Konkordančnik za govorni korpus GOS

Darinka Verdonik,* Ana Zwitter Vitez,† Miro Romih‡, Simon Krek‡

* Fakulteta za elektrotehniko, računalništvo in informatiko, Univerza v Mariboru
Smetanova 17, 2000 Maribor
darinka.verdonik@uni-mb.si
† Trojina, zavod za uporabno slovenistiko
Partizanska cesta 5, 4220 Škofja Loka
ana.zwitter@guest.arnes.si
‡ Amebis, d. o. o., Kamnik
Bakovnik 3, 1241 Kamnik
miro.romih@amebis.si, simon.krek@guest.arnes.si

Povzetek

Prispevek predstavlja funkcije konkordančnika GOS za iskanje po referenčnem govornem korpusu slovenščine. Konkordančnik bo od oktobra 2010 na voljo na spletnem naslovu www.korpus-gos.net zainteresiranim uporabnikom. Bistvena prednost konkordančnika pred drugimi sorodnimi jezikovnimi orodji je, da ponuja povezavo med zvokom in zapisom: vsako konkordanco lahko tudi poslušamo v kontekstu. Konkordančnik izkorišča podatke, vsebovane v korpusnem gradivu, tudi za zelo poglobljeno, specifično iskanje zahtevnih uporabnikov, hkrati pa se skuša približati manj zahtevnim potencialnim uporabnikom z enostavnim in hitro usvojljivim načinom uporabe.

Concordancer for the speech corpus GOS

The paper presents the functions of the GOS concordancer designed for browsing the Slovenian reference speech corpus. The concordancer will be available on the web address www.korpus-gos.net from October 2010. The main advantage of the concordancer compared to other similar language tools is its link between the sound and the transcription: every concordance can be listened to in its context. The concordancer uses all the metadata from the corpus to enable complex search function for highly-demanding users, but at the same time tries to attract less demanding potential users with a simple, user friendly interface.

1. Uvod

Potreba po izdelavi referenčnega govornega korpusa slovenščine je bila v zadnjem desetletju v slovenskem jezikoslovnem prostoru večkrat eksplicitno izražena. Prvi je na to opozoril Stabej, ko je pri predstavitvi besedilnovrstne sestave korpusa FIDA poudaril, da bi bil "seveda v slovenskem prostoru še bolj dragocen korpus, ki bi vseboval tudi govorjena besedila" (Stabej, 1998: 100); ideja je bila podrobneje predstavljena l. 2000 (Stabej, Vitez, 2000: 79). Tudi Weiss (2001: 422) je poudaril nujnost vključevanja govorjenih besedil v elektronsko zbirko. Gorjanc je pozival k začetku gradnje: "Čim prej bi bilo treba oblikovati skupino, ki bi začela s pripravami govornega dela korpusa." (Gorjanc, 2005: 53) Tudi v okviru dialektoloških študij je novo tisočletje vzbudilo pričakovanja po govornem korpusu: "/…/ širitev korpusa na spontani nejavni govor vzbuja upanje na drugačne čase" (Kenda Jež, 2004: 271), v okviru govornih tehnologij pa je bila potreba po vključitvi spontanega govora v raziskave izražena med drugim v Verdonik (2006: 40).

Teoretična izhodišča gradnje govornega korpusa, preizkušena na manjšem učnem govornem korpusu, so bila izdelana l. 2007 (Zemljarič Miklavčič, 2007; 2008). Leta 2008 so bila v okviru projekta *Sporazumevanje v slovenskem jeziku (SSJ, 2008−2013)*[1] zagotovljena sredstva za izgradnjo govornega korpusa slovenščine v obsegu 1 milijona besed ali 110 ur govora[2], s čimer so bili natanko desetletje po prvem pozivu h gradnji izpolnjeni vsi pogoji za začetek. Tekom dela je korpus dobil ime GOS (korpus GOvorjene Slovenščine) in bo konec leta 2010 v celoti dokončan in na voljo uporabnikom. Prve predstavitve njegove zasnove najdemo v Zemljarič et al. (2009) in Zwitter Vitez et al. (2009).

Enako kot sam korpus je za njegovo širšo dostopnost in uporabo pomemben tudi spletni vmesnik – konkordančnik – za iskanje po korpusu. Zapise iz govornega korpusa bi seveda lahko vključili tudi v konkordančnik za pisni korpus (za slovenščino npr. konkordančniki korpusov Nova beseda ali FidaPLUS), vendar bi s tem izgubili mnoge posebnosti govornega materiala, zlasti povezavo med zvokom in zapisom, govorni korpus GOS pa ima še eno pomembno posebnost, ki bi se v konkordačniku pisnega korpusa prav tako izgubila: dvojni zapis govora, v pogovorni in standardizirani različici.

Sofinanciranje projekta izdelave konkordančnika, prirejenega posebej za govorni korpus GOS, je jeseni 2009 odobrilo Ministrstvo za visoko šolstvo, znanost in tehnologijo ob sofinanciranju iz Evropskega regionalnega sklada, in tako bo od jeseni 2010 na spletnem naslovu www.korpus-gos.net na voljo konkordančnik GOS, s katerim bo mogoče spočetka iskati po delu korpusnega gradiva, po celotnem gradivu zaključenega korpusa GOS pa od januarja 2011.

V tem prispevku se bomo po uvodnem kratkem pregledu nekaterih sorodnih tujih konkordančnikov in

[1] http://www.slovenscina.eu/

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

skupin potencialnih uporabnikov osredotočili predvsem na predstavitev iskalnih funkcij in informacij o gradivu, ki jih bo ponujal konkordančnik GOS.

2. Pregled nekaterih tujih sorodnih konkordančnikov

Britanski nacionalni korpus (http://www.natcorp.ox.ac.uk/) je prosto dostopen na spletu, vendar govorna podsekcija zajema samo transkripcije brez izvornih zvočnih posnetkov. Obsega 10 milijonov besed. Iskanje po korpusu je med drugim omogočeno z vmesnikom XAIRA, s pomočjo katerega dobimo informacije o frekvencah besed, slovničnih struktur in kolokatorjev ter primerih konkordanc. Iskanje omogoča po celotnem korpusu ali omejenem naboru gradiva.

Tudi pri ruskem nacionalnem korpusu (http://www.ruscorpora.ru/en/index.html) je govorna podsekcija korpusa dostopna samo z zapisom, brez zvoka. Iskanje je mogoče po obliki, kanalu, slovničnih ali semantičnih značilnostih oz. po metapodatkih o govorcih in diskurzu.

Spletni konkordančnik omogoča tudi brskanje po francoskem Corpus de la parole (http://www.corpusdelaparole.culture.fr/). Iščemo lahko po jeziku (zastopani so namreč vsi govorjeni jeziki v Franciji), po besedah ali frazah ali po metapodatkih o diskurzu in govorcih.

Pregled še ostalih tujih korpusov kaže, da tako kot navedeni večinoma[3] ne omogočajo dostopa do zvočnih posnetkov, pri iskanju pa izkoriščajo možnosti, ki jih ponuja korpusno gradivo (poleg osnovnega iskanja po besedah/frazah še iskanje po jezikovnih oznakah na različnih ravneh ter metapodatkih o govorcih in diskurzih).

3. Uporabniki konkordančnika GOS

Pri zasnovi slovenskega spletnega konkordančnika za iskanje po govornem korpusu GOS smo upoštevali predvidene potrebe naslednjih potencialnih skupin uporabnikov konkordančnika:

- raziskovalci – zahtevni uporabniki, ki jim je treba omogočiti čim več iskalnih funkcij in podatkov o gradivu
- izobraževanje – v izobraževanju (pri učenju slovenskega jezika) so pomembni predvsem enostavnost iskanja, kratka, jasna in priročna navodila ter grafična privlačnost konkordančnika
- poklici v stiku z govorom (razni pisci, tolmači in prevajalci, poklicni govorci...) – po eni strani enostavnost in privlačnost uporabe, po drugi strani pa vseeno ustrezno pester nabor iskalnih funkcij, zlasti glede na metapodatke o govorcih in diskurzih

Glede na ta predvidevanja ter glede na podatke, ki jih vključuje korpusno gradivo, in z navezavo na predvideni prenovljeni konkordančnik za pisni korpus slovenskega jezika, ki nastaja prav tako v okviru projekta Sporazumevanje v slovenskem jeziku, bo imel konkordančnik GOS iskalne funkcije, kot so opisane v sekciji 4.

4. Iskalne funkcije konkordančnika GOS

Konkordančnik podpira dva osnovna tipa prikaza rezultatov: v obliki konkordančnega niza ali v obliki seznama. Več o prikazu rezultatov je v naslednjem poglavju, v nadaljevanju pa so opisane iskalne funkcije konkordančnika, ki jih delimo na:

- iskanje po konkordančnem nizu (enostavno iskanje, napredno iskanje)
- iskanje po seznamu

Enostavno iskanje omogoča naslednje načine iskanja:

Enostavno iskanje	
po pogovornem zapisu	***po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"
filter po govorcih in diskurzih	

Tabela 1: Enostavno iskanje.

Z enostavnim iskanjem lahko iščemo poljubno po pogovornem ali po standardiziranem zapisu govora v korpusu. Razlika med enim in drugim zapisom je naslednja:

- **Pogovorni zapis:** Govor je zapisan v veljavnem slovenskem črkopisu (ni fonetični zapis) in upoštevane so veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pa je pri tem kolikor mogoče verno predstavljena glasovna podoba govora. Nekaj primerov: *tud, neki, mislm, prov, navm, najraj b vidu...*
- **Standardizirani zapis:** Pri pretvorbi v standardizirani zapis so odpravljene glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Ciljna oblika je knjižna različica istega leksema. Na drugih jezikovnih ravneh besede niso spremenjene. Če določenega leksema ni v knjižni normi, je ohranjen v obliki, ki se pojavlja v govoru. Nekaj primerov: *tudi, nekaj, mislim, prav, ne bom, najraje bi videl, štengica, sva šle...*

Primer iste izjave v pogovornem in standardiziranem zapisu:

P: a	*veš*	*boš*	*mov*	*veš*	*kok*	*boš*
S: a	*veš*	*boš*	*imel*	*veš*	*koliko*	*boš*

P: traku	*zgubu*	*al*	*kva*	*maš*	*not*
S: traku	*izgubil*	*ali*	*kaj*	*imaš*	*notri*

Pri pogovornem zapisu je mogoče samo iskanje po kanalu besede (iščemo samo in točno tisto obliko, ki smo ji vpisali), medtem ko je pri iskanju po standardiziranem zapisu privzeto iskanje po kanalu leme, kar pomeni, da se iskana beseda avtomatsko pretvori v osnovno obliko in se prikažejo rezultati za vse oblike te besede. Šele če vpišemo iskani niz med narekovaji, iščemo po kanalu besede, torej samo in točno tisto obliko, ki smo ji vpisali.

[3] Poslušanje zadetkov omogoča npr. nemški vmesnik (http://dsav-oeff.ids-mannheim.de/DSAv/SUCHMASK.H_TM), mnogi drugi pa ne, npr. češki - http://ucnk.ff.cuni. cz/english/hledat_v_cnk.php, italijanski - http://badip.uni-graz.at/, poljski – http://korpus.ia.uni.lodz.pl/conversational/...

Potem ko se prikažejo rezultati iskanja, lahko le-te **filtriramo po tipih diskurzov/govorcev:** ta funkcija omogoča iskanje znotraj vključenih metapodatkov o diskurzih in govorcih.

Metapodatki o govorcih vključujejo naslednje izbire:

- spol
- starost
- izobrazba
- regionalna pripadnost:
    - enotna (posameznik je vse življenje preživel v eni regiji)
    - razpršena (posameznik je del življenja [vsaj eno leto: šolanje, služba, selitev...] preživel v kateri drugi regiji/-ah)
- prvi jezik (slovenski ali tuji)

Regionalna pripadnost je v korpusu GOS označena glede na večja regionalna mestna središča, h katerim gravitira posamezno področje in ki sovpadajo z registrskimi območji v Sloveniji (MB, MS, SG, CE, KK, NM, LJ, KR, PO, GO, KP), oz. glede na pripadnost regijam zunaj Slovenije (zamejski Slovenci: Italija, Avstrija, Madžarska, življenje v tujini). V skladu s temi označbami so tudi iskalne možnosti v konkordančniku.

Metapodatki o diskurzih vključujejo izbire po:

- regiji snemanja
- letu snemanja
- klasifikaciji diskurza, kot prikazuje tabela 3

Tip diskurza	Kanal	Govorni dogodek
javni informativno-izobraževalni	televizija	novinarski prispevek
		moderirani pogovor
	radio	moderirani program
		moderirani pogovor
	osebni stik	osnovnošolska učna ura
		srednješolska učna ura
		tečaj
		javno predavanje
		fakultetno predavanje
javni razvedrilni	televizija	moderirani pogovor
		moderirana oddaja
		resničnostni šov
		športni prenos
	radio	moderirani program
		moderirani pogovor
nejavni nezasebni	osebni stik	formalni delovni sestanek
		neformalni delovni sestanek
		konzultacija na fakulteti
		storitev
		razgovor
	telefon	storitev
nejavni zasebni	osebni stik	pogovor v družini
		pogovor med prijatelji/znanci
	telefon	pogovor v družini
		pogovor med prijatelji/znanci

Tabela 3: Pregled iskanj glede na klasifikacijo diskurzov.

Za zahtevnejše uporabnike je na voljo napredno iskanje:

Napredno iskanje	
iskanje po bližini (1 in 2)	
iskanje po oblikoslovnih oznakah (samo 2)	
1 po pogovornem zapisu	***2 po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"
filter po govorcih in diskurzih	

Tabela 2: Napredno iskanje.

Pri naprednem iskanju lahko poleg iskalnih funkcij, ki jih omogoča že enostavno iskanje (torej iskanje po pogovornem ali po standardiziranem zapisu in filtriranje rezultatov glede na podatke o govorcih in/ali podatke o diskurzih), uporabljamo še dve funkciji:

- **Iskanje po bližini:** Omogoča iskanje glede na bližino/nebližino drugih besed. Iščemo lahko v okolici do 5 besed.
- **Iskanje po oblikoslovnih oznakah:** Samo pri iskanju po standardiziranem zapisu lahko iskano besedo tudi dodatno omejimo z besedno vrsto in ostalimi oblikoslovnimi oznakami (iščemo po kanalu oblikoslovnih oznak). Ker bo korpus GOS v prvi fazi samo avtomatsko označen z označevalnikom, naučenim na pisnih besedilih, še ni znano, kolikšna bo natančnost kanala z oblikoslovnimi oznakami.

Drugi način iskanja po korpusu GOS je iskalna funkcija seznam:

Seznam	
iskanje z nadomestnimi znaki (1 in 2)	
iskanje po oblikoslovnih oznakah (samo 2)	
1 po pogovornem zapisu	***2 po standardiziranem zapisu***
samo po besedah	privzeto po lemah
	med narekovaji "po besedah"

Tabela 4: Pregled iskalne funkcije seznam.

Funkcija seznam omogoča iskanje z nadomestnimi znaki, in sicer naslednjimi:

- znak * nadomešča poljubno število znakov

- znak ? nadomešča en znak

Enako kot pri iskanju po konkordančnem nizu je mogoče izbirati med iskanjem po pogovornem in po standardiziranem zapisu, pri slednjem je na voljo tudi filtriranje po oblikoslovnih oznakah. Filtriranje po tipih diskurzov in govorcev v prvi fazi ne bo omogočeno, je pa predvidena tovrstna nadgradnja naknadno.

Rezultati iskanja pri seznamu niso prikazani v obliki konkordanc, tako kot pri iskanju po konkordančnem nizu, ampak v obliki seznama besed s podatki o frekventnosti posamezne besede in njeni standardni obliki.

Iskalna načina seznam in konkordančni niz sta med seboj povezana: s klikom na besedo v seznamu se v zavihku konkordančni niz izdela ustrezen konkordančni niz, ki vsebuje primere rabe v kontekstu za besedo s seznama.

5. Prikaz zadetkov in podatki o njih

Rezultati iskanja se prikažejo:

- v obliki seznama konkordanc pri iskanju po konkordančnem nizu
- v obliki seznama besed s podatki o frekvenci in standardizirani obliki pri iskanju po seznamu

Rezultati iskanja pri konkordančnem nizu so:

- prikazani v pogovornem zapisu
- označeni glede na tip diskurza (JI – javni informativno-izobraževalni, JR – javni razvedrilni, NN – nejavni nezasebni, NZ – nejavni zasebni)

Desno od vsake konkordance je zvočnik. S klikom nanj lahko poslušamo eno ali več izjav, v kateri(h) je bil izgovorjen iskani niz.

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

- razširjeni kontekst v pogovornem zapisu (tj. izjava ali izjave, v kateri(h) je bil izgovorjen iskani niz + 1 predhodna + 1 naslednja izjava)
- podatki o diskurzu, iz katerega je konkordanca (tip diskurza, kanal, opis govornega dogodka, regija, kjer je potekal diskurz, datum in čas poteka diskurza, vir posnetka, opis diskurza)
- podatki o govorcu (spol, starost, izobrazba, regionalna pripadnost, prvi jezik)

Če želimo še več informacij, imamo na voljo:

- da poslušamo razširjeni kontekst
- da pogledamo standardizirani zapis razširjenega konteksta
- da pogledamo korpusne oznake (lema, oblikoslovne oznake)
- da shranimo podatke v odložišče

Rezultate lahko tudi:

- sortiramo
- izvozimo

Pri urejanju rezultatov iskanja lahko le-te razvrstimo:

- glede na jedrno besedo/besede
- glede na levo sobesedilo
- glede na desno sobesedilo

6. Zaključek

Prispevek predstavlja funkcije konkordančnika GOS za iskanje po referenčnem govornem korpusu slovenščine. Konkordančnik bo od oktobra 2010 na voljo na spletnem naslovu www.korpus-gos.net zainteresiranim uporabnikom. Bistvena prednost konkordančnika pred drugimi sorodnimi jezikovnimi orodji je, da ponuja povezavo med zvokom in zapisom: vsako konkordanco lahko tudi poslušamo v kontekstu. Konkordančnik izkorišča podatke, vsebovane v korpusnem gradivu, tudi za zelo poglobljeno, specifično iskanje zahtevnih uporabnikov, hkrati pa se skuša približati manj zahtevnim potencialnim uporabnikom z enostavnim in hitro usvojljivim načinom uporabe. Ob tem bo seveda opremljen s priročnimi navodili, priročniki, videom ... o uporabi in bo na voljo tudi v angleški različici.

V prihodnosti bo poleg funkcionalnosti konkordančnika za uporabo korpusa GOS ključna seveda tudi sama obsežnost in opremljenost korpusnega gradiva. Korpus bo ob zaključku v okviru sedanjega financiranja obsegal 1 mio. besed. V tujini referenčni govorni korpusi že rastejo na 10 mio. besed ali več (angleški, poljski, nizozemski, portugalski...), in tudi za slovenščino je ob sedanjem obsegu gradiva izraz »referenčni« le pogojno upravičen, saj je v 1 mio. besed avtentičnih pogovorov nemogoče zajeti vso pestrost govorjene slovenščine. Za rabo v jezikoslovju in jezikovnih tehnologijah pa bi si poleg večjega obsega želeli tudi dodano opremljenost gradiva, npr. skladenjsko označevanje, fonetični zapis ipd.

7. Literatura

Gorjanc, V., 2005. *Uvod v korpusno jezikoslovje*. Domžale, Izolit.

Kenda Jež, K., 2004. Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju. Kržišnik, E. (ur.), *Obdobja 22*. Ljubljana, Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko: 263−276.

Stabej, M., Vitez, P., 2000. KGB (korpus govorjenih besedil) v slovenščini. *Informacijska družba IS'2000: Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan.

Stabej, M., 1998. Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6.

Verdonik, D., 2006. *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora*. Doktorska disertacija, Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Weiss, P., 2001. Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU: utemeljitev. *Jezikoslovni zapiski* 7, 1−2. Ljubljana, Založba ZRC: 419−428.

Zemljarič Miklavčič, J., 2007. *Načela oblikovanja govornega korpusa za slovenščino*. Doktorska disertacija, Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Zemljarič Miklavčič, J., 2008. *Govorni korpusi*. Univerza v Ljubljani, Filozofska fakulteta.

Zemljarič Miklavčič, J., Stabej, M., Krek, S., Zwitter Vitez, A., 2009. Kaj in zakaj v referenčni govorni korpus slovenščine. Stabej, M. (ur.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani: 437–442

Zwitter Vitez, A., Zemljarič Miklavčič, J., Stabej, M., Krek, S., 2009. Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Stabej, M. (ur.), *Obdobja 28: Infrastruktura slovenščine in slovenistike*. Ljubljana, Znanstvena založba Filozofske fakultete Univerze v Ljubljani: 437–442