Govorni korpus kot lektorjev priročnik

Darinka Verdonik
darinka.verdonik@uni-mb.si
Univerza v Mariboru, Fakulteta za elektrotehniko, računalništvo in informatiko

V letu 2011 smo dobili dolgo napovedovan in pričakovan govorni korpus slovenščine v obsegu milijon besed oz. 110 ur govora. Potem ko je bil dan v uporabo, so seveda ostale želje po njegovi rasti in razvoju, hkrati pa se je začelo zastavljati tudi bolj praktično vprašanje: kako ga približati tistim, ki jim je namenjen. Ena od skupin uporabnikov so tudi lektorji, ki imajo pri svojem delu tako ali drugače stik z govorjenim jezikom – bodisi so to lektorji v radijskih ali televizijskih hišah, gledališču ipd. bodisi lektorji, ki pri lektoriranju naletijo na besedila, ki zajemajo besedje ali slog tudi iz govorjenega jezika (leposlovje, kolumne idr.).

Lektorji si lahko pri svojem delu pomagajo s pripomočki, ki bi jih razdelila v tri skupine. V prvi so normativni priročniki – razni slovarji in slovnica/-e. V drugo skupino bi uvrstila različne slogovne priročnike, ki so tradicionalno tudi del lektorjeve knjižnice, je pa treba opozoriti, da v nasprotju z normativnimi priročniki praviloma nimajo formalnega institucionalnega zaledja. Njihova splošna veljavnost je tako manj zanesljiva. V tretjo skupino pa štejem sodobne korpusne vire, s čimer ne mislim samo urejenih korpusov, kot so Fida+, Nova beseda, Evrokorpus idr., ampak tudi spletne brskalnike (Google, Yahoo! ...), ki jih lahko lektor prav tako uporabi kot vir informacij o jezikovni rabi.

Ko razmišljamo o govornem korpusu kot lektorjevem pripomočku, je smiselno gledati nanj kot na predstavnika tretje skupine priročnikov. Ti so bistveno različni od ostalih priročnikov v tem, da prinašajo uporabniku samo informacijo o vsakdanji rabi jezika v določenih kontekstih, ne vključujejo pa nobenega priporočila ali napotila – uporabnik mora povsem samostojno sprejeti odločitev o jezikovnem vprašanju, na katero je naletel. Uporabni so predvsem v primerih, ko lektor naleti na primere, za katere ne najde odgovora v normativnih priročnih, bodisi zato, ker ti niso vedno ažurni, bodisi zato, ker ne obravnavajo vseh posebnosti, ki se lahko pojavijo v jeziku. Ko govorimo o govorjenem jeziku in govornem korpusu, je ta vidik pomembnejši kot pri kateremkoli drugem korpusu, saj ima slovenski lektor govorjenega jezika le malo priročnikov, na katere se lahko opre. V nadaljevanju bom navedla nekaj primerov, kako se lahko opre na referenčni govorni korpus slovenščine GOS, ob tem pa tudi izpostavila najpomembnejše vidike govornega korpusa.

Korpus GOS

Referenčni govorni korpus slovenskega jezika GOS je bil izdelan v okviru projekta Sporazumevanje v slovenskem jeziku, spletni konkordančnik, ki omogoča prost dostop in iskanje po govornem korpusu GOS, pa v okviru projekta e-vsebin Spletni konkordančnik za nacionalni govorni korpus slovenskega jezika. Vsebuje transkripcije različnih govornih dogodkov v skupnem obsegu zapisa 1 mio. besed in je uravnotežen v smislu, da zajema vzorčne primere različnih govornih situacij in različnih govorjenih besedil, da zajema demografsko reprezentativen vzorec govorcev in da zajema predvsem tiste govorne situacije, v katerih so uporabniki jezika najbolj pogosto produktivno-receptivno udeleženi. Za uporabnike je prosto dostopen na naslovu www.korpus-gos.net, lektor torej za njegovo uporabo, tako kot za mnoge ostale pomembnejše slovenske korpusne vire, potrebuje le računalnik z dostopom v splet.

Primer 1: Izgovorjava

Vzemimo za prvi primer besedo »premier«. Normativni priročniki predvidevajo dve enakovredni izgovorjavi, »premjêr« in »premjé«. Če se kot lektor želimo prepričati, katera varianta je pogostejša, lahko po GOS-u poiščemo vse pojavitve te besede. Dobimo 73 zadetkov. Če želimo, lahko izbor filtriramo, tako da izberemo na primer samo tiste pojavitve, ki se pojavijo v informativno-izobraževalnih vsebinah na radiu in televiziji, kjer lahko pričakujemo najbolj zborno različico slovenščine v rabi. Tako skrčimo zadetke na 33.

Ta primer nas opozori na prvo pomembno lastnost govornega korpusa, to je avdio material. Vsako izjavo v korpusu GOS lahko prek spletnega vmesnika tudi poslušamo, kar je zelo pomembno za marsikateri vidik uporabe korpusa, toda – avdio gradiva ne moremo avtomatsko statistično analizirati.

Zadetke lahko poslušamo in na podlagi tega seveda tudi delamo manjšo statistiko, če želimo, vendar je to časovno zahtevno in nikoli ne bomo mogli analizirati res velikega števila primerov, kot to naredimo strojno.

Primer 2: Pogovorne različice
Za drugi primer izberimo besedico »lahko«. Če se kot lektor govorjenega jezika srečamo s pogovorno zvrstjo, nam bo prišlo prav, da lahko preverimo, kakšne so realizacije besed v govorjenem jeziku. Za besedico »lahko« najdemo naslednje realizacije, navedeno od najbolj do najmanj pogoste: lahko, loh, lah, lahk, lohk, lehko, lahku, lohku, leko, uohk, lohka, lahka, lejko, lahka, lehku, lek, lohko, uohk, lh, leh, lahke, uhka, lako, lahke ...

Ta primer opozori med drugim na pomembno lastnost govorjenega jezika, ki je v korpusu GOS izpostavljena: številne besede imajo različne izgovorne različice, odvisno od regije in okoliščin jezikovne rabe. Da bi to lastnost ohranili vidno in omogočili njeno preučevanje, je zapis govora dvojen: pogovorni, ki sledi logiki »zapiši, kot slišiš«, in standardizirani, ki sledi logiki »zapiši, kot pišemo«. GOS-ov spletni iskalnik omogoča iskanje po enem ali drugem zapisu.

Primer 3: Govorjeno izrazje
S tretjim primerom naj opozorim na govorjeno izrazje, s tem mislim tisto, ki ga slovarji slovenskega jezika do zdaj ali niso zabeležili ali pa ga označujejo kot slengovsko, pogovorno, narečno ipd. Za tovrstno izrazje samo v pisnih virih ne najdemo dovolj informacij za ustrezen slovarski opis. Pogost tak izraz je na primer beseda »fora«. V nekaterih okoljih ali rabah bo za isti pomen uporabljena beseda »finta«. Če ima lektor opraviti z besedilom, v katerem se avtor izraža v pogovorni zvrsti, ga bo morda zanimalo, kateri od teh izrazov je pogostejši, v katerih regijah in ali resnično nastopata oba izraza v enakem pomenu ali ne. Opis teh izrazov v obstoječih slovarjih je zelo skop in dopolnitev vedenja o njih s pomočjo korpusa bo lektorju omogočila bolj samozavestno in strokovno odločitev.

Zvrstnost govorjenega jezika
Lektor govorjenega jezika se bo pogosto spraševal o t. i. zborni slovenščini. Kje naj jo išče po korpusu GOS? GOS v skladu z doktrino korpusnega pristopa kolikor le mogoče ne vrednoti zbranega gradiva, zato tudi ne pripisuje zvrstnih in funkcijskih oznak. Ima pa bogat nabor informacij o gradivu, ki se kolikor mogoče opirajo na dejstva, in spletni vmesnik, ki omogoča natančno navigacijo med temi informacijami. Tako lahko zadetke iskanja filtriramo glede na situacijo govorjenja, kanal snemanja, regijo in leto snemanja, opis govornega dogodka ter lastnosti govorca (spol, starost, izobrazba itd.). Uporabnik zato resda ne more izbrati gradiva, ki bi bilo označeno kot zborna zvrst, lahko pa na primer filtrira zadetke samo na tiste, ki so izrečeni v javnem informativno-izobraževalnem diskurzu na radiu in televiziji, ali – če ga zanima narečje – na zasebni diskurz ter želeno regijo govorcev ipd. S takim izborom se kar najbolj približa tisti situaciji, v kateri bo uporabljeno oz. ki jo imitira besedilo, ki ga lektorira.

Obseg govornega korpusa
Ob koncu ne moremo mimo opozorila uporabnikom korpusa glede obsega in vrste gradiva, po katerem iščemo v korpusu GOS. Zadetki za tukaj izbrane primere so se gibali med 20 in 100, le za primer »lahko« poskočijo na 4000. Za primerjavo: »lahko« ima v pisnem korpusu Fidaplus čez mejnih 100.000 zadetkov. Medtem ko se referenčni pisni korpusi merijo v sto milijonih besed, se referenčni govorni korpusi merijo samo v milijonih besed – ne le pri nas, ampak tudi v svetu. Vendar je tako milijon besed, kolikor jih trenutno obsega GOS, kot tudi pet ali deset milijonov še vedno zelo malo za reprezentativno sliko govorjenega jezika. Ker zbiranje večje količine ovira izredna časovna in finančna zahtevnost takih projektov in ker ni videti, da bi v kratkem obseg govornih korpusov bistveno poskočil, je še veliko bolj kot pri pisnih korpusih potrebna visoka odgovornost uporabnikov, da pravilno interpretirajo rezultate. To zahteva med drugim dobro poznavanje gradiva, po katerem iščemo, in prav zato zadetke v korpusu GOS spremlja legenda z informacijami o zadetkih in podroben opis gradiva pod posebnim zavihkom. Ustrezna interpretacija rezultatov, čeprav ne za raziskovalne, ampak povsem praktične namene, je mogoča samo skupaj s temi informacijami.