KORPUS GOS IN NJEGOVA UPORABA V RAZISKOVALNE, DIDAKTIČNE IN LJUBITELJSKE NAMENE

Ana Zwitter Vitez
Trojina, Zavod za uporabno slovenistiko, Ljubljana

UDK 808=163.6:81'42:004.738.52

Korpus GOS je zbirka avtentičnih posnetkov govorjene slovenščine, prosto dostopna na spletnem naslovu www.korpus-gos.si. Gre za interdisciplinarni rezultat projekta Sporazumevanje v slovenskem jeziku, ki sloni na analizi diskurza, sociolingvističnih in medijskih raziskavah, programiranju in procesiranju govornih signalov. Namenjen je poučevanju slovenščine, jezikoslovnim in jezikovnotehnološkim raziskavam, poklicnim govorcem in piscem besedil, pa tudi, nenazadnje, uporabnikom, ki doslej o jeziku sploh niso razmišljali.

govorni korpus, uravnoteženost korpusov, transkribiranje posnetkov, uporaba korpusov

The Spoken Corpus of Slovene GOS is a collection of authentic recordings including speakers from all over Slovenia and available online on the website www.korpus-gos.net. This achievement represents one of the results of the project Communication in Slovene and is based on the interdisciplinary research (sociolinguistic, media and discourse analysis, computer processing and sound engineering). The corpus GOS can be used for language teaching, linguistic analysis, language technologies, professional speakers and, nevertheless, for the wider audience.

spoken corpus, balanced corpus, speech transcription, use of language corpora

1 Uvod

Od oktobra 2010 je na naslovu www.korpus-gos.net mogoče najti korpus govorjene slovenščine z imenom korpus GOS. Elektronska zbirka transkribiranih posnetkov, zgrajena v okviru projekta Sporazumevanje v slovenskem jeziku, zajema milijon besed oziroma 120 ur posnetkov. Pri tem smo želeli zajeti:

– demografsko reprezentativen vzorec govorcev,
– najpogostejše govorne situacije.

Korpus GOS je plod združenih moči sociolingvističnih in medijskih raziskav, analize diskurza, računalniškega programiranja in procesiranja govornih signalov. Kot edinstven vir avtentičnih primerov govorjenih besedil iz najrazličnejših komunikacijskih situacij je korpus namenjen spoznavanju realne spontane govorne komunikacije v slovenščini, kar je pomembno za poučevanje slovenščine, poklicne govorce in snovalce besedil, jezikoslovne raziskave na področju sociolingvistike, pragmatike in analize diskurza ter jezikovnotehnološke raziskave za potrebe sinteze in avtomatskega razpoznavanja govora. Poleg omenjenih namenov uporabe ima korpus GOS še en nadvse pomemben cilj: zbuditi zanimanje uporabnika, ki sicer o jeziku sploh ne razmišlja, ter prispevati k njegovemu veselju in ustvarjalnosti pri vsakodnevnem govornem sporazumevanju.

V nadaljevanju prispevka bomo predstavili, kako so različna področja sodelovala pri gradnji korpusa GOS, in razmislili o tem, kako bodo cilji korpusa govorjene slovenščine dejansko uresničeni.

2 Zajem gradiva

Osnovni cilj zajema govorjenih besedil sta bili čim boljša uravnoteženost in reprezentativnost končne elektronske zbirke. Zato smo kombinirali demografske in besedilnovrstne kriterije.

Demografske kriterije predstavljajo dejavniki, ki najverjetneje (vendar še ne dokazano) najbolj vplivajo na razlike v govoru: spol, starost, izobrazba in regionalna pripadnost govorcev (ne pa tudi socialni status, saj je za slovenske razmere težko določljiv).[1] Podatke za določitev demografskih kriterijev smo pridobili pri Statističnem uradu Republike Slovenije.[2]

Pri besedilnovrstnih kriterijih smo upoštevali javnost diskurza in prenosnik. Javni diskurz je odprt za širšo javnost in naslavlja veliko skupino ljudi. Nejavni diskurz predstavljajo zasebni pogovori (v okviru družine, prijateljev) in nezasebni pogovori (v uradih, trgovinah in profesionalnem življenju). V javni diskurz smo vključili razvedrilne in informativno-izobraževalne medijske vsebine, pa tudi pedagoški diskurz v različnih izobraževalnih ustanovah.

Zajeli smo štiri različne prenosnike govorjenih besedil: radio in televizija (javni diskurz) ter osebni stik in telefon (pedagoški in nejavni diskurz).

3 Označevanje in transkribiranje posnetkov

Vsak posnetek je opremljen z naslednjimi podatki:

1. **podatki o govorcih** (spol, starost, izobrazba, regijska pripadnost, prvi jezik),
2. **podatki o posnetku** (vrsta diskurza, prenosnik, regija, leto snemanja),
3. **zapis govora** (transkripcija).

3.1 Zapis govora

Pri pregledu obstoječih praks transkribiranja smo ugotovili, da pri nas ni poenotenega načina zapisovanja, vendar obstajajo tri ključne tendence:

- dialektologija večinoma sledi fonetični transkripciji z diakritičnimi znamenji,[3]
- pragmatične raziskave (Smolej 2006) sicer uporabljajo slovenski črkopis, vendar zapisujejo pojave moderne vokalne redukcije ter druge pogovorne in narečne prvine govorjene slovenščine (npr. *maš čevle*),
- v jezikovnotehnološki praksi (Zemljarič Miklavčič 2007) se uporablja poknjiženi zapis govorjenega jezika, iz katerega pogovorne prvine niso več razvidne.

Pri transkribiranju posnetkov korpusa GOS smo želeli najboljše razmerje med hitrostjo postopka zapisovanja, berljivostjo končnih transkripcij in učinkovitostjo iskanja želene besedne oblike. Pri tem smo upoštevali, da je slovenščina močno dialektalno in socialnozvrstno razčlenjen jezik, ter želeli ohraniti čim več različnih oblik določene besede. Zato smo govor zapisali na dva različna načina: pri »pogovornem zapisu« skušamo kar se da zvesto ujeti dejansko akustično podobo izgovorjenih besed; pri »standardiziranem zapisu« dobi več različic neke besedne oblike (npr. *imam*, *mam*, *jemam*) krovno standardizirano obliko (npr. *imam*).

S pogovornim zapisom tako omogočimo dober vpogled v besedje in oblike govorjenega jezika, s standardiziranim zapisom omogočimo boljše iskalne možnosti, za povsem natančno zvočno podobo govora pa je vedno na razpolago tudi posnetek izjave.

Primer 1: Besedilo, zapisano s pogovornim (Pog.) in standardiziranim zapisom (Stan.)

Pog.	*a*	*lohk*	*mal*	*pošpegamo*
Stan.	*a*	*lahko*	*malo*	*pošpegamo*

1 Več o kriterijih zajema besedil je v Zemljarič Miklavčič idr. (2009).
2 www.stat.si
3 http://ola.zrc-sazu.si

4 Elektronska baza

V zadnji fazi gradnje korpusa GOS je velik del odgovornosti slonel na programiranju. Tekstovni del korpusa je tako zapisan v računalniškem standardu XML in v skladu s priporočili TEI (Text Encoding Iniciative). Zahtevnejši uporabniki si ga lahko v tej obliki snamejo s spletne strani www.korpus-gos.net. Opis sheme XML je vključen v paket za prenos korpusa oziroma dostopen na strani http://nl.ijs.si/ssj/gos/schema/tei_gos_doc.pdf.

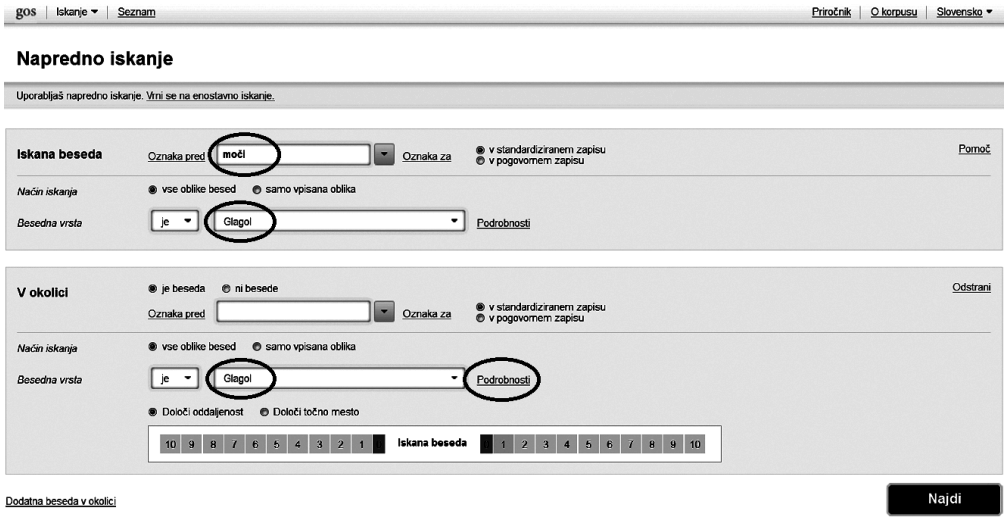
5 Primeri rabe korpusa GOS

Primeri rabe korpusa GOS bodo osredotočeni na tiste situacije, v katerih se govorni korpusi še ne uporabljajo: pouk slovenščine in študij govorjenih žanrov za poklicne govorce. Uporabe korpusa za potrebe jezikoslovnih raziskav in jezikovnih tehnologij ne bomo posebej predstavljali. Predpostavljamo namreč, da jezikoslovci, programerji in elektroinženirji dobro vedo, kako uporabljati govorne korpuse za potrebe sociolingvističnih, pragmatičnih, dialektoloških in prevodoslovnih analiz ali pri sintezi in avtomatskem razpoznavanju govora.[4]

5.1 Učenje jezikov

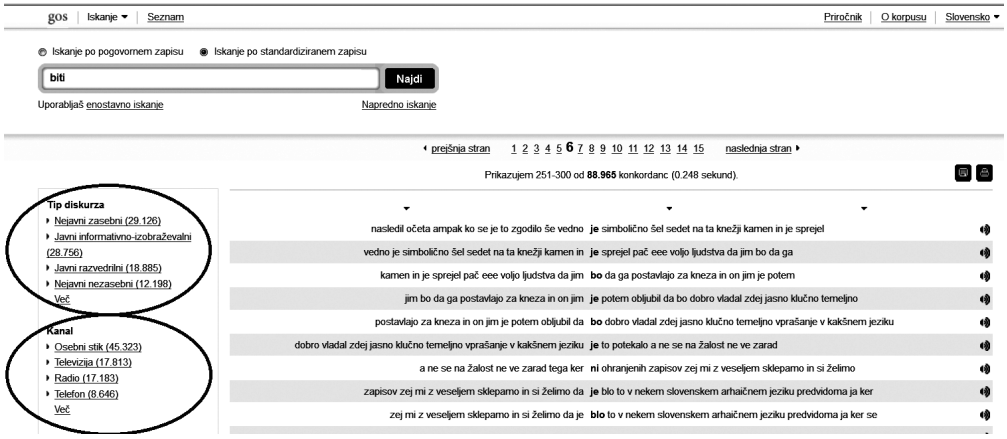
Eden izmed pomembnejših ciljev je uporaba korpusa GOS pri pouku slovenščine kot neprvega jezika, saj »vnaprej napisani dialogi redko odsevajo nepredvidljivost in dinamiko konverzacije, zato imajo učenci manj možnosti, da bi usvojili ustrezni nabor jezikovnih sredstev, ki jih bodo potrebovali v neznanih situacijah izven učilnice« (Burns 2001: 21). Tudi I. Ferbežar (2003: 35–36) ugotavlja, da je jezikovna raba »odvisna od različnih situacijskih dejavnikov: prostora, časa, vloge (odnosa med udeležencema pogovora, njune starosti, statusa ipd.), teme, sporazumevalnega namena in seveda sobesedila«, vendar »obstoječi učbeniki za slovenščino kot neprvi jezik vse te dejavnike premalo upoštevajo«.

Pri pouku slovenščine za materne govorce je lahko korpus GOS koristen vir avtentičnih pogovorov pri usvajanju razlik med formalnimi in neformalnimi situacijami. Predstavljajmo si, da petnajst minut po začetku ure v razred stopi učenec, ki svojo zamudo pojasni



Slika 1

4 Sinteza in avtomatsko razpoznavanje govora je do zdaj večinoma slonelo na idealiziranem gradivu dekontekstualiziranih izjav, posnetih v idealnem studijskem okolju in z (domnevno) idealnimi govorci.



Slika 2

z izjavo *Oprostite, sem mogu k zobarju*. Učencu (in razredu) želimo pojasniti neprimernost uporabljene strukture. Uporabimo funkcijo napredno iskanje, ki omogoča iskanje po posameznih oblikah sestavljene strukture. V naprednem iskanju pod prvo iskano besedo vpišemo *moči* in pri oznaki besedna vrsta označimo, da gre za glagol. V podrobnostih označimo, naj bo glagol nezanikan (ker je problematična samo struktura v trdilni obliki). Nato kliknemo na možnost beseda v okolici in označimo, naj prvi besedi sledi glagol v nedoločniku (Slika 1).

Tako iskanje bo vrnilo rezultate tipa *kolko dražje bi mogle bit*, *združenje bi moglo rečt mi ne bomo nič delali*, *ga je po mojem moglo ful ful bolet*. Hiter pregled situacij, v katerih se ta struktura pojavlja najpogosteje, nam pove, da gre za zasebni diskurz. Glede na dejstvo, da je šola ustanova formalnega diskurza, je lahko to za učenca dober dokaz, da bi bila v njegovem primeru ustreznejša raba strukture *morati* + *nedoločnik*.

5.2 Poklicni govorci in snovalci besedil

Predstavljajmo si situacijo, v kateri želi igralec ali pisec literarnega dela usvojiti diskurz določenega profila govorca za potrebe svojega lika. V tem primeru bo zagotovo koristno filtriranje besedil glede na tip situacije in profil udeležencev. Smiselno je izbrati zelo pogosto besedo, ki nastopa v vsaki govorni situaciji, na primer vse oblike glagola *biti*. Izberemo enostavno iskanje po standardiziranem zapisu, vpišemo *biti* in vidimo, da je rezultatov res veliko (okrog 90.000; Slika 2).

Levo od rezultatov so filtri diskurzov in govorcev. Pri tipu diskurza izberemo zasebni diskurz, pri regiji snemanja Dolenjska, pri podatkih o govorcu pa izberemo moške med 35. in 59. letom s končano srednješolsko izobrazbo. Dobimo približno 65 izjav izbranega profila govorca v izbrani govorni situaciji, s katerimi si lahko pomagajo poklicni govorci in snovalci besedil (igralci, scenaristi, novinarji, prevajalci).

6 Prvi odzivi na korpus GOS

Prve odzive uporabnikov med učitelji slovenščine,[5] med raziskovalci s področja jezikovnih tehnologij[6] ter med bodočimi

5 Predavanje na Filozofski fakulteti Univerze v Mariboru, predavanje na Filozofski Fakulteti Univerze v Ljubljani, konferenca Sirikt 2010.

poklicnimi govorci in snovalci besedil[7] lahko strnemo v naslednje komentarje in vprašanja:

– komaj čakam, da korpus GOS uporabim pri svojem delu,
– kdaj bodo na voljo obsežnejše delavnice za delo s korpusom GOS,
– kdaj bo zbirka nadgrajena z dodatnim naborom besedil (da bo po njej na primer relevantno tudi iskanje kolokacij).

7 Zaključek

Korpus govorjene slovenščine kot edinstvena elektronska zbirka avtentičnih govorjenih besedil v najrazličnejših govornih situacijah predstavlja jezikovni vir, ki uporabnika postavi v dejavno vlogo raziskovalca izgovorjave, besedišča in struktur govorjene slovenščine v domačem in zasebnem okolju, pa tudi govorjene slovenščine, ki jo zahtevajo bolj formalni govorni položaji.

Glede na prve odzive strokovne javnosti lahko zaključimo, da se je korpus GOS v slovenskem raziskovalno-strokovnem prostoru že prijel. Zdi pa se smiselno, da ostanemo previdni pri pričakovanjih glede moči demokratizacije tega novega dosežka jezikovnih tehnologij. V okviru programa mWomen[8] so med muslimanske ženske razdelili na tisoče mobilnih telefonov, da bi jim zagotovili bolj demokratičen status (več stikov z vrstnicami ter boljše možnosti za zaposlitev in izobraževanje), vendar so mobilni telefoni končali v rokah njihovih soprogov, ki jim jih posodijo le, ko morajo na pot, da jih lahko bolje nadzorujejo.

V primeru korpusa GOS seveda zlorabe ne morejo biti tako radikalne, vendar bi bilo žalostno, če bi tak jezikovni vir ostal zgolj v rabi znanstvenoraziskovalne elite. Nam bo s korpusom GOS uspelo zbuditi zanimanje širšega kroga uporabnikov? Bo korpusu GOS uspelo zmanjšati frustracije maternih govorcev slovenščine glede njihovih jezikovnih zmožnosti in zbuditi veselje pri njihovem vsakodnevnem govornem sporazumevanju? Če bomo nekega dne dosegli ta cilj – morda tudi s pomočjo težko priborjenega prostega dostopa na spletu in uporabniku prijaznega vmesnika – bo to velika reč.

Literatura

BURNS, Anne, 2001: *I see what you mean: Using Spoken Discourse in the Classroom*. Sydney: National Centre for English Language Teaching and Research.

FERBEŽAR, Ina, 2003: Dialog med ciljem in metodo. *Jezik in slovstvo* 48/1. 31–44

KRNEL, Dušan, 2008: Uporaba informacijsko-komunikacijske tehnologije (IKT) pri pouku v nižjih razredih osnovne šole. *Naravoslovna solnica* 13/1. 6–9.

SMOLEJ, Mojca, 2006. *Vpliv besedilne vrste na uresničitev skladenjskih struktur: Primer narativnih besedil v vsakdanjem spontanem govoru. Doktorska disertacija*. Ljubljana: Filozofska fakulteta.

STABEJ, Marko, VITEZ, Primož, 2000: KGB (korpus govorjenih besedil) v slovenščini. *Informacijska družba IS'2000: Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan.

VERDONIK, Darinka, 2007: *Jezikovni elementi spontanosti v govoru*. Maribor: Slavistično društvo Maribor.

ZEMLJARIČ MIKLAVČIČ, Jana, 2009: *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete.

ZEMLJARIČ MIKLAVČIČ, Jana idr., 2009: Kaj in zakaj v referenčni govorni korpus slovenščine. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike. Obdobja 28*. Ljubljana: Znanstvena založba Filozofske fakultete.

ZWITTER VITEZ, Ana 2011: À la recherche de la parole perdue: l'exemple du corpus national du slovène parlé. Noelle Serpollet (ur.): *Colloque Transcrire, Ecrire, Formaliser II.* Orléans: Presses universitaires d'Orléans.

ZWITTER VITEZ, Ana idr., 2009: Načela transkribiranja in označevanja posnetkov v refe-

6 Konferenca Jezikovne tehnologije na Institutu Jožef Stefan 2010, konferenca Transcrire, Ecrire, Formaliser v Orleansu 2011.

8 www.mwomen.org

ZWITTER VITEZ, Ana idr., 2009: Načela transkribiranja in označevanja posnetkov v referenčnem govornem korpusu slovenščine. Marko Stabej (ur.): *Infrastruktura slovenščine in slovenistike*. *Obdobja 28.* Ljubljana: Znanstvena založba Filozofske fakultete.

www.korpus-gos.net
www.stat.si
http://ola.zrc-sazu.si