Kaj in zakaj v referenčni govorni korpus slovenščine

Jana Zemljarič Miklavčič, Marko Stabej, Simon Krek, Ana Zwitter Vitez

Eden izmed ciljev nacionalnega projekta *Sporazumevanje v slovenskem jeziku* je nastajajoči referenčni korpus govorjene slovenščine. V prispevku bo predstavljena shema za zajem gradiv, ki temelji na demografskih in besedilnovrstnih kriterijih. Pojasnjena bodo izbrana izhodiščna razmerja zastopanosti govorcev in besedilnih vrst, ki po prepričanju avtorjev v okviru danih možnosti zagotavljajo referenčnost in uravnoteženost govornega korpusa.

One of the goals of the "Communication in Slovene" project is the compilation of the reference corpus of spoken Slovene. The paper describes the design of spoken sample collection which focuses on demographic and text genre criteria. Basic speaker-text genre ratios will be described which in the opinion of the authors ensure that the referential and balanced character of the spoken corpus is maintained, within the limits of the project framework.

Ključne besede:
govorni korpus, referenčni korpus, govorjeni jezik, demografsko uravnotežen korpus, besedilnovrstno uravnotežen korpus
spoken corpus, reference corpora, spoken language, demographically balanced corpus, typologically balanced corpus

1 Govorni korpus slovenščine – čemu in za koga

Govorno sporazumevanje je primarna oblika sporazumevanja tako s stališča posameznika kot s stališča družbe, poleg tega je oblika, v kateri se jezik najpogosteje udejanja. Kljub temu v raziskanosti zaostaja za pisnim jezikom, po eni strani zaradi tradicionalno prestižne vloge knjižnega jezika, po drugi strani pa zaradi akustične pojavnosti, ki otežuje zbiranje gradiva in njegovo analizo. Večje raziskave govora je omogočil šele razvoj digitalnih tehnologij v zadnjih desetletjih: digitalno snemanje in shranjevanje podatkov, možnost urejanja in analize z računalnikom – vse to prinaša mnoga sodobna orodja za raziskovanje govora. Uporabnost govornih korpusov je usmerjena tudi v prihodnost, v razvoj govornih tehnologij (razpoznave/sinteze govora ...) in z njimi povezanih aplikacij. Ne nazadnje pa predstavlja govorni korpus tudi nadvse dragocen vir kulturne dediščine za naše zanamce. Številne jezikovne skupnosti so že zgradile govorne korpuse in s tem omogočile raziskovanje govorjenega jezika, gradnja referenčnega govornega korpusa pa je stekla tudi na Slovenskem.

Potreba po izdelavi govornega korpusa 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 in 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). 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, s čimer so bili natanko desetletje po prvem pozivu h gradnji izpolnjeni vsi pogoji za začetek.

2 Govorni korpusi

Začetki gradnje sodobnih govornih korpusov segajo v devetdeseta leta prejšnjega stoletja. Dolgo najvplivnejši referenčni vir za gradnjo govornih korpusov je bil *Britanski nacionalni korpus*, ki vključuje govorni podkorpus velikosti 10 milijonov besed;[2] zgrajen je bil v letih 1990–1994, sestavljen pa je iz demografsko in kontekstualno uravnoteženega dela. Drugi britanski referenčni govorni korpus je *Bank of English***,** ki pa je zelo skopo dokumentiran.[3] Med neangleškimi korpusi lahko omenimo *Češki govorni korpus,*[4] ki ga sestavlja več enot skupne velikosti čez 3 milijone besed in ga postopno gradijo že več kot desetletje. Enega večjih pravkar začetih projektov gradnje govornega korpusa predstavlja *Poljski nacionalni korpus* (Przepiorkowski et al. 2008), ki bo vključeval govorni podkorpus javnega govora v obsegu 30 milijonov besed, 3 milijone besed pa naj bi obsegal podkorpus vsakdanjih pogovorov. *Švedski govorni korpus*[5] obsega 1,4 milijone besed, *Nizozemski govorni korpus*[6] pa skoraj 9 milijonov besed. *Korpus govorjene italijanščine*[7] obsega 100 ur govora ali okrog 900.000 besed, drugi večji govorni korpus za italijanščino LABLITA[8] pa sestavlja govor odraslih in govor otrok v starosti od 15 do 36 mesecev. *Referenčni korpus sodobne portugalščine*[9] vključuje govorni podkorpus približno 2,5 mio. besed, gradivo pa je bilo zajeto v vseh portugalsko govorečih deželah po svetu. Navedli smo le nekaj najzgodnejših in največjih govornih korpusov, iz kratkega pregleda pa je razvidno, da je govorni korpus danes nujni jezikovni vir oziroma standard za raziskovanje govorjenega jezika.

Tudi za slovenski jezik že obstaja nekaj srednje velikih in manjših govornih zbirk in korpusov. Zajemajo televizijske dnevnoinformativne in pogovorne oddaje (BNSI Broadcast News, Broadcast News Speech Database), parlamentarni diskurz (transkripcije razprav iz Državnega zbora v FidiPLUS ter baza Sloparl), telefonske pogovore s turističnimi agencijami (Turdis) ter osnutek referenčnega korpusa (Zemljarič Miklavčič 2007). Poleg navedenih obstaja še nekaj drugih govornih baz, ki pa ne vključujejo avtentičnega govora, in nekaj transkripcij sicer avtentičnega govora, ki pa niso urejene tako, da bi omogočale avtomatsko iskanje brez dodatnega urejanja.

3 Temeljna izhodišča za zajem besedil

Referenčni govorni korpus je namenjen raziskavam govorjenega jezika (in jezika nasploh). Velikost načrtovanega korpusa za slovenščino (1 milijon besed) izhaja iz razpoložljivih finančnih sredstev in časa. Avtorji se zavedamo, da je zaradi omejene velikosti referenčnost korpusa pogojna, vendar smo prepričani, da bo dobra uravnoteženost gradiva omogočila

[1] http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx
[2] http://www.natcorp.ox.ac.uk/corpus/creating.xml
[3] http://mycobuild.com/about-collins-corpus.aspx
[4] http://ucnk.ff.cuni.cz/english/index.html
[5] http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3
[6] http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm
[7] http://www.alphabit.net/Glottophilia/2007/02/clips-corpus-of-spoken-italian.html
[8] http://lablita.dit.unifi.it/corpora/descriptions/lablita/
[9] http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php

relevantne raziskave govorjenega jezika, hkrati pa verjamemo, da se bo korpus v prihodnosti nadgrajeval in širil.

Na zasnovo govornega korpusa pomembno vpliva njegova namembnost znotraj projekta *SSJ*, kjer predstavlja govorni podkorpus referenčnega korpusa, ta pa bo vir za izdelavo leksikalne podatkovne baze s podatki o frekvenci, pomenski strukturi in zgledi rabe ter izdelavo pedagoške korpusne slovnice. Seveda naj bi govorni korpus v čim večji meri omogočal tudi korpusne raziskave, neodvisne od projekta SSJ, zlasti v uporabnem in teoretičnem jezikoslovju.

Glede na navedena izhodišča ima gradnja govornega korpusa tri osnovne cilje:

- zajeti vzorčne primere različnih govorjenih diskurzov v različnih situacijah,
- zajeti govorjeni diskurz demografsko reprezentativnega vzorca govorcev,
- zajeti predvsem tiste govorne situacije, v katerih so uporabniki jezika najpogosteje produktivno-receptivno udeleženi.

Pri zbiranju gradiva si prizadevamo ohraniti avtentično dolžino diskurzov. Zajeti želimo diskurze, ki potekajo v naravnem govornem okolju in niso zrežirani. Pri tem pa je treba upoštevati nekatere ovire: s pravnega in etičnega vidika je zajemanje gradiva omejeno, ker je pred snemanjem po zakonu potrebno zagotoviti soglasje vseh govorcev. Izkušnje kažejo, da vsi govorci niso pripravljeni za sodelovanje, poleg tega se praviloma vsaj na začetku snemanja vedejo drugače kot sicer. Za snemanje je treba ustrezno namestiti snemalne naprave, kar prav tako vpliva na avtentičnost diskurza. Ker je za transkripcijo potreben kvaliteten zvok, je problematično snemanje v okoljih z veliko šumi ali z velikim številom govorcev. Brana ali na pamet naučena besedila v govorni korpus ne bodo zajeta, pogoj za vključitev je namreč najmanj delna spontanost.

Določene omejitve obstajajo tudi glede govorcev: govor otrok in mladostnikov se pogosto zbira ločeno od govora odraslih. Ker bo govorni korpus SSJ vir za izdelavo pedagoške slovnice, bo ne glede na to pomemben del korpusa zajemal tudi šolski govor otrok od 10. leta naprej. Zajet bo tudi delež govorcev, za katere slovenščina ni prvi jezik, pri čemer bomo skušali upoštevati realno demografsko sestavo govorcev slovenščine. Prav tako bo zajet ustrezni delež govorcev, ki živijo v zamejstvu; govor Slovencev po svetu pa v prvi fazi ne bo vključen v korpus.

3 Kriteriji za zajem gradiva v korpus SSJ

Glede na namen govornega korpusa smo se odločili, da pri zajemu kombiniramo demografske in besedilnovrstne kriterije. Demografski kriteriji so uravnoteženi na podlagi najnovejših podatkov Statističnega urada Republike Slovenije (SURS), besedilnovrstni pa deloma na podlagi hipotez, predvsem pa tako, da so ustrezno zastopani glede na namen govornega korpusa.

3.1 Demografski kriteriji

Demografski kriteriji, ki jih upoštevamo, so: spol, starost, izobrazba in regijski izvor. Gre za kriterije, ki glede na obstoječe raziskave v slovenskem jezikoslovju (prim. Zemljarič Miklavčič 2007) in glede na hipotetična predvidevanja najbolj vplivajo na razlike v govoru.

Pri tem ni upoštevan kriterij socialnega statusa, ker je za slovenske razmere težko določljiv in vprašljiv.

Demografski kriteriji in izhodiščna razmerja v odstotkih:[10]

1. **prvi jezik:**
   a. slovenščina 98 %
   b. drugi jeziki 2 %

2. **država bivanja:**
   a. Slovenija 97 %
   b. Avstrija 1 %
   c. Italija 1 %
   d. Madžarska 1 %

3. **spol:**
   a. moški 50 %
   b. ženski 50 %

4. **starost:**
   a. do 34 let 40 %
   b. nad 35 let 60 %

5. **dosežena izobrazba:**
   a. nižja (osnovna in srednja šola) 70 %
   b. višja (več kot srednja šola) 30 %

6. **regijska pripadnost:**

Regijsko pripadnost označujemo glede na večja regionalna mestna središča, h katerim gravitira posamezno področje in ki sovpadajo z registrskimi območji v Sloveniji.

A. zasebni diskurz:
   a. govor JZ Slovenije brez ljubljanske regije (NM, KK, LJ, KR, GO, PO, KP) – 35 %
   b. govor ljubljanske regije – 25 %
   c. govor SV Slovenije brez mariborske regije (MS, SG, CE) – 25 %
   d. govor mariborske regije – 15 %

B. javni in nejavni nezasebni diskurz:
   a. JZ Slovenija: 60 %
   b. SV Slovenija: 40 %

3.2 Besedilnovrstni kriteriji

Definiranje besedilnovrstnih kriterijev je težavnejše kot definiranje demografskih kriterijev. Možni kriteriji so struktura besedila (monolog, dialog/multilog), okoliščine (javna besedila, zasebna besedila), govorni položaj (formalni, neformalni), prenosnik (osebni stik, telefon, avdio, video), namernost, tematika itd. Vendar v jezikovni rabi pogosto prihaja do prekrivanja

[10] Odstotki pri demografskih kriterijih so izraženi glede na tisti del gradiva, ki bo demografsko uravnotežen, tj. zasebni diskurz, razen če je drugače navedeno.

več lastnosti (npr. prevladuje monolog, ki pa občasno preide v dialog), tako da gradiva ni mogoče nedvoumno razvrščati v kategorije. Ob upoštevanju zastavljenih ciljev govornega korpusa, potreb korpusnih raziskav in statistične reprezentativnosti posameznih skupin glede na predvideni obseg korpusa smo se odločili samo za dva osrednja besedilnovrstna kriterija: javnost diskurza in prenosnik.

Besedilnovrstni kriteriji in izhodiščna razmerja v odstotkih:[11]

1. **javnost:**
   a. javni diskurz 60 %
      i. razvedrilni 20 %
      ii. nerazvedrilni 40 %
   b. nejavni diskurz 40 %
      i. nezasebni 15 %
      ii. zasebni 25 %

2. **prenosnik:**
   a. osebni stik 50 %
   b. telefon 10 %
   c. radio 20 %
   d. televizija 20 %

Za javni diskurz štejemo diskurz, ki je odprt za širšo javnost ali naslavlja veliko skupino ljudi. Nejavni diskurz ločimo na zasebni diskurz (v okviru družine, prijateljev, znancev), nejavni nezasebni diskurz pa vključuje uradne in poluradne diskurze (v uradih, trgovinah, ob storitvah, v profesionalnem življenju ipd.). V javnem diskurzu ločimo medijske vsebine, ki so predvsem razvedrilne, ter medijske vsebine, ki so predvsem informativne/izobraževalne/družbene.

3.3 Dodatni kriteriji za zajemanje pedagoškega diskurza

V skladu s cilji govornega korpusa v okviru projekta SSJ bo znaten del korpusa zajemal pedagoški diskurz v osnovni in srednji šoli. Ta del bo zasnovan kot podkorpus, ki sledi spodaj predstavljenim dodatnim kriterijem za zajem. V besedilnovrstni klasifikaciji zgoraj je zajet v kategoriji javni nerazvedrilni diskurz, zajema pa 15 % celotnega korpusa. Kriteriji so uravnani glede na podatke Ministrstva za šolstvo in šport RS in SURS.

1. stopnja šolanja:
   a. osnovna šola 8 %[12]
      - 2. triletje 4 %
      - 3. triletje 4 %
   b. srednja šola 7 %
      - gimnazije 3 %
      - nižje in srednje poklicno, srednje strokovno in poklicno-tehniško izobraževanje 4 %
2. regija:
   a. JZ 9 %

[11] Odstotki pri besedilnovrstnih kriterijih so izraženi glede na celoten korpus.
[12] Odstotki so izraženi glede na celoten korpus.

b. SV 6 %

3. učni predmet:
   a. naravoslovni in tehnični predmeti 7,5 %
   b. družboslovni in humanistični predmeti 7,5 %

3.4 Skupna preglednica zajetih tipov govorjenega diskurza

Sestava celotnega korpusa (odstotki so izraženi glede na celoten korpus):

OKOLIŠČINE	**%**	**NAMEN**	**%**	**PRENOSNIK**	**%**	**REGIJA**	**%**
javni govor	60 %	informativno-izobraževalni	40 %	TV[13]	10 %	SV	4
						JZ	6
				radio	10 %	SV	4
						JZ	6
				osebni stik	20 %	SV	8
						JZ	12
		razvedrilni	20 %	TV	10 %	SV	4
						JZ	6
				radio	10 %	SV	4
						JZ	6
							60,00
nejavni govor	40 %	nezasebni	15 %	telefon	5 %	SV	2,00
						JZ	3,00
				osebni stik	10 %	SV	4,00
						JZ	6,00
		zasebni[14]	25 %	telefon	5 %	SV[1]	1,25
						MB[1]	0,75
						JZ[1]	1,75
						LJ[1]	1,25
				osebni stik	20 %	Italija	1,00
						Avstrija	1,00
						Madžarska	1,00
						Neslovenci	2,00
						SV[1]	3,75
						MB[1]	2,25
						JZ[1]	5,25
						LJ[1]	3,75
SKUPAJ:	**100%**				**100,00**		**40,00**

4 Zaključek

[13] Seznam TV in radijskih oddaj, ki jih bomo zajeli v GK, zajema predvsem programe in oddaje z najvišjo gledanostjo.
[14] Govorci v zasebnem diskurzu, razen tujejezičnih govorcev in Slovencev v zamejstvu, bodo uravnoteženi glede na spol, starost in izobrazbo.

Navedena shema je po prepričanju avtorjev v okviru danih možnosti takšna, da zagotavlja visoko stopnjo referenčnosti in uravnoteženosti govornega korpusa, hkrati pa služi namenom korpusa znotraj projekta *Sporazumevanje v slovenskem jeziku*. Specifikacije, v katerih je bila definirana shema zajema (v nadaljevanju pa tudi načela transkribiranja in urejanja gradiva, ki bodo predstavljena v drugih prispevkih) so bile izdelane do aprila 2009. Gradnja korpusa – zbiranje in transkribiranje gradiv – je stekla spomladi 2009, končana pa bo decembra 2010. Predvidoma od takrat naprej bo referenčni govorni korpus tudi dostopen raziskovalni in drugi zainteresirani javnosti, ki ga, kot smo videli, že nestrpno pričakuje.

5 Literatura

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

Kenda Jež, Karmen, 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.

Przepiorkowski, A., Gorski, R., Lewandowska-Tomaszczyk, B., Lazinski, M., 2008: Towards the national corpus of Polish. *Proceedings of 6th Language Resources and Evaluation Conference*, Maroko, Marakeš.

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

Stabej, Marko, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. *Jezikovne tehnologije* (tematska št., ur. Z. Kačič). 96–106.

Verdonik, Darinka, Rojc, M., 2006: Are you ready for a call? - Spontaneous conversations in tourism for speech-to-speech translation systems. *5th International Conference on Language Resources and Evaluation*, Genova, Italija.

Verdonik, Darinka, 2006: *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora.* Doktorska disertacija. Mentorja: M. Stabej in Z. Kačič. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Weiss, Peter, 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, Jana, 2007: *Načela oblikovanja govornega korpusa za slovenščino*. Doktorska disertacija. Mentorja: M. Stabej in V. Gorjanc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

Žgank, A., T. Rotovnik, D. Verdonik, Z. Kačič, 2004: Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. *Informacijska družba IS'2004: Jezikovne tehnologije*. 94−98.

Žgank, A., Rotovnik, T., Grašič, M., Kos, M., Vlaj, D., Kačič, Z., 2006: Slovenska govorna baza parlamentarnih razprav za avtomatsko razpoznavanje govora. *Informacijska družba IS'2006: Jezikovne tehnologije*. Ljubljana, Institut Jožef Stefan. 115−118.

Žibert, J., Mihelič, F., 2004: Development, evaluation and automatic segmentation of Slovenian Broadcast News Speech Database. *Informacijska družba IS'2004: Jezikovne tehnologije.* Ljubljana, Institut Jožef Stefan. 94−97.