Korpusne oznake za opis konteksta govornih dogodkov v slovenskih govornih korpusih

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

Zaradi časovno in finančno zahtevne priprave govornega korpusa je ob zasnovi potreben temeljit razmislek o njegovi sestavi in kategorizaciji beleženih metapodatkov. Raznoliki govorni dogodki, vključeni v nacionalni referenčni korpus, naj bi v čim večji meri odražali raznolikost sodobnega govorjenega jezika. Zanimalo nas bo, na kakšen način kategorizirati oznake za opis konteksta govornih dogodkov, da bi to reprezentativnost dosegli, ne da bi se popolnoma odrekli medsebojni primerljivosti podatkov. Premišljena zasnova nam omogoča, da je ob kasnejših korpusnih nadgradnjah potrebnih čim manj časovno zamudnih prilagoditev oznak. Izvedli bomo primerjalno analizo domačih in tujih govornih korpusov, s katero bomo kritično ovrednotili štiri temeljne kategorije oznak za opis konteksta govorne situacije. Pregledali bomo zasnovo tujih referenčnih govornih korpusov FOLK, BNC2014, ORAL2013, Nizozemskega govornega korpusa in C-ORAL-ROM ter jih primerjali z aktualnim referenčnim korpusom govorjene slovenščine Gos 2.1. Problematizirali bomo izbrane oznake in izpostavili težavnejša mesta, ki bi zahtevala dodatne premisleke in potencialno prekategorizacijo v prihodnje.

Ključne besede: govorni korpusi, zasnova korpusa, govorni dogodki, kategorizacija oznak

Bizjak, A.: Korpusne oznake za opis konteksta govornih dogodkov v slovenskih govornih korpusih. Slovenščina 2.0, 12(1): 54–94.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2024.1.54-94

1 Uvod

Govorni jezikovni viri so pomembni ne le za jezikoslovno raziskovanje sodobnega govorjenega jezika, temveč tudi za področje razvoja govornih tehnologij in usposabljanja modelov strojnega učenja. V vse bolj digitalizirani družbi so nepogrešljivi pri razvoju razpoznave in sinteze govora, analize sentimenta, prevajalnikov ter velikih jezikovnih modelov. Govorni jezikovni viri, med njimi tudi govorni korpusi, poleg multimodalnih posnetkov in zapisa govora vsebujejo metapodatke o posnetkih, govorcih in govornih dogodkih (npr. lokacija in čas snemanja, spol in starost govorca, tip govora in govornega dogodka). Izbor metapodatkov, ki jih beležimo, med drugim omogoča preverjanje uravnoteženosti in reprezentativnosti govornega vira glede na govorce in govorne situacije. Vsebinske odločitve, katere kategorije metapodatkov pri zasnovi korpusa zajeti in kako podrobno jih opisati, so odvisne od vrste gradiva in namena govornega vira. Tako se posledično ob združevanju različnih govornih jezikovnih virov pojavljajo težave, izhajajoče iz vsebinske nezdružljivosti zabeleženih metapodatkov (Verdonik idr., 2022).

Prav govorni korpusi so tisti, ki med drugim omogočajo primerjalne analize različnih jezikovnih rab glede na regionalno ali dialektološko obarvanost, čas in prostor, demografijo govorcev, družbeno-kulturni kontekst itd. Obsežno množico korpusnih podatkov najpogosteje analiziramo s pomočjo konkordančnikov, napredne avtomatske analitične tehnike, kot je rudarjenje besedil, pa nam omogočajo, da iz velikih količin podatkov izluščimo vzorce in informacije, ki niso takoj razvidni. Posamezne tehnike rudarjenja besedil, kot je npr. tematsko modeliranje v sklopu parlamentarnih razprav (Pretnar Žagar idr., 2022), lahko uporabimo za identifikacijo tematik (Chizhik in Sergeyev, 2021), modeliranje argumentacije (Petukhova idr., 2015), za analizo čustvene zaznamovanosti govora (Rheault idr., 2016) ali stališč govorcev (Abercrombie in Batista-Navarro, 2020). Govorni korpusi raziskovalcem omogočajo razumevanje procesiranja naravnega jezika, izvedbo pragmatičnih ali sociolingvističnih analiz gradiva (Gorjanc in Fišer, 2013), kjer je treba zabeležiti čim več podatkov o kontekstu, ter pravorečne, žanrske in leksikološke analize – korpusne raziskave postajajo vse pomembnejše na različnih raziskovalnih področjih (Vintar, 2010). Po Čermáku lahko

jezikovne podatke delimo na zunanje, neobdelane (pred vključitvijo v korpus) in na notranje, obdelane podatke (zdaj že del korpusa), za katere je značilno poenoteno označevanje (Gorjanc in Krek, 2005). Če so oznake za opis govornih situacij nedosledne, premalo natančne in nekoherentne, imajo raziskovalci in drugi uporabniki težave pri uporabi gradiva ali pa jih celo zavedemo k neustreznemu iskanju po korpusu oz. k neustreznim raziskovalnim rezultatom. Temeljit premislek o sestavi govornega korpusa in kategorizaciji govornih dogodkov, ki naj v čim večji meri odražajo najrazličnejše kontekste, v katerih govorci komunicirajo, pa je ključnega pomena tudi zato, ker je priprava govornega korpusa časovno in finančno izjemno zahtevna.

Kategorije, navezujoče se na situacijski kontekst in udeležence, so pri zasnovi govornih korpusov najzahtevnejši tip kategorij. Hkrati namreč odražajo raznolike značilnosti konteksta v širšem pomenu kot tudi značilnosti govorcev, vključenih v točno določeni govorni dogodek. Še zlasti so problematične podkategorije, navezujoče se na tematiko, predmet pogovora itd., saj ni na voljo nobene splošno priznane taksonomije, ki bi jo lahko upoštevali, zaradi česar se zdijo tako rekoč neskončne (Cermák, 2009). Z dodatnim izzivom se soočimo že pri samem poimenovanju kategorije, ki opisuje govorni dogodek, saj so v rabi različni izrazi (tip ali opis govornega dogodka, domena, žanr itd.). Pojem govorni dogodek, uporabljen v korpusu Gos 1.1, izhaja iz sociolingvistike (Verdonik, 2013), natančneje iz dela o etnografiji komunikacije D. Hymesa (Hymes, 1974) in njegove opredelitve govornega dogodka (*speech event*). V korpusu Artur (Verdonik idr., 2023) se kategorija vrsta govornega dogodka nanaša na najvišjo raven delitve govora (npr. na javni, nejavni, brani govor), kar v Gos 2.1 ustreza kategoriji tip govora (npr. javni razvedrilni, nejavni zasebni). Kategoriji, ki je v nekaterih tujih korpusih poimenovana kot domena (*domain*), v Arturju ustreza kategorija opis govornega dogodka (npr. okrogla miza, seja državnega zbora, prosti dialog med dvema sogovornikoma), v korpusi Gos 2.1 pa tip govornega dogodka (npr. intervju, predavanje, pogovor med prijatelji/znanci). V korpusu FOLK je za isti namen uporabljeno poimenovanje *Gattung*, za razlikovanje med zasebnim, javnim in institucionalnim govorom pa kategorija *Interaktionsdomäne*. V korpusu BNC2014 je za opis različnih tipov uporabljeno poimenovanje *conversation type*,

za podrobnejši ročni vsebinski opis govornega dogodka pa *activity description*. V korpusu ORAL2013 je bil izbran izraz *typ komunikační situace*, v Nizozemskem govornem korpusu *component*, v C-ORAL-ROM pa *genre*. V tem prispevku bomo uporabljali poimenovanja iz najaktualnejše različice referenčnega govornega korpusa pri nas, tj. korpusa Gos 2.1. Za taksonomijo na najvišji ravni delitve govora bomo uporabljali kategorijo tip govora (npr. informativno-izobraževalni govor), za poimenovanja različnih govornih dogodkov pa tip govornega dogodka (npr. osnovnošolska učna ura). Za prosti in podrobnejši opis govornega dogodka bomo uporabili kategorijo opis dogodka (npr. splošno predavanje za prvi letnik prevajalstva) ter kategorijo kanal za označevanje načina zajemanja ali reprodukcije zvočnih podatkov.

Zaradi pomanjkanja poglobljenega premisleka o ustreznosti kategorizacije metapodatkov v slovenskih govornih korpusih je cilj tega prispevka kritično oceniti nabor oznak, ki v korpusih opisujejo kontekst govorne situacije. V procesu vključevanja posnetkov iz korpusov Gos 1.1, Gos VideoLectures in Artur v nadgrajeno različico korpusa Gos 2.1 smo zaznali nekaj nedoslednosti in neusklajenosti, predvsem pa manko kakršne koli razprave na to temo. Z namenom kritičnega pretresa izbranih oznak bomo v 2. poglavju najprej pregledali zasnovo petih tujih referenčnih govornih korpusov in korpusa Gos 2.1. V 3. poglavju bomo predstavili uporabljeno metodologijo, v 4. poglavju pa skušali identificirati kritična mesta izbranih oznak v korpusu Gos 2.1 ter analizirati njihovo konsistentnost v primerjavi z gradivom, kasneje dodanim iz korpusov Gos 1.1, Gos VideoLectures in korpusa Artur. Osredotočili se bomo na primerjalno analizo štirih temeljnih kategorij za opis konteksta govornih dogodkov, to so tip govora (*taxonomy*), tip govornega dogodka (*domain*), opis govornega dogodka (*title*) in kanal (*channel*). Primerjali bomo razlike med odločitvami snovalcev domačih in tujih govornih virov ter analizirali, ali so bolj naklonjeni prostim vnosom ali sistematičnemu beleženju podatkov z uporabo vnaprej opredeljenih odgovorov. V 5. poglavju bomo predstavljene rezultate kritično ovrednotili ter podali potencialne razrešitve identificiranih šibkih mest, ki bi jih bilo pri zasnovi oz. nadgradnji govornih korpusov za slovenščino smiselno upoštevati v prihodnje. Glavne ugotovitve bomo povzeli v 6. poglavju.

2 Pregled področja

Na področju kategorizacije oznak za opis konteksta govornih dogodkov je bila izvedena raziskava (Kopřivová idr., 2019), ki ponuja celovit pregled najrelevantnejših kriterijev, ki jih je pri zasnovi govornega korpusa smiselno upoštevati. Ob upoštevanju različnih teoretičnih pristopov in ob pregledu obstoječih govornih korpusov je bil pripravljen pragmatično utemeljen nabor govornih dogodkov, ki si ne prizadeva biti univerzalen ali dokončen, temveč služi kot vodilo in konceptualni okvir za spodbujanje upoštevanja raznolikosti govornih dogodkov. V ospredju so kriteriji, o katerih lahko sklepamo neposredno iz situacijskega konteksta govornih dogodkov, ne da bi bilo treba za podatke zaprositi udeležence. Dvoumni izrazi, kot je npr. „spontano“, s pomenom neformalno ali nepripravljeno, niso bili vključeni. Kriterije za kategorizacijo govornih dogodkov so avtorji raziskave (Kopřivová idr., 2019) razvrstili glede na kategorijo, povezano z lokacijo snemanja (*setting-related criteria*), in kategorijo, povezano z udeleženci. V prvo sodijo kriteriji, kot so stopnja uradnosti, ki je odvisna od družbene vloge govorca in tega, ali govor poteka v imenu institucije/na slovesnosti (npr. ravnateljev govor na začetku šolskega leta, rojstnodnevna zdravica). Naslednji kriteriji so stopnja javnosti, ki je odvisna od velikosti občinstva in dostopnosti za ožjo ali širšo javnost (npr. politična razprava na TV vs. usposabljanje na delovnem mestu vs. pogovor z odvetnikom); posredovanje komunikacije (govorci so lahko fizično prisotni na istem mestu ali pa govorijo na daljavo preko telefona ali komunikacijskih platform, kot je Skype) ter sinhronost, ki je odvisna od tega, ali pogovor poteka sinhrono v istem času za oba govorca ali pa je čas produkcije govora različen od časa njegove percepcije (npr. posnetki na TV in spletu, ki ne potekajo v živo).

Med kriterije, povezane z udeleženci, so uvrščeni število (aktivnih) govorcev (monolog vs. dialog); stopnja pripravljenosti, ki je odvisna od tega, ali govorec vnaprej pozna tematiko in namen pogovora ali pa je nepripravljen ter odgovore tvori sproti; število naslovnikov (eden ali več); stopnja naslovnikove aktivnosti (dialog vs. občinstvo pri snemanju oddaje v živo vs. gledalci pred televizijskim zaslonom); odnos med udeleženci, ki je odvisen od stopnje skupnega znanja in izkušenj (npr.

razlikovanje med prijatelji, strokovnimi sodelavci in neznanci); stopnja medsebojnega poznavanja ali zaupnosti (družinski člani in prijatelji vs. udeleženci razgovora za zaposlitev); simetrija družbenih vlog, ki je odvisna npr. od starosti in izobrazbe (pogovor med prijatelji iste starosti vs. pogovor med nadrejenim in podrejenim) in socialno-demografske značilnosti (npr. spol, starost, izobrazba, kraj bivanja, poklic itd.).

Da bi s pomočjo primerjalne analize kritično ocenili nabor oznak, ki v slovenskih govornih korpusih opisujejo kontekst govorne situacije, bomo v nadaljevanju najprej pregledali zasnovo petih tujih referenčnih govornih korpusov. Ker zaradi časovnih omejitev ne moremo analizirati vseh, smo izbrali prepoznavnejše in široko uporabljene referenčne govorne korpuse, ki niso domensko omejeni in torej pokrivajo širok spekter govorjenega jezika, hkrati pa so relevantni po svojem obsegu in vplivu, ki so ga po izdaji imeli na oblikovanje drugih korpusov. Eden od kriterijev za izbor je bila tudi dostopnost relevantne znanstvene literature o zasnovi govornega korpusa in njegovih oznakah. Za referenčne korpuse je namreč metodologija gradnje, ki predvideva celo mrežo kriterijev, natančneje izdelana (Gorjanc, 2005). Ker je »kakovost korpusa določena z avtentičnostjo besedil« (prav tam), smo v analizo vključili čim bolj avtentičen spontani govor. Izbrali smo evropske govorne korpuse, ki so nam jezikovnokulturno in znanstvenoraziskovalno blizu, kar nam omogoča bolj neposredno primerjavo in potencialno implementacijo primerov dobrih praks. V analizo smo želeli zajeti govorjeni jezik različnih jezikovnih skupin, da bi v čim večji meri izključili možnost vpliva le-te na označevanje konteksta govornih dogodkov. Analizirali bomo korpuse FOLK, BNC2014, ORAL2013, Nizozemski govorni korpus in C-ORAL-ROM. Slednji je še zlasti ustrezen za kontrastivne študije, saj gre za primerljivi korpus (Gorjanc in Fišer, 2013), ki vključuje primerljiva besedila v štirih romanskih jezikih.

2.1 Korpus FOLK

Korpus FOLK (Forschungs- und Lehrkorpus gesprochenes Deutsch)[1] je korpus pogovorov v nemškem jeziku, obsegajoč širok nabor interakcij v zasebnem, institucionalnem in javnem okolju (Schmidt, 2014).

1 https://agd.ids-mannheim.de/folk.shtml

Opazujemo in beležimo lahko raznolike značilnosti govornih dogodkov, pri čemer za opredelitev določenega govornega dogodka vse seveda niso enako relevantne. Deppermann in Hartung (2012) podajata pregled značilnosti, ki so bile v začetnih fazah snovanja korpusa FOLK del širšega premisleka, posamezne značilnosti pa nato vključene tudi v samo izgradnjo korpusa. Med analizirane značilnosti v fazi načrtovanja sodijo tip govornega dogodka (*Gattung*) in družbeni sektorji, v katere so ti razvrščeni (npr. izobraževanje, gospodarstvo, medicina, prosti čas, religija, politika, umetnost in mediji). Na govorne dogodke vplivajo tudi družbene vloge govorcev v smislu pravic in obveznosti, ki so konstitutivne za njihov vstop v zasebne ali institucionalne interakcije in ki se ne tvorijo šele skozi pogovor (npr. mati, otrok v pogovorih za družinsko mizo; pogovor v krogu ožjih prijateljev, sodnik/obtoženec v sodnih postopkih). Nadaljnja analizirana značilnost je, da so govorni dogodki za udeležence različno dostopni. Razlikujemo zaprte govorne dogodke (intimne/sorodstvene vezi ali pogojevanje dostopa s specifičnimi institucionalnimi vlogami: npr. šepet v postelji, posvet zaprtega tipa), (delno) javne dogodke (npr. sestanek fakultete) in javne govorne dogodke (govor v politični kampanji). Pomembna je tudi zaupnost med udeleženci (neznanci ob prvem stiku, znanci, zaupni odnosi med prijatelji in družino ali mešani odnosi).

Kriterij institucionalnosti govorne dogodke deli glede na to, ali so vloge udeležencev v pogovoru vezane na institucije (npr. svetovalec in svetovanec v svetovalnem razgovoru). Stopnja priprave loči spontane govorne dogodke brez priprave (pogovor na zabavi), vnaprej pripravljene (razgovor za službo), vnaprej oblikovane govorne dogodke (molitev) in glasno branje (novice). Govorni dogodek je lahko vezan na tematiko, kar je odvisno od vrste govornega dogodka in ne konkretnega primera (npr. tematsko fiksirani govorni dogodki, kot sta televizijska politična razprava in sodna obravnava; dogodki, vezani na določeno tematsko področje, kot sta pogovor med zdravnikom in pacientom ali pogovorna oddaja; in nefiksirani govorni dogodki, kot sta pogovor za družinsko mizo in pogovor v gostilni). Opredeljen je lahko tudi namen pogovora (npr. postavitev diagnoze, psihološko svetovanje, pritoževanje znotraj družinske razprave). Pri določitvi opazovanih metapodatkov je treba biti pozoren na število udeležencev – diadni

pogovor (npr. psihoanaliza), triadni (mediacijska pogajanja) ali večosebni pogovor (npr. skupinska terapija), glede na menjavo govorca pa razlikujemo monolog in dialog. Smiselno je upoštevati prisotnost oz. odsotnost občinstva in njihovo vlogo (npr. pogovori z udeleženci, ki primarno niso verbalno aktivni – primer pogovorne oddaje, kjer je občinstvo razpršeno: razlikovanje med gosti pogovorne oddaje, gosti v studiu in televizijskimi gledalci).

Ena od opazovanih značilnosti je lahko tudi medij oz. tehnični vidik prenosa (npr. pogovor iz oči v oči, telefonski pogovor, posredovano preko množičnih medijev). Pri pogovornih oddajah lahko opazujemo notranji komunikacijski krog, kjer bi kot kanal opredelili osebni stik, in zunanji komunikacijski krog, kjer gre za prenos preko množičnih medijev. Naslednji značilnosti sta kraj, ki podrobneje opredeljuje vrsto govornega dogodka (npr. pridiga v cerkvi, družinski pogovor v zasebnem domu, sodna obravnava na sodišču), in čas (npr. pogovor pri zajtrku, ponedeljkov jutranji krožek v osnovni šoli, novoletni govor predsednika). Dodatna značilnost, vezana na čas, je časovna omejenost, ki razlikuje omejeno trajanje govornega dogodka od neomejenega oz. od trajanja, ki ni vnaprej določeno (institucionalni in javni pogovori so npr. skoraj vedno časovno omejeni). Kot zadnjo značilnost obravnavata praktično referenco/sklicevanje (*empraktischer Bezug*), ki opredeljuje govorne dogodke, pri katerih govorjenje ni v središču, ampak ima le dopolnilno oz. koordinacijsko vlogo (Deppermann in Hartung, 2012).

Pri sprejemanju odločitev o tem, katere metapodatke bomo v govornem korpusu dejansko beležili, je treba upoštevati splošno sprejeto soglasje na področju korpusnega jezikoslovja, in sicer da reprezentativnost v statističnem smislu ne more biti cilj zasnove korpusa (Lemnitzer in Zinsmeister v Deppermann in Hartung, 2012). Prvi izziv pri pripravi nacionalnega govornega korpusa, kot navajata Deppermann in Hartung (2012), se pojavi pri sami opredelitvi osnovne populacije, ki naj bi jo vzorčili. Kvantitativno sestavo jezikovne resničnosti je glede na številne parametre izjemno zahtevno smiselno oceniti, npr. koliko voznikov med vožnjo preklinja in kako pogosto, koliko jih poje in kako pogosto. Poleg tega se pri vsakodnevnem sporazumevanju nenehno pojavljajo novi govorni dogodki, kot je npr. hiphop dvoboj, obstoječi pa se preoblikujejo ali izginjajo. Nejasno je tudi, po katerih kriterijih naj

bi posamezne govorne dogodke utežili, npr. dialog vs. skupinska debata; pogovor v živo vs. komuniciranje preko množičnih medijev, kjer ima peščica posameznikov aktivno, množica pa pasivno vlogo. Eden od nadaljnjih izzivov, identificiranih ob pripravi korpusa FOLK (Deppermann in Hartung, 2012), je, da ne obstaja celovit seznam vseh govornih dogodkov, ki se pojavljajo v komuniciranju, saj je znanje o tem razpršeno, velikokrat intuitivno in nikjer v celoti popisano. Zato je iluzorno pričakovati, da bi lahko v korpus vključili vse mogoče kombinacije tako govornih dogodkov kot vseh kategorij empiričnih metapodatkov o govorcih, kot so spol, starost itd. Spodaj navajamo pregled ključnih parametrov, ki so bili vključeni v zasnovo korpusa FOLK (Kaiser, 2018).

Tabela 1: *Stratifikacijski parametri v korpusu FOLK*

	Parameter	Vrednosti
Vodilna stratifikacija	Interakcijska domena	zasebno institucionalno javno drugo
	Življenjsko področje (*Lebensbereich*)	**zasebno**: zasebno (ni določeno) **institucionalno**: izobraževanje, organi/uprava, medpoklicno sporazumevanje, klubske/društvene dejavnosti (*Vereinsleben*), religija/cerkev, kultura (umetnost/zabava/šport), storitvene dejavnosti, medicina/zdravje **javno**: politika, zabava, znanost, gospodarstvo **drugo**: drugo (ni določeno)
	Aktivnosti	**zasebno**: usmerjeno v aktivnosti[2]: odprt vnos (npr. obnavljanje, načrtovanje dopusta); neusmerjeno v aktivnosti **institucionalno**: usmerjeno v aktivnosti: odprt vnos (npr. sestanek, učna ura vožnje) **javno**: usmerjeno v aktivnosti: odprt vnos (npr. mediacija, panelna razprava) **drugo**: usmerjeno v aktivnosti: *maptask*, intervju, eksperimentalna igra

2 Ta oznaka pomeni, da potek aktivnosti določajo npr. neke vnaprej znane tematike ali specifični tipi nalog.

	Parameter	Vrednosti
Dodatno	Medij	iz oči v oči telefon posredovano preko množičnih medijev
	Število udeležencev	natančna specifikacija delitev na interakcijo z dvema, tremi ali več osebami
	Občinstvo	da ne odprto polje
	Zaupnost	poznan neznan zaupen drugo
	Družbene vloge in odnosi	odprt vnos
	Praktična referenca	da ne neznano/mešano
	Jezik(i) interakcije	odprt vnos

Najpogostejše oznake v korpusu FOLK, ki so bile preko prostega vnosa zabeležene za opis tipa govornega dogodka na področju zasebne interakcije, so pogovor ob kavi, pogovor med partnerjema, pogovor v družini, pogovor med prijatelji, pogovor med študenti, pogovor na počitnicah/izletu, pogovor med gospodinjenjem, odrasli igrajo družabne igre, igranje iger z otroki in branje otrokom. Na področju interakcij v šoli/na univerzi gre za učne ure na zasebni srednji šoli, učne ure na poklicni šoli, ustne izpite na univerzi ter povratne informacije med učitelji. Najpogostejši govorni dogodki na področju interakcije na delovnem mestu so sestanek, menjava izmene v bolnišnici, usposabljanje v dobrodelni organizaciji in pogovor na policijski postaji, na področju javne interakcije pa mediacijski pogovori. Drugi tipi interakcije so še navigacijske naloge (*maptask*) in biografski intervjuji (Schmidt, 2014).

2.2 Britanski nacionalni korpus

Na pobudo britanske vlade konec 80. in v začetku 90. let prejšnjega stoletja je v sodelovanju akademskih institucij z gospodarstvom nastal

obsežen Britanski nacionalni korpus (BNC)[3], katerega 10-odstotni delež sestavlja govorjeni jezik, tj. okoli 10 milijonov besed (Zemljarič Miklavčič, 2008), izmed katerih jih približno polovica sodi v spontani govor. BNC je enojezikovni, vzorčni, mešani, sinhroni korpus, ki pa navkljub zavidanja vredni uravnoteženosti predstavlja zgolj zamrznjeno sliko angleščine, kakršna je bila v rabi nekako od sredine 70. do sredine 90. let prejšnjega stoletja (Gorjanc, 2005).

Eden najvplivnejših referenčnih virov za gradnjo govornih korpusov danes je Britanski nacionalni korpus 2014 (BNC2014)[4], katerega govorni podkorpus vključuje več kot 11 milijonov besed. Sestavljen je iz demografsko in kontekstualno uravnoteženega dela (Zemljarič Miklavčič idr., 2015). Ob izgradnji govornega korpusa so snovalci sprejeli odločitev, da bodo zbirali podatke, ki se pojavljajo zgolj v neformalnih kontekstih, saj so ugotovili, da obstaja večja potreba po uporabi in raziskovanju tovrstnih podatkov (Love idr., 2017). Govorni del BNC2014 vsebuje prepise posnetih pogovorov, ki so jih med letoma 2012 in 2016 zbrali predstavniki britanske javnosti. Pogovori so bili posneti v neformalnem okolju (običajno doma) in so potekali med prijatelji in družinskimi člani. Inovativni vidik korpusa je, da so govorci svoje pogovore posneli z vgrajeno napravo za snemanje zvoka v svojih pametnih telefonih. Korpus obsega 1 251 pogovorov, v katerih je sodelovalo skupaj 672 govorcev.[5]

Love in sodelavci (2018) so ugotovili, da obstajajo metapodatki, na osnovi katerih lahko besedila razvrstimo v posamezne kategorije. Takšni metapodatki so število govorcev, snemalno obdobje, leto snemanja in podatki o transkriptorjih. Po drugi strani pa nekaterih metapodatkov, zabeleženih na ravni besedila, ni mogoče uporabiti za kategorizacijo besedil, saj so lahko (in pogosto so) za vsako besedilo različni. Med njimi so opis aktivnosti (možnost prostega vnosa v BNC2014 – snemalci so bili pozvani, naj sami opišejo, kaj se je med snemanjem dogajalo, npr. „par se sprehaja po podeželju in se pogovarja“, njihovi opisi pa so dokumentirani kot dobesedni zapisi); razmerje med govorci (snemalci so na vnaprej

3 http://www.natcorp.ox.ac.uk/
4 http://corpora.lancs.ac.uk/bnc2014/
5 http://corpora.lancs.ac.uk/bnc2014/

določenem seznamu potencialnih razmerij med govorci izbrali eno od možnosti, npr. ožja družina, partnerji, najbližji prijatelji; prijatelji, širši družinski krog; kolegi; znanci); ID-ji govorcev; dolžina posnetka in lokacija posnetka. Nadaljnji metapodatki so še izbrani opisi, značilni za tip pogovora (seznam izbirnih polj, na katerem so sodelujoči lahko izbrali več govornih dejanj, ki so na posnetku, npr. razpravljanje in pojasnjevanje, poizvedovanje, pritoževanje, zahtevanje, vabljenje, svetovanje, razglašanje, pripovedovanje anekdot, dogovarjanje, opravičevanje, kupovanje/prodaja, pripovedovanje šal) in teme (možnost prostega vnosa, kamor so snemalci lahko zapisali vse teme, ki so bile zajete v pogovoru, npr. računalniško programiranje, hrana, vino, savne, odpiranje daril), katerih zapisi so bili dobesedno dokumentirani (Love idr., 2018). Metapodatki o snemanju, ki so jih v celoti izpolnili avtorji posnetkov po svoji lastni presoji, kot na primer posamezne teme pogovora, so, kot rečeno, dobesedno objavljeni v dokumentaciji o korpusu, brez poskusov shematizacije ali standardizacije (Love idr., 2017).

Ker je govor v demografskem delu korpusa pretežno vsakdanji/neformalen, je bil v kontekstualnem delu korpusa namerno dodan nabor jezikoslovno motiviranih besedilnih vrst (Burnard, 2000). Na najvišji ravni tipologije je bila izvedena delitev na štiri enako velike kontekstualno utemeljene kategorije, od katerih je vsaka razdeljena na podkategoriji monolog (v obsegu 40 %) in dialog (v obsegu 60 %):

- **izobraževanje/informiranje**: predavanja, razprave, učne predstavitve, novičarski komentarji, interakcije v razredu
- **poslovanje/gospodarstvo**: pogovori in intervjuji v podjetjih, pogovori s člani sindikata, prodajne predstavitve, poslovni sestanki, posveti (na področju zdravstva, prava, gospodarstva, stroke)
- **javno/institucionalno**: politični govori, pridige, javne razprave, seje sveta, verska srečanja, parlamentarni in sodni postopki
- **prosti čas**: govori, predvajani športni komentarji, pogovori znotraj klubov, predvajane pogovorne oddaje in telefonski pogovori, klubska srečanja

Znotraj vsake podkategorije je bil opredeljen razpon besedilnih vrst, ki ni bil fiksno določen, temveč zasnovan dovolj prožno, da je omogočal vključitev dodatnih besedilnih vrst. Splošen cilj je bil doseči uravnotežen izbor besedilnih vrst ob upoštevanju kategorij, kot so regija, spol in tema. Druge lastnosti, kot je namen, so bile dodane na osnovi naknadnih presoj (Burnard, 2000).

2.3 Korpus ORAL2013

Kot smo v uvodu že navedli, so pri oblikovanju govornih korpusov kategorije, navezujoče se na situacijski kontekst in udeležence, najzapletenejša vrsta kategorij. V Tabeli 2 je prikazanih 12 besedilnih značilnosti oz. atributov, predstavljenih v obliki binarnih opozicij in poskusno razdeljenih v štiri večje skupine. Pri konkretnem posnetku ali besedilu so navedene besedilne značilnosti lahko bodisi prisotne (plus +) bodisi odsotne (minus –), kar pa ne pomeni, da včasih ne obstaja tudi tretja možnost in da je razlikovanje med njimi vedno povsem jasno in enostavno določljivo. Če je vseh 12 kategorij prisotnih (+), naj bi šlo za prototipsko govorjeni jezik, v obratnem primeru za pisna besedila, vendar pogosto prihaja do prekrivanj (Cermák, 2009).

Tabela 2: *Razvrstitev 12 besedilnih značilnosti*

PLUS (+)	**MINUS (–)**
Izvor besedila:	
govorjeno (tj. izvorno)	brano
dialog (tj. izvorno, tipično)	monolog
Medosebni, družbeni odnosi med govorci in fizični kontekst:	
poznavanje/bližina (prijatelji, družina)	govorci si niso blizu
enakost govorcev (družbena, poklicna)	neenakost
zasebno (nejavno)	javno
neformalno	formalno
interaktivno	enosmerno (predavanje, govor, ukazi)
prisotnost (fizična bližina)	oddaljenost (npr. telefon)
eden-na-enega	eden-na-več (občinstvo)
Pristop k temi/situaciji:	
spontano (improvizirano)	pripravljeno (bolj ali manj pripravljeno)

PLUS (+)	MINUS (–)
sproščeno (neformalno)	ustaljeno (*regular*)/uradno (obrednost, protokol)
Zavedanje o snemanju:	
govorec se snemanja ne zaveda	govorec se snemanja zaveda

Korpus ORAL2013[6] je zasnovan kot reprezentativni korpus spontane govorjene češčine, ki se uporablja v neformalnih komunikacijskih situacijah. Vsebuje 291 ur posnetkov in več kot 130 000 besed.

Dejavnike, ki vplivajo na naravo govorjenega jezika, lahko za namen sestave korpusa razdelimo v dve glavni skupini. V prvo skupino uvrščamo predvsem situacijske dejavnike, ki so pomembni za zagotavljanje neformalnosti in so zato merilo za pridobitev posnetkov in njihovo vključitev v korpus. Pri tem imajo ključno vlogo neformalnost komunikacijske situacije ter njena nejavna in neuradna narava; drugi dejavniki vključujejo zlasti zasebno okolje, dialoškost izjav, fizično prisotnost govorcev in njihovo tesno medsebojno povezanost, nepripravljenost in spontanost. Vsi posnetki, vključeni v korpus ORAL2013, izpolnjujejo navedena merila, kar zagotavlja kar največjo prototipičnost zajetih govorjenih posnetkov. Najpogostejši govorni dogodki v korpusu so obisk, pogovor doma, v restavraciji, med skupno dejavnostjo itd. Podatkov iz telefonskih intervjujev, komunikacije prek Skypa ali drugih podobnih situacij namenoma niso zbirali. Glavni cilj je bil ohraniti čim večjo avtentičnost govorcev in njihovih govorov, kar je seveda pogojeno z naravnim okoljem. Zato govorci o snemanju običajno niso bili obveščeni vnaprej, temveč šele po končanem snemanju (Lucie idr., 2015).

2.4 Nizozemski govorni korpus

Korpus govorjene nizozemščine (Spoken Dutch Corpus)[7] ni bil zgrajen za poseben namen ali v interesu (ene same) točno določene skupine uporabnikov. Obsega okoli 800 ur posnetkov in 9 milijonov besed. Glavni parameter, na katerem je korpus zasnovan, je družbeno-situacijsko okolje, v katerem se jezik uporablja. Na osnovi tega lahko

6 https://wiki.korpus.cz/doku.php/en:cnk:oral2013
7 https://lands.let.ru.nl/cgn/doc_English/topics/project/pro_info.htm

razlikujemo več kategorij, ki jih lahko opredelimo glede na situacijske značilnosti, kot so sporazumevalni cilj, medij, število govorcev ter odnos med govorcem in poslušalcem (Oostdijk, 2000).

Tabela 3: *Taksonomija Nizozemskega govornega korpusa*

dialog/ multilog	zasebno		spontano	neposreden stik	pogovor (iz oči v oči)
					intervju
				na daljavo	telefonski pogovor
					poslovna transakcija/posel
	javno	predvajano	bolj ali manj pripravljeno		intervju in diskusija
		nepredvajano (*non-broadcast*)	spontano		diskusija, debata, sestanek
					predavanje
monolog	zasebno		bolj ali manj pripravljeno		opis slike
	javno		predvajano	spontano	spontani komentar
				pripravljeno	novice, programi o aktualnem dogajanju
					informativni bilten
					komentar
			nepredvajano	pripravljeno	predavanje, govor
					brano besedilo

Pri določanju obsega posameznih kategorij korpusa so snovalci korpusa upoštevali več vidikov. Zaradi velike potrebe po podatkih o spontanem govorjenem jeziku je bilo zbiranje naklonjeno nepripravljenemu, spontanemu govoru. Ker interakcija velja za tipično značilnost govorjene komunikacije, so bili obširneje zastopani dialogi in multilogi. Da bi zajeli razlike med posameznimi jezikovnimi zvrstmi v smislu številčnosti njihovih variacij, so bolj heterogene zvrsti v korpusu na splošno zastopane z večjim številom vzorcev kot bolj homogene. O dolžini posnetkov, ki se razlikuje od kategorije do kategorije, zaradi česar je nemogoče napovedati optimalno dolžino posnetka, so se pogosto odločali na osnovi intuitivne presoje. Ob tem ne gre zanemariti dejstva, da je posamezne vrste podatkov lažje zbrati kot druge. Da bi zadostili potrebam določenih skupin uporabnikov, je bilo treba za posamezne kategorije zbrati minimalno količino podatkov, kar še posebej velja za kategorije, ki se uporabljajo za razvoj tehnoloških aplikacij, tj. telefonski pogovori in brana besedila (Oostdijk, 2000).

2.5 C-ORAL-ROM

Korpus C-ORAL-ROM[8] uporablja dve različni strategiji vzorčenja, eno za predstavitev govorjenega jezika v formalnem kontekstu in drugo za neformalni del. Žanr (oz. področje rabe) je z zaprtim spustnim seznamom strogo opredeljen samo pri formalnem govoru, kar pa ne velja za neformalni govor, kjer je omogočen prosti vnos (Cresti idr., 2004).

Tabela 4: *Parametri za opis govorjenega jezika v shematski zasnovi korpusa C-ORAL-ROM*

Kategorija *(type)*	Pod-kategorija	Pod-pod-kategorija
neformalno	zasebno	dialog in multilog
neformalno	zasebno	monolog
neformalno	javno	dialog in multilog
neformalno	javno	monolog
formalno	naravni kontekst (osebni stik)	politični govor, politična razprava; pridiga; poučevanje; strokovna razprava; konferenca; gospodarstvo; pravo
formalno	mediji	pogovorna oddaja; znanstveni tisk; reportaža; intervju; šport; novice; vremenska napoved
	telefon	zasebni pogovor; interakcija človek–stroj

V nasprotju s formalnimi situacijami je za množico situacij, v katerih se uporablja neformalni jezik, značilno, da je odprta in da noben kontekst zanjo ni bolj tipičen od drugega. Zato je bila pri zasnovi korpusa C-ORAL-ROM sprejeta odločitev, da žanri in področja rabe niso vnaprej izrecno opredeljeni. Na ta način je bila na teoretični ravni zagotovljena možnost, da je v korpus vključen kateri koli pomemben žanr (oz. tip govornega dogodka) oz. da noben žanr ni bil obsojen na nično verjetnost pojavitve (Cresti in Moneglia, 2005).

Vzorčenje štirih romanskih jezikovnih virov za korpus C-ORAL-ROM je temeljilo na nizu spremenljivih parametrov, ki tvorijo semiološko in sociološko strukturo korpusa spontanega govora. Ti so dialoška struktura (monologi, dialogi, multilogi); družbeni kontekst rabe, saj so posnetki pogovorov znotraj družinskega in zasebnega življenja ločeni od tistih, ki potekajo v javnosti; kanal (interakcije iz oči v oči, posnetki medijske produkcije in telefonski posnetki), področje rabe,

8 http://www.elda.org/en/proj/coralrom.html

pri čemer gre za področje družbenega okolja, aktivnosti in poklicev, kot so pravo, gospodarstvo, raziskovanje, poučevanje, cerkev itd. Opazovani parametri so še register, saj so posnetki s standardnim jezikom ločeni od posnetkov, na katerih prevladuje neformalna, nestandardna raba jezika, in metapodatki o govorcu, preko katerih se beležijo glavne sociolingvistične značilnosti govorcev, kot so starost, spol, izobrazba, poklic in geografski izvor. Formalni register opredeljujejo dejavniki, kot so javna raba govora, profesionalna raba govora v skladu z družbenimi vlogami, ki jih ima govorec v skupnosti, in namera o izvedbi govorjenega besedila, ki predvideva obravnavo določene tematike, argumentacijo, zaključke itd. Za formalni govor velja, da je lahko tudi spontan, če ta ni (delno) vnaprej pripravljen (Cresti in Moneglia, 2005).

2.6 GOS 2.1

Prvi referenčni govorni korpus za slovenščino, tj. korpus Gos, je izšel leta 2011 (Verdonik idr., 2013). Različica Gos 1.1[9] obsega približno 120 ur transkribiranega govora, posnetega v različnih situacijah: radijske in televizijske oddaje, učne ure in predavanja, zasebni pogovori med prijatelji ali v družini, delovni sestanki, posvetovanja, prodajni pogovori itd. Gos 1.1 vsebuje več kot milijon besed. Med letoma 2017 in 2021 je bil kot dodatek h korpusu Gos izdan korpus Gos VideoLectures (Verdonik, 2018). Govor v korpusu Gos VideoLectures 4.2[10] je javni akademski govor, ki obsega 22 ur v obliki 55 javnih predavanj, dostopnih prek spletnega portala Videolectures.net. Artur 1.0[11] je govorna podatkovna zbirka, zasnovana za potrebe avtomatske razpoznave govora za slovenski jezik. Podatkovna baza vsebuje 1067 ur govora, od katerih je transkribiranih 884 ur. Bazo sestavlja 573 ur branega govora, 62 ur javnega govora (npr. medijski posnetki, spletne konference, delavnice, spletna predavanja), 74 ur nejavnega oz. zasebnega govora (npr. monologi, dialogi kot sproščeni pogovori ali prosti govor o različnih temah) ter 201 ura parlamentarnega govora.

Podatki iz korpusa Artur so bili uporabljeni za nadgradnjo slovenskega referenčnega govornega korpusa Gos (Verdonik idr., 2013). V

9 https://www.clarin.si/repository/xmlui/handle/11356/1438
10 https://www.clarin.si/repository/xmlui/handle/11356/1222
11 https://www.clarin.si/repository/xmlui/handle/11356/1776

zadnjo različico Gos 2.1[12] je vključenih 185 ur iz korpusa Artur 1.0, in sicer vsi transkribirani terenski posnetki javnega in nejavnega govora. Da bi bili podatki čim bolj uravnoteženi, je v sklopu parlamentarnega govora v Gos 2.1 vključenih največ 4000 besed na posameznega govorca. Ti izbrani posnetki so bili skupaj z vsem gradivom Gos Videolectures 4.2 in Gos 1.1 združeni v korpus Gos 2.1. Referenčni korpus Gos 2.1 tako obsega približno 300 ur govora in 2,4 milijona besed. Na voljo je v repozitoriju CLARIN.SI in v novem korpusnem konkordančniku[13]. Brani govor iz korpusa Artur, ki ni bil vključen v referenčni govorni korpus Gos 2.1, pomeni dragoceno podatkovno bazo za fonetične in fonemske raziskave v prihodnje (Verdonik, 2021).

3 Metodologija

S ciljem kritično ovrednotiti nabor oznak za opis konteksta govorne situacije v govornih korpusih Gos 1.1, Gos VideoLectures, Artur in Gos 2.1 bomo skušali odgovoriti na dve temeljni raziskovalni vprašanji. Najprej nas bo zanimalo, kako je nabor oznak primerljiv z izbranimi tujimi referenčnimi govornimi viri, nato pa bomo v izbranih kategorijah oznak skušali identificirati kritična mesta, ki bi zahtevala premislek o njihovi potencialni prekategorizaciji v prihodnje. V primerjalni analizi petih referenčnih tujih govornih korpusov, FOLK, BNC2014, ORAL2013, Nizozemski govorni korpus ter C-ORAL-ROM, in govornih korpusov za govorjeno slovenščino se bomo posvetili štirim kategorijam oznak, ki se v največji meri navezujejo na kontekst govorne situacije. To so tip govora, tip govornega dogodka, njegov podrobnejši opis in kanal. Zanimali nas bodo shematska zasnova primerjanih korpusov, vsebinsko-formalne razlike med posameznimi oznakami ter strategija pristopa k zajemu metapodatkov. Ugotavljali bomo, ali so katere od oznak nezadovoljivo zastopane, nekonsistentne, redundantne, dvoumne ali nejasno poimenovane. Nazadnje bomo nabor oznak ovrednotili še v smislu njihove odprtosti za zajem kar najobsežnejše množice različnih govornih dogodkov v primerjavi z rabo zaprtih spustnih seznamov z omejenim številom vnaprej določenih oznak, ki beleženim metapodatkom

12 https://www.clarin.si/repository/xmlui/handle/11356/1863
13 https://viri.cjvt.si/gos/

omogočajo večjo sistematičnost in medsebojno primerljivost.

Pri analizi zajetih metapodatkov o kontekstu govornih dogodkov v korpusih govorjene slovenščine se bomo osredotočili tudi na postopek preslikave oznak, ki so bile opravljene ob pripravi aktualne različice Gos 2.1. Da bi bili podatki medsebojno čim bolj primerljivi in kompatibilni, so bile posamezne oznake iz korpusov Gos 1.1, Gos Videolectures in Artur ob prenosu v korpus Gos 2.1 prekategorizirane, preimenovane ali dodane. V ta namen bomo analizirali več kot 115 avdio posnetkov, pri katerih smo v predhodno opravljeni primerjalni analizi identificirali težavnejša mesta. Na osnovi vsebine avdio posnetkov bomo oblikovali smernice za morebitne prekategorizacije ali dopolnitve oznak, ki jih bomo predlagali v diskusiji. Pri analizi bomo poleg avdio posnetkov, njihovih zapisov in pregleda tuje literature upoštevali tudi razpoložljivo dokumentacijo o zasnovi in metapodatkih obravnavanih govornih korpusov.

4 Rezultati

Najprej si bomo ogledali, kako je nabor oznak v slovenskih govornih korpusih primerljiv z izbranimi tujimi referenčnimi govornimi viri, v nadaljevanju pa bomo skušali identificirati kritična mesta izbranih oznak in analizirati, iz katerih vidikov nabor oznak ustrezno opisuje kontekst govorne situacije, kateri vidiki pa zahtevajo morebitne dopolnitve.

4.1 Primerjava nabora oznak za opis govorne situacije v domačih in tujih govornih korpusih

Ker so se snovalci korpusov BNC2014 in ORAL2013 odločili, da bodo v govorni korpus vključili zgolj posnetke, ki se pojavljajo v neformalnih kontekstih, neposredna primerjava s tipi govora v korpusu Gos 2.1 ni mogoča. Zanimivejše je opažanje, da so nekateri posnetki, ki so v Gos 2.1 označeni kot nejavni nezasebni govor, v korpusu FOLK uvrščeni v kategorijo institucionalno, v korpusu BNC2014 pa v kategorijo javno/institucionalno. V primerjavi korpusa FOLK s korpusom Gos (Kaiser, 2018) je izpostavljeno, da je komuniciranje, ki je v korpusu FOLK opredeljeno kot institucionalno, v Gos ex negativo opredeljeno kot

nejavno nezasebno.[14] Interakcije na področju izobraževanja, ki v korpusu FOLK sodijo na področje institucionalnega, so v korpusu Gos »iz nejasnih razlogov označene kot javne«.[15] Avtor primerjave se kritično opredeli tudi do kategorije javni govor, znotraj katere pogreša posnetke telefonskih pogovorov. Telefonski pogovori na delovnem mestu so v korpusu FOLK kategorizirani kot institucionalni, v korpusu Gos pa kot nejavni nezasebni (Kaiser, 2018). V nadaljevanju bomo izpostavili še mejno kategorizacijo parlamentarnega govora v Gos 2.1 in jo primerjali z izbranimi rešitvami v tujih korpusih.

V korpusu C-ORAL-ROM je taksonomija v večji meri primerljiva s korpusom Gos 2.1. Poleg zasebnih posnetkov, ki se odvijajo v domačem okolju, so v omenjenem korpusu znotraj neformalnega govora vključeni še posnetki, ki se odvijajo v javnem življenju. V korpusu Gos 2.1 pa so pri nejavnem govoru poleg zasebnih posnetkov vključeni še posnetki nezasebnega govora, navezujoči se na posnetke interakcij na delovnem mestu in v izobraževalnih institucijah. Glede na analizo tujih govornih korpusov bi bilo smiselno razmisliti, da bi morda že na ravni taksonomije v referenčne korpuse za govorjeno slovenščino v prihodnje vključili razlikovanje med družbenimi sektorji oz. področji, v katerih se govorni dogodek odvija (npr. gospodarstvo, politika, zabava, mediji, izobraževanje). Ideja v slovenskem prostoru ni nova, saj se v kontekstu splošnih načel gradnje korpusov, tako pisnih kot govornih, znotraj lastnosti oz. kategorije **besedilni kontekst** kot potencialne oznake omeni izobraževanje, dom, delo in prosti čas (Gorjanc in Logar, 2007).

14 Kot primer za tovrsten tip govornega dogodka navajamo formalni delovni sestanek.
15 Sem sodita na primer osnovnošolska in srednješolska učna ura.

Tabela 5: *Pregled taksonomije oz. tipov govora v izbranih tujih korpusih*

FOLK	zasebno	institucionalno		javno	drugo
	ni opredeljeno	izobraževanje, uprava, medpoklicno sporazumevanje, klubsko/društveno življenje, religija, kultura (zabava, umetnost, šport), storitvene dejavnosti, medicina		politika, zabava, znanost, gospodarstvo	ni opredeljeno
BNC2014	demografski del	kontekstualni del			
	pretežno vsakdanji/neformalen govor	izobraževanje, informiranje	poslovanje, gospodarstvo	javno, institucionalno	prosti čas
		predavanja, učne predstavitve, novičarski komentarji, interakcije v razredu, možnost dodatnega vnosa	pogovori in intervjuji v podjetjih, prodajne predstavitve, poslovni sestanki, posveti, možnost dodatnega vnosa	politični govori, pridige, javne razprave, seje sveta, verska srečanja, sodni postopki, možnost dodatnega vnosa	govori, predvajani športni komentarji, pogovorne oddaje, telefonski pogovori, klubska srečanja, možnost dodatnega vnosa
ORAL2013	neformalno (nejavno)				
Nizozemski govorni korpus	zasebno			javno	
				predvajano (preko množičnih občil) *broadcast*	
				nepredvajano (preko množičnih občil) *non-broadcast*	
C-ORAL-ROM	formalno	neformalno			
		zasebno	javno		
			predavanje univerzitetnega profesorja, diskusija v pogovorni oddaji itd.		

Tipi govornih dogodkov, vključenih v korpus FOLK, so precej podrobno opredeljeni, saj so bili pridobljeni preko prostega vnosa, npr. pogovor med opravljanjem gospodinjskih opravil; odrasli igrajo družabne igre; učna ura na zasebni srednji šoli; ustni izpit na univerzi; menjava izmene v bolnišnici; mediacijski pogovor; biografski intervju itd. (Schmidt, 2014). Slednje potrjujejo tudi izkušnje pri pripravi korpusa BNC2014. Pri pridobivanju metapodatkov so različne značilnosti, nanašajoče se na tip govornega dogodka, kategorizirali v čim manjši meri, saj so ugotovili, da so te lahko (in pogosto so) za vsako besedilo drugačne (Love idr., 2018). Če namreč želimo čim bolj celovito prikazati spontani govor, je bistveno, da z merili za sestavo korpusa omogočamo kar najobsežnejšo raznolikost govornih kontekstov in da si prizadevamo v kar najmanjši meri vplivati na opredelitev kategorije govornega dogodka (Cresti in Moneglia, 2005). Na osnovi pregleda zasnove tujih korpusov se pri poimenovanju tipov govornega dogodka zdi torej smiselno, da zapisovalce metapodatkov v čim manjši meri usmerjamo. Omogočiti jim je treba, da v odprtem, prostem poimenovanju zajamejo čim več vsebinskih podatkov o govornem dogodku, tako da uporabnik korpusa že na osnovi tipa govornega dogodka pridobi informacijo o dogajanju na posnetku.

Glede na zgornjo ugotovitev so tipi govornega dogodka v Gos 2.1 ustrezno kategorizirani, saj odslikavajo raznolik spekter govorjenega jezika. V poimenovanjih posameznih tipov govornega dogodka lahko razberemo celo podatek o družbenem kontekstu (npr. fakultetno predavanje), številu udeležencev (npr. prosti dialog med dvema sogovornikoma) in stopnji pripravljenosti (npr. prosti monološki govor). Sistematizacija tipov govornih dogodkov je uporabna predvsem zaradi medsebojne primerljivosti metapodatkov in možnosti hitrih analiz. Vendar pa je pri tem priporočljivo, da ob vnaprej opredeljenih spustnih seznamih dodamo vsaj še oznako *drugo*. Druga možnost je, da pri beleženju tipa govornega dogodka dopuščamo prost vnos, po zaključenem procesu pridobivanja posnetkov pa te smiselno poimenujemo po posameznih tipih ali pa sistematiziramo zgolj proste opise, pridobljene pod oznako *drugo*.

Nujno se zdi ohraniti dodatno kategorijo, opis govornega dogodka, kjer označevalci metapodatkov na kratko prosto opišejo kontekst

govorne situacije. Takšen način je bil implementiran že v prvi različici korpusa Gos 1.1, podobno rešitev pa najdemo tudi v nekaterih tujih korpusih (npr. BNC2014). S prostim opisom prihodnjim raziskovalcem in drugim uporabnikom zagotovimo čim bolj natančne in celovite vsebinske informacije za raziskovalno delo ali razvoj novih rešitev.

Tabela 6: *Pregled kategorije tip govornega dogodka (domain) v izbranih tujih korpusih*

FOLK	neusmerjeno v aktivnosti	usmerjeno v aktivnosti: **prosti vnos:** načrtovanje dopusta, učna ura vožnje, panelna razprava itd.
BNC2014	demografski del:[16] prosti opis dogodka oz. kratek naslov	
	prosti vnos: par se sprehaja po podeželju in se pogovarja; pogovor o odnosih na kavi s prijatelji; pogovor s sostanovalci med pripravo obeda itd.	
ORAL2013	najpogostejši govorni dogodki: obisk, pogovor doma, v restavraciji, med skupno dejavnostjo itd.	
Nizozemski govorni korpus	opredeljenih je **14 kategorij**: pogovor (iz oči v oči); intervju; telefonski pogovor; poslovna transakcija oz. izmenjava/posel; intervju in diskusija; diskusija, debata, sestanek; predavanje; opis slike; spontani komentar; novinarsko poročilo; programi, ki poročajo o aktualnem dogajanju; informativni bilten; komentar; predavanje, govor; brano besedilo	
C-ORAL-ROM	neformalni govor	formalni govor
	prosti vnos	politični govor; politična razprava; pridiga; poučevanje; strokovna razprava; konferenca; gospodarstvo; pravo; pogovorna oddaja; znanstveni tisk; reportaža; intervju; šport; novice; vremenska napoved

Pri kategoriji kanal je treba upoštevati razlikovanje med govorci in končnimi naslovniki oz. uporabniki. Kanal lahko opredelimo glede na to, kdo so najaktivnejši oz. primarni govorci ali pa glede na to, komu je posnetek namenjen. Kanal radio in televizija na primer določajo končni naslovniki, ne pa osebni stik govorcev v studiu. Na to razlikovanje opozarjata tudi Deppermann in Hartung (2012), ki predlagata razlikovanje med notranjim in zunanjim komunikacijskim krogom. Pri pogovorni oddaji lahko kanal opredelimo glede na notranji komunikacijski krog kot osebni stik, medtem ko ga glede na zunanji komunikacijski krog opredelimo kot množični mediji. Govorce ločujemo na več nivojih: vabljeni gostje pogovorne oddaje (primarno verbalno aktivni

16 Razpon besedilnih vrst znotraj štirih krovnih kontekstualno utemeljenih kategorij ni bil fiksno določen, s čimer je omogočal vključitev dodatnih besedilnih vrst (Burnard, 2000).

govorci), občinstvo v studiu in gledalci pred domačimi TV-zasloni. Tovrstno stopenjsko razvrščanje občinstva je v korpusu FOLK omogočil prost vnos kot dodatna možnost poleg beleženja prisotnosti ali odsotnosti občinstva. Za označevanje samega kanala so bile vnaprej ponujene možnosti: iz oči v oči/osebni stik, telefon in posredovano preko množičnih medijev (+ mešano). Zabeleženi kanal v korpusih BNC2014 in ORAL2013 je osebni stik, v Nizozemskem govornem korpusu se mu pridružuje še oznaka na daljavo (npr. telefonski pogovor). Formalni del govornega korpusa C-ORAL-ROM pri kategoriji kanal razlikuje naravno okolje oz. osebni stik, telefon in množični mediji. Kanali, označeni v korpusu Gos 1.1, so televizija, radio, osebni stik in telefon, v korpusu Gos 2.1 pa mu je zlasti zaradi posnetkov, dodanih iz korpusa Artur, ki so bili posneti v času pandemije covida-19, dodan kanal internet. Pred tem internet v nasprotju s telefonom, osebnim stikom, avdiem in videem še ni bil upoštevan kot poseben prenosnik, saj naj bi bil dojet zgolj kot kanal za prenos zvoka in/ali slike (Zemljarič Miklavčič, 2008).

4.2 Analiza različnih vidikov nabora oznak za opis govorne situacije v slovenskih govornih korpusih

Taksonomija korpusa Gos 2.1, ki je utemeljena na **tipih govora** iz korpusa Gos 1.1, se deli na javni govor in nejavni govor. V javni govor sodita informativno-izobraževalni in razvedrilni govor, v nejavnega pa nezasebni in zasebni govor. Posnetki, ki so bili v korpus Gos 2.1 vključeni iz korpusa Gos VideoLectures, so bili označeni kot informativno-izobraževalni govor. Tako so bili označeni tudi posnetki javnega govora, vključeni v Gos 2.1 iz korpusa Artur, medtem ko so bili posnetki nejavnega govora iz Arturja označeni kot nejavni zasebni govor. Kot omenjeno, je nekoliko diskutabilen parlamentarni govor, ki je bil ob vključitvi iz korpusa Artur v Gos 2.1 označen kot nejavni nezasebni govor. Razmisliti bi bilo smiselno o njegovi prekategorizaciji v javni govor, podrobneje v informativno-izobraževalni govor. Argumentacijo za tovrstno določitev lahko najdemo v korpusu BNC2014, kjer so parlamentarni in sodni postopki ter seje sveta uvrščeni med javna/institucionalna besedila, in v korpusu FOLK, kjer področje „politike“ spada k javni interakcijski domeni.

Ob nadgradnji korpusa Gos v različico 2.1 je bil pripravljen interni predlog za prestrukturiranje oznak znotraj kategorije, imenovane **tip govornega dogodka**. Oznake posnetkov javnega, zlasti medijskega govora, so bile združene v skupine s podobnimi lastnostmi. Predlagane oznake, ki natančneje od izvirnih pojasnjujejo kontekst govornega dogodka, so televizijski novinarski prispevek namesto novinarski prispevek (npr. *POP TV, Preverjeno*), televizijski športni prenos namesto zgolj športni prenos, moderirani radijski program namesto moderirani program (*npr. Val 202, Aktualno*) in jezikovni tečaj namesto tečaj. Precej splošna in za končnega uporabnika nejasna oznaka moderirani pogovor naj bi bila prekategorizirana v televizijski intervju/soočenje (npr. *RTV Slovenija, Odmevi*), televizijsko razvedrilno oddajo (npr. *POP TV, As ti tud not padu*) in radijski intervju (npr. *Ognjišče, Naš gost*). Posamezni posnetki, v Gos 1.1 opredeljeni kot moderirani program, naj bi v skladu s predlaganim prestrukturiranjem sodili k radijskim intervjujem, namesto moderirane oddaje in resničnostnega šova pa naj bi bilo v Gos 2.1 uvedeno poenoteno preimenovanje televizijska razvedrilna oddaja (*npr. POP TV, Vzemi ali pusti*). Implementacija predlaganih sprememb bi vsekakor pripomogla k večji informativnosti in nedvoumnosti metapodatkov in k izboljšanju uporabniške izkušnje. Preimenovanje bi bilo smotrno tudi zato, ker posamezni tipi govornega dogodka s svojo novo oznako hkrati implicirajo tudi že kanal (npr. televizijski novinarski prispevek, radijski intervju).

Posnetkom iz korpusa Gos VideoLectures, ki tipov govornih dogodkov niso imeli beleženih, je bila ob njihovi vključitvi v Gos 2.1 dodana enotna oznaka javno predavanje.[17] Posnetki, dodani iz korpusa Artur, so ohranili svoj izvorni tip govornega dogodka (npr. seja državnega zbora, novinarska konferenca), vendar je treba opozoriti, da v Gos 2.1 niso bili vključeni posnetki, ki še niso transkribirani (npr. seminarji in predavanja, katerih vir sta Arnes in Univerza v Mariboru). Gos 2.0, dostopen v konkordančniku na CJVT,[18] vključuje spodaj navedene tipe

17 V konkordančniku za korpus Gos 2.0, dostopnem na CJVT, jim je bila dodana oznaka predavanje.

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

govornih dogodkov[19] (nejasno je, zakaj se kot tip govornega dogodka dvakrat pojavi oznaka moderirani pogovor):

Seja državnega zbora
Pogovor med prijatelji/znanci
Predavanje
Spletni dogodek
Intervju
Prosti monološki govor
Moderirani pogovor
Pogovor v družini
Prosti dialog med dvema sogovornikoma
Moderirani program
Razlaganje in opisovanje
Moderirani pogovor
Okrogla miza
Osnovnošolska učna ura
Srednješolska učna ura
Storitev
Fakultetno predavanje
Novinarska konferenca
Neformalni delovni sestanek
Športni prenos
Novinarski prispevek
Prodaja/trgovina
Formalni delovni sestanek
Moderirana oddaja
Konzultacija
Svetovanje
Javno predavanje
Informacije
Resničnostni šov
Formalni razgovor
Tečaj
Tajništvo
Nagovor na dogodku

Skozi analizo smo identificirali nekaj diskutabilnih oznak za opis tipa govornega dogodka v korpusu Gos 2.1, za katere bomo v diskusiji predlagali nekaj potencialnih rešitev oz. predlogov za prekategorizacijo metapodatkov v prihodnjih nadgradnjah slovenskih govornih korpusov. Pri javnem govoru v nadaljevanju predlagamo preimenovanje oznak predavanje, intervju, spletni dogodek in (osnovnošolska, srednješolska) učna ura. Tip govornega dogodka predavanje je nejasno označen kar s tremi različnimi oznakami, in sicer se poleg oznak javno predavanje in predavanje v konkordančniku Gos 2.0 pojavlja še oznaka fakultetno predavanje.

Pri nejavnem nezasebnem govoru ugotavljamo, da bi posamezne tipe govornih dogodkov lahko združili, spet druge pa podrobneje vsebinsko opredelili. Mednje sodijo neformalni delovni sestanek, storitev, prodaja/trgovina, informacije, svetovanje in tajništvo. Cilj prekategorizacije tipov govornih dogodkov bi vsekakor moral biti poenotenje, ko gre za govorne situacije v skoraj identičnih kontekstih, saj na ta način dosežemo večjo primerljivost in strukturiranost metapodatkov. V dodatnem opisu govornih dogodkov pa naj zapisovalci prosto in podrobneje zabeležijo podatke o tem, kdo je sodeloval v govornem dogodku,

19 Posamezne oznake v korpusu Gos 2.0 v konkordančniku na CJVT in v korpusu Gos 2.1 v repozitoriju Clarin niso poenotene, kot je npr. oznaka akademski, družboslovje vs. fakultetno predavanje; gimnazija, družboslovje vs. srednješolska učna ura; javno predavanje vs. predavanje.

kdaj in kje se je ta odvijal in kakšne so bile glavne teme ali namen pogovora.

Pri nejavnem zasebnem govoru je analiza pokazala nekaj nekonsistentnosti pri tipih govornega dogodka, označenih kot prosti dialog med dvema sogovornikoma, razlaganje in opisovanje ter prosti monološki govor. Glavni vzrok za nekonsistentno poimenovanje posnetkov z oznako razlaganje in opisovanje je ne povsem usklajen proces pridobivanja posnetkov na več fakultetah.[20] Med snemanjem so se govorcem na zaslonu sproti izpisovala vprašanja, ki so določala potek njihovega govora. Nekaj primerov tovrstnih vprašanj:

- Kateri film vam je bil všeč in zakaj?
- Na kratko opišite zgodbo tega filma.
- Kateri izletniški kraj vam je všeč in zakaj?
- Opišite pot do tega kraja.
- Opišite pripravo jedi, ki vam je všeč.
- Narekujte primer vabila na rojstnodnevno zabavo.

V primerjavi z zgoraj opisanim načinom so imeli govorci pri snemanju posnetkov z oznako prosti monološki govor več svobode pri izbiri tematike. Vendar pa tudi tukaj monologi niso povsem prosti, saj so jih s (pod)vprašanji usmerjali snemalci ali pa so se vprašanja izpisovala na zaslonu. Za boljše razumevanje konteksta snemalne situacije navajamo del navodil:

a) prosto narekovanje pametnim napravam,[21] brez predloge. Govorec si je moral predstavljati, kot da narekuje pametnemu telefonu ali drugi pametni napravi, ta pa njegovo besedilo zapisuje (npr. prosto narekovanje SMS-sporočil, govorno iskanje po spletu). Isti govorec je lahko v enem posnetku uporabil več različnih tem:

    - Kako pripraviš svojo najljubšo jed?
    - Po kateri poti greš od doma do npr. šole, fakultete, službe ali kakšne turistične destinacije?

20 Koordiniranje snemanja nejavnega govora za govorno bazo Artur je potekalo na FE UL in FERI UM.

21 Koordiniranje snemanja je potekalo na FERI UM.

- Pametni napravi narekuj vabilo na dogodek s podatki kaj, kdaj, kje.
- Pametnemu asistentu narekuj katero koli drugo primerno temo (vsebovati ne sme sovražnega ali žaljivega govora).

b) prosti monološki govor, delno voden s sprotnimi kratkimi podvprašanji s strani moderatorja.[22] Monološki govor je sproti usmerjal moderator ali pa so bile okvirne tematike govorcem predlagane vnaprej. Na posnetkih se pojavljajo poljubne kombinacije teh tematik, včasih pa tudi povsem prosti monolog o poljubnih temah, kot je npr. opis postopka izgradnje objekta in priprave dokumentacije zanj, opis nakupovanja ali nastanka glasbenega banda itd.

Pri kategoriji **Opis govornega dogodka** gre za prosti opis govornega dogodka po presoji snemalca oz. zapisovalca metapodatkov. Prosti opisi iz korpusa Gos 1.1 so bili v nespremenjeni obliki preslikani v aktualno različico korpusa Gos 2.1, enako velja za naslove predavanj iz korpusa Gos VideoLectures. Pri izdelavi korpusa Artur zbiranje metapodatkov za to kategorijo ni bilo predvideno, zato bi jo bilo treba posnetkom dodati ročno v eni od prihodnjih nadgradenj. S tem bi podatke poenotili ter raziskovalcem in drugim uporabnikom referenčnega korpusa Gos 2.1 omogočili pridobivanje natančnejših informacij o kontekstu govornega dogodka. Nekateri primeri takšnih opisov iz Gos 2.1 so obnavljanje zavarovalne police za hišo; glasbena oddaja o domači glasbi z gosti in nagradno igro ter predavanje ob dnevu fakultete o najnovejših odkritjih v vesolju.

Ker želimo, da referenčni korpus čim bolj celovito odslikava raznolikost spontanega govora, moramo po vzoru tujih korpusov dopuščati kar najbolj prosto opredelitev oznak za opis konteksta govornega dogodka. V primeru rabe formalnega jezika namreč lahko predvidimo vrsto tipičnih govornih dogodkov, v katerih je glede na določen družbeno-zgodovinski kontekst formalna raba jezika prednostna, kar pa ne velja za področje neformalnega govora. Formalni govor je zato mogoče učinkovito prepoznati tako, da opredelimo njegove najbolj tipične kontekste rabe. V nasprotju s tem pa je množica situacij, v katerih se uporablja neformalni jezik, odprta in tipov govornega dogodka ne moremo opredeliti s

22 Koordiniranje snemanja je potekalo na FE UL.

pomočjo seznama tipičnih kontekstov rabe. Noben kontekst ni namreč bolj tipičen od drugega. Z natančno opredelitvijo žanra (oz. tipa govornega dogodka) je nedvomno mogoče doseči višjo stopnjo primerljivosti podatkov, vendar pa lahko s tem izgubimo oz. izločimo pomembna področja rabe neformalnega govora, kjer so posamezni žanri še vedno v veliki meri neraziskani (Cresti in Moneglia, 2005).

Oznake znotraj kategorije **kanal** so pri preslikavi iz korpusov Gos 1.1 in Gos VideoLectures v Gos 2.1 ostale nespremenjene. Kanala televizija in radio sta v korpusu Gos 2.1 določena glede na končne uporabnike, medtem ko sta kanala osebni stik in telefon opredeljena glede na govorce. Kategorija kanal je bila dodana pri posnetkih iz korpusa Artur, saj metapodatki tam niso bili beleženi. Pri govornih dogodkih seja državnega zbora, okrogla miza, nagovor na dogodku, novinarska konferenca, prosti dialog med dvema sogovornikoma in prosti monološki govor je bil dodan kanal osebni stik. Pri govornem dogodku intervju je bil kanal pri preslikavi oznak opredeljen kot radio, pri intervjujih, domensko poimenovanih kot spletni dogodek, pa je kanal označen kot internet. Pri slednjih gre pretežno za intervjuje, posnete z namenom objave na spletu, podobni primeri bi lahko bili tudi podkasti. Pri govornih situacijah, kot sta okrogla miza in novinarska konferenca, je priporočljivo razmisliti o podrobnejši podkategorizaciji udeležencev, in sicer na primarne govorce, to so vabljeni govorniki, za katere je tipičen kanal osebni stik, od fizično prisotnega (bolj ali manj aktivnega) občinstva ter uporabnikov, ki dogodek spremljajo ali preko televizije ali v digitalnem formatu na internetu.

5 Diskusija

Na osnovi primerjalne analize s tujimi govornimi korpusi in pregleda obstoječega nabora oznak v slovenskih govornih korpusih podajamo nekaj predlogov za potencialno prekategorizacijo metapodatkov za opis konteksta govorne situacije v prihodnje. Znotraj kategorije **tip govornega dogodka** bi zaradi večje medsebojne primerljivosti podatkov oznaki javno predavanje in predavanje poenotili v oznako javno predavanje. Slednja namreč izraža, da je predavanje namenjeno širšemu občinstvu. Oznaka fakultetno predavanje je po našem mnenju

ustrezna, saj natančneje opredeljuje tip predavanja. Glede na to, ali je predavanje predvajano preko interneta ali poteka v živo, se primerno določi kanal, zaradi vse večje digitalizacije pa bi bilo vendarle smiselno razmisliti tudi o vpeljavi nove oznake spletno predavanje. Skoraj vsi intervjuji v korpusu Gos 2.1 so radijski intervjuji, zato bi jih tako tudi preimenovali. Za intervjuje, predvajane preko interneta, bi bilo smiselno preimenovanje v spletne intervjuje oz. spletne pogovore. Slednja oznaka je primerna tudi za posnetke, na katerih več udeležencev debatira preko spletne platforme (npr. *posnetki STA kluba*).

Kot navedeno, je oznaka spletni dogodek ena od najbolj neustreznih oznak za poimenovanje tipa govornega dogodka v korpusu. Ob nadgradnji korpusa Gos 2.1 je bila prenesena iz korpusa Artur in podaja informacijo o kanalu, ne pa o govornem dogodku samem. Razlog za to odločitev je treba razumeti v duhu časa epidemije covida-19, ko je večina posnetih dogodkov za bazo Artur potekala preko interneta in se je oznaka spletni dogodek zdela ustrezna. V času po epidemiji ugotavljamo, da bi bilo treba posnetke ponovno analizirati in za vsak govorni dogodek posebej na novo določiti oznako. Za posnetke, katerih vir je STA, predlagamo preimenovanje v spletni posvet ali spletno soočenje (*npr. predstavitev kandidatov za predsednika Atletske zveze Slovenije*) in (moderirani) spletni pogovor (*npr. o rakavih bolnikih in pripravah na pandemijo covida-19*). Za posnetke, ki sta jih v korpus Artur prispevala Univerza v Mariboru in ZRC SAZU, predlagamo oznaki spletna konferenca in spletna (panelna) razprava, nekatere dodatne možnosti so še video navodilo (*tutorial*), spletna delavnica in spletni seminar (*webinar*).

Premisliti je treba tudi, ali tip govornega dogodka, kot sta osnovnošolska in srednješolska učna ura, sodi v javni govor ali bi bilo morda bolj ustrezno, da ga vključimo v nejavni nezasebni govor. V korpusu FOLK spada oznaka izobraževanje na področje, ki se imenuje institucionalno, v korpusu BNC2014 pa na področje, imenovano izobraževanje/informiranje. Ustreznejše poimenovanje za neformalni delovni sestanek, ki označuje posnetke, na katerih se zaposleni v trgovini (*npr. lovski ali ribiški*) sproščeno pogovarjajo, bi morda bilo pogovor med sodelavci. Po vzoru tujih korpusov, ki s poimenovanjem tipov govornih dogodkov precej natančno opredeljujejo vsebino govornega dogodka, in z namenom doseganja čim večje obvestilnosti bi predlagali

prekategorizacijo oznake storitev v prodajna predstavitev, pogovor med uslužbencem in stranko ter v inštruiranje. Sem spadajo posnetki iz korpusa Gos 1.1, kot so pogovor arhitekta s stranko o izgradnji bazena, fizioterapevtke s pacientko in pogovor v frizerskem salonu. Glede na vse večjo digitalizacijo delovnih procesov bi med metapodatke lahko umestili dodatno oznako, in sicer spletni sestanek. V primeru, da bi na višji ravni taksonomije dodali predlagani podatek o družbenem sektorju, kot je npr. gospodarstvo, storitvene dejavnosti ali trgovina, bi tip govornega dogodka prodaja/trgovina preimenovali v (telefonski) pogovor med prodajalcem in stranko. Gre za govorne dogodke, kot so pogovor med prodajalcem prevlek in stranko, nakup rož v cvetličarni, (telefonski) pogovor med stranko in prodajalcem v zlatarni ali pogovor med stranko in uslužbenko ponudnika telekomunikacijskih storitev. Posamezni posnetki, v Gos 1.1 označeni kot informacije, so vsebinsko skoraj identični dogodkom, označenim kot prodaja/trgovina, zato bi jih bilo smiselno poenotiti v (telefonski) pogovor med prodajalcem/informatorjem in stranko, podrobnejše podatke o govornih dogodkih, kot so bančna uslužbenka posreduje informacije o ponudbi banke ali podajanje informacij s strani klicnega centra podjetja, ki trži telekomunikacijske storitve, pa bi navedli v kategoriji opis govornega dogodka. Oba posnetka z oznako svetovanje bi vsebinsko natančneje opredelili kot (telefonski) pogovor med svetovalcem in stranko (npr. posnetek o svetovanju glede energetsko varčne gradnje), oba posnetka z oznako tajništvo pa kot telefonski pogovor med sodelavci (npr. telefonski pogovor med tajnico in predavateljem).

Če bi želeli tipe govornega dogodka v različnih jezikovnih virih, kot sta Artur in Gos 2.1, čim bolj poenotiti, bi prosti dialog med dvema sogovornikoma preimenovali v pogovor v družini ali pogovor med prijatelji/znanci. Če se za preimenovanje ne odločimo, je navedba med dvema sogovornikoma v vsakem primeru redundantna, saj dialog že nakazuje število udeležencev. Za oznako razlaganje in opisovanje predlagamo, da se poimenovanje poenoti v vodeni monološki govor (oz. vodeni monolog). Posamezni posnetki nejavnega govora v korpusu Artur, ki so sicer vsebinsko primerljivi, so bili namreč zaradi različnih načinov snemanja na več fakultetah nekonsistentno poimenovani. V primerjavi s posnetki z oznako razlaganje in opisovanje so imeli govorci

pri snemanju posnetkov z oznako prosti monološki govor več svobode pri izbiri tematike, vendar, kot smo že videli, tudi tu ne gre za povsem proste monologe. Na osnovi tega bi bilo smiselno tip govornega dogodka preimenovati v delno vodeni monološki govor (oz. delno vodeni monolog).

Tabela 7: *Predlog preimenovanja oznak za kategorijo tip govornega dogodka*

Obstoječe poimenovanje	**Alternativna poimenovanja**
Javni govor	
• javno predavanje • predavanje	javno predavanje
nova oznaka	spletno predavanje
intervju	• radijski intervju • spletni intervju • spletni pogovor
spletni dogodek	• spletni posvet • spletno soočenje • (moderirani) spletni pogovor • spletna konferenca • spletna (panelna) razprava • video navodilo (tutorial) • spletni seminar (webinar)
Nejavni nezasebni govor	
neformalni delovni sestanek	pogovor med sodelavci
storitev	• prodajna predstavitev • pogovor med uslužbencem in stranko • inštruiranje
nova oznaka	spletni sestanek
prodaja/trgovina	(telefonski) pogovor med prodajalcem in stranko
informacije	• (telefonski) pogovor med prodajalcem in stranko • (telefonski) pogovor med informatorjem in stranko
svetovanje	(telefonski) pogovor med svetovalcem in stranko
tajništvo	telefonski pogovor med sodelavci
Nejavni zasebni govor	
prosti dialog med dvema sogovornikoma	• pogovor v družini • pogovor med prijatelji/znanci • prosti dialog
razlaganje in opisovanje	• vodeni monološki govor • vodeni monolog
prosti monološki govor	• delno vodeni monološki govor • delno vodeni monolog

Ugotovili smo, da je pri kategoriji **opis govornega dogodka** priporočljivo dopustiti, da zapisovalci v kratkem zapisu s svojimi besedami podrobneje opišejo kontekst govornega dogodka. V korpusu Artur opisi govornega dogodka niso bili beleženi, zato bi jih bilo v prihodnje smiselno ročno dodati. Za govorni tip prosti monološki govor bi bilo v opisu smiselno izpostaviti pogosto ponavljajoče se teme:

- Opis priprave (najljubše) jedi
- Narekovanje ali podajanje ukazov pametni napravi
- Govorno iskanje po spletu
- Opis poti
- Opis poteka (delovnega) dne
- Opis dela na vrtu
- Opis vsebine najljubšega filma
- Opis najljubšega kraja in njegovih znamenitosti itd.

Na osnovi analize zasnov tujih govornih korpusov podajamo najprej v tabelarični in nato še v opisni obliki nekaj usmeritev za vključitev priporočenih sklopov podatkov, pri čemer ni nujno, da vsak opis govornega dogodka zajame vse sklope. Zapisovalce metapodatkov lahko usmerimo in jim priporočimo zapis podatkov, ki so za opis govorne situacije še posebej relevantni, hkrati pa jim dopuščamo kritično možnost vnosa glede na konkreten govorni dogodek in glede na že zabeležene metapodatke.

Tabela 8: *Predlog atributov za prosti opis govornega dogodka*

Atribut	Opis	Primeri
Zaupnost med govorci	Označuje stopnjo medsebojnega poznavanja med govorci.	družinski člani, prijatelji, neznanci
(Družbene) vloge udeležencev	Označuje vnaprej določene pravice in obveznosti, konstitutivne za različen tip govorčevih interakcij.	mati in otrok, sodnik in obtoženec, moderator in intervjuvanec
Namen oz. cilj	Označuje namen govornega dogodka.	izmenjava mnenj, prepričevanje, postavitev diagnoze, pritoževanje, pripovedovanje šal
Tematika	Označuje glavno in/ali dodatne teme govornega dogodka.	računalniško programiranje, kulinarika

Atribut	Opis	Primeri
Stopnja pripravljenosti	Označuje stopnjo vnaprejšnje pripravljenosti vsebine in/ali strukture govornega dogodka.	(delno) spontani, (delno) vodeni, vnaprej pripravljeni
Interaktivnost oz. vloga občinstva	Označuje morebitno prisotnost občinstva in možnost njegovega vključevanja v govor.	vabljeni gostje – primarni govorci, občinstvo v studiu, gledalci v domačem okolju
Število aktivnih govorcev	Označuje število govorcev, ki sodelujejo v govornem dogodku.	monolog, dialog, multilog
Kraj, čas	Označuje kraj in čas govornega dogodka.	v šoli, v tihem studiu, ponedeljkov jutranji krožek

Ena od pomembnih oznak je seznanjenost govorcev oz. medsebojna zaupnost, ki opredeljuje, kako dobro se govorci med seboj poznajo. V korpusu FOLK je opredeljenih več oznak z različnimi stopnjami zaupnosti, kot so npr. poznan, neznan in zaupen. V korpusu BNC2014 je bil implementiran zaprt spustni seznam z vnaprej določenimi oznakami, kot so ožja družina, partnerji, najbližji prijatelji vs. prijatelji in širši družinski krog, medtem ko so bili v korpus ORAL2013 vključeni zgolj govorci, ki so medsebojno tesno povezani, se dobro poznajo ali so intimni. Na govorni dogodek vplivajo tudi (družbene) vloge udeležencev. Te se, kot v korpusu FOLK, navezujejo na pravice in obveznosti, konstitutivne za govorčev vstop v zasebne ali institucionalne interakcije, zanje pa je značilno, da se ne tvorijo šele med samim govornim dogodkom, temveč so znane vnaprej. Takšne vloge so npr. mati in otrok v domačem okolju, sodnik in obtoženec v sodnem postopku, moderator in intervjuvanec, učitelj in učenec, uradnik in stranka itd.

Korpusa BNC2014 in Nizozemski govorni korpus v opisu navajata tudi namen oz. cilj govornega dogodka. Sem lahko uvrščamo izmenjavo mnenj, pogajanje, prepričevanje, pojasnjevanje, postavitev diagnoze, psihološko svetovanje, pritoževanje, poizvedovanje, opravičevanje, pripovedovanje šal itd. Po vzoru korpusa BNC20214 sodi v opis govorne situacije tudi oznaka tematike pogovora (omemba glavne in dodatnih tem), kot sta npr. računalniško programiranje in kulinarika. Govorne dogodke lahko razlikujemo glede na to, ali so vezani na specifično tematiko/področje ali pa so tematsko neopredeljeni. Govorni dogodek je odvisen tudi od stopnje pripravljenosti, saj je dogodek

(delno) spontan, (delno) voden ali pa moderiranje poteka v živo in sproti. Njegova struktura je lahko določena, kar pomeni, da se govorci nanj pripravijo vnaprej ali pa se jim vprašanja sproti izpisujejo oz. se jih sproti glasovno usmerja. Korpus ORAL2013 vsebuje izključno spontani, nepripravljeni govor, Nizozemski govorni korpus pa ločuje oznake spontani, pripravljeni in bolj ali manj pripravljeni.

Naslednja oznaka, ki podrobneje določa govorni dogodek, je interaktivnost oz. vloga občinstva. Navesti je priporočljivo morebitno prisotnost občinstva in v nadaljevanju opredeliti, ali ima občinstvo možnost takojšnjega in interaktivnega podajanja povratnih informacij ali ne. Če doslej metapodatka o številu in menjavi aktivnih govorcev (tj. tistih, ki kaj povedo) še nismo zajeli, ga je treba navesti vsaj na tem mestu (npr. monolog, dialog, multilog). Ključne informacije o kontekstu govornega dogodka pridobimo tudi iz oznak, kot sta kraj in čas govornega dogodka. Korpusa BNC2014 in C-ORAL-ROM snemalcem omogočata prosti opis kraja (npr. v tihem studiu ali pridiga v cerkvi) in časa (npr. ponedeljkov jutranji krožek v osnovni šoli), korpus BNC2014 pa je za leto snemanja uporabil spustni seznam. Beležimo lahko celo časovno omejenost, kjer razlikujemo govorne dogodke, ki so časovno omejeni (npr. govorilne ure), od tistih, katerih trajanje je neomejeno.

Kar se kategorije **kanal** tiče, bi bilo smiselno razlikovati med tem, ali je dogodek prvenstveno namenjen izvedbi v živo in je na spletu objavljen zgolj njegov posnetek (v tem primeru bi kanal označili glede na govorce oz. govorni dogodek) ali pa je prvenstveno namenjen za objavo na spletu in se v živo odvija zgolj zato, da je posnetek bolj zanimiv ali avtentičen, na primer v prisotnosti občinstva (v tem primeru bi kanal označili glede na končne uporabnike, torej kot internet). Takšen primer so seje državnega zbora, kjer bi kanal glede na končne uporabnike lahko bil opredeljen kot televizija ali internet, saj so seje na voljo tudi preko spletnih strani Parlameter[23]. Podobna dilema se pojavi tudi pri predavanjih, objavljenih na spletni platformi VideoLectures. V Gos 2.1 je kanal v obeh primerih označen kot osebni stik, čeprav bi ga glede na končne uporabnike lahko opredelili tudi kot internet. Posamezna netranskribirana predavanja, ki sicer niso bila dodana v korpus Gos 2.1, najdemo pa jih v korpusu Artur, so bila zaradi epidemije

23 https://parlameter.si/

covida-19 od samega začetka snemalnega procesa načrtovana kot spletna predavanja. Smiselno je torej, da je oznaka za kanal pri tovrstnih posnetkih internet in ne osebni stik. Takšni govorni dogodki so na primer nekatera predavanja, objavljena na Arnesovem spletišču in spletna predavanja za zaposlene na Univerzi v Mariboru.

6 Zaključek

Da bi kritično ovrednotili nabor oznak za opis konteksta govorne situacije, smo v raziskavi primerjali nabor oznak v slovenskih govornih korpusih z izbranimi rešitvami v tujih referenčnih govornih virih. Z analizo oznak v izbranih kategorijah smo identificirali posamezna kritična mesta in zanje podali nekaj usmeritev in predlogov za potencialno prekategorizacijo. Dotaknili smo se tudi internih priporočil za prestrukturiranje oznak znotraj kategorije tip govornega dogodka, ki so bila oblikovana ob nadgradnji korpusa Gos v različico 2.1. Ker zgolj delna implementacija omenjenih priporočil pomeni nekonsistentnost označevanja, bi bilo temu v prihodnje priporočljivo nameniti dodatno pozornost.

V zvezi s poimenovanjem tipov govornih dogodkov predlagamo čim manj usmerjanja pri beleženju metapodatkov, kar omogoča celovitejši odraz raznolikosti sodobnega govorjenega jezika. Zapisovalcem omogočimo prosti vnos ali pa na spustnih seznamih, ki povečujejo medsebojno primerljivost podatkov, dodamo vsaj še oznako *drugo*. Gre za kategorije, kakršni sta tudi besedilna vrsta in govorni položaj, ki so v kontekstu korpusnega jezikoslovja (pisni in govorni korpusi) razumljene kot »jezikovno in kulturno najbolj specifične« (Gorjanc in Logar, 2007). Z vnaprejšnjo kategorizacijo oznak tvegamo, da posamezne govorne dogodke izpustimo ali pa jih nezadostno in nekoherentno opišemo. Pri posnetkih, označenih kot spletni dogodek, predlagamo ponovno analizo in določitev nove oznake, saj trenutna podaja informacijo o kanalu in ne o govornem dogodku. Razmislek o potencialni prekategorizaciji je potreben tudi pri posnetkih, ki so v Gos 2.1 označeni kot nejavni nezasebni govor, v tujih korpusih pa uvrščeni v kategorijo institucionalno (FOLK) ali javno/institucionalno (BNC2014). Podobno velja za javne govorne dogodke, kot so npr. osnovnošolske in srednješolske učne ure, za katere je najbrž primernejša oznaka nejavni

nezasebni tip govora. Druga možnost je, da na ravni taksonomije dodamo diferenciacijo po (družbenih) področjih, kot so gospodarstvo, politika, zabava, mediji in izobraževanje. Na področje politike bi uvrstili parlamentarni govor, pri katerem po vzoru tujih govornih korpusov podajamo pobudo za prekategorizacijo v javni govor. Nekaj predlogov za alternativna preimenovanja oznak kategorije tip govornega dogodka podajamo v diskusiji v Tabeli 7, strnjen pregled priporočenih smernic za opis govornega dogodka pa v Tabeli 8. Ročne opise govornega dogodka bi bilo v prihodnje treba dodati vsem posnetkom, ki so bili ob nadgradnji v korpus Gos 2.1 dodani iz korpusa Artur. Kanal lahko diferenciramo glede na (fizično) prisotnost ali odsotnost občinstva ali glede na stopnjo aktivnosti govorcev. Pri posnetkih sej državnega zbora in predavanj s spletišča VideoLectures bi tako kanal lahko opredelili kot televizija/internet in ne kot osebni stik.

Zaradi hitrega razvoja velikih jezikovnih modelov, ki temeljijo na obsežnih zbirkah besedil, bo vse večjo vlogo pridobivala avtomatska identifikacija žanrov (*automatic genre identification*), ki bo avtomatsko pridobljena besedila, značilna za velike spletne korpuse, obogatila z oznakami o tipu govornega dogodka (npr. promocijsko, pravno besedilo). Obetavne rezultate nudijo že veliki jezikovni modeli sami, ki za avtomatsko identifikacijo žanrov niso učeni (*zero-shot*). To velja npr. za modela GPT-3.5 in GPT-4, ki sta se izkazala tudi na področju zunajdomenskih in večjezičnih virov ter pri identifikaciji žanrov znotraj manjših jezikov, kot je slovenščina. Vendar se je treba zavedati njihovih omejitev, kot sta zamuden ročni pregled dobljenih rezultatov in skrita arhitektura, ki onemogoča poznavanje kriterijev za razvrščanje. Za učinkovito rabo teh modelov potrebujemo strokovno znanje o tvorjenju učinkovitih ukazov (*prompt engineering*), pri nekaterih, kot so Falcon, pa tudi visoko zmogljivo strojno opremo. Vse to otežuje dostopnost raziskav posameznikom in organizacijam z omejenimi viri (Kuzman et al., 2023). Natančnejše in hitreje generirane rezultate zagotavlja prosto dostopen, na XLM-RoBERTa temelječ in preko ročno označenih podatkov prilagojen (*fine-tuned*) žanrski klasifikator (*genre classifier*),[24] ki omogoča avtomatsko identifikacijo tipov govornih dogodkov v številnih jezikih. V sklopu raziskave je bil podan tudi zanimiv

24 https://huggingface.co/classla/xlm-roberta-base-multilingual-text-genre-classifier

predlog za enotno žanrsko shemo za združevanje žanrsko raznolikih podatkovnih baz s pomočjo naslednjih oznak: informacije/razlaga (npr. raziskovalni članek, biografija), navodila (npr. recept, tehnična pomoč), pravni tekst (npr. licenca za programsko opremo, pogodba), novice (npr. športno poročilo, policijsko poročilo), mnenje/argumentacija (npr. blog, politična propaganda), promocija (npr. oglas, vabilo na dogodek), proza/lirika (npr. pesem, šala), forum (npr. forum, vprašanja in odgovori – Q&A) in drugo (Kuzman et al., 2023). Ne glede na to, ali tipe govornih dogodkov označujemo avtomatsko ali ročno, pa si je treba prizadevati, da zabeleženi metapodatki čim celoviteje odražajo sodobno govorjeno slovenščino v vseh njenih najsubtilnejših različicah in pestrosti kontekstualno pogojenih pojavitev.

Zahvala

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

Literatura

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. *Journal of Computational Social Science, 3*(1), 245–270.

Burnard, L. (2000). The British national corpus users reference guide. In: Oxford University Computing Services Oxford.

Cermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics, 14*, 113–123. Pridobljeno s https://www.jbe-platform.com/content/journals/10.1075/ijcl.14.1.07cer

Chizhik, A. V., & Sergeyev, D. A. (2021). Exploring the Parliamentary Discourse of the Russian Federation Using Topic Modeling Approach. International Conference on Digital Transformation and Global Society.

Cresti, E., & Moneglia, M. (2005). C-ORAL-ROM: *Integrated Reference Corpora for Spoken Romance Languages*. John Benjamins Publishing Company. Pridobljeno s https://books.google.si/books?id=ybc5AAAAQBAJ

Cresti, E., Nascimento, F. B. d., Moreno-Sandoval, A., Véronis, J., Martin, P., & Choukri, K. (2004). The C-ORAL-ROM CORPUS. A Multilingual Resource

of Spontaneous Speech for Romance Languages. International Conference on Language Resources and Evaluation.

Deppermann, A., & Hartung, M. (2012). Was gehört in ein nationales Gesprächskorpus? : Kriterien, Probleme und Prioritäten der Stratifikation des "Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim).

Gorjanc, V. (2005). *Uvod v korpusno jezikoslovje*. Izolit.

Gorjanc, V., & Fišer, D. (2013). *Korpusna analiza* (2. izd. ed.). Znanstvena založba Filozofske fakultete.

Gorjanc, V., & Krek, S. (Ur.). (2005). *Študije o korpusnem jezikoslovju: zbornik* (1. izd. ed., Vol. 130). Krtina.

Gorjanc, V., & Logar, N. (2007). Od splošnih do specializiranih korpusov-načela gradnje glede na njihov namen. *Razvoj slovenskega strokovnega jezika*, 637–650. Pridobljeno s https://doi.org/https://repozitorij.uni-lj.si/Dokument.php?id=182750&lang=slv

Hymes, D. H. (1974). Foundations in Sociolinguistics: An Ethnographic Approach.

Kaiser, J. (2018). Zur Stratifikation des FOLK-Korpus: Konzeption und Strategien. *Gesprächsforschung*, *19*, 515–552. Pridobljeno s https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/8668/file/Kaiser_Zur_Stratifikation_des_FOLK-Korpus_2018.pdf

Kopřivová, M., Komrsková, Z., Poukarová, P., & Lukeš, D. (2019). Relevant Criteria for Selection of Spoken Data: Theory Meets Practice. *Journal of Linguistics/Jazykovedný casopis*, *70*(2), 324–335. doi: 10.2478/jazcas-2019-0062

Kuzman, T., Mozetič, I., & Ljubešić, N. (2023). Automatic Genre Identification for Robust Enrichment of Massive Text Collections: Investigation of Classification Methods in the Era of Large Language Models. *Machine Learning and Knowledge Extraction*, *5*(3), 1149-1175. Pridobljeno s https://www.mdpi.com/2504-4990/5/3/59

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, *22*, 319–344. Pridobljeno s https://www.jbe-platform.com/content/journals/10.1075/ijcl.22.3.02lov?crawler=true

Love, R., Hawtin, A., & Hardie, A. (2018). *The British National Corpus 2014: User Manual and Reference Guide (version 1.1)*. ESRC Centre for Corpus Approaches to Social Science.

Lucie, B., Michal, K., & Martina, W. (2015). Korpus spontánní mluvené češtiny ORAL2013.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and First Evaluation. International Conference on Language Resources and Evaluation,

Petukhova, V., Malchanau, A., & Bunt, H. (2015). Modelling argumentation in parliamentary debates. Proceedings of the 15th Workshop on Computational Models of Natural Argument, Principles and Practice of Multi-Agent Systmes Conference (PRIMA 2015), Bertinoro, Italy,

Pretnar Žagar, A., Pahor de Maiti, K., & Fišer, D. (2022). What's on the agenda?: topic modelling parliamentary debates before and during the COVID-19 pandemic = Kaj je na dnevnem redu?. Pridobljeno s https://sidih.github.io/agenda/index-sl.html

Rheault, L., Beelen, K., Cochrane, C., & Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PloS one*, *11*(12), e0168843.

Schmidt, T. C. (2014). The Research and Teaching Corpus of Spoken German — FOLK. International Conference on Language Resources and Evaluation,

Verdonik, D. (2013). Koncept konteksta v jezikoslovnih in diskurznih teorijah. *Slavistična revija*, *61*(4), 631–650. Pridobljeno s https://srl.si/sql_pdf/SRL_2013_4_08.pdf

Verdonik, D. (2018). Korpus in baza Gos Videolectures.

Verdonik, D. (2021). *Govorni viri za pravorečje*. 1. slovenski pravorečni posvet. Pridobljeno s https://www.sazu.si/uploads/files/publikacije21/Rared2RAZPRAVE.pdf

Verdonik, D., Bizjak, A., Žgank, A., Bernjak, M., Antloga, Š., Majhenič, S., Čakš, P., ..., & Bordon, D. (2023). *ASR database ARTUR 1.0 (audio)*. Faculty of Electrical Engineering and Computer Science, University. Pridobljeno s https://www.clarin.si/repository/xmlui/handle/11356/1776

Verdonik, D., Bizjak, A., Žgank, A., & Dobrišek, S. (2022). Metapodatki o posnetkih in govorcih v govornih virih: primer baze Artur. Pridobljeno s https://nl.ijs.si/jtdh22/pdf/JTDH2022_Verdonik-et-al_Metapodatki-o-posnetkih-in-govorcih-v-govornih-virih-primer-baze-Artur.pdf

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., & Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, *47*, 1031–1048. Pridobljeno s https://link.springer.com/content/pdf/10.1007/s10579-013-9216-5.pdf

Vintar, Š. (Ur.). (2010). *Slovenske korpusne raziskave* (1. natis ed.). Znanstvena založba Filozofske fakultete.

Zemljarič Miklavčič, J. (2008). *Govorni korpusi* (1. natis ed.). Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

Zemljarič Miklavčič, J., Stabej, M., Krek, S., & Zwitter Vitez, A. (2015). *Kaj in zakaj v referenčni govorni korpus slovenščine*. Pridobljeno s http://www.korpus-gos.net/Content/Static/Kaj_in_zakaj_v_referencni_govorni_korpus_slovenscine.pdf

Corpus annotations for describing the context of speech events in Slovene speech corpora

The time-consuming and costly preparation of a speech corpus requires careful consideration of its composition and the categorization of the recorded metadata at the time of its design. The variety of speech events included in the national reference corpus should reflect the diversity of contemporary spoken language as much as possible. We will be interested in how to categorize the annotations used to describe the context of the speech events in order to achieve this representativeness without completely giving up the intercomparability of the data. A thoughtful design allows us to minimize the time-consuming tag adjustments required in subsequent corpus upgrades. We will carry out a comparative analysis of domestic and foreign speech corpora to critically evaluate the four basic categories of annotations used to describe the context of a speech situation. We will review the design of the foreign reference speech corpora FOLK, BNC2014, ORAL2013, the Spoken Dutch Corpus and C-ORAL-ROM and compare them with the current reference corpus of spoken Slovene Gos 2.1. We will problematize the selected annotations and highlight the more problematic areas that would require further consideration and potential future re-categorization.

Keywords: speech corpora, corpus design, speech events, categorization of annotations