Klara Šumenjak
Znanstveno-raziskovalno središče Univerze na Primorskem
klara.sumenjak@zrs.upr.si

Zasnova dialektološkega korpusa na primeru govora Koprive na Krasu

V prispevku je predstavljena metodološka zasnova dialektološkega korpusa GOKO *(*Govorni korpus Koprive na Krasu*). Predstavljena so nekatera metodološka vprašanja in rešitve, potrebne pri načrtovanju dialektološkega korpusa. Na koncu je predstavljena tudi uporabnost takega korpusa, saj ni namenjen le dialektologom, ampak tudi širši javnosti in v izobraževalne namene.*

1 Uvod

Korpusno jezikoslovje je študij in opis jezika na podlagi empiričnega gradiva, ki vključuje tudi oblikovanje metodologije za gradnjo korpusov in njihovo dejansko gradnjo. Gorjanc definira korpus kot »računalniško zbirko besedil oz. delov besedil, zbranih po enotnih kriterijih za namene različnih, predvsem jezikoslovnih raziskav«. Poleg tega še ugotavlja, da »pri terminu *korpus* gre za elektronske, torej računalniško berljive besedilne zbirke, ki so (a) enovite, (b) notranje strukturirane in (c) standardno označene glede na namen samega korpusa v skladu z obstoječimi standardi za njihovo gradnjo« (2005: 7). Čeprav so korpusi v svetu začeli nastajati okoli leta 1960, je za Slovenijo prelomna letnica 1997, ko se je začela gradnja prvega referenčnega korpusa FIDA,[1] ki mu je sledilo še nekaj splošnih ali specializiranih korpusov. Ravno avtorji prvega slovenskega referenčnega korpusa so opozarjali na vedno večjo potrebo po izdelavi korpusa govorjene slovenščine, ki bi omogočil bolj celostno analizo slovenskega jezika (Stabej in Vitez 2000; Zemljarič Miklavčič 2011: 21). Čeprav so se vsi zavedali obsežnosti in težavnosti projekta, je ta stekel, zato imamo tudi v Sloveniji referenčni govorni korpus slovenskega jezika GOS, izdelan v

1 Korpus FIDA je referenčni sinhroni korpus, v katerega so vključena besedila (cca 100 milijonov besed), nastala v letih 1994–2000.

okviru projekta *Sporazumevanje v slovenskem jeziku* (*www.korpus-gos.net/Support.aspx/About*), ki lahko predstavlja izhodišče za dialektološki korpus, saj gre v obeh primerih za zapis govorjenega jezika.

2 Dialektološki korpus

Četudi lahko govorjeni jezik razumemo kot primarno obliko jezikovnega sporazumevanja, so se slovenski jezikoslovci z raziskovanjem spontanega govora s pomočjo korpusne metode začeli ukvarjati relativno pozno. Najprej se naučimo jezika okolja, narečja, šele kasneje knjižnega jezika, ki je normiran in zato v primerjavi z narečji precej bolj statičen. Narečja pa se precej hitro spreminjajo na vseh jezikovnih ravninah, zato je bistvenega pomena prav njihovo zapisovanje in ohranjanje. Prvi se je že leta 1841 začel ukvarjati z znanstvenimi raziskavami slovenskih narečij ruski jezikoslovec Sreznjevski, ki je izdal prispevek *O narečjih slavjanskih.* Po tem so bili sestavljeni različni korpusi pisnih besedil za slovenščino, npr. referenčni korpusi (*FIDA*, *Beseda*, *Nova beseda* idr.), vzporedni korpusi (vzporedni angleško-slovenski korpus *ELAN*, vzporedni angleško-slovenski korpus *TRANS*, *Evrokorpus* idr.), specializirani korpusi (korpus *Verbalni napadi na JNA* (VAYNA), korpus vojaških besedil *Grizold*, večjezični turistični korpus *TURK* idr.) in tudi nekateri govorni korpusi. Leta 2011 je nastal prvi referenčni korpus govorjene slovenščine *GOS*. Kljub temu da se je raziskovanje slovenskih narečij začelo že pred približno 160 leti, v Sloveniji (še) nimamo dialektološkega korpusa (v pomenu, ki ga navaja Gorjanc). Čeprav obstajajo potrebe za njegov nastanek, je to zaenkrat še neizvedljivo. Do danes nimamo v Sloveniji niti dialektološkega korpusa, ki bi nastal na podlagi govora vsaj ene vasi, kaj šele referenčnega dialektološkega korpusa. Zamisel za sestavo se nam je porodila, ko smo hoteli opisati govor Koprive na Krasu na bolj inovativen, sodoben način, s pomočjo tehnologije, ki je sedaj na voljo. Nismo želeli, da to postane, kot mnogi drugi, zgodovinski dokument, ampak smo korpus želeli ponuditi najširši javnosti, to je uporabnikom spleta – od tod ideja za dialektološki korpus *GOKO* (*Govorni korpus Koprive na Krasu*).

V nadaljevanju članka bomo prikazali zasnovo dialektološkega korpusa govora Koprive na Krasu. Glede na to, da je to pionirsko delo, je bil naš namen predvsem poiskati najustreznejšo metodologijo, ki je bistvenega pomena za gradnjo dobrega korpusa, in sploh preveriti, ali je tak korpus mogoč. V naslednjih podpoglavjih je predstavljen del[2] metodoloških vprašanj, pomembnih za gradnjo korpusa. Če bi se predstavljeni metodološki pristop izkazal za učinkovitega, bi seveda lahko na enak način v podatkovno zbirko uredili tudi ostala slovenska narečja.

2.1 Obseg korpusa

Obseg vsakega korpusa je odvisen od njegovega namena (več o tem glej npr. Vintar 2008, Logar 2009, Mikolič in Beguš 2011). Če bi želeli predstaviti slovenski referenčni dialektološki korpus, bi moral biti veliko obsežnejši, primerljiv vsaj z obsegom

2 Zaradi omejenosti s prostorom bo celotna metodološka zasnova korpusa predstavljena ob drugi priložnosti.

korpusa *GOS*. Glede na to, da gre pri korpusu *GOKO* za govorni korpus, ki je po Gorjančevi definiciji sorazmerno majhen, a jezikoslovno izjemno bogato označen (2004: 8), bo začetni modelni korpus obsegal približno 10.000 besed, kar je zaenkrat dobro izhodišče za izdelavo ustrezne metodologije za gradnjo korpusa pa tudi za analizo zbranega gradiva. Menimo, da je za poskus gradnje takega korpusa število besed reprezentativno, saj nam poleg analize govora na glasoslovni in leksikalni ravnini, ki ne potrebujeta tako velikega vzorca besed, omogoči tudi analizo na oblikoslovni in skladenjski ravnini. Glavna prednost korpusa je, da se bo lahko neprestano večal. Vanj bomo lahko poljubno dodajali nova besedila, saj je korpus – kot jezik – živ organizem, ki se neprestano spreminja in vase sprejema nove besede in tako celovitejše odraža podobo jezika skozi čas.

2.2 Avtorske pravice

Že v začetni fazi načrtovanja gradnje korpusa moramo biti pozorni na avtorske pravice, potrebne za uporabo in objavo besedil, saj se lahko zgodi, da gradivo že posnamemo, prepišemo, a nato ne dobimo soglasja avtorja za uporabo. Zelo dobro je pridobiti dovoljenje za objavo tako transkribiranih besedil kot zvočnih posnetkov, saj je programska oprema toliko napredovala, da nam omogoča multimedijsko predstavitev, kjer lahko besedilo hkrati poslušamo in beremo.

Zelo je pomembno, da informante, preden začnemo z zbiranjem gradiva, vprašamo za dovoljenje za objavo zbranega gradiva, čeprav je znano, da se ljudje, ko vedo, da so snemani, vedejo nekoliko drugače kakor običajno, saj niso sproščeni, počutijo se »opazovani«. Prvotna zamisel je bila, da bi informantom pred pričetkom terenske raziskave povedali, da jih bomo ves mesec snemali. Tako bi se izognili vsakokratni zadregi ob vklopu snemalnika, ki je danes lahko že tako majhen, da je za informante nemoteč. Gradivo smo nameravali zbirati en mesec, praviloma vsak dan, da bi vaščani to lažje sprejeli in se navadili na izpraševalčevo prisotnost.

2.3 Zbiranje gradiva

Odločili smo se za dva načina pridobivanja gradiva, in sicer za snemanje spontanega govora ter za vodenje usmerjene vprašalnice. Izbrani informanti (gl. Izbira informantov) so bili snemani v najrazličnejših govornih položajih, med vsakdanjimi opravili, pogovorom po telefonu, s sosedo, z možem … Zaradi lažjega zapisovanja zbranega gradiva smo se odločili, da v korpus vključimo predvsem njihove pripovedi, ki so jih delili z izpraševalko. Prosili smo jih, naj nam opišejo razna opravila,[3] naj nam povedo, kako skuhajo tipične jedi, kako obdelujejo zemljo, povprašali pa smo jih tudi o starih običajih, o njihovem življenju nekoč in tako dobili dovolj obsežno in raznoliko zbirko besedil, ki smo jih vključili v korpus. Usmerjena vprašalnica bo izdelana naknadno, ko bo že opravljena oblikoslovna analiza zbranega gradiva. Možno je, da v besedilih spontanega govora ne bomo dobili določenih slovničnih oblik besed (npr. sam. v im. mn. I. ž. skl.), zato bomo te vključili v usmerjeno vprašalnico.

3 Izpraševalkina vprašanja in komentarji so zapisani v oglatem oklepaju in ne bodo vključeni v korpusno analizo, saj njen govor ni koprivski.

2.4 Izbira informantov

Pri vsaki narečni raziskavi je izbira informantov zelo pomembna. Pomembno je, da se jih za opis določenega govora uporabi več, saj se v nasprotnem primeru pojavi nevarnost, da začnemo zapisovati informantov idiolekt. Ker je bila dialektološka raziskava[4] Koprive že opravljena, imamo v vasi že nekaj ustreznih in preverjenih informantov. Čeprav nekateri dialektologi menijo, da je bolje, če informant ni izobražen, sami na podlagi lastnih terenskih izkušenj v Koprivi menimo, da je dober informant lahko tudi nekdo, ki ima končano višjo/visoko šolo. Velikokrat je to učitelj ali pa visoko izobražena oseba, ki se zaveda razlik med knjižnim in narečnim jezikom, vendar zavestno neguje svoje narečje. Ponavadi so ti odgovori informantov uporabni za zapis leksike, saj so za analizo skladenjskih vzorcev najverjetneje preveč pod vplivom knjižnega jezika. Za raziskavo smo izbrali 6 informantov, starih med 70 in 90 let. Da bo korpus čim bolj uravnovešen, smo v raziskavo vključili 3 ženske in 3 moške. Pomemben kriterij pri izbiri informantov je tudi, da so vse življenje živeli v Koprivi in tako na njihov govor ni vplival kateri drug krajevni govor (prim. Vprašalnica za SLA: informant/informator).

2.4.1 Demografsko vzorčenje

Če bi šlo za splošni govorni korpus, bi morali z demografsko metodo statistično določiti vzorec govorcev, ki bi ustrezal celotni izbrani populaciji. Kriteriji, ki se jih lahko upošteva pri vzorčenju govorcev, so: spol, starost, regijska pripadnost, etnična pripadnost, izobrazba, poklic in socialni status.

Poleg teh obstajajo tudi drugi demografski kriteriji, ki so pomembnejši za gradnjo specifičnih korpusov – npr. kraj bivanja, kraj rojstva, verska pripadnost idr. (Zemljarič Miklavčič 2008: 59). Za naš korpus lahko zanemarimo regijsko in etnično pripadnost, saj sta pri vseh vaščanih enaki. Tudi socialni status informantov, ki bodo vključeni v raziskavo, je enak. V korpusu bomo posebej označili spol, starost, izobrazbo in poklic informantov, dodali pa bomo še informacijo, od kod so starši in (morebitni) mož/žena informanta, saj lahko ta dejstva vplivajo na spreminjanje določenih značilnosti govora.

2.5 Transkripcija[5]

V slovenski dialektologiji se uporablja t. i. nova nacionalna transkripcija, ki sledi osnovnim načelom transkripcije *Slovanskega lingvističnega atlasa*. Slovenski jezikoslovci še niso uskladili načina zapisovanja govorjenega besedila oziroma njegovega standardiziranja, ko gre za natančnost fonetičnega zapisa (Ivančič Kutin 2011: 80–81). Ker želimo, da je korpus namenjen tudi širši javnosti, smo zbrano gradivo zapisali v treh različicah, ki bodo opremljene tudi z zvočnim posnetkom, in sicer: a) v fonetičnem zapisu; b) v poenostavljenem zapisu in c) v poknjiženi varianti.

4 Šlo je za zapis 1989 besed vprašalnice za *NASIK – Narečni atlas slovenske Istre in Krasa* (1. 1. 2007–31. 12. 2009) – temeljni raziskovalni projekt na UP ZRS, vodja Goran Filipi.

5 Natančneje o transkripciji v naslednjem članku.

V primerjavi s priporočili TEI (*Text Encoding Initiative*) smo se odločili, da bomo zanemarili nekatere kinezične dogodke (kimanje, skomiganje z rameni ...), saj bi ti otežili gradnjo korpusa in niso (razen pri skladenjski analizi) relevantni za opis koprivskega govora, ohranili pa smo premore in zvoke obotavljanja (prim. Zemljarič Miklavčič 2008: 97–99).

2.5.1 Komentirani primeri transkripcije besedila v korpusu

Fonetični zapis je zapisan v skladu s pravili slovenske fonetične transkripcije in je nastal na podlagi fonološkega opisa govora Koprive.

Primer fonetičnega zapisa:

> 'Tədi 'jəst, 'vi̯ẹ̯š, 'kərko k'rət, kə bi <po...> 'taka t'ri ku'zice m'li̯ẹ̯ka, u̯'sək 'dan jə <pəršl...>, u̯'sək d'ruγi 'dan jə 'pəršla, kə 'təkrət so <nar> bren'γarli </nar>, so 'rekli.

V **poenostavljenem zapisu** moramo biti pozorni na stopnjo poenostavljanja, saj ne želimo, da se izgubijo določene fonetične značilnosti, reprezentativne za koprivski govor. Prepis pa moramo vseeno toliko poenostaviti, da bo primeren tudi za zainteresirano javnost. Odločili smo se, da ohranimo vse diftonge, ne upoštevamo pa kvalitete glasov. Nezveneča in zveneča velarna pripornika, značilna za kraško narečje, *x* in *γ* smo zamenjali s *h* in *g*. Ohranili smo vse polglasnike in vsa naglasna mesta (označena z nevtralnim ' nad naglašenim samoglasnikom), česar zaradi omejenosti s časom in materialnimi sredstvi niso upoštevali zapisovalci *GOS-a*. Seznam poenostavljenih glasov: u̯ → w; i̯ → j; γ → ü; ẹ̯ → e; γ → g; x → h.

Primer poenostavljenega zapisa:

> Tədi jəst, viẹš, kərko krət, kə bi <po...> təka tri kuzi̯ce mli̯ẹka, wsək dən jə <pəršl...>,wsək drùgi dən jə pə́ršla, kə təkrət so <nar> brengə́rli </nar>, so rẹ́kli.

Poknjižena varianta bo uporabnikom korpusa vidna le, če bodo to možnost izbrali, saj v njej ne bodo ohranjene značilnosti krajevnega govora na glasoslovni in oblikoslovni ravnini. Pomembna bo kot izhodišče za iskanje po korpusu (izraze iščemo prek njihove knjižne različice) in kot izhodišče za izdelavo diferencialnega slovarja.

Vse tri variante vsebujejo tudi posebne oznake besed, npr. *lai* = lastno ime, *nar* = narečno … (več v naslednjem članku). Takih besed računalnik ne bo avtomatizirano označeval, ker so posebnosti, ki imajo lahko v knjižnem jeziku npr. drugačen spol (prim. *lepa komunska šterna* = lep občinski vodnjak), zato se jih bo označevalo ročno. Tovrstne besede (sem spadajo tudi izposojenke) bodo grafično ločene od ostalih besed, saj predvidevamo, da ne bodo razumljive vsem uporabnikom korpusa, zato se bo s pomikom kurzorja nanje v oblačku izpisal njihov pomen.

Primer poknjižene variante:

> Tudi jaz, veš, koliko krat, ker bi <po...> take tri kozice mleka, vsak dan je <prišl...>, vsak drugi dan je prišla, ker takrat so <nar> prodajati </nar>, so rekli.

Korpus je, zaenkrat še v testni obliki, dostopen na spletnem naslovu *http://jt.upr.si/GOKO/*. Ker je bilo doslej vanj vnesenega še zelo malo gradiva, omogoča le iskanje po nekaj zapisanih besedah, npr. po vezniku *in*.

3 Načrt nadaljnjega dela

V članku je bil predstavljen le del metodoloških vprašanj za gradnjo korpusa GOKO, ki jih je bilo treba rešiti. Zavedamo se, da gre za pilotsko raziskavo in da je zato možnost napak velika, vendar upamo, da bomo na ta način spodbudili druge dialektologe, da začnejo razmišljati o večjem – referenčnem dialektološkem korpusu. Tak korpus ne bi bil pomemben zgolj za jezikoslovce in dialektologe, ampak tudi za širšo javnost in za izobraževalne namene. V poznavanju in uporabi novih jezikovnih tehnologij, kot so korpusi, raziskovalci vidimo priložnost, da uporabnikom predstavimo jezikovno raznolikost slovenskega jezika in razcepljenost na številna narečja in govore ter jih ozavestimo o pomembnosti ohranjanja narečij. S pomočjo korpusa bi lahko potencialni uporabniki na podlagi avtentičnih govorjenih besedil neposredno spoznavali slovenska narečja in tako razvijali svojo jezikovno kompetenco. Tak korpus pa lahko služi tudi kot osnova za nadaljnja raziskovanja govora, tako za študente kot za raziskovalce (prim. Zwitter Vitez in Krapež Vodopivec 2011).

4 Literatura

GORJANC, Vojko, 2005: *Uvod v korpusno jezikoslovje.* Domžale: Izolit.

LOGAR, Nataša, 2009: Korpusi v terminografiji: umik potrebe po introspektivni presoji. Nina Ledinek, Mojca Žagar Karer in Marjeta Humar (ur.): *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC SAZU. 319–328.

MIKOLIČ, Vesna in BEGUŠ, Ana, 2011: Meddisciplinarne pomenotvorne zmožnosti v procesih (de)terminologizacije turistične terminologije. Simona Kranjc (ur.): *Meddisciplinarnost v slovenistiki*. Ljubljana: Filozofska fakulteta (Obdobja, 30). 313–319.

STABEJ, Marko in VITEZ, Primož, 2000: KGB (korpus govorjenih besedil) v slovenščini. Cene Bavec idr. (ur.): *Informacijska družba IS'2000: zbornik 3. mednarodne multikonference*. 79–81.

VINTAR, Špela, 2008. *Terminologija: terminološka veda in računalniško podprta terminografija*. Ljubljana: Filozofska fakulteta.

ZEMLJARIČ MIKLAVČIČ, Jana, 2008: *Govorni korpusi*. Ljubljana: Filozofska fakulteta.

ZWITTER VITEZ, Ana in KRAPEŽ VODOPIVEC, Irena, 2011: Korpus govorjene slovenščine (GOS) za kakovostno in prijazno učno uro. Andreja Bačnik idr. (ur): *Mednarodna konferenca Splet izobraževanja in raziskovanja z IKT - SIRIKT.* 309–314. <http://prispevki.sirikt.si/datoteke/sirikt2011_zbornik.pdf>. Dostop 12. junija 2012.

Klara Šumenjak[1], *Jernej Vičič*[2]
[1] *Znanstveno-raziskovalno središče Univerze na Primorskem*
[2] *Inštitut Andrej Marušič Univerze na Primorskem*
klara.sumenjak@zrs.upr.si, jernej.vicic@upr.si

Izzivi pri izdelavi dialektološkega korpusa *GOKO*

V prispevku sta avtorja (jezikoslovka in računalniški strokovnjak) predstavila potrebe in izzive pri izdelavi dialektološkega korpusa GOKO. *Opisala sta, pri čem in na kakšen način sta morala sodelovati, predstavila nekatere probleme pri gradnji korpusa in podala rešitve, na koncu pa predstavila možnost nadgradnje korpusa* GOKO *v referenčni dialektološki koprus, ki bi bil pomemben doprinos za slovenske raziskovalce in širšo javnost.*

1 Uvod

Za nas je bil velik izziv povezati raziskovanje dialektologije in korpusnega jezikoslovja, saj tega (v Sloveniji) ni počel še nihče.[1] Kot izhodišče za gradnjo *Govornega korpusa Koprive na Krasu – GOKO*[2] smo uporabili delo Jane Zemljarič Miklavčič iz leta 2008 *Govorni korpusi* (predelana doktorska disertacija, v okviru katere je predstavila načela za gradnjo govornega korpusa slovenščine), saj sta si metodologiji gradnje govornega in dialektološkega korpusa v izhodišču zelo podobni. Za izoblikovanje dobre metodologije za gradnjo dialektološkega korpusa pa je potrebno natančno poznavanje kraja in značilnosti govora, ki ga bomo opisali, treba je poznati osnove korpusnega jezikoslovja in jezikovnih tehnologij, ki metodologijo prenašajo v prakso. V prispevku bosta avtorja natančneje predstavila, na katerih mestih in na kakšen način morata jezikoslovec in računalničar sodelovati, da se zasnova dialektološkega korpusa realizira, in kateri so morebitni problemi in rešitve pri realizaciji.

1 Carmen Kenda Jež je leta 2007 na konferenci *Slovenska narečja med sistemom in rabo* predstavila prispevek *Kako do slovenskega narečnega korpusa*, vendar je objavljen le povzetek.
2 Korpus *GOKO* je natančneje predstavljen v prejšnjem članku.

2 Potreba po sodelovanju

Izdelava dialektološkega korpusa je multidisciplinarno opravilo. Jezikoslovec osnuje metodologijo korpusa in poda svoje zahteve, predstavljene v sledečih podpoglavjih (na primer kakšno fonetično pisavo potrebuje, katere so posebne oznake, ki jih bo v korpusu potreboval, katere zvočne posnetke bi želel vključiti), uresniči pa jih računalničar.

2.1 ZRCola

Pri oblikovanju načel transkribiranja besedil je sicer treba upoštevati mednarodne standarde in že obstoječa priporočila – v mednarodnem jezikoslovju prevladuje tendenca, da se uporablja fonetična transkripcija s simboli IPA (*International Phonetic Alphabet*) (Zemljarič Miklavčič 2008: 106, IPA 1999) – vendar je za našo dialektološko raziskavo potrebna slovenska fonetična transkripcija, ki vsebuje veliko znakov, ki jih IPA nima. Ravno zaradi specifike slovenske fonetične transkripcije je dr. Peter Weiss razvil vnašalni sistem ZRCola (Weiss 2012), ki je bil na Znanstvenoraziskovalnem centru SAZU v Ljubljani razvit za jezikoslovne, predvsem dialektološke potrebe in deluje v Microsoftovem programu Word v operacijskem sistemu Windows. Temelji na standardu unikod (unicode) (The Unicode Consortium 2012).

Uporaba sistema je dvojna:

- sistem omogoča vnos posebnih znakov (prim. *č, ž, š, į, ų, ɣ*) in celo posebnih zgodovinskih znakov v dajnčici in metelčici;
- sistem omogoča prikaz posebnih znakov s priloženo pisavo ZRCola (Weiss 2012).

Sistem za vnos znakov je v rabi med dialektologi, pisavo pa je poleg urejevalnika besedil mogoče uporabiti tudi v ostalih programih, ki uporabljajo (oziroma vsaj podpirajo) Microsoftov format pisav (*TrueType*), torej tudi v vseh modernejših spletnih brskalnikih.

Pisavo smo uporabili pri oblikovanju spletnih strani, ki prikazujejo vsebino korpusa. Slika 1 kaže del kode CSS (*Cascading Style Sheets*), ki vključuje pisavo ZRCola. Poleg te kode potrebujemo še datoteko s pisavo, ki jo shranimo na strežnik. Datoteka je dostopna na strežniku SAZU (Weiss 2012).

```
@font-face {
 font-family:''00 ZRCola'';
 src: url('zrcola.ttf');
 src: local('@'),
     url('zrcola.ttf') format('opentype');
```

Slika 1: Del kode CSS, ki vključuje pisavo ZRCola

Brskalniki na sistemih, ki že imajo naloženo pisavo ZRCola, bi pravilno prikazali strani tudi brez teh oznak, ostali sistemi pa ob prvem ogledu naložijo samo pisavo in pravilno prikažejo posebne znake.

2.2 Transkripcija (fonetična, poenostavljena, poknjižena)[3]

Pomemben kriterij pri izdelavi načel transkribiranja govora je namen korpusa. Korpus *GOKO* bo namenjen jezikoslovcem (dialektologom) in tudi širši javnosti, zato smo se odločili, da bomo zbrano gradivo zapisali v treh različicah, ki bodo opremljene tudi z zvočnimi posnetki:

a) **fonetični zapis**: upoštevajoč vse glasoslovne variante govora Koprive na Krasu in zapis s slovensko fonetično transkripcijo – ta oblika bo namenjena naši analizi govora in drugim jezikoslovcem in dialektologom, za katere je relevanten natančen fonetični zapis narečja;

b) **poenostavljeni zapis**, kjer smo skušali obdržati temeljne glasoslovne značilnosti krajevnega govora; pomembno vprašanje, ki se pojavlja na tem mestu, je, do kolikšne mere poenostaviti besedilo; če bi popolnoma zanemarili zapis glasoslovnih značilnosti, namreč ne bi mogli več govoriti o koprivskem dialektološkem korpusu, ampak zgolj o korpusu, ki ima le nekaj zanimivih leksikalnih in skladenjskih narečnih potez, zato smo se odločili, da pri transkripciji ohranimo večino glasoslovnih posebnosti koprivskega govora, vendar jih zapišemo z znaki, ki so bližji večini laičnih uporabnikov;

c) **poknjižena varianta**, kjer bodo besede (ne pa tudi besedne zveze in stavki) zapisane v slovenskem standardnem jeziku – taka različica bo nujna za iskanje po korpusu in uporabljena kot izhodišče za diferencialni slovar koprivskega govora.

2.3 Označevanje

Referenčni govorni korpus slovenskega jezika *GOS* (Zwitter Vitez idr. 2009) je bil izdelan v okviru projekta *Sporazumevanje v slovenskem jeziku*. Korpus *GOS* je prosto dostopna zbirka posnetkov in transkripcij govorjene slovenščine.

V okviru projekta so nastale tudi smernice za zbiranje, transkripcijo in označevanje (Zwitter Vitez idr. 2009) ter opis gradiv (Verdonik in Erjavec 2010).

Pri snovanju korpusa *GOKO* smo upoštevali predlagane smernice (Verdonik in Erjavec: 2010), ki smo jim dodali lastne razširitve, opisane v nadaljevanju razdelka.

Osnovne oznake, ki so uporabljene v korpusu *GOS*, smo razširili z lastnimi oznakami, ki opisujejo posebnosti predstavljenega korpusa. Spodaj je predstavljen seznam oznak, ki smo jih potrebovali za gradnjo korpusa, in njihovih pomenov. Oznake so že tako prirejene, da jih lahko vključimo v standariziran zapis korpusa, kot je TEI-P5 (TEI 2008). Pri standardizaciji zapisa besedila se opisane oznake prevedejo v standardne TEI-oznake. Spletni vmesnik bo omogočal prikazovanje oziroma skrivanje teh oznak, saj lahko zaradi prevelikega števila podatkov pride do zastranitve besedila.

• **<nar>** začetek narečnega izraza ali besedne zveze, **</nar>** konec narečnega izraza ali besedne zveze. Kriterij za uvrščanje pojmov med narečno je bil, da se izraza ne najde med gesli v *Slovarju slovenskega knjižnega jezika*, če pa se tam nahaja, ne sme biti označen z nar., pog. ali imeti drugačen pomen, kot je v naših besedilih. Oznaka

3 Glej tudi prejšnji članek.

ne bo vidna uporabnikom korpusa – tekst, označen z *narečno*, bo vizualno ločen od ostalega besedila.

Primer: *[...] in ˈbum jə* **<*nar*>** *ˈratəu̯* **</*nar*>***, ˈveš, kəˈdai̯ [...]***

• **<lai>** začetek lastnega imena, **</lai>** konec lastnega imena. Pod lastna imena so uvrščena osebna, zemljepisna in stvarna lastna imena. Oznaka ne bo vidna uporabnikom korpusa – tekst, označen z *lastno ime*, bo vizualno ločen od ostalega besedila.

Primer: *[...] sˈmo ˈmi̯e̯li* **<*lai*>** *Šˈtanjəu̯* **</*lai*>** *təmˈle, bˈlizi in* **<*lai*>** *Iˈpavo* **</*lai*>** *[...]*

• **<neraz/>** nerazumljivo. Pri zapisovanju govora ni vedno mogoče ugotoviti, kaj je govorec povedal. Zavedamo se, da bi se ta del lahko enostavno izpustil, vendar smo želeli čim bolj natančno označiti korpus, zato je na mestih, kjer zapisovalka ni razumela, kaj je pripovedovalec povedal (v nekaterih primerih gre zgolj za eno besedo, pri nekaterih pa tudi za besedno zvezo ali stavek), zapisan **<?>**. Uporabnik bo imel možnost izbrati, ali naj se mu ta oznaka prikaže v korpusu ali ne.

Primer: *[...]raˈzuməš, in* **<?>** *ˈtir jə bu ˈnutər [...]*

• **<nedokon value='mam...'/>** začetek nedokončane besede. V korpusu nismo želeli enačiti pojava, ko je celotna beseda ali besedna zveza nerazumljena, in pojava, ko informant ni izrekel besede do konca. Tukaj gre največkrat za samopopravljanje govorca, ko želi najprej povedati določeno stvar, nato pa si premisli in pove drugo, ali pa ko za določeno informacijo želi dodati še eno, kar je izvidno iz sledečega primera primera: informant je želel takoj povedati, da gre za sedem metrov, vendar je dodal nedoločni člen »enih«, s katerim izraža približnost informacije. To smo v korpusu označili tako, da smo zapisali izrečene glasove, ki pa niso bili izgovorjeni kot beseda, prim. **<nedokon value=**'si̯e̯...'**/>**, ko je hotel informant izreči besedo sedem. Uporabnik bo imel možnost izbrati, ali naj se mu ta oznaka prikaže v korpusu ali ne.

Primer: *[...] ne ˈvi̯e̯m,* **<nedokon value=**'si̯e̯...'**/>** *ˈanix ˈsi̯e̯dəm ˈmetrou̯ [...]*

Primer, viden uporabniku: *[...] ne ˈvi̯e̯m,* <si̯e̯...> *ˈanix ˈsi̯e̯dəm ˈmetrou̯ [...]*

• **<polglas/>** obotavljanje, uporabnik korpusa bo to videl zapisano s polglasnikom in dolžino əː. Uporabnik bo imel možnost izbrati, ali naj se mu ta oznaka prikaže v korpusu ali ne.

Primer: *[...] in* **<*polglas*/>** *tu jə* **<*polglas*/>** *ˈčudno [...]*

Primer, viden uporabniku: *[...] in* əː *tu jə* əː *ˈčudno [...]*

• **<nav>** začetek navedka, **</nav>** konec navedka, uporabnik korpusa bo to videl zapisano z » in «. Na začetku gradnje korpusa nismo nameravali ločevati premega govora znotraj pripovedovanja, vendar smo ugotovili, da je informant, ko je navajal

koga drugega, v nekaterih primerih spremenil ton pripovedovanja, uporabil pa je tudi besede, netipične za koprivski govor, saj je najverjetneje želel posneti protagonista zgodbe. Zaradi ugotovljenega premi govor ne bo vključen v glasoslovno, oblikoslovno in skladenjsko analizo in prav zato je zapisan kot ločena skladenjska enota, ne pa v skladu s pravopisnimi pravili.

> Primer: *[...] u̯ˈrati pˈrɛ̨t γosˈtilno,* ***<nav>*** *Pˈridiˌpˈridiˌ* ***</nav>*** *ˈjəs ˈtečem če ˈγor,* ***<nav>*** *Oˌˈrauŋo pˈrau̯ˌ da si pˈrišla, se γˈremo iγˈrat u̯ saˈlon.* ***</nav>*** *in ˈjəs sə γˈrem [...]*

> Primer, viden uporabniku: *[...] u̯ˈrati pˈrɛ̨t γosˈtilno,* ***»****Pˈridiˌpˈridiˌ****«*** *ˈjəs ˈtečem če ˈγor,* ***»****Oˌˈrauŋoˌpˈrau̯ˌ da si pˈrišla, se γˈremo iγˈrat u̯ saˈlon.****«*** *in ˈjəs sə γˈrem [...]*

• [ ] v oglatem oklepaju so označene vse izjave in vsi neverbalni glasovi, ki jih izreče izpraševalka. Zaradi njene različne govorne pripadnosti te izjave niso vključene v korpusno analizo. Uporabnik bo imel možnost izbrati, ali naj se mu ta oznaka prikaže v korpusu ali ne.

> Primer: *[...] ˈmisləm, də ˈvi̯ɛ̨š, kəˈku so ˈtəkrət* ***[Aha]*** *ˈpisəli, ne,* ***[Aha]*** *šteˈvilkə in <əː> iˈmi̯ɛ̨* ***[Aha]****.*

• **<smeh/>** neverbalni glasovi so sicer lahko označeni opisno <smeh> ali pa kot del fonetičnega prepisa besedila <ha ha ha>. Zaradi poenostavljanja in poenotenja smo se odločili, da neverbalni glas označimo z oznako <smeh/>. Uporabnik bo imel možnost izbrati, ali naj se mu ta oznaka prikaže v korpusu ali ne.

> Primer: *In ˈtəkrət, al ˈpei̯ ˈtisti <nar> ˈbot </nar>* ***<smeh/>*** *<nar> zəsˈtopəš </nar> [...]*

• **<premor/>** odločili smo se, da s to oznako označimo premor, ki traja več kot 2 sekundi, nanaša pa se na tihi premor – brez kakršnega koli zvoka v ozadju.

> Primer: *[...] də jə bu tˈrau̯nik sˈpi̯ɛ̨t* ***<premor/>*** *in <əː> tu jə <əː>[...]*

• _ podčrtaj povezuje dve besedi. Besedilo v vseh treh oblikah mora biti poravnano, na tak način izenačujemo število besed.

> Primer fonetične transkripcije: *bəl_velik*; primer poknjižene variante: *večji*

2.4 Posnetki

Na terenu je bilo zbranih za približno 10 ur posnetkov, v korpus pa bo zajetih okrog 90 minut. Glede na to, da gre za pilotsko raziskavo, smo skušali v *GOKO* zajeti besedila, ki so čim bolj jasna in vsebujejo čim manj interferenc s strani izpraševalke. Zaradi enostavnejše manipulacije so posnetki razdeljeni na manjše dele, shranjeni so v formatu MP3.

2.5 Lokacija

Natančna določitev lokacije govorjenega besedila omogoča umestitev besedil na geografske informacijske sisteme, GIS (Chang 2007). Poleg osnovne uporabe korpusa je ključna tudi predstavitev v vizualni obliki – kartografiranje s sistemom GIS, ki ponuja odlično izhodišče za sistematične primerjalne analize med posameznimi frazeološkimi sistemi glede na funkcijsko in socialno zvrstnost tako znotraj slovenskega jezika, kot tudi primerjalne analize s frazeološkimi sistemi drugih jezikov (sosednjih in bolj oddaljenih – kontrastivna analiza); areali frazemov namreč lahko pokažejo na vplivanje med jezikovnimi sistemi.

Prvi del kartiranja korpusa narečnih izrazov obsega zasnovo, pripravo in izdelavo korpusa, ki mora biti združljiva s standardiziranimi zapisi orodij GIS. To pomeni, da je treba bazi pri vsakem vnosu podati geografsko referenco, ki jo orodje GIS prepozna in izraz ustrezno umesti na predhodno določen izsek površja in v končni fazi tudi na standardiziran kartni prostor.

```
<geoDecl datum=''WGS84''/>
```

Slika 2: Deklaracija standarda za zapis lokacije

Izbrali smo oznako lokacije po standardu WGS84 (*World Geodetic System*) (NIMA 1984), ki je shranjena v znački **<geo>**. Predstavljena je z dvema številoma, ki opisujeta zemljepisno širino in dolžino po opisanem standardu. Slika 2 kaže dodatno oznako, ki jo umestimo v glavo korpusa.

Slika 3 kaže primer natančnega opisa lokacije, kjer je bilo gradivo zbrano, dodana je tudi datumska oznaka. Za opisan korpus je tako natančno označevanje vprašljivo, saj so vsa gradiva zbrana v isti vasi, zanimiva je le temporalna oznaka, vendar tako standardizirano označevanje omogoča umestitev korpusa v druge projekte in umestitev v sisteme GIS.

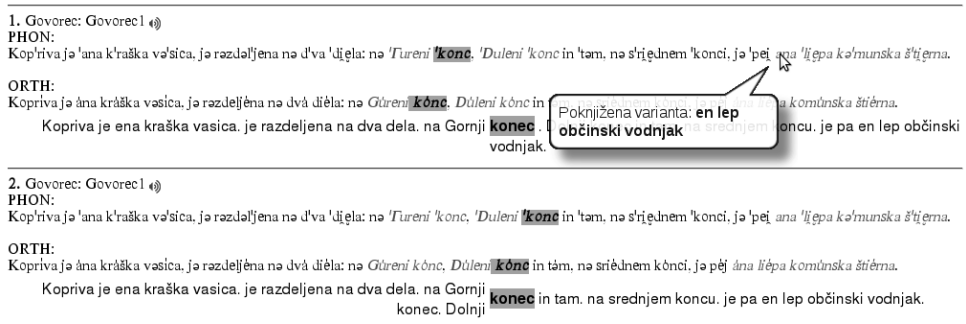
```
<settingDesc>
       <place>
        <placeName>
         <settlement>Kopriva</settlement>
         <region>Obalno-kraška</region>
         <country>Slovenija</country>
        </placeName>
        <location>
         <geo>
          45.781876,13.833693
         </geo>
        </location>
       </place>
       <setting>
        <date when=''2009''/>
       </setting>
      </settingDesc>
```

Slika 3: Opis natančne lokacije nastanka gradiva z datumsko označbo

2.6 Spletni vmesnik

V okviru predstavljenega projekta je bil izdelan spletni vmesnik. Iskanje je možno le po pokniževeni različici besedila, saj je ta edina dovolj standardizirana, da nam omogoča enolične rezultate iskanj. Rezultati iskanja so prikazani v standarni obliki kolokacij z možnostjo pregleda po vzporednih zapisih v ostalih dveh oblikah ter s povezavo na zvočni zapis, ki vsebuje najdeno vsebino z vnaprej določeno dolžino konteksta. Vse tri oblike besedila, fonetična, poenostavljena in pokniževena, so poravnane in najdenemu besedilu v pokniževeni obliki se dopišeta še vzporedni besedili v ostalih dveh oblikah.

Spletni vmesnik se pri iskanju in nadzoru podatkov (*data querying and management*) zanaša na IMS Open Corpus Workbench (*OpenCWB*) (Christ 1994). Tako lahko iščemo po podatkih z jezikom CQP (Christ in Shulze 1996) ali pa tudi s poenostavljenim vmesnikom. Korpus je označen v skladu s smernicami *JOS* (Erjavec idr. 2010), kar pomeni, da je tudi iskanje po tem korpusu podobno iskanju po korpusu *JOS* in derivatih, kot je *GOS*. Sistem je primeren tudi za veliko večje korpuse, kar omogoča širitev v bodoče.



Slika 4: Primer prikaza s spletnim vmesnikom, iskana beseda je »konec«

Slika 4 prikazuje primer izpisa rezultatov iskanja besede »konec«, glede na to, da gre za dialektološki korpus, je najprej izpisan fonetični prepis, nato poenostavljena varianta, na zadnje pa pokniževena varianta – uporabnik bo imel možnost izbrati, katero od različic želi videti (lahko samo eno, dve ali vse tri variante). Za pravilen prikaz fonetične in poenostavljene variante potrebujemo pisavo ZRCola, ki se samodejno namesti med ogledom strani (prvi ogled je počasnejši).

2.7 Možnosti napak

Največja možnost napak je pri načinu transkripcije in doslednosti zapisovanja. Glede na to, da je prepisovanje ročno, je treba biti zelo natančen, da se v katerem izmed treh zapisov ne pozabi katere besede ali ločila (kar bi lahko porušilo celoten korpus) saj, kot že povedano, gre za avtomatsko pretvarjanje.

3 Sklep

Namen projekta je bil dvojen: izdelava dialektoloških gradiv korpusa *GOKO* ter orodij in smernic za lažjo izdelavo novih gradiv. Korpus bo prosto dostopen prek spletnega vmesnika. Orodja in smernice bodo olajšali gradnjo novih sorodnih korpusov oz. omogočili nadgradnjo obstoječega, ki bi lahko sčasoma postal referenčni dialektološki korpus. To bi se doseglo tako, da bi se slovenski dialektologi med seboj povezali in svoje zbrano gradivo uredili skladno s priporočili za *GOKO*. V taki obliki bi ga lahko nato vključili v korpus, saj je *GOKO* zgrajen tako, da omogoča enostavno dodajanje novih podatkov. Tak korpus bi bil edinstvena priložnost, da bi imeli na enem mestu zbranih veliko število zapisov in posnetkov govorov, ki se zelo hitro spreminjajo ali celo izumirajo. Dodana vrednost takega korpusa je v tem, da bi širši javnosti, tudi učencem, dijakom in študentom omogočil neposredni stik s paleto različnih slovenskih govorov ter ponudil neštete možnosti njihovega nadaljnjega raziskovanja na vseh jezikovnih ravninah.

4 Literatura

CHANG, Kang-Tsung, 2007: *Introduction to Geographic Information System*. 4th edition. McGraw Hill.

CHRIST, Oliver, 1994: *A Modular and Flexible Architecture for an Integrated Corpus Query System*. Budimpešta: COMPLEX'94.

CHRIST, Oliver in SHULZE, Bruno Maximilian, 1996: *The Cqp User's Manual*. Stuttgart: Universität Stuttgart.

ERJAVEC, Tomaž, GORJANC, Vojko in STABEJ, Marko, 1998: Korpus FIDA. Tomaž Erjavec in Jerneja Gros (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan.

ERJAVEC, Tomaž idr., 2010: The JOS Linguistically Tagged Corpus of Slovene. *Proceedings*. Malta: International Conference on Language Resources and Evaluation (7).

IPA, 1999: *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

KENDA JEŽ, Carmen, 2007: Kako do slovenskega narečnega korpusa. Vera Smole (ur.): *Slovenska narečja med sistemom in rabo.* Ljubljana: Filozofska fakulteta (Obdobja, 26). 16–17.

NIMA, 1984: *Department of Defense World Geodetic System 1984, Its Definition and Relationships With Local Geodetic Systems*. Technical Report TR8350.2: National Geospatial-Intelligence Agency.

BURNARD, Lou in BAUMAN, Syd (ur.), 2008: *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford idr.: TEI Consortium.

The Unicode Consortium, 2012: *The Unicode Standard*. Unicode Consortium, Mountain View, CA: <http://www.unicode.org/versions/Unicode6.0.0/>. Dostop 8. junija 2012.

VERDONIK, Darinka in ERJAVEC, Tomaž, 2010: *Opis tipov podatkov v govornem korpusu GOS*. <http://nl.ijs.si/ssj/gos/A3-OpisPodatkov-GOS+TEI.pdf>. Dostop 8. junija 2012.

WEISS, Peter, 2012: *Vnašalni sistem ZRCola*. Znanstvenoraziskovalni center SAZU: <http://www.zrc-sazu.si>. Dostop 8. junija 2012.

ZEMLJARIČ MIKLAVČIČ, Jana, 2008: *Govorni korpusi*. Ljubljana: Filozofska fakulteta.

ZWITTER VITEZ, Ana idr., 2009: Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike*. Ljubljana: Filozofska fakulteta (Obdobja, 28). 437–442.