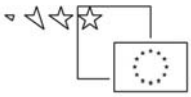
REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT



Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

SPORAZUMEVANJE V SLOVENSKEM JEZIKU – GOVORNI KORPUS

Vrsta dokumenta: Splošne specifikacije zbiranja gradiva

Datum: 31. marec 2009
Verzija: 1.0

KAZALO

1 UVOD

Namen tega dokumenta je definirati kriterije za zajem gradiva za govorni korpus (GK) in toleranco odstopanj od teh kriterijev, določiti govorne situacije za snemanje, želene karakteristike govorcev ter podatke o zajetem gradivu, ki bodo vključeni v korpus.

V svetu se v zadnjih desetletjih med drugimi jezikovnimi viri gradijo tudi številni govorni korpusi.[1] V nadaljevanju na kratko navajamo nekatere po obsegu in namenu najbolj primerljive, predvsem evropske projekte gradnje govornih korpusov ter nekatere najbolj relevantne slovenske korpuse in publikacije.

1.1 Pregled nekaterih tujih primerljivih projektov

Korpus britanske angleščine **(British National Corpus – BNC)** vključuje govorni podkorpus v obsegu 10 mio. besed (http://www.natcorp.ox.ac.uk/corpus/creating.xml). Ta je sestavljen iz demografskega dela, ki je uravnotežen po demografskih kriterijih: družbeni status, spol, regionalna razpršenost (govorci z 38 različnih lokacij) in starost. Po spolu, družbenem statusu in starosti so bile različne skupine enakomerno zastopane. Drugi del govornega podkorpusa je kontekstno usmerjen, in sicer skuša v enakomernih deležih zajeti 4 kategorije družbenega konteksta: izobraževanje in informiranje, poslovni dogodki, institucionalni in javni dogodki, kot so maše, politični govori, parlamentarna zasedanja, ter zabava (športni komentarji, klubska srečanja, nagovori ob večerji ...).

V angleščini obstaja vsaj še en za nas referenčen govorni korpus – **Bank of English** – ki pa je zelo skopo dokumentiran (http://mycobuild.com/about-collins-corpus.aspx). Osnovno vodilo avtorjev je kvantiteta in ažurnost korpusa, tako da mu mesečno dodajajo novo gradivo.

Češki govorni korpus (http://ucnk.ff.cuni.cz/english/index.html) je sestavljen iz treh enot: praškega govornega korpusa (PMK, zajema predvsem govor Prage in bližnje okolice, obseg je 675.000 besed), brnskega govornega korpusa (BMK, zajema predvsem govor Brna in bližnje oklice, obseg je 490.000 besed) ter ORAL2006 (zajema govor čeških narečnih področij, obseg je 1 mio. besed). Vsi trije podkorpusi se opirajo na sociolingvistične kriterije: starost (od 20 do 35, nad 35), spol, izobrazba (osnovna ali višja) in formalnost govora (formalni – predvsem monolog, neformalni – predvsem dialog).

Enega največjih in pravkar začetih projektov gradnje govornega korpusa predstavlja nacionalni **poljski** korpus (National Corpus of Polish – NKJP; Przepiorkowski et al., 2008), ki bo vključeval govorni podkorpus v obsegu 30 mio. besed (javni govori, parlamentarne debate, televizijske oddaje, pogovorne oddaje, radijski intervjuji, dnevnoinformativne oddaje, 3 mio. besed pa naj bi obseg podkorpus vsakdanjih pogovorov).

Švedski govorni korpus (Goeteborg Spoken Language Corpus – GSLC; http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3) obsega 1,4 mio. besed. Sestavljajo ga samo posnetki spontanega govora. Bolj kot demografska uravnoteženost jih je zanimala pokritost čim večjega spektra različnih besedil, tako da vključuje več kot 25 družbenih aktivnosti.

[1] Prim. npr. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/index2.html.

Nizozemski govorni korpus (Spoken Dutch Corpus/Corpus Gesproken Nederlands – CGN) obsega skoraj 9 mio. besed. Natančni podatki o zasnovi korpusa so dostopni na http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm. V prepleteni strukturi ločijo 14 kategorij: monolog – multilog/dialog, javno – zasebno, javno nadalje na radio in televizijo, brano – spontano, formalno – neformalno in na koncu besedilne tipe: pogovor, intervju, telefonski pogovor, poslovni pogovor, intervju in diskusija, diskusija/debata/srečanje, predavanje, opisi slik, spontan komentar, poročilo in dnevnoinformativni program, novice, komentar, govori, brano besedilo. Pri tistih kategorijah, kjer je bilo smiselno, so upoštevali tudi demografske značilnosti (spol, starost, regijski izvor, socialno-ekonomski status) kot kriterij za vzorčenje.

Korpus govorjene **estonščine** (Hennoste et al., 2008) obsega okoli 1 mio. besed. Kriteriji za zajem gradiva so: družbena in narečna pripadnost govorcev, dialog – monolog, stopnja spontanosti, prenosnik (osebni stik, telefon, množični medij) in stopnja formalnosti s štirimi podkategorijami: razmerje med sogovorniki (poznani, nepoznani), vloga sogovornikov (privatna oseba, predstavnik institucije), scena (zasebni prostor, uradni prostor), namen interakcije (udeležba, informiranje). Vsebina korpusa so tako predvsem institucionalni ali delno institucionalni pogovori, manj pa zasebni pogovori.

Korpus govorjene **italijanščine** (CLIPS; http://www.alphabit.net/Glottophilia/2007/02/clips-corpus-of-spoken-italian.html) obsega 100 ur govora, kar bi znašalo okoli 900.000 besed. Posnetki so narejeni v 15 mestnih regijskih središčih po Italiji. Za vsako od teh središč so zajeti mediji (radio in televizija), dialogi, brani govor profesionalnih in neprofesionalnih govorcev in telefonski pogovori s simulirano hotelsko recepcijo. Eden večjih korpusov za italijanščino je tudi LABLITA (http://lablita.dit.unifi.it/corpora/descriptions/lablita/), ki je razdeljen v dva podkorpusa: govor odraslih govorcev v obsegu 640.000 besed vključuje pogovore v osebnem stiku, govor prek masovnih medijev in govor prek telefona. Pogovori v osebnem stiku so nadalje uravnoteženi glede na družbeni kontekst (družina, privatno, javno), vrsto interakcije (vodena, prosta) in strukturo komunikacijskega dogodka (monolog, dialog, pogovor). Drugi podkoprus LABLITA vključuje govor zgodnjega učenja italijanščine (pri otrocih v starosti od 15 do 36 mesecev).

Referenčni korpus sodobne **portugalščine** (Reference Corpus of Contemporary Portugeese – CRPC; http://www.clul.ul.pt/english/sectores/linguistica_de_corpus/projecto_crpc.php) vključuje govorni podkorpus približno 2,5 mio. besed. Pokriva portugalščino, kot je v rabi v vseh portugalsko govorečih deželah po svetu.

1.2 Pregled obstoječih slovenskih korpusov in publikacij

Za slovenski jezik obstaja nekaj srednje velikih in manjših specializiranih govornih korpusov, ki zajemajo:

- televizijske dnevnoinformativne in pogovorne oddaje
  - BNSI Broadcast News, 72 ur (Žgank et al., 2004)
  - Broadcast News Speech Database, 34 ur (Žibert, Mihelič, 2004)
- parlamentarni govor:
  - določen delež korpusa FidaPlus predstavljajo transkripcije razprav iz državnega zbora, ki pa so prilagojene drugim, neraziskovalnim namenom in niso vedno dobesedne; obsegajo ca. 2 mio. besed

    - baza Sloparl vključuje prav tako transkripcije razprav iz državnega zbora, ki pa se trenutno urejajo v raziskovalne namene; obsega ca. 100 ur govora (Žgank et al., 2006)
- telefonske pogovore s turističnimi agencijami, hotelsko recepcijo in turistično pisarno – Turdis, obseg ca. 30.000 besed (Verdonik, Rojc, 2006)
- govorjene diskurze različnih tipov, osnutek referenčnega korpusa, obsega ca. 15.000 besed (Zemljarič Miklavčič, Stabej, 2005; Zemljarič Miklavčič, 2006)

Poleg navedenih zasledimo še nekaj drugih govornih baz, ki pa ne vključujejo avtentičnih govorjenih diskurzov, in nekaj transkripcij sicer avtentičnih govorjenih diskurzov, nastalih predvsem v okviru različnih diplomskih, magistrskih in doktorskih projektov, ki pa niso urejene do te mere, da bi omogočale avtomatsko iskanje brez dodatnega urejanja.

Od publikacij je za definiranje specifikacij za gradnjo referenčnega govornega korpusa slovenskega jezika relevantna predvsem disertacija, katere cilj je bil definirati načela za gradnjo govornega korpusa slovenščine (Zemljarič Miklavčič, 2007), in iz nje izhajajoče publikacije (Zemljarič Miklavčič, 2006; Zemljarič Miklavčič, Stabej, 2005), prvi grob osnutek korpusa govorjenih besedil pa je bil predstavljen že v (Stabej, Vitez, 2000).

2 CILJI

Govorni korpus (GK) je v splošnem namenjen korpusnim raziskavam govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah. Naravnan je v kar se da referenčni zajem govorjenih diskurzov, vendar je pri tem omejen z obsegom ca. 1 mio. besed.

GK bo predstavljal govorni podkorpus referenčnega korpusa slovenskega jezika v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ). GK bo v projektu SSJ govorni vir za:

- izdelavo leksikalne podatkovne baze s podatki o frekvenci, pomenski strukturi in z zgledi rabe,
- izdelavo pedagoške korpusne slovnice.

Sekundarno bo GK omogočal tudi druge korpusne raziskave, zlasti v leksikografiji, skladnji, analizi diskurza/konverzacijski analizi in govornih tehnologijah, potencialno pa tudi v sociolingvistiki in tradicionalnem opisnem jezikoslovju.

Za slovenski jezik je značilna precejšnja razlika med pisno in govorjeno rabo jezika. Pri tem je govorjena raba jezika, zlasti pa spontana govorjena raba jezika, bistveno manj raziskana kot pisna raba jezika. GK je nujni začetni korak k boljšemu poznavanju spontane govorjene rabe jezika.

GK sledi trem osnovnim ciljem:

1. **Zajeti vzorčne primere različnih govornih situacij in različnih govorjenih besedil**.
2. **Zajeti govorjeni diskurz demografsko reprezentativnega vzorca govorcev slovenskega jezika.**
3. **Zajeti predvsem tiste govorne situacije, v katerih so uporabniki jezika najbolj pogosto produktivno-receptivno udeleženi.**

V GK želimo kolikor mogoče ohraniti avtentično dolžino govorjenih diskurzov.

V GK želimo zajeti predvsem avtentične govorjene diskurze, tj. diskurze, ki potekajo v naravnem okolju in niso zrežirani ali umetno sproženi. Ta kriterij je pomembnejši kot popolna demografska in besedilnovrstna uravnoteženost.

2.1 Pravno-etični vidiki zajemanja gradiva

S pravnega in etičnega vidika je zajemanje gradiva za govorni korpus omejeno predvsem zato, ker je treba zagotoviti predhodno seznanitev in soglasje vseh govorcev, katerih govor bo posnet za namene GK. Izkušnje pri tem kažejo, da vsi govorci niso vedno pripravljeni za takšno sodelovanje, še pomembneje pa je, da se govorci praviloma vsaj v začetku vedejo drugače kot sicer, če vedo, da se njihov govor snema.

2.2 Tehnični vidiki zajemanja gradiva

Za zajemanje gradiva je potrebno ustrezno namestiti in aktivirati snemalno tehnično opremo (mikrofone, nosilce podatkov ipd.). Te aktivnosti lahko zmotijo običajen potek diskurza. Nadalje tudi s tem, ko namestimo mikrofon, govorce spomnimo, da bo njihov govor posnet, kar lahko vpliva na njihovo vedenje. Ker je za natančno transkripcijo

potreben kvaliteten zvok, pa je s tehničnih vidikov zelo težavno in vprašljivo snemanje v zelo šumnih okoljih ali v skupinah, kjer je prisotnih zelo veliko govorcev.

2.3 Opredelitev glede spontanosti govora

Govorne situacije, v katerih je govorjeni jezik prevladujoče le govorna predstavitev pisnega besedila (brano ali naučeno besedilo), v GK ne bodo zajete. Čeprav lahko pričakujemo, da je tudi brani diskurz različen od pisnega diskurza (tj. predpostavljamo, da pisci besedila, ki bodo posredovana prek govornega prenosnika, oblikujejo nekoliko drugače kot besedila, ki bodo posredovana prek pisnega prenosnika), pa je ta razlika bistveno manjša kot pri spontanem diskurzu, poleg tega že obstaja nekaj korpusov, ki beležijo predvsem brani ali delno spontani govorjeni diskurz v slovenskem jeziku (npr. diskurz po televiziji: BNSI Broadcast News, diskurz v parlamentu: Sloparl). Ker je GK po obsegu zelo omejen, doslejšnji korpusni viri za slovenščino pa so najskromnejši ravno na področju spontanega govorjenega diskurza, je **kriterij za zajem v GK prevladujoča vsaj delna ali popolna spontanost**, tj. da govorec vsaj delno sproti tvori besedilo (npr. ima vnaprej pripravljene samo začetne izjave, je bran samo del diskurza) oz. da v celoti sproti tvori besedilo (kot je značilno za zasebni diskurz).

2.4 Opredelitev glede govora mladoletnikov

Govor otrok in mladostnikov se pogosto zbira ločeno od govora odraslih govorcev. Toda pomemben del govorne komunikacije odraslih je tudi govor z otroki in mladostniki (zlasti v okviru družine). Ker bo GK tudi jezikovni vir za izdelavo pedagoške slovnice, pa bo pomemben del korpusa zajemal tudi šolski diskurz za otroke od 10. leta starosti naprej. Govorci, mlajši od 18 let, tako ne bodo popolnoma izključeni iz GK.

2.5 Opredelitev glede prvega jezika govorcev

V GK bo zajet tudi določen delež govorcev, za katere slovenščina ni prvi oz. materni jezik, pri čemer bomo skušali upoštevati realno demografsko sestavo govorcev slovenščine.

2.6 Opredelitev glede govora Slovencev v zamejstvu in po svetu

V GK bo zajet tudi določen delež govorcev, ki živijo v zamejstvu. Govor Slovencev, ki živijo po svetu, v prvi fazi ne bo vključen v GK.

3 METODA

Iz pregleda tujih podobnih projektov vidimo, da se avtorji odločajo za različne pristope k definiranju kriterijev: nekateri poudarjajo predvsem lastnosti govorcev, kot so spol, starost, izobrazba ipd., kar imenujemo s skupnim terminom demografski kriteriji (npr. češki korpus); drugi poudarjajo zlasti različnost zajetih besedil in pri tem upoštevajo spontanost, število udeležencev, formalnost, tipologijo govorjenih besedil itd., kar bomo imenovali s skupnim terminom besedilnovrstni kriteriji (enako Zemljarič Miklavčič, 2007), ta pristop je uporabljen npr. pri švedskem korpusu. Večina projektov pa kombinira oba pristopa na različne načine: v dveh ločenih podkorpusih (BNC), izhajajo iz besedilnovrstnih kriterijev in vključujejo demografske, kjer je to smiselno (nizozemski korpus), izhajajo iz regijske razdeljenosti in nato vključujejo besedilnovrstne kriterije (italijanski korpus).

Iz zastavljenih ciljev GK sledi, da je ustrezna metoda tista, pri kateri kombiniramo demografske in besedilnovrstne kriterije.

Demografski kriteriji so spol, starost, izobrazba, regijska pripadnost, socialni status, višina prihodkov ipd., v zvezi z jezikom tudi prvi in neprvi jezik.

Besedilnovrstni kriteriji so javnost/nejavnost govora, formalnost/neformalnost govora, število sogovornikov v diskurzu, prenosnik, namen diskurza, tematika ipd.

Demografski in besedilnovrstni kriteriji za zajem v korpus bodo definirani tako, da upoštevajo zastavljene cilje GK, tuje izkušnje na tem področju, specifike slovenskega okolja, tehnične in pravne vidike zajemanja gradiva, potrebe korpusnih raziskav in statistično reprezentativnost posameznih skupin glede na predvideni obseg korpusa. Demografski kriteriji za zajem bodo uravnoteženi na podlagi najnovejših podatkov Statističnega urada Republike Slovenije ter v skladu z definiranimi cilji GK. Besedilnovrstni kriteriji bodo uravnoteženi tako, da bo bodo ustrezno zastopani glede na definirane cilje GK.

4 KRITERIJI ZA ZAJEM GRADIVA

4.1 Demografski kriteriji

Demografski kriteriji so: spol, starost, dosežena 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 izvora, ker je (z današnjega stališča) za slovenske razmere težko določljiv in vprašljiv.

Demografske kriterije bi vsekakor lahko še bistveno bolj natančno specificirali, vendar korpusno zbiranje gradiva zahteva čim bolj robustno in enostavno shemo, zato se omejujemo le na najbolj bistvene dejavnike.

Demografski kriteriji so uravnoteženi glede na najnovejše podatke SURS in v skladu s cilji GK ter za večje skupine zaokroženi na deset odstotkov. Bolj podrobna uravnotežitev ni smiselna, saj je pri snemanju v avtentičnem okolju demografske kriterije najtežje nadzorovati in lahko pričakujemo tudi večja odstopanja od zastavljenih razmerij.

Demografski kriteriji in izhodiščna razmerja v odstotkih:

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 govor:
   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 nezasebni govor:
a. JZ Slovenija: 60%
b. SV Slovenija: 40%

4.2 Besedilnovrstni kriteriji

Kriteriji, na podlagi katerih klasificiramo govorjene diskurze v tipe, so lahko zelo različni (npr. klasifikacija na podlagi namernosti, prenosnika, funkcije, strukture diskurza, tematike, socialne zvrstnosti itd.). Večinoma se avtorji pri definiranju besedilnovrstnih kriterijev za zajem gradiva v reprezentativni govorni korpus odločajo za kombinacijo več kriterijev. Zemljarič Miklavčič (2007: 96) za slovenščino predlaga naslednje besedilnovrstne kriterije: struktura besedila (monolog, dialog/multilog), okoliščine (javna besedila, zasebna besedila), govorni položaj (formalni, neformalni), prenosnik (osebni stik, telefon, avdio, video).

Definiranje besedilnovrstnih kriterijev je težavnejše kot definiranje demografskih kriterijev, saj pogosto govorjenim diskurzom ne moremo nedvoumno določiti neke lastnosti (npr. odločitev, ali je neki diskurz spontan ali nespontan, formalen ali neformalen) oziroma se lahko v istem diskurzu meša več lastnosti (npr. prevladuje monolog, ki pa občasno preide v dialog), tako da je lahko smiselnost posameznih faktorjev kot kriterijev za zajem celo vprašljiva. Z upoštevanjem zastavljenih ciljev GK, tujih izkušenj, specifik slovenskega okolja, potreb korpusnih raziskav in statistične reprezentativnosti posameznih skupin glede na predviden obseg korpusa ugotavljamo naslednje:

Cilj, da zajamemo predvsem spontane ali delno spontane govorne situacije, pomembno vpliva na strukturo zajetih besedil: monologi so namreč praviloma vnaprej pripravljena, brana besedila v javnem diskurzu. Večina diskurzov v GK bo torej dialoških ali multiloških, zato ni smiselno, da faktor monolog/dialog uvrstimo med kriterije, ki naj zagotovijo uravnoteženi zajem besedil.

Ker je cilj GK, da zajame predvsem spontana besedila in le v manjši meri pripravljena ali delno brana, tudi stopnje spontanosti ni smiselno posebej izpostavljati kot kriterij za zajem. Poleg tega je lahko stopnja spontanosti zelo težko določljiva, zlasti v javnem govoru, in lahko o njej sklepamo samo iz besedila.

Formalnost kot kriterij se nanaša predvsem na izbiro jezikovnih sredstev in za posamezne tipe govornih situacij je zelo težko vnaprej predvidevati te izbire, sploh ker na tem področju še ni bilo veliko raziskav v slovenskem jeziku. Zato tudi formalnosti ne uvrstimo med besedilnovrstne kriterije.

Namen in tematika besedil sta bila pretežno klasificirana ali za tuje kulturno okolje ali za pisni diskurz in sheme niso enostavno prenosljive na govorjeni diskurz v slovenskem jeziku.

Kot bistvena besedilnovrstna kriterija tako ostaneta javnost diskurza in prenosnik. Njuna podrobnejša klasifikacija in izhodiščna razmerja v odstotkih so naslednja:

1. javnost:
    a. javni diskurz 60%
        i. razvedrilni 33%
        ii. nerazvedrilni 67%
    b. nejavni diskurz
        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 tisti diskurz, ki je odprt za širšo javnost ali naslavlja veliko skupino ljudi, vsak drugi diskurz šteje za nejavni.

Nejavni diskurz nadalje ločimo na zasebni diskurz, tj. diskurz v zasebnem življenju posameznikov (v okviru družine, prijateljev, znancev). Nejavni nezasebni diskurz vključuje različne uradne in poluradne diskurze (v uradih, trgovinah, ob storitvah, v profesionalnem življenju ipd.).

V javnem diskurzu ločimo medijske vsebine razvedrilnega programa in drug javni diskurz, katerega namen je predvsem razvedrilen (razvedrilni javni diskurz), ter medijske vsebine informativnega/izobraževalnega/kulturnega programa in drug javni diskurz, katerega namen je predvsem informativen/izobraževalen/socialen (nerazvedrilni oz. informativno-izobraževalni javni diskurz).

4.2.1 Dodatni kriteriji za zajemanje pedagoškega diskurza

V skladu s cilji GK bo znaten del korpusa zajemal pedagoški diskurz v osnovni in srednji šoli. Ta del bo zasnovan kot podkorpus GK, ki sledi spodaj predstavljen dodatnim kriterijem za zajem. V besedilnovrstni klasifikaciji zgoraj je zajet v kategorijah javni nerazvedrilni oz. informativno-izobraževalni diskurz, prenosnik je osebni stik. Kriteriji so uravnani glede na podatke MŠŠ in SURS.

1. stopnja šolanja:
    a. osnovna šola 55%
        - 2. triletje 50%
        - 3. triletje 50%
    b. srednja šola 45%
        - gimnazije 40%
        - nižje in srednje poklicno, srednje strokovno in poklicno-tehniško izobraževanje 60%
2. regija:
    a. JZ 60%
    b. SV 40%

3. učni predmet:
    a. naravoslovni in tehnični predmeti 50%
    b. družboslovni in humanistični predmeti 50%

4.3 Shema pokrivanja definiranih besedilnovrstnih in demografskih kriterijev

Za celoten korpus:

OKOLIŠČINE	%	NAMEN	%	PRENOSNIK	%	REGIJA	%
javni diskurz	60%	informativno-izobraževalni	40%	tv[2]	10%	SV	4%
						JZ	6%
				radio[3]	10%	SV	4%
						JZ	6%
				osebni stik	20%	SV	8%
						JZ	12%
		razvedrilni	20%	tv[2]	10%	SV	4%
						JZ	6%
				radio[3]	10%	SV	4%
						JZ	6%
							60,00%
nejavni diskurz	40%	nezasebni	15%	telefon	5%	SV	2,00%
						JZ	3,00%
				osebni stik	10%	SV	4,00%
						JZ	6,00%
		zasebni	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%**

[1] Se uravnoteži glede na spol, starost in izobrazbo v skladu z demografskimi kriteriji, definiranimi v 4.1.
[2] Seznam televizijskih vsebin, ki jih bomo skušali zajeti v GK, je natančneje opredeljen v prilogi 1 tega dokumenta in zajema najbolj gledane televizijske programe in oddaje.
[3] Seznam radijskih vsebin, ki jih bomo skušali zajeti v GK, je natančneje opredeljen v prilogi 2 tega dokumenta in zajema najbolj poslušane radijske postaje in vsebine po posameznih slovenskih regijah.

Za podkorpus šolskega diskurza znotraj GK (javni informativno-izobraževalni diskurz, osebni stik, obsega 15% celotnega korpusa):

STOPNJA ŠOLANJA I	%	STOPNJA ŠOLANJA II	%	REGIJA	%	PREDMET	%
osnovna šola	8%	2. triletje	4%	JZ	2,4%	družb.	1,2%
						nar.	1,2%
				SV	1,6%	družb.	0,8%
						nar.	0,8%
		3. triletje	4%	JZ	2,4%	družb.	1,2%

						nar.	1,2%
				SV	1,6%	družb.	0,8%
						nar.	0,8%
							8,00%
srednja šola	7%	gimnazija	3%	JZ	1,8%	družb.	0,9%
						nar.	0,9%
				SV	1,2%	družb.	0,6%
						nar.	0,6%
		poklic.-strok. šola	4%	JZ	2,4%	družb.	1,2%
						nar.	1,2%
				SV	1,6%	družb.	0,8%
						nar.	0,8%
					15,00%		**7,00%**

4.4 Toleranca odstopanj

V poglavju 4.1 in 4.2 navedena razmerja med demografskimi in besedilnovrstnimi kriteriji so »idealna« in pričakujemo, da bodo dejanska razmerja med njimi v GK nekoliko drugačna, saj teh kriterijev pri snemanju avtentičnih diskurzov (kar je eden osrednjih ciljev GK in pomembnejša zahteva kot popolna demografska in besedilnovrstna uravnotežitev) ni mogoče popolnoma nadzorovati.

Zato štejemo, da GK izpolnjuje kriterije teh specifikacij, če so le-ti uresničeni v okviru 30-odstotnega odstopanja relativno, ter zelo dobro izpolnjuje kriterije teh specifikacij, če so uresničeni v okviru 10-odstotnega odstopanja relativno.

Prav tako štejemo, da GK izpolnjuje kriterije teh specifikacij, če samo določen del gradiva vključuje podatke, na podlagi katerih je mogoče preveriti pokritost definiranih kriterijev, in ta del GK izkazuje ustrezno zastopanost (glej prejšnji odstavek) definiranih kriterijev.

5 PODATKI O ZAJETEM GRADIVU

Ob zajemanju gradiva se bodo zajemali naslednji podatki o govorcih in diskurzih:

5.1 Podatki o govorcih

1. Spol: a) m b) ž

2. Starost: a) do 10 b) 10 do 14 c) 15 do 18 d) 19 do 24 e) 25 do 34 f) 35 do 59 g) nad 60

3. Regionalna pripadnost:
a) MS b) MB c) SG d) CE e) NM f) KK g) LJ h) KR i) GO j) PO k) KP
l) Italija m) Avstrija n) Madžarska o) tujec p) mešano r) nedoločno

4. Izobrazba: a) osnovna šola ali manj b) srednja šola c) višja ali visoka šola e) fakulteta ali več

5. Prvi jezik:
a) slovenščina b) angleščina c) nemščina d) italijanščina
e) madžarščina f) južnoslovanski jeziki (brez slovenščine) g) albanščina
i) drugi: slovanski germanski romanski neindoevropski jeziki

5.2 Podatki o diskurzih

1. Vrsta institucije, v okviru katere je potekal diskurz

2. Opis govornega dogodka

3. Kraj odvijanja/predvajanja diskurza

4. Čas odvijanja/predvajanja diskurza

5. Število aktivnih udeležencev

5.4 Podatki o snemanju

1. Kraj in čas snemanja
2. Snemalec

6 VAROVANJE OSEBNIH PODATKOV

Po Zakonu o varstvu osebnih podatkov (ZVOP-1, UL RS 86/4) je osebni podatek katerikoli podatek, ki se nanaša na posameznika, ne glede na obliko, v kateri je izražen. Ob zajemanju gradiva za GK bodo s posebnim obrazcem zbrani tudi osebni podatki govorcev, ki bodo prispevali v GK, in sicer spol, starost, izobrazba, prvi jezik, kraj in ulica stalnega bivališča (brez hišne številke), kraj šolanja in kraj zaposlitve. Od tega je podatek o prvem jeziku po ZVOP-1 občutljivi osebni podatek. Vsi podatki bodo v zbirko osebnih podatkov v GK vključeni tako, da bodo anonimizirani, kar pomeni, da jih ne bo več mogoče povezati s posameznikom.

Vsi posamezniki, ki bodo imeli stik z osebnimi podatki govorcev in bodo delali z gradivom, bodo podpisali izjavo o varovanju osebnih podatkov in gradiva skladno z ZVOP-1, upravljavec teh podatkov pa bo ravnal v skladu z določili o zavarovanju osebnih podatkov po ZVOP-1.

V zvezi z zajemanjem, varstvom in zavarovanjem osebnih podatkov ter za samo privolitev govorcev v snemanje njihovih diskurzov bo pripravljen obrazec s pisnim privoljenjem, ki ga bodo vsi govorci, udeleženci posnetih govornih dogodkov, podpisali pred snemanjem, v nasprotnem primeru snemanje ne bo izvedeno. Govorjeni diskurzi v medijih (radio, TV) bodo predvidoma zajeti iz arhivov medijskih hiš, pri čemer bo podpisana pogodba, ki bo določala obveznosti obeh pogodbenih strank.

7 LITERATURA

Hennoste, T., Gerassimenko, O., Kasterpalu, R., Koit, M., Raabis, A., Strandson, K., 2008: From human communication to intelligent user interfaces: corpora od spoken Estonian. V: Proceedings of 6th Language Resources and Evaluation Conference, Maroko, Marakeš.

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

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

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

Zemljarič Miklavčič, J., Stabej, M., 2005: Building a pilot spoken corpus. V: Garabik, R. (ur.): Computer Treatment of Slavic and East European Languages. Slovaška, Bratislava. 229-240.

Zemljarič Miklavčič, J., 2006: Korpus govorjene slovenščine. V: Erjavec, T., Žganec Gros, J. (ur.). Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006. Ljubljana: Institut Jožef Stefan. 124-127.

Zemljarič Miklavčič, J., 2007: Načela oblikovanja govornega korpusa za slovenščino. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Žgank, A., T. Rotovnik, D. Verdonik, Z. Kačič, 2004: Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. V: 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. V: 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. V: Informacijska družba IS'2004: Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan. 94-97.

Priloga 1: Okvirni seznam televizijskih programov in oddaj za zajem v GK

TV Slovenija
Informativne oddaje:
- Dnevnik
- Odmevi
- Tednik
- Polemika

Zabavne/razvedrilne oddaje:
- Piramida
- NLP s Tjašo Železnik in Klemnom Slakonjo
- športni prenos

POP TV
Informativne oddaje:
- 24ur
- Trenja
- Preverjeno

Zabavne/razvedrilne oddaje:
- As ti tud not padu?!
- Kmetija
- Vzemi ali pusti
- Desetka

Priloga 2: Okvirni seznam radijskih postaj in vsebin za zajem v GK

Val 202
Informativne vsebine:
- Aktualna tema
- Vroči mikrofon

Zabavne/razvedrilne vsebine:
- Dobro jutro
- Izbor popevke tedna

Slovenija 1
Informativne vsebine:
- Studio ob 17h
- Intervju

Radio Ognjišče
Informativne vsebine:
- Naš gost
- Aktualni pogovori

Radio Center
Zabavne/razvedrilne vsebine:
- jutranji program: Polona Požgan in Sašo Papp

Radio Maribor
Informativne vsebine:
- Radijska tribuna

Radio City Maribor
Zabavne/razvedrilne vsebine:
- jutranji program

Radio Murski val
Informativne vsebine:
- Za zdravje
- Intervju

Zabavne/razvedrilne vsebine:
- Geza se zeza

Radio Štajerski val
Informativne vsebine:
- tedenski intervju

Koroški radio
Informativne vsebine:
- Sobotni gost
- Zdravje za življenje

Zabavne/razvedrilne vsebine:
- Šalter

Radio Alfa

Zabavne/razvedrilne vsebine:
- Odbilo je enaindvajseto

Radio Hit
Informativne vsebine:
- Na obisku

Zabavne/razvedrilne vsebine:
- Hitova budilka

Radio Aktual
Informativne vsebine:
- Na sredi z Majdo Juvan

Zabavne/razvedrilne vsebine:
- Malo morgen, Racman

Radio Kranj
Informativne vsebine:
- Se res poznamo?

Radio Belvi
Zabavne/razvedrilne vsebine:
- Belvijevo jutro

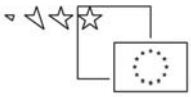
Radio Koper
Informativne vsebine:
- Aktualno
- Portret

Radio Capris
Zabavne/razvedrilne vsebine:
- Kikiriki

Radio Sraka
Informativne vsebine:
- sobota dopoldne

Radio Krka
Zabavne/razvedrilne vsebine:
- nagradne igre

REPUBLIKA SLOVENIJA
MINISTRSTVO ZA ŠOLSTVO IN ŠPORT



Naložba v vašo prihodnost
OPERACIJO DELNO FINANCIRA EVROPSKA UNIJA
Evropski socialni sklad

SPORAZUMEVANJE V SLOVENSKEM JEZIKU – GOVORNI KORPUS

Vrsta dokumenta: Specifikacije transkribiranja gradiva

Datum: 31. marec 2009
Verzija: 1.0

KAZALO

1 UVOD

Namen tega dokumenta je izbor orodja za transkribiranje in definiranje pravil za transkribiranje zvočnega gradiva za govorni korpus (GK). Dokument je vsebinsko nadaljevanje specifikacij za zbiranje gradiva za GK in specificira drugo fazo izgradnje GK.

1.1 Obstoječi mednarodni standardi in prakse

1.1.1 EAGLES

Evropska iniciativa EAGLES (Expert Advisory Group on Language Engineering Standards; http://www.ilc.cnr.it/EAGLES/home.html) je nastala leta 1993 na pobudo Evropske komisije z namenom, da pospeši oblikovanje skupnih standardov za izdelavo obsežnih jezikovnih virov. Njihova priporočila so zato za gradnjo GK nadvse relevantna.

Pri izdelavi priporočil za govorjena besedila (EAGLES, 1996) upoštevajo obstoječe prakse (priporočila TEI, NERC, SpeechDat, tradicijo konverzacijske analize idr.) in skušajo najti skupne elemente različnih tradicij transkribiranja.

V povzetku priporočajo označevanje naslednjih elementov: glasovnih polleksikalnih (eee, mhm, aha itd.) in neleksikalnih (smeh, cmokanje, kašljanje, kihanje, jok, zehanje itd.) enot, negovornih nekomunikacijskih dogodkov, identitete govorca, menjavanja govorcev, hkratnega govora, izpustitev pri branih besedilih, samopopravljanj, besednih fragmentov in nerazumljivih fragmentov.

Pri zapisu govora ločujejo tri ravni:
S1 – ortografska predstavitev besedila
S2 – fonemska predstavitev besed v citatni obliki (tj. v obliki, kot so besede izgovorjene v izolaciji)
S3 – fonetična transkripcija, ki predstavlja dejansko glasovno podobo izjave

Za ortografski zapis priporočajo, da je narejen v standardni (tj. knjižni) normi in da so vse enote (tudi okrajšave, številke, črkovanja ipd.) polno izpisane. To priporočilo temelji na predpostavki, da obstaja možnost avtomatske povezave med ortografskim (S1) in fonemskim (S2) zapisom.

Za fonemski in fonetični zapis priporočajo uporabo fonetične abecede SAMPA oz. X-SAMPA.

Na prozodični ravni priporočajo vsaj označevanje izjav in premorov.

1.1.2 TEI

TEI (Text Encoding Iniciative; http://www.tei-c.org/index.xml) je organizacija, katere namen je definiranje standardov za kodiranje besedil. Njihova priporočila (TEI, 2007; http://www.tei-c.org/Guidelines/index.xml) v posebnem (osmem) poglavju obsežno obravnavajo tudi transkribiranje govora. Priporočila EAGLES-a že upoštevajo bistvena priporočila TEI.

Glede samega kodiranja transkripcij pa je pomembno izpostaviti, da so standardi TEI izraženi v danes široko uporabljanem kodifikacijskem jeziku XML, čeprav sicer kodna shema TEI ni odvisna od tega jezika.

1.2 Dosedanje prakse pri gradnji slovenskih govornih korpusov

V slovenskem prostoru zasledimo tri usmeritve bistveno različnih praks transkribiranja daljših besedil:

1. V slovenski dialektologiji se običajno uporablja t. i. tradicionalna slovenska fonetična transkripcija oz. fonetična transkripcija OLA (Slovanski lingvistični atlas) z dodanimi različnimi diakritičnimi znaki, npr. krožec pod zvočnikom za silabem, pika pod e ali o za ozki izgovor vokala, strešica pod e ali o za nevtralni izgovor, e in o brez diakritičnega znamenja za široka vokala, dvopičje za črko za dolgi vokal itd. (Zorko, 1995).

2. V besediloslovnih in pragmatičnih raziskavah govorjene slovenščine se večinoma uporablja ortografski zapis pogovornega jezika. Skupnega standarda pri tem sicer še ni, na podlagi zapisov govorjenih besedil in besedilnih fragmentov nekaterih avtorjev transkripcij (npr. Kranjc, 1999; Krajnc, 2005; Smolej, 2006) pa lahko ugotavljamo naslednji temeljni skupni značilnosti:
- Uporabljen je slovenski knjižni črkopis brez dodatnih posebnih znakov za različne glasove (npr. za polglasnik, dvoustnični u, diftonge ipd.). Izjemoma najdemo dodan poseben znak za nekatere rabe polglasnika (npr. *sevðda*, Krajnc, 2005).
- Zapis ponazarja pojave moderne vokalne redukcije in drugih pogovornih in narečnih prvin govorjene slovenščine (npr. *maš čevle, to je zloml, jes mam, notr držim, hitr kuple, beu avto, neki, k bi lohk, je poneso, ceno svojih življen, tko je blo, maš polhen kufer, mislm*). Seveda se pri različnih avtorjih pojavljajo tudi nekoliko različne rešitve pri zapisu nekaterih pogovornih oblik (npr. *jes* proti *js* za *jaz*, *neč* proti *nč* za *nič*, različni zapisi polglasnika v vlogi diskurznega označevalca: *ee, {eee}, ə* itd.).

Podoben zapis pogovornega jezika se zadnji čas vse širše uporabljajo v spletnih forumih, klepetalnicah, blogih in drugih zapisih, ki jih uporabniki objavljajo v spletu, kot tudi v nekaterih literarnih delih. Tako npr. pišejo *čeprov, niti najmnj, se nej, pusti pr mer, nč ne rečm, tut js, kšni prjatlci, s kero kol, pejt spat, js sm zlo, odpravt…* (http://www.cveka.com, 22. 12. 2008) oz. *najraj b vidu, če bse mi, čeb šu sam pa bmene pustu, tud jest, čeu pa ta, mi nau hotu povedat, ker neb mel*... (Welsh, 1997); *razmete, rejsan, nej pred sebo, pred bougon, san ges živa, tou ka znam, pejneze, fkraj, z menof...* (Šarotar, 2007).

3. V jezikovnotehnološki praksi (Žgank et al., 2004; 2006; Zemljarič Miklavčič, 2007) se uporablja v glavnem pokrnjižen zapis govorjenega jezika, iz katerega niso več vidni pogovorni in narečni pojavi, kot je npr. moderna vokalna redukcija. Dopuščeno je le omejeno število določenih najpogostejših odstopanj, npr. kratki nedoločnik, *k* za veznike *ki, ko, ker*, *pol* v pomenu *potem* itd. V tej smeri je nastal tudi poskus definiranja pravil za transkribiranje govornih korpusov (Zemljarič Miklavčič, 2007). Verdonik (2006) pa poleg pokrnjiženega zapisa dodaja fonetični zapis v abecedi SAMPA, vendar samo v primerih, ko določena besedna oblika ne sovpada s knjižno normo.

O tem, katere vrste podatkov in kako podrobne podatke o samem diskurzu in govoru vključujejo transkripcije, so se avtorji odločali različno, glede na cilje in potrebe posameznih transkripcij.

2 ORODJE ZA TRANSKRIBIRANJE

Orodje za transkribiranje mora za namene izgradnje GK izpolnjevati naslednje zahteve:

1. je uporabniško prijazno za transkribiranje
2. podpira dolge zvočne datoteke
3. podpira šumnike
4. omogoča segmentacijo zvočnega signala na poljubne enote na enostaven način
5. informacije o povezavi med posameznimi segmenti in pripadajočimi deli transkripcije lahko vnašamo na enostaven način
6. podpira vnos metaoznak za opis različnih pragmatičnih, akustičnih in drugih dogodkov
7. podpira vnos podatkov o govorcih, zajetih s popisnim obrazcem
8. podpira vnos besedilnovrstnih podatkov o diskurzu

2.1 Primerjave orodij za transkribiranje v literaturi

Obstaja več orodij, veliko tudi prosto dostopnih, ki jih lahko uporabimo za transkribiranje zvočnih posnetkov, vendar nobeno ni bilo narejeno specifično za namene, kot jih imamo pri gradnji GK. V literaturi najdemo tudi nekaj primerjav in vrednotenj različnih orodij.

Garg et al. (2004) med drugim primerjajo naslednja orodja za transkribiranje: Praat, Transcriber, TASX in Anvil. Zaključijo, da je za splošno transkripcijo, kot jo bomo izvajali tudi pri GK, najprimernejši Transcriber.

Rohlfing et al. (2006) primerjajo prednosti in slabosti naslednjih orodij za multimodalno označevanje avdio in video posnetkov: Anvil, ELAN, Exmaralda, TASX in MacVista. V zaključku ugotavljajo, da je bilo vsako od naštetih orodij razvito v različne namene, zato je tudi izbor orodja odvisen predvsem od potreb. Od naštetih so za naše cilje potencialno primerni ELAN, Anvil in Exmaralda. Orodij, ki so izdelana samo za transkribiranje, ne vključijo v primerjavo.

Zemljarič Miklavčič (2007) podrobneje primerja orodja Transcriber, Praat in WinPitch. Opozarja na pomanjkljivost Transcriberja, da ne omogoča hkratnega zapisa več kot dveh govorcev – če torej hkrati govorijo trije ali več govorcev, lahko zapišemo govor samo dveh. Za Praat ugotavlja, da te pomanjkljivosti nima, hkrati pa navaja, da je po avtoričinih izkušnjah »v najboljšem primeru še mogoče transkribirati besedilo, ki ga hkrati izrečejo trije govorci; če govorijo več kot trije naenkrat, je običajno mogoče transkribirati samo posamezne fragmente iz posameznih izjav« (Zemljarič Miklavčič, 2007: 137). Tudi verzija WinPitchPro programa WinPitch omogoča transkribiranje, vendar v primerjavi s Transcriberjem in Praatom ni prosto dostopen. Avtorica sklene, da se za ortografsko transkribiranje zdita najprimernejša programa Transcriber in Praat.

2.2 Testiranje orodij za transkribiranje

V slovenskem prostoru že obstaja nekaj specializiranih govornih korpusov, za izdelavo katerih je bilo uporabljeno transkripcijsko orodje: v glavnem je bil to Transcriber (za baze BNSI Broadcast News (Žgank et al., 2004), Broadcast News Speech Database (Žibert, Mihelič, 2004), Turdis (Verdonik, Rojc, 2006), Sloparl (Žgank et al., 2006)), v

enem primeru pa Transcriber in Praat (Zemljarič Miklavčič, Stabej, 2005; Zemljarič Miklavčič, 2006).

Na podlagi tega dejstva in na podlagi informacij iz literature se odločimo, da bomo pred izborom praktično preskusili naslednja orodja: Transcriber, Praat, Exmaralda in ELAN.

Nobeno od njih ne izpolnjuje naših potreb v celoti.

ELAN: ELAN je namenjen predvsem označevanju zvočnih in video posnetkov, za uporabo pri transkribiranju pa ni preveč uporabniško prijazen, čeprav je tudi mogoče. Prav tako ni posebej uporabniško prijazen za določanje in spreminjanje mej segmentov. Vnašanje podatkov o govorcih in diskurzih je mogoče samo v obliki sledi. Po naši oceni je orodje bolj kot za samo transkribiranje govora primerno za označevanje pri gradnji multimodalnih govornih baz, kjer so označene tudi geste, mimika, kretnje ipd. Posebnost ELAN-a je, da omogoča med drugim uvoz datotek, zapisanih s programom Transcriber.

Exmaralda: to orodje edino podpira vnos poljubnih podatkov o govorcih in diskurzih, kar je gotovo njegova pozitivna lastnost za naše namene. Osrednja pomanjkljivost pa je slaba povezava med transkripcijo in zvočnim posnetkom: funkcije za pomikanje po zvočnem signalu so premalo natančne, kar lahko bistveno upočasni transkribiranje, vzpostavljanje povezave med posameznimi segmenti in pripadajočimi deli transkripcije pa je s tem orodjem zamudno. Zaradi tega Exmaralda ni posebej primerna za naše namene.

Praat: Praat je v osnovi namenjen akustičnim analizam, vendar omogoča tudi transkribiranje daljših zvočnih datotek. Njegova prednost v primerjavi s Transcriberjem je, da omogoča zapis hkratnega govora več kot dveh govorcev. Tudi predvajanje zvočnega posnetka in segmentiranje signala na manjše enote je enostavno. Glavna pomanjkljivost je, da zaradi številnih funkcij akustične analize ni najbolj uporabniško prijazen, če ga želimo uporabljati samo za namen transkribiranja. Prav tako ne podpira vnosa podatkov o govorcih in diskurzih ter vnosa metaoznak. V nasprotju z ostalimi programi Praatova izhodna datoteka ni v XML formatu, ampak v posebni Praatovi skripti (ki pa je sicer široko podprta v drugih transkripcijskih orodjih).

Transcriber: to orodje je za naše namene najbolj uporabniško prijazno in tudi po svoji zasnovi najbližje: narejeno je bilo namreč za namene transkribiranja televizijskih informativnih oddaj. Dodatna prednost je, da v slovenskem prostoru že obstaja praksa njegove uporabe. Povezava med zvočnim posnetkom in transkripcijo je izredno dobra, po zvočnih posnetkih se enostavno premikamo in posnetek enostavno segmentiramo. Orodje tudi podpira vnos metaoznak v transkripcijo, prav tako vnos nekaterih podatkov o govorcih in diskurzu, čeprav za naše namene nekoliko preveč omejeno. Njegova največja pomanjkljivost pa je, da ne omogoča zapisa hkratnega govora več kot dveh govorcev.

2.3 Izbor orodja za transkribiranje in kodni format

Potem ko smo pretehtali prednosti in slabosti testiranih orodij, smo se odločili, da naše cilje najbolje izpolnjuje **Transcriber,** s tem da bomo vanj vključili tudi slovenski črkovalnik. Za vnos podatkov o govorcih in diskurzu bomo uporabili Excel ali drug podoben program. Za drugi nivo zapisa govora bomo uporabili Open Office ali drug primeren urejevalnik besedil.

Delovni kodni format bo CP1250, končni kodni format pa UTF-8.

3 CILJI

Govorni korpus (GK) je v splošnem namenjen korpusnim raziskavam govorjene podobe slovenskega jezika v najrazličnejših govornih situacijah. Predstavljal bo govorni podkorpus referenčnega korpusa slovenskega jezika v okviru projekta Sporazumevanje v slovenskem jeziku (SSJ). GK bo v projektu SSJ govorni vir za:

- izdelavo leksikalne podatkovne baze s podatki o frekvenci, pomenski strukturi in z zgledi rabe,
- izdelavo pedagoške korpusne slovnice.

Sekundarno bo GK omogočal tudi druge korpusne raziskave, zlasti v leksikografiji, skladnji, analizi diskurza/konverzacijski analizi in govornih tehnologijah, potencialno pa tudi v sociolingvistiki in tradicionalnem opisnem jezikoslovju.

Zaradi obsega dela (1 mio. besed) in zaradi zagotavljanja homogenosti transkripcij bodo transkripcije kar se da enostavne in skrčene na zapisovanje za namene GK najnujnejših podatkov, in sicer:

1. transkripcije vključujejo najpomembnejše kontekstne informacije, zlasti vse tiste, ki so pridobljene ob samem zbiranju gradiva in so pomembne kot potencialni iskalni pogoj,
2. transkripcije vključujejo besedilnovrstno razvrstitev diskurzov, ki bo omogočala omejevanje iskanja na različne tipe govorjenih besedil,
3. transkripcije vsebujejo samo pragmatično[2] najpomembnejše informacije o strukturi diskurza,
4. sam zapis govora je tak, da omogoča čim hitrejše transkribiranje,
5. zapis govora kar najbolj nazorno predstavlja dejansko govorjeno podobo diskurza,
6. zapis govora omogoča avtomatsko iskanje po besednih oblikah z enako oblikoslovno in semantično vlogo, a različnimi glasovnimi podobami,
7. zapis govora vsebuje samo pragmatično najpomembnejše informacije o nebesednih in nejezikovnih zvokih, ki so pomembni za uporabnikovo boljše razumevanje poteka diskurza.

GK je mišljen predvsem kot jezikovni govorni vir, zato transkripcije ne bodo vključevale informacij o kretnjah, mimiki, gestah in drugih nezvočnih spremljevalnih kanalih govornega sporazumevanja.

[2] Za pragmatično pomembne štejemo tiste dogodke, ki imajo vidnejši učinek na diskurz, pri čemer mislimo tako jezikovno rabo kot socialna razmerja in kognitivno dimenzijo.

4 PRAVILA TRANSKRIBIRANJA

Pravila transkribiranja bodo v skladu z zastavljenimi cilji GK definirana na naslednjih ravneh:

1. podatki o govorcih
2. podatki o diskurzu
3. struktura diskurzov
4. zapis govora
5. nebesedni in nejezikovni zvoki ter prozodija

4.1 Podatki o govorcih

1. identifikacijska koda govorca

2. spol:
   - o moški
   - o ženski
   - o nedoločno

3. starost:
   - o do 10
   - o 10 do 14
   - o 15 do 18
   - o 19 do 24
   - o 25 do 34
   - o 35 do 59
   - o nad 60
   - o nedoločno (ko ni podatka)

4. regionalna pripadnost govorca glede na registrsko območje (če posamezen govorec zaradi daljšega bivanja na različnih območjih čuti pripadnost različnim regijam, označimo vse ustrezne regije):

SV:
   - o MS
   - o MB
   - o SG
   - o CE

JZ:
   - o NM
   - o KK
   - o LJ
   - o KR
   - o GO
   - o PO
   - o KP

ostale:
   - o Italija
   - o Avstrija
   - o Madžarska
   - o tujina

- nedoločno (ko ni podatka)

5. izobrazba:
   - OŠ ali manj
   - SŠ
   - višja ali visoka šola
   - fakulteta ali več
   - nedoločno (ko ni podatka)
6. prvi jezik:
   - slovenščina
   - angleščina
   - nemščina
   - italijanščina
   - madžarščina
   - južnoslovanski jeziki (brez slovenščine)
   - albanščina
   - drugi:
     - slovanski
     - germanski
     - romanski
     - neindoevropski
   - nedoločno

4.2 Podatki o diskurzu

1. identifikacijska koda diskurza

2. dolžina posnetka v minutah in sekundah

3. tip diskurza:
   a. javni informativno-izobraževalni
   b. javni razvedrilni
   c. nejavni nezasebni
   d. nejavni zasebni

4. vrsta situacije:
   a. *televizija*
   b. *radio*
   c. *osnovna šola*
   d. *srednja šola*
   e. *fakulteta*
   f. *telefon*
   g. *osebni stik*

5. opis diskurza:
   a. [ime televizijske hiše], [ime oddaje]
   b. [ime radijske postaje], [ime oddaje]
   c. [letnik], [vrsta predmeta] – za šolski in akademski diskurz
   d. [tip institucije, kjer poteka nejavni nezasebni diskurz], npr. nepremičninska agencija, turistična agencija, upravna enota, trgovina, podjetje itd.
   e. [tip interakcije, med katerimi poteka nejavni zasebni diskurz], npr. doma/družina, doma/prijatelji, delovno mesto/sodelavci itd.

6. regija (kjer poteka diskurz):
    a. za javni diskurz:
        i. SV Slovenija
        ii. JZ Slovenija
        iii. celotna Slovenija
    b. za nejavni diskurz:
        i. MS
        ii. MB
        iii. SG
        iv. CE
        v. LJ
        vi. NM
        vii. PO
        viii. KR
        ix. KK
        x. KP
        xi. GO
        xii. Italija
        xiii. Avstrija
        xiv. Madžarska
        xv. Neslovenci
        xvi. nedoločno
7. vir:
    a. ko je vir posnetka nekdo drug, npr. medijska hiša, klicni center itd., navedemo ime institucije
    b. ko smo posnetek posneli sami izključno za GK, navedemo »terenski posnetek«

8. kraj: zemljepisni kraj (mesto, naselje, vas itd.), kjer je potekal diskurz

9. čas: datum in okvirna ura, ko je potekal diskurz

10. udeleženci: št. aktivnih udeležencev (tistih, ki kaj povejo)

11. opis govornega dogodka: opišemo najvažnejše kontekstne značilnosti (npr. tema, namen pogovora, razmerja med udeleženci...)

4.3 Podatki o transkribiranju

Za vsako transkripcijo se kronološko in po fazah dela beležijo podatki o datumu in osebi, ki je delala ali kakor koli spreminjala transkripcijo.

4.4 Struktura diskurzov

4.4.1 Izjava

Izjava je osnovna enota govora, ki približno ustreza pojmu povedi v pisnem jeziku. Določimo jo tako, da je prozodično, semantično in skladenjsko zaokrožena enota.

Pri zelo strnjenem govoru smo pozorni, da izjavo označimo na takem mestu, da je v signalu mogoče nedvoumno označiti mejo med izjavama, ne da bi pri tem odrezali konec prejšnje ali začetek nove izjave.

Kjer zgornji vodili omogočata različne odločitve, imajo prednost krajše enote.

4.4.2 Vloga

Vloga: pomeni govor enega govorca, dokler ga ne prekine drug govorec. Vloga je lahko sestavljena iz ene ali več izjav. Za vsako vlogo z identifikacijsko kodo govorca označimo, kdo je govorec.

Hkratni govor:
Kot začetek hkratnega govora označimo začetek izjave, v kateri se vključi drug govorec, ne glede na to, ali se vključi na začetku ali sredi izjave.

Kot konec hkratnega govora označimo konec zadnje izjave, v kateri se pojavlja hkratni govor, ne glede na to, ali se hkratni govor konča že sredi izjave.

Oporne signale obravnavamo enako kot hkratni govor.

V hkratnem govoru ne označujemo, točno kateri deli besedila so izgovorjeni hkrati.

Če govorijo hkrati več kot dva govorca, zapisujemo največ govor dveh govorcev, če je razumljiv, sicer takih segmentov sploh ne zapisujemo.

Daljše prekinitve:
Premore v govoru, daljše kot 1,5 sek., označimo kot prazno izjavo z oznako »premor«. Na tak način označujemo tudi reklamne bloke, novinarske prispevke, glasbene točke in druge prekinitve tistega dela diskurza, ki ga transkribiramo, v televizijskih in radijskih oddajah.

4.5 Zapis govora

Za namene jezikovnotehnološke uporabe korpusa je zelo priporočljivo, da je govor zapisan v knjižni normi, kot izhaja tudi iz jezikovnotehnološke prakse transkribiranja govora v slovenščini. Vendar na ta način izredno popačimo resnično jezikovno podobo, v kateri je veliko redukcij glasov ter neknjižnih besednih oblik in besed, tako da iz samega zapisa uporabnik dobi zelo nenatančen vtis o diskurzu ter nepopoln oz. napačen vtis o jezikovni podobi. Zato se odločimo za dva nivoja zapisa: pogovornega in knjižnega. Pogovorni zapis sledi smernicam transkribiranja govorjenih besedil, ki se oblikujejo v besediloslovnih in pragmatičnih raziskavah, ter praksi pogovornega pisanja v nekaterih spletnih in leposlovnih besedilih. Tak zapis poteka veliko hitreje, kot bi potekal fonetični zapis z abecedo SAMPA, in uporabnikom korpusa na enostaven način in v poznanem črkopisu predstavi govor. Tak način zapisovanja bo tudi omogočil za cilje GK zelo pomembne raziskave najbolj tipičnih neknjižnih besednih oblik in njihovih oblikoskladenjskih vlog. Na drugem nivoju bo dodan poknjižen zapis, katerega osrednji namen bo izboljšati in razširiti korpusne iskalne možnosti.

4.5.1 Prvi nivo zapisa govora – pogovorni zapis

Osrednje vodilo: Govor zapisujemo v veljavnem slovenskem črkopisu in upoštevamo veljavne strategije predstavljanja posameznih glasov z določenimi črkami. Upoštevaje omejitve, ki izhajajo predvsem iz omejenega nabora črk, pa pri tem kolikor mogoče verno predstavimo glasovno podobo govora.

Podrobna pravila zapisa:

1. Redukcije:
    a. Glasov, ki niso izgovorjeni, ne zapisujemo, npr. *tud, neki, tko, mam, čevli*...
    b. Polglasnika ne zapisujemo posebej pri:
        i. zvočnikih r, l, m, n: *sn, pr, mislm, hitr, zloml, prjatlci...*
        ii. enoglasovnih predlogih, členkih ipd.: *s, z, d...* (tudi če so izgovorjeni zložno, s polglasnikom)
        iii. enozložnih besedah: *js, nč*...
    c. Polglasnik lahko zapisujemo z »e« v dvo- ali večzložnih besedah, npr. *kešni (kakšni),* razen pred zvočniki m, n, r, l (*zloml, mislm, hitr*...).
    d. Zapisovanje oblik pomožnega glagola »biti«:
        i. redukcije »bi« v »b« zapisujemo kot samostojno besedo, npr. *ne b (ne bi), če b (če bi), pa b mene (pa bi mene)*, *najraj b vidu...*
        ii. redukcije in premene oblik za prihodnjik (*bom, boš, bo*...) zapisujemo na naslednji način: *čev (če bo), navm (ne bom), nav (ne bo)*...

2. Premen po zvenečnosti v pisavi ne upoštevamo (*tud dobr, tud tak, grandž scena*...).

3. Zvočnik dvoustnični v (ni nosilec zloga) zapisujemo s črko »v« (*prov, nav, navm, odpravt, davn*...) oz. tudi z »l«, če tako izhaja iz knjižne norme (*kosil, mel*). Če je u samoglasniški, tj. je nosilec zloga (tudi če gre za predlog *v,* izgovorjen samoglasniško), ga pišemo s črko »u« (*pršu, vidu, u tem delu*...).

4. Diftonge in druge pokrajinsko specifične foneme, ki jih ni v knjižnem jeziku, pišemo z najbližjimi ustreznimi črkami, odvisno tudi od izgovorjave v konkretnih primerih, npr. *»ej«, »ov«, »je«; »u«* ali tudi *»i«* za u s preglasom; *»h«* ali tudi *»g«* za zveneči primorski h; *»r«* za mehkonebni koroški r itd.

5. Podaljšan polglasnik ali zvočnik *m* ali *n* in njihove kombinacije, ki pogosto zapolnjujejo premore v govoru, pišemo s tremi črkami, in sicer: *eee, eem, een, nnn, mmm...* Druge medmete zapišemo z nizom črk, ki najbolje ustreza dejanski izgovorjavi. Trajanja medmetov ne označujemo posebej.

6. Zloženke: kadar čutimo, da gre za eno besedno enoto, jih pišemo skupaj in brez vezaja (ne glede na to, ali gre za podredno ali prireden zloženko), če ne predstavljajo ene besedne enote oz. gre za zvezo prislova in pridevnika, ju zapišemo s presledkom kot dve besedi.

7. Besedni fragmenti (prekinjene besede ipd.): označimo s praznim oklepajem stično za besedo, npr. *lju()*.

8. Ločila: jih ne uporabljamo, izjema sta:
    a. vprašaj za vprašanja,

b. klicaj za izrazito ukazujoč govor oz. ob zavpitju ali vzkliku,

z namenom, da si uporabniki lažje predstavljajo zvočno podobo govora in lažje razumejo pomen. Vprašaj in klicaj pišemo stično.

9. Izjave: začenjamo z malo, ne z veliko začetnico.

10. Lastna imena:
    a. Domača lastna imena: zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico skladno s pravopisom, npr. *Delo, Brežice*. Večbesedna lastna imena dodatno označimo z zavitimi oklepaji (npr. *{Novo mesto}, {Lenart v Slovenskih goricah}, {Ministrstvo za kulturo Republike Slovenije}, {Občina Starše}, {Osnovna šola Ivana Cankarja}* itd.).
    b. Tuja lastna imena: zapisujemo tako, kot so izgovorjena, vendar z veliko začetnico, npr. *Bler, Hjuston.* Če so večbesedna, jih označimo z zavitimi oklepaji, npr. *{Nju Jork}, {Los Endželes}*.

11. Citatne besede, ki niso lastno ime: pišemo tako, kot so izgovorjene.

12. Osebni podatki o samih govorcih (tudi pri javnem diskurzu): jih anonimiziramo tako, da označimo samo vrsto podatka, in sicer na naslednji način:
    a. *[ime],*
    b. *[priimek],*
    c. *[ulica],*
    d. *[št],*
    e. *[tel],*
    f. *[email],*
    g. *[EMŠO],*
    h. *[DŠ],*
    i. *[TRR].*

    Prav tako anonimiziramo osebne podatke o osebah, ki sicer niso prisotne v diskurzu, so pa vseeno omenjene, če gre za nejavne osebnosti.

    Normalno, kot lastno ime, pa zapišemo imena javnih osebnosti, ki so omenjena v diskurzu, npr. imena politikov, športnikov, novinarjev in voditeljev, umetnikov in drugih kulturnih delavcev ter ostalih medijsko opaznih osebnosti.

    Kraj bivališča in poštna številka ne štejeta za osebni podatek in ju ne anonimiziramo.

    Za vse osebne podatke, ki jih anonimiziramo, dodatno označimo tisti del zvočnega posnetka, v katerem je izgovorjen osebni podatek oz. zaporedoma več osebnih podatkov, tako da je v transkripciji zapis časovnih mej, ki določajo osebni podatek v zvočnem posnetku (uporabimo sled background/other).

13. Kratice:
    a. Pišemo tako, kot so izgovorjene, vendar skupaj, če gre za eno kratico, npr. *erteve, teve, trr*.
    b. Če je kratica lastno ime, jo pišemo z veliko začetnico, npr. *Sazuja, Tevetri* itd.

14. Številke: vedno izpišemo z besedo.

15. Nerazumljive ali nerazločljive besede ali daljše govorne enote: označimo z oznako:
    a. »neraz«.
    Trajanja ne označujemo.

16. Petje:
    a. Če je npr. zapet le kratek verz ali par besed, zapeto besedilo vključimo v zapis govora in ga ne označujemo posebej.
    b. Če je zapet daljši odsek (več verzov, kitica, cel refren, cela pesem...), ne transkribiramo, ampak označimo kot premor (prazna izjava).

17. Daljše premore (več kot 1,5 sekunde) označimo kot prazno izjavo brez besedila in brez govorca, označimo jo z oznako:
    a. »premor«.
    Ne opisujemo posebej, kaj se je v premoru dogajalo (npr. tišina, prevezovanje, reklame, poudarjanje povedanega ipd.).

4.5.2 Drugi nivo zapisa govora – knjižni zapis

Potem ko bo opravljen prvi nivo zapisa govora po pravilih, navedenih v 4.5.1, bo dodan drugi nivo zapisa govora, ki bo vsaki besedi pripisal najbližjo knjižno besedno obliko, z namenom, da se omogočijo boljše iskalne možnosti.

Osrednje vodilo:

Pri pretvorbi pogovornega zapisa v knjižni zapis odpravimo glasoslovne premene, ki so prisotne pri posamezni besedni obliki, ob upoštevanju pogostosti rabe. Izhodišče je knjižna različica istega leksema. Na drugih jezikovnih ravneh besed ne spreminjamo. Za ločevanje, kdaj gre za glasovno premeno in kdaj ne, se bodo ob primerih oblikovala načela dobre prakse. Če določenega leksema ni v knjižni normi, ga ohranimo v obliki, ki se pojavlja v govoru.

Podrobna pravila zapisa:

1. Če eno besedo na prvem nivoju zapisa pretvorimo v dve ali več besed na drugem nivoju, to ustrezno označimo.

2. Če po dve ali več besed na prvem nivoju zapisa pretvorimo v eno besedo na drugem nivoju, to ustrezno označimo.

3. Onomatopeje, medmete, besedne fragmente in druge glasove, za katere v knjižnem jeziku ni standardnega zapisa, pustimo zapisane tako, kot so bili zapisani na prvem nivoju zapisa.

4. Lapsuse v izgovorjavi, če so nedvoumni, odpravimo (npr. *indidualnih -> individualnih*).

5. Zloženke pišemo enako kot na prvem nivoju, samo skupaj ali narazen, brez vezajev.

6. Kratice zapišemo z velikimi črkami:

ce -> C
veveve -> WWW
es i -> SI
ha pe -> HP
s o n c h e k -> S O N C H E K
kagebe -> KGB
ertees -> RTS

7. Tuja lastna imena zapišemo po knjižni normi:

   {Kos Porta} -> {Cost Porta}
   {Šarm el Šejk} -> {Sharm El Sheikh}
   Atene -> Atene
   {vadi Mudžep Kinks Vej} -> {Wadi Mudžep Kings Vej}

8. Citatne občne besede zapišemo po knjižni normi, ali citatno ali v poslovenjenem zapisu – slednje zlasti v primerih, ko so poslovenjene oblike že izpričane v pisnih besedilih (korpusi, internet...). Vezajev ne pišemo, ampak se odločimo za zapis ali skupaj ali narazen.

   granč scena **->** *grunge* scena
   ofišl suporterja -> official supporterja
   rentakar -> rentacar
   sori -> sori
   tu mač -> tu mač
   pab -> pub

9. Ločil ne spreminjamo in ne dodajamo (tudi ne pik in vejic).

10. Začetki izjav ostanejo z malo začetnico.

4.6 Nebesedni in nejezikovni zvoki ter prozodija

4.6.1 Nebesedni zvoki

1. Onomatopeje (posnemanje zvokov iz narave): zapišemo jih z nizom črk, ki najbolje ustreza dejanski izgovorjavi.

2. Druge zvoke, ki nastanejo z govorili in prispevajo k vsebini diskurza, pa jih nikakor ne moremo ustrezno zapisati s črkami, lahko označimo z oznako:
   a. »glas«.

3. Smeh: ne označujemo trajanja, ampak samo označimo mesto, kjer se smeh začne, in sicer z oznako:
   - »smehgo«, če se smeji samo govorec,
   - »smehna«, če se smejijo samo naslovniki oz. občinstvo, in
   - »smehob«, če se smejijo oboji,
   - ne označujemo, kakšen je smeh, kako glasen je, kako dolgo traja, prav tako ga ne zapisujemo z medmeti, ki običajno označujejo smeh (npr. haha);
   - nejasnih primerov ne označimo kot smeh.

4. Druge pragmatično pomembne zvoke, ki nastanejo z govorili, kot so npr. jok, hlipanje oz. hlipajoč govor, zehanje, vzdih, odkašljanje ipd., označimo z oznako:
   a. »glas«
   na mestu v transkripciji, kjer se tak zvok začne.

5. Vdihov in izdihov, kašljanja, tleskov z jezikom in drugih zvokov, ki ne nosijo sporočila in niso pragmatično pomembni za diskurz, ne označujemo.

4.6.2 Nejezikovni zvoki

V to skupino štejemo zvoke, ki ne nastanejo z govorili, ampak je njihov izvor zunanji. Če je tak zvok pragmatično pomemben, npr. zvonjenje telefona prekine potek diskurza in eden od govorcev se odzove na klic, označimo to z oznako:

a. »zvok«

na mestu v transkripciji, kjer se tak zvok začne.

Če zvok ne vpliva na diskurz, ga ne označujemo (npr. v ozadju se pojavi glasba, vendar sogovorniki ne reagirajo na to, sliši se listanje, premikanje mikrofona...).

4.6.3 Prozodija

Prozodičnih lastnosti govora, kot so tempo, jakost, trajanje glasov, intonacija, ton, poudarjanje, krajši premori (manj kot 1,5 sek.) z namenom poudarjanja ipd., ne označujemo, predvsem zato, ker je pri takšnem obsegu gradiva te pojave zelo težko in zamudno dovolj natančno in homogeno označiti, hkrati pa za cilje GK, kot so definirani v poglavju 3, ti podatki niso zelo pomembni.

5 ZAGOTAVLJANJE KVALITETE

Kvaliteta transkripcij se bo zagotavljala na naslednje načine:

1. Z največjo možno stopnjo računalniške podpore ob transkribiranju:
   a. uporaba ustreznih programskih orodij (Transcriber, Excel, Open Office ipd.);
   b. uporaba črkovalnika;
   c. omogočanje izbir s klikom namesto tipkanja podatkov, kjer je to mogoče (obrazci s podatki o govorcu, obrazci s podatki o diskurzu, metaoznake itd.);
   d. uvedba spletnega mesta, kjer se bodo zbirali vzorčni primeri, primeri iz prakse ipd.

2. Z avtomatsko validacijo:
Avtomatsko validiranje XML-zapisa, vrste vnesenih dogodkov oz. metaoznak in dolžine sledi za označevanje osebnih podatkov.

3. Z ročnim preverjanjem:
   a. manjše število kontrolorjev bo opravljalo ročno kontrolo vseh ali večine transkripcij, pri čemer bodo vnašali potrebne popravke;
   b. manjše število sodelavcev bo preverjalo 10% naključno izbranih transkripcij; v kolikor bo število napak manjše od 8 za 1 minuto govora (tj. ca. 150 besed), kar je povprečno manj kot 5%, šteje, da so transkripcije ustrezne kvalitete, sicer se jih vrne v popravljanje prvotnemu transkriptorju in nato kontrolorju.

6 LITERATURA IN VIRI

Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J., Traum, D.R., 2004: Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus. Proceedings of 4th Internation Conference on Language Resources and Evaluation '04, Lizbona, Portugalska.

Krajnc, M., 2005: Besedilne značilnosti javne govorjene besede: Na gradivu sej mariborskega Mestnega sveta. Maribor: Slavistično društvo.

Kranjc, S., 1999: Razvoj govora predšolskih otrok. Ljubljana: Znanstveni inštitut Filozofske fakultete.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbara, I., Milde, J.T., Parrill, F., Rose, T., Schmidt, T., Sloetjes, H., Thies, A., Wllinghoff, S., 2006: Comparison of multimodal annotation tools – workshop report. Gespraechsforschung – Online-Zeitschrift zur verbalen Interaktion, 7, 99-123.

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

Šarotar, Dušan, 2007: Biljard v Dobrayu. Ljubljana: Študentska založba.

Verdonik, D., 2006: Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

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

Zemljarič Miklavčič, J., Stabej, M., 2005: Building a pilot spoken corpus. V: Garabik, R. (ur.): Computer Treatment of Slavic and East European Languages. Slovaška, Bratislava. 229-240.

Zemljarič Miklavčič, J., 2006: Korpus govorjene slovenščine. V: Erjavec, T., Žganec Gros, J. (ur.). Jezikovne tehnologije: zbornik 9. mednarodne multikonference Informacijska družba IS 2006. Ljubljana: Institut Jožef Stefan. 124-127.

Zemljarič Miklavčič, J., 2007: Načela oblikovanja govornega korpusa za slovenščino. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Zorko, Z., 1995: Narečna podoba Dravske doline. Maribor: Kulturni forum.

Žgank, A., T. Rotovnik, D. Verdonik, Z. Kačič, 2004: Baza Broadcast News za slovenski jezik (BNSI) in sistem za razpoznavanje tekočega govora. V: 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. V: 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. V: Informacijska družba IS'2004: Jezikovne tehnologije. Ljubljana, Institut Jožef Stefan. 94-97.

Welsh, I., 1997: Trainspotting. Prevod: A. Skubic. Ljubljana: DZS.