	0	3	31	20
nejavni zasebni	1	8	1	0	10	3
skupaj	124	91	1	9	225	22

Tabela 18: Rezultati iskanja za izraz *'pa tako naprej'*

tip diskurza	pa tko naprej	pa tk naprej	pa tak naprej	skupaj	na 100.000 besed
javni informativno-izobraževalni	6	0	1	7	2
javni razvedrilni	0	1	1	2	1
nejavni nezasebni	8	0	0	8	5
nejavni zasebni	8	0	0	8	3
skupaj	22	1	2	25	2

Tabela 19: Rezultati iskanja za izraz *'in tako dalje'*

tip diskurza	in tako dalje	in tak dale	skupaj	na 100.000 besed
javni informativno-izobraževalni	17	0	17	5
javni razvedrilni	6	2	8	3
nejavni nezasebni	3	0	3	2
nejavni zasebni	2	0	2	1
skupaj	28	2	30	3

Tabela 20: Rezultati iskanja za izraz *'pa tako dalje'*

tip diskurza	pa tak dale	pa tak dole	pa tak dalje	pa tko dalje	skupaj
javni informativno-izobraževalni	0	0	0	0	0
javni razvedrilni	1	1	0	0	**1**
nejavni nezasebni	2	0	0	0	**2**
nejavni zasebni	0	0	1	1	**2**
skupaj	**3**	**1**	**1**	**1**	**6**

Rezultati kažejo, da je raba tovrstnih izrazov v javnem diskurzu precej večja kot v nejavnem. Različica z veznikom *pa* je veliko redkejša kot različica z veznikom *in* ter se več pojavlja v nejavnem diskurzu kot različica z *in*. Reducirane oblike (s *tko, tk, tak* – 134 rab) so presenetljivo v skupnem seštevku skoraj enako pogoste kot nereducirane (s *tako* – 152 rab). Razvrstitev zadetkov glede na regije kaže, da so različice s *tko* in *tk* značilne za osrednjo, južno in zahodno Slovenijo, s *tak* pa za severno in vzhodno Slovenijo. Izraz *in/pa tako naprej* (250 rab) je v splošnem veliko bolj pogost kot *in/pa tako dalje* (36 rab).

6 Sklepne misli

Preden se v sklepnih mislih o korpusu Gos ozremo naprej, se vrnimo na začetek in s pridobljenimi izkušnjami še enkrat preletimo nekatera vprašanja zasnove korpusa in konkordančnika.

Besedilnovrstni in demografski kriteriji za uravnoteženost gradiva

V zasnovi je Gos ločen na besedilnovrstno uravnotežen del (nezasebni diskurz) ter na demografsko uravnotežen del (zasebni diskurz), s tem da je zelo groba demografska delitev na dve regiji (SV in JZ) upoštevana tudi pri nezasebnem diskurzu. Izkušnje pri zbiranju gradiv so v zvezi s temi kriteriji pokazale naslednje potencialno šibkejše točke specifikacij:

a) Televizijski in radijski diskurz lahko regionalno ločimo glede na kraj studia, v katerem se odvija, ali glede na območje, kjer je slišen/viden. Ker tega na začetku nismo posebej opredelili, smo se pri zbiranju gradiva odločili za drugo varianto, saj tudi poslušalci/gledalci, čeprav pasivno, sodelujejo v radijskem in televizijskem diskurzu. Posledično smo morali popraviti zastavljeno regionalno delitev, saj najbolj poslušani/gledani mediji pogosto oddajajo po celotni Sloveniji.

b) Javni diskurz v osebnem stiku vključuje predvsem predavanja in šolske ure, kjer praviloma prevladuje en govorec (predavatelj ali učitelj). Določene demografske kriterije bi lahko prenesli tudi na ta del gradiv.

c) V nezasebnem nejavnem diskurzu je nabor različnih pogovornih situacij zelo širok, npr. pogovori v trgovini, ob raznih storitvah, svetovanjih, posredovanju informacij, na sestankih, konzultacijah ipd., s tem da je določene zelo težko posneti zaradi občutljivosti informacij. Ob zbiranju gradiva za Gos smo skušali zajeti čim več različnih situacij. V pomoč bi bil podrobnejši predhodni razmislek o tovrstnih situacijah, k uravnoteženosti pa bi pripomogla primerna, robustna shema za njihov zajem. Upoštevanje demografskih kriterijev v tem sklopu bi bilo morda smiselno, bi pa moralo biti omejeno (npr. mogoča bi bila bolj podrobna regionalna opredelitev, upoštevali bi lahko tudi kriterij spola govorca, ostale demografske kriterije pa bi bilo treba premisliti posebej za ta tip diskurza).

č) Zasebni diskurz: delo na terenu je pokazalo, da je zajemanje zasebnih diskurzov bistveno lažje, kot smo pričakovali ob začetku. Ker se prav v tem segmentu kaže največja pestrost (in nepoznavanje) govorjene slovenščine, bi bilo treba ponovno premisliti, ali ne bi ta segment zajemal nekoliko večji delež v korpusu.

Zbiranje podatkov o gradivu ter njihova predstavitev

Pri snemanju so bili zbrani podatki o govorcih in o situaciji, v kateri je potekal diskurz. Ker smo želeli ohraniti čim več informacij, so zbrani podatki bolj natančni, kot so kategorije, ki jih predvidevajo specifikacije za zajemanje gradiva (npr. medtem ko specifikacije ločijo samo višjo in nižjo izobrazbo, je v zbranih podatkih to razdeljeno naprej na štiri stopnje). Komentar zahtevata predvsem dva tipa podatkov: regija in tip govornega dogodka.

a) Regija je zabeležena dvakrat: kot regija snemanja in kot regija govorca, kar ustreza dejanskemu stanju. V praksi je regija snemanja lahko dvojna pri telefonskem pogovoru, če sta sogovornika v različnih krajih. Označevanje regije snemanja pri terenskih posnetkih in pri medijskem diskurzu ni povsem enotno. Ker smo pri medijskem diskurzu kot regijo snemanja upoštevali območje, kjer medij oddaja, smo ločili tri regije: celotna Slovenija, SV Slovenija, JZ Slovenija, pri terenskih posnetkih pa je regija snemanja opredeljena bolj podrobno (mariborska, ljubljanska, novomeška...). To ustreza resničnemu stanju, lahko pa zmede uporabnika korpusa. S tega stališča bi ta podatek zahteval ponoven premislek.

b) Govorec ima lahko označenih več regij, če je živel dlje časa v različnih. V Gosu pri tem ni določena prioriteta v smislu regija rojstva, regija študija, regija dela... Pri gradnji konkordančnika se je pokazalo, da bi bila praktična na primer ločitev na primarno regijo (kje je govorec odraščal) ter eno ali več sekundarnih regij.

c) Določanje tipa govornega dogodka je zahtevalo največ interpretacije in se je skozi gradnjo Gosa večkrat spreminjalo. Končen nabor tipov govornega dogodka, ki je viden tudi v konkordančniku (pogovor med prijatelji/znanci, pogovor v družini, moderirani program, moderirani pogovor, srednješolska učna ura...), je bil določen šele, ko smo imeli pregled nad celotnim gradivom. Zaradi velike odvisnosti od interpretacije označevalca korpusa je ta podatek, čeprav nadvse zanimiv, najmanj zanesljiv in bi zahteval dodaten natančnejši premislek že v zasnovi korpusa.

Pogovorni zapis vs. standardizirani zapis

Pri transkribiranju gradiva je bila sprejeta odločitev za dvonivojski zapis: pogovorni in standardizirani. Odločitev se je skozi gradnjo in uporabo korpusa potrdila za odlično vsaj z dveh vidikov: (1) definirati standardizirani zapis je bila ena najtežjih nalog zasnove korpusa in zdi se nemogoča naloga, kako predvideti vnaprej vse probleme in priučiti veliko število transkriptorjev, ki so transkribirali gradivo v prvem koraku, da bi enotno zapisovali; (2) z dvonivojskim zapisom smo dobili pomemben vpogled v pogovorne različice besed. Iz povedanega je že jasno, da pri standardizaciji zapisa nismo mogli vnaprej predvideti vseh težav in bi zato bilo smiselno pred morebitno nadgradnjo korpusa pregledati problematične primere in pripraviti morebitne popravke.

Oblikoslovne oznake

V korpusu Gos so bile oblikoslovne oznake dodane avtomatsko, vzete so iz nabora JOS in naučene na pisnih besedilih. To je časovno in finančno nezahtevna rešitev, ji pa vsekakor ne moremo pripisati enake zanesljivosti kot ročno pregledanim oznakam. Težava so na primer podaljšani polglasnik in drugi zapolnjevalci vrzeli, onomatopeje in besedni fragmenti – nič od tega ne znajo na pisnih besedilih naučeni algoritmi ustrezno označiti. Prav tako je treba kritično premisliti, ali resnično lahko brez sprememb prenesemo obstoječe nabore oblikoslovnih oznak, za katere se ve, da so nastajale z mislijo in izkušnjami iz pisnih besedil, na spontani govorjeni jezik. V zvezi s tem je poleg klasifikacije prej navedenih primerov treba razmisliti na primer še o klasifikaciji diskurznih označevalcev, kot so *ja, glejte, veste, dobro, v redu*.

Konkordančnik

O Gosovem konkordančniku v času nastajanja tega poglavja prav veliko žal še ne moremo komentirati, saj šele začenja svojo pot med uporabnike in njihovih odzivov še ni. V njegovo nastajanje je bilo vloženega veliko truda in premisleka s strani več ljudi, vse z osrednjo željo, narediti uporabniku prijazen vmesnik. S stališča strokovnega uporabnika pa lahko avtorici ob zaključku vendarle že ugotoviva nekaj prednosti in nekaj manjkajočih funkcij, ki bi bile uporabne. Pomembne prednosti so možnost poslušanja zadetkov, podrobna opremljenost zadetkov s kontekstnimi informacijami ter prikaz podatkov o pogostosti zadetkov glede na različne značilnosti diskurzov in govorcev (ki deluje hkrati tudi kot filter). Med funkcijami pa pogrešamo brskanje po korpusu (kar je bilo zaradi omejitev projekta opuščeno). Razliko med iskanjem (angl. search) in brskanjem (angl. browse) mislimo v pomenu, da po korpusu iščemo glede na specificirano besedo ali izraz, brskamo pa po zapisih, ki se prikažejo glede na specificirane atribute govorca in/ali diskurza. Funkcija brskanja bi

torej omogočila, da bi prebirali izbrane korpusne zapise, kar bi bilo zanimivo za vsakogar, ki bi se šele želel seznaniti z značilnostmi določenega tipa govora. Primer konkordančnika, ki omogoča obe funkciji, je michiganski korpus akademske govorjene angleščine (MICASE[62]). Druga funkcija, ki bi si je še želeli, je razdvoumljanje zadetkov. Včasih iskalnega pogoja ne moremo postaviti tako, da bi dobili samo želene zadetke, in jih moramo zato ročno izločiti. S tem pa tudi podatki o pogostosti zadetkov v posameznih kategorijah diskurza in govorcev niso več ustrezni, saj izločanja ustreznih zadetkov ne moremo opraviti v konkordančniku, ampak le z naknadno obdelavo podatkov na svojem računalniku.

Potem ko smo še enkrat od začetka premislili in pokomentirali določene rešitve v Gosu, se ozrimo naprej. Ko je govora o korpusih, je odgovor skoraj vedno enak: več gradiva, več oznak. Jasno, oboje bi si za Gos, ki je po obsegu majhen korpus, še kako želeli. Na omejitve, ki izhajajo iz njegovega obsega, moramo biti nadvse pozorni tudi kot uporabniki: marsičesa v korpusu ni, marsikdaj so rezultati povezani samo z enim ali dvema govorcema in seveda posledično nereprezentativni. Gos v sedanjem obsegu nam lahko ponudi odgovore le o najbolj splošnih lastnostih govorjene slovenščine kot celote, odgovore o njenih številnih različicah pa le za najpogostejše izraze.

Toda bolj (in prej) kot o omejitvah in nadaljnjem delu razmislimo o njegovi uporabi v obliki in obsegu, kot ga ima zdaj. Kdo lahko z njim kaj počne? Zdi se, da je to vprašanje, čeprav redko izrečeno, ves čas nekje v ozadju in da odgovore še iščemo. Čeprav jih nekaj ponudi že peto poglavje, naj nanizamo še nekaj zamisli v zvezi s tem.

a) Jezikoslovje: prvi in osrednji uporabnik korpusa je gotovo slovenistično jezikoslovje. Gos je lahko vir iztočnic za slovar govorjenega jezika ali vir informacij za dopolnitev splošnega slovarja z informacijami o govorjenem jeziku. Ker v zasebnem diskurzu posega v vse konce slovenskih regij, je lahko v pomoč tudi pri dialektoloških študijah. Tudi besedotvorje lahko iz njega izlušči uporabne podatke. Do neke mere je morda v pomoč tudi pri glasoslovnih raziskavah, čeprav bi bil za širšo uporabo v teh potreben fonemski zapis. Enako bi za širšo uporabo v skladenjskih raziskavah potrebovali dodane skladenjske oznake.

b) Analiza diskurza: ker je gradivo dobro opremljeno z dodatnimi informacijami o kontekstu in z zvokom, so mogoče mnoge kvalitativne raziskave na posameznih odsekih iz korpusa (npr. z metodo konverzacijske analize, za analizo govornih dejanj, implikacij itd.), vendar je pri tem potreben dostop do izvornega

62 http://quod.lib.umich.edu/m/micase/).

gradiva. Prek konkordančnika pa so mogoče razne kvantitativne raziskave, nekaj smo jih nanizali v poglavju 5.

c) Jezikovne tehnologije: najbolj neposredna uporaba Gosa bi bila za nadgradnjo jezikovnih modelov s komponento za spontani govorjeni jezik. Enako kot v jezikoslovju je Gos tudi vir besedja za slovarje, ki se uporabljajo v raznih aplikacijah jezikovnih tehnologij.

č) Učenje slovenščine: za učenje slovenščine v šolah in tečajih za tujce je Gos med drugim vir primerov, tako da ni več potrebe po uporabi izmišljenih primerov v didaktične namene. Učence morda pritegne tudi kot zabava ob poslušanju avtentičnih primerov govorjene slovenščine.

d) Lektoriranje: lektorji govorjenega jezika bodo v Gosu na primer želeli preveriti, v katerih situacijah in kako pogosto se rabijo določeni izrazi, poiskali pogovorne različice kakšne besede, primerjali različne izgovorjave posameznih besed/izrazov ipd.

e) Literatura, film, gledališče, prevajanje: Gos je izreden vir za spoznavanje značilnosti spontanega govorjenega jezika – zaenkrat sicer omejen s funkcijami konkordančnika, nekoliko širši vpogled omogoča z brskanjem po izvorni datoteki korpusa, ki pa ni uporabniško tako prijazna.

To je samo nekaj zamisli – nedvomno se bodo uporabniki domislili uporabnosti korpusa še za marsikaj drugega, če se bodo le zavedali njegove prisotnosti. Mi pa s tem zaključimo sprehajanje skozi nastajanje in uporabo korpusa govorjene slovenščine. Za konec dodajmo le še eno misel: želeli bi si, da bi Gos predstavljal korak naprej v spoznavanju in neobremenjenem sprejemanju vsakdanje govorjene slovenščine, kot jo govorimo vsi in vsakdo na svoj način, hkrati pa spodbudil željo po boljšem obvladovanju njenih številnih različic.

Povzetek

Korpus Gos je nastal po desetletju razmišljanj in priprav, potem ko je našel svoje mesto v okviru projekta Sporazumevanje v slovenskem jeziku. Pri njegovem snovanju smo sledili trem ciljem: (1) zajeti vzorčne primere različnih govornih situacij in različnih govorjenih diskurzov; (2) zajeti govorjeni diskurz demografsko reprezentativnega vzorca govorcev slovenskega jezika; (3) zajeti predvsem tiste govorne situacije, v katerih so uporabniki jezika najbolj pogosto produktivno-receptivno udeleženi.

S temi cilji v mislih smo definirali demografske (prvi jezik, država bivanja, spol, starost, regijska pripadnost) in besedilnovrstne kriterije (javnost, prenosnik) za zajem gradiv.

Za zajemanje gradiv je bilo treba pripraviti vzorce pogodb in izjav, s katerimi smo skladno z zakonskimi zahtevami reševali prenose avtorskih pravic in varovanje osebnih podatkov. Izbrana je bila uporaba v skladu z licenco Creative Commons: "priznanje avtorstva" + "nekomercialno" + "deljenje pod istimi pogoji". Prav tako je bilo treba najti rešitve glede tehnične opreme za snemanje kot tudi za organizacijo in koordinacijo dela. Poleg specifikacij za zajemanje gradiv so bili potrebni tudi razmisleki, kako posnetke govora zapisati. Ob upoštevanju ciljev korpusa, mednarodnih standardov in praks v slovenskem okolju sta bila definirana dva nivoja transkribiranja, pogovorni in standardizirani zapis. V pogovornem zapisu zapisujemo govor na način, zapiši, kot slišiš. V standardiziranem zapisu zapisujemo govor na način, zapiši, kot pišemo. Transkripcije so bile izdelane z orodjem Transcriber.

V korpus Gos so bili zajeti naslednji tipi govornih dogodkov: vsebinsko zaključeni odseki iz moderiranih vsebin in pogovorov na radiu in televiziji, novinarski prispevki, odseki iz moderiranih oddaj, resničnostnih šovov in športnih prenosov; posnetki osnovnošolskih in srednješolskih učnih ur, predavanj na fakulteti in javnih predavanj; posnetki govora na formalnih in neformalnih delovnih sestankih, konzultacijah na fakulteti, ob raznih storitvah, v formalnih razgovorih, ob prodaji in v trgovini, ob svetovanju, posredovanju informacij itd.; posnetki pogovorov v družini ter med prijatelji/znanci. Vse transkripcije posnetkov so opremljene s podrobnimi informacijami o okoliščinah snemanja in o govorcih. Tekstovni del korpusa je zapisan v standardu XML in v skladu s priporočili TEI ter je javno dostopen.

Korpus je dostopen prek spletnega iskalnika – konkordančnika, ki je postavljen na spletnem mestu www.korpus-gos.net. Zasnova konkordančnika temelji na izčiščenosti in intuitivnem ravnanju uporabnika, hkrati pa omogoča izkoriščanje kompleksnih podatkov o govorcih in diskurzih, primerno za zahtevnejše uporabnike. Konkordančnik nudi dva tipa iskanja: osnovno iskanje, pri katerem se pojavijo zadetki

v obliki konkordanc (izpisi tistih mest, kjer se zadetek pojavi), ter iskanje po seznamu, pri katerem se izpišejo vse realizirane oblike iskane besede v obliki seznama ter s številčnim podatkom o pogostosti rabe.

Prvi primeri uporabe korpusa potrdijo nekatere že predpostavljene ugotovitve o diskurznih označevalcih *aha* in *aja*, pridobljene iz drugih (manjših in specifičnih) gradiv, hkrati pa ponudijo tudi nekaj novih informacij o rabi teh izrazov v posameznih tipih diskurza ter v gručah z drugimi diskurznimi označevalci. Prav tako z uporabo korpusa pridobimo zanimiv vpogled v pogovorne različice prislova *lahko* in njegovo rabo v različnih regijah. Zanimive so tudi ugotovitve o rabi izraza *in tako naprej*, ki v nasprotju s prevladujočo rabo podobnih izrazov v neformalnih situacijah prevladuje v formalnih govornih situacijah.

Delo zaključujemo z željo, da bi korpus Gos predstavljal korak naprej v spoznavanju in neobremenjenem sprejemanju vsakdanje govorjene slovenščine, kot jo govorimo vsi in vsakdo na svoj način, hkrati pa spodbudil željo po boljšem obvladovanju njenih številnih različic.

Summary

The monograph The corpus of spoken Slovene - Gos presents the planning and the design of the first reference corpus of spoken Slovene, as well as the application (concordancer) used to access it. The Gos corpus is an electronic collection of recordings of speakers from all over Slovenia in their most common speech situations. The material includes 110 hours of recordings taken from the media, educational institutions, and personal situations. The recordings have been transcribed, segmented into smaller units and made available online through a user-friendly interface on the website www.korpus-gos.net.

This unique collection of authentic spoken texts in Slovene can be used:

- as a resource for corpus-based language guides,
- for linguistic research and language technologies,
- for teaching Slovene as a mother tongue or as a foreign language,
- by professionals whose work involves a great deal of spoken communication (journalists, moderators, actors)

Taking account of the principal goals of the presented source, the authors describe the process of planning the Gos corpus, the corpus materials and the transcription system based on the pronunciation-based and on the standardized level. The user-friendly interface allows searching for different words, dialects, speakers' profiles and speech situations, thus promoting a better understanding of different demographic and genre layers of Slovene as a mother tongue or as a foreign language.

The monograph presents an original method by which the authors have managed to collect million words of material, secure the copyrights, and protect the speakers' personal information. That is why the Gos corpus is the only spoken corpus in the world that allows free access to authentic recordings to any interested users.

Literatura

1. Blakemore, Diane, 2005: Discourse Markers. Horn, L.R., Ward, G. (ur.): The Handbook of Pragmatics. Oxford: Blackwell. 221-240.
2. Erjavec, Tomaž, Krek, Simon, Arhar, Špela, Fišer, Darja, Ledinek, Nina, Saksida, Amanda, Sivec, Breda, Trebar, Blaž, 2010: Oblikoskladenjske specifikacije JOS, V1.1. Dostopno na: http://nl.ijs.si/jos/msd/html-sl/ (28. 10. 2011).
3. Fraser, Bruce, 1999: What are Discourse Markers? Journal of Pragmatics 31. 931–52.
4. Garg, Saurabh, Martinovski, Bilyana, Robinson, Susan, Stephan, Jens, Tetreault, Joel, Traum David, 2004: Evaluation of Transcription and Annotation tools for a Multi-modal, Multi-party dialogue corpus. V: Proceedings of Fourth International Conference on Language Resources and Evaluation, Lizbona: lrec. 2163-2166.
5. Gorjanc, Vojko, 2005: Uvod v korpusno jezikoslovje. Domžale, Izolit.
6. Hennoste, Tiit, Gerassimenko, Olga, Kasterpalu, Rina, Koit, Mare, Raabis, Andriela, Strandson, Krista, 2008: From human communication to intelligent user interfaces: corpora od spoken Estonian. V: Proceedings of 6th Language Resources and Evaluation Conference, Maroko, Marakeš.
7. Kenda Jež, Karmen, 2004: Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju. Kržišnik, E. (ur.): Obdobja 22. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko. 263–276.
8. Krajnc, Mira, 2005: Besedilne značilnosti javne govorjene besede: Na gradivu sej mariborskega Mestnega sveta. Maribor: Slavistično društvo.
9. Kranjc, Simona, 1999: Razvoj govora predšolskih otrok. Ljubljana: Znanstveni inštitut Filozofske fakultete.
10. Overstreet, Marian, 2005: And stuff und so: Investigating pragmatic expressions in English and German. Journal of Pragmatics 37. 1845–1864.
11. Przepiórkowski, Adam, Górski Rafal, Lewandowska-Tomaszczyk, Barbara, Lazinski, Marek, 2008: Towards the national corpus of Polish. V: Proceedings of 6th Language Resources and Evaluation Conference, Marrakech. Rohlfing. 827-830.
12. Rohlfing, Katharina, Loehr, Daniel, Duncan, Susan, Brown, Amanda, Franklin, Amy, Kimbara, Irene, Milde, Jan-Torsten, Parrill, Fey, Rose, Travis, Schmidt, Thomas, Sloetjes, Han Sloetjes, Thies, Alexandra, Wllinghoff, Sandra, 2006: Comparison of multimodal annotation tools – workshop report. V: Gespraechsforschung – Online-Zeitschrift zur verbalen Interaktion, 7. 99-123.
13. Schourup, Lawrence, 1999: Discourse Markers. Lingua 107. 227–65.
14. Smolej, Mojca, 2006: Vpliv besedilne vrste na uresničitev skladenjskih struktur: Primer narativnih besedil v vsakdanjem spontanem govoru. Doktorska disertacija. Ljubljana: Filozofska fakulteta.
15. Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. Uporabno jezikoslovje 6. Jezikovne tehnologije (tematska št., ur. Z. Kačič). 96–106.
16. Stabej, Marko, Vitez, Primož, 2000: KGB (korpus govorjenih besedil) v slovenščini. Informacijska družba IS'2000: Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan.
17. Šarotar, Dušan, 2007: Biljard v Dobrayu. Ljubljana: Študentska založba.
18. Verdonik, Darinka, 2006: Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora. Doktorska disertacija. Mentorja: M. Stabej in Z. Kačič. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.
19. Verdonik, D., Rojc, M., 2006: Are you ready for a call? - Spontaneous conversations in tourism for speech-to-speech translation systems. V: 5th International Conference on Language Resources and Evaluation, Genova, Italija.
20. Verdonik, Darinka, 2006: Mhm, ja, no, dobro, glejte, eee ...: diskurzni označevalci v telefonskih pogovorih. Jezik in slovstvo 51/2. 19-36.
21. Verdonik, Darinka, 2007: Jezikovni elementi spontanosti v pogovoru: Diskurzni označevalci in popravljanja. Maribor: Slavistično društvo Maribor.
22. Verdonik, Darinka, 2008: Označevanje vrste diskurznih označevalcev. V: Erjavec, Tomaž (ur.), Žganec Gros, Jerneja (ur.): Zbornik Šeste konference Jezikovne tehnologije/Zbornik 11. mednarodne multikonference Informacijska družba - IS 2008. Ljubljana: Institut Jožef Stefan. 25-28.
23. Verdonik, Darinka, 2010: Vpliv komunikacijskih žanrov na rabo diskurznih označevalcev. V: Vintar, Špela (ur.): Slovenske korpusne raziskave. Ljubljana: Znanstvena založba Filozofske fakultete. 88-108.
24. Verdonik, Darinka, 2011: Govorni korpus kot lektorjev priročnik. V: Krakar Vogel, Boža (ur.). Slavistika v regijah - Maribor (Zbornik Slavističnega društva Slovenije, 22). Ljubljana: Zveza društev Slavistično društvo Slovenije. 171-173.

25. **Vitez,** Primož, **Zwitter Vitez**, Ana, 2004: Problem prozodične analize spontanega govora. V: Jezik in slovstvo xlix/6. 3-24.
26. **Weiss,** Peter, 2001: Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša zrc sazu: utemeljitev. Jezikoslovni zapiski 7, 1–2. Ljubljana: Založba zrc. 419–428.
27. Zakon o informacijskem pooblaščencu (ZInfP, ul rs 113/05).
28. Zakon o spremembah in dopolnitvah Zakona o ustavnem sodišču (ZUstS-A, el rs 51/07).
29. Zakon o spremembah in dopolnitvah Zakona o varstvu osebnih podatkov (zvop-1A, ul rs 67/07).
30. Zakon o varstvu dokumentarnega in arhivskega gradiva ter arhivih (zvdagam, ul rs 30/2006).
31. Zakon o varstvu osebnih podatkov (zvop-1, ul rs 86/4).
32. **Zemljarič Miklavčič,** Jana, **Stabej,** Marko, 2005: Building a pilot spoken corpus. Garabik, R. (ur.): Computer Treatment of Slavic and East European Languages. Slovaška, Bratislava. 229-240.
33. **Zemljarič Miklavčič,** Jana, 2006: Korpus govorjene slovenščine. V: Erjavec, T., Žganec Gros, J. (ur.). Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba is 2006. Ljubljana: Institut Jožef Stefan. 124-127.
34. **Zemljarič Miklavčič,** Jana, 2007: Načela oblikovanja govornega korpusa za slovenščino. Doktorska disertacija. Mentorja: M. Stabej in V. Gorjanc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.
35. **Zemljarič Miklavčič,** Jana, 2008: Govorni korpusi. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta.
36. **Zemljarič Miklavčič,** Jana, **Stabej,** Marko, **Krek**, Simon, **Zwitter Vitez**, Ana, 2009: Kaj in zakaj v referenčni govorni korpus slovenščine. Stabej, M. (ur.), Obdobja 28: Infrastruktura slovenščine in slovenistike. Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 437–442.
37. **Zorko,** Zinka, 1995: Narečna podoba Dravske doline. Maribor: Kulturni forum.
38. **Zwitter Vitez,** Ana, **Zemljarič Miklavčič**, Jana, **Stabej**, Marko, **Krek**, Simon, 2009: Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Stabej, M. (ur.), Obdobja 28, Infrastruktura slovenščine in slovenistike, Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko. 437-442.
39. **Zwitter Vitez,** Ana, **Krapš Vodopivec**, Irena, 2011: Korpus govorjene slovenščine (gos) za kakovostno in prijazno učno uro. Bačnik, A. et al. (ur.), Sirikt 2011. Ljubljana: Miška. 309-314.
40. **Zwitter Vitez,** Ana, 2011: Korpus Gos in njegova uporaba v raziskovalne, didaktične in ljubiteljske namene. Kranjc, S. (ur.), Obdobja 30: Meddisciplinarnost v slovenistiki, Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko. 559-564.
41. **Žgank,** Andrej, **Rotovnik**, Tomaž, **Verdonik**, Darinka, **Kačič**, Zdravko, 2004: Baza Broadcast News za slovenski jezik (bnsi) in sistem za razpoznavanje tekočega govora. Informacijska družba is'2004: Jezikovne tehnologije. 94–98.
42. **Žgank,** Andrej, **Rotovnik**, Tomaž, **Grašič**, Matej, **Kos**, Marko, **Vlaj**, Damjan, **Kačič**, Zdravko, 2006: Slovenska govorna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. V: Informacijska družba is'2006: Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan. 115-118.
43. **Žibert,** Janez, **Mihelič**, France, 2004: Development, evaluation and automatic segmentation of Slovenian Broadcast News Speech Database. V: Informacijska družba is'2004: Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan. 94-97.
44. **Welsh,** Irvine, 1997: Trainspotting. Prevod: A. Skubic. Ljubljana: dzs.

Seznam uporabljenih spletnih strani

1. Bank of English, http://mycobuild.com/about-collins-corpus.aspx (27. 7. 2010)
2. BNC, http://www.natcorp.ox.ac.uk/corpus/creating.xml (27. 7. 2010)
3. BYU-BNC, British National Corpus, http://corpus.byu.edu/bnc/ (15. 11. 2011)
4. CLIPS, korpus govorjene italijanščine, http://www.clips.unina.it/it/ (27. 7. 2010)
5. CLUL, Reference Corpus of Contemporary Portuguese (CRPC), http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php (27. 7. 2010)
6. Collins Corpus z Bank of English, http://mycobuild.com/about-collins-corpus.aspx (11. 11. 2011)
7. Corpus de la parole, http://corpusdelaparole.in2p3.fr/ (8. 11. 2011)
8. Corpus del Español, http://www.corpusdelespanol.org/ (8. 11. 2011)
9. CREA – Corpus de Referencia del Español Actual, http://corpus.rae.es/creanet.html (8. 11. 2011)
10. Creative Commons, http://creativecommons.org/licenses/by-nc-sa/2.5/si/legalcode (22. 11. 2011)
11. CRPC – Reference Corpus of Contemporary Portugeese, http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php (11. 11. 2011)
12. Česky narodni korpus, http://ucnk.ff.cuni.cz/ (27. 7. 2010)
13. Čveka forum, http://www.cveka.com (22. 12. 2008)
14. Dereko, http://www.ids-mannheim.de/kl/projekte/dereko/ (8. 11. 2011)
15. Deutsches Spracharchiv, http://dsav-oeff.ids-mannheim.de/DSAv/ (8. 11. 2011)
16. EAGLES (Expert Advisory Group on Language Engineering Standards), http://www.ilc.cnr.it/EAGLES/home.html (27. 7. 2010)
17. Fidaplus, http://www.fidaplus.net/ (8. 11. 2011)
18. Gigafida, http://www.gigafida.net/ (23. 11. 2011)
19. Goeteborg Spoken Language Corpus (GSLC), http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3 (27. 7. 2010)
20. Korpus Gos, www.korpus-gos.net (31. 10. 2010)
21. LABLITA Corpus, http://lablita.dit.unifi.it/corpora/descriptions/lablita/ (11. 11. 2011)
22. Nova beseda, http://bos.zrc-sazu.si/s_beseda.html (8. 11. 2011)
23. Poljski govorni korpus, http://korpus.ia.uni.lodz.pl/conversational/ (9. 8. 2010)
24. Russian National Corpus, http://www.ruscorpora.ru/en/index.html (8. 11. 2011)
25. SAMPA, http://www.phon.ucl.ac.uk/home/sampa/ (22. 11. 2011)
26. Spoken Dutch Corpus/Corpus Gesproken Nederlands (CGN), http://lands.let.kun.nl/cgn/ehome.htm (27. 7. 2010)
27. Sporazumevanje v slovenskem jeziku (SSJ), www.slovenscina.eu (27. 7. 2010)
28. TEI (Text Encoding Iniciative), http://www.tei-c.org/index.xml (27. 7. 2010)
29. TEI priporočila, http://www.tei-c.org/Guidelines/index.xml (27. 7. 2010)

Stvarno kazalo

Priloga: Korpus Gos v številkah

tip diskurza	kanal	tip govornega dogodka	število besed	%
javni informativno-izobraževalni	radio	moderirani program	34958	3,384861
javni informativno-izobraževalni	radio	moderirani pogovor	59578	5,76873
javni informativno-izobraževalni	televizija	novinarski prispevek	20650	1,999467
javni informativno-izobraževalni	televizija	moderirani pogovor	81613	7,902302
javni informativno-izobraževalni	osebni stik	osnovnošolska učna ura	59925	5,802329
javni informativno-izobraževalni	osebni stik	srednješolska učna ura	56087	5,430709
javni informativno-izobraževalni	osebni stik	tečaj	2877	0,27857
javni informativno-izobraževalni	osebni stik	fakultetno predavanje	30956	2,997361
javni informativno-izobraževalni	osebni stik	javno predavanje	12905	1,249546
javni razvedrilni	radio	moderirani program	101255	9,804168
javni razvedrilni	radio	moderirani pogovor	21897	2,12021
javni razvedrilni	televizija	moderirani pogovor	52492	5,082617
javni razvedrilni	televizija	moderirana oddaja	20475	1,982523
javni razvedrilni	televizija	resničnostni šov	10500	1,016678
javni razvedrilni	televizija	športni prenos	22146	2,14432
nejavni nezasebni	osebni stik	formalni delovni sestanek	21320	2,064341
nejavni nezasebni	osebni stik	neformalni delovni sestanek	32820	3,177846
nejavni nezasebni	osebni stik	konzultacija	16885	1,634916
nejavni nezasebni	osebni stik	storitev	31877	3,086539
nejavni nezasebni	osebni stik	formalni razgovor	6066	0,58735
nejavni nezasebni	osebni stik	prodaja/trgovina	11019	1,066931
nejavni nezasebni	telefon	neformalni delovni sestanek	930	0,090049
nejavni nezasebni	telefon	prodaja/trgovina	6989	0,67672
nejavni nezasebni	telefon	svetovanje	13359	1,293505
nejavni nezasebni	telefon	informacije	11474	1,110987
nejavni nezasebni	telefon	tajništvo	732	0,070877
nejavni zasebni	osebni stik	pogovor v družini	90171	8,730943
nejavni zasebni	osebni stik	pogovor med prijatelji/znanci	132736	12,85236
nejavni zasebni	telefon	pogovor v družini	16308	1,579047
nejavni zasebni	telefon	pogovor med prijatelji/znanci	51775	5,013193
			1032775	100

podatki o govorcih	kategorija	število govorcev	število besed
starost	do 10	1	1832
starost	10 do 14	1	31
starost	15 do 18	5	3807
starost	19 do 24	76	129657
starost	25 do 34	28	44606
starost	35 do 59	53	75558
starost	nad 60	22	31648
starost	neznano	0	3851
spol	moški	90	141331
spol	ženski	96	145808
spol	neznano	0	3851
izobrazba	oš ali manj	21	26923
izobrazba	srednja šola	112	169736
izobrazba	višja ali visoka šola	20	27451
izobrazba	fakulteta ali več	31	56810
izobrazba	neznano	2	10070
primarna regija	celjska	16	23524
primarna regija	novogoriška	13	6825
primarna regija	krška	17	24373
primarna regija	koprska	11	4252
primarna regija	kranjska	10	15728
primarna regija	ljubljanska	43	75849
primarna regija	mariborska	22	35430
primarna regija	murskosoboška	12	23922
primarna regija	novomeška	9	20693
primarna regija	postojnska	2	3155
primarna regija	slovengraška	14	14901
primarna regija	Avstrija	6	10792
primarna regija	Italija	5	13488
primarna regija	tujina	5	12403
primarna regija	neznano	1	5655
prvi jezik	slovenščina	179	273001
prvi jezik	angleščina	1	1509
prvi jezik	italijanščina	1	226
prvi jezik	južnoslovanski	3	7827
prvi jezik	drugi romanski	2	4576
prvi jezik	neznano	0	3851