Korpus govorjene slovenščine

Jana Zemljarič Miklavčič

Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta, Univerza v Ljubljani
Aškerčeva 2, 1000 Ljubljana

Povzetek

Učni korpus govorjene slovenščine je bil zgrajen kot teoretična in praktična podlaga za gradnjo večjega govornega korpusa slovenščine, ki naj bi dopolnjeval korpus pisnih besedil FidaPlus. Učni korpus sestavljajo digitalni posnetki spontanega govora, ki so bili zbrani po različnih taksonomskih in demografskih kriterijih. Posnetki so bili transkribirani v razširjeni ortografski transkripciji. Med transkribiranjem so bila določena načela transkribiranja in označevanja. Učni korpus je v obliki, ki omogoča iskanje po korpusu, z geslom dostopen na spletnem naslovu http://torvald.aksis.uib.no/talem/jana/s9.html; dostopna so cela besedila ter konkordance in kolokacije posameznih besed. Za iskanje po korpusu so na voljo različni kriteriji, transkripcije pa so povezane z zvočnimi posnetki.

Spoken Corpus of Slovene

A pilot corpus of spoken Slovene has been compiled to establish a theoretical and empirical foundation for building a large spoken corpus of Slovene, which is planned to complement the written FidaPLUS corpus. Pilot corpus is based on digital recordings of spontaneous speech, collected according to different contextual and demographic criteria. The recordings have been transcribed in enriched orthographic transcription. During actual transcription work, the transcription and annotation standards have been outlined. Pilot spoken corpus is available in searchable form at http://torvald.aksis.uib.no/talem/jana/s9.html (authorization needed); the whole texts are accessible, as well as concordances and collocations of single words. Different criteria could be used for searching the corpus. In each case, transcriptions are linked to sound files.

1. Uvod

V času mojega trimesečnega študijskega bivanja na Oddelku za kulturo, jezik in jezikovne tehnologije[1] na Univerzi v Bergnu na Norveškem je nastal manjši korpus govorjene slovenščine, ki naj bi služil kot učni korpus za gradnjo govorne komponente referenčnega korpusa. Namen gradnje učnega korpusa je bil spoznati metode zbiranja, shranjevanja in dokumentiranja govorjenih besedil, razviti in testirati načela transkribiranja, določiti in testirati korpusne oznake ter pokazati nekatere možnosti za uporabo govornega korpusa.

2. Korpusi govorjenega jezika

Korpusi govorjenega jezika so računalniške zbirke transkribiranih posnetkov spontanega govora; razlikujemo jih od korpusov govora, kjer gre običajno za studijske posnetke izoliranih izjav, ki nastajajo za potrebe fonetično-fonoloških raziskav in govornih tehnologij (Gorjanc 2005: 8). Korpusi govorjenega jezika so izrednega pomena za raziskovanje jezika, predvsem njegovih slovnično-leksikalnih lastnosti: uporabljajo se za jezikovne opise, za preverjanje hipotez o jeziku, kot jezikovni vir pri poučevanju in učenju tujega jezika, pa tudi pri raziskavah, ki zadevajo sintezo in razpoznavanje govora (Verdonik 2006: 7). Govorni korpusi so v nasprotju z realno jezikovno produkcijo mnogo manjši od pisnih, ker jih je izredno težko graditi; največji doslej zgrajeni govorni korpusi so govorna komponenta BNC, govorna komponenta BoE (obe velikosti okrog 10 milijonov besed) in Nizozemski govorni korpus (8,3 milijona besed), ki je najmlajši izmed njih (2004) in ima edini transkripcije povezane z zvočnimi posnetki.[2]

V ortografaskem prepisu govora se izgubijo številne značilnosti originalnega govornega dogodka. Izgubo teh informacij je mogoče vsaj deloma nadomestiti s prozdičnimi oznakami in fonetično transkripcijo, pomembna pa je tudi sočasna dostopnost zvočnih posnetkov. Stopnja transkripcije je odvisna od namembnosti korpusa; govorni korpusi so zaradi velikanske količine gradiva običajno transkribirani v ortografski transkripciji, na manjšem delu korpusa pa je lahko narejena tudi fonetična transkripcija.

3. Korpus govorjene slovenščine

3.1. Zajem besedil v korpus

Vprašanje reprezentativnosti korpusa se zastavlja na začetku vsakega razmišljanja o gradnji korpusa in je podlaga za vse kasnejše postopke pri uresničevanju gradnje. Če želimo iz korpusa dobiti relevantne podatke o jeziku, mora korpus kot vzorec izkazovati podobne lastnosti kot govorjeni jezik celotne populacije, ki jo želi predstavljati. Reprezentativnost govornih korpusov se običajno dosega s kombinacijo demografske in besedilnovrstne metode zbiranja (Crowdy 1993: 260); pri prvi izberemo reprezentativni vzorec govorcev celotne populacije (glede na demografske lastnosti, kot so spol, starost, izobrazba, regijski izvor), pri drugi pa besedila zajemamo na podlagi besedilnovrstne taksonomije govorjenih besedil.

Pri gradnji učnega korpusa zaradi časovnih in "človeških" omejitev ni bilo mogoče v celoti upoštevati načel za reprezentativni in uravnoteženi zajem besedil. Kljub temu sem se pri zbiranju gradiva trudila, da bi se besedila čim bolj razlikovala po lastnostih, na katerih je zgrajena taksonomija govorjenih besedil, in da so se govorci čim bolj razlikovali po demografskih lastnostih. Korpus sestavlja 7 različno dolgih posnetkov v skupni dolžini 89 minut (15.000 besed), v njem pa je sodelovalo 20 govorcev; popis demografskih lastnosti govorcev je podan v Tabeli 1.

[1] http://www.aksis.uib.no/

[2] http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm#sample1

Kriterij	Porazdelitev glede na določene kategorije						?
Spol	ženske: 11			moški: 9			–
Starost	30 let ali manj: 8			30 let ali več: 11			1
Regija	LjO: 13	SZ: 0	SV: 3	Z: 0	J in JZ: 2	drugo: 1	1
Izobrazba	končana OŠ: 3		končana SŠ: 4		končana univ.: 12		–
Prvi jezik	slovenski: 19			drugo: 1			–

Tabela 1: Demografske lastnosti govorcev učnega korpusa

Razmerja znotraj posameznih kontekstualnih kriterijev so naslednja:
- dialogi (in multilogi) proti monologom: 94 % : 6 %,
- zasebna besedila proti javnim: 19,5 % : 80,5 %,
- neformalna besedila proti formalnim: 35,5 % : 64,5 %,
- besedila, posneta v osebnem stiku, proti besedilom s prenosnikom: 31 % : 69 %.

Velikost in struktura korpusa sta zadoščali za določitev korpusnih oznak, zadovoljivi pa sta bili tudi za določitev transkripcijskih standardov. Žal pa se pri tej velikosti ni bilo mogoče niti približati reprezentativnosti govornega korpusa; tega dejstva ne bi bistveno spremenila niti dvakratna ali celo trikratna količina gradiva. Glede na izkušnje pri gradnjah drugih govornih korpusov predvidevam, da bi bilo za slovenščino smiselno načrtovati govorni korpus velikosti 1 milijon besed, to pa je po mojem mnenju tudi največ, kar je v doglednem času mogoče doseči.

3.2. Transkribiranje učnega korpusa

Besedila učnega korpusa govorjene slovenščine so členjena na izjave, ki so omejene bodisi s premorom bodisi z menjavo govorcev. Izjave so transkribirane po priporočilih EAGLES (19996) v ortografski transkripciji brez ločil in brez velikih začetnic na začetku povedi. Velike začetnice imajo lastna imena, kadar ne gre za osebne podatke; ti so nadomeščeni z nevtralnimi oznakami. Besede, ki imajo znano (slovarsko) pisno obliko, so skoraj dosledno zapisane z upoštevanjem pisne norme; pri tem sem se zgledovala po drugih referenčnih govornih korpusih[3] ter sledila osnovni filozofiji delovne skupine EAGLES za govorjena besedila, ki temelji na načelu, naj bo pri transkribiranju spontanih govorjenih besedil v čim večji meri upoštevan standardni zapis besed, vse nestandardne oblike v transkripciji pa naj bodo jasno označene.[4] Zapisovanje govora z ortografsko transkripcijo postane problematično, kadar besede nimajo ustaljene pisne oblike ali od nje v govoru zelo odstopajo. Pri zapisovanju teh besed je treba iskati rešitve, ki morajo biti abstrahirane in transparentne. V nadaljevanju navajam nekaj primerov besed brez ustaljene pisne oblike (v kurzivu):

- *ene* tri ure smo čakali,
- a veš, *un* Michael, *uni* fjordi, *un* Grega,
- vsi *ta* glavni fjordi,
- a *pol* gremo,
- a *čmo* pogledat,
- to je *tle*, to je *tlele*,
- *ə* (polglasnik),
- citatne besede (*imamo mi posla i bez toga)*,
- narečne/žargonske/slengovske besede (*pležuh, tošel, bičikleta, jabčki)*.

Sicer pa so bile za označevanje lastnosti govora in za opis nejezikovnih dogodkov v učnem korpusu uporabljene oznake podane v Tabeli 2

Oznaka	Pomen
<pavza>	premor
<ime>	nadomešča osebno lastno ime
[besedilo]	prekrivni govor
<neraz>	nerazumljivi govor
<?>*besedilo*</?>	nezanesljiva transkripcija
<repeat>	ponavljanje
=	napačni začetek
<tj:>*besedilo*</tj>	besedilo, izgovorjeno v tujem jeziku
<nst>*besedilo*</nst>	beseda brez ustaljene pisne oblike[5]
<nv>smeh</nv>	neverbalni dogodki
 besedilo</br>	brano besedilo
(*opis*)	neverbalni zvoki v ozadju

Tabela 2: Oznake učnega korpusa

Za transkribiranje učnega korpusa sta bili uporabljeni dve transkripcijski orodji, Transcriber in Praat; obe orodji ob nadaljnjem procesiranju omogočata neposredno povezavo transkripcij in zvočnih posnetkov. V primeru gradnje večjega govornega korpusa bi bilo verjetno bolje uporabljati Praat, ker omogoča tudi akustične analize govora.

3.3. Konvertiranje učnega korpusa

Transkripcije je v korpus s konkordančnikom in s povezavo med transkripcijami in zvočnimi posnetki konvertiral Knut Hofland na Univerzi v Bergnu. Kot vidimo v Sliki 1, obe transkripcijski orodji, Praat in Transcriber, vsaki transkribirani izjavi z veliko natančnostjo pripišeta začetno in končno časovno oznako.

[3] Npr. govorni komponenti BNC in BoE, Nizozemskem govornem korpusu itd.

[4] http://www.ilc.cnr.it/EAGLES96/spokentx/node24.html#csor

[5] Ta oznaka je bila v učnem korpusu večkrat po nepotrebnem uporabljena; žal na oznake korpusa, ki je na strežniku Univerze v Bergnu, po izteku štipendije ne morem več vplivati.

```
<Turn speaker="spk2" startTime="325.991"
endTime="331.468">
<Sync time="325.991"/>
Ru&lt;lz&gt; Rupel je ne vem on je tak
svetovljan on je brihten človek
</Turn>
<Turn speaker="spk1 spk2"
startTime="331.468" endTime="334.563">
<Sync time="331.468"/>
<Who nb="1"/>
&lt;nst&gt;kurca&lt;/ nst&gt; je [brihten
&lt;neraz&gt;]
<Who nb="2"/>
[je je&lt;repeat/&gt; je je&lt;repeat/&gt;]
Rupel je hud veš
</Turn>
<Turn speaker="spk1" startTime="334.563"
endTime="347.261">
<Sync time="334.563"/>
takrat ko so bili Pankrti je zapisal v eno
revijo
<Sync time="337.567"/>
```

Slika 1. Izjave s pripisanim začetnim in končnim časom

Program razdeli časovni odsek med začetkom in koncem izjave s številom besed in naredi interpolacijo časa za vsako besedo znotraj izjave; na ta način je dosežena dokaj natančna sinhronizacija zvoka in transkripcije. Razumljivo je, da pri tem prihaja tudi do zamikov, zato pri poslušanju izjav v konkordancah ne slišimo vedno samo tistega, kar bi želeli.

Iskalna platforma korpusa omogoča enostavno iskanje po demografskih in/ali kontekstualnih kriterijih, ki so bili predvideni ob načrtovanju korpusa: po spolu govorca, izobrazbi, regiji, prvem jeziku, odnosu med govorci, pa tudi po nekaterih kontekstualnih kriterijih (skrivaj posneta besedila, tip in struktura besedila, okoliščine in prenosnik).[6]

Iskalno okno omogoča iskanje ene, dveh ali treh sosednjih besed in izpis njihovih konkordanc. Iščemo lahko cele besede, lahko pa le začetne ali končne dele besed. Poleg tega okno omogoča tudi prilagajanje sobesedila v konkordančnem izpisu in spreminjanje dolžine predvajanega zvočnega posnetka levo in desno od besed(e) v konkordanci.

4. Iskanje po korpusu

Besedila UKGS so dostopna na dva načina: kot celote ali preko konkordančnika. Dostopnost besedil, ki jih je mogoče v celoti ali po delih tudi poslušati, je za nekatere jezikoslovne analize zelo pomembna. Slika 2 prikazuje primer transkribiranega in označenega besedila, segmentiranega na izjave.

Običajnejši dostop do gradiva v korpusu je preko konkordančnika, kjer kot rezultat iskanja dobimo konkordančni niz: primer vidimo v Sliki 3. Na začetku vsake vrstice je šifra posnetka, iz katerega je vzeta izjava (npr. R06), in šifra govorca izjave (G17). S klikom na šifro prikličemo glavo besedila, kjer lahko preberemo podatke o okoliščinah nastanka besedila in demografske podatke o govorcu.

L	G17:	[ja {neraz} Mobi je izgubil zdaj]
	G16:	[{neraz} zgubil] je to kar so že podpisali ne
L	G16:	in potem predsednik ta ə Vege [ne]
	G17:	[ja]
L	G16:	əm predsednik uprave jaz ne vem kdo {neraz} saj {nst} nima veze {/nst}
L		əm Američan je skratka ne pač ə
L		se pizdi ne kako lahko zdaj Mobitel əm se zoperstavi Združenim narodom ne
L	G17:	{shift=vpr} a [res] {/shift=vpr}
	G16:	[Američan] a veš pizda
L	G16:	ne [oni ki imajo] {pavza} [ne {neraz}] potem pa oni ves zgrožen {shift=vpr} kako lahko {/shift=vpr} ne
	G17:	[ja ja zastopim] ki se itak ne šmirglajo [Združenih narodov]
L	G17:	[{nv} smeh {/nv}]

Slika 2. Del transkribiranega besedila

R04--G15	imam celo <nst> kle </nst>	**ne vem**	+G03+ <pavza>
R05--G11	to so pa tako široke da	**ne vem**	+G11+ kot da bi
R05--G11	[pobiramo te besede] ne pa	**ne vem**	+G12+ [ja ja]
R05--G11	podvozja stanejo okoli	**ne vem**	+G12+ [pet
R07--G20	pa pa <repet/> sem pozabil	**ne vem**	+G19+ ja ja
R06--G16	ima ąe ali kaj saj	**ne vem**	[meni se zdi ja da
R03--G08	najhujša +G08+ najlepša	**ne vem**	auv= vedno
R02--G02	in in in <repet/>	**ne vem**	besede in
R02--G03	əm mogoče bomo naredili	**ne vem**	dve ali tri poletne
R02--G02	recimo ljudje ki so	**ne vem**	dve uri na na
R06--G17	</okr> ali kje l'e saj	**ne vem**	ja ə če bi
R02--G02	tudi druge oblike recimo	**ne vem**	glede seminarja
R06--G17	ə </shift=vpr> pač <pavza>	**ne vem**	jaz zdaj teh izraz
R06--G16	a veš +G17+ [<neraz>	**ne vem**	kako jim to
R06--G17	<nst> pizda </nst> to pa	**ne vem**	ja čisto so
R05--G11	[<nv> smeh </nv>] +G11+	**ne vem**	kaj ąe ja to je to

Slika 3. Konkordančni niz in kolokacija besed "ne vem"

Med najbolj dragocene podatke v referenčnih korpusih sodijo podatki o frekvenci pojavljanja in sopojavljanja besed. Nekateri statistični podatki so dostopni tudi za učni korpus: v korpusu je okrog 15.000 pojavnic in 3118 različnic (o pojavnicah in različnicah prim. Gorjanc 2005); od tega se jih dve tretjini pojavi samo enkrat, kar naj bi bilo za govorne korpuse značilno. Najvišjo frekvenco v učnem korpusu ima oblika glagola biti "je", in sicer skoraj 500. Slika 4 prikazuje besede z najvišjo frekvenco pojavljanja v učnem korpusu.

1	498	35.422	je
2	425	30.230	ne
3	358	25.464	ə
4	313	22.263	pa
5	297	21.125	in
6	284	20.201	se
7	270	19.205	da
8	268	19.063	to
9	265	18.849	ja
10	264	18.778	v

[6] Oznake kriterijev so v angleščini, ki je bila sporazumevalni jezik med menoj in Knutom Hofmanom na Univerzi v Bergnu.

11	186	13.230	na
12	143	10.171	tudi
13	130	9.247	za
14	115	8.180	ki
15	106	7.540	so
16	105	7.469	tako
17	105	7.469	mhm
18	98	6.971	kaj
19	88	6.259	a
20	86	6.117	še
21	84	5.975	če
22	78	5.548	zdaj

Slika 4. Seznam besed z najvišjo frekvenco v učnem korpusu

V prvem stolpcu seznama je zaporedna številka (glede na frekvanco pojavljanja), v drugem stolpcu absolutna frekvenca (število pojavitev v korpusu), v tretjem stolpcu pa relativna frekvenca (število pojavitev na 1000 besed).

5. Zaključek

Učni korpus govorjene slovenščine je zaenkrat z geslom dostopen na spletnem naslovu http://torvald.aksis.uib.no/talem/jana/s9.html (geslo je mogoče dobiti na naslovu jana.zemljaric@ff.uni-lj.si). Največji pomanjkljivosti korpusa sta njegova neuravnoteženost in nereprezentativnost, kar je treba upoštevati pri morebitni nadaljnji uporabi korpusnih podatkov. Namen gradnje učnega korpusa je bil vendarle dosežen, saj so bila ob gradnji razvita načela zajemanja govorjenih besedil, predstavljene so bile različne možnosti transkribiranja spontanega govora, izbrane in preizkušene možnosti označevanja, prikazati pa je mogoče tudi nekatere možnosti uporabe govornega korpusa. Za nekatere oznake in načela korpusa so bili že podani predlogi za izboljšave, ki bodo upoštevani ob morebitni gradnji večjega govornega korpusa.

6. Literatura

Burnard, L., 2000. *Where did we go wrong? A retrospective look at the design of the BNC*. SILFI 6th International Conference, "Spoken Italian", Congress Proccedings. Duisburg, 28. 6.–2. 7. 2000. <http://users.ox.ac.uk/~lou/wip/silfitalk.html>

Crowdy, S., 1993. Spoken Corpus Design. *Literary and Linguistics Computing 8/4*. Oxford University Press. 259–265.

Crowdy, S., 1994. Spoken Corpus Transcription. *Literary and Linguistics Computing 9/1*. Oxford University Press. 25–28.

EAGLES preliminary recommendations on Spoken Texts, 1996. EAGLES (Expert Advisory Group on Language Engineering Standards) Spoken Language Working Group. <http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html>

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

Leech, G., G. Myers in J. Thomas (ur.), 1995. *Spoken English on Computer. Transcription, mark-up and application*. New York: Longman Publishing.

Llisteri, J., 1996. *Preliminary recommendations on Spoken Texts*. EAGLES (Expert Advisory Group on Language Engineering Standards). Version of May, 1996. http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html

Stabej, M., in P. Vitez, 2000. KGB (korpus govorjenih besedil) v slovenščini. V: ERJAVEC in GROS (ur.), *Jezikovne tehnologije za slovenski jezik*. Ljubljana: 79–81.

The spoken Dutsch corpus project, 2004. <http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm>

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

Zemljarič Miklavčič, J., 2004. Taksonomija besedilnih tipov za gradnjo govornega korpusa. V E. Kržišnik (ur.): *Aktualizacija jezikovnozvrstne teorije na slovenskem: Členitev jezikovne resničnosti. Obdobja 22*. Ljubljana: Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta Univerze v Ljubljani.