Univerza v Ljubljani
FILOZOFSKA FAKULTETA

PREVODOSLOVJE
IN UPORABNO
JEZIKOSLOVJE

Jana Zemljarič Miklavčič

GOVORNI KORPUSI

Jana Zemljarič Miklavčič

GOVORNI KORPUSI

GOVORNI KORPUSI
Zbirka Prevodoslovje in uporabno jezikoslovje

Avtorica: dr. Jana Zemljarič Miklavčič
Recenzenta: dr. Marko Stabej, dr. Vojko Gorjanc
Povzetek prevedla: dr. Agnes Pisanski Peterlin

Izdala: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
Založil: Oddelek za prevajalstvo
Za založbo: Roman Kuhar, dekan Filozofske fakultete
Oblikovanje: Bons, d. o. o.

Prva izdaja, elektronska izdaja
Publikacija je brezplačna.
Publikacija je dostopna na: https://e-knjige.ff.uni-lj.si
DOI: 10.4312/9789612379902

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID=292908800

ISBN 978-961-237-989-6 (epub)
ISBN 978-961-237-990-2 (pdf)

Valentini

Kazalo vsebine

Predgovor

Raziskovanje govorjenega jezika in korpusno jezikoslovje sta vznemirljivi in aktualni področji sodobnega jezikoslovja, zato je velik izziv povezati ti dve temi v okviru znanstvene monografije. Govorno sporazumevanje je v marsičem primarna oblika sporazumevanja, tako s stališča posameznika kot s stališča človeštva, poleg tega je oblika, v kateri se jezik najpogosteje udejanji. Kljub temu v raziskanosti zaostaja za pisnim jezikom, predvsem zaradi svoje akustične narave, ki jo je (bilo) težko ujeti, shraniti in analizirati. Večje raziskave govora je omogočil šele razvoj računalniških tehnologij v zadnjem desetletju ali dveh: digitalno snemanje, shranjevanje velikih količin podatkov, možnost urejanja in analize z računalnikom – vse to se je združilo v obliki računalniških korpusov, ki predstavljajo sodobno orodje za raziskovanje govora. Govorni korpusi so v okviru sicer relativno mlade jezikoslovne veje – korpusnega jezikoslovja – nekaj, kar je novejše, manj znano, težje izvedljivo, vendar nujno potrebno. Številne jezikovne skupnosti so že zgradile govorne korpuse in s tem omogočile raziskovanje govorjenega jezika; zgrajeni so korpusi za angleščino (britansko, ameriško in mednarodno), nizozemščino, švedščino, gradijo jih na Norveškem, Češkem, v Nemčiji, na Madžarskem, načrtujejo pa jih še številni drugi narodi, med drugimi načrtujemo govorno komponento referenčnega korpusa tudi pri nas.

Govorni korpusi pa ne bodo poravnali samo dolgov na področju raziskovanja govorjenega jezika, ampak je njihova uporabnost usmerjena tudi v sedanjost in prihodnost – v razvoj govornih tehnologij. Raziskave avtomatskega pretvarjanja govora v zapis in obratno ter strojnega prevajanja so se znašle na točki, ko na podlagi studijskih posnetkov govora in brez korpusov spontanega govora ne morejo več napredovati. Govorni korpusi predstavljajo nujni pogoj za razvoj govornih tehnologij, kar je ključnega pomena za vsak jezik, ki želi ohraniti svojo suverenost in avtonomijo na vseh področjih sporazumevanja, tudi kot jezik hitro razvijajočih se računalniških govornih aplikacij.

Prve spodbude in načrti za gradnjo govornega korpusa na Slovenskem so prišli s strani avtorjev prvega slovenskega referenčnega korpusa Fida, saj je za celovito raziskovanje posameznega jezika poleg pisnega nujno zgraditi tudi korpus avtentičnih govorjenih besedil. Pobuda je v natančno desetih letih (1998–2008) prerasla v konkretni načrt za gradnjo referenčnega govornega korpusa za slovenščino v okviru projekta Sporazumevanje v slovenskem jeziku,[1] prispevek k načrtovani gradnji pa je tudi pričujoča monografija.

Sama sem se nad korpusi navdušila zato, ker gre za področje, kjer se stikajo humanistika in informacijske tehnologije. Ko sem pripravljala doktorsko disertacijo, se

[1] http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx

mi je v okviru štipendije Marie Curie (6. evropski okvirni program) pod mentorstvom prof. dr. Konraada de Smedta ponudila priložnost za študij na Oddelku za kulturo, jezik in jezikovne tehnologije (AKSIS)[2] na Univerzi v Bergnu na Norveškem. Tam sem lahko dostopala do večine znanih govornih korpusov, s pomočjo norveških kolegov pa tudi zgradila manjši učni korpus govorjene slovenščine, ki omogoča računalniško obdelavo jezikovnih podatkov.[3] Učni korpus naj bi omogočil čim boljšo verifikacijo hipotez in načel, oblikovanih v teoretičnih poglavjih disertacije. Z njim se je bilo mogoče učiti, kako snemati in shranjevati govorjena besedila, kako jih transkribirati in označevati, hkrati pa je bilo mogoče nakazati nekatere možnosti za uporabo korpusa – opis in analizo govorjene slovenščine.

Pričujoča knjiga *Govorni korpusi* je v veliki meri nastala na podlagi doktorske disertacije *Načela gradnje govornega korpusa slovenščine*, pa tudi nekaterih kasnejših raziskav, ki so nastale na podlagi učnega korpusa. Obsežnega dela v zvezi z načrtovanjem govornega korpusa ne bi mogla opraviti brez sodelovanja kolegov, strokovnjakov s področja korpusnega jezikoslovja. Zato se posebej zahvaljujem dr. Marku Stabeju in dr. Vojku Gorjancu za mentorstvo in dolgoletne spodbude v obliki strokovnih debat in prijateljskih srečanj, Knutu Hoflandu pa za tehnično podporo pri gradnji učnega korpusa; zahvaljujem se tudi kolegom, ki so sodelovali s prispevki spontanega govora, in nenazadnje Agenciji za raziskovalno dejavnost RS ter Oddelku za prevajalstvo FF UL, ki sta omogočila natis te knjige. Upam, da bo razprava dosegla svoj namen in konstruktivno prispevala h gradnji referenčnega govornega korpusa ter s tem k raziskovanju slovenskega jezika.

Avtorica

[2] http://www.aksis.uib.no/
[3] http://torvald.hit.uib.no/talem/jana/

1 Uvod

1.1 RAZISKOVANJE GOVORJENEGA JEZIKA NA SLOVENSKEM

Govorjeni slovenščini je bilo doslej, razen v dialektologiji, v primerjavi s pisnim jezikom posvečene manj jezikoslovne pozornosti: raziskav, ki bi temeljile na analizi spontanega javnega in zasebnega govora, ni bilo veliko, za kar obstajajo tudi utemeljeni razlogi. Med najstarejše znane modernejše[4] razprave o govorjenem jeziku sodijo *Vprašanja govorjenega jezika* (Pogorelec 1965). Tu so že zaznane lastnosti govorjenega jezika, kot jih opažamo še danes, in upoštevani zunanji dejavniki, ki vplivajo na govor: »Kadar govorimo, govori z nami vse okolje in se pridružuje našemu izrazu: ljudje, s katerimi se pogovarjamo ali jim kaj pripovedujemo, prizorišče, kjer govorimo, naš temperament si pomaga s kretnjami, in da ne bi učinkovali prisiljeno, priredimo temu tudi jezik, v katerem govorimo« (Pogorelec 1965, 301). Navedenih značilnosti govorjenega jezika v času nastanka razprave ni bilo mogoče preverjati na večji količini avtentičnega gradiva, kar je omejevalo raziskovanje in analizo govorjenega jezika še tudi kasneje.

Kljub temu je v zadnjem desetletju nastalo nekaj odmevnejših razprav o govorjenem jeziku; mednje sodijo *Govor celjskega predmestja Gaberje* (Škofic-Guzej 1998), *Jezikovno načrtovanje govorjenega jezika pri Slovencih* (Pogorelec 1998), *Podoba govorjenega slovenskega knjižnega jezika v Slovenskem pravopisu* (Tivadar in Jurgec 2001), *Govorjeni knjižni jezik v televizijskih dnevnoinformativnih oddajah: študija primera* (Verovnik 2004), *Nekateri vidiki zvrstnosti govorjenega diskurza s stališča poslušalca* (Vogel 2004), *Podoba in funkcija govorjenega knjižnega jezika glede na neknjižne zvrsti* (Tivadar 2004) in druge.[5] Vsem razpravam je skupno, da so nastale na podlagi teoretičnega razmisleka in na analizi omejene količine avtentičnega gradiva ter ob zelo omejenih možnostih obdelave (štetje primerov ipd.). Tako npr. Škofic-Guzej zaključuje svojo razpravo z ugotovitvijo, da bi bilo »za natančno predstavitev uporabljenih jezikovnih vezov in prikaz medzvrstnega preklapljanja v različnih govornih položajih ter za natančnejši pregled značilnosti sistemov govorjenega jezika teh informatork in nato slovenskih pogovornih jezikov sploh seveda potrebno obdelati in natančneje razčleniti veliko več govornih dogodkov« (Škofic-Guzej 1994, 577). Pogorelec pa ugotavlja, da je govor kljub »prvenstveni vlogi v

[4] Nekateri starejši pogledi na govorjeno slovenščino so navedeni npr. pri Pogorelec 1998 in Tivadar 2004.

[5] Zdi se, da so povečanemu zanimanju za raziskovanje govorjenega jezika v 21. stoletju pot utirale diplomske naloge, ki so nastale v zadnjem desetletju 20. stoletja – npr. Mihaela Bregant, *Mariborski pogovorni jezik*, Maribor, 1991; Marjeta Longyka, *Nekatere prvine pogovornega jezika v Ljubljani*, Ljubljana, 1994; Andrej Skubic, *Geografsko-socialna pogojenost govorca, funkcija in vsebina besedila ter okoliščine govornega dogodka kot dejavniki jezikovne zvrstnosti*, Ljubljana, 1994; Mario Galunič, *Struktura povedi (problem zloženosti) govorjenega knjižnega jezik*a, Ljubljana, 1995; Hotimir Tivadar, *Govorjeni knjižni jezik – njegovo normiranje in uresničevanje*, Praga, Ljubljana, 1998; Nataša Hribar, *Govorjeni jezik politikov (razčlenitev besedil z vidika skladenjske strukture in konferenčnosti)*, Ljubljana, 2000.

jezikovnem dogajanju« ostal »nemara tudi zaradi težav pri uvodnem določanju in zbiranju gradiva bolj ali manj neobdelan« (Pogorelec 1998, 59).

Prav v zadnjih letih je pri raziskovanju govorjenih besedil prišlo do bistvenega izboljšanja možnosti za zbiranje in obdelavo gradiva. Različne avdio- in videonaprave, ki so postale širše dostopne, omogočajo snemanje zvočnega gradiva oz. govora, digitalne tehnike omogočajo shranjevanje gradiva na računalnik, računalniška orodja pa njegovo analizo. Tako sta nastali (vsaj) dve razpravi, ki sta izkoristili v dokumentacijske namene posneta in transkribirana govorjena besedila (razprave v Državnem zboru RS in seje Mestnega sveta Maribora), in sicer dve magistrski nalogi, *Skladenjska razčlenitev sodobnega slovenskega parlamentarnega jezika* (Hribar 2003) in *Besediloslovne značilnosti pokrajinskega pogovornega jezika (na gradivu mariborščine)* (Krajnc 2004).[6]

Novejši pristop v raziskovanju spontanega govora predstavlja razprava, v kateri avtorja s sodobnimi metodami in tehnologijo raziskujeta prozodične lastnosti spontanega govora (Vitez in Zwitter Vitez, 2004). Korak naprej predstavljata tudi doktorski disertaciji, ki temeljita na gradivu, posnetem posebej za namen raziskav, pri transkribiranju posnetkov pa ohranjata specifične (tudi fonetične in deloma prozodične) lastnosti govora; to sta disertaciji *Vpliv besedilne vrste na uresničitev skladenjskih struktur (primer narativnih besedil v vsakdanjem spontanem govoru)* Mojce Smolej (2006) in *Analiza diskurza kot podpora sistemom strojnega simultanega prevajanja govora* Darinke Verdonik (2006).[7] S specifikami govorjenega jezika se ukvarajata tudi disertaciji *Zasebni dvogovori* (Krajnc Ivič 2008) in *Kakovost in trajanje samoglasnikov v govorjenem knjižnem jeziku* (Tivadar 2008).

Kljub naraščajočemu številu razprav lahko sklenem, da v slovenskem jezikoslovju druge polovice dvajsetega in v začetku enaindvajsetega stoletja teorija govorjenega jezika v polnem pomenu besede še ni bila razvita; jezikoslovci so opazovali predvsem knjižni jezik (*langue*), pri raziskovanju govora pa, kadar ni šlo za narečne raziskave, iskali predvsem sistemske lastnosti v govoru ali odstope od knjižnega jezika, torej odnose na ravni *langue-parole*. Celovitejši pristop k raziskovanju govorjenega jezika bi oz. bo omogočil šele referenčni korpus z reprezentativno in uravnoteženo govorno komponento.

[6] Kasneje tudi monografija – *Besedilne značilnosti javne govorjene besede: na gradivu sej mariborskega Mestnega sveta* (Krajnc Ivič 2005).

[7] Kasneje monografija - *Jezikovni elementi spontanosti v pogovoru* (Verdonik 2007).

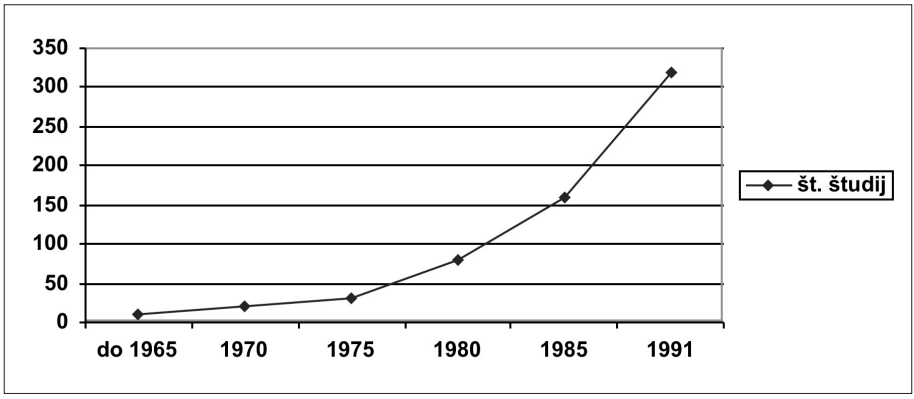
1.2 RAZVOJ KORPUSNEGA JEZIKOSLOVJA

Nov pristop k raziskovanju in opisovanju jezika se je izoblikoval skozi novo jezikoslovno vejo, za katero se je ustalilo ime korpusno jezikoslovje. Danes si pod pojmom *korpusni jezikoslovec* predstavljamo raziskovalca, ki načrtuje gradnjo korpusa ali pa svoje jezikoslovne študije razvija na podlagi raziskav korpusa, v kombinaciji z drugimi jezikoslovnimi vejami. Korpusno jezikoslovje je študij in opisovanje jezika na podlagi empiričnega gradiva, v ožjem smislu pa tudi oblikovanje metodologije za gradnjo korpusov in njihova dejanska gradnja, tudi na Slovenskem »oblikujoče se metodološko in teoretično jedro drugačnega pristopa k opazovanju, opisovanju in nenazadnje tudi predpisovanju jezika« (Stabej 2003b, 24).

V zgodovini jezikoslovja naletimo na izraz »zgodnje« korpusno jezikoslovje (McHenry in Wilson, *Early Corpus Linguistics*).[8] Pojem se nanaša še na čas pred Chomskim, ko so se posamezni raziskovalci lotevali jezikoslovnih študij na avtentičnem jezikovnem gradivu. Med tovrstne raziskave iz prve polovice 20. stoletja sodi npr. raziskovanje otroškega govora, pa tudi raziskave jezika ameriških Indijancev. Chomsky je v petdesetih letih preusmeril pozornost jezikoslovja stran od empiričnih študij, vendar so nekateri jezikoslovci tudi v obdobju prevlade Chomskega nadaljevali s pionirskim delom na področju korpusov. Tako je npr. Randolph Quirk leta 1959 zasnoval prvi predračunalniški korpus pisne in govorjene angleščine *Survey of English Usage* (SEU). John Sinclair je leta 1960 vodil gradnjo prve zbirke govorjenih besedil, ki je bila shranjena na računalniku (Sinclair 1995, 99), kar sam postavlja na začetek korpusnega jezikoslovja. Leta 1961 sta William Nelson Francis in Henry Kučera začela dve desetletji trajajoč projekt gradnje prvega računalniškega korpusa Brown. Ti raziskovalci so bili takrat sicer v manjšini, vendar so kmalu dobili številne posnemovalce, med prvimi Jana Svartvika, ki je l. 1975 govorno komponento korpusa SEU pretvoril v digitalno obliko, primerno za računalniško branje, in s tem začel gradnjo prvega govornega korpusa.

V tem času so računalniki začeli postajati glavna podpora korpusnega jezikoslovja in komaj je bilo mogoče slutiti, kakšne spremembe in kolikšne nove možnosti bo razvoj računalniške tehnologije vnesel v raziskovanje jezika. Vzpon korpusnega jezikoslovja je bil bliskovit, kakor kaže tudi spodnja tabela, ki prikazuje naraščanje števila raziskav s področja korpusnega jezikoslovja v obdobju 1965–1991:

[8] http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/corpus1/1fra1.htm



Slika 1: Razvoj korpusnega jezikoslovja v obdobju 1965–1991 (Johansson 1991, 313)[9]

Za obdobje 1990–1998 je bibliografija študij, ki so nastale na podlagi korpusov angleškega jezika, zbrana na spletni strani ICAME;[10] obsega 47 strani oz. čez 700 bibliografskih naslovov. V zadnjem desetletju v središču zanimanja korpusnega jezikoslovja ni več samo angleščina, ampak se je gradnja korpusov razširila med številne druge svetovne jezike.

Kljub temu, da je korpusni pristop v jezikoslovju dandanes najširše priznan (ali pa morda ravno zaradi tega), pa to ne pomeni, da ga vsi raziskovalci razumejo na enak način. V evropskem jezikoslovju je v zadnjem desetletju postala »razširjena navada, da se rezultate jezikoslovnih analiz zavaruje s sklicevanjem na korpusnojezikoslovne postopke« (Teubert 1999; V: Gorjanc in Krek 2005b, 103). Pozornost raziskovalcev je posvečena vprašanjem, kako lahko s korpusi in njihovo analizo podpremo klasično jezikoslovje – »strukturalistično jezikoslovje se podkrepljuje s korpusnimi podatki« (Teubert 1999, prav tam) – pogled, ki korpusno jezikoslovje vidi samo kot metodo raziskovanja, ne pa kot samostojno jezikoslovno vedo z lastnim teoretičnim izhodiščem. Nasprotno mnenje je, da se korpus sicer lahko uporablja za potrjevanje in ovrednotenje hipotez o jeziku, lahko pa se ga uporablja kot izhodišče za gradnjo nove jezikovne teorije (Tognini-Bonelli 2001, 65). V primerih, ko se korpus uporablja za potrditev hipotez in jezikovnih opisov, ki so mnogo starejši od samih korpusov, je to t. i. *delni korpusni pristop* (Gorjanc 2005; V: Gorjanc in Krek 2005b, 185). *Popolni korpusni pristop* pa narekuje uporabo korpusa neodvisno od uveljavljenih jezikoslovnih teorij in interpretacij; pri tovrstnem razumevanju korpusnega jezikoslovja »opazovanje jezikovnih podatkov vodi v oblikovanje hipotez, nadalje v posploševanje in nazadnje v oblikovanje teoretičnih stališč oz. trditev«

[9] Avtor povzema aktualno stanje v času nastanka članka.
[10] *International Computer Archive of Modern and Medieval English,* http://khnt.hit.uib.no/icame/manuals/icambib3.htm.

(Tognini-Bonelli 2001, 85). Vloga jezikoslovca v tem procesu je izjemnega pomena: v vseh fazah sodeluje s svojim znanjem, izkušnjami in zmožnostjo interpretiranja (Tognini-Bonelli 2001, prav tam). Podobna je definicija pravega korpusnega pristopa pri Teubertu (V: Gorjanc in Krek 2005b, 108): »Korpusno jezikoslovje širi naše jezikoslovno znanje s tem, da kombinira tri postopke: identifikacijo jezikovnih podatkov v korpusu, korelacijo jezikovnih podatkov s pomočjo statističnih metod in na koncu (intelektualno) interpretacijo rezultatov.«

Glede na to, da gre pri teoriji govorjenega jezika za celostno dokaj neraziskano področje, je govorjeni jezik pravi poligon za popolni korpusni pristop. V tem smislu je treba razumeti tudi Teubertovo ugotovitev, da »je analiza govorjenega jezika tista, ki je prva zasidrala korpusno jezikoslovje kot disciplino, samostojno tudi v svoji teoretični težnji« (Teubert 1999; V: Gorjanc in Krek 2005b, 105). Tudi sama namen gradnje govornega korpusa razumem v tem smislu – šele na zbranem in računalniško urejenem gradivu – referenčnem govornem korpusu – bo mogoče postavljati nove hipoteze, jih analizirati in interpretirati.[11]

1.3 SLOVENSKO KORPUSNO JEZIKOSLOVJE

Področje računalniške obdelave jezikovnih podatkov se je na Slovenskem začelo razvijati v začetku 80. let 20. stoletja; ohranjena sta zbornika II. (1982, ur. P. Tancig) in III. (1985, ur. T. Erjavec) znanstvenega srečanja *Računalniška obdelava lingvističnih podatkov*. Pobuda je bila v celoti v rokah računalničarjev, slovenske jezikoslovne stroke pa se je komajda dotaknila.[12] Gradnja jezikovnih korpusov se je na Slovenskem začela z vključitvijo v projekt MULTEXT-EAST (Multilingual Text Tools and Corpora for Central and Eastern European, 1996), v okviru katerega je nastal vzporedni večjezični pisni korpus, ki vsebuje Orwellov roman *1984* in nekaj drugih pisnih besedil ter vzporedni korpus govora; ob projektu je nastal prvi slovenski strokovni članek, ki podrobneje opisuje jezikovne korpuse in njihove lastnosti (Erjavec 1996/97).

Za začetek korpusnega jezikoslovja na Slovenskem lahko imamo leto 1997, ko se je začela gradnja prvega slovenskega referenčnega korpusa FIDA,[13] vzporedno z gradnjo pa je raslo število razprav o zajemu besedil v korpus, označevanju besedil

[11] Potrditev, da je ta pot pri raziskovanju govorjenega jezika prava, sta navsezadnje tudi že omenjeni disertaciji D. Verdonik in M. Smolej, ki v nekem smislu sledita principu popolnega korpusnega pristopa.

[12] Tomo Korošec 1982: Uporabnost računalniških konkordanc v lingvističnih in literarnih raziskavah. V: *Zbornik II. znanstvenega srečanja Računalniška obdelava lingvističnih podatkov*. Ljubljana: Institut Jožef Stefan. 405–415.

[13] V projektu, ki je trajal dve leti, so sodelovali štirje partnerji – Filozofska fakulteta, Institut Jožef Stefan, DZS in Amebis (Stabej 1998, Erjavec 1998b, Gorjanc 1999); finančno breme projekta je bilo predvsem na slednjih dveh (ekonomskih) partnerjih, zato korpus ni bil prosto dostopen.

in iskanju po korpusu. Leta 1998 je bila organizirana prva konferenca za področje jezikovnih tehnologij, ob njej pa izdan zbornik prispevkov *Jezikovne tehnologije za slovenski jezik* (ur. Tomaž Erjavec in Jerneja Gros), istega leta pa je bilo področje pokrito tudi s tematsko številko revije *Uporabno jezikoslovje* (ur. Zdravko Kačič), izdano ob II. kongresu Društva za uporabno jezikoslovje Slovenije; obe publikaciji med drugim prinašata tudi razprave s področja korpusnega jezikoslovja. Pet let kasneje (2003) je bilo dogajanje na področju jezikovnih tehnologij, predvsem tistega dela, ki se vsebinsko povezuje z uporabno slovenistiko, podrobneje predstavljeno v tematski dvojni številki *Jezika in slovstva* pod naslovom Jezikovne tehnologije za slovenščino (ur. Vojko Gorjanc).

V obdobju zadnjih desetih let je na Slovenskem nastalo – glede na število govorcev in razpoložljivost raziskovalnih potencialov – kar lepo število splošnih in specializiranih korpusov, in sicer:

- referenčni korpus slovenskega jezika FIDA (http://www.fida.net/slo/index.html),
- korpus slovenskega jezika NOVA BESEDA (http://bos.zrc-sazu.si),
- vzporedni angleško-slovenski korpus ELAN (http://nl.ijs.si/elan/),
- vzporedni angleško-slovenski korpus TRANS (http://nl2.ijs.si//index-bi.html)
- korpus prevodov evropske zakonodaje EVROKORPUS (http://www.sigov.si/evrokorpus),
- referenčni korpus slovenskega jezika FIDAPLUS (http://www.fidaplus.net),
- korpus besedil odnosov z javnostmi KoRP (http://www.korp.fdv.uni-lj.si/).

Prvo desetletje slovenskega korpusnega jezikoslovja je bilo zaznamovano predvsem z gradnjo korpusov. V letu 2005 je izšel prvi na korpusu (FIDA) temelječi angleško-slovenski slovar. Dokončani so bili trije raziskovalni projekti, ki so spodbujali delo na področju korpusnega jezikoslovja, in sicer *Jezikovni viri za slovenščino* (2003–2005, nosilec M. Stabej), *Zasnova na korpusu temelječih slovarskih in slovničnih opisov slovenskega jezika* (2004–2006, nosilec V. Gorjanc) in *Oblikovanje slovenskega korpusnega omrežja* (2004–2006, nosilec M. Stabej). Večina raziskovalne energije je bilo usmerjene v oblikovanje in izboljševanje jezikovne infrastrukture ter v povezovanje obstoječih aplikacij. Raziskovalci so se posvečali tudi specifičnim jezikovnim opisom; med prvimi sta bili disertaciji *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov* (Gorjanc 2002) in *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov* (Vintar

2003) ter diplomska naloga Špele Arhar *Gradnja specializiranega korpusa* (2004), kasneje pa še disertaciji *Načela gradnje govornega korpusa slovenščine* (Zemljarič Miklavčič 2007) in *Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah* (Logar 2007). Izdelani in načrtovani jezikovni viri skupaj s specifičnimi opisi predstavljajo izhodišče za celovito prenovo slovenskega jezikovnega opisa; med prve na korpusih temelječe jezikovne opise sodita znanstveni monografiji *Stalne besedne zveze v slovenščini: korpusni pristop* (Gantar 2007) in *Terminologija* (Vintar 2008).

1.4 NAČRT ZA GRADNJO GOVORNEGA KORPUSA SLOVENŠČINE

Potreba po izdelavi korpusa govorjene slovenščine je bila v zadnjih letih v slovenskem jezikoslovnem prostoru in v drugih vedah, ki se stikajo s korpusnim jezikoslovjem (npr. govorne tehnologije) večkrat eksplicitno izražena. Prvi je na to opozoril Stabej, ko je pri prvi 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). V prvi fazi gradnje FIDE so se govornemu korpusu morali odreči, predvsem zaradi finančnih, časovnih in človeških omejitev, razlogi pa so bili tudi jezikoslovno-teoretične narave:

> »Vsaj za slovensko jezikoslovje se zdi, da na vprašanja govorjenega jezika doslej nikakor ni bilo dovolj pozorno in da bi bilo treba še pred oblikovanjem korpusa govorjenih besedil opraviti nekaj vzorčnih in temeljnih raziskav na različnih področjih govorjene komunikacije, od besediloslovne in pragmatično-jezikoslovne analize diskurza ter sociolingvističnih raziskav do sodobnih raziskav glasovne, naglasne in intonacijske podobe slovenskega govora. Šele spoznanja takih raziskav bi lahko vzpostavila po eni strani kvaliteten okvir za zajem podatkov, po drugi strani pa odgovorila na zapletena vprašanja načina transkripcije govorjenih besedil, kar je za korpus odločilnega pomena« (Stabej 1998, 100).

Istega leta so tudi na ljubljanski Fakulteti za elektrotehniko strokovnjaki za govorne tehnologije ob predstavitvi govorne zbirke GOPOLIS[14] omenjali »pomanjkljivosti zbirke, ki so posledica pomanjkanja splošnega znanja o slovenskem govorjenem jeziku« (Dobovišek idr. 1998, 105). Simona Kranjc s člankom o otroškem govorjenem jeziku poskuša »utemeljiti potrebo po načrtnem zbiranju govorjenih besedil in njihovo vključitev v korpus slovenskega jezika« (Kranjc 1998, 109), zaključi pa z

[14] Podrobnejša predstavitev GOPOLIS-a je v pogl. 2.2.14, *Slovenske govorne zbirke*.

mislijo, da bi »vključevanje govorjenih besedil v korpus pomenilo tudi prvi korak v njihovo intenzivnejše raziskovanje« (Kranjc 1998, 111). Gorjanc pri napovedi nadgradnje Korpusa FIDA posebej izpostavi nujnost vključevanja transkripcij govora v korpus, še pred tem pa bi bilo po njegovem mnenju treba »izoblikovati metodološka izhodišča gradnje podkorpusa govora, izhajajoč iz deloma že oblikovanih predvsem za angleški jezik, vendar z nujnim upoštevanjem specifike slovenščine« (Gorjanc 1999, 55). Ideja o gradnji korpusa govorjenih besedil (KGB) je bila podrobneje predstavljena na 2. konferenci *Jezikovne tehnologije za slovenščino 2000*, oblikovanje KGB pa naj bi »prispevalo h kvalitetnejšim analizam slovenskega govorjenega jezika, po drugi strani pa tudi k razvoju govornih jezikovnih tehnologij« (Stabej in Vitez 2000, 79). Weiss ob predstavitvi načrta za gradnjo Slovenskega nacionalnega korpusa poudarja nujnost vključevanja govorjenih besedil: »Potreba po čim večji popolnosti elektronske zbirke bo spodbudila zbiranje tistih virov, ki zbiralcev in upraviteljev zbirk doslej niso zanimali. Tako so za jezikoslovno delo recimo zelo pomembni ustni viri« (Weiss 2001, 422). Gorjanc je bil v svoji disertaciji glede gradnje govornega korpusa še bolj konkreten: »Čim prej bi bilo treba oblikovati skupino, ki bi začela s pripravami govornega dela korpusa« (Gorjanc 2002, 71). Tudi v okviru dialektoloških študij je novo tisočletje vzbudilo pričakovanja po govornem korpusu: »Predvidena govorna zbirka slovenskih besedil /…/ in metodološka izhodišča za korpus govorjenih besedil /…/, ki obljubljajo širitev korpusa na spontani nejavni govor, vzbujajo upanje na drugačne čase /…/« (Kenda Jež 2004, 271).

Potreba po gradnji govornega korpusa je bila eksplicitno izražena tudi v okviru govornih tehnologij, in sicer v doktorski disertaciji, ki je nastala v sodelovanju Oddelka za slovenistiko ljubljanske Filozofske fakultete in mariborske Fakultete za elektrotehniko, računalništvo in informatiko: »Temeljna teza te disertacije /je/, da se je treba pri razvoju govornih tehnologij, ki bi uspešno procesirale pogovorni govor, nasloniti na tiste veje jezikoslovja, ki preučujejo spontan govorjeni diskurz, in to v vsakdanji jezikovni rabi« (Verdonik 2006, 40). Gradnja govornega korpusa za slovenščino je bila v slovenskem jezikoslovju prvič konkretneje načrtovana v okviru raziskovalnega projekta *Oblikovanje slovenskega korpusnega omrežja* (2004–2006), in sicer v smislu oblikovanja teoretičnih izhodišč za gradnjo korpusa; izhodišča so bila podrobneje predstavljena v disertaciji *Načela gradnje govornega korpusa slovenščine* (Zemljarič Miklavčič 2007). Leta 2008 je stekel projekt večjega obsega *Sporazumevanje v slovenskem jeziku* (2008–2013),[15] v okviru katerega so zagotovljena tudi sredstva za izgradnjo enomilijonskega govornega korpusa za slovenščino, kar pomeni dejanski začetek gradnje govornega korpusa. Preden je bilo mogoče začeti z gradnjo govornega korpusa, je bilo potrebno izoblikovati teoretična izhodišča za gradnjo. Posamični cilji uresničevanja namena so bili:

[15] http://www.slovenscina.eu/Vsebine/Sl/Domov/Domov.aspx

- določiti shemo za zajem besedil,
- določiti načela transkribiranja govora,
- izdelati priporočila za označevanje govornega korpusa in
- zgraditi učni korpus za testiranje postavljenih načel.

Z izpolnitvijo zastavljenih ciljev se s pričujočo razpravo vključujem v začetek postopkov za gradnjo referenčnega govornega korpusa na Sovenskem.

1.5 TEMELJNI POJMI

Preden začnem razpravljati o govornih korpusih, moram definirati nekaj temeljnih pojmov, kakor jih razumem v okviru te razprave. Prvi sklop se nanaša na izraze, povezane z govorjenim jezikom, drugi sklop pa na korpusno terminologijo. Posebej v prvem sklopu v slovenskem jezikoslovju poimenovanja in pojmovanja niso poenotena.

Govorjeni jezik razumem kot vrsto jezika, ki jo definira prenosnik, torej tisto, kar govorimo oz. slišimo, v nasprotju s pisnim jezikom. Vendar ta definicija v zvezi z govornimi korpusi ni dovolj precizna: opredeliti se je treba tudi do stopnje spontanosti govora. V zvezi z gradnjo referenčnega korpusa nas zanima predvsem *spontani govor*, pri čemer gre lahko za različne stopnje spontanosti, od nepripravljenega do pripravljenega govora. Branih besedil v tem smislu ne razumemo kot pravi govorjeni jezik, saj gre samo za oralizacijo pisnega jezika. Pri zbiranju in shranjevanju govorjenega jezika za namen gradnje referenčnega govornega korpusa branih besedil ne iščemo načrtno, če pa se v govorjenem jeziku pojavijo, jih posebej označimo.

Kot enoto govorjenega jezika uporabljam izraz *govorjeno besedilo*. Poimenovanje v slovenskem jezikoslovju ni novo, ni pa tudi vsestransko uveljavljeno. Zveza je bila uporabljena v naslovu prvega načrta za gradnjo govornega korpusa (Stabej in Vitez 2000, 79). Izraz prav tako že v naslovu navede Vogel (2004, 453),[16] uporabljen pa je bil tudi kot naslov enega izmed sklopov v zborniku Aktualizacija jezikovnozvrstne teorije na slovenskem (Kržišnik (ur.) 2004), in sicer v zvezi govorjeno besedilo/govorjeni diskurz. Da se ta dva pojma v slovenskem jezikoslovju pogosto zamenjujeta, opozarja že Kovačič (1994, 6): »V teoretičnih razpravah termin 'diskurz' tekmuje z 'besedilom'.« Kranjc (1996/1997, 307) npr. pri raziskovanju otroškega govora uporablja zvezo »govorjeni diskurz«, v zvezi z govornimi korpusi pa govori o »govorjenih besedilih« (Kranjc 1998, 109). Smolej (2006,

[16] Zanimivo je, da je izraz v kazalu zamenjan z "govorjeni diskurz".

16) uporablja »besedilo v spontanem govoru«, čeprav ga sopomensko zamenjuje tudi z izrazom diskurz: »/…/ si bomo ogledali vrste besedil (diskurzov), ki smo jih posneli /…/«; Zuljan Kumar (2005, 23) uporablja izraz »govorjeni diskurz«, pri čemer diskurz razume kot »komunikacijo v procesu«, besedilo pa kot »produkt tega procesa«. Verdonik (2006, 35) pa ločuje pisni in govorjeni diskurz, slednjega poimenuje »pogovor«. Sklenem lahko, da se izraza govorjeno besedilo in (govorjeni) diskurz v nekaterih definicijah pomensko prekrivata, vendar pa prvi izraz sodi bolj v domeno korpusnega jezikoslovja, drugi pa bolj v domeno analize diskurza.

Pojem »govorjeno besedilo« je pravzaprav težko definirati. Tognini Bonelli raziskuje razmerje med besedilom in korpusom. Ugotavlja, da glede na to, da je korpus zbirka besedil, korpusno analizo lahko razumemo kot raziskovanje jezika, kakor je ta realiziran v besedilih. To pomeni, da korpusno jezikoslovje izhaja iz istih predpostavk kot besediloslovje, da je namreč besedilo glavni nosilec pomena, a gre kljub temu za dva pristopa k raziskovanju, ki se razlikujeta v več pogledih (Tognini Bonelli 2001, 3):

BESEDILO	KORPUS
beremo [poslušamo] kot celoto	beremo [poslušamo] kot fragment
beremo horizontalno [poslušamo linearno]	beremo vertikalno [poslušamo izseke]
beremo [poslušamo] zaradi vsebine	beremo [poslušamo] zaradi vzorcev
beremo [poslušamo] kot enkratni dogodek	iščemo ponavljajoče se dogodke
beremo [poslušamo] kot individualno dejanje	beremo [poslušamo] kot primere družbene prakse
je primer parole	omogoča vpogled v langage
koherenten komunikacijski dogodek	nekoherenten komunikacijski dogodek

Tabela 1: Besedilo in korpus (Tognini Bonelli 2001, 3)

Pri definiranju besedila si lahko si pomagamo z eno izmed aktualnih definicij, npr. po Beaugrandu in Dresslerju ga označimo kot »kot komunikacijski dogodek, ki ustreza sedmim merilom tekstualnosti« (de Beaugrande in Dressler 1981, 3), ti pa so kohezija, koherenca, namernost, sprejemljivost, informativnost, situacijskost in medbesedilnost. Vendar ta definicija ne pove, na kakšen način je govorjeno besedilo »materializirano«. Transkribirano govorjeno besedilo je po definiciji TEI »transkripcija niza govorjenega besedila, ki ga je zaradi določenih razlogov mogoče imeti za samostojno enoto in ga obravnavati kot zaključeno besedilo« (prim. Johansson 1995, 86). Zanimivo je, da govori o »nizu« govorjenega besedila, kar je

kar ustrezno, pa še vedno zelo nedoločeno poimenovanje. Za potrebe te raziskave govorjeno besedilo v »materialnem smislu« lahko označimo kot dogodek, ki se začne, ko se govorjenje začne, in konča, ko se govorjenje neha, torej zaključen akustični dogodek.

Govorjena besedila se razlikujejo med seboj; tista s skupnimi lastnostmi se uvrščajo v isto *besedilno vrsto*. Poimenovanje mi v okviru raziskave pomeni skupino besedil, ki imajo dovolj skupnih lastnosti (glede strukture, namena, prenosnika, števila udeleženih govorcev, govornega položaja, okoliščin, vsebine itd.), da jih lahko uvrstimo v skupno vrsto; pomeni mi tudi nek končni element delitve, saj so znotraj besedilne vrste samo še posamezna besedila. Besedilne vrste so npr. intervju, okrogla miza (pogovor na izbrano temo), družabna konverzacija, predavanje, zdravica, pridiga itd. Vzporedno s pojmom besedilna vrsta se pojavlja tudi poimenovanje tip besedila; tako je npr. tudi pri Gorjancu (2005, 31), kjer se sicer v *Stvarnem kazalu* pojavlja samo *besedilna vrsta*, znotraj besedila pa tudi *tip besedila*: »BNC je v želji po vsestranski uravnoteženosti korpusa fiksiral razmerja med posameznimi tipi besedil /…/«. V raziskavi uporabljam obe poimenovanji, *vrsta* in *tip*, prvega v smislu skupine govorjenih besedil z več skupnimi lastnostmi, drugega pa pri razlikovanju govorjenih besedil glede na posamezni kriterij,[17] npr. monologi proti dialogom, telefonski pogovor proti pogovoru v osebnem stiku, formalno besedilo proti neformalnemu, javno proti nejavnemu ipd.[18]

Klasifikacijo govorjenih besedil s stališča poslušalca najdemo pri Vogel (2004, 461); razlikuje besedilne vrste glede na prenosnik, poslušalčevo vlogo, vnaprejšnjo pripravljenost besedila, poslušalčev namen in predvideni poslušalčev odziv. Taksonomija govorjenih besedil, narejena v smislu načrtovanja gradnje govornega korpusa, je že bila objavljena (Zemljarič Miklavčič 2004, 503–522); tam predlagane rešitve sem nekoliko spremenila in dopolnila.

V poimenovanjih, ki izhajajo iz korpusnega jezikoslovja, sledim uveljavljeni slovenski terminologiji (prim. Erjavec 1996/97, Gorjanc 2005, Vintar 2003 idr.). Za namen te raziskave je treba najprej definirati pojem *referenčnega korpusa*; to so korpusi, ki naj bi »predstavili celovito podobo nekega jezika« (Gorjanc 2005, 8). Predstavljajo izhodišče za temeljne jezikoslovne raziskave predvsem s področja slovnice in slovarja. So večjega[19] obsega, za njihovo gradnjo pa se predvideva mreža kriterijev za zajemanje raznoterih besedil glede na vrsto predvsem besedilo-

[17] Kakor lahko razlikovanje razumemo tudi pri Gorjancu (2005).

[18] V korpusu FidaPLUS (http://www.fidaplus.net/Info/Info_index.html) so besedila razvrščena glede na zvrst (neumetnostna in umetnostna besedila) in na tip (internetno, časopisno, knjižno, revialno gradivo).

[19] Velikost je težko natančneje opredeliti, poleg tega pa se razumevanje velikosti spreminja praktično iz dneva v dan. Za slovenske in svetovne razmere je danes velik referenčni korpus FidaPLUS, ki obsega 621 milijonov besed (pojavnic).

slovnih in sociolingvističnih kriterijev. Referenčni korpusi večinoma vključujejo tudi transkripcijo govora, sicer pa se *govorni korpusi* zaradi bistveno drugačne metodologije gradnje oblikujejo samostojno, največkrat kot podkorpusi referenčnih korpusov (govorni komponenti *BNC*, *The Bank of English*), lahko pa tudi kot samostojni korpusi (Nizozemski govorni korpus). To so transkribirani posnetki spontanega govora; za (referenčne) govorne korpuse so zanimive predvsem slovnično-leksikalne lastnosti (Gorjanc 2005, 8), prozodične pa le, če je korpus ustrezno označen; ne gre torej za korpuse, namenjene raziskovanju fonetičnih lastnosti govora, ampak za zajetje posebnosti govorne komunikacije (Atkins in drugi 1992, 2). Korpusi za potrebe fonetično-fonoloških raziskav in do pred kratkim tudi govornih tehnologij nastajajo drugače, kot studijski posnetki, zajemajo pa največkrat samo izbrane (in prebrane) stavke; imenujemo jih korpusi govora ali *govorne zbirke* (Gorjanc 2005, 8).

Temeljni pojem, povezan z referenčnimi korpusi v jezikoslovju, je *reprezentativnost* korpusa. Gre za vprašanje, kako določiti in uravnotežiti količino raznoterih besedil v korpusu, da bo korpus čim boljši približek celovite podobe jezika, ki ga želimo opazovati. Načela za zajemanje besedil v korpus za doseganje čim večje reprezentativnosti in uravnoteženosti se pri različnih korpusih razlikujejo in bodo predstavljena v nadaljevanju.

Izraza *pojavnica* in *različnica* uporabljam v skladu s terminologijo, uveljavljeno pri Gorjancu (2005, 11): pojavnica (token) je vsaka izrazna enota, ki se pojavi v korpusu, različnica (type) pa vsaka beseda, ki se pojavi v korpusu. V korpusu govorjenih besedil pričakujemo nekatere različnice, ki jih v pisnem korpusu ni, poleg tega pa med različnice uvrščam tudi polverbalne izraze (mhm, ə).

V razpravi večkrat uporabljam kratico *UKGS* za Učni korpus govorjene slovenščine, kratico *KGS* pa uporabljam za korpus govorjeni besedil, in sicer za govorni podkorpus referenčnega korpusa slovenskega jezika. Vsi navedki posnetkov oz. transkripcij, ki se pojavljajo v besedilu, so vzeti iz Učnega korpusa, zapisani pa so v tipografiji Curier, da se že na pogled ločijo od ostalega besedila.

2 Govorni korpusi

2.1 UVOD

Težko je natančno ugotoviti, kako daleč v zgodovino jezikoslovja se moramo ozreti, če hočemo odkriti prve predhodnike govornih korpusov. John Sinclair je leta 1960 vodil gradnjo prve zbirke govorjenih besedil velikosti 220.000 besed, ki je bila shranjena na računalniku (Sinclair 1995, 99),[20] kar sam postavlja na začetek korpusnega jezikoslovja. Glede na to, da je bil prvi računalniški korpus (pisnega jezika) Brown zgrajen leta 1960, je moralo preteči še natanko 20 let do gradnje prvega govornega korpusa (korpus London-Lund, 1980). Osemdeseta in prva polovica devetdesetih let so bila zaznamovana z velikim napredkom in razvojem na področju korpusnega jezikoslovja (tudi zaradi tehnološkega razvoja), zgrajeni so bili nekateri znameniti govorni korpusi, oblikovane delovne skupine, ki so načrtovale enotno označevanje korpusov, vzporedno pa je raslo tudi število bibliografskih enot s področja korpusnega jezikoslovja. Zgrajeni so bili govorni korpusi za vse različice angleškega jezika, izdelana priporočila za označevanje korpusnih dokumentov ter izdelani in preizkušeni programi za avtomatsko in polavtomatsko označevanje korpusov. Ko je bilo področje angleščine z govornimi korpusi že dobro »pokrito«, je nastopilo obdobje, ko so začeli nastajati govorni korpusi tudi za druge jezike. Številne jezikoslovne skupine na nacionalni ali večnacionalni ravni so se združile ob velikih korpusnih projektih. Neangleška jezikovna okolja, predvsem tista, kjer je jezik tudi sredstvo nacionalne identitete, so pogosto zaznamovana še z drugačnim razmerjem do jezika; tu je lahko pisna norma še bolj določujoča in je zato gradnja govornega korpusa še težavnejša. Kot bomo videli v nadaljevanju, je novejšim načrtovalcem korpusov časovni zaostanek petih ali desetih let prinesel tudi veliko tehnološko prednost: v tem času je postalo mogoče in celo dokaj enostavno sinhronizirati zvok s transkripcijo. Večina govornih korpusov za angleščino se sedaj sooča z novimi razmerami, v katerih so njihovi korpusi korak za najsodobnejšimi; videli bomo, da težave poskušajo reševati na različne načine.

V nadaljevanju so predstavljeni nekateri večji oz. bolj znani (dostopni) delujoči govorni korpusi (navedeni so v zaporedju glede na začetek gradnje):

1.	Korpus London-Lund (britanska angleščina)
2.	Korpus Lancaster/IBM, kasneje MARSEC (britanska angleščina)
3.	Korpus Santa Barbara (ameriška angleščina)
4.	Korpus ICE (Mednarodni korpus angleščine)
5.	Govorna komponenta korpusa BNC (britanska angleščina)

[20] Kasneje je bil vključen v Cobuildov korpus.

6.	Govorna komponenta korpusa *Bank of English* (britanska angleščina)
7.	Govorna komponenta korpusa ČNK (češčina)
8.	Budimpeštanski sociolingvistični intervjuji (madžarščina)
9.	Korpus COLT (britanska angleščina najstnikov)
10.	Švedski govorni korpus
11.	Nizozemski govorni korpus
12.	C-ORAL-ROM (korpusni paket francoščine, italijanščine, španščine in portugalščine)

Predstavljeni bodo našteti korpusi – njihova velikost, potek gradnje, namen, sestava, transkripcijska načela, dostopnost in uporabnost. Nekatere komponente bodo predstavljene tudi v poglavjih 3, *Zajem besedil v govorni korpus* in 4, *Transkribiranje in označevanje govornih korpusov*.

2.2 GOVORNI KORPUSI

2.2.1 Korpus govorjene angleščine London-Lund

Korpus London-Lund (*London-Lund Corpus*, LLC) je najstarejši govorni korpus in prva računalniška zbirka govorjenih besedil. Nastal je na podlagi korpusa *Survey of English Usage* (SEU), ki sodi med prve korpuse sploh, še v obdobje predračunalniških besedilnih zbirk. Korpus SEU je l. 1959 zasnoval Randolph Quirk, sestavlja pa ga 200 enako dolgih besedil, 100 pisnih in 100 govornih, skupaj 1 milijon besed. Korpus je bil namenjen za preučevanje govorjene in pisane britanske angleščine odraslih govorcev in je služil kot vir za slovnični opis britanske angleščine.

Govorjena besedila korpusa SEU so bila transkribirana, nato pa sta bili obe komponenti (govorna in pisna) prozodično in oblikoskladensjko označeni. Vsa besedila so bila nato razdeljena na listke s po 17 vrsticami označenega besedila, s po štirimi vrsticami prekrivnimi s predhodnim listkom in štirimi prekrivnimi z naslednjim listkom (predračunalniški korpus!). Zaradi različnih težav, predvsem pa zaradi zahtevnosti transkripcije je bilo delo v celoti dokončano šele konec osemdesetih let.

Leta 1975 je Jan Svartvik na univerzi v Lundu na Švedskem začel sestrski projekt korpusa SEU. Njegov namen je bil govorno komponento korpusa SEU prene-

sti v digitalno obliko, primerno za računalniško branje. 87 do tedaj zbranih in transkribiranih besedil korpusa SEU je bilo prenesenih v računalniško obliko in v začetku leta 1980 je računalniška verzija govorne komponente korpusa SEU – korpus London-Lund – zakrožila med zainteresiranimi znanstveniki po celem svetu. Kasneje je bilo dodanih še preostalih 13 besedil, tako da popolna verzija (LLC:c)[21] vsebuje 100 transkribiranih besedil govornega dela korpusa SEU. Gradivo korpusa London-Lund sedaj obstaja v treh različicah: na listkovnem gradivu korpusa SEU, na CD-romu in v knjižni[22] obliki.

Naslednja shema prikazuje delež različnih tipov besedil v osnovni verziji korpusa London-Lund (87 besedil):

TIP BESEDILA	Dialog	Neposredni stik govorcev	Zasebno	Posneto brez vednosti govor.	Radio in TV	Št. besedil	Št. vseh besedil
Neposredni pogovor	+	+	+	+	–	34	46
	+	+	+	–	–	12	
Telef. pogovor	+	–	+	+	–	10	10
Diskusije, intervjuji, debate	+	+	–	–	+	12	12
Javni nepripr. govor	+	+	–	–	–	3	12
Komentar	–	+	–	–	+	2	
Demonstracija	–	–	–	–	+	7	
Javni pripravljeni govor	–	+	–	–	–	7	7
Skupaj	**71**	**70**	**56**	**44**	**21**	**87**	**87**

Tabela 2: Število besedilnih tipov v prvi (nepopolni) verziji korpusa London-Lund[23]

[21] *The complete London-Lund Corpus*, celotni korpus London-Lund.
[22] *A Corpus of English Conversation,* ur. J. Svartvik in R. Quirk. Lund Studies in English, Lund: Liber/Gleerups, 1980. 893 strani.
[23] http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

Na podlagi SEU je nastalo več kot 200 strokovnih in znanstvenih razprav,[24] med drugim tudi slovnica angleškega jezika *A comprehensive grammar of the English language*.[25]

2.2.2 Korpus govorjene angleščine Lancester/IBM (MARSEC)

Leta 1984 se je začel projekt gradnje govornega korpusa britanske angleščine (*Spoken English on Computer*, SEC), in sicer v sodelovanju univerze v Lancastru in znanstvenega centra za raziskave govora podjetja IBM. Zgrajen je bil razmeroma majhen korpus, ki vsebuje nekaj čez 52.000 besed govorjene britanske angleščine, vendar je v času svojega nastanka predstavljal velik napredek na področju jezikovnih tehnologij. Korpus je nastal z namenom, da bi služil predvsem raziskavam na področju sinteze in analize govora. Namen uporabe korpusa vedno narekuje njegovo sestavo in oznake, zato so bila besedila zbrana tako, da bi predstavljala dober model za raziskavo govora: velik delež besedil predstavljajo posnetki z radia BBC. Če se posamezni posnetek ni dobro slišal ali je vseboval veliko t. i. prekrivnega govora (ko govori več govorcev hkrati), besedila niso uvrstili v korpus. Izločeni so bili tudi posnetki govorcev, ki so imeli prepoznaven narečni naglas. Razmerje med moškimi in ženskami ni bilo uravnoteženo, govorci pa so bili predvsem radijski napovedovalci in profesorji. Kasneje se je korpus dejansko uporabljal za raziskave govora v IBM-ovih raziskovalnih laboratorijih, pa tudi kot učno gradivo pri študiju fonetike na univerzi v Lancastru.[26]

Besedilnovrstna sestava in način označevanja besedil sta sledila standardom, ki sta jih postavila dva starejša korpusa, korpus Brown in korpus Lancaster/Oslo/Bergen (LOB). Korpus Brown je enomilijonska zbirka pisnih besedil ameriške angleščine, namenjen pa je bil jezikoslovnim raziskavam, kar je za označevanje bistvenega pomena. Korpus LOB je britanska zrcalna podoba korpusa Brown, dograjen pa je bil l. 1978 v sodelovanju univerz v Lancastru in Oslu ter Centra za računalniške tehnologije v Bergnu. Oba korpusa, Brown in LOB, sta sestavljena iz 500 besedilnih vzorcev s po 2000 besedami. Korpus Lancaster/IBM je, kljub temu, da je govorni korpus, poskušal prevzeti besedilne kategorije korpusa LOB, kolikor je bilo to mogoče.

V korpus Lancaster/IBM so bila govorjena besedila uvrščena v naslednji sestavi in razmerju:[27]

[24] http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
[25] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik. London, New York: Longman, 1985.
[26] http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM
[27] ICAME Collection (CD-ROM v lasti Univerze v Bergnu).

Besedilna vrsta	%
Komentar	17
Novica	10
Predavanje, namenjeno širšemu občinstvu	8
Predavanje, namenjeno specializiranemu občinstvu	14
Versko radijsko besedilo	3
Finance	9
Literarna proza	14
Poezija	2
Dialog	13
Reklama	3
Razno	6

Tabela 3: Zgradba govornega korpusa Lancaster/IBM

Kot je razvidno, dialoška besedila zavzemajo samo 13 odstotkov vseh besedil, kar bi bilo za potrebe referenčnega korpusa nesprejemljivo, saj dialog predstavlja večinski del govorjenih besedil, za fonetične študije pa je dialog zaradi prepletenosti in prekrivanja izjav manj primeren. Vseh govorjenih besedil v korpusu je 53, dolga pa so od 1 do 24 minut (to je pomembna razlika glede na korpus London-Lund, kjer so bila vsa besedila približno enako dolga, kar je bilo treba deloma tudi umetno skonstruirati, torej besedila »rezati«). Večina besedil je označenih z datumom posnetka, običajno so z imenom in priimkom navedeni tudi govorci, poleg tega pa so bile urejene avtorske pravice z BBC in z vsemi sodelujočimi govorci; tudi metodologijo urejanja avtorskih pravic so povzeli po korpusu Brown (Kennedy 1998, 27).

Za jezikoslovne raziskave je korpus Lancaster/IBM še danes zanimiv predvsem zaradi petih različnih verzij, v katerih obstaja: zvočni posnetki, transkripcija brez ločil, ortografska transkripcija, prozodična transkripcija in oblikoskladenjsko označena verzija; tri različne transkripcije korpusa Lancaster/IBM bodo podrobneje predstavljene v poglavju 4, *Transkribiranje in označevanje govorjenih besedil*.

Besedila korpusa so bila slovnično označena z označevalnikom CLAWS1 (kasneje z verzijo CLAWS2), izdelanim za korpus LOB. Program je bil narejen za označevanje pisnih besedil, zato je bil lahko uporabljen samo na ortografski transkripciji govorjenih besedil korpusa. Program je deloval v petih korakih, pri čemer so bili vsi razen zadnjega avtomatski:[28]

[28] ICAME Collection (CD-ROM v lasti Univerze v Bergnu).

1. Besedilo je najprej konvertiral v navpični format, tako da je bila vsaka beseda (tudi ločilo) v svoji vrsti.
2. Vsem besedam je pripisal slovnične oznake, ki bi jim lahko ustrezale (oznake je program pripisal tako, da je besedo poiskal v leksikonu; če besede ni našel, jo je uvrstil na poseben seznam).
3. Tretja faza je bilo označevanje stalnih besednih zvez; če je program tako strukturo prepoznal (našel v leksikonu), jo je označil kot idiom (namesto vsake besede posebej).
4. Četrta faza je bila disambiguacija; program je v primeru, da je bilo besedi pripisanih več slovničnih oznak, izbral pravo, in sicer na podlagi sobesedilnega okolja.
5. V zadnji fazi je bilo treba vse oznake ročno pregledati in popraviti napake.

Slovnično označena besedila korpusa Lancaster/IBM so bila shranjena v dveh oblikah, navpični in vodoravni. Navpična verzija zavzema mnogo več prostora pri shranjevanju (kar je v osemdesetih letih še predstavljalo problem), je pa veliko bolj uporabna pri nadaljnjih analizah. Navpična in vodoravna verzija sta predstavljeni na spodnjih izsekih (številke označujejo kategorijo in številko besedila, nadalje vrstico, velike tiskane črke pa v obeh verzijah slovnično oznako):

A01	5	010	JJ	good
A01	5	020	NN	morning
A01	5	021	.	.
A01	5	022	---	----------
A01	5	030	AP	more
A01	5	040	NN	news
A01	5	050	IN	about
A01	5	060	ATI	the
A01	5	070	NPT	Reverend
A01	5	080	NP	Sun
A01	5	090	NP	Myung
A01	5	100	NP	Moon
A01	5	101	,	,
A01	5	110	NN	founder
A01	6	010	IN&	of
A01	6	020	ATI	the
A01	6	030	NNP	Unification

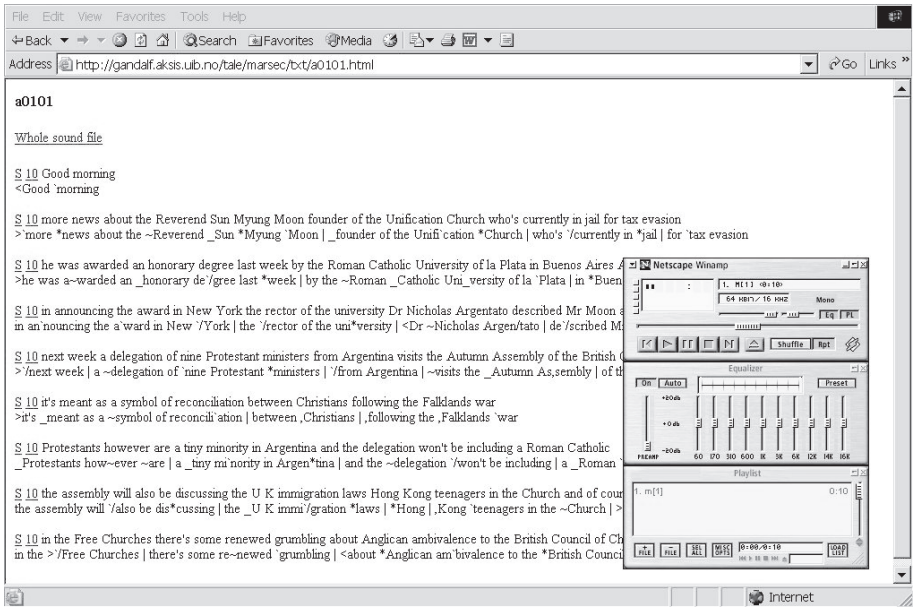
Tabela 4: Navpična verzija slovnično označenega besedila korpusa Lancaster/IBM[29]

[29] ICAME Collection (CD-ROM v lasti Univerze v Bergnu).

A01	5 ^ good_JJ morning_NN ._. ^ more_AP news_NN about_IN the_ATI
A01	5 Reverend_NPT Sun_NP Myung_NP Moon_NP ,_, founder_NN
A01	6 of_IN the_ATI Unification_NNP church_NN ,_, who_WP 's_BEZ
A01	6 currently_RB in_IN jail_NN for_IN tax_NN evasion_NN :_:

Tabela 5: Vodoravna verzija slovnično označenega besedila korpusa Lancaster/IBM[30]

Korpus Lancaster/IBM je bil v letih 1992–1994 nadgrajen; narejena je bila digitalizacija avdio-posnetkov, ti pa so bili shranjeni na CD-ROM. Prvotna verzija namreč ni vsebovala akustičnih posnetkov oz. ti raziskovalcem niso bili dostopni. Za novo izdajo korpusa, ki se je takrat tudi preimenoval v MARSEC (*Machine Readable Spoken English Corpus*), so prvotni verziji dodali zvočne posnetke in program za sinhronizacijo zvoka in transkripcij (Knowles 1995, 209). Tako jim je drugi najstarejši govorni korpus uspelo aktualizirati in prilagoditi sodobnemu tehnološkemu razvoju.



Slika 2: Govorni korpus MARSEC (nadgrajeni Lancaster/IBM)[31]

[30] ICAME Collection (CD-ROM v lasti Univerze v Bergnu).
[31] http://gandalf.aksis.uib.no/tale/marsec/txt/a0101.html

2.2.3 Korpus govorjene ameriške angleščine

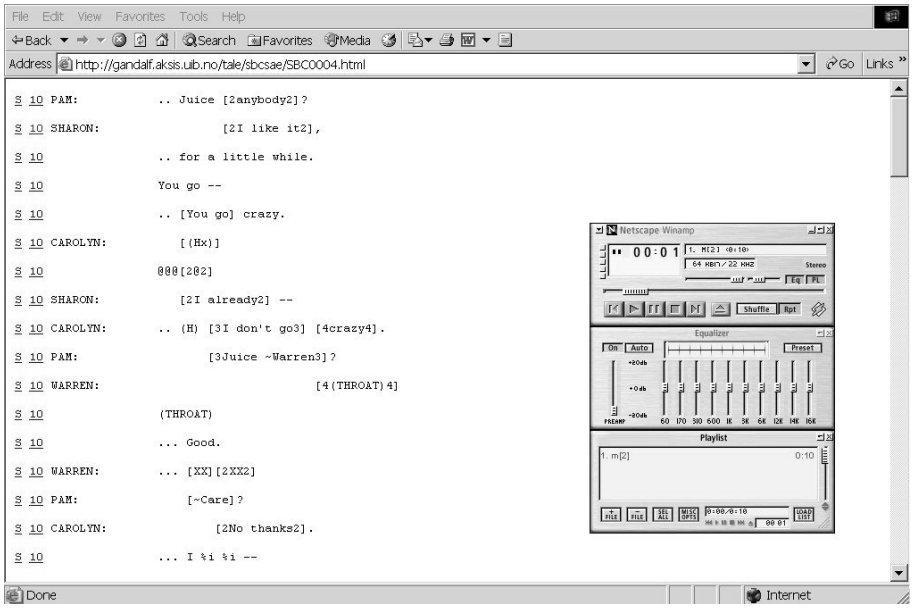
Konec osemdesetih let se je začel projekt gradnje korpusa govorjene ameriške angleščine, ki naj bi približno ustrezal korpusu London-Lund, s čimer bi tudi na področju govora dobili primerljiva korpusa za britansko in ameriško angleščino. Korpus je začel nastajati na Univerzi Santa Barbara v Kaliforniji. Pri projektu so sodelovali trije raziskovalci, korpus pa naj bi bil izdelan v treh letih. Njegovo velikost je narekovala takratna zmogljivost CD-roma, saj naj bi bil to nosilec korpusa (in sicer zvočnih posnetkov in transkripcije). Po takratnih izračunih bi lahko na CD-rom shranili do 15 ur zvočnih posnetkov in njihovo transkribirano verzijo ter konkordančnik, to pa naj bi pomenilo velikost korpusa pribl. 200.000 besed. Načrtovalci so nameravali vsa transkribirana besedila objaviti tudi v knjigi (Chafe idr. 1991, 66 in naprej) .

Korpus naj bi sestavljalo 30 polurnih posnetkov sporazumevanja v »standardni ameriški angleščini«. Pojem se je nanašal na jezikovne različice, ki se govorijo v formalnih in neformalnih položajih (pogovarjanje, kreganje, opravljanje, pogovori na delovnem mestu, sestanki, predavanja, politični govori, pravljice za lahko noč, pogrebi, poroke itn.), v različnih okoliščinah, z različnimi regionalnimi naglasi; združujejo jih skupna slovnična pravila in strukture, ki so med vsemi govorjenimi različicami ameriške angleščine še najbližje pisnim besedilom, čeprav se od njih še vedno razlikujejo na sto in en način (Chafe in drugi 1991, 68). Ker se pojem standardna angleščina nanaša na številne jezikovne variante, ne pripada nobeni posebni skupini govorcev. Govorce so nameravali izbrati tako, da bi pokrivali različne kategorije spola, starosti, etnične pripadnosti, poklica in izobrazbe, torej različne demografske kategorije.

Korpus je sestavljen izključno iz dialoških besedil, posnetih z vednostjo govorcev. Na ta način je postal primerljiv s podkorpusom korpusa London-Lund, »konverzacija v osebnem stiku«, ki obsega 250.000 besed. Primerljiva bi bila tudi po transkripciji, saj je bila tudi za korpus govorjene ameriške angleščine predvidena prozodična transkripcija z označenimi tonskimi enotami, tonskim potekom, premori in naglasi. Vsaka tonska enota naj bi bila povezana z ustreznim zvočnim posnetkom in na zahtevo takoj dostopna, kar bi bil pomemben korak v razvoju korpusnega jezikoslovja.

Leta 2000 je združenje LDC (*Linguistic data consortium*) objavilo, da je za 75 dolarjev mogoče kupiti prvi del korpusa ameriške angleščine na treh CD-romih. Prvi del korpusa sestavlja 14 govornih datotek, dolgih med petnajst in trideset minut. Vsaka datoteka ima pripeto ustrezno transkripcijo, v kateri so posamezne

izgovorjene fraze časovno označene in na ta način povezane z zvočnim posnetkom. Na prvem CD-ju korpusa Santa Barbara je torej že bila izvedena sinhronizacija zvoka in transkripcije. Osebna imena, naslovi in telefonske številke so bili zaradi zagotavljanja anonimnosti govorcev izbrisani tako iz transkripcij kot z zvočnih posnetkov, kjer so bili filtrirani, tako da so postali neprepoznavni.



Slika 3: Govorni korpus Santa Barbara[32]

Leta 2003 je izšel drugi del korpusa Santa Barbara (1 DVD nosilec), leta 2004 tretji (1 DVD nosilec) in 2005 četrti del.[33] Del korpusa Santa Barbara je tudi del Mednarodnega korpusa angleščine (ICE) in predstavlja njegovo ameriško komponento.

2.2.4 Mednarodni korpus angleščine (ICE)

Projekt Mednarodni korpus angleščine (*International Corpus of English*),[34] ki se je prav tako začel konec osemdesetih let, je predvideval izdelavo paralelnih enomilijonskih korpusov angleščine v različnih državah, kjer je angleščina uradni jezik

[32] Štiri dele korpusa Santa Barbara je že mogoče kupiti; verzija na sliki je v lasti Univerze v Bergnu (http://gandalf.aksis.uib.no/tale/sbcsae/SBC0004.html).

[33] http://www.linguistics.ucsb.edu/research/sbcorpus.html

[34] http://www.ucl.ac.uk/english-usage/ice/

(ali eden izmed uradnih jezikov). Pobudnik projekta je bil Sidney Greenbaum, ki je l. 1975 prevzel uredništvo korpusa SEU in gradnjo prvega (neračunalniškega) govornega korpusa pripeljal do konca. Potreba, ki je narekovala gradnjo mednarodnega korpusa angleščine, je izhajala iz dejstva, da takrat še ni bilo korpusa govorjene ameriške angleščine; zaradi tega so bile mogoče samo primerjave med govorjeno in pisno britansko angleščino ter pisnima britansko in ameriško angleščino, ne pa primerjave med britansko in ameriško govorno ter ameriško govorno in pisno angleščino. Začetna ideja je bila torej zgraditi paralelna govorna korpusa – britanskega in ameriškega. Pri tem korpus London-Lund ni mogel več služiti za britansko govorno različico, saj so bila besedila v njem prestara – njihovo zbiranje se je začelo že v petdesetih letih – za primerjalne študije pa so načrtovalci korpusa potrebovali sodobna besedila, zbrana približno v enakem času. Ker sta v tem času že obstajala še dva korpusa po vzoru korpusa Brown, korpus pisne avstralske angleščine (korpus Macquarie) in korpus angleščine v Indiji (korpus Kolhapur), povsod pa je manjkala govorna komponenta, je prvotni načrt prerasel v idejo, da bi k sodelovanju povabili še druge zainteresirane. V projektu ICE se je združilo 15 nacionalnih timov: VB, ZDA, Južna Afrika, Avstralija, Kanada, Vzhodna Afrika (Kenija, Tanzanija in Zambija), Hong Kong, Indija, Irska, Jamajka, Malezija, Nova Zelandija, Filipini, Singapur in Šri Lanka (Greenbaum 1996, 8).

L. 1989 oblikovana svetovalna skupina korpusa ICE je izoblikovala osnovna načela gradnje korpusa: vsak nacionalni korpus naj bi bil enomilijonski, in sicer pisni in govorni, sestavljalo pa naj bi ga 500 besedil s po pribl. dva tisoč besedami (besedil ne bi rezali natančno pri tej meji, ampak v bližini, na kakšni primerni vsebinski zarezi). V posamezni besedilni kategoriji bi moralo biti najmanj 10 besedil (20.000 besed), kasneje pa je bilo dodano še načelo, da je 300 besedil govornih, 200 pa pisnih. Če bi kakšna skupina v svoj nacionalni korpus želela vključiti besedila v celoti ali dodati kakšno kategorijo besedil, ki je ICE ni določil, je bilo to mogoče, vendar s tistim delom korpusa ne bi mogli vstopiti v skupni ICE. Ob tem naj bi nastajali tudi trije vzporedni korpusi: (1) korpus prevodov v angleščino iz jezikov EU, (2) govorni korpus mednarodne komunikacije, v kateri sodelujejo govorci, katerih prvi jezik ni angleščina, in (3) besedila, ki se uporabljajo za učenje angleščine kot tujega jezika (Greenbaum 1991, 85).

V zvezi z vrstami govorjenih besedil je bilo dogovorjeno, da bodo sestavili korpuse t. i. izobraženske angleščine: govorci naj bi bili stari najmanj 18 let in naj bi imeli dokončano najmanj srednjo šolo, izobraževali pa naj bi se v angleščini. S tem naj bi dosegli vsaj približno usklajenost in primerljivost vseh korpusov. Vključevali bi dialoška in monološka govorjena besedila, posneta v osebnem stiku ali po prenosniku (telefon). Vrsta vprašanj pa se je med pridruženimi strokovnjaki

sprožala tudi v zvezi z načeli transkripcije korpusa ICE oz. njegovih nacionalnih enot. Navsezadnje je prevladala odločitev za ortografsko transkripcijo (brez ločil, a z označevanjem krajših in daljših premorov), čeprav se je koordinator projekta Sidney Greenbaum zavzemal za prozodično transkripcijo. Tovrstno transkribiranje bi bilo pri tolikšnem številu sodelujočih le težko konsistentno, saj je bilo že iz preteklih izkušenj jasno, da tudi po dva usklajena fonetika označujeta neenotno. Dogovorjeno pa je bilo, da bodo ob transkripcijah na voljo tudi zvočni posnetki, tako da bodo na korpusih mogoče nadaljnje prozodične ali fonetične obdelave.

Gradnja mednarodnega korpusa angleščine poteka počasneje, kot je bilo predvideno. Po smrti Sidneya Greenbauma projekt vodi Gerald Nelson, leta 2008 pa je bilo v okviru projekta zgrajenih in dostopnih 7 korpusov – korpusi Hong Konga, Vzhodne Afrike, Velike Britanije, Indije, Nove Zelandije, Filipinov in Singapurja, dostopni pa so na CD-romih in deloma na internetu.[35]

2.2.5 Govorna komponenta Britanskega nacionalnega korpusa

Konec osemdesetih in v začetku devetdesetih let so v Veliki Britaniji zasnovali korpusni projekt velikih razsežnosti, Britanski nacionalni korpus BNC (*British National Corpus*).[36] S strani britanske vlade je prišla pobuda za gradnjo jezikovnih virov za angleščino, pri tem pa naj bi se glede na deloma prekrivne interese spodbudilo akademsko-industrijsko sodelovanje. K projektu so pristopile univerzitetne institucije in komercialni partnerji (Oxford University Press (vodilni partner), Longman Group UK Ltd., Chambers Harrap, Britanska knjižnica, Univerza v Oxfordu in Univerza v Lancastru), vlada pa je z različnih oddelkov za projekt v treh letih namenila 1,5 milijona britanskih funtov, kar naj bi zadoščalo za celotno kritje znanstveno-raziskovalnega dela in za 50-odstotno kritje dela komercialnih partnerjev (BNC User Reference Guide 2000, 2). 100-milijonski korpus je bil zgrajen v letih 1990–1994.

Zahtevna gradnja korpusa BNC, »proizvodna linija« (»*BNC Sausage machine*«, Burnard 2000, 3), je tekla tako, da je bila po odsekih razdeljena med institucije in partnerje: pisna besedila so zbirali v založbah Oxford University Press in Chambers, besedila za govorni del v založbi Longman, preoblikovanje besedil v enotno računalniško obliko je potekalo v raziskovalnem centru Univerze v Oxfordu, slovnično označevanje na Univerzi v Lancastru (tudi pri BNC z označevalnikom

[35] http://www.ucl.ac.uk/english-usage//ice/index.htm
[36] http://www.natcorp.ox.ac.uk/

CLAWS, ki je bil uporabljen že za korpus Lancaster-Oslo-Bergen in Lancaster/IBM), končno skupno generiranje oznak pa spet na Univerzi v Oxfordu. Pri gradnji britanskega nacionalnega korpusa so uporabili vse znanje in izkušnje, pridobljene ob prejšnjih korpusnih projektih; to in pa dejstvo, da je večino sredstev prispevala britanska vlada, je tudi opravičevalo besedo »nacionalni« v imenu korpusa (BNC User Reference Guide 2000, 1). BNC ob svojem nastanku ni bil več največji britanski korpus, saj ga je v tem smislu dohitel korpus Bank of English (BoE), vendar je imel tudi BNC nekaj prednosti: bil je skrbneje uravnotežen in dostopen širokemu krogu uporabnikov. Ti lastnosti skupaj z velikostjo korpusu BNC še danes zagotavljata posebno mesto med obstoječimi korpusi.

BNC vsebuje pribl. 10 odstotkov govorjenih besedil, to je podkorpus v velikosti 10 milijonov besed. Razmerje med govornimi in pisnimi besedili je bilo določeno na podlagi ekonomske logike, saj so izračunali, da staneta zbiranje in transkripcija enega milijona besed spontanega govora vsaj desetkrat več kot priključitev enega milijona besed iz časopisa v pisni korpus. Tako govorna kot pisna komponenta korpusa izkazujeta visoko uravnoteženost tipov besedil, ki bo predstavljena v nadaljevanju, vendar pa ravno zaradi tega tudi statično sliko jezika iz določenega obdobja (1991–1994); dandanes gredo težnje in zahteve v jezikoslovju v smeri večje dinamičnosti korpusa (priključevanja novih, časovno aktualiziranih besedil), tudi na račun uravnoteženosti, vendar pa tudi znotraj dinamičnih korpusov obstajajo uravnoteženi podkorpusi tipa BNC za posebne raziskave.

Pri zajemanju besedil so načrtovalci korpusa BNC prvič uporabili metodo demografske klasifikacije govorcev, ki so jo prevzeli iz sociоloških raziskav, in jo kombinirali s kontekstualno metodo zbiranja (BNC User Reference Guide 2000, 4.1), uporabljeno že pri gradnji starejših korpusov; obe metodi bosta predstavljeni v poglavju 3, *Zajem besedil v govorni korpus.*

Britanski nacionalni korpus je bil v devetdesetih letih prejšnjega stoletja najvplivnejši referenčni vir za gradnjo korpusov. S svojimi načeli gradnje ter z industrijskimi razsežnostmi produkcije je postavil temelje novemu obdobju, v katerem so se jezikovne tehnologije premaknile z akademskega obrobja v jedro informacijske družbe (Burnard 2000, 1).

2.2.6 Govorna komponenta korpusa *The Bank of English*

Začetek gradnje Banke angleščine (The Bank of English, BoE)[37] sega v osemdese-

[37] http://www.collins.co.uk/books.aspx?group=153

ta leta prejšnjega stoletja, ob njem pa sta se združila založniško podjetje Collins in Univerza v Birminghamu. Načrt je predvideval gradnjo 200-milijonskega pisnega in govornega korpusa, projekt pa je vodil John Sinclair. Glavnina označevanja je potekala v letih 1993–1994. Tudi v tem primeru je realizacija prerasla načrte, saj je spletna stran Banke angleščine leta 2005 dosegla velikost 524 tisoč besed. Vendar pa so dostopni podatki o korpusu zelo skromni; ni mogoče razbrati, kolikšen delež besedil predstavljajo govorjena besedila, prav tako ni taksonomije besedil, zajetih v korpus, kar je npr. v popolnem nasprotju z obsežno javno dostopno dokumentacijo korpusa BNC. Iz kasnejših Sinclairjevih izjav je znano, da je bil sam zagovornik kvantitete v korpusu, v smislu, da velika količina podatkov sama po sebi zagotavlja uravnoteženost korpusa.[38] Korpus je na voljo predvsem leksikografom in jezikoslovcem založbe Collins, del korpusa pa je prosto dostopen tudi na internetu *(Collocation and Concordance Demostration)*;[39] ta del korpusa dosega velikost 56 milijonov besed, sestavljajo pa ga trije podkorpusi: britanska pisna in brana angleščina (36 milijonov), ameriška pisna in brana angleščina (10 milijonov) in britanska govorjena angleščina (10 milijonov besed). Sklepamo lahko, da je govorna komponenta Collinsovega korpusa najmanj tako velika kot govorna komponenta BNC. Na podlagi korpusa, ki je predvsem zaradi založniških interesov skrbno čuvana skrivnost, je med drugim v uredništvu Johna Sinclaira nastal prvi na korpusu temelječi slovar angleškega jezika (19871).[40]

2.2.7 Češki govorni korpusi

Med slovanskimi narodi so gradnjo velikega jezikovnega korpusa prvi realizirali Čehi. Pri načrtovanju novega slovarja češkega knjižnega jezika (zadnji je nastal v letih 1960–1971) so se raziskovalci zavedali, da s staro metodo izpisovanja na listke ne bodo mogli zaobjeti jezika zadnjih tridesetih let, zato se je v začetku devetdesetih let rodila ideja o gradnji računalniškega korpusa (Čermák 1997, 186). L. 1994 je bil na Filozofski fakulteti v Pragi ustanovljen Oddelek za Češki nacionalni korpus (ČNK), kar je pomenilo tudi odlično osnovo za razvoj korpusnega jezikoslovja kot posebne znanstvene discipline. Oddelek je prerasel v Inštitut za ČNK, v katerem je sodelovalo devet raziskovalnih in izobraževalnih institucij iz Prage in Brna; izoblikoval se je znanstveno-raziskovalni tim, ki se je v marsičem zgledoval po korpusnih centrih in timih v tujini (Čermák 1997, 187), hkrati pa razvijal lastno raziskovalno dejavnost in gradil korpus. Oteževalna okoliščina je bila, da so bili pridruženi partnerji zgolj akademski, z eno samo izjemo založniške hiše, kar je predstavljalo veliko oviro pri zbiranju potrebnih sredstev; večino

[38] »I have an important policy in spoken language collection /…/, and that is to put quantity first« (Sinclair 1995, 101).
[39] http://www.collins.co.uk/Corpus/CorpusSearch.aspx
[40] Krek 2004, 4.

je prispevala država, del tudi omenjena založniška hiša, sicer pa, po Čermáku, v poslovnem svetu ni bilo interesa za gradnjo korpusa.

Pri gradnji Češkega nacionalnega korpusa so se raziskovalci deloma oprli na smernice v znamenitem članku *Corpus Design Criteria* (Atkins et al. 1992), deloma pa so razvili lastne kriterije glede na specifične lastnosti češkega jezika in glede na razpoložljiva sredstva. Korpus je sestavljen iz pisnega dela, govornega dela, diahronega korpusa ter paralelnega korpusa.[41] Govorno komponento sestavljajo štirje ločeni korpusi, Praški govorni korpus (Pražský mluvený korpus, PMK), velikosti 675 tisoč besed, Brnski govorni korpus (BMK), velikosti pol milijona besed ter po enomilijonska korpusa ORAL 2006 (besedila, zbrana v letih 2002–2006) in ORAL 2008 (besedila, zbrana v letih 2002–2007, različna od besedil v korpusu ORAL 2006). Vsi štirje korpusi so transkribirani po enakih načelih, v t. i. modificirani verziji ortografske transkripcije. To pomeni precejšnjo prilagoditev pisni normi, poskušajo pa vendarle ohraniti najbolj tipične in najpogostejše specifične lastnosti govorjenega jezika: v zapisu ohranjajo nekatere regijske variante (`d´ole, vz´adu, kamen, zrouna`), pa tudi pri besedah, ki v množični rabi pri izgovorjavi odstopajo od zapisane oblike, upoštevajo pri zapisu govorno varianto (`sem = jsem, pudu = p°ujdu, von v´yde = vyjde, j´a si to vemu = vezmu, dyt', kani˘cka, ˘reben, kerej, pr´aznej, muskej`); transkripcije zato vsebujejo transkripcijske dvojnice (Kopřivová 2005, 140).

Za uravnoteženost korpusov so upoštevani štirje sociolingvistični kategoriji: spol, starost, izobrazba in govorni položaj. Znotraj vsake kategorije sta bili dve podkategoriji:

- **spol:** moški ali ženski (M/Ž),
- **starost:** mlajši (med 20 in 35 let) ali starejši (več kot 35 let) (Iunior/Vetus),
- **izobrazba:** osnovna (končana osnovna ali srednja šola) ali višja (končana višja ali visoka šola),
- **govor:** formalni (predvsem monolog) ali neformalni (predvsem dialog).

Regijskega izvora govorcev sestavljavci korpusov sprva niso imeli za kriterij za zajem besedil, tudi pri korpusih ORAL ne: »Naš osnovni cilj je zbiranje posnetkov tipičnega govorjenega jezika; zajeti hočemo običajno rabo govorjenega jezika, to pomeni jezik, kakršen se uporablja v vsakdanjih situacijah. Zato se ne oziramo na zajem posameznih narečij ali splošne češčine«[42] (Kopřivová 2005, 138); v korpusu

[41] http://ucnk.ff.cuni.cz/
[42] *Obecná čeština.*

ORAL 2008 so bili govorci vendarle uravnoteženi tudi glede na regijski izvor, in sicer v štirih skupinah (centralna, sevrovzhodna, jugozahodna Češka in obmejna področja).

2.2.8 Budimpeštanski sociolingvistični intervjuji

Za raziskovanje madžarskega govorjenega jezika je bilo na Raziskovalnem inštitutu za jezikoslovje (Madžarska akademija znanosti) v letih 1987–89 v Budimpešti posnetih 250 intervjujev, dolgih od dve do tri ure. Govorci za intervjuje so bili izbrani naključno, kasneje pa je bilo iz zbirke izbranih 50 govorcev, ki so bili razvrščeni po poklicu, kar naj bi opredeljevalo določeno demografsko stratifikacijo (po deset učiteljev, prodajalcev, navadnih delavcev, univerzitetnih študentov in praktikantov). Transkribiranje in označevanje intervjujev je bilo končano l. 2002 – transkribirana zbirka obsega 3.135.764 znakov. Z ustreznim procesiranjem podatkov je tako nastal prvi madžarski govorni korpus.[43] Zbirka se imenuje Budimpeštanski sociolingvistični intervjuji (Verzija 2) in zaenkrat ni del Madžarskega nacionalnega korpusa,[44] ki je bil sicer prav tako zgrajen na Madžarski akademiji znanosti, obsega pa 187,6 milijonov besed.[45]

2.2.9 Govorni korpus najstniške angleščine (COLT)

Ideja za gradnjo korpusa najstniške londonske angleščine (*The Bergen Corpus of London Teenage Language,* COLT) se je rodila na Oddelku za angleški jezik na Univerzi v Bergnu na Norveškem. Novembra 1992 so v sodelovanju z Oddelkom za kulturo, jezik in informacijske tehnologije (AKSIS) organizirali delavnico, na kateri so jim vodilni korpusni lingvisti predstavili »umetnost gradnje in rabe korpusov« (Stenstrom et al 2002, 2).[46] Norveški raziskovalni svet (*Norwegian Research Council*) je za namen gradnje korpusa polovično zaposlil enega raziskovalca. Drugi finančni viri so omogočali honorarno zaposlovanje podiplomskih študentov. Prvotni načrti za gradnjo polmilijonskega korpusa so se kmalu izkazali za nerealistične in jih v načrtovanih okvirih ni bilo mogoče izpeljati. Sodelavcem je uspelo dobiti podporo drugih institucij in finančnih virov: strokovnjaki iz založbe Longman, ki so transkribirali že BNC, so za COLT naredili ortografsko

[43] http://www.nytud.hu/depts/socio/index.html

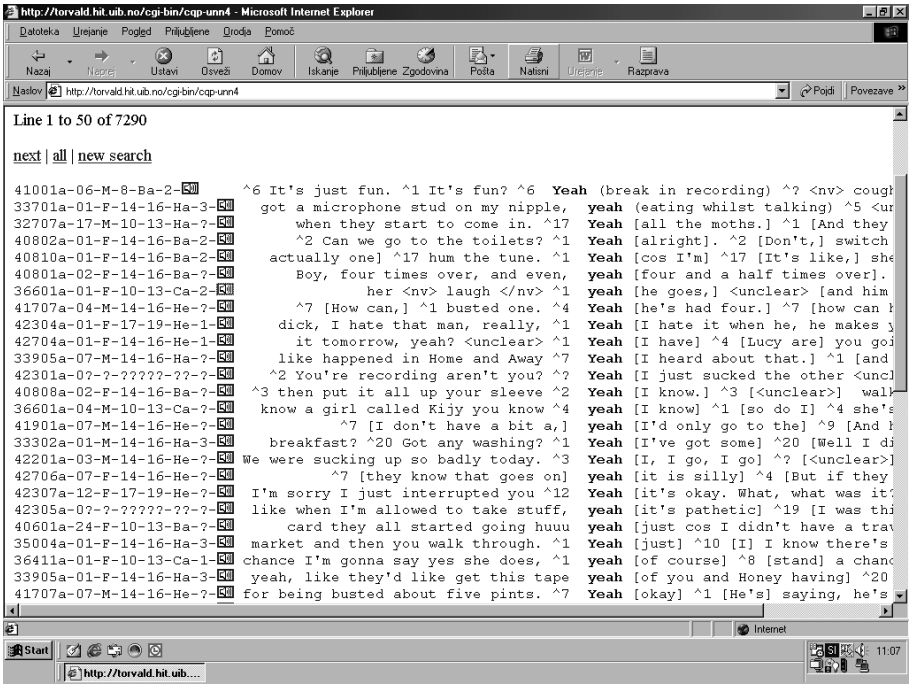
[44] http://corpus.nytud.hu/mnsz/index_eng.html

[45] Sestavljajo ga pisna besedila, razdeljena v pet podkorpusov: tiskani mediji, literatura, znanost, pravna besedila in neformalna besedila. Zadnji podkorpus sestavljajo besedila z madžarskih internetnih forumov, ki predstavljajo zelo spontano komunikacijo in so po mnenju sestavljalcev korpusa blizu govorjenemu jeziku.

[46] To so bili Jan Arts (projekt TOSCA), Paul Crowdy (BNC), Sidney Greenbaum (London-Lund in ICE), Stig Johansson (LOB), John Sinclair (COBUILD) in Jan Svartvik (London-Lund).

transkripcijo posnetkov (v zameno za podatke), na Oddelku za lingvistiko na Univerzi v Lancastru so besedila oblikoskladenjsko označili, projektu pa je tehnično podporo ves čas nudil tudi center HIT. Tako jim je vendarle uspelo zgraditi korpus velikosti pol milijona besed (enako kot korpus London-Lund), ki je bil zgrajen, transkribiran in označen po načelih BNC, kasneje pa BNC-ju tudi priključen kot podkorpus najstniškega govora.[47]

Za zbiranje govorjenih besedil so sodelavci projekta pridobili 33 najstnikov, starih od 13 do 17 let, iz različnih predelov Londona in z različnim socialnim izvorom. Snemanja so potekala dva tedna spomladi 1993 in en teden jeseni istega leta. Izbrane najstnike so prosili, naj 3 do 5 dni snemajo svoje pogovore s sovrstniki, po možnosti tako, da ti za snemanje ne bodo vedeli; vsak najstnik je dobil deset kaset, ki naj bi jih napolnil. Na koncu so sestavljavci dobili vrnjenih 30 kompletov ustreznih posnetkov (en komplet je bil prazen, eden preslabe kvalitete, en walkman pa se je izgubil). Najstniki so tudi dokaj zavzeto izpolnjevali identifikacijske liste govorcev, kar je za gradnjo in oznake govornega korpusa zelo pomembno.[48]



Slika 4: Korpus COLT

[47] http://torvald.aksis.uib.no/colt/
[48] http://www.hf.uib.no/i/Engelsk/COLT/facts.html

Sestavljavci korpusa so prvotno načrtovali enostavno prozodično transkripcijo, nekaj vmesnega med BNC-jevim modelom,[49] kjer so izjave transkribirane kot stavki (Crowdy 1991, 5), in modelom korpusa London-Lund, ki natančno označuje prozodično realizacijo govora. Čeprav se je v fazi načrtovanja model zdel dokaj enostaven in jasen, se je kasneje izkazalo, da bi bila njegova implementacija v danih okoliščinah nemogoča. Načrt je bil spremenjen in transkripcija je potekala v treh fazah: najprej so strokovnjaki založbe Longman vse posnetke transkribirali ortografsko; bergenski tim je še enkrat pregledal oz. primerjal posnetke z zapisanimi besedili.[50] Prvotnim načrtom se niso hoteli v celoti odpovedati, zato so 25 odstotkov korpusa označili po načelih prozodične transkripcije; korpus je bil dokončan l. 1998.

Korpus COLT je v komercialne namene dostopen v paketu, ki vsebuje 3 CD-rome (z zvočnimi posnetki, ortografsko transkripcijo posnetkov, prozodično transkripcijo, oblikoskladenjsko označenim besedilom in iskalnim programom). Korpus ima na spletnih straneh[51] zbrano in prosto dostopno dokumentacijo korpusa in bibliografijo raziskav, povezanih s korpusom, z geslom pa je dostopen tudi sam korpus.

2.2.10 Švedski govorni korpus

Korpus govorjene švedščine je bil zasnovan na Univerzi v Göteborgu že konec sedemdesetih let, dograjen pa je bil leta 1995, v velikosti 1,4 milijona besed. Sestava korpusa je temeljila na spoznanju, da se v različnih družbenih okoliščinah govori različne vrste govorjenega jezika, in sicer tako v pogledu regionalnega naglasa,[52] izgovorjave, besedišča in slovnice kot tudi komunikacijskih funkcij (Allwood et al. 2001, 1). Zato je bil cilj načrtovalcev korpusa vanj zajeti besedila, ki bi zajemala čim širši spekter družbenih okoliščin in sporazumevalnih dejavnosti. Korpus naj bi bil podoben novozelandskemu korpusu govorjene angleščine (del ICE), nekatere lastnosti pa naj bi ga povezovale tudi z BNC in s korpusom London-Lund.

Korpus sestavljajo izključno posnetki spontanega govora (Allwood et al. 2001, 1), in sicer je polovica samo zvočnih, polovica pa tudi video posnetkov; delež

[49] BNC-jev model je ortografska transkripcija z nekaterimi elementi prozodične transkripcije (vključuje premore in kazalce intonacije, pa tudi elipse, ponavljanja, ponovne začetke idr.) ter brez fonetičnih oznak, temelji pa na stavkom podobnih enotah (*sentence-like units*; Crowdy 1991, 5) in ne na tonskih enotah.

[50] Zanimivo je, da je velikost korpusa po pregledu bergenskih strokovnjakov narasla za skoraj 20 % (predvsem na račun razjasnjenih mest, ki so jih Longmanovi transkriptorji označili kot nejasna).

[51] http://torvald.aksis.uib.no/colt/

[52] V švedskem in norveškem jeziku velja, da naj bo govorec tudi v standardnem govoru regionalno prepoznaven, zato je njihov pogled na regionalni naglas drugačen kot npr. naš.

slednjih je daleč največji glede na vse obstoječe govorne korpuse. Sestavo korpusa prikazuje naslednja tabela:

Okoliščine/žanr	Št. posnetkov	Št. pojavnic	Trajanje
Cerkev (pogreb)	2	10.235	1:46:43
Ciljno naravnani dialog	26	15.347	2:05:20
Diskusija	36	255.261	28:36:32
Dražba	2	28.094	3:14:11
Formalni sestanek	14	202.850	24:20:16
Hotel	9	18.137	9:49:55
Igra vlog	3	8.055	0:57:16
Igre	2	10.316	1:17:01
Intervju	57	389.396	44:37:51
Konzultacija (optik itd.)	16	34.285	4:08:04
Neformalna konverzacija	19	87.087	8:19:41
Obnova besedila, članka	7	5.291	0:42:00
Potovalna agencija	40	39.881	6:00:10
Predavanje	2	14.667	1:38:00
Pripravljena diskusija	2	9.098	0:47:15
Sejmišče	16	14.116	1:21:22
Sestanek	2	45.484	6:00:58
Sodišče	6	33.723	3:58:33
Avtobus	1	1.348	0:13:37
Telefon	32	14.613	2:01:48
Terapevt/pacient	2	13.527	2:04:07
Tovarna	5	28.884	2:56:28?
Trg	4	12.175	3:55:07
Trgovina	54	50.497	10:34:33
TV	1	2.921	0:25:35
Večerja	5	30.144	2:49:54
Skupaj	365	1.375.432	174:42:23

Tabela 6: Zajem besedil v korpus govorjenega švedskega jezika[53]

[53] http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3&SUBPAGE=3&FILE=corpus_overview_brief

Pri zajemanju besedil v švedski govorni korpus se sestavljavci niso ozirali na demografsko sestavo govorcev, zelo pa so si prizadevali za veliko število različnih vrst besedil in znotraj njih za čim več gradiva. Zelo veliko pozornost so posvetili transkripcijskim standardom. Razvili so t. i. nevtralni göteborški transkripcijski standard (GTS) in ga poleg švedščine preizkusili še na kitajskem, arabskem, angleškem, španskem, bolgarskem in finskem jeziku (predstavljen je v poglavju 4.3.5, *Göteborška modificirana ortografska transkripcija*). Poleg nevtralnega so razvili tudi poseben standard za zapisovanje švedskega govorjenega jezika, t. i. modificirano standardno ortografijo (MSO), ki so jo uporabili za transkribiranje posnetkov korpusa.

Na podlagi raziskav švedskega korpusa je nastalo veliko razprav, za najodmevnejše veljajo razprave o razlikah v frekvencah besed med govorjenim in pisnim jezikom.[54] Sicer pa so načrti v zvezi s korpusom usmerjeni predvsem k povečevanju njegove velikosti, in sicer z dodajanjem besedil iz novih govornih situacij in z izenačevanjem količine gradiva znotraj posameznih situacij. Z razvijanjem programskih orodij bodo skušali korpus narediti dostopnejši (velika količina videoposnetkov delovanje korpusa zelo upočasnjuje, poleg tega pa zahteva veliko shranjevalnega prostora), sicer pa naj bi nadaljevali s kvantitativnimi in kvalitativnimi analizami, znotraj česar je najambicioznejši cilj slovnični opis govorjenega jezika (Allwood et al. 2001, 9).

2.2.11 Nizozemski govorni korpus

Korpus govorjenega nizozemskega jezika je bil projekt velikanskih razsežnosti, ki se je uradno začel 1. 6. 1998. Ob njem so združile moči številne vladne, izobraževalne in raziskovalne institucije na Nizozemskem in v Belgiji. Za načrtovani govorni korpus velikosti 10 milijonov besed (1000 ur posnetkov) je bil predviden proračun 4,6 milijona €; na koncu je bilo za nekoliko manjši korpus (8,9 milijona besed, 800 ur posnetkov) porabljenega nekoliko več denarja (4,9 milijona €).[55] Gradnja korpusa je imela zelo natančno organizacijsko strukturo, ki so jo sestavljali korpusni svet, nadzorni svet, management korpusa, delovne skupine za posamezne faze gradnje (zajem in zbiranje besedil, transkripcija, označevanja) in tajništvo. To je nazoren primer, kako vzorna mora biti organizacija tako velikega korpusnega projekta in koliko sredstev je potrebnih za uspešno realizacijo; projekt je bil zaključen v petih letih, 1. marca 2004.

[54] http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=5
[55] http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm

Projekt sta financirali nizozemska in belgijska vlada ter Nizozemska organizacija za znanstveno raziskovanje. Motivacija za gradnjo tako velikega govornega korpusa je bila naslednja: nizozemščino kot »enega najmanjših jezikov Evrope«[56] naj bi močno ogrožala angleščina, ki jo je dostopnost številnih jezikovnih virov utrdila na vodilnem mestu med svetovnimi jeziki, kot jezik mednarodne (poslovne) komunikacije pa si položaj še utrjuje. Dejstvo, da je bilo za nizozemščino pred desetimi leti dostopnih le malo relevantnih jezikovnih virov, je resno ogrožalo razvoj nizozemskih jezikovnih in govornih tehnologij, velik govorni korpus pa naj bi to situacijo bistveno izboljšal; gradnjo govornega korpusa je narekoval visoko izražen nacionalni interes, brez tega pa se dandanes tako velikega projekta verjetno ne more načrtovati.

Poleg interesa jezikovnih in govornih tehnologij naj bi korpus služil še drugim raziskovalnim interesom, predvsem jezikoslovcem, ki so do gradnje korpusa za opise jezika uporabljali predvsem pisne vire, učiteljem nizozemščine kot tujega jezika in izdelovalcem različnih gradiv za to področje, uporabljali pa naj bi ga tudi pri pouku nizozemščine kot prvega jezika.

Besedila so v skladu s financiranjem zajemali na Nizozemskem (2/3) in v Flandriji (1/3). Zajemanje je temeljilo na vnaprej določenih družbenih okoliščinah, v katerih se jezik uporablja, znotraj teh pa so bili upoštevani različni vidiki sporazumevanja, npr. sporazumevalni cilj, medij, število udeležencev in razmerje med govorcem/-ci in poslušalcem/-ci.[57] Kjer se je zdelo potrebno, so bile lastnosti govorcev (spol, starost, regija in socialni status) uporabljene kot (demografski) kriteriji zbiranja,[58] sicer pa so bili ti podatki zabeleženi samo kot metajezikovni opis v glavah transkripcij.

Korpus je transkribiran ortografsko, transkripcije pa so povezane z zvočnimi posnetki. Ortografska transkripcija je bila tudi izhodišče za lematizacijo in oblikoskladenjsko označevanje. Na delu korpusa (1 milijon besed) je bila narejena fonetična transkripcija, za ta del pa je bila narejena tudi povezava med transkripcijami in zvočnimi posnetki na ravni besede.[59] Nadalje je bil del korpusa (1 milijon besed) označen skladenjsko, ne nazadnje pa je bil manjši del korpusa (250.000 besed) označen tudi prozodično.

[56] Formulacija je z zgoraj navedene predstavitvene strani Korpusa; sicer je nizozemščina uradni jezik na Nizozemskem (15 milijonov govorcev), v Belgiji (Flandrija, 5,6 milijona govorcev), v južnoameriški državi Surinam (300.000 govorcev) in na Nizozemskih Antilih (240.000 govorcev).

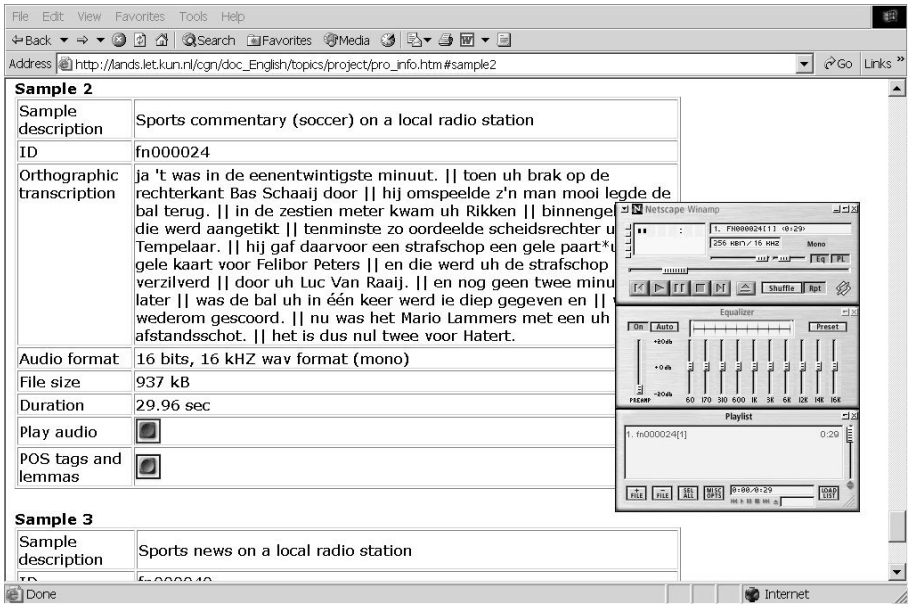
[57] Taksonomija besedilnih tipov bo podrobneje predstavljena v poglavju 3, *Zajem besedil v govorni korpus.*

[58] Npr. v okviru besedilne vrste »spontana konverzacija«.

[59] Sicer so običajno povezave narejene na ravni časovnih odsekov, ki ustrezajo izjavam (prim. poglavje 8.6, *Konvertiranje*, str. 187); korpus govorjene nizozemščine je za zdaj edini znani korpus na svetu, v katerem je izvedena tudi povezava na besedni ravni.

Lastnik korpusa je zdaj Nederlandse Taalunie,[60] Nizozemska jezikovna zveza, ki skrbi za promocijo in razvoj nizozemskega jezika (ustanovljena l. 1980). Korpus se lahko uporablja samo v raziskovalne namene in z dovoljenjem Zveze. Kdor kupi komercialno licenco, lahko korpus uporablja tudi za komercialne izdelke (npr. programi za razpoznavo govora, slovarji). Korpus je shranjen v obliki 33 DVD-jev, od tega so na 32 shranjeni zvočni posnetki.

Nizozemski govorni korpus ima na internetu dostopno tudi izjemno pregledno in natančno dokumentacijo o gradnji. Tam je mogoče slediti vsem stopnjam nastajanja korpusa (tudi spremembam, do katerih je prišlo med gradnjo), preštudirati organizacijsko shemo, taksonomijo besedilnih tipov, transkripcijske standarde in standarde označevanja ter obsežno bibliografijo razprav, korpus pa je mogoče v demo izvedbi tudi preizkusiti.[61]



Slika 5: Nizozemski govorni korpus (demo)

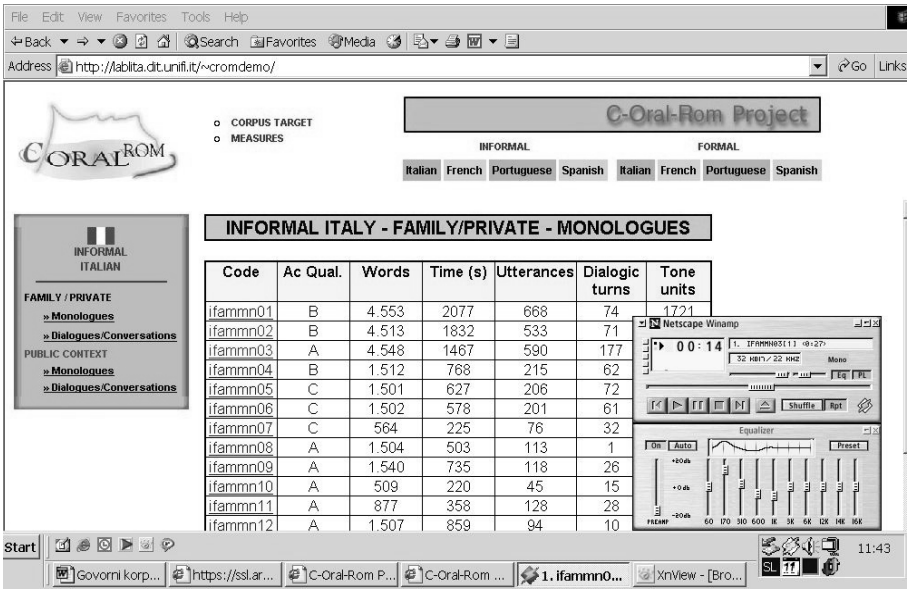
Po velikosti je nizozemski govorni korpus primerljiv z govorno komponento BNC; raznovrstne oznake, od ortografske do fonetične in prozodične transkripcije, oblikoskladenjske in skladenjske oznake, poleg tega pa še sinhroni dostop do zvočnih posnetkov pa nizozemski korpus trenutno uvrščajo med vodilne govorne korpuse na svetu.

[60] http://taalunieversum.org/en/about_us/
[61] http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm#sample2

2.2.12 C-ORAL-ROM

C-ORAL-ROM je zbirka štirih približno enakih korpusov spontanega govorjenega jezika, in sicer francoščine, italijanščine, portugalščine in španščine.[62] Projekt izdelave se je začel konec devetdesetih let, dokončan pa je bil leta 2003; vsak posamezni korpus vsebuje 300.000 besed. Glavni namen gradnje je bil zagotoviti zbirko podatkov govorjenega jezika za jezikoslovne raziskave, predvsem za potrebe govornih tehnologij. Korpus omogoča primerjave med štirimi romanskimi jeziki, pa tudi raziskave znotraj posameznega jezika. Transkripcije so prozodične, za označevanje pa je bil razvit poseben program WinPitch, ki se ga sedaj pogosto uporablja za transkribiranje ali za naknadno povezovanje transkripcij z zvočnimi posnetki.

Zajem besedil v korpuse je bil vnaprej določen in naj bi zagotavljal določeno stopnjo primerljivosti korpusov. Polovica besedil je bila zbranih v neformalnih, druga polovica pa v formalnih okoliščinah, znotraj tega je 15 odstotkov medijskega govora. Demografske lastnosti govorcev niso vplivale na zajem besedil v korpus in med posameznimi korpusi tudi niso primerljive, so pa vedno označene kot metajezikovni podatki v glavi korpusa (starost, spol, izobrazba in regija govorca).



Slika 6: Korpusni paket C-ORAL-ROM (demo) [63]

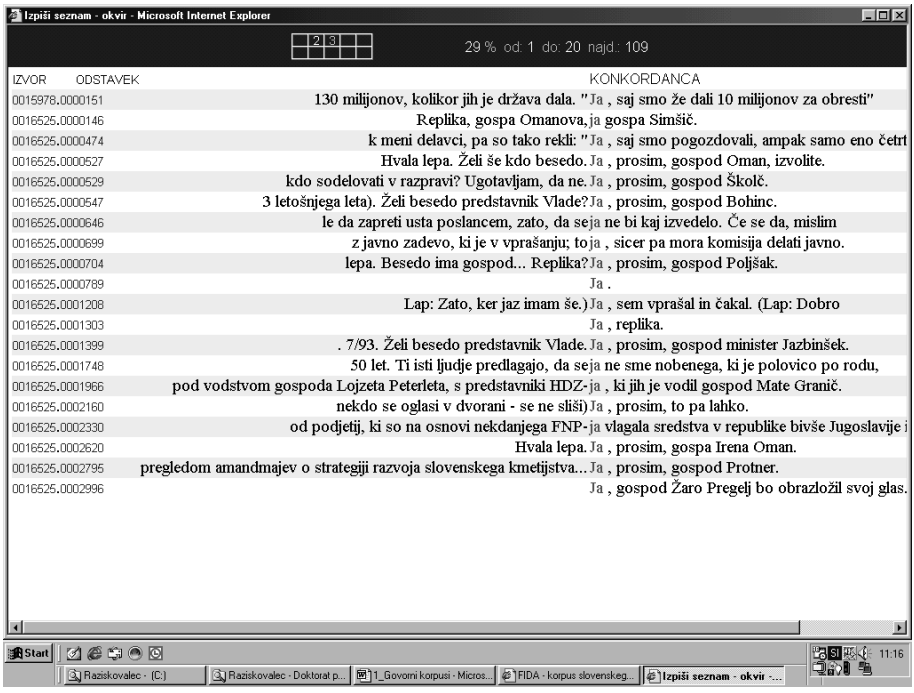
[62] http://lablita.dit.unifi.it/coralrom/index.html
[63] http://lablita.dit.unifi.it/~cromdemo/

Novost, ki jo korpusni paket C-ORAL-ROM prinaša na področje govornih korpusov, je predvsem njegova izrecna namembnost za raziskave na področju govornih tehnologij.[64] Sistemi za razpoznavo govora morajo biti zaradi zahtev trga in konkurenčnosti vedno bolj usposobljeni za razumevanje spontanega govora, sintetizatorji pa morajo vedno bolje posnemati naravni govor; za gradnjo takih sistemov postajajo korpusi spontanega govorjenega jezika nujno potrebne baze podatkov tudi za strokovnjake s področja govornih tehnologij.

2.2.13 Govorna komponenta korpusa FIDA

V korpus slovenskega jezika FIDA je bil vključen tudi 5-odstotni delež transkribiranih govorjenih besedil. Obseg »govornega podkorpusa« je relativno velik, več kot dva milijona besed (2.041.453), vendar pa vključuje samo eno vrsto besedil, in sicer transkripcije razprav iz Državnega zbora RS, ki so bile narejene v popolnoma druge namene.[65] Transkripcije so ortografske, brez prozodičnih oznak, v celoti prilagojene pisnemu jeziku, to je segmentirane na povedi. Poleg tega so transkripcijska načela popolnoma nedokumentirana, kar je glede na namen transkribiranja razumljivo. Pri analizah govornega podkorpusa FIDE je treba upoštevati njegovo besedilnovrstno sestavo, zaradi katere ga ne moremo imeti za uravnoteženega ali reprezentativnega, vendar pa podkorpus kljub temu omogoča določene relevantne vpoglede v nekatere vidike govorjenega jezika oziroma govornega sporazumevanja. Transkripcije parlamentarnih razprav bi v ustrezno prilagojenem transkripcijskem standardu lahko bile tudi del morebitnega uravnoteženega govornega korpusa slovenskega jezika.

[64] http://lablita.dit.unifi.it/coralrom/objectives.html
[65] Transkripcije sej DZ RS so vključene tudi v korpus FidaPLUS, vendar tam obsegajo manj kot 1 % korpusa.



Slika 7: Konkordance iskalnega niza *ja* iz govornega podkorpusa FIDA

2.2.14 Slovenske govorne zbirke

Predstavila bom tudi nekaj slovenskih govornih zbirk, čeprav niso pravi govorni korpusi. Tudi ta gre za načrtno zbiranje, shranjevanje in transkribiranje posnetkov govora, v nekaterih primerih celo s sledjo avtentičnosti.

Baza izgovorjav SNABI je bila izdelana (1994–1998) na Fakulteti za elektrotehniko, računalništvo in informatiko Univerze v Mariboru, in sicer za gradnjo sistemov za avtomatsko prepoznavanje govora po telefonu. Vsebuje govorni signal telefonske kvalitete 82 govorcev in govorni signal studijske kvalitete 56 govorcev. Vsak govorec je izgovoril v povprečju 200 stavkov, 80 izoliranih besed, številske nize in abecedo. Govorni signal je bil segmentiran (členjen na izjave), označen in fonetično transkribiran (Kačič in Horvat 1998, 101–102).

Podatkovna zbirka govora GOPOLIS (GOvorjena POizvedovanja o Letalskih Informacijah) je nastala v drugi polovici devetdesetih let na Fakulteti za elektrotehniko Univerze v Ljubljani. Narejena je bila za razvoj sistema za razpoznavo

govora in krmilnika dialoga pri govornih poizvedbah o letalskih informacijah. Podlaga za gradnjo korpusa je bilo 15 ur telefonskih pogovorov med poizvedovalci in telefonisti Adrie Airways. Iz pogovorov so bili nato izbrani najbolj tipični stavki, ki so jih v studiu prebrali izbrani govorci (25 moških in 25 žensk); ti posnetki sestavljajo govorno zbirko GOPOLIS. Za govorni korpus slovenskega jezika bi bila seveda zanimiva poizvedovanja sama, torej telefonski pogovori, ki so služili za podlago Gopolisa. Vendar pa avtorske pravice za posnetke niso bile urejene, vsaj ne na način, kot to zahteva gradnja govornega korpusa (dovoljenje za objavo), zato vključitev teh pogovorov v govorni korpus tudi teoretično ne bi prišla v poštev.

SPEECH-DAT je bil evropski projekt gradnje govornih zbirk za razvoj sistemov telefonskega govornega dialoga za delo v realnem okolju. Slovenska govorna zbirka, ki je primerljiva z govornimi zbirkami drugih evropskih jezikov v okviru istega projekta, vsebuje govor 1000 govorcev, posnetih preko telefonske linije.

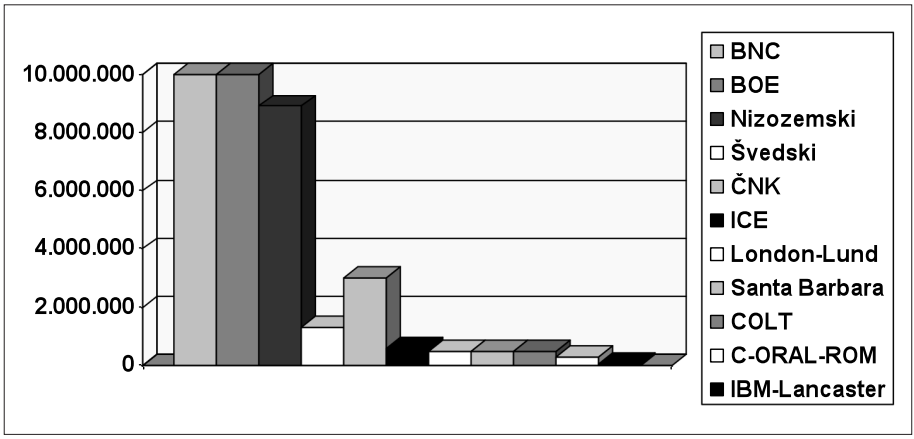
Tudi govorna zbirka POLIDAT je bila zgrajena za razvoj avtomatskih telefonskih sistemov govornega dialoga. Pri gradnji so poskušali v čim večji meri slediti priporočilom in kriterijem, ki so bili definirani v okviru evropskega projekta SpeechDat II, nanašali pa so se predvsem na demografsko sestavo govorcev.[66] V korpus so sicer uvrščene izjave 1000 govorcev (izbranih izmed 1400), posneti govor pa predstavljajo skoraj v celoti vnaprej pripravljena brana besedila, majhen delež spontanega govora pa odgovori izbranih govorcev na vprašanja o letu rojstva, o tem, koliko je ura in podobno. Zanimivo je stališče sestavljavcev korpusa, da bo v prihodnje za zagotavljanje konkurenčnosti potrebno razvijati sisteme »za razpoznavanje tekočega ali celo spontanega govora« (Zögling Markuš in drugi 2000, 95), kar potrjuje hipotezo, da so govorni korpusi nujno potreben jezikovni vir tudi za razvijanje sistemov za analizo in sintezo govora.

2.3 PRIMERJAVE IN DRUGI GOVORNI KORPUSI

Zgoraj je bilo podrobneje predstavljenih 12 tujih govornih korpusov, ki so nastali v preteklih desetletjih (1980–2007), in sicer tistih, ki so nastali med prvimi, ki sodijo med največje ali pa so v času svojega nastanka pomenili korak naprej v tehnološkem razvoju. Ugotovili smo lahko, da se korpusi med seboj razlikujejo po namenu, kar vpliva na njihovo velikost, sestavo in transkripcijski standard. Med njimi je več kot polovica samostojnih korpusov, nekateri pa so govorne komponente referenčnih korpusov. Kljub razlikam pa so korpusi vendarle tudi primer-

[66] Izbrani kriteriji bodo podrobneje predstavljeni v poglavju 3, *Zajem besedil v govorni korpus*.

ljivi; podobnostim lahko sledimo predvsem pri opazovanju procesov gradnje, ki si sledijo v enakem zaporedju in kjer si vsak tim na svoj način prizadeva doseči svoj cilj – zgraditi zbirko avtentičnih jezikovnih podatkov, ki bo omogočala raziskovanje govorjenega jezika. V spodnji tabeli so obravnavani govorni korpusi urejeni po velikosti:



Slika 8: Primerjava govornih korpusov po velikosti

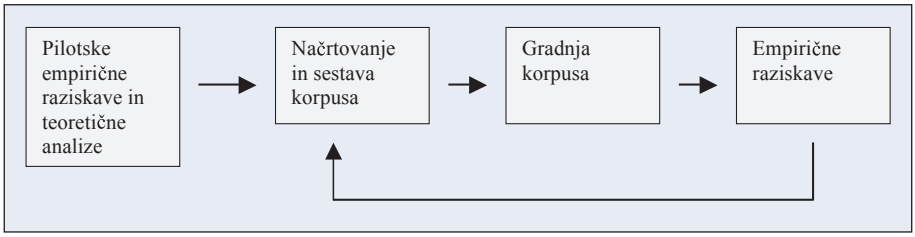
Med govornimi korpusi, ki jih nisem podrobneje predstavila, velja omeniti mdr. korpus MICASE (*Michigan Corpus of Academic Spoken English*, http://www.hti.umich.edu/m/micase/, 1.848.364 besed), CANCODE (*Cambridge and Nottingham Corpus of Discourse in English,* 5 milijonov besed), korpus COLA (korpus najstnikov Madrida, Buenos Airesa in Santiaga de Chile, http://www.colam.tk/), korpus TOSCA (Univerza v Nijmegenu, http://taalunieversum.org/taal/technologie/onderzoeksgroep_tosca_nijmegen/index.php), BySoc (danski govorni korpus, http://www.id.cbs.dk/ ~pjuel/cgi-bin/BySoc_ID/index.cgi?EeNnGg, 1,3 milijona besed) in dva manjša bolgarska govorna korpusa (http://www.hf.uio.no/east/bulg/mat/).

Vpogled v gradnjo obstoječih govornih korpusov nam lahko pomaga kot izhodišče pri oblikovanju lastnih načel za zajemanje besedil v govorni korpus slovenščine, kar bo tema naslednjega poglavja.

3 Zajem besedil v govorni korpus

3.1 KRITERIJI ZAJEMANJA

Gradnja korpusa mora biti skrbno načrtovana, določajo pa jo poleg namena korpusa tudi finančne zmožnosti in razpoložljivi človeški viri (raziskovalci, zbiralci gradiva, transkriptorji, označevalci itd.). Med gradnjo se zlahka pojavijo okoliščine, ki zahtevajo spremembe v načrtovanju korpusa. Niti gradnja pilotskega ali učnega korpusa ne more zagotoviti popolnoma trdnih in nespremenljivih kriterijev zajemanja, vendar to tudi ni potrebno. Pomembno je, da se načrtovalci korpusa zavedajo, da je gradnja korpusa ciklični proces, ki ob sprotnem evalviranju lahko zahteva spreminjanje začetnih izhodišč, pa tudi proces, ki se v idealnih okoliščinah ne bi nikoli končal, saj bi morali korpus nenehno dopolnjevati in aktualizirati.



Slika 9: Gradnja korpusa kot ciklični proces (Biber 1993, 256)

Določene teoretične raziskave in predpostavke morajo biti narejene pred začetkom gradnje govornega korpusa. Najprej je treba določiti okoliščine in dejavnike, za katere predvidevamo, da pogojujejo razlike v uporabljenih jezikovnih znakih/strukturah govorne produkcije določene govorne skupnosti, določiti pa je treba tudi lingvistične in sociolingvistične komponente, ki jih nameravamo s korpusom raziskovati in analizirati (Biber 1993, 243). Te komponente bodo določale tipe besedil in govorcev; biti pa morajo dobro označene in dokumentirane v glavi dokumenta, saj bodo predstavljale kriterije za iskanje po korpusu.

Pri načrtovanju govornega korpusa kot dela referenčnega korpusa je temeljno prizadevanje namenjeno doseganju reprezentativnosti in uravnoteženosti korpusa. Problem predstavlja dejstvo, da ne obstajajo objektivna merila, po katerih bi lahko izmerili celotno govorno produkcijo neke jezikovne skupnosti. Mogoče je sestaviti izčrpni seznam besedilnih vrst in z jezikoslovnimi raziskavami poskušati določiti kvantitativna razmerja med njimi (Crowdy 1995, 224), vendar je velika verjetnost, da bomo med gradnjo korpusa želeli ali morali ta razmerja spreminjati.

Reprezentativnost govornega korpusa temelji na določitvi čim večjega števila različnih besedilnih vrst in na številu vzorcev znotraj posamezne besedilne vrste,

poleg tega pa tudi na določitvi reprezentativnega vzorca vseh govorcev. V nadaljevanju si bomo podrobneje ogledali vse naštete dejavnike.

3.1.1 Velikost korpusa

John Sinclair je o velikosti govornih korpusov razmišljal takole: »Pred leti sem imel govorni korpus velikosti 220.000 besed in takrat se mi je zdel velik, ne, zdel se mi je ogromen, in tak se je zdel tudi računalniku« (Sinclair 1995, 103). Sicer pojem velikosti korpusa variira glede na namen korpusa: če gre za jezik v specifični funkciji, je lahko že manjši korpus dovolj reprezentativen, drugače pa je, če želimo predstavljati splošnejši tip jezika, takrat je na prvem mestu kvantiteta (Sinclair 1995, 101).

Pri pregledovanju obstoječih govornih korpusov smo videli, da je njihova velikost postopno naraščala; sredi devetdesetih let je bila dosežena količina 10 milijonov besed, ki jo danes vsaj približno dosegajo trije korpusi na svetu, govorni komponenti BNC in BoE ter Nizozemski govorni korpus. Zdi se, da je zaenkrat to največja količina gradiva, ki jo je mogoče zbrati z metodami, orodji in programi, ki so danes na voljo; morda se bo obseg govornih korpusov povečal, ko bodo sistemi za razpoznavo govora zmogli olajšati transkribiranje.[67] Naslednji po velikosti je sedaj Češki govorni korpus, saj je skupna velikost vseh štirih govornih podkorpusov čez 3 milijone besed. Za ameriški korpus Santa Barbara sicer načrtujejo 5 milijonov besed, vendar kljub desetletnemu intenzivnemu delu zastavljenega cilja še niso dosegli. Najbolj številčno skupino glede na velikost predstavljajo korpusi velikosti 500 do 600 tisoč besed: London-Lund, 15 načrtovanih korpusov znotraj ICE in COLT; načrtovalci korpusa štirih romanskih jezikov C-ORAL-ROM so se odločili za velikost 300.000 besed za vsak posamezni korpus.

3.1.2 Metode zajemanja besedil

Za govorne korpuse, ki so del referenčnih korpusov, načeloma velja, da je treba vanje vključiti čim več različnih besedilnih vrst in znotraj njih čim več besedil. V specializiranih korpusih je namen ožje določen in zato nabor besedilnih vrst bistveno ožji; tak je npr. korpus MICASE[68] (*Michigan Corpus of Academic Spoken English*), v katerem je pozornost osredotočena samo na akademski diskurz, zato je taksonomija tudi znotraj vrste lahko mnogo bolj podrobna in tudi kvantitativno natančneje določena.

[67] Danes si je težko predstavljati, da bi lahko programska oprema avtomatsko transkribirala govor, kjer se prekriva več govorcev; kljub temu je teoretično v perspektivi mogoče tudi to.

[68] Velikost 1.848.364 besed; dostopen na http://www.hti.umich.edu/micase/.

Pri gradnji reprezentativnih in uravnoteženih govornih korpusov obstajajo različne možnosti kategorizacije besedilnih vrst. Temeljijo lahko na besedilni produkciji, besedilni recepciji ali na besedilu samem (Biber 1993, 245). Prvi dve kategorizaciji temeljita na demografskih komponentah: iz celotne populacije so izbrani posamezniki, ki predstavljajo reprezentativni vzorec celote, nato pa se skuša zajeti njihov jezik v celoti (pisni in/ali govorjeni). Načrtovanje korpusa na ravni produkcije bi zajemalo besedila, ki jih ti posamezniki tvorijo, načrtovanje na ravni recepcije pa besedila, ki jih sprejemajo. Oba pristopa iščeta odgovor na vprašanje o dejanski rabi jezika znotraj določene populacije.

3.1.2.1 Demografsko vzorčenje

Za zajem besedil z demografsko metodo je treba statistično določiti vzorec govorcev, ki ustreza celotni izbrani populaciji.[69] Kriteriji, ki se jih pri vzorčenju govorcev lahko upošteva, so:

- spol,
- starost,
- regijska pripadnost,
- etnična pripadnost,
- izobrazba,
- poklic,
- socialni status
- in drugo.

Obstajajo še drugi demografski kriteriji, ki pridejo redkeje v poštev in so morda pomembni v specifičnih jezikovnih ali kulturnih okoljih, npr. kraj bivanja, kraj rojstva, verska pripadnost itd. Poleg teh pa obstajajo še demografski kazalci, ki ne karakterizirajo skupin uporabnikov jezika, ampak individualne uporabnike (Biber 1993, 256), npr. osebnostne lastnosti, osebni interesi, osebna prepričanja in razpoloženje; omenjene lastnosti se običajno pri gradnji referenčnega korpusa zanemarijo.

Našteti demografski kriteriji lahko pri gradnji korpusa služijo samo kot izhodišče. Vsak kriterij posebej zahteva znotraj posamezne jezikovne skupnosti poseben premislek in kot bomo videli v nadaljevanju, odpira številna vprašanja. Poleg tega ne moremo pričakovati, da bo korpus, zgrajen na demografskih kriterijih, zajel vse znane tipe besedil: večina govorcev nekaterih besedilnih tipov nikoli ali skoraj nikoli ne tvori, nekaterih pa tudi nikoli ali skoraj nikoli ne sprejema. Če bi bil

[69] Lahko celotna populacija govorcev nekega jezika, lahko npr. samo odrasli ali samo mladostniki itd.

referenčni korpus zgrajen samo na podlagi demografske stratifikacije govorcev, bi predvidoma vseboval približno 90 odstotkov besedil spontane konverzacije, 3 odstotke pisem in zapiskov, preostalih 7 odstotkov pa bi vsebovalo vsa ostala besedila, npr. literarna in znanstvena besedila itd. (Biber 1993, 247). Tak korpus bi odslikoval absolutno frekvenco (tvorjenih) besedil znotraj celotne produkcije, ne bi pa bil reprezentativen v smislu prikazovanja tistih jezikovnih prvin, ki v numeričnem smislu ne dosegajo visoke pojavnosti, imajo pa velik doseg (slišanost oz. branost). Korpus, ki želi izkazovati raznolikost jezikovnih prvin v različnih besedilnih vrstah, mora demografsko zbiranje podatkov kombinirati z drugimi metodami, predvsem z zajemanjem na podlagi besedilnovrstne tipologije.

3.1.2.2 Besedilnovrstno vzorčenje

Korpus, ki ima v izhodišču kategorizacije besedilo, bo ob skrbnem načrtovanju reprezentativno zastopal vse besedilne tipe, vendar ne v razmerjih, kot se pojavljajo v okviru celotne govorne produkcije. Kot smo videli pri znanih obstoječih govornih korpusih, so t. i. besedilnovrstni kriteriji za zajem besedil v korpus, ki temeljijo na različnih vidikih govorjenih besedil, najpogosteje naslednji:

- stopnja spontanosti: spontani, pripravljeni, brani govor,
- struktura besedila: monolog, dialog, multilog,
- namen besedila,
- okoliščine: zasebno, javno,
- govorni položaj: formalni, neformalni,
- prenosnik: osebni stik, telefon, mediji.

Tudi tu se načrtovalcu korpusa zastavljajo različna vprašanja. Za raziskovanje avtentičnega govorjenega jezika so najbolj dragoceni posnetki spontanega govora, zato se jih jezikoslovci trudijo čim več zajeti v korpus. Težko se je odločiti za hierarhijo med posameznimi kriteriji, prav tako pa tudi za kvantitativna razmerja komponent znotraj posameznih kriterijev. Gre za vprašanja, na katera si mora vsaka ekipa načrtovalcev korpusa odgovoriti sama, saj univerzalnih načrtov za gradnjo govornega korpusa ni. Pri tem lahko pomagajo ustrezne jezikoslovne analize, če obstajajo, deloma pa se načrtovalci opirajo na svoje hipoteze, ki jih morajo v procesu gradnje po potrebi tudi korigirati.

3.1.3 Avtorske pravice

Pridobivanje avtorskih pravic za uporabo in objavo besedil v korpusu je tesno povezano z načrtovanjem in začetno fazo gradnje korpusa. S tovrstno problematiko se ukvarjajo načrtovalci korpusov že od samega začetka, zdi pa se, da dobiva nove in nove razsežnosti. V zvezi s tem je zanimiva izkušnja načrtovalcev korpusa BNC. Na podlagi izkušenj s korpusi prve generacije, katerih uporaba je bila zaradi neurejenih avtorskih pravic izredno omejena, so se hoteli dobro zavarovati pred morebitnimi težavami. Z zelo natančnimi postopki so problem avtorskih pravic v času gradnje zadovoljivo rešili in s tem omogočili široko dostopnost korpusa: avtorje pisnih in govornih besedil so prosili za pisno dovoljenje za uporabo besedil v korpusu, pri tem pa so jim zagotavljali popolno anonimnost in uporabo korpusa izključno v raziskovalne namene. Če se avtorji besedil z objavo niso strinjali oz. niso podpisali dovoljenja, besedilo ni bilo vključeno v korpus. V zvezi s tem so morali načrtovalci rešiti tudi problem, kaj storiti, če se imena govorcev pojavljajo v besedilu kot referenca; najprej so razmišljali, da bi jih nadomestili z alternativnimi, lingvistično podobnimi imeni, vendar so kasneje namero opustili in lastna imena večinoma izločili iz besedil. Sredi devetdesetih let se je zdelo, da je problem avtorskih pravic korpusa BNC v celoti rešen.

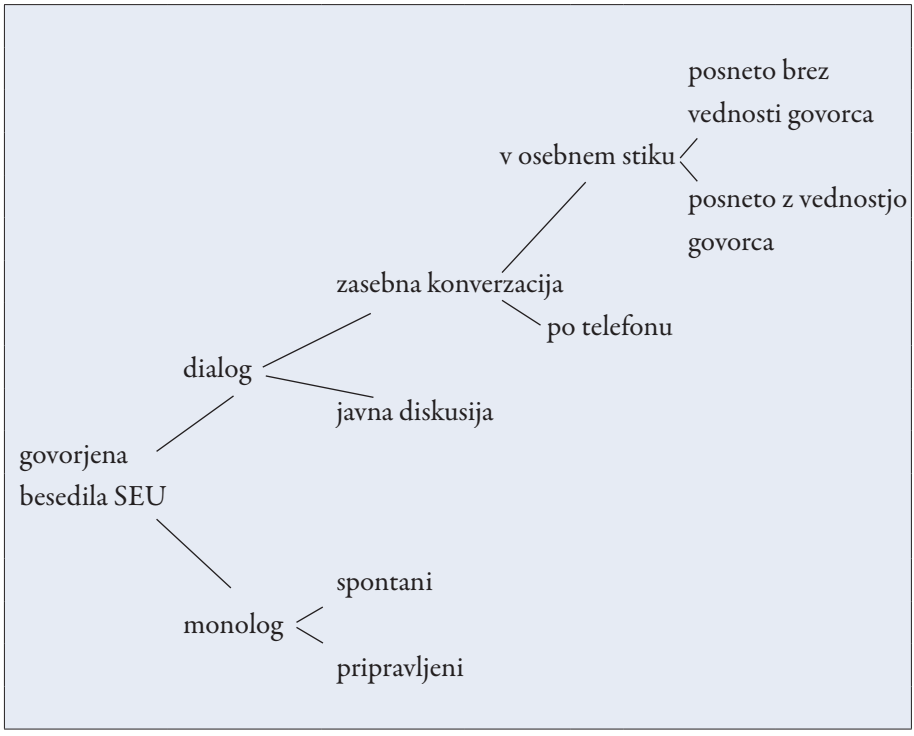
Vendar se je z leti pokazalo, da ni tako. Govorci besedil so bili naprošeni samo za dovoljenje za objavo transkribiranih besedil, ker takrat nihče ni pomislil na možnost objave zvočnih posnetkov. Kasneje, ko je programska oprema tako napredovala, da bi v revidirani verziji korpusa lahko začeli razmišljati o sinhronizaciji zvočnih posnetkov s transkripcijami (prim. Burnard 2000, 14), se je izkazalo, da tega brez dovoljenj za objavo zvočnih posnetkov ne morejo storiti. Naknadno je bilo dovoljenja praktično nemogoče pridobiti; dobili bi jih lahko od večine »izbranih« govorcev, ki so besedila snemali, ne pa od njihovih sogovorcev, ki jih zaradi postopkov anonimizacije ni bilo več mogoče odkriti. Govorna komponenta korpusa BNC tako ostaja dostopna samo v transkribirani obliki.

Gre torej za problematiko, pri kateri morajo načrtovalci korpusa skrbno postopati in razmišljati tudi vnaprej. Vsekakor je treba pridobiti pisna dovoljenja govorcev tako za uporabo zvočnih posnetkov kot transkripcij. Poleg tega je treba v okviru državne zakonodaje preveriti vse morebitne zadržke in pridobiti morebitna dovoljenja. Pri načrtovanju govornega korpusa je treba predvideti tudi morebitna dovoljenja za objavo video posnetkov, nenazadnje pa v okviru državne zakonodaje tudi preveriti, kako je z dovoljenji za posnetke, objavljene v javnih medijih: ali dovoljenje medija zadostuje za vključitev v korpus ali je treba pridobiti še individualna dovoljenja posameznih govorcev.

3.2 Obstoječe sheme zajemanja besedil v govorne korpuse

3.2.1 Besedilnovrstna sestava korpusa London-Lund

V nadaljevanju si bomo ogledali, kako so zgrajeni nekateri najbolj znani govorni korpusi. Najprej bom predstavila besedilnovrstno sestavo najstarejšega govornega korpusa London-Lund. Randolph Quirk je pri snovanju predračunalniškega korpusa SEU skušal vanj zajeti vse vrste govorjenih besedil, ki se pojavljajo v realnem govoru. Sto besedil je bilo razporejenih v naslednje kategorije:



Slika 10: Shematski prikaz tipologije besedil v govorni komponenti korpusa SEU

Hierarhično je bila na vrhu delitev besedil na dialoge in monologe. Dialogi so se naprej delili glede na okoliščine na zasebne in javne. »Javna diskusija« je dialog, ki se odvija pred poslušalci, ki se vanj ne vključujejo. To je lahko npr. intervju pred poslušalci ali diskusija/omizje/intervju na radiu oz. televiziji (Greenbaum 2003,

2). »Spontani monolog« je relativno nepripravljen govor, npr. športni komentar, komentar različnih dogodkov (npr. državne proslave), pa tudi nepripravljen govor v parlamentarni razpravi in podobno. »Pripravljeni monolog« naj bi bil bližje pisnemu jeziku, vendar še vedno vključuje določeno mero spontanosti; to so npr. predavanja, pridige, sodni govori itd. (Greenbaum: prav tam).

Podrobnejši popis vseh zajetih besedil je objavljen v knjižni izdaji korpusa,[70] dostopen pa je tudi na spletnih straneh korpusa London-Lund; nekateri izseki iz originalnega popisa besedil so prikazani v nadaljevanju.[71]

```
THE LONDON-LUND CORPUS OF SPOKEN ENGLISH
INFORMATION ABOUT THE TEXTS

S.1 - S.3 35 texts, c. 175,000 words. Spontaneous, surreptitiously
   recorded conversations between intimates and distants.
            S.1-2: Conversations between equals
            S.3:   Conversations between disparates

-----------------------------------------------------------------

S.5       13 texts, c. 65,000 words.
S.5.1-7:  Non-surreptitious public conversation between equals
          (radio discussions)
S.5.8-11: Non-surreptitious private conversations between equals
S.5.12:   Committee meeting
S.5.13:   Academic meeting
```

Slika 11: Izsek iz originalnega popisa besedil korpusa SEU (London-Lund)[72]

Iz predstavljenih delov je razvidno, da gre za natančen in nazoren popis besedil, vključenih v korpus. Najprej so označene večje enote, ki označujejo vrste besedil, S.1–S.3 npr. spontani, skrivaj posneti dialogi (35 besedil, 35 odstotkov vseh besedil). Označeno je tudi število besedil in število besed v posamezni enoti, sledijo pa oznake posameznih besedil, ki imajo poleg tehnične oznake (npr. S.11.1) tudi oznako, ki opredeljuje tematiko in pragmatično funkcijo besedila (navzkrižno

[70] J. Svartvik in R. Quirk (ur.), *A Corpus of English Conversation*, Lund Studies in English 56, Lund: Gleerups/Liber, 1980.

[71] Pri opisu besedil v originalnem besedilu se pojavljajo pojmi *intimates*, *distants*, *equals*, *disparates*, ki pa niso podrobneje pojasnjeni oz. definirani. Iz popisa govorcev sklepam, da gre pri *equals* za osebe, ki so v enakovrednem družbenem položaju (npr. pogovor med dvema univerzitetnima profesorjema, dvema tajnicama ipd.), v nasprotju z *disparates*, ki so v neenakovrednem družbenem položaju.

[72] http://helmer.aksis.uib.no/icame/london-lund/

zasliševanje, pridiga, predavanje itd.). Razmerje med dialogom in monologom korpusa SEU je 76 : 24 odstotkov.

Tako kot opis besedil obstaja tudi popis govorcev korpusa SEU. Veže se na vsako posamezno besedilo (npr. S.1.1), vsebuje pa datum posnetka in demografske informacije o govorcih. Pri govorcih razlikujejo spol, izobrazbo in starost. Del originalnega popisa govorcev je prikazan na spodnji sliki:

```
INFORMATION ABOUT THE SPEAKERS

Text number      Speakers
Year of          When speaker identity symbol differs from
recording        the original SEU version, the latter is
                 given in brackets.

S.1.1            A          male academic, age c. 44
(1964)           B          male academic, age c. 60

S.1.2:
S.1.2            A          male academic, age c. 43
(1963)           B          male academic, age c. 42
S.1.2a           A          male academic, age c. 45
(1965)           B          male academic, age 41
                 CAL        telephone caller
S.1.2b           A          male academic, age 45
(1965)           B          male academic, age 36
```

Slika 12: Izsek iz originalnega popisa govorcev korpusa SEU (London-Lund)[73]

Vidimo lahko, da so bili demografski podatki govorcev upoštevani že pri gradnji prvega govornega korpusa, čeprav niso služili kot kriterij za zajem besedil v korpus in zato tudi niso bili uravnoteženi glede na celotno populacijo. Vsekakor jih je mogoče in potrebno upoštevati pri analizah, ki temeljijo na korpusu. Ker so bili govorci korpusa večinoma visoko izobraženi, je danes korpus London-Lund v mednarodnem arhivu korpusov angleščine ICAME[74] označen kot »korpus, ki vsebuje primere govora izobražencev britanske angleščine«.

[73] http://helmer.aksis.uib.no/icame/london-lund/
[74] *International Computer Archive of Modern and Medieval English*, http://icame.uib.no/.

Čeprav so se metode zajemanja besedil v govorni korpus po objavi obsežne dokumentacije o gradnji korpusa BNC zelo spremenile, pa je zanimivo, da so v zadnjem času spet nastali nekateri veliki govorni korpusi, ki v izhodišče zajemanja postavljajo samo besedilo. Tako je bil npr. pri gradnji korpusa govorjene nizozemščine (2004) osnovni princip za zajem besedil določitev okoliščin, v katerih se jezik uporablja, podrobnosti pa so predstavljene v nadaljevanju.

3.2.2 Besedilnovrstna sestava Nizozemskega govornega korpusa

Načrtovana shema za zajem besedil v govorni korpus nizozemščine je bila naslednja (številke pomenijo načrtovano število besed):

<table>
<tr><td rowspan="6">dialog/
multilog
8.110.000</td><td colspan="2" rowspan="2">zasebni
6.635.000</td><td rowspan="2">brez
pisne
predloge</td><td>neposredno</td><td>spontana konverzacija</td></tr>
<tr><td>intervjuji</td></tr>
<tr><td colspan="2" rowspan="2"></td><td rowspan="2">na daljavo</td><td>telefon. konverzacija</td></tr>
<tr><td>poslovne transakcije</td></tr>
<tr><td rowspan="2">javni
1.475.000</td><td>v medijih</td><td colspan="2">s pisno predlogo</td><td>intervjuji in diskusije</td></tr>
<tr><td>neposredno</td><td colspan="2">brez pisne predloge</td><td>diskusije, sestanki, debate</td></tr>
<tr><td colspan="4"></td><td>predavanja</td></tr>
<tr><td rowspan="6">monolog
1.890.000</td><td colspan="2">zasebni, 40.000</td><td colspan="2">s pisno predlogo</td><td>opis slik</td></tr>
<tr><td rowspan="5">javni
1.850.000</td><td rowspan="3">v medijih</td><td colspan="2">brez pisne predloge</td><td>spontani komentar</td></tr>
<tr><td rowspan="2">s pisno predlogo</td><td>poročila</td></tr>
<tr><td>novice</td></tr>
<tr><td rowspan="2">neposredno</td><td rowspan="2">s pisno predlogo (pripravljen)</td><td>komentarji</td></tr>
<tr><td>predavanja, govori</td></tr>
<tr><td colspan="4"></td><td>brano besedilo</td></tr>
</table>

Slika 13: Načrtovana taksonomija besedil v govornem korpusu nizozemščine (1998)[75]

[75] http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm

V skrajnem desnem stolpcu tabele lahko razberemo, da so načrtovali 14 vrst besedil. Število besed se je od vrste do vrste razlikovalo glede na to, v kolikšni meri je bila besedilna vrsta navznoter diferencirana. V končni verziji korpusa je bil nabor besedilnih vrst nekoliko spremenjen, morda lahko v spremembah zaznamo premike v smeri poenostavitve zajema:

1.	Spontana konverzacija (neposredni stik)
2.	Intervjuji z učitelji nizozemščine
3.	Spontani telefonski pogovori (posneti v različnih centralah)
4.	Spontani telefonski pogovori
5.	Simulirana poslovna pogajanja
6.	Intervjuji/diskusije/debate (v medijih)
7.	(Politične) diskusije/debate/sestanki (neposredni stik)
8.	Učne ure v razredu
9.	Komentarji "v živo" (v medijih)
10.	Novice/reportaže (v medijih)
11.	Poročila (v medijih)
12.	Komentarji/pregledi (v medijih)
13.	Slavnostni/žalni govori
14.	Predavanja/seminarji
15.	Brana besedila

Slika 14: Realizirana taksonomija besedil v govornem korpusu nizozemščine (2004)[76]

Podobno je bila tudi taksonomija besedil korpusa C-ORAL-ROM (2003) zgrajena samo na besedilnih kriterijih, medtem ko razlike med govorci niso bile upoštevane kot kriterij za zajem besedil; informacije o starosti govorcev, spolu, izobrazbi in geografskem izvoru so bile vedno označene kot metajezikovni podatki v glavi besedila.

[76] http://lands.let.kun.nl/cgn/doc_English/topics/design/design.htm

3.2.3 Uveljavitev demografske klasifikacije govorcev

3.2.3.1 Demografska komponenta BNC

Pri gradnji govorne komponente korpusa BNC so kot kriterij za zajem besedil v korpus prvič uporabili demografsko klasifikacijo govorcev. Govorno komponento korpusa BNC je v celoti izdelalo založniško podjetje Longman, sestavljena pa je iz t. i. demografske komponente (glede na razlikovanje govorcev) in kontekstualne komponente (glede na razlikovanje vrst besedil). Za demografsko komponento govornega korpusa so s pomočjo statističnih metod določili reprezentativni vzorec govorcev britanske angleščine glede na spol, starost, regijsko pripadnost in socialni razred. Nadaljevanje projekta je prevzelo podjetje za raziskavo tržišča, ki je pridobilo 153 prostovoljcev (starejših od 15 let), ki so predstavljali reprezentativni vzorec govorcev britanske angleščine.[77] Med govorci je bilo približno enako število moških in žensk, bili so z 38 različnih geografskih področij in enakomerno razporejeni v tri regijske skupine – severno, osrednjo in južno, v štiri starostne skupine in v štiri socialne razrede. Kasneje so v korpus dodali še posnetke najstnikov, starih 16 let ali manj, ki so jih zbrali v okviru korpusnega projekta COLT. Tako je nastal t. i. demografski del govornega korpusa BNC (imenovan tudi konverzacijski podkorpus), ki obsega 4.206.058 besed, to je pribl. 40 odstotkov celotne govorne komponente korpusa BNC.

Spodnja tabela prikazuje razporeditev gradiva, zbranega po demografski klasifikaciji:[78]

Spol	Št. govorcev	Št. besed	%
Neznano	5	16.151	0.38
Moški	73	1.730.592	41,14
Ženske	75	2.459.315	58,47

Starostna skupina	Št. govorcev	Št. besed	%
0-14	26	265.382	6,30
15-24	36	660.847	15,71
25-34	29	848.162	20,16
35-44	22	839.622	19,96
45-59	20	957.382	22,76
60+	20	634.663	15,08

[77] Načrtovalci korpusa so si želeli vključiti večji vzorec govorcev, npr. 1000, vendar to zaradi finančnih in časovnih omejitev ni bilo mogoče. Izračunali so, da bi bil vzorec 100 govorcev že lahko reprezentativen, vendar so zaradi kombiniranja štirih demografskih kriterijev v ustreznih razmerjih morali vzorec povečati na 153 govorcev (*British National Corpus User Reference Guide* 2000, 4).

[78] *British National Corpus User Reference Guide* 2000, 4.1.1.

Regijska pripadnost[79]			
Severno območje		2.765.729	26,74
Osrednje območje		2.471.184	23,89
Južno območje		4.658.232	45,04
Neopredeljeno		446.584	4,31
Socialni razred[80]			
Neznano	7	37.363	0,88
AB	59	1.363.571	32,41
C1	36	1.097.023	26,08
C2	31	1.080.654	25,69
DE	20	627.447	14,91

Tabela 7: Razporeditev govorcev v demografski komponenti BNC

»Izbrani govorci«[81] so od dva do sedem dni snemali vse svoje pogovore. Na ta način je bilo posnetih skupno pribl. 700 ur konverzacije, od česar je bilo pribl. 630 ur vključenih v demografski del govornega korpusa. Številke v tabelah, ki izkazujejo dokajšnjo uravnoteženost, se nanašajo na izbrane govorce. Po statistični obdelavi celotnega konverzacijskega podkorpusa pa se je izkazalo, da so se ta razmerja precej porušila ali zameglila, kar je postalo eden največkrat kritiziranih delov korpusa BNC. Tako se je npr. razmerje med moškimi in ženskami, ki je bilo v izhodišču zelo uravnoteženo (73 moških, 75 žensk, 5 neznanih), močneje prevesilo na stran žensk, ko so prešteli vse govorce v konverzacijskem podkorpusu (536 moških, 561 žensk), in se praktično porušilo, ko so prešteli vse izgovorjene besede glede na spol govorcev (1.714.443 moški, 2.593.452 ženske). To pomeni, da je bilo na vsakih sto besed podkorpusa, ki so jih izgovorili moški, izgovorjenih 151 besed, ki so jih izgovorile ženske (Rayson et al. 1997, 136). Takšno prešteva nje govorcev in besed se zdi mogoče na prvi pogled brezpredmetno, vendar ni tako. Če govorce klasificiramo na moške in ženske, je to gotovo z namenom, da bi raziskovali morebitne razlike v govoru enih in drugih; za take raziskave je treba imeti v vzorcu reprezentativno razmerje skupin govorcev, sicer nam podatek,

[79] Navedeni podatki za regijsko pripadnost besedil so skupni za demografsko in kontekstualno komponento (*British National Corpus User Reference Guide* 2000, 4.2.3).

[80] Populacijska stratifikacija socialnih razredov v Veliki Britaniji: A: visoki srednji razred (3 %), B: srednji razred (15 %), C1: nižji srednji razred (23 %), C2: delavski razred s kvalifikacijami (27 %), D: delavski razred (18 %), E: prejemniki socialnih pomoči (14 %) (prim. Gorjanc 2002, 46). Standardna populacijska stratifikacija socialnih razredov v Veliki Britaniji razrede opredeljuje glede na poklic nosilca gospodinjstva, pri čemer naj bi npr. razred A predstavljali najvišji vodilni delavci, razred E pa poleg delavcev najnižjih kategorij še prejemniki državne podpore in vdove brez lastnega zaslužka. Rayson (1997, 151, op. 7) navaja poklicno tipologijo teh razredov; ta naj bi se pogosto uporabljala pri raziskavah tržišča in pri drugem raziskovalnem delu; kljub temu je to le ena izmed obstoječih stratifikacijskih raziskav za Veliko Britanijo.

[81] V angleškem besedilu so prostovoljce, ki so se vključili v projekt snemanja konverzacijskega podkorpusa, poimenovali *respondents*, in jih tako ločili od ostalih govorcev, vključenih v pogovor, imenovanih *speakers*.

da so npr. moški govorci besedo *fuck* v korpusu uporabili 1.401-krat, ženske pa 325-krat, ne da relevantne informacije o rabi te besede, prav tako ne podatek, da so ženske besedo *she* izgovorile 22.623-krat, moški pa 7.134-krat (raziskava o demografski diferenciaciji rabe besed pri britanskih govorcih, Rayson, Leech in Hodges 1997).[82]

Čeprav so imeli izbrani govorci nalogo, da natančno popišejo demografske podatke svojih sogovorcev, je bilo pri tako veliki količini gradiva to razumljivo težko v celoti in natančno izvesti. Kasnejše analize korpusa so pokazale, da označevanje ni bilo izvedeno v enaki meri (do enake natančnosti) za vse govorce in za vsa besedila. Tako je spol znan za 81 odstotkov govorcev, starost za 42 odstotkov, socialni razred za 10 odstotkov, izobrazba za 9 odstotkov in prvi jezik za 58 odstotkov govorcev; 12 odstotkov govorcev je v celoti neidentificiranih. Za kategorijo »socialni razred«, ki je že tako ali tako označena pri izredno nizkem odstotku govorcev, naj bi veljalo, da se da na oznako pravzaprav v resnici zanesti le pri 153 izbranih govorcih (Berglund 1999, 45), kar odstotek še bistveno zmanjša. Vse našteto je treba pri uporabi podatkov iz korpusa upoštevati.

Namen gradnje korpusa BNC je bil v korpus zajeti jezik, ki je reprezentativen za celotno izbrano populacijo. Ob tem se je sestavljavcem že na začetku zastavljalo vprašanje, kateri jezik je pravzaprav reprezentativen, tisti, ki ga sprejemamo (beremo in poslušamo), ali tisti, ki ga produciramo (pišemo in govorimo). Kot »dobri anglosaksonski pragmatiki so se odločili, da si bodo prizadevali upoštevati obe perspektivi« (Burnard 2000, 5), in so poskušali določiti kriterije zbiranja besedil tako, da bi v korpus zajeli »vse znane oblike besedil glede na vse znane in določljive kriterije« (Burnard: prav tam).

Govorjena besedila, zbrana z demografsko metodo, so sicer v (relativno) reprezentativnem razmerju predstavljala celotno populacijo govorcev, poleg tega konverzacija v vsakdanjem okolju v resnici predstavlja največji (čeprav ni znano, kolikšen natančno) delež govorjenih besedil v realnosti. So pa iz nabora izpadla besedila, ki jih izbranih 153 govorcev ni produciralo, npr. predavanja, pridige, sodni govori, televizijski intervjuji, to so besedila, ki jih večina govorcev predvsem sliši, producira pa jih manjšina; gre za besedila, ki zaradi okoliščin, v katerih so govorjena, in govorcev, ki jih govorijo, pogosto obveljajo za v jezikovnem smislu prestižna.[83] Demografski del govorne komponente korpusa BNC so zato dopolnili s kon-

[82] Razmerje so raziskovali še naprej in ugotovili, da je bila nasprotno v kontekstualnem delu govorne komponente tehtnica močno premaknjena v nasprotno smer, saj so kar 3.199.812 besed izgovorili moški, ženske pa le 660.071 besed (Berglund 1999, 49, op. 7); razmerje vseh besed glede na spol v govornem delu BNC je 4.914.255 za moške in 3.253.523 za ženske.

[83] Lahko predpostavljamo (raziskav o tej problematiki zaenkrat ni na voljo), da imajo določeni govorci vpliv na druge ali obratno, da si nekateri govorci želijo posnemati druge (morda izpostavljene) govorce; vprašanje je, v kolikšni meri in na kateri ravni je to mogoče – na ravni retorične organiziranosti, leksike, naglasa …

tekstualnim delom, v katerega je bilo vključenih 757 besedil, skupaj 6.135.671 besed oz. pribl. 60 odstotkov govorne komponente korpusa BNC. V izhodišču taksonomije je bila delitev na štiri enako velika tematska področja (spodnja tabela) – izobraževanje in informiranje, poslovna/poklicna komunikacija, javni oz. institucionalni govor ter zabava in prosti čas. Vsako področje je bilo navznoter razdeljeno na monologe (40 %) in dialoge (60 %), kar pomeni, da monologi znotraj posameznega področja predstavljajo 10 odstotkov v celotnem kontekstualnem delu govorne komponente BNC. Znotraj teh določil so bila besedila razvrščena v različne besedilne vrste, število vrst znotraj področij pa ni bilo vnaprej določeno in se je v končni realizaciji tudi precej razlikovalo:

Področje	Besedilna vrsta	Št.besedil	Št. besed	%
Izobraževanje in informiranje	predavanja	169	1.633.303	26,61
	komentarji			
	interakcija v razredu			
Poslovna/poklicna komunikacija	govori v podjetjih, intervjuji	131	1.285.938	20,95
	govori na kongresih, konferencah			
	prodaja			
	poslovni sestanki			
	svetovanje (zdravniško, sodno itd.)			
Javni oz. institucionalni govor	politični govori	262	1.655.263	26,97
	pridige			
	javni/vladni govori in sestanki			
	sestanki lokalnih skupnosti			
	verska srečanja			
	parlamentarni govor			
	sodni postopki			
Zabava in prosti čas	govori	195	1.561.167	25,44
	športni komentarji			
	govori v klubih			
	TV in radio – kontaktne oddaje			
	klubski sestanki			

Tabela 8: Področja, besedilne vrste in razmerja med njimi v kontekstualnem delu BNC[84]

[84] *BNC Users Reference Guide 2000*, 4.2.1.1.

Metoda zbiranja oz. zajemanja besedil se je od vrste do vrste razlikovala, prav tako se je razlikovalo število besedil znotraj vrst (od 3 do 6). Število besedil znotraj področja se je gibalo med 131 in 162. Glede na to, da je demografsko zbrani del govorne komponente BNC obsegal samo dialoge, je bilo skupno razmerje med monologi in dialogi v korpusu BNC naslednje:[85]

Tip interakcije	Št. besedil	Št. besed	%
Monolog	212	1.578.614	15.26
Dialog	698	8.763.115	84.73

Tabela 9: Razmerje med monologom in dialogom v govorni komponenti BNC[86]

Tudi v kontekstualnem delu korpusa so, čeprav je bilo izhodišče za zbiranje gradiva drugačno, poskusili ohraniti enakomerno razmerje med govorci v smislu demografske klasifikacije. Posebna pozornost je bila namenjena spolu govorcev in regijski pripadnosti. Pri spolu so se znotraj vseh kategorij trudili variirati spol govorca; pri beleženju regijske pripadnosti pa je zanimivo, da je niso pripisovali govorcem, ampak kraju, kjer je bil posnetek narejen; tako so vnaprej izbirali univerze z različnih geografskih področij, enako tudi šole, podjetja, upoštevali so nacionalne in regionalne radijske in televizijske postaje, lokalne, regionalne in centralne veje oblasti itd.

Vsi navedeni podatki govorijo o tem, kako težko je uravnotežiti korpus glede na demografsko strukturo govorcev ter hkrati na vse vrste besedil in razmerja med njimi. Naj so se načrtovalci korpusa BNC v izhodišču še tako trudili, so se razmerja zaradi izjemne zahtevnosti zbiranja podatkov na koncu podirala. Kljub temu govorna komponenta BNC še danes velja za enega najbolj reprezentativnih in uravnoteženih govornih korpusov na svetu.

[85] *BNC Users Reference Guide* 2000, 4.2.3.

[86] Zanimivo je, kako so v dveh letih popravili statistično sliko korpusa BNC; Aston in Burnard 1998 (cit. po Gorjanc 2002, 45) navajata naslednje razmerje med dialogom in monologom:

Tip interakcije	št. besed	%
Monolog	1.932.225	18,64
Dialog	7.760.753	74,87
Neopredeljeno	672.486	6,48

Spremembe so nastale ob redakciji korpusa l. 1999, pet let po prvi objavi, ko je dokončno postal dostopen širšemu krogu uporabnikov (na začetku samo znotraj EU). Takrat je bilo iz korpusa odstranjenih pribl. 50 besedil, za katere ni bilo mogoče dobiti avtorskih pravic, poleg tega pa so bila nekatera besedila na novo označena oz. prešteta (Burnard 2000, 14).

3.2.3.2 Demografsko vzorčenje v govorni zbirki POLIDAT

Po objavi dokumentacije o gradnji korpusa BNC so začeli načrtovalci korpusov pogosto posnemati metode zajemanja besedil, kot jih je uvedel BNC, seveda z določenimi prilagoditvami. Tako so npr. govorno komponento ČNK, korpus praškega govora, zgradili samo na podlagi demografske klasifikacije govorcev, prav tako tudi korpus najstniške britanske angleščine COLT. Tudi slovenska govorna zbirka POLIDAT, ki je bila zgrajena po priporočilih, definiranih v okviru evropskega projekta SpeechDat II, je govorce razdelila glede na demografske kriterije. Načrtovalci so pri tem upoštevali spol in starost govorcev, njihovo regijsko pripadnost in izobrazbeno strukturo. Po spolu so govorce razdelili natančno na polovico. Kriterije za starost govorcev so razdelili, kot prikazuje spodnja tabela:

Najmanj 20 % govorcev	16–30 let
Najmanj 20 % govorcev	31–45 let
Najmanj 15 % govorcev	46–60 let

Tabela 10: Razdelitev govorcev glede na starost v govorni zbirki POLIDAT[87]

Pri razdelitvi na regijske skupine je evropsko priporočilo narekovalo, naj vsako izbrano narečno področje zajema najmanj 1 odstotek populacije. Sestavljavci govorne zbirke POLIDAT so pri določitvi narečij izhajali iz karte slovenskih narečij (Logar 1993) in dodali govor Maribora in Ljubljane. Razdelitev govorcev po izbranih področjih RS je bila naslednja:

Narečno področje	Št. govorcev	%
Panonsko	160	11,43
Štajersko	310	22,14
Koroško	50	3,57
Dolenjsko	190	13,57
Kočevsko	20	1,43
Rovtarsko	60	4,28
Gorenjsko	160	11,43
Primorsko	110	7,86
Maribor	110	7,86
Ljubljana	230	16,43

Tabela 11: Razdelitev govorcev glede na narečna področja v govorni zbirki POLIDAT[88]

[87] Zögling Markuš in drugi 2000, 96.
[88] Zögling Markuš in drugi 2000, 96.

Število govorcev s posameznega področja je bilo določeno glede na število prebivalcev na danem področju, predvidevamo torej, da gre za statistično reprezentativni vzorec glede na celotno populacijo (pri spolu ni tako, ker razmerje med moškimi in ženskami v RS ni ena proti ena). Seveda pa pri regijskem vzorcu problematika še ni do konca izčrpana, saj se npr. postavlja vprašanje, kdo se šteje za prebivalca določene regije in pod kakšnimi pogoji (preseljevanje itd.).

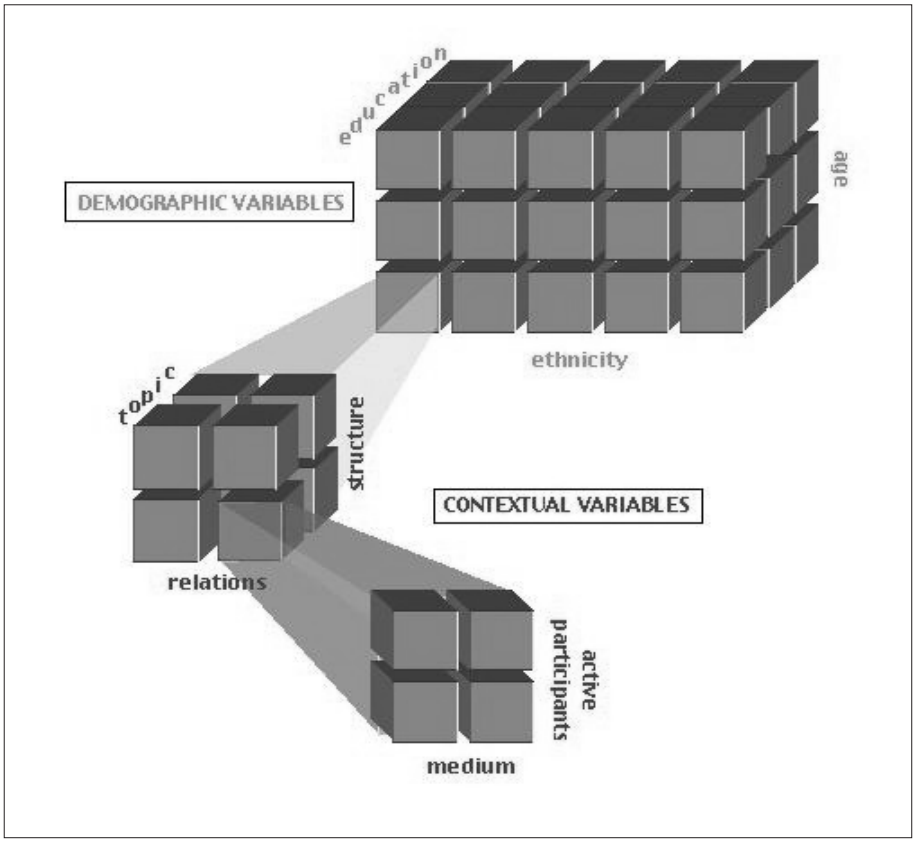
Pri izobrazbeni strukturi so v govorni zbirki POLIDAT postavili mejnike na končano osnovno šolo, poklicno šolo, srednjo šolo, visoko šolo, magisterij in doktorat, kar se zdi za gradnjo večjega korpusa nekoliko pretirana specifikacija, ker je prenatančna in s tem prezahtevna za zbiranje. In nenazadnje: ker so gradili govorno zbirko za vzpostavitev sistema za razpoznavo govora po telefonu, so za kriterij zbiranja uvedli tudi različne situacije, v katerih poteka telefonski pogovor: telefonska govorilnica, dom, pisarna, tovarna, javni prostor, cesta, vozilo in ostalo. S tem zadnjim kriterijem so dopolnili demografsko zbiranje podatkov in zagotovili raznovrstnost besedilnih vrst, čeprav v tem primeru ni šlo za običajno kombiniranje z besedilnovrstno taksonomijo, ampak za zagotavljanje besedil z različnimi šumi v ozadju, kar je glede na uporabni cilj govorne zbirke pričakovano.

3.2.4 Nov poskus kombinacije demografske in besedilnovrstne metode

Videli smo, da se demografska in besedilnovrstna metoda zajemanja podatkov pogosto kombinirata v smislu komplementarnosti. Seveda pa je mikavna misel, da bi ju poskusili združiti v eno samo metodo, ki bi zajem besedil gradila hkrati na demografski stratifikaciji govorcev in na besedilnovrstni taksonomiji. To idejo poskušajo realizirati v okviru gradnje korpusa govorjene izraelske hebrejščine (CoSIH).[89] Sestavljavci so se na podlagi preteklih izkušenj pri gradnji korpusov, predvsem BNC, odločili kombinirati demografsko in besedilnovrstno komponento govornega korpusa. Kot konceptualno orodje so si v ta namen zamislili večdimenzionalno celično matriko. Ogrodje je bila matrika velikosti 5 x 3 x 3, ki je temeljila na demografskih komponentah, in sicer etnični pripadnosti (5 kategorij), starosti in izobrazbi (vsaka po tri kategorije). Vsaka izmed dobljenih 45 celic je bila navznoter sestavljena iz matrike velikosti 2 x 2 x 2, ki so jo opredeljevali besedilnovrstni kriteriji (odnosi med govorci: formalni/neformalni, struktura pogovora: vodeni pogovor/interakcija, tema: zasebno, nezasebno). Znotraj tako koncipiranih celic pa je bila nova matrika 2 x 2, ki sta jo določali komponenti

[89] *The Corpus od Spoken Israeli Hebrew*, http://www.tau.ac.il/humanities/semitic/cosih.html.

monolog/dialog in medij: telefon/neposredno. Na ta način so snovalci korpusa hebrejščine sestavili matriko s 1000 celicami, ki si jih nameravali zapolniti z besedili; vsaka celica naj bi predvidoma vsebovala 5.000 besed, tako bi dobili 5-milijonski korpus. V idealnih razmerah bi to pomenilo, da bi bili demografsko reprezentativni predstavniki celotne populacije posneti v vseh besedilnovrstnih različicah. Pri načrtovanju korpusa so se zavedali, da bo idealno matriko težko v celoti zapolniti s predvidenimi tipi besedil, zato so v korpus vgradili tudi varovalne mehanizme, ki naj bi prispevali k reprezentativnosti.[90]



Slika 15: Matrična struktura zajema besedil v korpus govorjene izraelske hebrejščine[91]

[90] Bistvo varovalnega mehanizma je bil 5-odstotni del korpusa, ki ga je predstavljala samo kontekstualna komponenta, sestavljena iz besedil, ki se pri demografskem vzorčenju morda ne bi pojavila (mediji, parlament, sodišče), imajo pa velik doseg.

[91] http://www.tau.ac.il/humanities/semitic/cosih.html

Ni znano, do katere stopnje so načrtovalci korpusa projekt uspeli izvesti; spletne strani ne obnavljajo že od l. 2005, o načrtovanem korpusu govorjene izraelske hebrejščine pa ni novih poročil.

***

Pri načrtovanju govornega korpusa predstavlja študij gradnje drugih korpusov koristno izhodišče, vendar pa se nobene izmed metod zbiranja ne da brez sprememb prenesti v drugo okolje. Nekritično povzemanje znanih in že uporabljenih kriterijev za nabor besedil in njihov neposredni prenos na novo jezikovno situacijo bi bilo nespametno dejanje. Vsaka jezikovna situacija namreč izkazuje (poleg skupnih, splošnejših) tudi specifične lastnosti, ki morajo biti upoštevane pri gradnji reprezentativnega in uravnoteženega korpusa. Tako npr. britanske razdelitve celotnega govornega območja na severno, osrednjo in južno regijo ne moremo prenesti na slovensko govorno situacijo, saj glede na znane podatke o narečnih skupinah s to delitvijo ne bi dobili niti približno reprezentativnega vzorca govorcev. Enako nemogoč je prenos britanskega socialnega razreza družbe, ki temelji na poklicu nosilca gospodinjstva, na slovenske razmere, kjer pogosto ni mogoče opredeliti enega samega nosilca gospodinjstva. V nadaljevanju bom poskušala podati na slovenski jezikovni situaciji (in družbi) temelječa priporočila za zajem besedil v korpus govorjene slovenščine.

3.3 PREDLOG PRIPOROČIL ZA ZAJEM BESEDIL V KORPUS GOVORJENE SLOVENŠČINE

Splošna priporočila, ki zadevajo zajem besedil v reprezentativni uravnoteženi govorni korpus, lahko strnemo v naslednje točke:

- zanima nas predvsem spontani govor, zato ne načrtujemo zajemanja branih besedil,[92]
- glede na to, da je spontani govor v realnosti najpogostejša oblika govorjenih besedil, mora znotraj korpusa predstavljati večinski delež,
- nekateri besedilni tipi izkazujejo večjo notranjo diferenciranost (npr. dialogi v primerjavi z monologi), zato naj bodo v korpusu zastopani z večjim številom besedil,
- v korpusu ohranjamo naravno dolžino besedil.

To so načelna izhodišča, ki omogočajo dokaj fleksibilno načrtovanje zajema besedil. Znotraj tega se je treba odločiti, katere komponente bodo služile kot kriteriji

[92] Če se ta v govoru nenačrtovano vendarle pojavijo, jih ustrezno označimo.

zajemanja. Ob tem pa ne smemo pozabiti, da je gradnja korpusa ciklični proces, ki zahteva nenehno evalviranje strategije zajemanja in po potrebi spreminjanje začetnih izhodišč.

Za načrtovalce govornih korpusov sodijo med najtežje odločitve tiste, ki zahtevajo odgovor v številkah ali deležih. Preprosto zato, ker odgovori pogosto vsaj deloma slonijo na intuitivnih odločitvah, čeprav tudi te lahko izhajajo iz poznavanja jezikovne situacije, jezikoslovne problematike in vedenja o drugih korpusih. Načrtovalci se morajo odločiti, katere besedilne vrste izbrati, koliko besedil naj bo znotraj posamezne besedilne vrste in kakšna naj bodo razmerja med posameznimi besedilnimi tipi. Teh vprašanj v fazi načrtovanja korpusa načrtovalci ne smejo pustiti neodgovorjenih, saj bi oportunistična metoda zbiranja neizbežno vodila v neuravnoteženost in nereprezentativnost korpusa. Kljub konkretnemu načrtovanju korpusa pa je treba, kot že rečeno, kasneje dopuščati morebitna odstopanja od začetnih izhodišč, če bodo to zahtevale okoliščine zbiranja ali evalvacija v procesu gradnje korpusa.

3.3.1 Metode zajemanja

Glede na situacijo v slovenskem jezikovnem prostoru je najbolje upoštevati demografska in besedilnovrstna izhodišča pri zajemu besedil v korpus. Demografske lastnosti govorcev bi vsaj v delu korpusa morale biti zastopane tako, da bi govorci predstavljali reprezentativni vzorec celotne (izbrane) populacije. Za slovenščino obstaja nekaj raziskav, v katerih avtorji poskušajo analizirati razlike v govoru, ki nastanejo zaradi demografskih lastnosti govorcev, čeprav raziskave običajno niso nastale na podlagi velike količine avtentičnega gradiva; med take lahko štejemo npr. razprave *Sociolekti od izraza do pomena: kultiviranost, obrobje in eksces* (Skubic 2004), *Kdo govori kako* (Kržišnik 1997), *Seksizem kot jezikovnopolitični problem* (Stabej 1997), *Vpliv migracij na jezik in govor posameznika* (Guzej 1989/90) itd. Mnogo več je na Slovenskem dialektoloških študij, ki jih zaradi obsežnosti gradiva in včasih ozke specializiranosti na tem mestu ne bom navajala; med nekoliko bolj sintetičnimi pogledi naj navedem samo študiji *Narečje kot jezikovnozvrstna kategorija v sodobnem jezikoslovju* (Kenda Jež 2004) in *Nekaj resnic in zmot o narečjih v Sloveniji danes* (Smole 2004). Načrtovanje zajemanja besedil za slovenski govorni korpus samo na podlagi demografske klasifikacije govorcev pa ne bi dalo zadovoljivih rezultatov v smislu uravnoteženosti in reprezentativnosti, zato ga bo potrebno kombinirati z besedilnovrstno komponento.

V okviru demografske komponente se je treba najprej odločiti, katero populacijo zajeti v korpus. Kljub temu, da bi si želeli, da bi bil v korpusu zastopan

govor vseh govorcev slovenščine, bi bilo v prvi fazi gradnje verjetno treba sprejeti nekatere omejitve, v naslednjih fazah pa korpus dopolniti. Običajno se govor odraslih govorcev načrtuje ločeno od govora mladostnikov in otroškega govora.[93] Pri definiranju govorcev slovenščine bi se bilo treba opredeliti, ali vzorčimo samo govorce slovenščine kot prvega jezika ali upoštevamo tudi delež pripadnikov drugih jezikovnih skupnosti, za katere slovenščina ni prvi jezik, in v kolikšni meri.[94] Pri regijski pripadnosti govorcev se poleg regijske razdelitve osrednjega ozemlja zastavlja tudi vprašanje vključevanja govora govorcev, ki živijo zunaj meja slovenske države – v zamejstvu, zdomstvu ali izseljenstvu. Tudi v tem smislu na Slovenskem že obstaja precej raziskav, npr. *Usoda slovenskega jezika med Slovenci po svetu* (Šabec 2002), *Knjižnojezikovna norma v »argentinskoslovenski« Svobodni Sloveniji* (Kržišnik 2002), *Povojni generaciji Slovencev po svetu: narodnostna opredelitev, znanje in raba slovenščine* (Bešter 1996), *Slovenska jezikovna skupnost v Argentini: (socio)lingvistična analiza* (Špelko in Ban 2005), *Sociolingvistični problemi slovenske etnične skupnosti v Italiji* (Pogorelec 1989) idr. Seveda bi bil tudi za raziskovanje slovenskega govora v zamejstvu in diasporah (govorni) korpus neprecenljivega pomena, vendar bi bilo treba morebitni podkorpus zasnovati s posebnim premislekom in na podlagi temeljite sociolingvistične (pred)študije.

Predlagam, da bi slovenski referenčni govorni korpus sestavili iz dveh komplementarnih podkorpusov: prvega bi oblikovali na podlagi demografske klasifikacije govorcev (konverzacijski podkorpus), drugega pa na podlagi taksonomije besedilnih tipov (besedilnovrstni podkorpus). Razmerje med obema podkorpusoma je spet vprašanje, na katerega je težko odgovoriti. Korpus BNC na primer ima med demografsko in kontekstualno komponento razmerje 40 : 60 odstotkov. Demografska komponenta zajema spontani govor govorcev, za katerega je gradivo na splošno težje zbrati kot za besedilnovrstni podkorpus, poleg tega pa je gradivo tudi težje transkribirati.[95] Spontana konverzacija je torej za raziskovanje izredno dragocena, jo je pa težje pridobiti. Kljub temu bi si morali prizadevati za enakovredno razmerje med obema podkorpusoma oz. bi moral konverzacijski podkorpus obsegati vsaj 40 odstotkov celotnega korpusa, tako kot je to v primeru BNC. Če bi gradili npr. enomilijonski korpus govorjene slovenščine, bi to pomenilo konverzacijski podkorpus velikosti 400.000 do 500.000 besed, kar bi po velikosti ustrezalo svetovni »drugi ligi« govornih korpusov. Konverzacijski podkorpus bi znotraj celote lahko predstavljal celotno dialoško zasebno in ostalo nejavno produkcijo (v neposrednem stiku

[93] Raziskovanje otroškega govora v jezikoslovju že dolgo sodi v posebno znanstveno vejo (pri nas npr. Kranjc 1999, Fekonja 2004 idr.); seveda je lahko tudi otroški govor s svojimi specifičnimi lastnostmi posebna komponenta korpusa oz. podkorpus (kot npr. najstniški korpus COLT v okviru BNC).

[94] Tudi korpus usvajanja jezika kot tujega ali drugega jezika je za raziskovanje in učenje jezika izjemno pomemben podkorpus (prim. Ferbežar in Pirih Svetina 2004b in Stritar 2006).

[95] Če predpostavljamo, da je v besedilnovrstnem podkorpusu veliko medijskega in nasploh javnega govora, ki je bolj artikuliran, manj je prekrivanja govorcev itd.

in po telefonu), tako da teh tipov besedil v besedilnovrstni podkorpus ne bi bilo treba vključevati. V besedilnovrstni podkorpus bi tako vključili predvsem dialoge in monologe, govorjene v javnosti (RA, TV, šolske in druge javne ustanove, cerkev, sodišče itd.), dopolnili pa bi ga tudi z besedili, ki jih v konverzacijskem podkorpusu ne bi bilo dovolj, npr. monologi v zasebnih okoliščinah.

3.3.2 Konverzacijski podkorpus

Konverzacijski podkorpus ali demografska komponenta govornega korpusa obsega spontani govor reprezentativnega vzorca izbranih govorcev. Demografski kriteriji za zajem besedil v govorni korpus so tisti kriteriji, za katere na podlagi obstoječih jezikoslovnih razprav in deloma hipotez pričakujemo, da pogojujejo razlike v govoru. Šele naknadna analiza korpusnih podatkov pa bo lahko pokazala, ali te razlike res obstajajo in kakšne so.

3.3.2.1 Spol in starost govorcev

Pri prvem demografskem kriteriju, to je spolu, je treba za reprezentativni vzorec govorcev upoštevati razmerje med moškimi in ženskami v celotni izbrani populaciji; pri tem si lahko pomagamo s popisom prebivalstva.[96] Nasprotno pa je pri drugem kriteriju, starosti, treba hipotetično določiti starostne intervale, v katerih naj bi med govorci prihajalo do občutnejših razlik v govoru. BNC je npr. določil 6 starostnih intervalov po pribl. 10 let (in od 60 let naprej enotno), ČNK samo dva intervala – nad ali pod 35 let, C-ORAL-ROM pa štiri intervale (18–25, 25–40, 40–60 in več kot 60). Pri načrtovanju razdelitve na starostne intervale za zajem besedil za slovenščino se ne moremo opreti na obstoječe raziskave, moja lastna jezikovna izkušnja pa govori v prid češke razdelitve, ki je tudi zelo racionalna. Število govorcev znotraj posameznega starostnega intervala v reprezentativnem vzorcu mora ustrezati številu govorcev znotraj izbranega starostnega intervala v celotni populaciji.

3.3.2.2 Izobrazba govorcev

Podobno je z izobrazbo govorcev: število intervalov je lahko poljubno, dejanska analiza razlik bo mogoča šele na podlagi zbranega gradiva. C-ORAL-ROM ima razdelitev na tri intervale – končana osnovna šola, srednja šola ali univerza; kot

[96] Podatek glede na popis prebivalstva RS za l. 2006 znaša 986.982 moških in 1.023.395 žensk (http://www.stat.si/letopis/index_vsebina.asp?poglavje=4&leto=2007&jezik=si).

vemo, BNC razlik v izobrazbi ni vzel za demografski kriterij, ki bi vplival na razlike v govoru. V slovenskem okolju je mogoče predvidevati, da število let šolanja vpliva na govor posameznika; v zvezi s tem predlagam razdelitev na 3 izobrazbene kategorije: govorci s končano osnovno šolo ali manj (do 8 let šolanja) so v prvi, govorci s končano srednjo, višjo ali visoko šolo (9–16 let šolanja) v drugi kategoriji ter govorci z univerzitetno izobrazbo (več kot 17 let šolanja in študija) v tretji. Znotraj prve kategorije bi bilo glede na statistične podatke iz leta 2006[97] pribl. 480.000 govorcev (28 % prebivalstva nad 15 let), v drugi pribl. 1.100.000 govorcev (64 %) in v tretji 140.000 govorcev (8 %). Mejo med drugo in tretjo kategorijo bi bilo mogoče premakniti tudi tako, da bi bili v drugi samo govorci s končano srednjo šolo; v tem primeru bi imeli v drugi kategoriji 56 % govorcev (srednja šola), v tretji pa 17 % govorcev (višja, visoka, univerzitetna izobrazba).

3.3.2.3 Regijski izvor govorcev

Tudi pri regijski pripadnosti govorcev si je potrebno v izhodišču zastaviti kar nekaj vprašanj. Najprej je treba znotraj (in zunaj) slovenskega prostora določiti območja, iz katerih bomo vzorčili. Možnost, ki se pri tem ponuja (in ki je bila npr. v veliki meri privzeta pri oblikovanju govorne zbirke SpeechDat) je razdelitev, ki se ujema z razdelitvijo na osem slovenskih narečnih skupin. Vendar v tej razdelitvi pogrešamo govor osrednje Slovenije oz. Ljubljane z okolico, ki ne pripada nobeni narečni skupini in ki ima zaradi koncentracije medijev in izobraževalnih institucij zelo močan vpliv na vse govorce slovenščine. Kot smo lahko videli, so pri oblikovanju SpeechData tudi govor Maribora z okolico izključili iz štajerskega narečja in ga dodali kot posebno kategorijo. Poleg tega vzorčenje vseh narečij ni namen referenčnega korpusa, saj ne nastaja z namenom raziskovanja specifičnih prozodičnih, naglasnih in fonetičnih lastnosti govora, ampak predvsem z namenom raziskovanja tipične leksike in struktur v govorjenem jeziku. Zato bi za regijsko vzorčenje predlagala nekoliko drugačno delitev, ki temelji na predpostavki, da se v bazenih okoli večjih mestnih središč (posledično izobraževalnih in drugih ustanov ter medijev) izoblikuje nek regionalni jezik, sicer navznoter heterogen, po vendarle z nekimi skupnimi prepoznavnimi lastnostmi.[98] Predlagam regijsko razdelitev na 5 skupin: osrednjo, SV, SZ, J in JV ter Z govorno skupino. Tu navajam zgolj grobo razdelitev na regijske skupine, ki jim pripada po eno ali več mestnih središč; podrobnejša določitev mej med regijami bi zahtevala podrobnejši študij problematike, kar pa presega namen tega dela.

[97] http://www.stat.si/letopis/index_vsebina.asp?poglavje=6&leto=2007&jezik=si

[98] Potrditev za takšno razmišljanje najdemo tudi v nekaterih sodobnih pogledih na narečja in regije (prim. Kenda Jež 2004).

Drugo vprašanje, ki se prav tako zastavlja v zvezi z regijsko pripadnostjo, je, ali govorce opredeljujemo glede na kraj rojstva, kraj bivanja ali celo kraj dela (šolanja) ter katero regijo upoštevamo pri preseljevanju. Pri zbiranju gradiva za govorno komponento ČNK, ki trenutno poteka na Češkem, npr. označujejo, če je le mogoče, kraj in regijo rojstva govorca, kraj bivanja med šolanjem in kraj bivanja v času nastanka posnetka. V slovenskih razmerah lahko v povprečju pri govorcu pričakujemo po dve regijski oznaki: prva zadeva kraj rojstva in osnovno šolanje, druga pa morebitno selitev zaradi nadaljevanja šolanja, poklica ali drugih osebnih razlogov. Zaradi racionalizacije bi torej lahko pri identifikaciji govorca vnašali dva podatka, ki zadevata regijsko pripadnost: regijski izvor in bivanje v času nastanka posnetka.

3.3.2.4 Družbeni status govorcev

Kriterij družbenega statusa, ki je bil privzet v nekaterih korpusih (BNC, COLT), je specifičen za angleško družbo, kjer je socialna razslojenost družbe izrazito povezana z govorico. Britanski model je na slovenske razmere neprenosljiv, saj ni jasno, kako definirati družbeni status prebivalstva: glede na življenjski standard, mesečne prihodke, položaj v družbi ali kako drugače; vprašanje je tudi, kako tovrstni status vpliva na razlike v govoru posameznikov.[99] Pri gradnji korpusa COLT (korpus najstniške angleščine) so socialni status govorcev določili na podlagi identifikacijskih listov govorcev, ki pa so jih izpolnjevali samo t. i. izbrani govorci, tako da je socialni položaj znotraj COLT-a določen samo zanje in za njihove starše (50 odstotkov vseh govorcev COLT-a). Podatek o socialnem položaju je bil določen na podlagi treh informacij z identifikacijskega lista govorca: podatka o tem, kje govorec živi (*residental area*), poklica staršev in podatka o tem, ali so starši zaposleni ali ne, metoda izračunavanja pa je bila prevzeta od statističnega urada Velike Britanije. Govorci so bili razporejeni v tri socialne razrede – nižjega, srednjega in visokega. Na Slovenskem poznamo razvrščanje v socialne razrede npr. pri izračunavanju cenzusa pri vpisovanju otroka v vrtec ali pri izračunavanju razredov za dohodninsko osnovo, vendar se zdi pridobivanje tovrstnih podatkov v lingvistične namene praktično neizvedljivo in tudi nesmiselno.[100]

[99] Prim. Skubic 2004.

[100] Zanimivo razpravljanje o socialnem izvoru najdemo pri Cindriču (2002, 30). V raziskavi o slovenskem izobraženstvu v preteklosti raziskuje mdr. socialni izvor študentov s Kranjske, vpisanih na dunajsko univerzo v prvi polovici 19. stoletja. O socialnem položaju študentov sklepa iz podatkov o stanu oz. poklicni dejavnosti staršev (očeta), po katerih so študente spraševali na vpisnem listu. Na eni strani so označevali pripadnost k skupini, in sicer meščanski, plemiški, kmečki stan, na drugi strani pa poklicno oz. službeno dejavnost – učiteljski, uradniški, vojaški, cehovski stan, lastnik podjetja, veletrgovec, imetnik nepremičnine itd. Ta zgodovinska klasifikacija slojev je na današnje razmere neprenosljiva, zanimivo pa je, da so tudi pri gradnji COLT-a za določanje socialnega izvora govorcev uporabili kriterij poklica staršev, prim. zgoraj.

3.3.2.5 Končni predlog demografskih kriterijev

Priporočila za zajem besedil v korpus govorjene slovenščine na podlagi demografskih kriterijev govorcev so:

Spol:	moški			ženske		
Starost:	35 let in manj (mlajši)			36 let in več (starejši)		
Dosežena izobrazba:	osnovna šola ali manj		srednja šola		višja+visoka+univ.	
Regijski izvor:	osrednja Sl.	S in SZ	Z	SV	J in JV	drugo
Prvi jezik:	slovenski					drugo

Tabela 12: Predlog kriterijev za demografsko klasifikacijo govorcev KGS

Poleg navedenih kriterijev, ki bi služili tudi kot omejitveni kriteriji za iskanje po korpusu (lahko se npr. išče samo znotraj populacije, stare več kot 35 let, s končano višjo ali univerzitetno izobrazbo, ženskega spola), bi bili pri identifikaciji govorcev v glavi korpusa navedeni in dostopni tudi drugi podatki, npr. o poklicu govorcev, o morebitnem bivanju v drugi regiji ipd.

3.3.3 Besedilnovrstni podkorpus

Drugo komponento korpusa govorjene slovenščine gradimo na podlagi sistematizacije vseh znanih vrst govorjenih besedil. Besedilnovrstne kriterije za zajem govorjenih besedil, ki smo jih v različnih razmerjih ter v različnih kombinacijah lahko opazovali v obravnavanih tujih govornih korpusih, lahko strnemo v naslednji seznam:

- stopnja spontanosti: spontani, pripravljeni, brani govor,
- prevladujoča struktura besedila: monolog, dialog, multilog,
- okoliščine: javno, zasebno, uradno, neuradno,
- govorni položaj: formalni, neformalni,
- prenosnik: osebni stik, telefon, mediji (avdio, video), internet,
- okoliščine snemanja: z vednostjo govorcev, brez vednosti govorcev,
- namen besedila,
- tematika besedila.

Kriteriji z medsebojnim prepletanjem predstavljajo mrežo za zajem besedil v besedilnovrstno komponento korpusa. Večina kriterijev ima meje med posameznimi kategorijami prehodne, zato je besedila pogosto težko uvrstiti v posamezno kategorijo. Ekipa načrtovalcev se mora odločiti, katere kriterije bo izbrala in v kakšno medsebojno razmerje jih bo postavila. Pri tem je pomembno, da so vsi izbrani kriteriji zastopani z dovolj velikim številom besedil in da se njihovo medsebojno razmerje čim bolj približa oceni razmerja v realnosti. V nadaljevanju si bomo ogledali problematiko posameznih kriterijev.

3.3.3.1 Stopnja spontanosti govora

Pri stopnji pripravljenosti govora razlikujemo spontani in pripravljeni govor, vendar so stopnje pripravljenosti lahko različne in jih je s stališča recepcije zelo težko ali celo nemogoče določiti. Zato predlagam, da bi pri morebitni gradnji korpusa razlikovali samo med spontanim govorom in govorom, ki je bran, torej branim besedilom.[101]

3.3.3.2 Monologi in dialogi

V govornem korpusu morajo biti nedvomno zastopani dialogi in monologi. Vprašanje je, ali lahko ostanemo pri poimenovanju dialog za vse, kar ni monolog, torej tudi za multiloge. Zaradi nekaterih jezikovnih analiz bi se zdelo bolje imeti dialoge in multiloge ločene, posebej ker tovrstno označevanje opravi urednik korpusa in je popolnoma enostavno, pa vendarle predstavlja dodatno možnost iskanja po korpusu. Bolj problematična je lahko včasih odločitev, v katero kategorijo posamezno besedilo sploh uvrstiti. Intervju na primer je teoretično primer klasičnega dialoga, realizacija pa je lahko taka, da prva oseba zastavi samo eno kratko vprašanje, druga oseba pa odgovori s petminutnim monologom. Da bi se izognili neustrezni klasifikaciji, bi morda lahko imeli za kriterij, da mora npr. vsaka izmed oseb izgovoriti več kot 5 odstotkov besed celotnega besedila, da lahko govorimo o dialogu. Pri odločanju o prevladujoči strukturi besedila moramo torej upoštevati število izgovorjenih besed in število izmenjav govorcev; končno oznako o prevladujoči strukturi besedila bi moral postaviti urednik, kriteriji pa bi se lahko dokončno izoblikovali šele med samim delom.

Nadalje se zastavlja vprašanje kvantitativnega razmerja med dialogi in monologi. Mogoče bi ga bilo opredeliti tudi naknadno, ko bi bila vsa besedila zbrana, glede na izkušnje ob gradnji drugih korpusov pa bi razmerje lahko načrtovali v okviru 75

[101] Branih besedil v govorni korpus ne vključujemo načrtno, se pa včasih pojavijo med spontanim govorom, npr. na predavanjih itd.

odstotkov proti 25 odstotkov v korist dialogov oz. multilogov (korpus SEU 76 : 24 odstotkov, Nizozemski govorni korpus 81 : 19 odstotkov, BNC 85 : 15 odstotkov). Delež monologov v realnosti je verjetno še nižji, vendar pa imajo nekateri monologi velik doseg (npr. pedagoški, politični govor) in zato večjo vlogo s stališča recepcije.

3.3.3.3 Javni in zasebni govor

Naslednje vprašanje, ki se zastavlja znotraj besedilnovrstne komponente, je vprašanje *javnosti/zasebnosti* besedil. Pri gradnji korpusa SEU so definirali pojem javnosti kot »govor, ki se odvija pred poslušalci, ki se vanj ne vključujejo« (Greenbaum 2003, 2); definicijo prevzemajo tudi številni drugi raziskovalci. Problem nastane, če si predstavljamo npr. sestanek delovnega kolektiva, v katerega se kot govorci vključujejo vsi prisotni (torej ni »poslušalcev«, da bi mu po zgornji definiciji lahko pripisali status javnosti); sestanka vseeno ne moremo imeti za zasebno besedilo.[102] V slovenski jezikoslovni teoriji sta pojma zasebno/javno (besedilo) definirana glede na naslovnika (Bešter idr. 1999, 80): »glede na to, ali je naslovnik besedila posameznik ali javnost, so besedila *zasebna* ali *javna*.« Za namen te raziskave pojem *javno* razumem širše, ne samo tisto, kar je namenjeno (širši) javnosti, ampak vse, kar ni zasebno, torej *nezasebno*. Tu bi bilo mogoče uvesti še kategorijo *uradno/neuradno*, ki se ne prekriva s pojmoma *javno/zasebno*, ampak imamo uradna in neuradna besedila tako znotraj javnih kot znotraj zasebnih besedil. V slovenskem jezikoslovju na podoben razmislek naletimo pri novejši klasifikaciji zvrsti govorjenih besedil (Vogel 2004, 461), kjer avtorica razlikuje med »javnimi govornimi nastopi, javnimi pogovori (pogovori pred javnostjo), uradnimi zasebnimi pogovori in neuradnimi zasebnimi pogovori.«. Definicijo uradnega/neuradnega besedila imamo tudi pri Bešter idr. (1999, 80): »Glede na to, ali je bilo besedilo tvorjeno v enakovrednem ali neenakovrednem družbenem razmerju med sporočevalcem in naslovnikom oz. ali kateri od udeležencev govori v imenu ustanove, je besedilo *neuradno* ali *uradno*;« definicija se mi v prvem delu ne zdi ustrezna, saj govori o neenakovrednem družbenem položaju govorcev, kar ne pogojuje nujno uradnega besedila, pač pa je to odvisno tudi od okoliščin. Lahko si namreč predstavljamo isto skupino ljudi, ki so tesni sodelavci v kolektivu: magnetogram njihovega kolegija bo povsem uradno besedilo, njihov jutranji klepet ob kavi pa popolnoma neuradno besedilo. Poleg tega v slovenski jezikoslovni teoriji razmerje med govorcema pogosteje označujemo z izrazoma *formalno/neformalno*, kot bomo videli v nadaljevanju. Da je definicija javnega/zasebnega v govoru nujno potrebna, ne pa tudi zadostna, potrjuje tudi Kranjc (1996/97, 308): »Poleg javnosti/zasebnosti je treba upoštevati tudi družbeno distanco med sogovorcema, njuno starost in to, ali sta prijatelja ali le bežna znanca, skratka, pozorni

[102] Pojem zaseben-bna-o je v SSKJ razložen kot "nanašajoč se na posameznika kot neuradno osebo".

moramo biti na sociolingvistične danosti /.../.« Avtorica v bistvu govori o formalnem in neformalnem govornem položaju znotraj javnega in zasebnega govora, kar sicer v obratnem sorazmerju dokazujeta tudi izjava »Govorne položaje ločujemo na formalne in neformalne. V obeh vrstah govornih položajev pa se lahko odvija javni ali zasebni diskurz (Kranjc 1996/97, 308).«

Gradnjo govornega korpusa bi bilo racionalno načrtovati tako, da bi bila večina zasebnega govora zbrana znotraj demografske komponente, in sicer uradnih in neuradnih besedil, saj bi se izbrani govorci posneli tudi v govornih položajih, ki jih označimo za uradne (v pogovoru z osebo, ki predstavlja institucijo).[103] Nasprotno oz. komplementarno bi v besedilnovrstno komponento korpusa zajemali predvsem javna besedila; taka podkorpusa bi omogočala tudi študij razlik med javno in zasebno komunikacijo.

3.3.3.4 Stopnja formalnosti besedil

Sigley v študiji o formalnosti besedil (1997, 206–208) pojasnjuje, da pojem formalnosti lahko razumemo kot situacijski kontekst (tu lahko opazujemo vzrok), lahko pa gre za jezikovno dejstvo (tu opazujemo jezikovno izbiro, ki je posledica situacijske formalnosti, pri nas npr. vikanje). Obenem pa velja, da jezikovni podatki ne merijo oz. ne izkazujejo vedno ustrezno formalnosti situacije, in obratno, na podlagi formalnosti situacije ne moremo zanesljivo predvidevati, kakšne učinke bo imela na jezik govorcev. Sigley nadalje navaja situacijske dejavnike, ki vplivajo na stopnjo formalnosti:

Formalno		Neformalno
visoka, pomembna	*družbena vrednost dogodka*	nizka, nepomembna
vnaprej določena	*struktura govornega dogodka*	prosta
javno, institucionalno	*okolje*	domače, zasebno
tujci socialne vloge	*udeleženci,* *odnos, vloge*	intimen odnos, individualne osebe
informativni	*namen dogodka*	interakcija
abstraktna, specifična, vnaprej določena	*tema*	zasebna, splošna, nedoločena

Tabela 13: Situacijski dejavniki formalnosti[104]

[103] Tu se lahko pojavi problem snemanja; v tem primeru bi morali uradna zasebna besedila zajemati v okviru besedilnovrstne komponente korpusa.
[104] Sigley 1997, 209.

Gre za zapleten preplet mehanizmov, ki jih nezavedno pregledujejo človeški možgani, ko se človek odloča za izbiro jezikovnih sredstev. Kar zadeva slovensko jezikovno situacijo, lahko k Sigleyjevi shemi dodamo vsaj še starost udeležencev, ki v še tako neformalnih okoliščinah lahko vpliva na izbiro formalnega jezikovnega sredstva – vikanja. Pojem formalnosti se lahko nanaša bodisi na situacijo oz. okoliščine bodisi na razmerje med govorcema. Glede na to, da sem v okviru te razprave okoliščine že definirala s pojmi javno/zasebno in neuradno/uradno, bom kriterij formalnosti uporabila za označevanje razmerja med govorci, pri čemer je to razmerje določeno s starostjo govorcev ter njunim medsebojnim odnosom, pa tudi z okoliščinami (npr. če se govorca ne poznata, če eden zastopa uradno ustanovo itd.)[105] Formalnost/neformalnost govornega položaja po Kranjc (1996/1997, 309) določa »socialna razdalja med govorcema ter njun status in vloga v družbi.« Vprašanje je, v kolikšni meri se med seboj prekrivajo pojmi *uradno/neuradno* in *formalno/neformalno*; v slovenski jezikoslovni teoriji, kot smo videli, ni enotnega pojmovanja teh izrazov. Načeloma naj bi uradne okoliščine pogojevale večjo mero formalnosti govora, neuradne okoliščine pa ravno nasprotno. Obenem pa vemo, da v realnem govoru ni vedno tako: analize empiričnega gradiva bodo lahko pokazale, kdaj in kakšna so odstopanja, morda pa tudi, na kakšen način odstopanja vplivajo na sporazumevanje. Pričakujemo lahko tudi, da bodo tudi relevantne študije o stopnji formalnosti govora lahko nastale šele na podlagi govornega korpusa, kar pravzaprav velja tudi za vse ostale kriterije za zajem besedil. V fazi načrtovanja korpusa pa pričakujemo, da bosta dovolj veliki in dobro uravnoteženi demografska in besedilnovrstna komponenta govornega korpusa zagotavljali, da bodo v besedilih zastopana vsa jezikovna sredstva, ki se pojavljajo v govornih položajih z različno stopnjo formalnosti.

3.3.3.5 Tajnost snemanja

Zavedanje govorcev, da se njihovo govorjenje snema, zagotovo vpliva na njihovo govorno produkcijo; najbolj avtentične posnetke govora zato dobimo, če snemamo na skrivaj. Je pa tako postopanje v marsikaterem pogledu lahko sporno, na Slovenskem npr. prepovedano in kaznivo,[106] zato so ga pri gradnji govornih korpusov v preteklosti uporabljali le redko in v omejenem obsegu. Na podlagi posnetkov spontanega govora, narejenih za učni govorni korpus slovenščine, lahko vsaj ohlapno sklepam, da govorci dokaj hitro pozabijo na mikrofon in se v posameznih trenutkih nanj spet spomnijo, kar je razvidno iz njihovih reakcij. Zaradi zakonskih določil moramo biti pripravljeni na to, da bo večina posnetkov v govornem korpusu posnetih z vednostjo govorcev.

[105] Čehi pri gradnji govorne komponente korpusa odnos med govorcema označujejo s tremi stopnjami: govorca se ne poznata, govorca se poznata, govorca sta prijatelja (Kopřivová 2005, 139).

[106] Kazenski zakonik RS (KZ-UPB1), http://www.uradni-list.si/1/objava.jsp?urlid=200495&stevilka=4208.

3.3.3.6 Namen besedil

Tudi namen oz. funkcija besedila je lahko kriterij za zajem besedil v korpus. Pri gradnji korpusa BNC je bil namen v izhodišču besedilnovrstne tipologije; uokvirjen je bil v štiri kategorije: izobraževanje in informiranje (1), poslovna/poklicna komunikacija (2), javni oz. institucionalni govor (3) in prosti čas (4). Tipologija je na prvi pogled preprosta, v resnici pa nejasna. Pravzaprav se tu prekrivata namen besedila in situacija: če prva kategorija v resnici sloni na namenu besedila, pa je tretja vezana na konkretno okolje, v katerem sporazumevanje poteka, četrta pa na okoliščine. Na slovensko jezikovno situacijo se mi zdi besedilnovrstna tipologija BNC težko prenosljiva: univerzitetno predavanje npr. sodi v prvo in tretjo kategorijo, parlamentarna razprava v drugo in tretjo, javna okrogla miza o mestnem vodovodu v prvo, drugo in tretjo kategorijo itd. Tradicionalna slovenska zvrstna teorija sicer pozna poimenovanje *funkcijske zvrsti* (Toporišič 2000, 27), kategorijo, znotraj katere naj bi se besedila razvrščala glede na »uporabnostni namen«: umetnostna, publicistična, praktičnosporazumevalna in strokovna besedila. Ker omenjena teorija temelji izključno na pisnih besedilih, je za namen gradnje govornih korpusov neuporabna. Tudi če bi jo na silo transformirali na govorjena besedila, bi dobili praktično-sporazumevalna besedila, jezik medijev, strokovni jezik in umetnostni jezik; jezik medijev ni namen besedila, ampak prenosnik, strokovna govorjena besedila pa v realnosti predstavljajo zelo majhen (čeprav dragoceni) delež, o umetnostnem spontanem govoru pa sploh lahko govorimo le izjemoma.

Novejša slovenska funkcijska teorija v izhodišče delitve postavlja pragmatično funkcijo besedil. Skubic (1994/95) npr. glede na to, katera funkcija prevladuje v besedilu, ločuje štiri pragmatične funkcije govora, in sicer 1. znanstveni govor (spoznavna funkcija), 2. sporočanjskovplivanjski govor (ta poleg tradicionalne domene praktičnosporazumevalnega in publicističnega pokriva tudi dobršni del strokovnega govora, njegova funkcija je perlokucijska, torej usmerjena k doseganju praktičnih učinkov pri naslovniku in poslušalstvu, tudi k učinku spremembe vednosti (Skubic 1994/95, 156)), 3. konvencionalnoperformativni oz. uradni govor (performativna funkcija) in umetnostni govor (ki ima »umetnostno« funkcijo).[107] Tudi Skubičeva teorija funkcijskih zvrsti je za klasifikacijo besedilnih vrst za gradnjo govornega korpusa neuporabna, saj bi večina govorjenih besedil sodila v drugo (in tretjo) kategorijo, kar nam pri razvrščanju ne pomaga veliko.

O namenu besedil je v zvezi z gradnjo korpusov težko govoriti, saj imajo besedila lahko hkrati več namenov. Enako ali še bolj to velja za temo besedila, ki

[107] S kakšnim *namenom* tvorec ustvarja umetnostno besedilo ostaja tudi pri Skubicu odprto filozofsko vprašanje.

se v velikem delu govorjenih besedil lahko popolnoma poljubno spreminja. Za izhodišče besedilnovrstne klasifikacije besedil z namenom gradnje govornega korpusa je tako bolje upoštevati prej naštete kriterije, in sicer prevladujočo strukturo besedila, okoliščine, govorni položaj (odnos med govorcema/govorci) in prenosnik.

3.3.3.7 Končni predlog besedilnovrstnih kriterijev

Obravnavani kriteriji za zajem besedil v besedilnovrstno komponento referenčnega govornega korpusa so v povzeti v spodnji tabeli:

Kriterij	Kategorije			
struktura besedila	multilog	dialog	monolog	
okoliščine	javna besedila			
govorni položaj	neformalni		formalni	
prenosnik[108]	osebni stik	telefon	avdio	video
okoliščine snemanja	posneto z vednostjo govorca		brez vednosti govorca	

Tabela 14: Predlog kriterijev za zajem besedil v besedilnovrstno komponento KGS

Pri načrtovanju zajemanja javnih besedil na radiu in TV si lahko pomagamo tudi s podatki Statističnega urada RS, ki vodi podatke o deležu predvajanih oddaj na slovenskih radijskih in televizijskih oddajah, [109] in sicer po tematskih sklopih glasba, razvedrilo (kontaktne oddaje, kvizi, tekmovanja, humor), igrani dramski program, šport, dnevnoinformativni program, aktualnoinformativni program, oddaje za mladostnike, oddaje za izseljence, oddaje za kmetijce, sklop oddaj umetnost-humanistične vede-znanost, oddaje za izobraževanje, religija in drugo; navedena kategorizacija je zanimiva tudi za načrtovanje zajema besedil v korpus govorjenega jezika.

[108] Internet pri gradnji govornega korpusa ni upoštevan kot poseben prenosnik, saj gre pravzaprav samo za kanal za prenos zvoka in/ali slike.

[109] http://www.stat.si/letopis/index_vsebina.asp?poglavje=8&leto=2004&jezik=si; obstajajo tudi podatki o poslušanosti in gledanosti posameznih oddaj oz. RA- in TV-hiš.

3.3.4 Formalno-pravni vidiki gradnje govornega korpusa

Zakon o avtorskih in sorodnih pravicah (ZASP)[110] v 5. členu določa t. i. Varovana dela, in sicer določa:

> (1) Avtorska dela so individualne intelektualne stvaritve s področja književnosti, znanosti in umetnosti, ki so na kakršenkoli način izražene, če ni s tem zakonom drugače določeno.
>
> (2) Za avtorska dela veljajo zlasti:
> 1. govorjena dela, kot npr. govori, pridige, predavanja;
> /.../.

Iz navedenega je jasno, da se govorjena besedila uvrščajo med avtorska dela in da jih je treba tako tudi obravnavati; za vsako nadaljnjo uporabo govorjenih besedil je treba pridobiti dovoljenje avtorja. Nadalje pa Kazenski zakonik RS (KZ-UPB1)[111] tudi ne dovoljuje neupravičenega zvočnega snemanja:

> **Neupravičeno prisluškovanje in zvočno snemanje**
> **148. člen**
>
> (1) Kdor neupravičeno s posebnimi napravami prisluškuje pogovoru ali izjavi, ki mu nista namenjena, ali ju zvočno snema, ali kdor takšen pogovor ali takšno izjavo neposredno prenaša tretji osebi, ali ji takšen posnetek predvaja ali kako drugače omogoči, da se z njim neposredno seznani, se kaznuje z denarno kaznijo ali z zaporom do enega leta.
> (2) Enako se kaznuje, kdor zvočno snema njemu namenjeno zaupno izjavo drugega brez njegove privolitve z namenom, da bi takšno izjavo zlorabil, ali kdor takšno izjavo neposredno prenaša tretji osebi ali ji takšen posnetek predvaja ali ji kako drugače omogoči, da se z njim neposredno seznani.
> (3) Če stori dejanje iz prvega ali drugega odstavka tega člena uradna oseba z zlorabo uradnega položaja ali uradnih pravic ali poskusi to storiti, se kaznuje z zaporom od treh mesecev do petih let.
> (4) Pregon za dejanje iz prvega odstavka tega člena se začne na predlog, za dejanje iz drugega odstavka pa na zasebno tožbo.

[110] http://zakonodaja.gov.si/rpsi/r03/predpis_ZAKO403.html
[111] http://www.uradni-list.si/1/objava.jsp?urlid=200495&stevilka=4208

Ob gradnji večjega govornega korpusa bi zaradi občutljivosti problematike pri pripravi dokumentov za urejanje avtorskih pravic govorcev morali sodelovati strokovnjaki-pravniki, saj bi neurejenost na tem področju lahko ogrozila obstoj oz. uporabnost korpusa.

3.4 Končni predlog priporočil za zajem besedil v KGS

V zaključku bodo predstavljena priporočila za zajem besedil v govorni korpus slovenščine, ki izhajajo iz zgoraj povedanega in ki predstavljajo eno izmed možnosti za zajem besedil v govorni korpus. Z metodo zbiranja, ki jo predlagam, bi dosegli visoko stopnjo reprezentativnosti in uravnoteženosti govornega korpusa. Besedila bi zbirali s komplementarnim kombiniranjem demografske in besedilnovrstne metode. Besedila za demografsko komponento bi zbrali s pomočjo reprezentativnega vzorca govorcev. Kriteriji, na podlagi katerih bi izbrali reprezentativni vzorec govorcev, so: spol, starost, dosežena izobrazba, regijski izvor in prvi jezik. Gre za kriterije, ki glede na obstoječe raziskave v slovenskem jezikoslovju in glede na hipotetična predvidevanja najbolj vplivajo na razlike v govoru; pri tem npr. ni upoštevan kriterij socialnega izvora, ker je (vsaj z današnjega stališča) za slovenske razmere nedoločljiv in nesmiseln. Z demografsko metodo bi zbirali predvsem zasebna in nezasebna dialoška in monološka besedila, govorjena v osebnem stiku ali po telefonu; zbrana besedila bi obsegala približno polovico celotnega govornega korpusa.

Predlagani kriteriji za zbiranje besedil znotraj besedilnovrstne komponente so struktura besedila (monolog, dialog, multilog), okoliščine, v katerih je besedilo govorjeno (javna/zasebna besedila), govorni položaj (formalni/neformalni) in prenosnik (osebni stik, telefon, avdio ali video nosilec). Z besedilnovrstno metodo bi zbirali predvsem javne dialoge in monologe (govorjene v osebnem stiku, po radiu, TV ali telefonu, z višjo ali nižjo stopnjo formalnosti); znotraj besedilnovrstne metode bi lahko dopolnili tudi besedila, ki bi jih z demografskim zbiranjem zbrali manj, kot je bilo načrtovano.

Načrtovana mreža predvideva dovolj širok okvir za zajemanje besedilnih vrst, ki se pojavljajo v realnem govoru. Nisem želela izdelati togega sistema, ki bi predvideval zajem vseh teoretično možnih besedilnih vrst, če se ta v realnem govoru le redko ali izjemoma pojavljajo, npr. zasebni monolog na avdio posnetku, zasebni dialog na videu, javni monolog po telefonu ipd. Predvidljivi primeri besedilnih vrst sistema ne zapirajo, ampak predstavljajo izhodišče za zbiranje; načrt mora biti dovolj prožen in odprt za dopolnjevanje in spremembe med zbiranjem.

Večina besedil bo posnetih z vednostjo govorcev, ker tako predvideva slovenska zakonodaja, le izjemoma se lahko zgodi, da kdo izmed govorcev s snemanjem ni seznanjen (se nepredvideno vključi v pogovor).

4 Označevanje in transkribiranje govorjenih besedil

Naslednje temeljno vprašanje, ki se zastavlja načrtovalcem govornega korpusa, je zapisovanje govorjenih besedil. Pri oblikovanju načel za transkripcijo je nujno in smiselno upoštevati mednarodne standarde in priporočila, znotraj teh pa iskati najprimernejše rešitve in prilagoditve; problematika zadeva predvsem zapis v pisnem jeziku nestandardiziranih besedilnih delov, polverbalnih (mhm, ə) in neverbalnih zvokov (smeh, kašelj), nedokončanih besed, prekrivnega govora in premorov. Enako kot vsa načela glede gradnje korpusa so tudi načela za transkribiranje korpusa odvisna od namembnosti korpusa: za jezikoslovca, ki ga zanimajo predvsem ponavljajoči se (skladenjski) vzorci v jeziku, pogostnost pojavljanja ter pomeni besed, ki jih lahko izluščimo iz velike količine podatkov, »ne bo potrebe po zelo sofisticirani transkripciji in bo najpomembnejša količina in hitrost transkribiranja« (Thompson 2004). Fonetik potrebuje manjšo količino podatkov, vendar morajo biti ti mnogo bolj detajlno transkribirani v smislu prozodije in akustične realizacije ter nujno povezani z zvočnimi posnetki. Za sociolingvista in analitika diskurza so spet bistvene detajlne informacije o okoliščinah in sobesedilu itd. Za gradnjo govornih korpusov, ki so del referenčnih korpusov, je najpogosteje uporabljena modificirana ortografska transkripcija (Thompson 2004), ki vključuje tako zapis nekaterih prozodičnih lastnosti govora kot tudi nekatere vnaprej določene zapise govora, ki odstopajo od standardnega zapisa ali zapis šele vzpostavljajo. V tem smislu bo potekalo tudi načrtovanje priporočil za transkribiranje govorjene slovenščine, pri čemer tudi načrtovana povezava med transkripcijami in zvočnimi posnetki vpliva na izbor in količino oznak.

Predstavljena bodo priporočila strokovnih skupin za standardizacijo zapisa govorjenih besedil, nekaj transkripcijskih standardov, izdelanih ob gradnji govornih korpusov, in programska orodja za transkribiranje. Na tej osnovi bodo izdelana priporočila za zapis govorjenega jezika, pri čemer bodo upoštevani tudi redki primeri že obstoječih priporočil za zapisovanje govora v slovenščini.

4.1 PRIPOROČILA TEI ZA OZNAČEVANJE GOVORJENIH BESEDIL

TEI (Text Encoding Initiative)[112] je raziskovalni projekt, ki je prerasel v organizacijo, ustanovile pa so ga tri vodilne organizacije s področja jezikovnih tehnologij, *Association for Computational Linguistics, Association for Literary and Linguistic Computing* in *Association for Computers and the Humanities*. Raziskovalno in aplikativno delo delovnih skupin in komitejev TEI so (vsaj na začetku) financirali iz proračunskih sredstev ZDA, EU in Kanade, velik delež pa so v obliki brezplačnega dela prispevali tudi raziskovalci in strokovnjaki z različnih področij – jezikoslovja, raču-

[112] http://www.tei-c.org/

nalništva, dokumentalistike, zgodovine itd. Priporočila za označevanje govorjenih besedil TEI je izdelala skupina, ki so jo sestavljali Lou Burnard, Jane Edwards, Stig Johansson (vodja) in And Rosta; priporočila so objavili leta 1991 v publikaciji *Text Encoding Initiative, Spoken Text Working group, final report* (izdala Univerza v Oslu).

4.1.1 Enote govorjenega besedila

Transkribirano govorjeno besedilo je po definiciji TEI enako kot pisno korpusno besedilo razdeljeno na dve osnovni enoti, *telo* in *glavo* besedila. Telo je besedilo samo, in sicer:

<besedilo> je transkripcija niza govorjenega besedila, ki ga je zaradi določenih razlogov mogoče imeti za samostojno enoto in ga obravnavati kot zaključeno besedilo (Johansson 1995, 86).

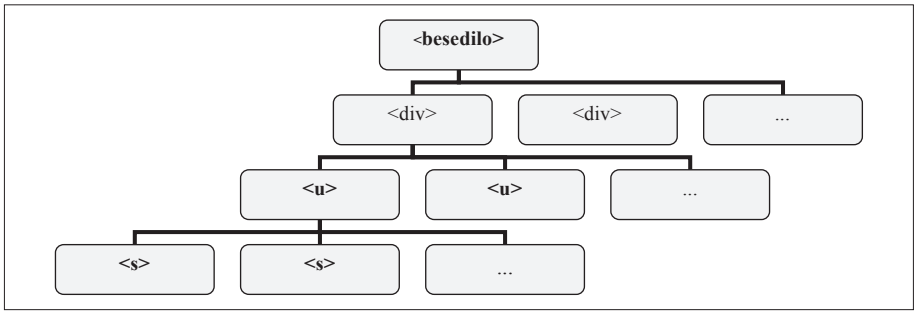
Podenota govorjenega besedila, ki jo danes priznava oz. sprejema večina transkripcijskih konvencij, je izjava (*utterance*), po splošnih priporočilih TEI:

<u> je segment diskurza, določen po skladenjskih, fonoloških ali prozodičnih načelih (Burnard 1995, 74)

in po posebnih priporočilih TEI za govorjeni jezik:

<u> je niz govorjenega besedila, ki ga običajno spredaj in zadaj zaznamuje (vsaj kratek) premor v govorjenju ali menjava govorcev (Johansson 1995, 87).

Shema TEI za segmentacijo govorjenih besedil omogoča štiristopenjsko segmentacijo:



Slika 16: Priporočila TEI za segmentiranje govorjenih besedil (Johansson 1995, 86)

Osnovni enoti sheme sta besedilo (<text>) in izjava (<utterance>): besedilo je lahko segmentirano samo na izjave. T. i. divizija (odsek) je ponujena kot možna, ne pa obvezna hierarhična stopnja med besedilom in izjavo, segment <s> pa je najmanjša (prav tako neobvezna) enota pri segmentiranju govorjenega besedila; segment predstavlja podenoto izjave, definirano po prozodičnih ali skladenjskih kriterijih.[113] Definicije oznak za segmentacijo so namenoma nekoliko ohlapne in nedoločne: vsak uporabnik te sheme (to je načrtovalec korpusa) mora vsako oznako sam natančneje definirati in definicije objaviti na dostopnem mestu (Johansson 1995, 87).

V korpusu BNC so prvič segmentirali govorjena besedila po priporočilih TEI. Najvišja enota segmentiranja govorjenih besedil je bil odsek (*division* , BNC Users Guide 2000, 7.1), sledijo pa v hierarhičnem razmerju izjava, stavek in beseda:

OZNAKA	POMEN
<stext>	govorjeno besedilo
	odsek (skupina izjav)
<u>	izjava
<s>	stavek
<w>	beseda

Slika 17: Strukturne oznake govorjenih besedil BNC

V demografski komponenti korpusa je <stext> oznaka za posneto gradivo posameznega izbranega govorca, <div1>, <div2> itd. pa oznake za posamezne konverzacije, ki jih je posnel ta govorec z različnimi sogovorniki. Vse transkripcije znotraj demografskega dela so nato členjene na izjave (<u>). V besedilnovrstnem delu korpusa, kjer ni bilo »izbranih govorcev«, tudi členitve na divizije ni bilo; vsako govorjeno besedilo je segmentirano neposredno na izjave (BNC Users Guide 2000, 7).

Izjave so v BNC sekvence govora, ki jih tvori en govorec (ali v posebnem primeru skupina govorcev), predstavljajo pa neprekinjeni del govora, sestavljen iz enega ali več stavkov, ki po pomenu spadajo skupaj.

[113] Oznake za segmentacijo besedila so skupne za pisna in govorjena besedila. Oznaka je rezervirana predvsem za segmentacijo pisnih besedil na poglavja, odstavke, kitice itd. in ima lahko različne hierarhične stopnje <div1>, <div2> itd.

4.1.2 Označevanje govorcev

Večina transkripcij, vsaj v okviru korpusnega jezikoslovja, do neke mere prevzema obliko dramskega besedila oz. scenarija, kjer je pred vsako izjavo zapisano, kdo jo izgovori. Načini, kako se govorci označujejo, se od korpusa do korpusa razlikujejo. TEI priporoča, da se govorce označuje s številkami in z oznako <who>, in sicer znotraj izjave (primer je iz korpusa BNC):

```
<u who=1>It's funny old day isn't it.</u>
<u who=2>Mm, it's not cold is it?...</u>[114]
```

Vrednost elementa <who> določa posameznega govorca, udeleženca komunikacijskaga dogodka, njegova identifikacija pa je podana v glavi besedila.[115] Označevanje govorcev je v formatih, ki so predvideni za nadaljnjo rabo in za branje, poenostavljeno. Kot primer si lahko ogledamo zapis besedila v korpusu najstniške britanske angleščine COLT:[116]

```
{sample 32407} header
{u 1-W1} Fuck , would be a good one. Everything's come
out my bag . ... Bloody way ! Aargh !
{u 2-W4} Come on Peter let's go and get on the bus .
{u 3-W1} Ah shit !
```

S klikom na glavo besedila (*header*) dobimo identifikacijo govorcev:

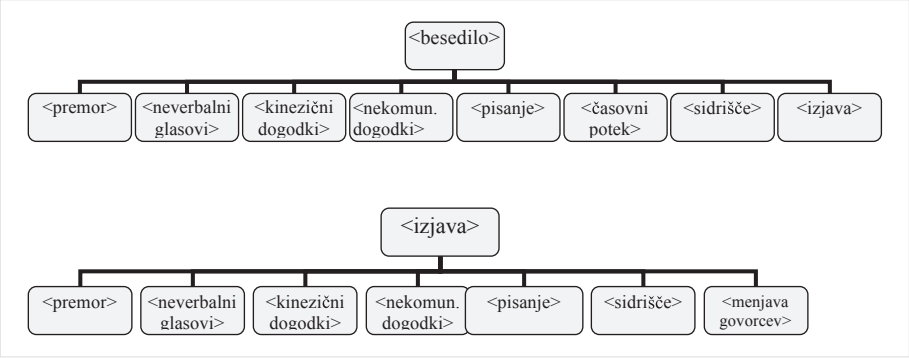
```
W1 gender=m age=14 dialect=London occupation=student
social group=2 ethnicity=white
W4 gender=m age=15 dialect=London occupation=student
social group=?
```

Kot bomo kasneje lahko videli tudi v kritikah, je bil največji problem priporočil TEI preveliko prilagajanje strojnemu zapisu, razumljivemu za računalnik. Ker ne obstaja program, ki bi avtomatsko pretvarjal strojni zapis (narejen po priporočilih TEI) v človeku prijazen zapis, si sestavljavci korpusov pomagajo na različne načine, večinoma pa zaenkrat poenostavljene zapise pripravljajo sami.

[114] Johansson 1995, 87.
[115] Gre za popis demografskih lastnosti govorca, kot so jih predvideli sestavljavci korpusa in kot jih je snemalec pogovora zabeležil, seveda v skladu z varovanjem osebnih podatkov.
[116] http://gandalf.hit.uib.no/c/t/32407.htm

4.1.3 Prozodične in neverbalne oznake govorjenih besedil

Transkriptorji zaznamujejo tudi akustične in druge dogodke, ki so lahko pomembni za interpretacijo govorjenega besedila. To so premori (tihi ali »zapolnjeni«), neverbalni (človeški) glasovi, mimika obraza in gibi, zvoki v ozadju in drugo. Oznake, kot jih na ravni besedila in izjave priporoča TEI, so prikazane v spodnji shemi in podrobneje razložene v nadaljevanju:



Slika 18: Prozodične in neverbalne oznake TEI pri transkripciji govorjenih besedil[117]

4.1.3.1 Premori

Premori (<*pause*>) se lahko pojavijo med posameznimi izjavami ali znotraj izjav. Če se premor pojavi znotraj izjave, ga pripišemo govorcu izjave, če pa se pojavi med izjavami, ga pripišemo vsem govorcem. Oznaka <*pause*> se nanaša samo na tihe premore (brez zvoka v ozadju), dolžina premora pa je po TEI lahko oznaki pripisana ali pa ne.

4.1.3.2 Neverbalni glasovi

Lahko so označeni opisno (kašljanje, nosljanje, grgranje) ali pa zapisani v ortografski obliki kot del besedila (*khm*). Oznaka TEI za tovrstne glasove je <*vocal*>, vedno pa ji je pripisana identifikacija govorca <*who*> in opis dogodka. Včasih se zgodi, da se neverbalni glasovi prekrivajo z izjavo, da torej tvorijo izjavo v celoti (npr. če govorec sodeluje v pogovoru samo z izjavo »mhm«); v takih primerih

[117] Johansson 1995, 88.

so v veljavi različne prakse zapisovanja, npr. kot da gre za izjavo ali samo za niz neverbalnih glasov.

4.1.3.3 Kinezični dogodki

V priporočilih TEI so kinezični dogodki (npr. pokimati, odkimati, skomigniti z rameni) označeni kot <*kinesic*>, pripisani so posameznemu govorcu in opisani. Primer:[118]

```
<u who=Jane>Have you read Vanity Fair</u>
<kinesic who=Lou desc=nod>
```

Zgoraj je kinezični dogodek pripisan osebi z imenom Lou in opisan kot kimanje; iz primera je tudi razvidno, da kinezični dogodki nimajo statusa izjave.

4.1.3.4 Nekomunikacijski dogodki

Izjave, neverbalne glasove in kinezične dogodke druži komunikacijska funkcija. Pri transkribiranju besedil pa transkriptorji naletijo tudi na nekomunikacijske zvoke, ki so lahko pomembni za interpretacijo besedila. Tovrstne oznake imajo podobno funkcijo kot didaskalije v dramskem besedilu. TEI priporoča za nekomunikacijske dogodke oznako <*event*> z dodanim opisom:[119]

```
<u who=Jan>This is just delicous</u><event desc=telephone
rings>
<u who=Kim>I'll get it</u>
```

Brez opisa dogodka, ki se je zgodil med obema izjavama (telefon zazvoni), bi bila komunikacija med obema osebama nerazumljiva oz. bi se druga izjava glede na prvo zdela nekoherentna. Če je potrebno, je dogodek lahko pripisan posameznemu govorcu. Meje med kinezičnimi in nekomunikacijskimi dogodki seveda ni vedno lahko določiti.

[118] Johansson 1995, 89.
[119] Johansson 1995, 89.

4.1.3.5 Pisanje

Besedilo v pisni obliki se lahko pojavi kot pomemben del govorne situacije, npr. na predavanju ali na sodišču. V takem primeru TEI priporoča oznako <*writing*> skupaj z oznako <*who*>, ki se nanaša na tvorca besedila. Če je brano besedilo popolnoma nerelevantno za potek govorjenja, je v transkripciji lahko označeno samo kot dogodek (<*event*>).[120]

4.1.3.6 Trajanje dogodkov

Različne transkripcijske sheme imajo različen odnos do označevanja trajanja posameznih dogodkov. Priporočila TEI dopuščajo pripis časovnega vrednotenja (*duration*) izjavam, premorom, neverbalnim glasovom ter kinezičnim in nekomunikacijskim dogodkom.

4.1.3.7 Časovni potek dogodkov

Razporeditev zvoka oz. glasov v času je v transkripciji prenesena na razporeditev v prostoru. Zaporedje besed v transkripciji mora torej čim natančneje predstavljati časovne sekvence zvoka oz. govora. Ta preprosta analogija se poruši, kadar se zgodi več akustičnih dogodkov hkrati, npr. kadar govori več govorcev hkrati. Prekrivni govor je v transkripcijskih konvencijah vedno označen, čeprav na različne načine: lahko je označen z zvezdicami ali z oglatimi oklepaji, lahko je indeksiran ali vertikalno poravnan v transkribiranih vrsticah. Oglejmo si nekaj primerov:

```
Tom: I used to smoke *a lot more than this* but I never
inhaled smoke
Bob: *you used to smoke*[121]
```

ali

```
Tom: I used to smoke [a lot more than this] but I never
inhaled smoke
Bob: [you used to smoke]
```

ali

```
Tom: I used to smoke [a lot more than this 1] but I never
inhaled smoke
Bob: [you used to smoke 1]
```

[120] Tu se lahko pojavi težava pri številnih govorjenih besedilih v medijih, ki lahko imajo pisno predlogo, pa transkriptor tega ne bo mogel z gotovostjo vedeti. Vendar se tudi to informacijo lahko dobi ob popisu govorca.

[121] Johansson 1995, 92.

ali

```
Tom: I used to smoke [a lot more than this] but I never
inhaled smoke
Bob:                    [you used to smoke]
```

Po priporočilih TEI bi moral biti prekrivni govora označen, kot prikazuje naslednji primer:[122]

```
<timeLine>
      <when id=T1>
      <when id=T2>
</timeLine>
...
<u who=Tom> I used to smoke <anchor synch=T1>a lot more
than  this <anchor synch=T2> but I never inhaled smoke</
u>
<u who=Bob><anchor synch=T1> you used to smoke<anchor
synch=T2>
```

V primerjavi z zgoraj navedenimi zapisi je zapis TEI izredno zahteven tako za zapisovanje kot za branje, pa kljub temu lahko iz zapisa razberemo, da sta govorni sekvenci med točkama T1 in T2 (sidrišči) na časovni premici izgovorjeni istočasno. Avtorji TEI so se ob pisanju priporočil zavedali, da je tovrstno zapisovanje v resnici stvar prihodnosti (Johansson 1995, 93), ker potrebna programska orodja za sinhronizacijo zvoka in zapisa ob pisanju priporočil še niso bila razvita. Z današnjega stališča se perspektiva spremeni, saj nekatera transkripcijska orodja (npr. Transcriber in Praat; prim. poglavji 4.4.1 in 4.4.2) sama avtomatsko merijo časovni potek in označujejo časovne segmente (običajno posameznih izjav), in to na osem ali več decimalk natančno; dopuščajo tudi vzporedno zapisovanje prekrivnega govora.

4.1.3.8 Fonetične oznake

Stopnja uvajanja fonetičnih oznak v transkripcije niha od nič do zelo natančne fonetične transkripcije. Ta je sicer manj običajna pri transkribiranju večjih količin govorjenega besedila, ali pa je izvedena samo na manjšem delu celotne transkripcije. Bolj običajna praksa v takih primerih je uporaba ortografske transkripcije z uvajanjem nekaterih prozodičnih oznak (Johansson 1995, 93).

[122] Johansson 1995, 92.

4.1.3.9 Druge prozodične oznake

Oznake za glasnost in hitrost govorjenja, kvaliteto glasu in tonski potek se pogosto zapisujejo v transkripcijah. Priporočilo TEI za njihovo označevanje je preprosto: oznaka <*shift*> nakazuje točko, kjer se objezikovno dogajanje začne, in tudi točko, kjer se konča. Če oznaka stoji sama, brez dodatnih pojasnil, pomeni vrnitev v normalno stanje govora. Na začetku je oznaka <*shift*> običajno dopolnjena z opisom dogodka oz. lastnosti (*feature*) in oznako (*new*), ki nakazuje, da gre za novo stanje. Primer:[123]

```
<u who=A><shift feature=loud new=f>It's not the end of
Chanuhaf</shift> in case you are interested</u>
```

Iz zapisa lahko razberemo, da je del izjave, ki je med oznakama <*shift*>, izgovorjen glasno, da je to nova situacija v poteku govorjenja in da nekje na sredini izjave glasnost pade na običajno raven.

Oznaka <*shift*> naj bi predstavljala univerzalni mehanizem za opis prozodičnih oznak, ki je primerljiv z omenjenimi obstoječimi shemami oznak, hkrati pa kot sistem odprt za dovolj natančen opis širokega spektra različnih objezikovnih dogodkov.

4.1.3.10 Uredniške opombe

Transkripcijske sheme običajno predvidevajo prostor za uredniške oz. za transkriptorjeve opombe. Med shematizirane zapise sodijo oznake za nerazumljive odseke govorjenega besedila, pa tudi druge. TEI priporoča oznako odprtega tipa <*note*>, kamor transkriptor lahko zapiše kakršnokoli opombo, ki se mu zdi pomembna za pravilno interpretiranje transkribiranega besedila.

4.1.4 Referenčni sistem

Vsako besedilo in tudi del besedila mora biti označen tako, da ga je mogoče uvrstiti v širši besedilni kontekst in določiti njegov izvor. Po priporočilih TEI se za zapisana in govorjena besedila uporablja enotni sistem označevanja: referenčni sistem besedil je običajno zgrajen na pripisovanju vrednosti »id« in »n« (*name*) posameznim besedilom.

[123] Johansson 1995, 94.

4.1.5 Jezikoslovno označevanje besedila

Oblikoskladenjsko in/ali skladenjsko označevanje bistveno poveča in razširi možnosti uporabe korpusa, predvsem v smislu jezikoslovnih analiz; oznake se od korpusa do korpusa razlikujejo, TEI pa v zvezi s tem priporoča, da so oznake za govorjena in pisna besedila v referenčnih korpusih enake.

4.1.6 Kritika priporočil TEI

Na delavnici »Računalniški govorni korpusi«, ki je potekala l. 1993 na univerzi v Lancastru, je Jonathan Payne, eden od sodelavcev korpusnega projekta COBUILD, primerjal priporočila TEI s transkripcijsko shemo, uporabljeno pri gradnji korpusa COBUILD. Ugotovil je, da so »priporočila TEI v veliki meri primerljiva z obstoječimi transkripcijskimi realizacijami« in da bi bilo v večini primerov zelo enostavno pretvoriti »računalniku prijazna priporočila TEI v uporabniku prijazne sisteme označevanja« (Johansson 95: 95). Kljub tej načeloma spodbudni oceni pa je bil Payne v nekaterih točkah tudi precej kritičen do priporočil TEI. Avtorjem je npr. očital, da so njihova priporočila uporabnejša za tiste, ki se odločajo za zahtevnejše oblike transkribiranja, manj prozornosti pa je posvečene zahtevam tistih, ki za svoje namene potrebujejo enostavnejšo obliko transkripcije, npr. ortografsko. Priporočilom TEI je očital tudi, da so preveč natančna pri označevanju neverbalnih dogodkov, ki »transkriptorjev ne skrbijo preveč, saj se ti ukvarjajo predvsem z govornim signalom«, po drugi strani pa so premalo natančna pri določanju, kaj naj bo vključeno v glavo besedila. Na večino Paynovih očitkov so avtorji priporočil odgovarjali v smislu filozofije TEI, da ponujajo precej podroben nabor oznak, uporabniki pa se morajo sami odločiti, kaj od ponujenega bodo izbrali; nekatere Paynove pripombe so bile upoštevane pri pripravi kasnejših verzij priporočil TEI.

Zelo kritičen do priporočil je bil tudi John Sinclair, vodja gradnje korpusa COBUILD. V svoji kritiki (na delavnici v Lancastru) zapisal, da skupina »TEI ni naredila preveč velike usluge sebi in družbi, ko je javno predstavila oznake, ki se jih sploh ne bi smelo videti oz. jih lahko vidi samo računalnik« (Sinclair 1995, 107). Transkriptorji pa so ljudje in tako bo po Sinclairu ostalo še kar nekaj časa. V danih okoliščinah je težko pričakovati, da bodo ljudje zmogli zapisovati besedila po priporočilih TEI. »Dejstvo, da zna Lou Burnard[124] tako pisati, sodi v Guinessovo knjigo rekordov, ne velja pa nujno za nas ostale« (Sinclair 1995, 107). Vsekakor bi, zaključuje Sinclair, dober pretvorbeni računalniški program stvari postavil na svoje mesto.

[124] Glavni avtor priporočil TEI.

4.1.7 Sklep

Avtorji priporočil TEI so se ob objavi zavedali, da je shema oznak zelo kompleksna, vendar je taka z razlogi:

- da lahko dovolj natančno označi vse sestavine besedila in jih s tem naredi transparentne;
- da omogoči natančno računalniško obdelavo besedila.

Dober softverski program bi moral znati pretvoriti besedilo, označeno po priporočilih TEI in primerno za računalniško obdelavo, v format, primernejši za pisanje oz. branje (in obratno). Z današnjega stališča, skoraj dvajset let po objavi priporočil TEI, lahko ugotavljamo, da:

- govorni korpusi večinoma upoštevajo priporočila TEI, vendar jih modificirajo vsak na svoj način;
- se z razvojem programov, ki omogočajo povezavo med zvokom in transkripcijo, torej simultano predvajanje v dveh kodih, označevanje transkribiranih besedil poenostavlja (akustične, prozodične in fonetične informacije so neposredno dostopne).

4.2 PRIPOROČILA EAGLES ZA OZNAČEVANJE GOVORJENIH BESEDIL

Evropska iniciativa EAGLES (*Expert Advisory Group on Language Engineering Standards*)[125] je nastala l. 1993 na pobudo Evropske komisije v okviru *XIII. programa za jezikoslovje in jezikovne tehnologije*. Glavni namen ustanovitve skupine je bil pospešiti oblikovanje skupnih standardov za izdelavo obsežnih jezikovnih virov. Ena izmed petih delovnih skupin je bila zadolžena za področje govorjenih jezikov (EAGLES *Spoken language working group*).[126] Preden je izdelala svoja priporočila, sta bila v korpusnem jezikoslovju že uveljavljena dva transkripcijska standarda, NERC in TEI.

Projekt NERC (*Network of European Reference Corpora*) je poskušal standardizirati korpusne zapise v Evropi. V končno poročilo projekta je bila vključena transkripcijska konvencija, ki je nastala v okviru gradnje govorne komponente korpusa COBUILD (1991). Transkripcijska sistemizacija NERC ima 4 stopnje:[127]

[125] http://www.ilc.cnr.it/EAGLES96/home.html

[126] Uradna stran iniciative EAGLES je bila zadnjič prenovljena l. 1996, ena izmed delovnih skupin je delovala še leta 1998, uradna stran delovne skupine za področje govorjenih jezikov pa ni več aktivna.

[127] http://www.ilc.cnr.it/EAGLES96/spokentx/node19.html

1. **stopnja:** ortografski zapis z vključeno minimalno uporabo ločil; zamenjave govorcev niso označene; vključuje konvencije za ortografski zapis govorjenih besedil in za vstavljanje ločil.
2. **stopnja:** razširjena ortografska transkripcija, z osnovnimi informacijami o govorcih, menjavah govorcev in z oznakami za neverbalne dogodke.
3. **stopnja:** vključuje vse informacije 2. stopnje, dodane pa so še intonacijske in interakcijske informacije: označene so tonske enote in prekrivni govor; transkriptorji morajo biti dobro izurjeni in tonski posnetki zelo kakovostni.
4. **stopnja:** vključuje vse informacije s 3. stopnje, dodane pa so še intonacijske, akustične in fonetične informacije; transkripcija je povezana s spektrogramom.

Delovna skupina NERC za gradnjo (referenčnih) govornih korpusov priporoča drugo stopnjo, to je t. i. obogateno ortografsko transkripcijo, ki vsebuje poleg ortografskega zapisa besedila tudi osnovne informacije o govorcih, prekrivnem govoru in neverbalnih dogodkih; obogatena ortografska transkripcija naj bi bila primerna za jezikoslovne študije, ki ne potrebujejo podatkov o intonaciji in fonetičnih lastnostih glasov.

Skupina EAGLES si je v svojih priporočilih, objavljenih v knjigi *EAGLES Handbook on Spoken Language Sistems* (1995) in leto kasneje deloma dopolnjenih v internetni različici *EAGLES preliminary recommendations on spoken texts*[128] »prizadevala poiskati pot, kako doseči neke vrste kompatibilnost med priporočili NERC in priporočili TEI, hkrati pa se je ozirala na vse druge znane in uporabljene transkripcijske standarde in iskala stične točke med njimi, za podlago svojim priporočilom«.[129] V Priporočilih je zapisano, da ne želijo biti obvezujoča, da jih je treba preizkusiti v praksi (čeprav temeljijo na obstoječih in realiziranih projektih) ter da pomenijo šele prvi korak k skupnemu naboru transkripcijskih oznak in konvencij, ki bo povečal in razširil uporabnost virov govorjenega jezika.

Za označevanje izjave (*utterance*) je delovna skupina EAGLES prevzela definicijo TEI, po kateri je izjava definirana kot niz govorjenega jezika, ki je običajno spredaj in zadaj omejen s premorom ali z zamenjavo govorcev; dodajajo pa, da kriteriji za definiranje izjave v okviru spontanega govora niso enoumno določeni in jasni niti v monologih niti v dialogih.

[128] EAGLES Spoken Language Working Group, 1996, http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html.
[129] http://www.ilc.cnr.it/EAGLES96/spokentx/node33.html

4.2.1 Stopnje transkripcij

V zvezi s transkribiranjem govorjenih besedil je delovna skupina EAGLES v posebnem poglavju priročnika *Handbook on Spoken Language Systems* opisala različne vrste transkripcij in z njimi povezane stopnje označevanja. Ortografska transkripcija ali transliteracija za zapisovanje govorjenih besedil uporablja standardni (knjižni) zapis besed. Fonemska transkripcija temelji na fonemih izbranega jezika, alofonska transkripcija uporablja različne oznake za foneme v različnih okoljih, fonetična transkripcija pa zapisuje individualno izgovorjavo besed posameznega govorca.[130]

Stopnje označevanja govornih korpusov, ki temeljijo na zgoraj navedenih transkripcijah, so naslednje:

1. **Ortografska raven**, na kateri je za zapis izgovorjenih besed uporabljen standardni zapis.
2. **Citatno-fonemska raven**, kjer so besede zapisane s fonemi, zapis pa izhaja iz zapisa besede v izolaciji (ne iz konteksta).
3. **Širša fonetična raven**, ki zapisuje foneme, pri čemer upošteva izgovor besed v kontekstu, v procesu govorjenja, kar vključuje tudi dodajanje ali izločanje fonemov iz besed ali transformacijo enega fonema v drugega.
4. **Ožja fonetična raven**, ki si prizadeva natančno zapisati, kaj in kako je govorec izgovoril, tudi z variantami fonemov.
5. **Akustično-fonetična raven** ločuje vse segmente govora, ki so prepoznavni kot samostojni segmenti na sliki akustičnega valovanja ali na spektrogramu.

Posebej so znotraj vsake transkripcijske ravni zapisani oz. označeni pojavi, kot so neverbalni glasovi govorcev, nekomunikacijski dogodki in prozodične lastnosti govora.

Ortografska transkripcija govorjenega besedila je torej zapisovanje izjav govorcev v standardnem zapisu (transliteracija); ta stopnja zapisa je običajna za govorne in pisne korpuse.[131] Posebnosti ortografskega transkribiranja po načelih EAGLES so:

[130] http://www.ilc.cnr.it/EAGLES96/spokentx/node18.html#SECTION00035100000000000000

[131] Ortografska stopnja je bila izbrana tudi za transkribiranje posnetkov v skupnem evropskem projektu SpeechDat pri izdelavi korpusa govornih poizvedovanj po letalskih informacijah.

1. reducirane oblike besed
   - priporočljivo je uporabljati reducirane oblike besed, kot se pojavljajo v standardnem slovarju,[132]
   - če je potrebno, se lahko zapisujejo tudi druge reducirane oblike besed, ki jih ne najdemo v standardnem slovarju,
   - zapis reduciranih oblik besed je priporočljiv, kadar imajo visoko frekvenco pojavljanja in kadar vključujejo izbris celega zloga,
2. dialektalne oblike
   - dialektalne oblike morajo biti v transkripciji označene,
3. številke
   - številke so transliterirane kot besede,
4. kratice, okrajšave in črkovanja
   - v transliteraciji kratice zapišemo kot kratice in jih označimo,
   - kratice, ki so izgovorjene kot besede, so tudi zapisane kot besede,
   - črkovanje v govoru mora biti v transkripciji označeno,
5. medmeti
   - medmeti naj bi bili označeni in zapisani na način, kot ga najdemo v slovarju.

Osnovna filozofija delovne skupine EAGLES za govorjena besedila torej »temelji na načelu, da naj bo pri transkribiranju govorjenih besedil v največji možni meri upoštevan standardni zapis besed, vse nestandardne oblike v transkripciji pa morajo biti jasno označene«.[133] Konkretna realizacija teh načel je specifična za vsako jezikovno situacijo posebej.

Fonemska in fonetična transkripcija sta navadno zapisani s simboli IPA (*International Phonetic Alphabet*); to je v jezikoslovju najpogosteje uporabljeni transkripcijski sistem, pa tudi standard, ki ga za zapis fonemskih in fonetičnih informacij priporočata TEI in NERC. Zapleteni sistem simbolov IPA je težko berljiv za računalnik, zato je bil v okviru evropskega projekta *Speech Assessment Methodology* (SAM) razvit računalniku prijazen transkripcijski sistem SAM Phonetic Alphabet (SAMPA), ki je v sistem ASCII prenesel simbole IPA, in sicer tako, da je primeren za fonetično transkribiranje številnih evropskih jezikov; adaptacija je bila narejena tudi za slovenščino (http://www.phon.ucl.ac.uk/home/sampa/slovenian.htm).

[132] V priporočilih EAGLES ni posebej pojasnjen pojem standardnega slovarja, lahko pa predvidevamo, da gre za slovar, ki v določenem jezikovnem okolju velja za jezikovni standard. V slovenščini imamo Slovar slovenskega knjižnega jezika (morda skupaj z Besediščem), čeprav je v primeru tega slovarja "že zaradi letnice izida (1970-1991) jasno, da ne more biti več relevanten vir podatkov o sodobnem slovenskem jeziku in normi sodobnega knjižnega jezika" (Gorjanc in drugi 2005, [3]).

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

4.2.2 Prozodične oznake in neverbalni dogodki

Proces prozodičnega označevanja lahko opišemo kot zapisovanje lingvistično relevantnih dogodkov, ki se v izjavi zgodijo na področju jakosti, hitrosti, višine, časovnega in tonskega poteka zvočnega signala. TEI med prozodične elemente vključuje označevanje premorov, poudarjanj, tonskih enot ali intonacijskih fraz ter oznako za »spremembo« (<shift>), ki se nanaša na spremembe v hitrosti, višini in intenzivnosti govorjenja ter v ritmu in barvi glasu.

Za označevanje pojavov, ki spremljajo govor, je delovna skupina EAGLES predlagala nabor znakov, t. i. »minimalni skupni nabor dogodkov, ki naj bi bili označeni pri transkripciji različnih tipov govorjenih besedil«:[134]

1. **Glasovni polverbalni dogodki:** oklevanja, neverbalna komunikacijska sredstva (mhm, əm, aja) idr.; priporočljivo jih je zapisovati v seznam in slediti ortografski transkripciji teh glasov, kadar je mogoča (če so zapisani v standardnem slovarju).
2. **Glasovni neverbalni dogodki:** kašljanje, kihanje, tleskanje z jezikom, dihanje in vsi drugi neverbalni akustični zvoki, ki jih proizvaja govorec.
3. **Nekomunikacijski dogodki:** akustični dogodki, ki se zgodijo v okolju med snemanjem govora in so zapisani na posnetku; lahko jih proizvajajo drugi govorci, lahko pa gre za povsem druge zvoke – zvonjenje telefona, zapiranje vrat, listanje papirjev itd.
4. **Identifikacija govorca:** EAGLES povzema identifikacijo govorca po TEI, s pripombo, da obstajajo tudi manj zapletene oblike označevanja.
5. **Menjava govorcev:** označevanje po priporočilih TEI, osnova za definicijo »izjave«.
6. **Prekrivni govor:** označevanje po priporočilih TEI.
7. **Izpusti branega besedila:** kadar obstaja pisna predloga govorjenega besedila, je priporočljivo v transkripciji označiti morebitne govorčeve izpuste besed ali delov besedila.
8. **Samopopravki:** lahko so eksplicitno nakazani (npr. z besedami mislim, oziroma), lahko pa so implicitni, izkazani kot ponavljanje besed ali kot t. i. ponovni (lažni, napačni) začetki.
9. **Besedni fragmenti:** so glasovi, ki pripadajo besedi, ki v prvem poskusu ni bila do konca izrečena, zato jo govorec ponovi; tovrstna oklevanja/napačne začetke je v transkripciji priporočljivo označiti.
10. **Nerazumljivi fragmenti:** deli besedila, besede ali deli besed, ki jih transkriptor ni razumel; včasih je koristno razlikovati med negotovo tran-

[134] http://www.ilc.cnr.it/EAGLES96/spokentx/node23.html

skripcijo (ugibanjem) in popolnoma nerazumljivim delom besedila.

V priporočilih EAGLES ostaja nedorečeno zelo pomembno vprašanje, to je vprašanje ločil. V priporočilih NERC ortografska transkripcija vključuje delno uporabo ločil, in sicer piko na mestu, kjer transkriptor čuti stavčno mejo, in veliko začetnico na začetku stavka. Nasprotno je stališče v priporočilih SpeechDat, ki v celoti odsvetujejo uporabo ločil v transkripciji. Stališče delovne skupine EAGLES je le rahlo nakazano v ugotovitvi, da postavljanje stavčnih mej v govorjenem besedilu ni nikoli enostavno in da je zato uporaba ločil v ortografski transkripciji lahko zelo zapletena in tudi kontroverzna naloga.

Sicer pa EAGLES povzema »izredno pomembno priporočilo« (*fundamental recommendation*, Sinclair 1993, 70)[135] po NERC-u, in sicer, naj bodo digitalne verzije vsakega posnetka vključene kot posebne komponente v korpus. Tako priporočila NERC in TEI kot priporočila EAGLES so namreč nastala v obdobju, ko obstoječe tehnologije še niso omogočale neposredne povezave zvoka in zapisa, čeprav so raziskave že dajale slutiti razvoj v to smer. Januarja 1996 je bil na konferenci delovne skupine EAGLES med prioritetne naloge uvrščen razvoj mehanizmov, ki bi omogočali hkratni dostop do zvočnega signala. Prvi poskusi v tej smeri so potekali na adaptaciji korpusa IBM/Lancester, poimenovanega tudi SEC (Spoken English Corpus) in kasneje MARSEC.[136]

4.3 Transkripcijski standardi

Podrobneje si bomo ogledali nekaj transkripcijskih shem, ki so bile narejene ob gradnji različnih korpusov. Čeprav se načrtovalci vsakega korpusa odločijo za svojo transkripcijsko shemo glede na namen korpusa in na specifične lastnosti jezika ter jezikovne situacije, je gotovo smiselno preučevati obstoječe transkripcijske modele in se učiti na njih. Predstavila bom nekatere starejše modele, ki so nastali še pred priporočili TEI, nato »klasični« BNC-jev transkripcijski model, ob katerem so nastajala priporočila TEI, nenazadnje pa še enega izmed sodobnejših modelov, kjer sinhronizacija zvoka in zapisa dopušča poenostavitev transkripcijskih shem.

[135] Sinclair, J., Text Representation: Written Language / Spoken Language, Chapter 3, NERC Report. V: Calzonari, N., Baker, M., Kruyt, P. G. (Ur.), *Towards a Network of European Reference Corpora.* Pisa: Giardini, 1993.

[136] Prvi korpus, ki je že v izhodišču predvidel sinhronizacijo zvoka in transkripcije in jo tudi realiziral, je bil korpus COLT, dokončan l. 1998 na Univerzi v Bergnu.

4.3.1 Prozodična transkripcija korpusa London-Lund

Besedila korpusa London-Lund, najstarejšega govornega korpusa, so bila transkribirana po načelih prozodične transkripcije, kar pomeni, da so bile v pisni obliki označene nadsegmentne lastnosti govora, to je tonski poteki, premori in poudarki. Korpus je bil narejen z namenom, da bi služil kot vir za slovnični opis britanske angleščine, odločitev za zahtevnejši tip transkripcije pa je zelo podaljšala njegovo nastajanje. Osnovna enota prozodičnega označevanja v korpusu je bila »tonska enota« (*tone unit*), ki je lahko vsebovala tudi podenote. V vsaki enoti je bil označen t. i. začetek/nastop (*onset*), to je prvi izraziti zlog v tonski enoti, tonski potek (rastoči, padajoči, izravnan, rastoče-padajoči itd.), relativna tonska višina, dve vrsti premora (krajši ali daljši) in dve stopnji poudarka (normalen ali močnejši). Označena je bila tudi identiteta govorcev, prekrivni govor, neverbalni in nekomunikacijski dogodki (smeh, kašelj, zvonjenje telefona itd.) in nerazumljive besede. Gre torej za model, ki je v najbolj abstrahirani obliki prisoten v vsaki sodobni transkripcijski shemi, seveda kadar gre za prozodične transkripcije.

Najznačilnejše prozodične oznake korpusa London-Lund so prikazane v spodnji tabeli:[137]

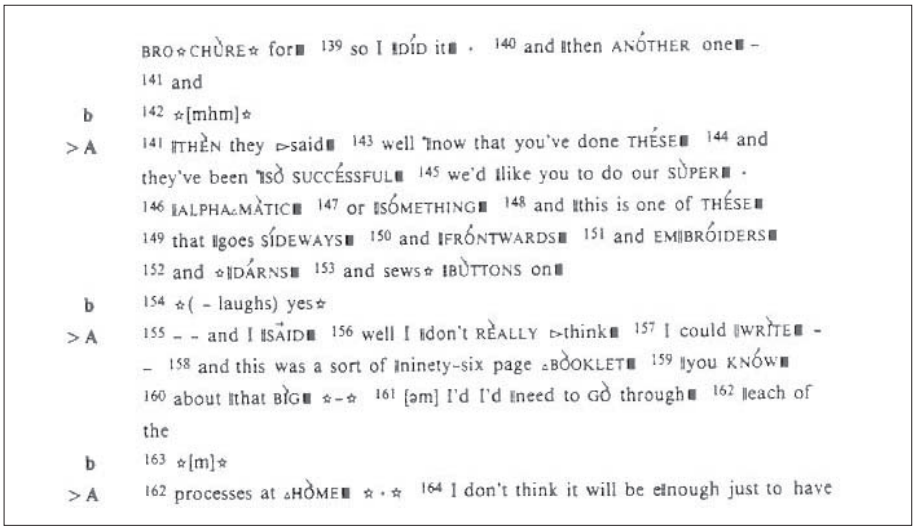
Simbol	Opis
številka (139, 140)	začetek tonske enote
■	konec tonske enote
‖	prvi izraziti (poudarjeni) zlog v enoti
[...]	podrejena tonska enota (podenota)
☆ … ☆	prekrivni govor
((...))	nerazumljivi govor
(...)	nekomunikacijski in neverbalni dogodki
à á â ā ě	padajoči, rastoči, rastoče-padajoči, izravnan, padajoče-rastoči ton
▹ ▵ △	relativna tonska višina: kontinuiranost, višje kot prejšnji zlog, višje kot prejšnji poudarjeni zlog, zelo visoko
– —	premor: krajši (dolžina kratkega zloga), daljši (dolžina poudarjenega zloga), kombinacije
’	močnejši poudarek

Tabela 15: Prozodične oznake korpusa London-Lund[138]

[137] Oznake se nanašajo na verzijo, natisnjeno v knjigi; oznake na listkovnem gradivu korpusa SEU se nekoliko razlikujejo od teh, prev tako oznake na CD-romu.

[138] *A Corpus of English Conversation*, ur. J. Svartvik in R. Quirk, 1980.

Oznake, uporabljene v transkribiranem besedilu, prikazuje naslednja slika:



Slika 19: Primer prozodične transkripcije korpusa London-Lund[139]

Podrobna identifikacija besedil in govorcev je dostopna v posebnih seznamih, v transkripciji pa je identiteta govorcev označena s črkami (A, b). Velika tiskana črka pomeni, da govorec ni vedel za snemanje, mala črka pa, da je vedel in da je bila njegova naloga spodbujanje poteka pogovora; besedila teh govorcev prozodično niso bila označena.

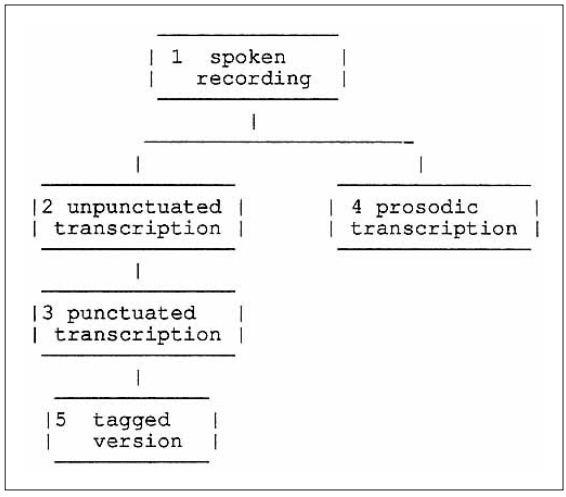
Prozodično transkribiranje je izredno zahtevna naloga, ki jo lahko uspešno opravijo samo dobro usposobljeni fonetiki. Označevanje je v (ne tako davni) preteklosti temeljilo bolj ali manj na individualni presoji označevalcev: sprememba tonskega poteka, njegova višina in podobno – vse to je bila stvar človeške interpretacije in ocene so se med seboj lahko zelo razlikovale. V primeru označevanja korpusa London-Lund sta bila označevalca dva. Za preizkus sta del besedil (pribl. 9 odstotkov) transkribirala oba; »prekrivne« transkripcije so služile za nadaljnje študije razlik med transkripcijo enega in drugega označevalca, izkazalo pa se je, da so bile razlike med njima precejšnje. Dandanes je z uporabo programskih orodij za analizo govora mogoče dosegati veliko večjo stopnjo objektivnosti, vendar pa lahko s temi programi učinkovito delajo samo usposobljeni fonetiki.

[139] *A Corpus of English Conversation*, ur. J. Svartvik in R. Quirk, 1980.

Prozodično označeni korpusi so tudi zelo zahtevni za nadaljnje računalniške obdelave, saj so oznake pogosto sredi besed, kar zmanjšuje robustnost korpusa. Problematične so lahko tudi oznake: če so te del standardnega nabora znakov, kot je bilo to za tonski potek v korpusu London-Lund, se ne razlikujejo dovolj od besedila samega. Če pa se uporabi nov nestandardni nabor oznak (primer korpusa Lancaster/IBM), se pogosto zgodi, da računalniki in tiskalniki oznak ne prepoznajo, kar zmanjšuje uporabnost korpusa. Tudi zato se za transkribiranje večjih govornih korpusov praviloma uporablja ortografska ali razširjena ortografska transkripcija, del korpusa pa je lahko tudi fonetično in akustično označen.

4.3.2 Večstopenjska transkripcija korpusa Lancaster/IBM

Korpus Lancaster/IBM je bil transkribiran na tri načine: kot transliteracija brez ločil, z ločili in kot prozodična transkripcija; tudi tu je bila odločitev pogojena z namenom korpusa, saj je nastal za potrebe razvoja na področju sinteze in analize govora. Spodnja shema prikazuje sosledja različnih verzij transkribiranja in označevanja korpusa Lancaster/IBM:



Slika 20: Sosledje transkribiranja in označevanja korpusa Lancaster/IBM[140]

[140] http://khnt.hit.uib.no/icame/manuals/sec/SAMP.HTM#1

Transkripcija brez ločil in prozodična transkripcija sta nastali neposredno iz zvočnih posnetkov, transkripcija z ločili (avtorji korpusa jo imenujejo ortografska transkripcija) je bila narejena na podlagi transkripcije brez ločil, oblikoskladenjsko označevanje pa je bilo izvedeno na podlagi transkripcije z ločili. Verzije od 1 do 4 so bile narejene ročno, verzija 5 pa polavtomatsko. Označevanje je potekalo na ortografsko transkribiranem besedilu, saj naj bi bila prozodična transkripcija prezahtevna za računalniško branje. V nadaljevanju so predstavljene vse tri transkripcije.

Transkripcija brez ločil poskuša posnemati govor, v katerem ločil seveda ni, so pa tonski poteki in premori, ki jih pri pisanju pogosto (ali vsaj nekatere izmed njih) zaznamujemo z ločili. Pri tovrstni transkripciji tudi ni velikih začetnic, ki bi zaznamovale začetek stavka; številke so zapisane s številkami v primerih, ko bi bile tako zapisane tudi v pisnem besedilu – hišne številke, telefonske številke, decimalne številke itd.; označene so tudi menjave govorcev. Transkripcija brez ločil je bila podlaga za ortografsko transkripcijo, pa tudi za prozodično, pri čemer sta slednji dve nastali neodvisno druga od druge. V transkripcijo brez ločil so ločila vstavili sodelavci z obeh sodelujočih institucij, Univerze v Lancastru in IBM-a, pri čemer jim zvočni posnetki niso bili na razpolago; ločila so morali vstavljati po smislu. To je bilo mogoče, če upoštevamo, da korpus sestavljajo predvsem radijske oddaje in (brana) univerzitetna predavanja, torej zelo malo spontanega govora. Le v redkih primerih so bili stavki dvoumni in se jih je dalo interpretirati na več načinov in v taki primerih so transkriptorji dobili pomoč zvočnega posnetka. Na spodnjih dveh izsekih je mogoče primerjati zapisa brez ločil in v t. i. ortografski transkripciji:[141]

```
good morning more news about the Reverend Sun Myung
Moon founder of the Unification church who's currently
in jail for tax evasion he was awarded an honorary de-
gree last week by the Roman Catholic University of la
Plata in Buenos Aires Argentina in announcing the award
in New York the rector of the university Dr Nicholas
Argentato described Mr Moon as a prophet of our time
```

Slika 21: Primer transkripcije brez ločil iz korpusa Lancaster/IBM

[141] *Manual of information to accompany the SEC corpus,* http://khnt.hit.uib.no/icame/manuals/sec/SAMP.HTM#1.

```
Good morning. More news about the Reverend Sun Myung
Moon, founder of the Unification church, who's currently
in jail for tax evasion: he was awarded an honorary
degree last week by the Roman Catholic University of
la Plata in Buenos Aires, Argentina. In announcing the
award in New York, the rector of the university, Dr Ni-
cholas Argentato, described Mr Moon as a prophet of our
time.
```

Slika 22: Primer ortografske transkripcije z ločili korpusa Lancaster/IBM

Tudi korpus Lancaster/IBM sta prozodično označila dva strokovnjaka, eden z univerze v Lancastru in eden iz laboratorija IBM. Kot osnovo za transkribiranje govora sta imela zvočne posnetke in transkripcijo brez ločil. Delala sta ločeno, pribl. 10 odstotkov korpusa pa sta transkribirala oba, da so lahko primerjali njuno prozodično označevanje. Osnovna enota prozodične transkripcije je bila tudi tu tonska enota (označena z dvema navpičnima črtama – glavna, in z eno črto – manj izrazita, podrejena). Tonski potek, ki je lahko visok, nizek, padajoč, rastoč, nizek rastoče-padajoč itd. (14 možnosti), je označen samo na naglašenih zlogih; kratek izsek prozodično transkribiranega besedila prikazuje spodnja slika:

```
#143Good `morning || #143`more news about the Rever-
end _Sun Myung Moon |_founder of the Unification Church
|who'scurrently in jail | for tax evasion || he was au-
warded an _honorary deigree last week |
```

Slika 23: Prozodična transkripcija besedila korpusa Lancaster/IBM[142]

Korpus Lancaster/IBM je sicer razmeroma majhen korpus govorjenega jezika (52.000 besed), s stališča referenčnosti tudi neuravnotežen (predvsem radijske oddaje), njegova največja prednost in pomembnost za nadaljnji razvoj korpusnega jezikoslovja pa so natančni zapisi govorjenih besedil v različnih oblikah.

[142] http://khnt.hit.uib.no/icame/manuals/sec/SAMP.HTM#1

4.3.3 Poenostavljena transkripcija korpusa COBUILD

Transkripcijska konvencija korpusa COBUILD je bila rezultat tridesetletnih izkušenj pri transkribiranju govorjenih besedil, podrobno pa je bila izdelana v letih 1991–1993 v sodelovanju med založbo COBUILD in forenzičnim laboratorijem, specializiranim za govor in jezik *JP French Associates* (Payne 1995, 203). Tudi v tem primeru je odločitev za transkripcijski standard narekovala namembnost korpusa: govorna komponenta referenčnega korpusa COBUILD je v novembru 1994 obsegala čez 10 milijonov besed in je bila primerljiva z velikostjo BNC; tako veliko količino posnetkov je bilo mogoče transkribirati samo z ortografsko transkripcijo. Za transkriptorje so bile izbrane osebe z dobrim znanjem tipkanja po nareku; te osebe so bile navajene v besedila med pisanjem vstavljati ločila, zato je bilo »logično, da so v COBUILD-u ločila vključili v transkripcijsko shemo« (Payne 1995, 203). Tako so govorjena besedila v transkripciji segmentirali v »funkcionalne stavke«, ki jih je težko jezikoslovno definirati, praktično pa transkriptorjem njihova določitev ni povzročala večjih problemov (Payne 1995, 204). Uporabljali so lahko samo piko in vprašaj; če ni bilo ne enega ne drugega, je to pomenilo nedokončano misel; klicajev, vejic, podpičij in dvopičij v transkripciji ni bilo.

V transkripcijah so bile ortografsko zapisane nekatere pogosto rabljene nestandardne oblike (*gonna* namesto *going to, 'cos* namesto *becouse*) in oblikovani seznami dovoljenih nestandardnih zapisov.[143] Izdelani so bili tudi seznami polverbalnih komunikacijskih sredstev (*erm, mm*), opisani neverbalni zvoki, označeni nerazumljivi deli besedila, premori in napačni začetki. Prekrivnega govora v COBUILDU niso posebej označevali, posebej če se je zgodil na koncu izjave enega govorca in na začetku izjave drugega govorca; če pa se je zgodil znotraj izjave posameznega govorca, so izjavo prekinjenega govorca razdelili na dve izjavi in vmes vstavili izjavo govorca, ki je spregovoril kasneje. V korpusih, ki nastajajo v zadnjem desetletju, takšno zapisovanje ne pride več v poštev: pozicija in dolžina prekrivnega govora sta vedno označeni, to omogočajo tudi transkripcijska orodja.

Transkripcijska shema korpusa COBUILD je dokaj preprosta, nikakor pa je ne gre podcenjevati, prej nasprotno, ker vemo, da je bila funkcionalna in da je govorni korpus služil svojemu namenu. Po Paynu (195: 207) je bila transkripcijska shema sestavljena z namenom, da bi korpus uporabljali za primerjavo med formalnim in neformalnim govorjenim jezikom ter med govorjenim in pisnim jezikom; hkrati so imeli njeni avtorji namen izdelati shemo, ki bi jo zlahka usvojili

[143] Celoten seznam dovoljenih nestandardnih oblik (*normalized forms from allowed list*; Payne 1995, 204) ni dostopen, prav tako niso znani natančni kriteriji, po katerih je bila beseda uvrščena na seznam; vsekakor je princip metodološko prenosljiv na druge jezike.

tudi transkriptorji-nespecialisti in ki bi jo uporabniki korpusa lahko brez večjih težav brali. Pretvoriti uporabniku prijazno shemo v standardizirano računalniku prijazno obliko pa je po Paynu (1995, 207) samo stvar enostavnega konverzijskega računalniškega programa.

4.3.4 Transkripcija BNC

Tudi 10-milijonska govorna komponenta referenčnega korpusa BNC je bila transkribirana v ortografski transkripciji. Načela za transkripcijo so nastajala sočasno in v povezavi s priporočili TEI, zato korpus BNC pomeni prvo praktično realizacijo teoretičnih predpostavk TEI. Za transkribirano besedilo to pomeni, da je bilo »bolj približano pisnemu besedilu kot aktualnemu zvočnemu zapisu« (Burnard 2000, 9). Relativno enostaven način transkribiranja je dopuščal, da so delo opravljali nelingvisti, ki so morali najprej opraviti nekajtedenski tečaj, ko so njihovo transkribiranje skrbno spremljali. Ni nepomembno, da so govor z določenega regionalnega področja vedno zapisovali transkriptorji, ki so bili tudi sami govorci tega področja. Tudi kasneje so transkripcije preverjali, in sicer so vsako peto transkripcijo primerjali z zvočnim zapisom, da bi zagotovili konsistentnost transkribiranja (Crowdy 1995, 228). Zbrani posnetki govorne komponente korpusa BNC so znašali 1200 ur, kasnejša analiza pa je pokazala, da je bila povprečna transkripcijska norma okrog 750 besed na uro (Crowdy 1994, 26), čeprav je ta izredno nihala v odvisnosti od kvalitete posnetka. Kljub nenehnemu usklajevanju je bilo nemogoče zagotoviti, da bi vsi transkriptorji glasove zapisovali na enak način, posebej kadar je šlo za zvoke, ki nimajo standardnega zapisa, npr. »əm«, »hm« itd., saj gre za subjektivno interpretacijo zvoka; ugotovljeno je bilo tudi, da celo posamezni transkriptor ne zapisuje vedno na enak način in z enako doslednostjo.

Vse transkripcije govorne komponente BNC so bile narejene v organizaciji založbe Longman. Od tam so bile poslane v Računalniški center Univerze v Oxfordu, kjer so bile narejene adaptacije transkripcij po priporočilih TEI. Ta podatek je zelo pomemben, ker pomeni, da transkriptorji niso zapisovali neposredno v zahtevnem sistemu TEI, ampak so bile oznake, ki so nujne za nadaljnjo računalniško obdelavo, dodane kasneje (kar pa seveda pomeni tudi en korak več pri gradnji korpusa); to pomeni, da je bil transkriptorjem prijazen kodifikacijski sistem v Oxfordu transformiran v SGML format,[144] kakršnega je predstavila TEI v svojih priporočilih.

Primer transkripcije, kakor so jo izdelali transkriptorji založbe Longman in preden je bila poslana v računalniški center na Oxfordu:

[144] *Standard Generalised Markup Language* je metajetzik, ki določa obliko in strukturo oznak, dodanih osnovnemu besedilu.

```
<1>  You gotta Radio Two with that. Bloody pirate station
     wouldn't you
```

Slika 24: Odlomek transkribiranega besedila korpusa BNC

Isto besedilo v SGML formatu:

```
<u who=d00011>
<s n=00011>
<event desc="radio on"><w PNP><pause dur=34>You
<w VVD>got<w TO0>ta <unclear><w NN1>Radio
<w CRD>Two <w PRP>with <w DT0>that <c PUN>.
<s n=00012>
```

Slika 25: Odlomek transkribiranega besedila korpusa BNC v SGML formatu[145]

Iz zgornjega primera je razvidno, zakaj so priporočila TEI pri nekaterih strokovnjakih naletela na tolikšen odpor. Besedilo je težko berljivo, težko pa ga je tudi zapisovati. V zvezi s tem so zanimiva načela glede ločil: izjava je v BNC definirana kot intonacijska enota in transkriptorji so imeli nalogo, da po intuiciji vnašajo v besedilo ustrezna ločila, omejena na piko, vejico, vprašaj in klicaj, s tem pa govorjeno besedilo segmentirajo v enote, podobne stavkom. Ločila so lahko vnašali samo na mesta, kjer je bilo to skladenjsko ustrezno, tudi če v govoru ni bilo nobenega premora (Crowdy 1994, 27), niso pa jih smeli vnašati na mesta, kjer so jih govorci sicer z glasom (ali premorom) nakazali, niso pa bila skladenjsko ustrezna.

Osnovna enota segmentiranja govorjenega besedila v BNC je bila menjava govorcev (*speaker turn*). Govorci so bili označeni s številkami v trikotnih oklepajih <1>, <2>, glede na to, v kakšnem zaporedju so se pojavljali v besedilu. Številka je bila povezana z govorčevo identifikacijo v glavi besedila. Govorci so bili označeni enoumno (če se je isti govorec pojavil na več posnetkih, je imel v transkripciji isto oznako), z namenom, da bi lahko opazovali govor posameznega govorca v različnih besedilnih vrstah. Če transkriptor govorca ni mogel identificirati, ga je označil z <?>.

Glede zapisovanja polverbalnih glasov[146] (*mhm*, *erm*) je bil sestavljen odprt seznam glasov. »Mnoge nestandardne oblike so bile zapisane ortografsko, kot npr.

[145] *BNC Users Guide* 2000, 7.3.
[146] V BNC *vocalised pauses*, pri EAGLES *semi-lexical events*.

dunno, *gonna*, *cos*« (Crowdy 1994, 27). Veljalo je tudi pravilo, da se tudi vse nestandardne oblike, ki se pojavijo kot gesla v splošnih slovarjih,[147] lahko v nestandardni obliki zapisane v transkripciji. Za t. i. nekomunikacijske zvoke (v BNC *contextual comment*) je veljalo, da se jih vpisuje v transkripcijo samo v primeru, če so kakorkoli relevantni za potek sporazumevanja.

Transkripcija BNC je zagotavljala anonimnost govorcem in osebam, omenjenim v besedilih, z uvedbo posebnih oznak, ki so nadomeščale osebna lastna imena, naslove in telefonske številke. »Imena javno znanih oseb in javno znani naslovi« niso bili spremenjeni (Crowdy 1994, 28), pri čemer vsaj iz dostopnih opisov ni mogoče razbrati opredelitve, v katerih primerih gre za javno znane osebe.

V zvezi s kraticami in okrajšavami je v transkripciji BNC veljalo načelo, da se zapisujejo z velikimi tiskanimi črkami in vmesnimi presledki, kadar so izgovorjene kot kratice (primer *PhD*, zapisano *P H D*), kadar pa so izgovorjene kot besede, se jih zapisuje z velikimi tiskanimi črkami brez presledkov (*NATO*).

4.3.5 Göteborška modificirana ortografska transkripcija

V 2. poglavju (Govorni korpusi) smo že omenili transkripcijski standard, ki je nastal ob gradnji 1,3-milijonskega korpusa govorjene švedščine. Standard sestavljata dva dela: prvi je göteborški transkripcijski standard (GTS), ki je neodvisen od jezika, vključuje pa oznake za zapisovanje izjav, prekrivnega govora, premorov, neverbalnih zvokov itd.; preizkušen je bil na kitajskem, arabskem, angleškem, španskem, bolgarskem in finskem jeziku. Drugi del je t. i. modificirana standardna ortografija, to pa je standard za zapisovanje govorjene švedščine (*Modified Standard Ortography*, MSO).

GTS prinaša poenostavljene oznake, se pa bistveno ne razlikuje od standardov, ki so bili že predstavljeni. Za transkribiranje govorjenih besedil je bila uporabljena standardna ortografija, razen v primerih, kadar v govoru obstaja več variant izgovora posamezne besede. Modificirana standardna ortografija je bližje govorjenemu jeziku kot standardna ortografija, vendar ni tako natančna, kot bi bila fonemska ali fonetična transkripcija (Allwood 1998, 3). Nekaj primerov:

- standardni zapis besede *jag* (jaz) je v MSO zapisan kot *jag* ali *ja*, ker se v govoru uporabljata obe varianti,
- standardni *och* (in) je lahko v MSO zapisan kot *å*, *och* ali *o*,
- standardni *är* (je) pa kot *e*, *ä* ali *är*.

[147] »*General dictionaries*«; Crowdy 1994, 27.

O tem, katere govorne variante bodo v MSO zapisovali, so se odločali do določene mere poljubno, po konsenzu; kasnejša refleksija je pokazala, da bi lahko zapisovali še več govornih variant (Allwood 1998, 3).

Kadar se je govorjena oblika besede od zapisane razlikovala samo po redukciji glasov iz zapisane oblike, so bili v MSO za zapis uporabljeni zaviti oklepaji, v katerih so bili zapisani manjkajoči glasovi; v primeru besede *jag* (jaz) je govorna varianta *ja* zapisana kot

ja{g}.

V primerih, ko je imela govorjena varianta besede enako obliko zapisa kot kašna druga standardna ali nestandardna oblika, so zapise razdvoumljali z numeričnimi indeksi. Tako npr. govorjeni *å* lahko pomeni *och* (in) ali *att* (člen za nedoločnik):

å1 – *och* (in)
å2 – (člen za nedoločnik).

Spodnja slika prikazuje primer transkribiranega besedila v standardu GTS:

§1. Small talk
$D: säger du de{t} ä{r} de{t} ä{r} de{t} så besvärlit då
$P: ja ja
$D: m // ha / de{t} kan ju bli så se{r} du
$P: < jaha >
@ <ingressive>
$D: du ta{r} den på morronen
$P: nej inte på MORRONEN kan ja{g} ju tar allti en promenad på förmiddan [1 å0]1 då vill ja{g} inte ha [2 den]2 medicinen å0 sen nä ja{g} kommer hem möjligtvis
$D: [1 {j}a]1
$D: [2 nä]2
$D: oh I see is it it is so troublesome then
$P: yes yes
$D: m // yes / it can be that way you see
$P < yes >
@ <ingressive >
$D: you take it in the morning
$P: no not in the MORNING I always take a walk before lunch [1 and]1 then I don't want [2 that]2 medicine and then when I get home possibly
$D: [1 yes]1
$D: [2 no]2

Slika 26: Primer transkribiranega besedila v švedskem govornem korpusu [148]

[148] Allwood in drugi 2000.

4.3.6 Transkripcijske konvencije na Slovenskem

V preteklosti je bilo transkribiranje povezano predvsem z zapisovanjem slovenskih dialektov; na tem področju je bilo opravljenega veliko transkripcijskega dela, izdelani pa so bili tudi določeni standardi zapisovanja. Manj raziskav je bilo narejenih v zvezi z zapisovanjem nedialektalnega govorjenega (pogovornega) jezika; v jezikoslovju je šlo za posamezne manjše raziskave, nekaj poskusov zapisa govora, pri katerih so včasih sodelovali tudi jezikoslovci, pa je bilo tudi na literarnem področju.[149]

Sredi 90. let se je na Slovenskem pojavila težnja po izdelavi standardov za zapis govora, ki naj bi poenotili raziskovalno delo na področju govornih tehnologij (Zemljak idr. 2002, 159). Nekaj zainteresiranih skupin raziskovalcev (FERI in Pedagoška fakulteta v Mariboru, Fakulteta za elektrotehniko in Inštitut Frana Ramovša v Ljubljani) je skupaj izdelalo načela za računalniški fonetični zapis slovenskega govora. Zapis temelji na mednarodni fonetični abecedi IPA, ta pa je za lažjo računalniško berljivost pretvorjena v mednarodno uveljavljeni sistem znakov MRPA (*Machine Readable Alphabet*). Sistem v celoti upošteva fonetične značilnosti slovenskega govora, kot so opisane v Slovenski slovnici (Toporišič 1976). Tako velja, da ima slovenščina 8 samoglasnikov, ki so dolgi ali kratki, naglašeni ali nenaglašeni, poleg tega pa 6 zvočnikov (z alofoni) in 16 nezvočnikov (z alofoni). V članku, kjer je transkripcijski standard predstavljen (Zemljak idr. 2002), so navedeni predvsem primeri iz slovnice, ni pa primerov transkripcij realnega govora.

V zadnjem času je v slovenskem jezikoslovju nastalo nekaj razprav, ki se z različnih zornih kotov ukvarjajo z govorjenim jezikom, pri tem pa (vsaj do določene mere) uporabljajo korpusni pristop.[150] Ker študije temeljijo na posnetkih spontanega govora, ki je transkribiran, lahko iz njih posredno razberemo tudi določena transkripcijska načela. Študije so nastale z različnimi nameni, zapis govorjene slovenščine pa nikjer ni bil v središču pozornosti.

Verdonik (2006, 69) pri zapisu govorjenega besedila sledi priporočilom EAGLES. Besedilo je zapisano »ortografsko, skladno s knjižnim standardom – to pomeni, da so tudi pogovorno, narečno, površno ipd. izgovorjene besede zapisane tako, kot je predvideno v knjižnem standardu, ne tako, kot so dejansko izgovorjene /…/« (Verdonik 2006, 69). »Izjema so naslednje besede:

[149] Branko Gradišnik, *Nekdo drug* (1990; avtor dodatnega besedila V. Gjurin); Irvine Walsh, *Trainspotting* (prev. Andrej Skubic, 1997); Andrej Skubic, *Fužinski bluz* (2001); Goran Vojnović: *Čefurji raus!* (2008).
[150] Vitez in Zwitter Vitez 2004, Verdonik 2005 in 2006, Smolej 2006a.

- pogovorni nedoločnik je prepisan brez končnega –i, če je tako izgovorjen,
- *mogli* v pomenu morali,
- *najdli* (od najti),
- *taki* (v pomenu tak, takšen),
- *pol* (v pomenu potem),
- *more* (namesto oblike mora),
- *večih, večim* (sklanjanje nesklonljive besede več).[151]

Besedam, ki niso izgovorjene, kot predvideva knjižni standard, sta dodana fonetični prepis (v mednarodni računalniški fonetični abecedi SAMPA) v oglatih oklepajih in oznaka +pron=* ali +pron=izg, npr.: tudi[t/u:t][+pron=*], kakšne[k/a:SnE][+pron=izg]« (Verdonik 2006, 69–70).

Pri Verdonik tako že najdemo zametek seznama dovoljenih nestandardnih zapisov, kakršne poznamo pri tujih govornih korpusih.[152] Ugotavljam lahko tudi, da bi seznam, za katerega so besede prispevali predvsem govorci iz štajerske narečne skupine, razen ene izrazito narečne prvine (*taki*) povsem ustrezal (oz. bi ga bilo treba še razširiti) tudi za osrednji slovenski govor, kateremu je pripadala večina govorcev na mojih posnetkih. Zelo zanimiva je tudi ideja o fonetičnem prepisu besed, ki v izgovoru odstopajo od pričakovanega standarda, čeprav se tu zastavlja nekaj vprašanj: prvič, kaj je pričakovani govorni standard oz. na kaj se transkriptorji pri tem lahko opirajo, in drugič, zdi se, da bi bilo za gradnjo velikega govornega korpusa zaradi obsežnosti gradiva tovrstno označevanje praktično nemogoče izvesti na celotnem gradivu, vsekakor pa bi bilo koristno vsaj na delu gradiva.[153]

Tudi Smolej (2006a) za zapis govora uporablja »knjižni oz. ortografski zapis. /…/ Kljub odločitvi za knjižni zapis pa so pri transkribiranju upoštevane nekatere izjeme, ki niso skladne s standardnim slovarjem (SSKJ) – odločili smo se za tiste izjeme, ki so skupne vsem govorcem posnetih besedil in kljub drugačnosti od knjižnega zapisa ne otežujejo berljivosti« (Smolej 2006a, 13). Upoštevane so npr.:

- izguba nenagl. končnice -o pri samostalnikih srednjega spola (*mlek*),
- izguba glasu i pri deležnikih množinske oblike moškega sploa (*igral, spal*) in izguba glasu o pri deležnikih srednjega spola (*dogajal*),
- izguba i-ja in u-ja v predponah in v osnovi (*prpravt, prjatu, drgač*),
- izguba končnice -i v dajalniku in mestniku (*men, pr men*),
- izguba nenaglašenega e in a v besedah (*clo, zlo*),
- ohranjeno ukanje v prednaglasnem zlogu (*utrok, purudnišnica*),

[151] Vse navedene besede so v transkripciji tudi posebej označene.
[152] *Allowed list* pri BNC.
[153] Za tovrstni prepis bi bilo treba k sodelovanju pritegniti fonetike.

- deležnik na -l za moško obliko se razen izjemoma zapisuje z –u (*naredu*) (Smolej 2006, 13–14).[154]

Čeprav je tovrstni zapis gotovo ustrezal namenu raziskave, pa se za namen gradnje govornega korpusa ne zdi dobro izhodišče, in to zaradi prevelikih nedoslednosti, saj ne gre niti za ortografski niti za fonetični zapis; v zapisanih besedah je le deloma upoštevana prilagoditev govoru (*purudnišnica* namesto *purudnišnca*, *prpravt* namesto *prpraut* itd.) Če bi se pri gradnji govornega korpusa odločili za fonetično transkripcijo (na delu korpusa ali na celoti), bi jo morali dosledno upoštevati. Sicer pa Smolej navaja (2006b) tudi seznam t. i. besedilnih aktualizatorjev, ki prav tako niso del knjižnega jezika, vsaj ne v zapisani obliki, bi pa zagotovo sodili na seznam dovoljenih oblik/besed. To so:

- en (sredstvo nedoločnega uvrščanja v besedilni svet),
- ta (v funkciji določnega člena in v deiktični vlogi),
- tist (v deiktični vlogi in drugo),
- un (splošno znana referenca in drugo).

Navedeni leksemi so tudi meni pri transkripciji vsaj na začetku povzročali težave; vsekakor za namene govornega korpusa predlagam zapis, ki je prilagojen govorjenemu izrazu, in uvrstitev na seznam dovoljenih oblik.

4.4 Transkripcijska orodja

V zadnjih petih letih je tehnološki razvoj tudi na področje transkribiranja prinesel pomembne novosti. Razvita so bila programska orodja, ki olajšajo transkribiranje, kar je lahko njihov končni razvojni cilj ali pa samo eden izmed stranskih produktov razvoja. Transkripcijska orodja omogočajo neposredno sinhronizacijo zvoka in zapisa, kar postaja nujna in nespregledljiva zahteva novo nastajajočih govornih korpusov. Transkripcijska orodja, ki bodo predstavljena v nadaljevanju, so že bila uporabljena pri gradnji različnih govornih korpusov, vsa tri pa so bila uporabljena tudi pri transkribiranju Učnega korpusa govorjene slovenščine (UKGS). Vsi primeri transkripcij v nadaljevanju so iz UKGS.

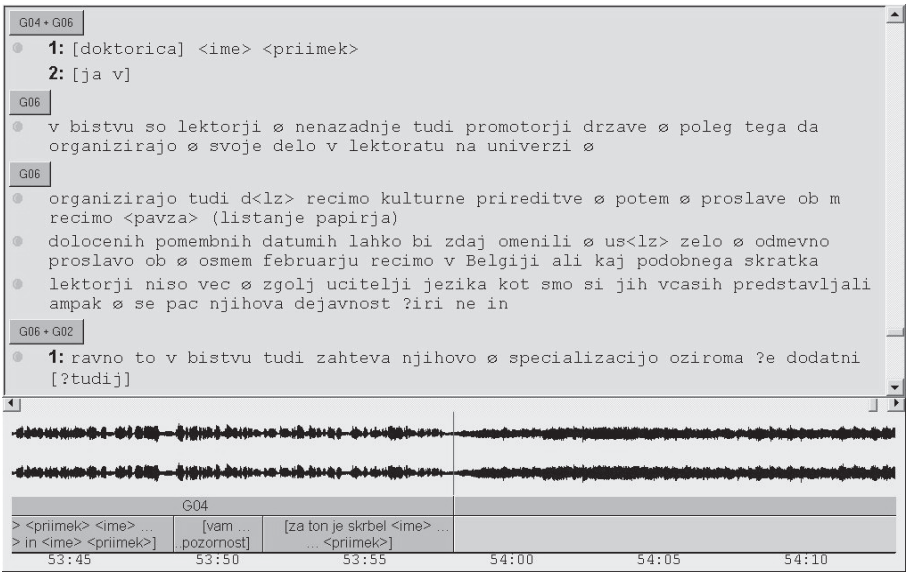
4.4.1 Transcriber

V okviru razvoja sistema za avtomatsko transkripcijo radijskih novic na Univerzi

[154] V nadaljevanju so navedene besede, ki se kljub drugačnemu izgovoru zapisujejo knjižno, npr. *jaz, brat, zdaj*, predlog *v* itd.

v Parizu (*Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur – LIMSI*), so potrebovali korpus zvočnega gradiva, na katerem bi se sistem učil. Za segmentiranje, označevanje in transkripcijo govora so izdelali programsko orodje *Transcriber* (avtor Claude Barras), prva verzija pa je bila registrirana l. 2001. Do danes je bil program že nekajkrat nadgrajen, vse njegove različice pa so prosto dostopne na internetu.[155]

Postopek transkribiranja v grafičnem okolju Transcriber se začne s segmentiranjem posnetka govora na krajše odseke. Segmenti govora ustrezajo t. i. izjavam, kot termin razumemo znotraj definicij TEI in EAGLES; meje med segmenti se postavljajo ob menjavah govorcev ali med premori. Ko imamo govorjeno besedilo segmentirano, se začne transkribiranje, pri čemer si lahko vsak segment poljubno velikokrat predvajamo. Segmentom v grafičnem okolju pripišemo govorce, hkrati pa med transkribiranjem zapisujemo tudi vse oznake vnaprej določene transkripcijske sheme.[156] Govorjeno besedilo transkribiramo neposredno v grafično okno, ki ga prikazuje spodnja slika:



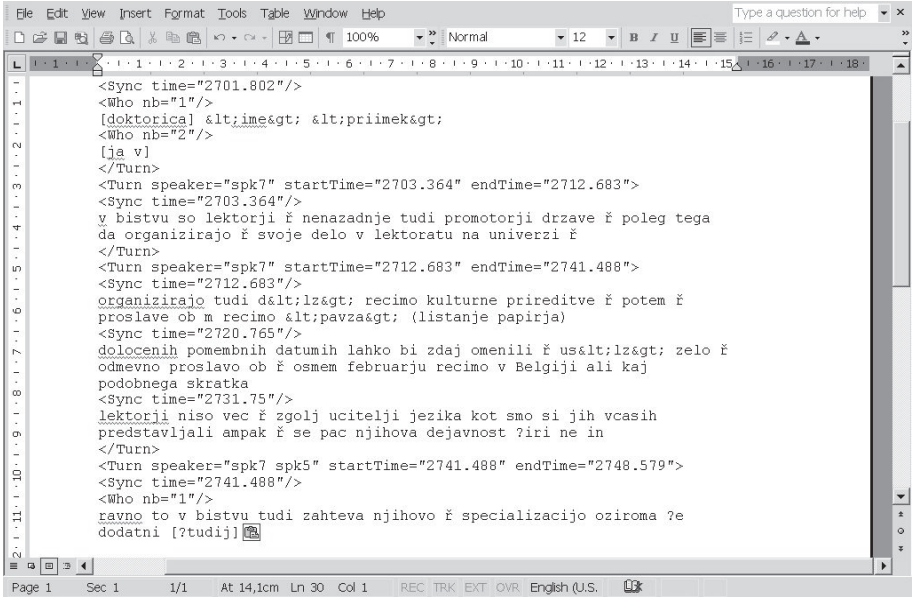
Slika 27: Programsko okno *Transcriber*, transkribiranje UKGS

Na spodnjem delu okna vidimo zvočni signal (v dveh vrstah, ker je bil posnetek narejen v stereo tehniki). Navpična črta potuje vodoravno po zvočnem signalu;

[155] http://trans.sourceforge.net/en/presentation.php
[156] Ta se med gradnjo korpusa lahko še deloma modificira.

kjer se signal prekine ali ošibi, je v govorjenju premor, in tam naredimo segmentno mejo. Kurzor lahko poljubno premikamo z računalniško miško in določamo, katere segmente želimo slišati: več izjav zaporedoma, samo eno izjavo ali manjši segment znotraj izjave, kar transkribiranje zelo olajša. Prvi trak pod zvočnim signalom prikazuje govorca (G04), trak pod njim pa njegove izjave, zapisane v ortografski transkripciji; vidimo lahko, da se meje med izjavami ujemajo z mesti v zvočnem signalu, kjer je amplituda nihanja majhna. Spodnja vrstica prikazuje časovni potek dogodkov.

Prednosti *Transkriberja* so naslednje: posamezne oznake lahko poljubno definiramo (npr. <nv>smeh</nv>, <ime>, <premor> itd.) in jih shranimo v seznam. Ko jih potrebujemo, jih izberemo iz seznama in nam jih ni treba vsakič znova ročno vpisovati; pri tem je treba spremeniti francoske transkripcijske standarde, ki so implementirani v program. Program avtomatsko zbira statistične podatke, npr. šteje zamenjave govorcev, izjave, pa tudi besede, poleg tega pa omogoča še statistične podatke o posameznih govorcih in njihovem deležu znotraj govorjenega besedila. Nadaljnja prednost *Transcriberja* (predvsem v primerjavi s transkripcijskim orodjem *Praat*) je, da nemoteno deluje tudi v primeru, ko so zvočni posnetki zelo dolgi (celo več ur), poleg tega pa lahko program bere zvočne posnetke v različnih formatih (*wav* in *mp3*).



Slika 28: Transkripcija, narejena v programu *Transcriber*, XML format[157]

[157] *Extensible Markup Language*, metajezik za označevanje dokumentov, ki je nadomestil SGML.

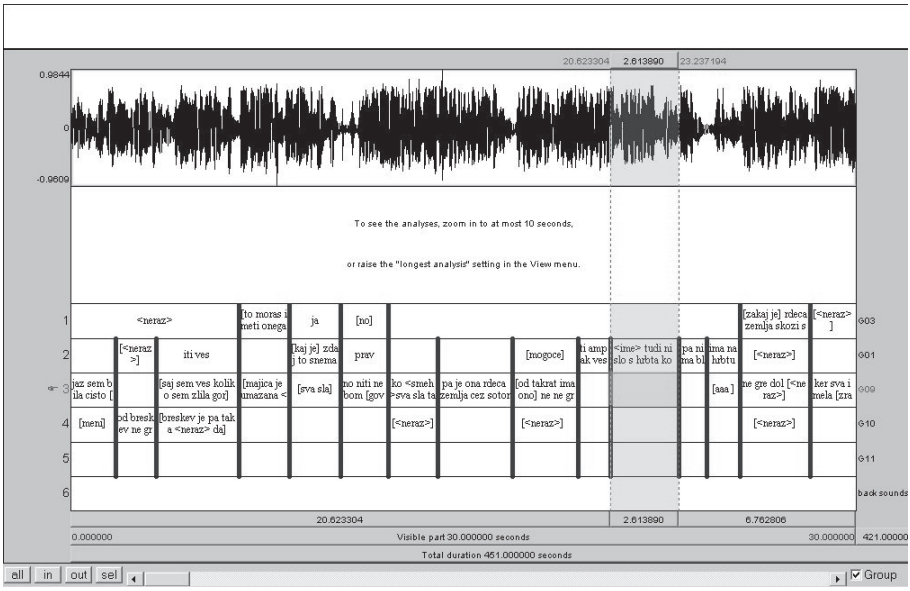
Velika pomanjkljivost *Transcriberja* je njegova omejena zmožnost zapisovanja prekrivnega govora. Realni čas govorjenja poteka linearno, tudi če govorita dva ali več govorcev hkrati, v transkripciji pa vsakemu govorcu ustreza samo ena vrstica. Če se v časovni enoti zgodita dva dogodka hkrati, je to v *Transcriberju* mogoče zapisati, saj program ponudi dve vzporedni vrstici v istem časovnem odseku. Če pa spregovori več oseb naenkrat, se tega ne da ustrezno zapisati. Transkripcijska praksa je pokazala, da se v primeru, ko sta v pogovoru udeležena več kot dva govorca, skoraj neobhodno zgodi, da spregovorijo vsi hkrati, tudi če gre za zelo formalne okoliščine in za javno nastopanje. To in pa dejstvo, da *Transcriber* ne omogoča fonetičnih analiz govora, sta lastnosti, ki lahko pretehtata pri odločitvi. Velja pa, da je za ortografsko transkribiranje radijskih novic (kjer načeloma ni prekrivnega govora več kot dveh oseb) *Transcriber* verjetno v tem trenutku najbolj primerno transkripcijsko orodje.[158]

4.4.2 Praat

Računalniški program *Praat* (nizoz. Govori!) je program za fonetične analize govora. Njegova avtorja sta Paul Boersma in David Weenink z Oddelka za fonetiko Univerze v Amsterdamu. Program je prosto dostopen na internetu in ga je mogoče prenesti na osebni računalnik;[159] avtorja in številni uporabniki po celem svetu program nenehno izboljšujejo in dopolnjujejo.

Praat se vedno pogosteje uporablja tudi za transkribiranje govorjenih besedil. Pri transkribiranju se programskega okna za fonetične analize ne uporablja, uporablja se samo okno s t. i. besedilno mrežo (Slika 29). Z vodoravnimi črtami v mreži so omejeni prostori, namenjeni posameznim govorcem, njihovo število pa je neomejeno. V zgornjem delu okna je grafično prikazano valovanje zvoka na posnetku. Resolucijo okna lahko spreminjamo, tako da nam celotno okno kaže npr. 10 sekund posnetka, 30 sekund ali nekaj minut; ustrezno se spreminja tudi slika valovanja. Spodnja slika prikazuje razpon posnetka v 30 sekundah. Iz slike valovanja zvoka je mogoče razbrati, kje so v govorjenju premori, kar transkriptorju zelo pomaga pri določanju mej med izjavami.

[158] *Transcriber* je uporabljala tudi D. Verdonik za transkribiranje govorne baze TURDIS, ki jo sestavljajo telefonski klici na turistične informacije.
[159] http://www.fon.hum.uva.nl/praat/



Slika 29: Programsko okno *Praat*, transkribiranje UKGS

Na sliki meje med izjavami vidimo kot navpične črte v spodnjem delu okna. Tako nastane mreža za besedilo, ki omogoča, da v vsako okence vpišemo pripadajočo izjavo posameznega govorca. Okenca v vodoravni vrsti predstavljajo izjave posameznega govorca, okenca v navpični vrsti pa izjave različnih govorcev, izrečene v istem časovnem obdobju. To seveda pomeni, da je *Praat* izredno primeren program za transkribiranje posnetkov, kjer sodeluje več govorcev in jih pogosto govori več naenkrat, kar se najpogosteje dogaja pri besedilih spontanega govora velike stopnje sproščenosti z več kot dvema govorcema. Po mojih izkušnjah je 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.

Ko je transkribiranje posnetka končano (lahko pa že prej), prenesemo transkribirano besedilo v poljubni urejevalnik besedil. Tu lahko transkripcijo še naprej spreminjamo, vendar spremembe v programu *Praat* ne bodo vidne. Konkretno v primeru slovenščine lahko npr. vnašamo sičnike in šumnike, ki jih programsko okolje *Praat* ne pozna.

Dodati je treba, da *Praat* v nasprotju s *Transcriberjem* ponuja tudi možnost akustičnih analiz transkribiranega besedila in je po ocenah fonetikov izvrstno orodje za fonetične analize. To pomeni, da je besedilo, transkribirano s Praatom, že

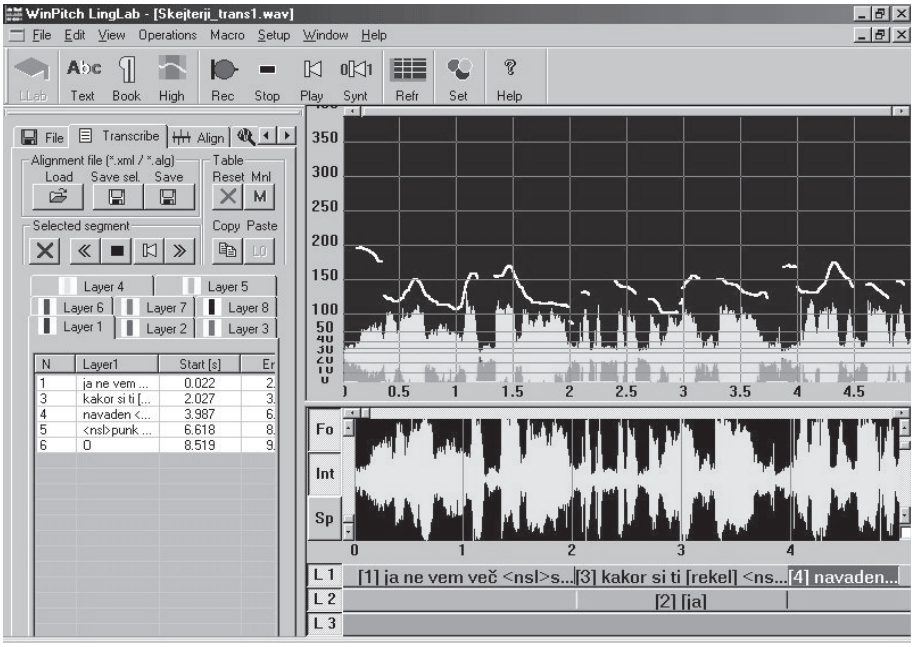
pripravljeno za nadaljnje analize, kar bi bilo pri morebitni gradnji govornega korpusa pomembno, če bi vsaj del gradiva želeli transkribirati tudi fonetično in prozodično.

4.4.3 WinPitch

Tudi program *WinPitch* lahko služi kot učinkovito orodje za transkribiranje govorjenih besedil. Program obstaja v treh različicah. *WinPitch Classic* je namenjen fonetičnim in fonološkim raziskavam, uporablja pa se ga tudi kot pripomoček pri govorni terapiji. *WinPitch TL* je namenjen učenju tujih jezikov; uporabljajo ga tako učitelji kot učeči se, in sicer pri učenju izgovorjave v tujem jeziku in govorjenja nasploh. Zaenkrat je najbolj učinkovit program za učenje in evalvacijo govora pri učenju jezika na daljavo, saj omogoča snemanje, predvajanje in analizo govora. Program je enostaven za uporabo, datoteke pa so zlahka prenosljive po elektronski pošti. *WinPitchPro* pa je verzija programa, ki z dodanimi tekstovnimi okni omogoča transkribiranje zvočnih posnetkov.

Kot transkripcijsko orodje lahko *Winpitch Corpus Pro* deluje na dva načina: v primeru, da je treba govorjeno besedilo transkribirati, uporabnik izbira segmente zvoka v oknu za analizo (na Sliki 30 desno zgoraj) ali pa v t. i. navigacijskem območju in jih ob predvajanju sproti zapisuje; na ta način avtomatsko gradi bazo podatkov (transkripcije), ki je kasneje lahko shranjena v XML formatu ali v Excelu. Druga možnost je, da besedilo že obstaja v zapisani obliki in ga želi uporabnik samo sinhronizirati z zvokom; v tem primeru si uporabnik odpre posebno okno z besedilom, nato pa si predvaja zvok in hkrati označuje (s klikanjem) besedilo, ki ustreza posameznim zvočnim segmentom. Slednja možnost je zelo uporabna v primeru, ko se nadgrajuje starejše govorne korpuse, kjer so bili zvočni posnetki narejeni še z analognimi napravami; analogne zvočne posnetke se najprej digitalizira, nato pa se jih sinhronizira z že obstoječimi transkripcijami.

Največja prednost programa *WinPitch* pred ostalimi transkripcijskimi programi je, da si uporabnik besedilo lahko predvaja z upočasnjeno hitrostjo, kar olajša transkribiranje. Zmanjšanje hitrosti govora na okoli 70 odstotkov naravne hitrosti še dopušča nemoteno transkribiranje, pri nižjih hitrostih pa je, vsaj po mojih izkušnjah, zvok že preveč deformiran.



Slika 30: Programsko okno *WinPitch*, transkribiranje UKGS

Program *WinPitch* je prosto dostopen na internetu[160] in ga je mogoče preizkusiti v vseh različicah, vendar brezplačna licenca za uporabo programa poteče po enaintridesetih dneh.

* * *

Predstavljena so bila tri transkripcijska orodja, ki se uporabljajo pri transkribiranju govornih korpusov oz. ki so dejansko že bila uporabljana za gradnjo velikih korpusov v zadnjih petih letih – *Transcriber* npr. pri gradnji španskega najstniškega korpusa (COLA[161]), *Praat* pri gradnji korpusa bergenskega najstniškega govora, *WinPitch* pa pri gradnji govornega korpusa C-ORAL-ROM. Za različne oblike analiz in obdelave govora obstajajo še druga programska orodja, npr. *Voice Walker 2.0*, ki je bil uporabljen pri gradnji ameriškega govornega korpusa Santa Barbara, in *Sound-Scriber*, ki je bil uporabljen pri gradnji korpusa MICASE; z nekaterimi orodji se je mogoče seznaniti na spletni strani *Speech analysis and transcription software*).[162] Pred odločitvijo o izbiri programskega orodja si je treba najprej odgovoriti na vprašanje, kakšne podatke imamo na razpolago in kaj na-

[160] http://www.winpitch.com/
[161] http://colam.org/publikasjoner/COLA-korpus-publ.html
[162] http://liceu.uab.es/~joaquim/phonetics/fon_anal_acus/herram_anal_acus.html

meravamo z njimi početi. Za ortografsko transkribiranje govora se trenutno zdita najprimernejša programa *Transcriber* in *Praat*, prvi predvsem za transkribiranje zvočnih posnetkov, kjer sta hkrati aktivna največ dva udeleženca, drugi pa za transkribiranje spontanega govora, kjer je udeleženih več govorcev. Če nameravamo korpus nameniti tudi za fonetične analize govora oz. ga deloma fonetično transkribirati, je najbolje transkribirati v programu *Praat*, ki omogoča fonetično analizo govora.

4.5 ZAKLJUČEK

Pri izdelavi načel za transkribiranje govorjene slovenščine in označevanje govorjenih besedil v korpusu sledimo priporočilom TEI in EAGLES, kolikor je to v primeru slovenščine mogoče in smiselno. Oznake korpusa se lahko v veliki meri prevzemajo, treba se je predvsem odločiti za to, katere oznake bodo vključene v korpus govorjene slovenščine, pač glede na to, kakšno uporabo korpusa predvidevamo. Bolj samostojno in neodvisno se morajo sestavljavci korpusa odločati pri določitvi načel za transkribiranje govorjenih besedil. Tudi tu je mogoče upoštevati nekatere znane transkripcijske standarde; načeloma se za transkribiranje za (referenčne) govorne korpuse uporablja ortografska transkripcija, pri čemer ostaja odprto temeljno vprašanje prilagoditve zapisa govorjenemu jeziku. V naslednjem poglavju bodo predstavljene nekatere možnosti transkribiranja govorjene slovenščine.

5 Predlog priporočil za transkribiranje besedil v govorni korpus

5.1 UVOD

Tako kot pri vseh pomembnih odločitvah tudi pri izbiri transkripcijskega standarda namen korpusa pogojuje izbiro. Kot smo videli, so govorni korpusi, ki so del referenčnih korpusov, najpogosteje transkribirani po načelih ortografske transkripcije (*EAGLES preliminary recommendations on Spoken Texts, BNC, The Bank of English, Nizozemski govorni korpus itd.)*. Tak korpus omogoča študij govorjenega jezika predvsem na ravni leksikografije in skladnje, omogoča analizo diskurza in sociolingvistične raziskave, primerjavo med pisnim in govorjenim jezikom, predstavlja jezikovni vir za razvoj orodij za sintezo in analizo govora ter za pripravo gradiv za učenje jezika kot tujega jezika, v širšem smislu pa so uporabniki govornega korpusa lahko tudi defektologi, pedagogi, sociologi, komunikologi in drugi. Fonetičnih študij tak korpus ne omogoča, vendar v povezavi z zvočnimi posnetki dopušča nadgradnjo – dopolnitev s fonetičnimi in prozodičnimi oznakami, običajno na manjšem delu korpusa.

Predlog priporočil za transkribiranje in označevanje govornega korpusa za slovenščino temelji na priporočilih TEI in EAGLES ter na izkušnjah pri gradnji drugih govornih korpusov, zapis pa je prilagojen slovenskemu jeziku in jezikovni situaciji. Izbira oznak seveda že pomeni hipotetično ugibanje o tem, kakšne analize bodo uporabniki želeli delati na korpusu. Prenatančno označevanje je lahko nesmiselno: povzroča zmedo in dodatno delo pri transkribiranju, prinaša pa tveganje, da označujemo stvari, ki jih nikoli nihče ne bo potreboval. Sinclair je tudi v zvezi s tem duhovito pripomnil, da mu v 30 letih korpusa ni niti enkrat uspelo označiti tako, kot bi želeli uporabniki: »uporabnik vedno hoče, da bi bil korpus označen nekoliko drugače, naslednji uporabnik pa spet nekoliko drugače …« (Sinclair 1995, 102).

V nadaljevanju je izdelan predlog za segmentiranje, zapisovanje in označevanje govora pri gradnji referenčnega govornega korpusa za slovenščino. Vsi primeri, ki dopolnjujejo teoretične predpostavke, so iz Učnega korpusa govorjene slovenščine.

5.2 SEGMENTIRANJE GOVORA

Večina pisnih besedil je strukturiranih po standardiziranih načelih in jih lahko razdelimo na manjše enote, kot so poglavja, odstavki, povedi in stavki. Naslovi, ločila in različna tipografska orodja še nadalje strukturirajo pisna besedila in jim dajejo prepoznavno obliko. Popolnoma drugače pa je pri govorjenih besedilih: tu gre za bolj ali manj strnjen dogodek v času, ki ga je težko razdeliti na manjše

enote. Kljub temu ga je treba za nadaljnje študije in analize na nek način segmentirati. Jezikoslovci se raje odločajo za segmentacijo govorjenih besedil na podlagi:

- **prozodičnih lastnosti** (tonske enote v korpusu London–Lund, intonacijske enote v korpusu Santa Barbara),
- **premorov v govoru in zamenjav govorcev** (priporočila TEI, EAGLES, COLT, Švedski govorni korpus).

Za definiranje izjave kot osnovne enote govora pri gradnji govornega korpusa je tudi v zvezi s slovenščino najbolje slediti priporočilom EAGLES in TEI: izjava je omejena s premorom in/ali menjavo govorcev.[163] Kadar gre za daljše monološke pasaže, kjer govorec ne dela premorov, poskušamo postaviti mejo med izjavami pri vdihu. Ta definicija izjave je bila v okviru korpusnih raziskav privzeta že pri Verdonik (2006, 50): »Praviloma gre za enoto govora, izgovorjeno med premoroma, običajno tudi med vdihom govorca.« Tako definirana izjava je predvsem prozodična enota govora, kljub temu pa je pogosto tudi »skladenjsko-semantično zaokrožena enota« (Verdonik 2006, 50), čeprav to ni pogoj. Vitez in Zwitter Vitez (2004, 8) ob analizi spontanega govora podobno ugotavljata, da je »opredelitev analiziranih govornih enot v osnovi usklajena z »negramatikalnimi« principi spontanega izrekanja, ki ne daje rezultatov v obliki idealnih stavkov ali stavčnih členov.« V nadaljevanju avtorja vzameta za (najmanjšo) enoto analize spontanega govora t. i. *govorjeni odstavek*, ki je določen z intonacijo,[164] poimenovanje pa je povzeto po analogiji z definicijo pojma odstavka v pisanju. Tu so osnovne enote govora torej prozodične enote, podobno kot npr. v korpusu London-Lund (4.3.1, *Prozodična transkripcija korpusa London-Lund*); prozodične enote so določene s pomočjo akustičnih meritev računalnika, kar je dandanes edini sprejemljivi način akustične analize. Govorni korpus bi predstavljal pravo izhodišče za širšo empirično prozodično analizo govora, ustrezno transkripcijo in označevanje pa je smiselno narediti na manjšem delu korpusa.

Izjava, kot sem jo definirala zgoraj (omejena s premorom in/ali menjavo govorcev), ni nujno skladenjsko-semantična enota, njen sporazumevalni namen pa nas pri gradnji korpusa niti ne zanima, zato je ni mogoče zamenjati s pojmom *izrek* iz pragmatike, ki je definiran kot »poved s komunikacijsko funkcijo« (Kranjc 1996/97, 307) ali kot »rezultat izrekanja oz. uresničitev govornega dejanja« (Bešter 1994, 45).[165]

[163] Podobno Kranjc (1996/97, 309) definira *vlogo* kot »strukturno enoto konverzacije; vloga je vse, kar govorec reče, preden začne govoriti drugi govorec.«

[164] Intonacijo avtorja razumeta kot celostni koncept prozodije, ki vključuje sočasno delovanje sprememb višine osnovnega tona, jakosti, hitrosti govora in premorov v diskurzu (Vitez in Zwitter Vitez 2004, 8).

[165] Vitez in Zwitter Vitez (2004, 6) na podoben način definirata *izrekanje* – kot »/…/ jezikovno (govorno) dejanje, ki v določenem jezikovnem in zunajjezikovnem kontekstu ustvarja govorne rezultate /…/.«

```
G17: ə če bi Slovenci volili {pavza} Kerryja ali Bu-
     sha jih je petindevetdeset procentov Slovencev
     je za Kerryja pet procentov pa [za Busha] {pav-
     za} [{neraz}] [ja]
G16: [ja pa] saj večina saj [vse] svetovne [države]
G17: večina [mislim evropske {neraz}]
G16: [cel svet je] cel s- cel svet je za Kerryja razen
     Amerike ma ja saj bomo videli saj ne bo dosti
     boljše a veš isti kurac bo po moje zdaj bo pri
     nas vse drugače ko bo [Janša]
G17: [{nv} smeh {/nv}] (R06)[166]
```

Slika 31: Segmentiranje na izjave v UKGS

Zgoraj je predstavljen primer segmentiranja govora v UKGS. Iz primerov je razvidno, da imajo nekatere izjave skladenjsko strukturo, ki na ravni pisnega besedila ustreza stavku, in razvidno semantično strukturo, druge pa nimajo ne tega ne onega, sporazumevanje pa kljub temu nemoteno poteka. Meje med izjavami bi bilo mogoče potegniti tudi drugače, pa to ne bi vplivalo na gradnjo korpusa. Transkripcijska orodja pri postavljanju mej med izjavami zelo pomagajo, saj na spektrogramu kažejo potek zvoka, iz katerega je mogoče razbrati, kje oz. kdaj bo najboljši trenutek za določitev meje med izjavami; orodja s svojim grafičnim okoljem (oknom, zaslonom) celo do neke mere sugerirajo dolžino izjav. *Transcriber* dovoljuje zapis daljših izjav, ki so dobro berljive tudi v transkripciji, kar je razumljivo, saj je bil izdelan za transkribiranje radijskih novic (po tipu monološka besedila z daljšimi izjavami; Slika 27). Po drugi strani pa orodje *Praat*, ki je primernejše za transkribiranje spontanih dialoških oz. multiloških besedil, transkriptorja sili v segmentiranje krajših izjav, saj jih zapisuje eno pod drugo na relativno skromno odmerjenem prostoru (Slika 29).

Postavljanje mej med izjavami je bolj zapleteno pri prekrivnem govoru, kjer mora transkriptor s pomočjo transkripcijskega orodja najti najboljšo rešitev, in sicer tako, da po možnosti ne potegne meje znotraj izjav, ki se po smislu zdijo celota, nikakor pa ne sme potegniti meje sredi besede katerega izmed govorcev.

[166] Oznake se nanašajo na transkribiranje učnega korpusa govorjene slovenščine; R06 pomeni, da gre za posnetek št. 7, G16 oz. G17 pa sta oznaki govorcev.

5.3 ZAPISOVANJE GOVORA

Tudi pri ortografski transkripciji govora se zastavlja vprašanje, v kolikšni meri se pri zapisu prilagajati pisnemu jeziku in kdaj se tudi v zapisu prilagoditi govorjenemu jeziku. Gre za odločitve, ki so v samem jedru načel za transkribiranje posameznega jezika in ki morajo biti obravnavane za vsak jezik posebej. V nadaljevanju je predstavljenih nekaj možnosti za zapisovanje govora v slovenščini.

5.3.1 Ortografska transkripcija

Če govor zapisujemo v ortografski transkripciji, besede zapisujemo v skladu z ustaljenim knjižnim zapisom, to je v skladu s kodificiranim zapisom (knjižnega) jezika; spodaj je primer takega zapisa iz UKGS:

```
G19: sem pa danes spila že kakšne tri kofete
G20: jaz sem ga dopoldne enega sem kuhal potem sem
     pa nič potem pa ko je enkrat ko imaš tukaj
     prištimano ko je enkrat potem {neraz} ne kaj
     pa križanka pa kavica pa cigaret pa vse sorte
     (R07)[167]
```

Primerjava s fonetično transkripcijo (poglavje 5.3.3) nam bo pokazala, da so pri popolni prilagoditvi pisni normi številne informacije ne samo izgubljene, ampak celo potvorjene (beseda pol je npr. zapisana kot potem, kle kot tukaj, k kot ko itd.). Kljub neposredni dostopnosti zvočnih posnetkov bi iskanje po korpusu s takšnim zapisom vodilo k neresničnim oz. neavtentičnim podatkom o jeziku. Tovrstni zapis je zato neustrezen, v zapisu je potrebna večja prilagoditev govorjenemu jeziku.

5.3.1.1 Ortografska transkripcija po švedskem modelu

Pri zapisu v ortografski transkripciji po švedskem modelu (prim. poglavje 4.3.5, *Göteborška modificirana ortografska transkripcija*) so v govoru reducirani glasovi v zapisu ohranjeni v zavitih oklepajih; primer iz UKGS bi bil zapisan, kot je prikazano spodaj:

[167] V oklepaju je identifikacijska oznaka posnetka.

```
G19: sem pa dan{e}s spila že ene tri kofete
G20: jaz sem ga dopoldne en{e}ga sem kuhal po{l} sem
     pa nič po{l} pa k{o} je enk{r}at k{o} {i}maš tle
     pr{i}štiman k{o} je enkat pol {neraz} ne ka{j}
     pa križanka pa kavica pa cigaret pa vse sorte[168]
     (R07)
```

Tovrstni zapis je bolje prilagojen govorjenemu jeziku, vendar za slovenščino ne najustreznejši, saj poleg popolne redukcije glasov poznamo v govoru tudi delne redukcije glasov (npr. prehajanje glasov v polglasnik) in prehajanje glasov v druge glasove, tako da samo z zapisovanjem popolne redukcije pri približevanju govoru ostanemo nekako na četrtini poti.

5.3.2 Fonemska transkripcija

V fonemski transkripciji zapisujemo vse slišane foneme:

```
G19: səm pa dôns spila že êne tri koféte
G20: jəs səm ga dopóudne ênga səm kuhou po səm pa
     nəč po pa kə je ênkat kə maš tlê prštiman kə je
     ênkat pol {neraz} ne ka pa križanka pa kavica pa
     cigarét pa vse sórte(R07)
```

Zapis v fonemski transkripciji je v mnogo večji meri prilagojen govorjenemu jeziku, kar je gotovo prednost takega zapisa. Slabša prilagoditev pisni normi pa otežuje iskanje po (referenčnem) korpusu in zmanjšuje robustnost korpusa. Rešitev, ki je že bila ponujena, je vkodiran dodatni zapis besed (lahko kot lema, kadar obstaja), ki ustreza ustaljenemu knjižnemu zapisu (kadar obstaja), kar bi razširilo možnosti iskanja po korpusu. Treba pa se je zavedati, da je fonetično transkribiranje zahtevno in zamudno opravilo, ki ga lahko opravijo samo usposobljeni strokovnjaki.

[168] Pri čemer transkriptor pogosto ugiba, kaj je tisto, česar se ne sliši; ali naj torej slišani "k" zapiše kot k{o}, k{i}, k{ar} itd.

5.3.3 Fonetična transkripcija

Za zapis govora na fonetični ravni lahko uporabimo standardni računalniški fonetični zapis simbolov za slovenski knjižni jezik MRPA (prim. poglavje 4.3.6, *Transkripcijske konvencije na Slovenskem*). Kot je bilo že omenjeno, zapis v celoti upošteva tradicionalno fonetično abecedo slovenskega knjižnega jezika z 8 samoglasniki, ki so dolgi ali kratki, naglašeni ali nenaglašeni, 6 zvočniki (z alofoni) in 16 nezvočniki (z alofoni). V nadaljevanju je predstavljen fonetični prepis govora iz UKGS:[169]

```
G19: s@m pa "dO:ns "spi:la Ze "E:ne "tri: ko"fe:te
G20: j@s s@m ga do"po:Udne "E:nga s@m "ku:hoU po s@m
     pa "n@tS po pa k@ je "E:Nkat k@ "ma:S "t_lE:
     p@r"Sti:man k@ je "E:Nkat  "pO:l'{neraz} ne ka
     pa "kri:ZaNka pa "ka:vica pa tsiga"re:t pa WsE
     "so:rte (R07)
```

Tu bi bilo mogoče polemizirati že z izhodiščno (Toporišičevo) razdelitvijo slovenskih glasov (pri jakostnem naglaševanju je po mojem mogoče problematizirati npr. razlikovanje med kratkimi in dolgimi glasovi), vendar to presega namen te razprave. Po drugi strani je fonetična transkripcija tudi preveč zamudna za transkribiranje velike količine govorjenih besedil, za namene referenčnega korpusa tudi nepotrebna. Mogoče in koristno bi jo bilo izdelati na manjšem delu korpusa, v sodelovanju s fonetiki, ki bi ob tem tudi ponovno premislili nekatere temeljne lastnosti slovenskih glasov.

5.4 RABA LOČIL

Ločila so po definiciji »/…/ nečrkovna pisna (grafična) znamenja, ki /.../ nam zaznamujejo tonski potek, premore, vrste stavkov, povedi ipd. (skladenjska raba) ter okrajšave besed in besedila, vrstilnost številnikov, kratnost prislovov, zapisanih s številkami ipd. (neskladenjska raba)« (SP 2001, 226). Ločila so značilnost pisnega in ne govorjenega jezika, njihova naloga pa je, vsaj v okviru skladenjske rabe, segmentiranje pisnih besedil. Pri transkribiranju govorjenega besedila ugotavljamo, da govorci z intonacijo oz. sploh s prozodičnimi sredstvi le redko oblikujejo skladenjske enote, kot bi jih pričakovali v pisnem bese-

[169] Z razlago posameznih fonetičnih simbolov glej Zemljak idr. 2002.

dilu. Vnašanje ločil v transkribirano besedilo najpogosteje poteka na podlagi intuicije, individualne interpretacije, individualnih pričakovanj in navad transkriptorja. Res pa je, da transkriptorja pri transkribiranju »srbijo prsti«, da bi vnašal ločila, ker je tako navajen pisati že vse življenje.[170] Številni sestavljavci korpusov se kljub temu odločajo za transkribiranje brez ločil, nenazadnje tudi priporočila EAGLES sugerirajo tak zapis. Obstaja pa tudi dokaj pogosto uporabljena metoda delnega vnašanja ločil (npr. samo končna ločila ali samo pika in vprašaj).

V nadaljevanju je navedenih nekaj primerov iz UKGS, ki ponazarjajo zapis besedila brez ločil, z ločili in z delno uporabo ločil.

5.4.1 Ortografska transkripcija z ločili

```
G19: Sem pa danes spila že ene tri kofete.
G20: Jaz sem ga dopoldne enega, sem kuhal, pol sem pa
     ... {pavza}
G20: Nič, pol pa, ko je enkrat – ko imaš tle prištimano,
     ko je enkrat, pol – {neraz} ne, kaj pa, križanka
     pa kavica pa cigaret pa vse sorte. (R07)
G16: Kar fajn: dala sva – za vsako sva dala petnajst
     jurjev, (telefon zazvoni), a veš.
G17: Ja? Ja, kaj je? A? Ja. (R06)
```

Transkribiranje z vsemi ločili, značilnimi za pisni jezik, ja neke vrste nasilje nad govorom, saj ga poskušamo umestiti v formo, ki ni njegova naravna pojavnost. Številni premori in oklevanja v govoru bi nenehno narekovali rabo pomišljajev, tropičij in drugih oznak prekinjene skladnje. Pri tem so seveda možne različne interpretacije in dopustna različna raba ločil, kar za konsistentnost zapisa ni dobro. Tudi izkušnje pri gradnji drugih korpusov kažejo v smer izogibanja tovrstnemu zapisu.

[170] Omenila sem že Sinclairovo pojasnilo v zvezi z gradnjo korpusa Cobuild, kjer so za transkribiranje najeli osebe z administrativno izobrazbo: "Transkriptorji so ljudje in dati jim je treba nalogo, ki jo bodo lahko opravili; če hočejo postavljati pike, naj jih postavljajo, da bodo le postavljene na pravem mestu" (Sinclair 1995, 107).

5.4.2 Ortografska transkripcija s končnimi ločili

Transkribiranje z uporabo končnih ločil je dokaj pogosta praksa pri gradnji govornih korpusov (*The Bank of English*, Korpus ameriške angleščine). In to kljub temu, da je »meja med enostavčno povedjo in večstavčno (priredno) povedjo v spontano govorjenem besedilu v primerjavi s pisnim (knjižnim) jezikom težje določljiva« (Smolej 2006, 15).

```
G19: Sem pa danes spila že ene tri kofete.
G20: Jaz sem ga dopoldne enega sem kuhal pol sem pa
     {pavza}
     Nič pol pa ko je enkrat ko imaš tle prištimano
     ko je enkrat pol {neraz} ne kaj pa križanka pa
     kavica pa cigaret pa vse sorte. (R07)
G16: Kar fajn dala sva za vsako sva dala petnajst ju-
     rjev (telefon zazvoni) a veš.
G17: Ja? Ja kaj je? A? Ja. (R06)
```

Tudi dodajanje samo končnih ločil je neke vrste interpretacija besedila. Res pa je, da ločila v nekaterih primerih zelo povečajo razumljivost zapisa; tako npr. v zadnji vrstici primera s posnetka R06, kjer si samo iz zapisanega lažje predstavljamo, kaj se je v resnici dogajalo, če so dodana še ločila, kot če jih ni. Tukaj z ločili pravzaprav označimo prozodične lastnosti govora. Zato bi bilo morda bolj upravičeno razmišljati o prozodičnem zapisu govora (vsaj na delu korpusa), kot je to npr. pri Vitez in Zwitter Vitez (2004), kot pa o transkribiranju s končnimi ločili.

5.4.3 Ortografska transkripcija brez ločil

Zapisovanje govora brez ločil je najbolj adekvaten zapis govora, pri čemer ne izgubimo nobene informacije, saj teh informacij v govoru ni; tak način je bil uporabljen npr. za transkribiranje BNC in Nizozemskega govornega korpusa.

```
G19: sem pa danes spila že ene tri kofete
G20: jaz sem ga dopoldne enega sem kuhal pol sem pa
     nič pol pa ko je enkrat ko imaš tle prištimano
     ko je enkrat pol
     {neraz} ne kaj pa križanka pa kavica pa cigaret
     pa vse sorte
     (R07)
G16: kar fajn dala sva za vsako sva dala petnajst ju-
     rjev
     (telefon zazvoni) a veš
G17: ja ja kaj je a ja (R06)
```

Za UKGS je bila uporabljena ortografska transkripcija brez ločil in enak koncept predlagam tudi za morebitno gradnjo večjega govornega korpusa, in sicer zaradi zgoraj naštetih razlogov. Obenem predlagam, da se del korpusa govorjenih besedil tudi prozodično označi (poudarki, intonacija, tonska višina, hitrost), in to na podlagi meritev, ki jih omogoča transkripcijsko orodje Praat.

5.5 RABA VELIKIH ZAČETNIC

Tudi velike začetnice so (pravo)pisna znamenja, ki jih govor ne pozna, njihova uporaba pa je povezana z rabo ločil. Velikih začetnic na začetku povedi v transkripcijah, kjer ni ločil, ni. Drugače je pri lastnih imenih, osebnih (tistih, ki niso odstranjena iz transkripcije), geografskih in stvarnih: tu se velika začetnica ohranja (prim. npr. Sliko 32).

5.6 Končni predlog priporočil za transkribiranje govorjene slovenščine

Za zapis govorjene slovenščine v referenčnem govornem korpusu, ki predvideva zapis velikih količin govora, kot osnovno načelo predlagam razširjeno ortografsko transkripcijo brez ločil. Pojem razširjene ortografske transkripcije se nanaša na prilagoditev zapisa nekaterim najbolj izrazitim značilnostim govora, ki jih lahko določimo na podlagi dosedanjih raziskav govorjenega jezika in izkušenj pri transkribiranju govora. Besedi, ki bi jo zapisali prilagojeno govoru, bi lahko v obliki XML kode pripisali tudi knjižni zapis (če ta obstaja),

ker bi s tem povečali možnosti za poizvedovanje po korpusu. Sicer pa bi te besede uvrščali na »seznam prilagojenih zapisov«. Na podlagi dveh oz. treh na korpusu temelječih raziskav govorjenega jezika,[171] ki izhajajo iz dveh različnih regij (Ljubljanske in Štajerske), sklepam, da bi bilo na seznamu prilagojenih zapisov veliko skupnih slovenskih govornih značilnosti, nekaj pa bi bilo tudi regionalnih (npr. *toti*).

Na tem mestu pravzaprav lahko predlagam samo osnutek seznama za osrednjo (Ljubljansko) regijo; za preostale regije bi bilo treba tudi s pomočjo jezikoslovnih strokovnjakov za posamezne regije narediti manjše področne raziskave in določiti osnovni nabor dovoljenih zapisov; vsekakor se tudi ta seznam dopolnjuje in uskljajuje med dejansko gradnjo korpusa. Osnutek predloga pisnih prilagoditev govoru je naslednji:

- nedoločnik se zapisuje brez končnega i-ja, kadar je tako izgovorjen
- deležnik na –l se v dvojinski in množinski obliki zapisuje tako, kot je izgovorjen (*smo delal*, *bova delale*),
- vezniki *ki*, *ko*, *ker* ali oziralni zaimek *ki* se zapišejo kot *k*, kadar so tako izgovorjeni
- določne pridevnike v imenovalniku in tožilniku se zapisuje brez končnega i-ja, kadar so tako izgovorjeni
- zapisujemo *pol* (v pomenu potem)
- zapisujemo členek *una,-a,-o* v vseh spolih, sklonih in številih (*un Michael*)
- zapisujemo členek *en,-a,-o* v vseh spolih, sklonih in številih (*ene tri ure*)
- govorni določni člen *ta* zapisujemo (*ta rdeč*)
- kadar je beseda spremenjena do take mere, da je ne moremo zapisati v pričakovani knjižni obliki (če sploh obstaja), zapišemo slišano besedo (*kle, klele, tle, tlele, čmo* itd.).

Redukcij sredi besede in spremenjenih glasov načeloma ne bi zapisovali, in sicer z namenom, da ohranimo čim večjo konsistentnost zapisa, za katerega predvidevamo, da ga bo kreiralo med deset in dvajset oseb. Predlagani način obeta, da bi zapis ohranil nekatere najbolj tipične (oblikoskladenjske in leksikalne) značilnosti govorjenega jezika, hkrati pa različnim zapisovalcem predpisoval dovolj jasna navodila za zapisovanje (in s tem zagotavljal konsistentnost zapisa). Predlagam tudi, da bi čim večji del referenčnega govornega korpusa (kolikor bi to dopuščale finančne in časovne omejitve) nadgradili s prozodično in morda tudi s fonetično transkripcijo; to je mogoče storiti v fazi gradnje referenčnega korpusa ali kot kasnejšo nadgradnjo.

[171] Verdonik 2006, Smolej 2006a in Zemljarič Miklavčič 2007.

Besedilo, ki ga pripravljamo za vključitev v korpus, dodatno označujemo zato, da bi z oznakami ujeli v zapis čim več značilnosti govorjenega jezika, pa tudi zato, da bi zavarovali osebne podatke govorcev in v govoru omenjenih oseb ter izločili nerazumljive fragmente govora. Osnova za nabor oznak so priporočila EAGLES, oznake pa so bile preizkušene pri transkribiranju Učnega korpusa govorjene slovenščine. Nekatere oznake so se že med uporabo pokazale kot problematične, zato jih lahko sicer najdemo v UKGS, vendar jih za uporabo pri morebitni gradnji referenčnega govornega korpusa odsvetujem. O podrobnostih govorim v nadaljevanju, vsi primeri so iz UKGS.

6 Predlog priporočil za označevanje besedil v govornem korpusu

6.1 OSEBNA IMENA IN DRUGI OSEBNI PODATKI

Vsi podatki, na podlagi katerih bi bilo mogoče iz besedila identificirati govorce ali osebe, o katerih se govori, morajo biti zaradi varovanja osebnih podatkov pri transkribiranju nadomeščeni z nevtralnimi oznakami, v zvočnem posnetku pa »prekriti« z nevtralnim zvokom; gre predvsem za imena, priimke, naslove in telefonske številke: <ime>, <priimek>, <drugo ime>,[172] <naslov>, <telefonska številka>. Osebna imena javnih oseb bi načeloma lahko ostala v besedilu (kot npr. v BNC), vendar je pri tem lahko zelo problematično ali pa vsaj ne enoumno določljivo, v katerih primerih gre za »javne osebe«. V tem smislu so za načrtovalce govornega korpusa pomembna nekatera določila Zakona o varstvu osebnih podatkov:[173]

> 1. člen
>
> Z varstvom osebnih podatkov se preprečujejo nezakoniti in neupravičeni posegi v zasebnost posameznika pri obdelavi osebnih podatkov, varovanju zbirk osebnih podatkov in uporabi le-teh. /.../
>
> V tem zakonu uporabljeni izrazi imajo naslednji pomen:
>
> 1. Osebni podatek - je katerikoli podatek, ki se nanaša na posameznika, ne glede na obliko, v kateri je izražen.
>
> 2. Posameznik - je določena ali določljiva fizična oseba, na katero se nanaša osebni podatek; fizična oseba je določljiva, če se jo lahko neposredno ali posredno identificira, predvsem s sklicevanjem na identifikacijsko številko ali na enega ali več dejavnikov, ki so značilni za njeno fizično, fiziološko, duševno, ekonomsko, kulturno ali družbeno identiteto, pri čemer način identifikacije ne povzroča velikih stroškov, nesorazmerno velikega napora ali ne zahteva veliko časa. /.../
>
> 18. Anonimiziranje - je takšna sprememba oblike osebnih podatkov, da jih ni več mogoče povezati s posameznikom ali je to mogoče le z nesorazmerno velikimi napori, stroški ali porabo časa.

Iz navedenega izhaja, da se osebnih podatkov, po katerih se lahko osebo identificira, v besedilih ne sme uporabljati, ampak je potrebno določiti postopek anonimizacije. Manj pa je razumljivo, katere (javne) osebe in v katerih primerih so iz tega izvzete; morda bi lahko obveljalo načelo ločevanja po kriteriju zajema zasebno/

[172] Vzdevki, izpeljanke iz priimkov ipd.
[173] http://zakonodaja.gov.si/rpsi/r06/predpis_ZAKO3906.html

javno: znotraj zasebne komunikacije se vsi osebni podatki anonimizirajo, znotraj javne komunikacije pa vsi podatki ostanejo nespremenjeni. Vsekakor bi morali v primeru gradnje govornega korpusa obstoječa zakonska določila pregledati in interpretirati pravni strokovnjaki.

6.2 NERAZUMLJIVI FRAGMENTI

Z oznako <neraz> transkriptor označi dele besedila, ki jih ne more transkribirati, ker jih ne razume – zaradi nerazločnega govora, prekrivnega govora ali hrupa v ozadju. Primer iz UKGS:

```
G16: to samo gledaš <pavza>
G17: <neraz> svojega denarja vredni ti Američani
     (ropot avtomobila) (R06)
```

Na take primere naletimo pri vseh transkripcijah govora, zato mora biti v naboru oznak predvidena oznaka za njihovo označevanje. Izkušnje pri transkribiranju korpusa COLT so pokazale, da se z dodatnim branjem oz. poslušanjem posnetkov velikost korpusa poveča skoraj za 20 odstotkov, predvsem na račun nerazumljivih delov (prim. op. 49).

6.3 NAPAČNI ZAČETKI

Napačni začetki so značilnost govorjenega jezika in ena izmed razlikovalnih prvin glede na pisni jezik. Termin označuje dogodek, ki se zgodi na ravni ene same besede (in ne sintaktične enote) – govorec besedo začne, pa je iz različnih razlogov ne dokonča – lahko je prekinjen, lahko si premisli in izbere drugo besedo, lahko se zmoti itd. V slovenski jezikoslovni teoriji je pojem večkrat obravnavan v okviru analize diskurza, poimenovanja pa se nekoliko razlikujejo; pri Kranjc (1999, 66) so to »napačni starti«, pri Hribar (2003, 24) pa »nedokončanost besede«, kar je zaznamovano s stičnim tropičjem, npr. »v Sloveniji popijemo enkrat več alkohola, kot je evropsko povprečje, enk… en… enkrat več ljudi umre zaradi posledic alkohola« (Hribar 2003, 206). Verdonik uporablja izraz »napačni začetek« (2006, 59), obravnava pa ga kot eno izmed oblik popravljanja; v Tabeli oznak (2006, 71) ima poimenovanje »nedokončana beseda« in oznako () na mestu prekinitve, npr.: »Samo malo, da po()« (Verdonik 2006, 70).

Pri transkribiranju besedila za korpus je treba napačne začetke označevati, ker gre največkrat za besedne fragmente; z oznako mdr. zaustavimo nadaljnje analize na fragmentih, ker niso smiselne in običajno tudi ne mogoče, hkrati pa oznaka razdvoumlja ugibanje uporabnikov korpusa, na kaj so naleteli. Sicer pa so napačni začetki lahko tudi predmet raziskave korpusa – zakaj prihaja do tega pojava, na katerih mestih v govoru, kdaj bolj pogosto itd.

```
                a Centru še z  e- z eno vrsto ljudi ə
              kjer lahko na v  en- kratkih nekaj straneh pač
   ni <repet/> tako običajni   fono- fenomen v Evropi
      koncu saj nikoli ne bo   gom- bomo perfektno govorili
        recimo ki imajo manj   govo- n- [govorcev] +G04+
       bo stvar <pavza> zelo   hi- zelo široko razvijala
     po slovensko na začetku   ho- mu hočejo pomagati ne
    rezervo ə dudlajo vzorce   i- in in <repet/> skladen-
študentka ko je diplomirala    i- v Nottinghamu v Veliki
    tega toliko da se moramo   ieo- oddvojiti oziroma ločiti
      namene posnetke svojih   iz- svojih izpitov ə da
        da se to financira da  j- je bilo to dim- mislim
         +G04+ [mhm] əm kako   j- pa je na poletni šoli
 najrazličnejšim publikam ə    j- se pripravlja in bo
       da razbijemo ta kliše   j- v bistvu je težek tako
         vaša izkušnja +G07+   [j-] [kako] to
        jezik in ə ni uradni   je- ne bo zdajle če bojo
 javnosti funkcionira in ti    je- smo poskušali oblikovati
      rekli in jaz mislim da   je- to je najbolj dragoceno
              +G04+ [mhm] da   je- ə bo treba po mojem
      slovenščino pa še malo   jəə- podrobneje jezikoslovje
```

Slika 32: Izbor napačnih začetkov v UKGS

6.4 PONOVITVE

Ponavljanje posameznih besed ali besednih zvez je značilnost govorjenega jezika. Pogosto se zgodi spontano in brez vidnega razloga ali posebnega namena, lahko pa ima natančno določeno funkcijo v govoru, in sicer poudarjanje. Podrobneje o tem razpravlja Verdonik (2006, 157), ki ima tudi ponavljanje za eno izmed strategij popravljanja, ločuje pa ponavljanje na ravni izraza (besede, več besed,

druge besedne oblike iste slovarske iztočnice) in ponavljanje na ravni fonema. V nadaljevanju so prikazani nekateri primeri ponavljanj iz UKGS:

1. kar fajn dala sva za vsako {repet} sva dala {/repet} petnajst jurjev (telefon zazvoni) a veš
2. razlika je za najširšo javnost težko <pavza> <repet>težko</repet> natančno dojemljiva
3. ampak mora biti to kar udeleženec opazi <repet>mora biti</repet> pa veliko bolj lahkotno
4. [ja zdaj mogoče bi res lahko] <repet>mogoče</repet> ločili
5. in to je uspelo ne tudi <repet>tudi uspelo</repet>
6. kar se tiče <repet>kar se tiče</repet> tega ne <repet>kar se tiče</repet>
7. ne pa da <repet>da</repet> na rezervo dudlajo vzorce i=<repet>in in</repet>; skladen= in <repet>in in</repet> ne vem besede in
8. ja ja <repet>ja ja</repet>

Ponovitev se lahko pojavi takoj (primera 2 in 8), pogosto pa so med ponovljenimi besedami in besednimi zvezami še druge besede (vsi ostali primeri). Kako daleč od »izvirnika« je lahko ponovljena beseda, da jo še imamo za ponovitev? Zdi se, da bi tu hitro lahko prišlo do subjektivnega odločanja transkriptorjev. Nadaljnja značilnost ponavljanja je, da se lahko ponovi samo ena beseda (primera 2 in 4), lahko zveza besed (3 in 6), tudi v zamenjanem vrstnem redu (1 in 5), ponovitev je lahko ena, lahko pa jih je več (7 in 8). Vprašanje je, kako ravnati v primerih, ko se beseda ponovi več kot enkrat, ali namreč označiti vsako ponovitev posebej ali vse skupaj. Odločiti bi se bilo treba tudi, ali na enak način označevati ponavljanje tudi v primerih, ko si ponovitve sledijo večkrat ena za drugo:

9. ful je veter pihal ful ful ful
10. in se zelo zelo intenzivno učijo
11. to si mislila ja bravo bravo

Tudi Verdonik ločuje vsaj dve vrsti ponavljanj: prva so instrument popravljanja, druga pa imajo funkcijo poudarjanja, potrjevanja, zagotavljanja. Če bi hoteli ločiti različne vrste ponavljanja, »bi morali analizirati govorčeve namere« (Verdonik 2006, 148). Pri tem gre že za analizo diskurza, ki bi lahko nastala na podlagi govornega korpusa, posebej če bi bila raziskovalcem dostopna tudi cela besedila

(in kjer je iskanje ponovitev relativno enostavno avtomatsko iskati). Poskus na učnem korpusu govorjene slovenščine je pokazal, da bi bilo ponavljanje težko enoznačno označevati, zato predlagam, da pri gradnji večjega govornega korpusa tega elementa spontanega govora ne označujemo; s tem raziskovanje ponavljanj prepustimo uporabnikom korpusa, ki si za namen svojih raziskav lahko korpus po svojih potrebah dodatno označijo ali pa za iskanje po korpusu uporabijo posebej prirejene programe.

6.5 POPRAVLJANJA

Popravljanja[174] so razširjena kategorija napačnih začetkov. Če se termin *napačni začetki* nanaša samo na nedokončane, prekinjene besede, se popravljanje nanaša na besede in besedne zveze, ki so do konca izgovorjene, vendar jih govorec nato zaradi določenega (a pogosto težko določljivega) razloga popravi oz. ponovno izreče – v drugačni obliki, drugačnem zaporedju, s spremenjeno besedo itd. Prikazani so nekateri primeri iz UKGS:

1. v začetku že skušaš- poskusiš govoriti- spregovoriti
2. kot danes ko je bila dosežena- dosežen konsenz
3. so tukaj navade kakšni əm ne vem kakšen je delovni čas
4. hočeš vse imeti iz ne vem od iz Obsessiona
5. mislim və- primerljiv seveda z ne vem s francoskim inštitutom

Kot je bilo že omenjeno, se s popravljanji v svoji analizi diskurza podrobneje ukvarja Verdonik (2006, 59–64 in 147–157), saj predstavlja ta značilnost spontanega govora precej težav pri razvoju sistemov strojnega simultanega prevajanja govora. Avtorica natančno in obsežno povzema tuje razprave o popravljanju, pri tem pa ugotavlja, da je »implicitna ali eksplicitna predpostavka večine raziskav popravljanj, da so popravljanja odmik od idealnega, na nek način pomanjkljivost spontanega govorjenega diskurza« (Verdonik 2006, 64). Kljub temu se zdi, da popravljanja lahko tako razumemo samo, če imamo za »idealno« neko skladenjsko strukturo, ki ustreza pisnemu jeziku. Vsekakor govorec s popravljanji delno razkriva psihološke in sociološke procese, ki potekajo ob tvorjenju govora (Verdonik 2006, 65).

S popravljanjem se v okviru skladenjske analize diskurza podrobneje ukvarja tudi Smolej (2006a, 2006b), kar dokazuje, da je popravljanje z različnih vidikov

[174] Ta termin uporablja tudi Verdonik (2006).

zanimiva značilnost spontanega govora. Označevanje popravljanja v korpusu je zahtevna naloga, predvsem zato, ker je popravljene besede včasih težko opredeliti kot popravke in gre pogosto za interpretiranje, kar onemogoča enoznačno transkribiranje. Če bi se zanj vendarle odločili, bi morali izbrati dovolj robusten način označevanja, ki bi uporabnike korpusa samo opozoril, da je v okolici oznake prišlo do dogodka, ki bi lahko bil popravljanje, izključeval pa bi vsakršno interpretacijo.

6.6 NESTANDARDNE BESEDE IN OBLIKE

Morebitno označevanje nestandardnih besed in oblik zahteva, da najprej definiramo, kaj je v jeziku standardno. S pojemom *jezikovnega standarda* se je v slovenskem jezikoslovju mdr. ukvarjal Skubic, ki ga definira na dva načina:

> »/…/ ima izraz potencialno dva pomena: (1) jezik določene *kakovosti*, ki jo je treba doseči (jezik kot merilo), ter (2) najbolj *običajna*, splošno uveljavljena različica jezika (standard kot običajnost). V prvem pomenu, ki je v literaturi najpogostejši, je ta izraz najbliže v slovenščini uveljavljenemu izrazu *knjižni jezik* /…/. Večkrat pa se v slovenskem jezikoslovju ta izraz uporablja v tem drugem smislu (običajni jezik, lahko npr. določene družbene skupine ali občila (Pogorelec 1967, 1982), in takrat z izrazom *knjižni jezik* seveda ni prekriven. Zdi se, da v zadnjem času pojem pravopisne kodifikacije pri mlajših, korpusno usmerjenih slovenskih jezikoslovcih implicira nekakšno sintezo obeh pomenov – ko naj bi jezikoslovec na podlagi dejansko najbolj razširjene rabe posplošil pravilo in ga povzdignil v kakovostno normo, torej iz standardnega jezika v drugem v standardni jezik v prvem pomenu« (Skubic 2005, 211).

Razumevanje standarda v prvem pomenu za govorjeni jezik pride v poštev samo v določenih govornih položajih. Za pisni jezik je še do nedavnega veljalo, da imamo en sam standard, to je knjižni standard; z razvojem novih tehnologij in vzpostavitvijo novih (elektronskih) komunikacijskih kanalov pa se je tudi to razumevanje relativiziralo. Sedaj se tako v govorjenem kot v pisnem jeziku standard v pomenu zahtevane kakovosti lahko pričakuje samo v določenih okoliščinah, v vseh drugih okoliščinah pa je določena raba lahko povsem *običajna,* (standardna v drugem pomenu besede), odstopa pa od meril, ki veljajo za standard v prvem pomenu besede. Odstopi od idealne (predpisane) normativne izreke (in zapisa) v teh položajih predstavljajo poglavitne prvine jezikovne realnosti (Stabej 2000, 80), in jih je kot take treba tudi razumeti, brez nenehnega sopostavljanja nasproti normativnemu predpisu.

Izkušnje pri gradnji govornih korpusov so pokazale, da težave pri transkribiranju predstavljajo besede, ki jih v pisnem jeziku načeloma ni oz. jih ni v standardu knjižnega jezika, lahko pa so, v določenih ali vseh okoliščinah, povsem sprejemljive za govorjeni jezik; enako velja tudi za novotvorbe ali enkratne tvorbe. Besede, ki jih transkriptor ne more ali ne zna brez pomislekov zapisati, se običajno[175] zbirajo v posebnem seznamu besed (in besednih oblik, mogoče tudi besednih zvez) brez ustaljene pisne oblike. Seznam sestavljajo transkriptorji in je skupen vsem transkriptorjem, dokončno pa ga oblikujejo uredniki korpusa.

Kriteriji, katero besedo uvrstiti na ta posebni seznam besed, so težko določljivi. Odločitev se transkriptorju ali uredniku korpusa vsiljuje na intuitivni ravni, in sicer kot posledica globoko ukoreninjene zavesti o knjižnem jeziku in njegovi normiranosti na ravni pisnega jezika. Če se želimo izogniti intuitivnemu kvalificiranju jezika,[176] je treba pred začetkom gradnje določiti čim bolj objektivne kriterije za odločanje o tem, kaj so nestandardne besede in oblike. Eden izmed možnih kriterijev nestandarda je »standardni« slovar: če besede ni v slovarju, je nestandardna. Vendar pa konkretna slovenska situacija takega ravnanja ne dopušča, glede na to, da zaradi znanih razlogov niti SSKJ niti pravopisni slovar ne predstavljata sodobnega stanja slovenskega jezika. Zato je v tem smislu boljša rešitev naslonitev na korpus slovenskega pisnega jezika FidaPLUS, ki je največji referenčni pisni korpus slovenskega jezika. Vsako izgovorjeno besedo, s katero bi imeli transkriptorji težave, bi lahko preverili v pisnem korpusu. Če bi se izkazalo, da beseda (vsaj v korpusu) še ni bila zapisana (ali morda samo z zelo omejenim številom pojavitev), bi jo dodali na seznam besed brez ustaljenega zapisa; pričakujemo lahko, da bodo na tem seznamu predvsm besede in oblike, ki bi jih tradicionalno označili za pogovorne, narečne in slengovske.

Tudi Verdonik označuje nekatere besede in oblike: »pogovorna oblikoslovnoskladenjska raba oz. pogovorni izraz« (Verdonik 2006, 71) je označen z zvezdico v oglatih oklepajih, npr. boste mi dal[*], tam en[*] hotelček, pa morate vedet[*], lahko pokličete čez enih[*] pet minut, če slučajno vejo[*] na ministrstvu, pa se mogoče pol[*] čujeva, te[*] pa dorečema[*] pozneje, se pomenima[*], toto[*] animacijo; šalter[*], ziher[*], fajn[*], glih[*], garantirat[*], duplih[*] rezervacij, brezveze[*], nonstop[*], gor[*] na spletni strani (vse Verdonik 2006, 180–223). Kot lahko vidimo, so označene besede nestandardni izrazi, seveda v odnosu do knjižnega standarda. Verdonik ne pojasnjuje kriterija, po katerem je nastal ta se-

[175] Npr. v primeru BNC in The Bank of English.

[176] Kar se je zgodilo meni pri transkribiranju za UKGS, ko sem številne besede označevala za nestandardne, npr. angažma, benz, cajt, carsko, dek, dila, dofilati, faks, fila, fora, frej, ful, furajo, gužva, v iber hudih, jabki, jabčki, jebala ... , pač po analogiji s pisno normo knjižnega jezika.

znam; pri t. i. pogovornih izrazih je predvidoma prevladal avtoričina jezikovna intuicija. Pri morebitni gradnji referenčnega govornega korpusa, kjer predvidevamo večje število transkriptorjev, ki morajo delovati skladno in kjer morajo biti načela zapisovanja govora povsem transparentna tudi za vsakega morebitnega uporabnika korpusa, pa je treba načela določiti bolj jasno. Zato predlagam, da ostane prilagajanje zapisa govoru na ravni besed in besednih oblik, kot je bilo predlagano v poglavju 5.6, *Končni predlog priporočil za transkribiranje govorjene slovenščine*, in da se v celoti odrečemo tako oznaki <nst> kot označevanju besed, ki naj ne bi sodile v knjižni jezik. Če se te besede v govorjenem jeziku pojavljajo, jih moramo po osnovni ideji korpusnega jezikoslovja o zbiranju in analizi avtentičnega gradiva tako tudi dokumentirati, in sicer brez oznake, ki že pomeni interpretacijo.

6.7 KRATICE IN OKRAJŠAVE

Označevanje kratic in okrajšav ni bilo problematično, v UKGS pa se jih je pojavilo samo nekaj:

- to je bilo preko <okr>alteja</okr> (R04)
- okej okej (R02)
- na <okr>emtiviju</okr> (R06)
- po <okr>tiviju</okr> bo (R06)
- na <okr>fədəvəju</okr> bo (R06)

V zgornjih primerih pravzaprav ni jasno, katere sploh še zaznavamo kot kratice, katere pa imamo za navadne besede.

6.8 PREKRIVNI GOVOR

Izrazita značilnost govorjenega besedila je prekrivni ali hkratni govor več govorcev. Gre za pojav, ki ga zaznamo praktično v vseh dialoških ali multiloških besedilih, ne glede na okoliščine ali stopnjo formalnosti, čeprav se količina prekrivnega govora v bolj formalnih okoliščinah običajno znižuje. Očitno je, da prekrivanje besedila v govoru do neke mere ne povzroča težav v sporazumevanju, od neke točke naprej pa je moteče; to se izkazuje v intervencijah sogovornikov ali poslušalcev ali v njihovem izražanju nerazumevanja.

Pri transkribiranju za govorne korpuse se prekrivni govor vedno označuje, čeprav na različne načine; v korpusu London-Lund je npr. označen z zvezdicami, v korpusih

BNC in COLT z oglatimi oklepaji. V slovenski jezikoslovni teoriji Kranjc uporablja izraz »prekrivajoči se govor« (1999, 65), Verdonik pa »hkratni govor« (2006, 69); označuje ga z oznakama [1] na začetku in [2] na koncu prekrivajočega se dela izjave. V UKGS sem za transkribiranje prekrivnega govora uporabila oglate oklepaje (na začetku in koncu prekrivnega dela), kar se je po mojem izkazalo za ustrezno:

```
+          [drugi] del publike +G04+ [mhm] recimo ki študira
+        govo- n- [govorcev] +G04+ [mhm] saj v končni fazi ne gre
+     posebej [dopovedovati] +G04+ [mhm] torej seminar je zasnovan
+                  [ljudi] v +G04+ [mhm] tujih državah da bodo
+     od ponedeljka do petka ne ja mhm] <neraz> pa naši semenir-
+        življenje [<neraz>] +G04+ [mhm] <pavza> [Slovenija] in to
+      kot za eno državo +??+ [mhm mhm] <pavza> mislim in to
+   precej na delovnem mestu +G04+ [mhm] ə doktor <ime> <priimek>
+  razsežnosti ə [Slovenije] +G04+ [mhm] ə je pa razlika
+        kulture +G07+ [seminarja] [mhm] ə ki prihaja i
```

Slika 33: Primeri označenega prekrivnega govora v UKGS

Transkribiranje prekrivnega govora običajno predstavlja težavo pri transkribiranju, posebej kadar govori več govorcev hkrati. Pri zapisovanju prekrivnega govora so transkripcijska orodja lahko v veliko pomoč, saj omogočajo, da se na mnogo lažji način doseže bistveno večji pregled nad prekrivnim govorom znotraj transkripcij. Doslej še nisem zasledila, da bi izpis prekrivnega govora podpirali iskalniki po korpusu; v konkordančnem izpisu lahko sicer vidimo, da je bilo hkrati s prikazanim besedilom govorjeno še neko drugo besedilo, če imamo simultani dostop do zvočnih posnetkov pa lahko to tudi slišimo. Posnetke, kjer prekrivni govor večjega števila govorcev zelo otežuje transkribiranje, raje izločimo iz korpusa: pomembnost takih posnetkov je za namen referenčnega korpusa glede na vloženo delo in porabljeni čas premajhna.

6.9 PREMORI V GOVORU

Premore v govorjenem besedilu sem v UKGS zapisovala z oznako <pavza>.[177] Premorov, krajših od 1 sekunde, nisem zapisovala. Daljše premore v govorje-

[177] Pri izbiri oznak za UKGS sem se, kadar je bilo mogoče izbirati, odločala za izraze, ki so nekoliko bolj mednarodno razumljivi. Razlog za to je bilo dejstvo, da je učni korpus nastajal na Univerzi v Bergnu v sodelovanju s strokovnjaki, ki slovenščine niso razumeli.

nju pa je treba v transkripcijo zapisovati zaradi različnih razlogov. Lahko jih označujemo z navedbo časa oz. trajanja v oklepaju (<pavza>(20) pomeni premor dolžine 20 sekund). Podatek je pomemben za uporabnika korpusa, ki se s tem izogne »poslušanju tišine« (kadar so zvočni posnetki dostopni), seveda pa so premori sami na sebi lahko tudi raziskovalno gradivo, saj lahko v govoru funkcionirajo kot sredstvo za segmentiranje govora, za poudarjanje in drugo. Za transkriptorja so pomembni vsi premori v govoru, tudi krajši, ki jih sicer ne označuje, so pa njegovo vodilo pri postavljanju mej med izjavami. Premorov pri transkribiranju s pomočjo transkripcijskih orodij sploh ni težko zaznavati, saj je oscilogram govora ves čas razviden na ekranu računalnika (prim. Sliko 27, Sliko 29, in Sliko 31).

6.10 DRUGE PROZODIČNE OZNAKE

Govor spremljajo različni akustični dogodki, ki lahko odločilno vplivajo na razumevanje vsebine in pomena, zato jih pri transkribiranju označujemo. Posebna vrsta oznak je namenjena dogodkom, ki so neposredno povezani z govorjenjem. To so oznake za premore v govorjenju, glasnost in hitrost govorjenja, kvaliteto in višino glasu, intonacijo, pa tudi za različne oblike deformiranja glasu ali oponašanja. Prozodija je v različnih korpusih različno označena, spet odvisno od namembnosti korpusa ter od razpoložljivih človeških in finančnih virov. Premori so označeni tako rekoč v vseh korpusih, drugače pa je s tonsko višino, potekom in jakostjo. Med obravnavanimi tujimi korpusi je zelo natančno prozodično označen korpus London-Lund, četrtina polmilijonskega korpusa COLT in 250.000 besed 8-milijonskega Nizozemskega govornega korpusa. Razlogi so znani: označevanja intonacije, višine in jakosti tona so izredno zahtevna opravila, ki zahtevajo posebno tehnično opremo, vrhunske strokovnjake in veliko časa; za referenčne korpuse pride tovrstno označevanje, ki je sicer izjemnega pomena za raziskovanje jezika, v poštev samo na manjšem delu gradiva.

V slovenski jezikoslovni teoriji sta v zadnjem času nastali dve razpravi, ki se ukvarjata z raziskovanjem prozodije spontanega govora. Vitez in Zwitter Vitez sta na manjšem korpusu zelo natančno, s pomočjo merilnih naprav, raziskovala prozodične lastnosti govorjene slovenščine – višino (in spremembe višine) osnovnega tona, jakost, hitrost govora in premore; njune ugotovitve in metodologija bi lahko služile za izhodišče pri morebitnem prozodičnem označevanju dela govornega korpusa. Pri zapisovanju parlamentarnega jezika je členitev besedila s premori, potek stavčne intonacije, register in hitrost govora upoštevala tudi Hribar (2003), vendar je označevala »po posluhu« (2003, 25), kar je za postavljanje ločil v tran-

skripcijo morda zadostovalo, za prozodično označevanje korpusa pa ne pride v poštev.

Poleg prozodičnih oznak, ki zaznamujejo tonsko višino, tonski potek, jakost in hitrost govora, obstajajo tudi prozodične oznake, ki zaznamujejo oblike deformiranja glasu. Po priporočilih TEI imajo te skupno oznako <*shift*>, sledi pa ji opis dogodka. To je po eni strani pomembno zaradi možnosti dopolnjevanja seznama prozodičnih oznak (in hkrati ohranjanja konsistentnosti), po drugi strani pa omogoča lažji pregled nad oznakami pri kasnejših analizah korpusnega gradiva. V nadaljevanju so predstavljeni primeri deformiranja glasu oz. odstopi od normalnega poteka glasu v UKGS:

```
<shift:   smeh>besedilo</shift >
<shift:   glasno>besedilo</shift >
<shift:   kričanje>besedilo</shift >
<shift:   šepetanje>besedilo</shift >
<shift:   razpotegnjeno>besedilo</shift >
<shift:   počasi>besedilo</shift >
<shift:   odsekano>besedilo</shift >
<shift:   oponašanje otroškega
          glasu>besedilo</shift >
<shift:   oponašanje dolenjskega narečja>
          besedilo</shift >
<shift:   govorjenje s polnimi
          usti>besedilo</shift >
```

Nekatere oznake sem izločila že med gradnjo UKGS; ostali so samo primeri prozodičnega označevanja, ki se nanašajo na jakost besedila, tonski potek (antikadenca) in govorjenje med smehom, vendar je bilo označevanje v teh primerih zelo nekonsistentno:

```
{shift=poud} ta je pa vredna da jo preslišite {/shift
     =poud}
{shift=vpr} a res {/shift=vpr}
{shift=smeh} ne moreš skriti {/shift=smeh}
```

Za morebitno gradnjo govorne komponente referenčnega korpusa predlagam prozodično označitev dela korpusa, pri čemer naj označevanje višine in jakosti tona ter tonskega poteka (intonacije) temelji na meritvah.

6.11 NEVERBALNI GLASOVI

Med govorjenjem zaznavamo tudi glasove in glasovne sklope, ki lahko imajo komunikacijsko funkcijo, pa niso besede ali vsaj ne polnopomenske besede; to so medmeti tipa *mhm*, *uaa*, *oh,* lahko pa tudi drugi glasovi, ki jih proizvaja človek med govorjenjem – smeh, tleskanje z jezikom, vzdihovanje, kašljanje ipd. Priporočila EAGLES razlikujejo polverbalne glasovne dogodke (*mhm*) od neverbalnih dogodkov (vzdih). Vseh omenjenih glasovnih sklopov ne moremo obravnavati na enak način. Mnoge med njimi najdemo kot iztočnice v SSKJ in v SP. V takem primeru EAGLES priporoča, da jih zapisujemo na slovarski način, brez posebnih oznak. Manj pogoste medmete, ki jih običajno ne zapisujemo, je potrebno vpisovati v seznam, ki ga uporabljajo transkriptorji za poenotenje. Neverbalne glasove, ki jim ni mogoče dati ortografske podobe, po priporočilih EAGLES in TEI označujemo s posebno oznako (npr. <nv>) in opisom.

Večino glasovnih sklopov, ki se kot besedna vrsta uvrščajo med medmete (ali člene), v okviru najsodobnejših tokov analize diskurza opazujemo v vlogi diskurznih označevalcev (prim. npr. Andersen 2000, Verdonik 2006). Verdonik (2006, 81–141) jih zapisuje kot *mhm*, *aha*, *aja*, *eee*[178] in *mmm*, redkeje pa tudi *nnn*, *eeeh*, *eeef*, *eeen*, *eeennneee* in *eeemmmeee*. Kranjc v analizi otroškega spontanega govora ugotavlja, da je medmetov v govoru 22–25 odstotkov, podrobneje pa se z njimi ne ukvarja, ker jo zanimajo predvsem polnopomenske besede. Hribar (2003) diskurznih označevalcev ne zapisuje. Smolej (2006) medmete zapisuje v zavitih oklepajih, način zapisovanja ni pojasnjen; iz gradiva lahko razberemo zapise {e} (verjetno polglasnik), {ee}, {em}, {aaa} itd.

V UKGS sem medmete zapisovala tako, kot sem jih slišala, in brez posebne oznake. Pojavili so se naslednji glasovni sklopi:

> ajd, ah, aha, aja, ane, ej, evo, hjah, ə, əm, jah, joj, m, ma, mah, mhm, no, oh in oja.

To seveda ni končni seznam, ker nikoli ne moremo predvideti, kateri glasovni

[178] Ta glas imenuje zavlečeni polglasnik (Verdonik 2006: 110).

sklopi se bodo še pojavili. V UKGS so se pojavili tudi neverbalni glasovi, ki jih ni bilo mogoče ortografsko zapisati, ampak jih je bilo treba opisati v okviru oznake:

- `<nv> vdih </nv>`
- `<nv> smeh </nv>`
- `<nv> uau </nv>`
- `<nv> nedoločljivi glasovi </nv>`
- `<nv> vzdih, ki izraža začudenje </nv>`
- `<nv> pihne skozi usta </nv>`

Zapisovanje neverbalnih komunikacijskih dogodkov v okviru govornega korpusa je nujno, ker prispeva k razumevanju besedila in je lahko tudi predmet samostojnih raziskav (npr. diskurzni označevalci), tako v okviru jezikoslovja (Smolej 2006a) kot v okviru govornih tehnologij (Verdonik 2006), saj gre za inherentne sestavine govora, ki se jih mora naučiti razumeti tudi računalnik. Zato za označevanje govornega korpusa priporočam natančno zapisovanje neverbalnih komunikacijskih glasov in dogodkov.

6.12 NEKOMUNIKACIJSKI GLASOVI

To so glasovi, ki potekajo v ozadju sporazumevanja ali med njim, vendar običajno nimajo neposrednega vpliva na potek govorjenja (izjemoma lahko tudi). Gre za zvoke, ki jih ne producirajo govorci.[179] Zapisujemo jih znotraj navadnih oklepajev; včasih pomagajo pri razumevanju poteka pogovora, najpogosteje pa zapis omogoča lažjo uporabo korpusa. V okviru UKGS so se pojavili naslednji nekomunikacijski zvoki:

- (premikanje mikrofona)
- (glasba)
- (ropot avtomobila)
- (šumenje papirja, listanje)
- (zvonjenje telefona)

Računalniški pretvorbeni program besed, ki opisujejo nekomunikacijske zvoke, ne sme obravnavati kot del korpusnega besedila; vse, kar je zapisano znotraj navadnih oklepajev (*besedilo*), mora biti izločeno iz nadaljnje računalniške obdelave.

Vprašanje je, kje v transkripciji se nekomunikacijski zvoki zapisujejo. Deloma odgovor sugerirajo že obstoječa transkripcijska orodja, saj npr. *Praat* predvideva

[179] Po tem se tudi ločijo od neverbalnih glasov, ki nimajo komunikacijske funkcije, npr. kihanje, kašljanje.

posebno vrsto v transkripciji ravno za zvoke v ozadju, *Transcriber* pa predvideva zapisovanje zvokov iz ozadja znotraj posameznih izjav. Vsak način ima svoje prednosti in slabosti. Če gre za relativno kratek zvok, ga je smiselno zapisati znotraj izjave ali ob izjavi; če pa gre za dlje časa trajajoči zvok (npr. glasba na radiu, ki spremlja celotni govorni dogodek), bomo to verjetno zapisali v opombo v glavi besedila ali morda v transkripcijo na začetku ali na koncu besedila.

6.13 BRANO BESEDILO

Znotraj govora se lahko pojavi besedilo, ki je brano; eden izmed govorcev ali več govorcev prebere odlomke besedila ali besedilo v celoti. V korpusih govora, ki so nastajali za potrebe govornih tehnologij, so bila govorjena besedila največkrat brana, za korpus spontanega govora pa niso reprezentativna, zato jih v transkripciji označimo. Računalniški program lahko dele transkripcij, označene kot <branje>*besedilo*</branje>, izloči iz nadaljnje obdelave, lahko pa jih shrani v poseben podkorpus, za morebitne primerjalne raziskave.

6.14 NEZANESLJIVA TRANSKRIPCIJA

Oznako <?>*besedilo*</?> transkriptor uporabi, kadar besede ne pozna in je ne zna zapisati, čeprav jo razločno sliši in jo lahko ortografsko transkribira. Oznaka se razlikuje od oznake <neraz>, s katero transkriptor označi del besedila, ki ga ne sliši dobro, zato ne razume in ne more zapisati. V primeru UKGS npr. nisem znala zapisati besede, ki jo je govorec verjetno izgovoril v katalonskem jeziku <?>karrotera</?>, angleškega imena trgovine, ki so jo omenjali govorci – v <?>Obsešnu</?>, in slengovskega poimenovanja oblačila <?>*baggy pants*</?> Oznaka za nezanesljivo transkripcijo je v pomoč transkriptorju, da se na takih delih ne zadržuje predolgo; urednik korpusa bi moral pri pregledovanju transkripcij poskušati razrešiti čim več takih primerov, lahko tudi s pomočjo področnih strokovnjakov. Tako je to lahko neke vrste začasna, redakcijska oznaka, ki pa se teoretično lahko pojavi tudi v korpusu, če tudi urednik ne zna razrešiti primera.

6.15 NEPREPOZNAVNI GOVOREC

Včasih se zgodi, da transkriptor posamezne izjave ne more pripisati določenemu govorcu. To se pogosto zgodi v primeru, ko je izjava kratka in nerazločna, npr. mhm, ko gre za smeh ali kaj podobnega. Možno pa je tudi, da se pogovarjajo lju-

dje istega spola in enake starosti ter s podobnimi glasovi, kar privede do popolne negotovosti transkriptorja, kdo kdaj govori; v takem primeru je najbolje to zapisati v opombo v glavi besedila. V korpusu tovrstna negotovost vpliva predvsem na statistiko korpusa; koristno je, če je v transkripciji označeno, kdaj identifikacija govorca ni mogoča ali ni zanesljiva. To se lahko naredi z nevtralno oznako <GX_1> namesto identifikacijske oznake govorca, npr. <G13>, produkcijo neznanih govorcev pa v demografski statistiki korpusa vodimo ločeno.

6.16 TABELA OZNAK

Na podlagi vsega navedenega predlagam naslednji nabor oznak za označevanje jezikovnih in drugih dogodkov v govoru:

Oznaka	Pomen
<p>, <p(5)>	kratek premor (pribl. 1 sekundo), premor (5 sekund)
<ime>, <naslov>, <tel. št.>	nadomeščajo osebnih podatkov
<nz>	napačni začetek
[besedilo]	prekrivni govor
<nv>smeh</nv>	neverbalni zvoki
<tj>besedilo</tj>	besedilo, izgovorjeno v tujem jeziku
 besedilo</br>	brano besedilo
<neraz>, <neraz(5)>	nerazumljivo, nerazumljivi govor (5 sekund)
(premikanje mikrofona)	nekomunikacijski zvoki

Tabela 16: Priporočen nabor oznak za označevanje KGS

Označevanje besedila poteka neposredno oz. hkrati s transkribiranjem. Govorno komponento referenčnega korpusa se transkribira ortografsko, tudi z vpisovanjem (dogovorjenih) besed brez ustaljenega zapisa, hkrati pa se vnašajo tudi zgoraj navedene oznake. Ko je transkribiranje končano, transkriptor doda še glavo besedila, nato pa posnetek in transkripcijo z oznakami in glavo odda uredniku v dodatni pregled in označevanje.

7 Oznake v glavah dokumentov

Oznake v glavah dokumentov v korpusu dodatno opisujejo govorjena besedila, in sicer z namenom, da bi uporabniku čim bolj približala njihovo avtentično podobo. V prejšnjih poglavjih sem obravnavala oznake, ki so del besedila in se jih v besedilo vključuje med transkribiranjem, oznake v glavah pa besedila opisujejo na bolj splošni ravni. To so npr. bibliografski podatki o besedilu pri pisnih besedilih, pri govorjenih besedilih pa demografski podatki o govorcih, podatki o kraju in času nastanka posnetka, o postopku transkripcije in podobno. Tudi pri tem označevanju sledimo mednarodnim standardom zapisovanja.

7.1 PRIPOROČILA TEI ZA OZNAKE V GLAVAH TRANSKRIBIRANIH DOKUMENTOV

Po filozofiji TEI so na najvišji ravni vsi korpusni dokumenti, pisni in govorjeni, zgrajeni enako: sestavljeni so iz glave <*header*> in besedila <*text*> (Burnard 1995, 72).

<besedilo>	je samostojno besedilo ali segment naravnega jezika v kakršnikoli obliki, ki ga lahko imamo za samostojno enoto pri nadaljnjem procesiranju;
<glava>	nosi opis besedila kot bibliografske enote (vir, sistem označevanja, jezik, situacija, v kateri je besedilo nastalo, udeleženci itd.).

Vsi dokumenti naj bi po priporočilih TEI imeli naslednje oznake:[180]

<id>	(unikatna) identifikacijska oznaka dokumenta
<n>	ime dokumenta
<lang>	jezik.

Glava korpusa po TEI vsebuje štiri vrhnje elemente (Erjavec 1998a, 126 in 1998b, 88–89):

<*fileDesc*>	bibliografski podatki o besedilu, navedba vira
<*encodingDesc*>	podatki o sestavi korpusa in o oznakah korpusa
<*profileDesc*>	nebibliografski podatki besedila, npr. informacije o udeležencih pogovora, okoliščinah, uvrstitev v taksonomijo, kvantitativni podatki (velikost korpusa, trajanje posnetka), datumi

[180] Burnard 1995, 76.

<*revisionDesc*>	opis sprememb elektronskega besedila (npr. če je besedilo lektorirano).

V glavi korpusa je shranjena dokumentacija korpusa kot celote: postopki označevanja, taksonomija besedil, velikost korpusa, bibliografija itd.; v glavi posameznega besedila pa so oznake, ki uporabnika pripeljejo do izvirnika besedila, oznake, uporabljene v besedilu, vrstne oznake besedil glede na taksonomijo korpusa itd.

Posebna skupina TEI za govorne korpuse (Johansson 1995, 97) je izoblikovala dodatna načela za označevanje govorjenih besedil. Značilno za glave govorjenih besedil je, da predvidevajo dve dodatni oznaki:

<*recordingStmt*>	podatki o zvočnih posnetkih
<*scriptStmt*>	podatki o transkripcijah.

Glava besedila vsebuje tudi podatke o udeležencih, okoliščinah snemanja, trajanju posnetka, opombah itd. Pomembno je, da je predviden prostor za vsako informacijo, ki bi jo transkriptor želel zapisati (Johansson 1995, 85).

Tudi v slovenskem korpusnem jezikoslovju so bila pri gradnji prvega referenčnega korpusa FIDA (in njegove nadgradnje) upoštevana priporočila TEI (prim. Erjavec 1998a, 1998b, 2003). Tako glava korpusa vsebuje podatke o oznakah korpusa, bibliografijo korpusa, kvantitativne podatke o velikosti korpusa in številu uporabljenih oznak, taksonomijo besedil ter jezike korpusa. Glave posameznih besedil pa vsebujejo popolno bibliografijo besedila, vključno z referenco na originalni vir, seznam oznak, uporabljenih v besedilu, ter uvrstitev v taksonomijo besedil. Priporočila TEI so izhodišče tudi za načrtovanje oznak v glavah govorjenih oz. transkribiranih besedil, nekaj konkretnih že realiziranih zgledov pa sledi v nadaljevanju.

7.2 OZNAKE V GLAVAH TRANSKRIBIRANIH BESEDIL

7.2.1 Govorna komponenta BNC

V korpusu BNC so prvič označevali glave dokumentov (<*header*>), saj je korpus nastajal sočasno s priporočili TEI. Podatke so pri gradnji korpusa različni partnerji zbirali, shranjevali in zapisovali na različne načine. Šele kasneje so bili ti podatki preneseni v glavo besedila, med njimi pa so zaradi različnega načina zbiranja velike razlike, kar je bilo kasneje predmet številnih kritik uporabnikov

korpusa (npr. Berglund 1999). Spodnja slika prikazuje primer identifikacijskih oznak govorca iz glave dokumenta v korpusu BNC (*BNC Users Guide* 2000, 8.3):

```
<person age="0" dialect="XLO" id="PS5A1"
role="self" sex="m" soc="C2">

<name>Terry</name>
<age>14</age>
<occupation>student</occupation>
<dialect>London</dialect>
</person>
```

Slika 34: Primer identifikacije govorca v glavi dokumenta BNC

V prvih dveh vrstah so kategorizirani podatki, ki se pri uporabi korpusa uporabljajo za izpise oz. iskanje po posameznih kategorijah; v nadaljevanju so podatki, ki osebo dodatno označujejo, vendar nabor teh podatkov od primera do primera zelo variira.

Pri korpusu BNC je bilo prvič izvedeno označevanje v SGML formatu na tako veliki količini transkribiranega gradiva. Nekatere oznake, npr. menjave govorcev, prekrivni govor, napačne začetke, nerazumljive besede, premore itd. so v besedilu označevali že transkriptorji (seveda ne v SGML formatu; njihove oznake so bile kasneje konvertirane), druge, npr. demografske informacije in klasifikacijske podatke pa so uredniki kasneje dodali v glave besedil (*BNC Users Guide* 2000, 8).

7.2.2 Švedski govorni korpus

Glave transkribiranih dokumentov v Švedskem govornem korpusu so drugačne od glav BNC. Nastale so kasneje, so pa precej enostavnejše:

```
@Begin
@Transcribed by: Madelaine Holsten, 19980413, Hans
Vappula, 19980414, CLAN'ed Cajsa Ottesjö, 20030226
@Participants: J Jesper the travel agency
```

```
clerk, P Pia a young female customer
@Dependent: com, eng
@Com: Transkriptionsnyckel: ( Tydlig stigton, ( ty-
dligt fall, °omger tystare tal°, betoning höjning,
be:toning sänkning, *markerar att något säjs med sk-
ratt i rösten*, avbrott-
@Filename: A8201011:Travel Agency Dialogue I
@Date: 19980316
@Anonymized: yes
@Access: Public
@Duration: 00:02:45
@Content:
@Blank
```

Slika 35: Glava besedila v Švedskem govornem korpusu[181]

Glava vsebuje imena transkriptorjev in datume transkripcij, ime dokumenta, datum nastanka posnetka, trajanje, dostopnost in informacijo o anonimizaciji (nadomeščanje osebnih lastnih imen in drugih osebnih podatkov z nevtralnimi oznakami). Demografskih podatkov govorcev praktično ni, samo zelo površne informacije (npr. *P=Pia, mlajša stranka ženskega spola*). Pri zajemanju besedil v Švedski govorni korpus se sestavljavci niso ozirali na demografsko sestavo govorcev, njihov zajem je temeljil na besedilnovrstni taksonomiji: prizadevali so si za čim večje število različnih besedilnih vrst (Tabela 6, str. 46). Tudi tako poenostavljene glave besedil so očitno lahko dobro služile svojemu namenu, saj je na podlagi raziskav korpusa nastalo veliko razprav, predvsem o razlikah med govorjenim in pisnim jezikom.[182] S tem se še enkrat potrjuje hipoteza, da je označevanje korpusa odvisno predvsem od namena korpusa in mora biti za vsak korpus posebej določeno.

7.3.3 Glava v COLT-u

Mnogo bolj zapletene so glave besedil v korpusu COLT, kjer so sestavljavci dosledno sledili priporočilom TEI oz. oznakam v korpusu BNC :

[181] http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3&SUBPAGE=6&FILE=coded_dialog
[182] http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=5

<REG>	S	*region: south*
% reference number // title		
<REF>	B140402	*number of text/file*
<TIT>	?	*title*
% date // time		
<DAT>	?	*date*
<TIM>	?	*time*
% recording // input device		
<DEV>	wlk	*walkman*
% duration of conversation		
<DUR>	6.23	*minutes.seconds in decimal code*
<WC>	1324	*word count*
% details of conversation // locality - county & town		
<CTY>	Greater London county	
<LOC>	Barnet	*borough where conversation takes place*
% setting // activity		
<SET>	classroom	*setting*
<ACT>	art lesson	*activity*
% type of discourse // superfield // subject // individual sub		
<TYP>	conversation	*type of discourse*
<SUP>	?	*superfield*
<SUB>	?	*subject*
<IND>	?	*individual subject*
% spontaneity factor // audience		
<SPN>	3	*3 indicates highest spontaneity*
<AUD>	4+	*number of participants and/or people present*
<BOB>	?	*miscellaneous information regarding the conversation*
% details of participant // identifier // name // gender		
<IDT1>	1	*1 is always the recruit carrying the walkman*

<NAM1>	Alex	*name of the recruit*
<GEN1>	m	*gender*
% age // first language // dialect		
<AGE1>	14	*age*
<LAN1>	BrE	*language*
<DIA1>	London	*dialect*
% occupation // education // social group		
<OCC1>	student	*occupation*
<EDU1>	still studying	*education*
<SOC1>	1	*social group*
% relationship to respondent // other relationships		
<REL1>	respondent	*indicates that this speaker is the recruit (respondent)*
<OTH1>	friend2 friend8 pupil2 *this speaker is a friend of speakers 2 and 8, and a pupil of speaker 12*	
<BOB1>	?	*miscellaneous information regarding the speaker*
% details of participant // identifier // name // gender *information about the next speaker*		
<IDT2>	2	
<NAM2>	Marc	
<GEN2>	m	
% age // first language // dialect		
<AGE2>	14	
<LAN2>	BrE	
<DIA2>	London	
% occupation // education // social group		
<OCC2>	student	
<EDU2>	still studying	
<SOC2>	?	
% relationship to respondent // other relationships		
<REL2>	friend	*speaker 2 is a friend of the recruit*
<OTH2>	friend3 friend8 friend9 pupil12 - *and a friend of speakers 3, 8, 9, also a pupil of speaker 12*	

```
<BOB2> ?
% details of participant // identifier // name // gender etc
<IDT8>                  8
<NAM8>                  Daniel
<GEN8> m
% age // first language // dialect
<AGE8>                  13
<LAN8>                  BrE
<DIA8>                  London
% occupation // education // social group
<OCC8>                  student
<EDU8>                  still studying
<SOC8> ?
% relationship to respondent // other relationships
<REL8>                  friend
<OTH8>                  friend2 pupil12
<BOB8> ?
% details of participant // identifier // name // gender
<IDT12>                 12
<NAM12>                 ?
<GEN12> m
% age // first language // dialect
<AGE12> 35
<LAN12> BrE
<DIA12> London
% occupation // education // social group
<OCC12> teacher
<EDU12> ?
<SOC12> ?
The Bergen Corpus of London Teenage Language
14 % relationship to respondent // other relationships
<REL12> teacher
<OTH12> teacher2 teacher8
<BOB12> ?
```

Slika 36: Glava besedila v COLT-u[183]

[183] http://torvald.aksis.uib.no/colt/cd/32401-cor.txt

Prvi del podatkov se nanaša na posnetek: naslov, datum, čas nastanka, trajanje posnetka, snemalna naprava, kraj, lokacija in okoliščine, v katerih je posnetek nastal, število udeležencev, stopnja spontanosti in tip besedila. V drugem delu so podrobni demografski podatki o udeležencih: ime, spol, starost, jezik, narečje, poklic, izobrazba, socialna skupina ter razmerje do drugih udeležencev pogovora. V primerjavi s švedskim korpusom pa npr. glava v COLT-u nima podatkov o transkriptorjih in nastanku transkripcije. Razumljivo je, da je demografskim podatkom govorcev posvečeno veliko pozornosti, saj je zajem besedil v korpus govorjene najstniške angleščine temeljil prav na teh podatkih (taksonomija besedil ni bila pomembna), korpus pa je bil kasneje kot podkorpus priključen demografski komponenti BNC.

7.3 PREDLOG PRIPOROČIL ZA OZNAKE V GLAVAH DOKUMENTOV KGS

7.3.1 Dokumentacija posnetkov

Za vsak posnetek, ki ga nameravamo vključiti v govorni korpus, mora biti takoj po snemanju izpolnjen Zbirni list podatkov besedila. Ta dokument vključuje vse podatke, ki so po presoji načrtovalcev korpusa pomembni za kasnejše korpusne analize; podatki bodo preneseni v glavo posameznega besedila – transkripcije. Pri sestavljanju nabora podatkov za vzorčni *Zbirni list podatkov besedila* za UKGS sem se deloma opirala na priporočila za označevanje govorjenih besedil TEI in realiziranih korpusnih projektov, izbrane kategorije pa slonijo na odločitvah, sprejetih v poglavju 3.4, *Končni predlog priporočil za zajem besedil v KGS*. V nadaljevanju je predstavljen zbirni list podatkov besedila enega izmed posnetkov učnega korpusa.

ZBIRNI LIST PODATKOV BESEDILA

Identifikacijska oznaka posnetka:184 R07

(Izpolni, kdor je besedilo posnel)
Datum snemanja: 7. 11. 2004
Kraj snemanja (regija in ime kraja): *Gorenjska, Kranj*
Prostor: *doma v dnevni sobi*
Posebne okoliščine:
Besedilo posnel/-a: *M. L.*

[184] Identifikacijsko oznako posnetku dodeli urednik korpusa.

Trajanje posnetka: *5.12 min*
Št. aktivnih udeležencev/govorcev:[185] *2*
Št. poslušalcev:
Posneto brez vednosti govorca: da ✓ ne
Opombe: _____

(Izpolni transkriptor:)

Identifikacijska oznaka transkripcije:[186] TransR07

Transkriptor/-ka: *Jana Zemljarič Miklavčič*
Datum (dokončane) transkripcije: *24. 11. 2004*
Transkriptorjeve opombe:_____

(Izpolni urednik korpusa:)
Transkripcijo pregledal/-a: *Jana Zemljarič Miklavčič*
Datum pregleda: *30. 11. 2004*
Vrsta besedila: *spontana konverzacija*

Prevladujoča struktura besedila:[187]	monolog	dialog ✓	multilog
Okoliščine:	javno		zasebno ✓
Govorni položaj:	formalni		neformalni ✓
Prenosnik:	osebni stik ✓	telefon	mediji

Slika 37: Zbirni list podatkov besedila (R07)

Zbirni list vsebuje podatke o nastanku posnetka, transkripciji in kontroli transkripcije ter uredniške tipološke oznake dokumenta. Prvi del, ki vsebuje podatke o zvočnem posnetku in okoliščinah snemanja, izpolni oseba, ki besedilo snema; vpis mora biti opravljen takoj po snemanju oz. kakor hitro je mogoče. Podatke o okoliščinah snemanja potrebuje urednik korpusa za določitev vrste in tipa besedila, prav tako podatke o številu aktivnih udeležencev in poslušalcev. Urednik korpusa zbirni list podatkov besedila skupaj s posnetkom v elektronski obliki odda transkriptorju, ki izpolni drugi sklop podatkov. Zadnji sklop podatkov vnese urednik sam, ko na podlagi podatkov zbirnega lista in po potrebi

[185] Načeloma bi morali biti za vsakega aktivnega govorca priloženi Identifikacijski listi in Dovoljenja za uporabo, sicer besedilo ne bi smelo biti vključeno v korpus.
[186] Tudi identifikacijsko oznako transkripcije določi urednik korpusa.
[187] Zadnje štiri kategorije so določene na podlagi besedilnovrstnih kriterijev za zajem besedil v KGS, Tabela 14, str. 87.

tudi ponovnega poslušanja posnetka in pregleda transkripcije določi tip besedila, prevladujočo strukturo, okoliščine in prenosnik. Vsi podatki se kasneje vnesejo v glavo dokumenta. Nekateri podatki lahko služijo tudi kot kriterij za iskanje po korpusu (prevladujoča struktura, okoliščine, prenosnik, tip besedila), njihova dostopnost pri iskanju po korpusu pa je odvisna od nadaljnjih uredniških odločitev glede namembnosti korpusa, izbire konkordančnika itd.

7.3.2 Dokumentacija o govorcih

V govornih korpusih se običajno vodi natančno evidenco demografskih podatkov govorcev, saj so ti pogosto pomemben kriterij za zajemanje v korpus; podatki lahko kasneje služijo tudi kot izhodišča za korpusne analize. Informacije o govorcih vedno zbira oseba, ki govorjeno besedilo snema, in sicer tako, da izpolni identifikacijske liste za vsakega govorca posebej (tudi zase); najbolje je, če identifikacijske liste izpolnijo kar govorci sami. Tudi za načrtovano gradnjo korpusa govorjene slovenščine bodo demografski podatki o govorcih predvidoma pomembni, v fazi načrtovanja korpusa pa se je treba odločiti, kateri demografski podatki so za slovenščino relevantni; o tem je tekla razprava v poglavju 3.3, *Predlog priporočil za zajem besedil v KGS*. Ko se odločimo, katere informacije želimo zbirati, moramo oblikovati vprašanja; odgovori so dveh tipov: zaprtega (kjer gre za omejeno število možnih odgovorov, vprašani izbira med vnaprej pripravljenimi odgovori, npr. spol ali regija govorca) in odprtega tipa (kjer je število možnih odgovorov nepredvidljivo, npr. poklic; v tem pogledu lahko kasneje urednik razvršča poklice v določene skupine, ali pa pusti odprte vse možnosti). V nadaljevanju je predstavljen identifikacijski list enega izmed govorcev učnega korpusa.

IDENTIFIKACIJSKI LIST GOVORCA

Identifikac. oznaka (ID) govorca:[188] _____
Identifikac. oznaka (ID) posnetka: _____

1. *Za vsakega aktivnega govorca mora biti izpolnjen identifikacijski list.*
2. *Vsak govorec mora biti takoj po končanem snemanju in po dodatnih pojasnilih zaprošen, da podpiše Dovoljenje za uporabo; če govorec dovoljenja ne podpiše, besedilo ne bo vključeno v korpus.*
3. *Vsi navedeni podatki so zaupne narave in bodo uporabljeni izključno v znanstveno-raziskovalne namene; vsem sodelujočim v pogovoru je zagotovljena popolna anonimnost, ker bodo osebni podatki izbrisani iz korpusa, tako iz posnetega kot iz transkribiranega besedila.*

(Izpolni, kdor je besedilo posnel)
Besedilo posnel: *I. F, N. D.*
Datum snemanja: *2004*

PODATKI O GOVORCU

Ime (brez priimka):	Natasa		
Spol:	ženski ✓	moški	
Leto rojstva:	*1976*		
Dosežena izobrazba:	osnovna	srednja	višja/visoka/univ ✓
Poklic/delo, ki ga opravlja govorec:	*urednica*		
Regija, kjer živi:[189]	Lj. z okolico	Dolenjska	Notranjska
	Gorenjska	Koroška	Štajerska
	Prekmurje	Primorska	Drugo: *Zasavje*[190]

Ali je kdaj živel v drugi regiji ali kje drugje? Če ja, kje _____ in kako dolgo _____ .

Prvi jezik:	slovenščina ✓	drugo: _____

Odnos do drugih udl. v pogovoru:	sorodnik	partnersko razmerje		
	prijatelj	znanec	sodelavec	neznanec ✓
	nadrejeni	podrejeni	stranka	
	drugo: _____			

Najlepša hvala za sodelovanje!

Slika 38: Identifikacijski list govorca G14 v UKGS

[188] Identifikacijsko oznako govorcu dodeli urednik korpusa.
[189] Kasneje sem spremenila oznake za regije, prim. Tabelo 12, str. 81.
[190] Zanimiva napaka pri izpolnjevanju obrazca: rubrika Drugo je bila mišljena za govorce, ki ne živijo v Sloveniji, oseba, ki je izpolnjevala identifikacijski list, pa svoje regije med navedenimi odgovori očitno ni našla in je vprašanje razumela drugače.

Tudi podatki o govorcih se transformirajo v glavo dokumenta, o njihovi dostopnosti za uporabnike korpusa pa odločajo uredniki korpusa.

7.3.3 Avtorizacija posnetkov in transkripcij

Pri gradnji korpusa je treba poskrbeti za pridobitev dovoljenj govorcev za uporabo posnetkov in transkripcij, drugače korpusnih podatkov ne bo mogoče splošno uporabljati. To pomeni, da je treba pridobiti pisna dovoljenja tako za uporabo zvočnih posnetkov kot za uporabo transkripcij. To velja samo za posnetke, ki so posneti v osebnem stiku ali po telefonu; pri posnetkih, ki so pridobljeni iz medijev (radio in TV), zadošča pridobitev dovoljenja medijske hiše. Kadar gre za dvomljive primere, komu so prepuščene avtorske pravice posnetka, je načeloma bolje pridobiti več kot manj dovoljenj.

Spodaj je primer dovoljenja za uporabo iz UKGS:

DOVOLJENJE ZA UPORABO – odstop avtorskih pravic

Podpisani/-a odstopam vse pravice za uporabo zvočnih posnetkov, na katerih je posnet moj govor, in vseh transkripcij teh posnetkov, v izključno uporabo za gradnjo učnega korpusa govorjene slovenščine.

S podpisom dajem dovoljenje, da se posnetki, na katerih sodelujem s svojim govorjenjem, in transkripcije teh posnetkov, uporabijo na način, kot mi je bilo pojasnjeno ob podpisu. Zagotovljeno mi je bilo, da bodo govorci v korpusu popolnoma anonimni in da se bodo posnetki in transkripcije uporabljali izključno v znanstvene in raziskovalne namene. Uporaba posnetkov in transkripcij v komercialne namene ni dovoljena, je pa v te namene dovoljena uporaba raziskovalnih izsledkov, ki izhajajo iz analiz korpusa.

IME IN PRIIMEK: _____

(VELIKE TISKANE ČRKE)

PODPIS: _____ Datum: _____

Transkripcija govora bo jezikoslovno označena, nato pa bosta posnetek in transkripcija računalniško obdelana in dodana v učni korpus govorjene slovenščine. Posamezna transkribirana besedila in korpus bodo dostopni

samo v elektronski obliki, shranjeni pa na spletnih straneh korpusa in z geslom dostopni raziskovalcem. Gesla raziskovalcem dodeljuje urednik korpusa.

Sodelavci korpusa upamo, da boste dovoljenje podpisali in nam ga izročili ter s tem prispevali k razvoju nacionalno pomembnega jezikovnega vira – korpusa govorjene slovenščine. Hvala!

Filozofska fakulteta Univerze v Ljubljani *Jana Zemljarič Miklavčič, urednica UKGS*

Slika 39: Dovoljenje za uporabo posnetkov in transkripcij v UKGS

S podpisom dokumenta govorci odstopijo avtorske pravice svojega govora in transkripcij uredniku oz. lastniku korpusa. Pri gradnji učnega korpusa sem za dovoljenje zaprosila približno polovico govorcev;[191] govorci, ki pristanejo na snemanje, s podpisom dovoljenja načeloma ne delajo težav. Pri posnetkih, ki so bili narejeni za predvajanje v javnosti (na radiu, TV, na CD-romu, ki je del učbenika itd.), dovoljenj za objavo s strani posameznih govorcev nisem zbirala, treba pa bi bilo dobiti dovoljenje lastnika posnetka, npr. Radia Slovenija.

7.3.4 Podatki v glavi dokumenta

Na podlagi zbirnih listov posnetkov in identifikacijskih listov govorcev uredniki izdelajo (in izpolnijo) glave posameznih dokumentov. Vanje shranijo podatke o posnetku, transkripciji, taksonomski uvrstitvi besedila in govorcih. Predstavljen je primer glave dokumenta iz UKGS:

```
DOKUMENT R02

Posnetek
Identifikacijska oznaka posnetka: R02
Delovno ime posnetka: Klepet v službi
```

[191] Nekateri posnetki so nastali prej, preden je bilo pripravljeno Dovoljenje za uporabo – odstop avtorskih pravic. .

Trajanje: **7.31 min**
Posnel(-a): **Jana Zemljarič M.**
Datum snemanja: **september 2004**
Kraj snemanja: **Ljubljana, pisarna**
Okoliščine: **odmor, klepet sodelavcev ob kavi**
Število govorcev: **5**
Posneto na skrivaj: **NE**
Opombe: **veliko prekrivnega govora več kot dveh oseb**

Transkripcija
Identifikacijska oznaka transkripcije: **TransPraat_R02**
Transkribiral/-a: **Jana Zemljarič M.**
Datum (dokončane) transkripcije: **27. oktober 2004**
Opombe:
Pregled transkripcije
Pregledal/-a:
Datum:
Opombe:

Kategorija besedila
Tip:[192] *spontantana konverzacija*
Prevladujoča struktura besedila: **multilog**
Okoliščine: **nejavno, nezasebno**
Govorni položaj: **neformalni**
Prenosnik: **osebni stik**

Govorci

ID	Spol	L.r.	Izobr.	Regija	Prvi j.	Odnos do posl.	Poklic
G03	ž	1966	U	Ljo	slov	neformalni	strok. delavec
G01	ž	1963	U	Ljo	slov	neformalni	strok. delavec
G09	ž	1979	U	D	slov	neformalni	strok. delavec
G10	ž	1967	S	G	slov	neformalni	admin. delavec
G15	ž	1979	U	Ljo	slov	neformalni	strok. delavec

Slika 40: Primer glave dokumenta iz UKGS

[192] Po vnaprej določenem naboru.

7.4 ZAKLJUČEK

Dokumentacija posnetkov in transkripcij za učni korpus govorjene slovenščine je bila dovolj dobro pripravljena, da jo je mogoče z manjšimi modifikacijami uporabiti tudi pri morebitni gradnji referenčnega korpusa govorjene slovenščine. Spremembe se nanašajo samo na spremenjene kriterije za zajem besedil v korpus; tako npr. se na Identifikacijskem listu govorca v rubriki regijski izvor zamenja klasično delitev na narečne skupine z novo delitvijo na pet regijskih skupin (prim. pogl. 3.3.2, *Konverzacijski podkorpus*). Za ustrezna sta se izkazala tudi Zbirni list podatkov besedila in Dovoljenje za uporabo – odstop avtorskih pravic.[193] Na podlagi navedenih dokumentov je mogoče sestaviti informativno glavo posameznega korpusnega dokumenta, ki podpira različne analize govorjenega jezika.

[193] Tega bi morali ob morebitni gradnji korpusa pregledati še pravniki.

8 Gradnja učnega korpusa govorjene slovenščine

Gradivo učnega korpusa govorjene slovenščine sem že v prejšnjih poglavjih uporabljala za ilustracijo teoretičnih izhodišč. V tem poglavju pa bo UKGS predstavljen kot samostojna aplikacija, in sicer najprej njegova gradnja in lastnosti.

8.1 ZBIRANJE GRADIVA

Gradivo za korpus sem zbirala že pred odhodom na Norveško, kjer je korpus nastal: en posnetek sem z digitalno snemalno napravo Sony ICD-MS515 naredila sama, tri posnetke, narejene z isto napravo za druge namene, mi je odstopila Ina Ferbežar, en posnetek pa sem dobila z Radia Slovenija. Ker se je v Bergnu izkazalo, da imam gradiva premalo, so mi v Ljubljani s snemalno napravo naredili tri dodatne posnetke spontanega govora, od katerih sta bila dva vključena v UKGS; enega je posnel Žiga Rangus, drugega Meta Lokar. Tako je bilo v korpus vključenih 7 posnetkov, ki so nastali v obdobju od maja 2004 do oktobra 2004, nekateri prav z namenom vključitve v korpus (3 posnetki), drugi pa z drugimi nameni (4 posnetki). Posnetke sem v celoti transkribirala in označila v Bergnu; o načelih transkribiranja in označevanja sem se, kolikor je bilo mogoče, posvetovala z mentorjem, prof. dr. Markom Stabejem, pa tudi s prof. dr. Bredo Pogorelec. Konzultacije na daljavo so bile za nastajanje učnega korpusa izjemnega pomena, razumljivo pa je, da so bile precej omejene. Tudi pomoč sodelavcev iz Bergna je lahko potekala, vsaj kar zadeva transkribiranje slovenskega govora, predvsem na načelni ravni.

8.2 ZAJEM BESEDIL

Pri gradnji učnega korpusa zaradi omejene količine gradiva in časa nisem mogla v celoti upoštevati vseh načel za zajem besedil, da bi bil korpus uravnotežen, kar navsezadnje niti ni bil moj namen. Kljub temu sem se pri zbiranju gradiva trudila, da so se besedila čim bolj razlikovala po lastnostih, na katerih je zgrajena taksonomija govorjenih besedil, da je bilo dovolj spontanega dialoga in da so se tudi govorci razlikovali po demografskih lastnostih. Pri posnetkih, ki so bili narejeni z namenom vključitve v korpus, so bili izdelani tudi zbirni listi besedil in listi z identifikacijskimi podatki govorcev, pri preostalih besedilih pa so bili naknadno zbrani podatki, ki so bili dostopni. Dokumentacija posnetkov in popis govorcev UKGS sta predstavljena v nadaljevanju.

8.2.1 Dokumentacija posnetkov

Učni korpus govorjene slovenščine sestavlja 7 posnetkov z opisanimi karakteristikami:

ID[194]	Trajanje min	Št. govorcev	Datum snemanja	Kraj snemanja	Lokacija snemanja	Okoliščine	Tajnost snem.
R01	2.17	2	maj 2004	Ljubljana	na univerzi	primer besedila za šol. učbenik	ne
R02	54.50	6	17. 7. 2004	Ljubljana	v radijskem studiu	radijska oddaja	ne
R03	3.58	2	maj 2004	Ljubljana	doma	primer besedila za šol. učbenik	ne
R04	7.31	5	9. 9. 2004	Ljubljana	v pisarni	klepet sodelavcev	ne
R05	3.23	5	junij 2004	Ljubljana	na parkirišču	primer besedila za šol. učb., sleng	ne
R06	11.54	3	24. 10. 2004	Ljubljana	v kavarni na prostem	prijatelja na kavi	ne
R07	5.12	2	november 2004	Kranj	doma	prijateljica na obisku	da

Tabela 17: Dokumentacija posnetkov UKGS

Posnetki so različno dolgi, njihova skupna dolžina je 89 minut. Dolžina posnetkov je prilagojena naravnemu poteku dogodkov; besedila so posneta bodisi v celoti, bodisi so prekinjena na mestu, ki ga ne občutimo kot konec dejanskega besedila (npr. v primeru, ko je na snemalni napravi zmanjkalo spominskega prostora, zmanjkalo baterij ali pa je snemalec zaradi kakšnih zunanjih okoliščin prekinil snemanje).

Iz tabele lahko razberemo, da je v posameznih posnetkih sodelovalo od 2 do 6 govorcev. To sicer pomeni, da med besedili ni bilo nobenega pravega monologa, vendar pa imata dva izmed posnetkov prevladujočo monološko strukturo.[195] Časovni razpon nastanka besedil je 6 mesecev. Med pomanjkljivosti UKGS pa vsekakor sodi njegova demografska sestava, saj so bila vsa besedila razen enega posneta v Ljubljani, pa tudi večina govorcev je bila iz Ljubljane in okolice.

[194] Identifikacijska oznaka posnetka.
[195] Pojem "prevladujoča monološka struktura" je razložen v poglavju 3.3.3, *Besedilnovrstni podkorpus*.

Okoliščine snemanja v sedmih posnetkih so bile precej različne, kar je obetavna podlaga za raznovrstnost govorjenih besedil v korpusu. Samo eno besedilo je bilo posneto na skrivaj, sicer pa so vsi govorci za snemanje vedeli; njihove reakcije so bile, kadar je šlo za snemanje spontanega dialoga, nekoliko zadržane ali negativne:

G01: <shift=vpr> [kaj je] zdaj to snema </shift=vpr>
G09: no niti ne bom [govorila]

Pri tem je treba upoštevati, da so bili vedno seznanjeni z namenom snemanja in da so bili skoraj vsi govorci v prijateljskem ali kolegialnem odnosu z »urednico korpusa«, kar je navsezadnje pretehtalo pri pristanku na snemanje. Sicer pa govorci snemanja (po)govora ne sprejemajo vedno zlahka, na kar morajo biti načrtovalci korpusa in snemalci pripravljeni.

G17: a to za faks ali kaj
G16: ne ena kolegica me je prosila da ji posnamem
G17: zdaj se snema {nv} smeh {/nv} (4 sec)
{smeh} ne moreš skriti {/smeh}
G16: ne
G17: {shift=vpr} ja in zakaj {/shift=vpr}
G16: doktorat dela pa rabi posnetke ne {pavza} a veš {pavza}
pa je prosila če lahko {pavza}

8.2.2 Popis govorcev

UKGS vsebuje govor 20 govorcev. Nekateri govorci se pojavijo v več različnih posnetkih, vendar so v vseh označeni z isto šifro. Pri gradnji večjega govornega korpusa bi se lahko zgodilo, da bi se ista oseba večkrat pojavila kot govorec v različnih posnetkih, pa to ne bi bilo ugotovljeno, zato bi ji bile znotraj korpusa pripisane različne identifikacijske oznake; tega verjetno ni mogoče preprečiti. Demografske lastnosti govorcev UKGS predstavlja naslednja tabela:

Št.	ID	Spol	Leto roj.	Starost	Izobrazba	Regija
1.	G01	Ž	1964	40	U	Ljo
2.	G02	M	1965	39	U	Ljo

3.	G03	Ž	1966	38	U	Ljo
4.	G04	Ž	1967	37	U	Ljo
5.	G05	Ž	1968	36	U	Ljo
6.	G06	Ž	1968	36	U	Ljo
7.	G07	M	1970?[196]	34?	U	Drugo
8.	G08	M	1933?	71?	U?	Ljo
9.	G09	Ž	1979	25	U	D
10.	G10	Ž	1967	37	S	G
11.	G11	M	sr. šola	17?	O	Ljo
12.	G12	M	sr. šola	17?	O	Ljo
13.	G13	M	sr. šola	17?	O	Ljo
14.	G14	Ž	1976	28	U	D
15.	G15	Ž	1979	25	U	Ljo
16.	G16	M	1978	26	S	Ljo
17.	G17	M	1978	26	S	Ljo
18.	G18	Ž	?[197]	?	?	?
19.	G19	Ž	1969	35	U	G
20.	G20	M	1948	56	S	G

Tabela 18: Dokumentacija govorcev UKGS

V UKGS sem zajemala govorce po spolu, starosti, izobrazbi in regiji, iz katere izhajajo. Regije sem označevala po tradicionalni narečni razdelitvi slovenskih ozemelj, ki sem jim dodala regijsko skupino Ljubljana z okolico (Ljo); za razdelitev na pet regijskih skupin sem se odločila kasneje.

8.3 Karakteristike UKGS

Med lastnosti korpusa sodijo njegova velikost in sestava glede na vnaprej določene demografske in besedilnovrstne kriterije. Po teh lastnostih vrednotimo tudi uravnoteženost korpusa glede na izbrano celotno produkcijo. Glede na predlagane demografske kriterije (Tabela 12, str. 81) lahko znotraj učnega kor-

[196] Kjer je pri podatku naveden vprašaj, podatek ni popolnoma zanesljiv.
[197] Kjer je namesto podatka naveden vprašaj, podatek ni znan in o njem ni mogoče niti ugibati; to se npr. zgodi, ko se v posnetek nepričakovano vključi neznana oseba, kot se je to zgodilo pri posnetku R06.

pusa govorjene slovenščine identificiramo naslednje skupine govorcev:

Kriterij	Porazdelitev glede na določene kategorije						?
Spol	ženske: 11			moški: 9			–
Starost	manj kot 35: 9			35 ali več: 10			1
Regija	LjO: 13	SZ: 0	SV: 3	Z: 0	J in JZ: 2	drugo: 1	1
Izobrazba	končana OŠ: 3		končana SŠ: 4		viš. šola ali več: 12		–
Prvi jezik	slovenski: 19			drugo: 1			–

Tabela 19: Porazdelitev govorcev v UKGS glede na izbrane demografske kriterije

Vidimo lahko, da je učni korpus vsaj približno uravnotežen glede na porazdelitev moških in žensk; v celotni populaciji vseh (ne samo odraslih) Slovencev je ta porazdelitev 48,8 proti 51,2 odstotka,[198] v učnem korpusu pa 45 proti 55 odstotkov. Solidna je razporeditev govorcev glede na starost, pa tudi na prvi jezik govorcev (v celotni populaciji za 12,3 odstotka govorcev slovenščina ni prvi jezik, v učnem korpusu za 5 odstotkov).[199] Neuravnotežena je porazdelitev govorcev glede na izobrazbo, saj ima v celotni populaciji le 15 odstotkov odraslih (nad 15 let) prebivalcev doseženo več kot srednjo izobrazbo,[200] v učnem korpusu pa kar 60 odstotkov. Za demografsko komponento korpusa, kjer je izobrazba eden izmed kriterijev za zajemanje besedil, je to nesprejemljivo visok odstotek, vendar pa v celotnem korpusu lahko pričakujemo, da bo odstotek ljudi z visoko izobrazbo višji kot samo v demografski komponenti, ker ljudje z višjo izobrazbo predvidoma producirajo večje število govorjenih besedil, ki dosegajo množično recepcijo. Najnižjo stopnjo uravnoteženosti učni korpus dosega glede na regijsko porazdelitev govorcev; vse predvidene regije sploh niso zastopane, kaj šele, da bi bile zastopane v uravnoteženem razmerju. Oportunistični način zajemanja besedil glede na regijsko pripadnost govorcev ima za posledico neuravnoteženost učnega korpusa v tem pogledu.

Učni korpus največjo uravnoteženost glede na celotno populacijo govorcev izkazuje po kriteriju spola. Vprašanje pa je, kaj se z načrtovanimi razmerji zgodi ob gradnji korpusa. Kakšno razmerje bi dobili, če bi npr. prešteli vse besede, ki jih v korpusu dejansko izgovorijo moški, in vse, ki jih izgovorijo ženske? Verjetno

[198] Popis prebivalstva RS za l. 2002, http://www.stat.si/popis2002/si/rezultati/rezultati_red.asp?ter=SLO&st=2.

[199] Vendar pa teh 5 % (to je en sam govorec, katerega prvi jezik je katalonščina) seveda nima reprezentativnega prvega jezika glede na celotno populacijo.

[200] Popis prebivalstva RS za l. 2002, http://www.stat.si/popis2002/si/rezultati/rezultati_red.asp?ter=SLO&st=10.

bi se odmaknili od izhodiščne uravnoteženosti. Enak razplet lahko pričakujemo tudi pri morebitni gradnji korpusa govorjene slovenščine. Vendar pa se iz tega pravzaprav naučimo, kako je treba v resnici razumeti pojem uravnoteženosti korpusa. Gre za izhodiščni načrt strukture korpusa, ki zagotavlja, da se bo v korpusu pojavilo dovolj različnih besedil različnih govorcev, ne moremo pa pričakovati, da se bodo izhodiščna razmerja ohranila tudi v končni podobi korpusa.

Glede na predlagane besedilnovrstne kriterije lahko v učnem korpusu opazujemo naslednjo strukturo:

ID	Trajanje	Št. pojavnic	Št. govorcev	Prevladujoča struktura	Okoliščine	Govorni položaj	Prenosnik	Vrsta besedila
R01	2. 17	275	2	monolog	javno	formalni	CD	intervju
R02	54.50	8583	5	multilog	javno	formalni	radio	okrogla miza
R03	3.58	572	2	monolog	javno	formalni	CD	intervju
R04	7.31	1681	5	multilog	nejavno	neformalni	osebni stik	spontana konverz.
R05	3.23	650	5	multilog	javno	(ne-) formalni	CD	intervju
R06	11.54	2024	3	dialog	zasebno	neformalni	osebni stik	spontana konverz.
R07	5.12	832	2	dialog	zasebno	neformalni	osebni stik	spontana konverz.

Tabela 20: Porazdelitev besedil v UKGS glede na izbrane besedilnovrstne kriterije

Razmerja znotraj posameznih besedilnovrstnih kriterijev, če za enoto merjenja upoštevamo število besed, so naslednja:

- dialogi (in multilogi) proti monologom 94 % : 6 %
- zasebna besedila proti javnim 19,5 % : 80,5 %
- neformalni govorni položaj proti formalnemu 35,5 % : 64,5 %
- besedila, posneta v osebnem stiku, proti besedilom s prenosnikom 31 % : 69 %
- skrivaj posneta besedila proti ostalim 5,6 % : 94,4 %

Pri transkribiranju UKGS se je potrdilo znano dejstvo, da je tudi ortografsko transkribiranje zelo zamudno delo. Pri zapisovanju posnetkov monologov, posebej z govorci, ki so vešči javnega nastopanja, sem dosegla največjo učinkovitost

transkribiranja, in sicer sem 1 minuto posnetka transkribirala in označevala 15 minut. Pri posnetkih spontanega multiloga, kjer je govorilo več oseb hkrati, pogosto nisem dosegla niti transkripcijske hitrosti 1 minuto posnetka v 1 uri. Moja povprečna hitrost transkribiranja je bila eno uro za 2 minuti posnetka, skupno število ur transkribiranja pa okrog 50.

Zdi se, da je dosežena velikost korpusa (15.000 pojavnic) zadoščala za določitev kriterijev za zajem besedil v korpus in za priporočila za nabor oznak za označevanje govorjenih besedil, zadovoljiva pa je bila tudi za določitev transkripcijskih standardov. V 100 minutah se je namreč zgodila večina pričakovanih[201] verbalnih in neverbalnih dogodkov. Prav tako je bilo mogoče z dokajšnjo zanesljivostjo izdelati nabor za oznake v glavi besedila, kar je povratno vplivalo na izboljšavo identifikacijskega lista govorcev in na zbirni list podatkov besedila. Žal pa se v danih razmerah ni bilo mogoče bolj približati uravnoteženosti korpusa. Učni korpus, čeprav nereprezentativen, vseeno ima svojo pragmatično funkcijo in je uporaben, če je metoda zbiranja besedil transparentna in če so dostopni podatki o lastnostih besedil; neuravnoteženo sestavo je treba upoštevati pri interpretaciji korpusnih podatkov. Vsekakor pa je gradnja učnega korpusa pomemben korak v procesu gradnje vsakega večjega korpusa.

8.4 TRANSKRIBIRANJE UKGS

Pri transkribiranju UKGS sem upoštevala priporočila TEI in EAGLES. Izjava <u> je omejena bodisi s premorom bodisi z menjavo govorcev. Načeloma lahko rečemo, da so izjave pogosto smiselne vsebinske enote. Meje med izjavami sem vedno poskušala postaviti tako, da bi bile izjave smiselne. To načelo je bilo kršeno v primerih, ko je govorilo več govorcev hkrati. Včasih meje ni bilo mogoče postaviti drugače kot z nasilno prekinitvijo vsaj enega izmed govorcev; v takem primeru sem prekinila izjavo tistega govorca, katerega izjava se je slabše slišala.

Izjave so transkribirane v ortografski transkripciji brez ločil; velike začetnice imajo samo lastna imena. V UKGS sem besede in oblike, ki imajo ustaljeno pisno podobo, skoraj dosledno zapisovala z upoštevanjem norme knjižnega jezika: tudi če je bil npr. nedoločnik izgovorjen reducirano, sem ga zapisala v pričakovani pisni obliki, torej s končnim –i, enako pri določnih pridevnikih, deležnikih itd. To načelo sem v poglavju 5.6, *Končni predlog priporočil za transkribiranje govorjene slovenščine,* spremenila in poskušala zapis bolj prilagoditi govoru. Transkribiranje za UKGS je tako potekalo brez težav, če so izgovorjene besede imele znano pisno

[201] S tem mislim vse dogodke, za katere oznake predvideva npr. govorna skupina EAGLES.

(knjižno) obliko. Težave so se pojavile na tistih mestih, kjer besede niso imele ustaljene pisne oblike ali so od nje v govoru zelo odstopale; v praksi to pomeni tuje (citatne) besede, narečne in izrazito pogovorne besede ter besede, ki v govoru močno odstopajo od knjižne oblike. Pri zapisovanju teh besed je bilo treba iskati drugačne rešitve.

Težave sem reševala z različnimi oznakami. Še najmanj problematično je bilo zapisovanje citatnih besed (tujih besed, popolnoma neprilagojenih slovenščini), ki so označene s

- <tj: norv>*besedilo*</tj>,

kar naj bi pomenilo, da gre za tuji jezik, za dvopičjem pa je jezik tudi naveden. Konkretni zgledi iz korpusa so:

- `čakaj zdaj pa še <tj: norv>norge</tj>`
- `<tj: hrv>imamo mi posla i bez toga</tj>`
- `<tj:katalon><?>karradera</?>/tj>.`[202]

V poglavju *Predlog priporočil za označevanje besedil v govornem korpusu*, Tabela 16, sem opustila navajanje tujega jezika, v katerem je citat naveden; po eni strani zato, ker ga predvidoma ne bo mogoče vedno identificirati, po drugi strani pa ta podatek transkripcijo precej dodatno obremeni.

Bolj problematično je bilo označevanje besed, ki nimajo ustaljene pisne podobe v knjižnem jeziku. Te besede so v UKGS dobile oznako

- <nst>*beseda*</nst>.

Oznako so dobile vse besede, ki so v SSKJ označene s kvalifikatorji kot pogovorne, nižje pogovorne, narečne, vulgarne itd., poleg tega pa še vse tiste besede ali oblike, ki jih ni v slovarju, npr.

- `jezus marija smo <nst>trajbali</nst>`

[202] Oznaka <?> pomeni, da gre za negotovo transkripcijo; transkriptor ni prepričan, če je pravilno zapisal besedo.

Spodnji izsek iz UKGS prikazuje del besedišča, označenega kot <nst>:

```
_ 🔊 +   samo brez velike- velikega  <nst> angažmaja
_ 🔊 +              [dražje] +G16+ [<nst> benz </nst>]
_ 🔊 +       [ma] ja kje sem imela  <nst> cajt </nst>
_ 🔊 + na univerzi v Bergnu] +G03+  <nst> carsko </nst>
_ 🔊 + pa čisto različne od recimo  <nst> dekov </nst>
_ 🔊 +  [<shift=vpr> temu se reče]  <nst> dila </nst>
_ 🔊 +          +G11+ temu se reče [<nst> dila </nst>]
_ 🔊 +   +G19+ samo to moraš hitro  <nst> dofilati
_ 🔊 +     (ropot mikrofona) +G03+ [<nst> evo </nst> ga
_ 🔊 +  +G03+ [<nst> evo </nst> ga  <nst> evo </nst>
_ 🔊 +        mi je že prinesla pa  <nst> evo </nst> ti
_ 🔊 +            +G03+ [edino] na  <nst> faksu </nst>
_ 🔊 +         <neraz> in jih s tem  <nst> fila </nst>
_ 🔊 +       je možno da ima kakšne  <nst> fore </nst>
_ 🔊 +        direktorček] ja +G16+ [<nst> frej </nst>]
_ 🔊 +        +G03+ [<neraz>] +G09+  <nst> ful </nst> je
_ 🔊 +   slabo mislim pokrajina je  <nst> ful </nst>
_ 🔊 +   kam pa potem] </??> +G03+  <nst> ful </nst>
_ 🔊 +  ful </nst> je veter [pihal  <nst> ful </nst>
_ 🔊 +    </nst> <nst> ful </nst>]  <nst> ful </nst>
_ 🔊 +     [pihal <nst> ful </nst>  <nst> ful </nst>]
_ 🔊 +   večinoma se <?> park </?>  <nst> fura </nst>
_ 🔊 +  rajderji </nst> ki za njih  <nst> furajo </nst>
_ 🔊 +     evo </nst> ti <smeh> pa  <nst> glih </nst>
_ 🔊 +           (2) mislim sem to  <nst> glih </nst>
_ 🔊 +       [mislim] ne glede na  <nst> glih </nst>
_ 🔊 + telefon </shift=vpr>] +G01+  <nst> glih </nst>
_ 🔊 +         pa fora +G16+ [ja]  <nst> glih </nst> v
_ 🔊 +           [<nv> smeh </nv>]  <nst> glih </nst>
_ 🔊 +      zaciklali </nst> v eno  <nst> gužvo </nst>
_ 🔊 +    <??> v <nst> iber </nst>  <nst> hudih </nst>
_ 🔊 +      [ja take] +G12+ <??> v  <nst> iber </nst>
_ 🔊 +           +G03+ [mhm] +G01+  <nst> jabki </nst>
_ 🔊 +        šalamuna ma ja da mu  <nst> jebala </nst>
```

Slika 41: Del besedišča z oznako <nst>

Označevanje z oznako <nst> v UKGS je bilo nepotrebno; nastalo je kot posledica globoko ukoreninjenosti ustaljene podobe (knjižnega) pisnega jezika v zavest transkriptorja (v tem primeru mene). Vsekakor sem na začetku imela težave zapisovati izrazito pogovorne besede in oblike. Šele sčasoma sem dokončno sprejela dejstvo, da gre za inherentne sestavine govora, ki jih z zapisom samo dokumentiramo, vsako označevanje oz. dodajanje kvalifikatorja pa že pomeni interpretacijo, kar je v nasprotju z namenom gradnje korpusa. Zato oznake <nst> nisem uvrstila na seznam oznak v Tabeli 16, *Priporočen nabor oznak za označevanje KGS*.

V nadaljevanju predstavljam še nekaj primerov, ki so mi pri transkribiranju povzročali težave:

- `mogoče imam celo <nst>kle</nst>`
- `no glej <nst>klele</nst> je Bergen`

V zgornjem primeru je nemogoče zapisati slovarsko ustreznico besed *kle* in *klele*, zato ju zapišemo, kot ju slišimo, in uvrstimo na seznam dovoljenih pogovornih besed.

Sestavljavci govornih korpusov pogosto poročajo o težavah pri določitvi besednih mej. V slovenščini se to vsaj v UKGS ni pogosto dogajalo, je pa vsaj en izrazit primer, ko se je treba odločiti o besedni meji:

- `ta drobna zemlja`
- `to je izhodišče za vse te ta glavne fjorde`

Vprašanje je, ali bomo določni člen zapisovali skupaj z besedo, ki jo določa, ali bo vmes presledek. V govoru gre gotovo za eno enoto, pisna tradicija, mdr. tudi SSKJ (V, 1991, 11), pa narekuje pisanje narazen. V prid pisanja narazen govori tudi posledično boljša možnost za iskanje po korpusu.

Nekaj težav mi je povzročalo tudi zapisovanje t. i. besedilnih aktualizatorjev (Smolej 2006 termin povzema po Vidovič-Muha 1996), to je izrazov, ki »pretvarjajo (spreminjajo) slovarsko vrednost leksema, pred katerim stojijo) v konkretno referenco (pomen) oz. opravljajo vlogo usmerjevalcev/kazalcev na konkretnost/nekonkretnost oz. splošnost predstavljane reference« (Smolej 2006, 159). Smolej podrobneje obravnava besedilne aktualizatorje *(en) tak, tist, un, en, nek, tale* in *ta*, pri čemer ugotavlja, da so obravnavani aktualizatorji v zvezi s pridevniki in samostalniki v spontanem govoru zelo pogosti, da pa bi bile te zveze v pisnih besedilih,

ki sledijo knjižni normi, opazne in zaznamovane (Smolej 2006, 161). V UKGS najdemo mdr. naslednje primere:

- una rdeča zemlja
- si slišal kaj je un Michael
- pol pa un ves zgrožen
- pa un kurc kva je že
- vsak dan smo šli za ene dve ure hodit
- tam na enem stojalu
- glih pred ene dvema mesecema

Zdi se nemogoče, da bi zgornje deikte in zaimke zapisovali v skladu s pisno normo, ker bi se s tem preveč oddaljili od avtentičnega govora. Zato jih zapisujemo, kot jih slišimo, kar je bilo priporočeno že v poglavju *Končni predlog priporočil za transkribiranje govorjene slovenščine*. Pač pa lahko pričakujemo, da se besedilni aktualizatorji po različnih pokrajinskih govorih razlikujejo, česar v UKGS zaradi nediferencirane regijske sestave nisem mogla zaznati. Če primerjamo besedila UKGS z besedili korpusa TURDIS-1 pri Verdonik (2006), v katerem so vsi govorci iz štajerske regije, aktualizatorja *un* npr. ne najdemo.

Nenazadnje mi je težave povzročalo tudi zapisovanje oziralnega zaimka *ki* in veznikov *ko* in *ker*, ki so bili pogosto izgovorjeni kot *k*:

- tipično za njih k se vedno na demokracijo zgovarjajo
- sem gledu k zdej ta no kaj je že

V UKGS sem zapis k-ja prilagodila knjižni normi, vendar sem morala včasih o zapisu tudi ugibati, saj ni bilo mogoče vedno enoznačno ugotoviti, na katero besedo se k nanaša. Zato sem v poglavju *Končni predlog priporočil za transkribiranje govorjene slovenščine* predlagala, da se zapis k-ja prilagodi govoru.

Prikazanih je bilo samo nekaj težavnih mest, ki so se pojavila med transkribiranjem govorjenih besedil za UKGS. Tudi izčrpnejši popis problematike iz UKGS seveda ne bi zajel vseh težav, ki bi se utegnile pojaviti pri transkribiranju večjega govornega korpusa. Zato moramo pričakovati, da bo skupina načrtovalcev korpusa oz. uredniški odbor nastale težave reševal tudi sproti.

8.5 NABOR OZNAK UKGS

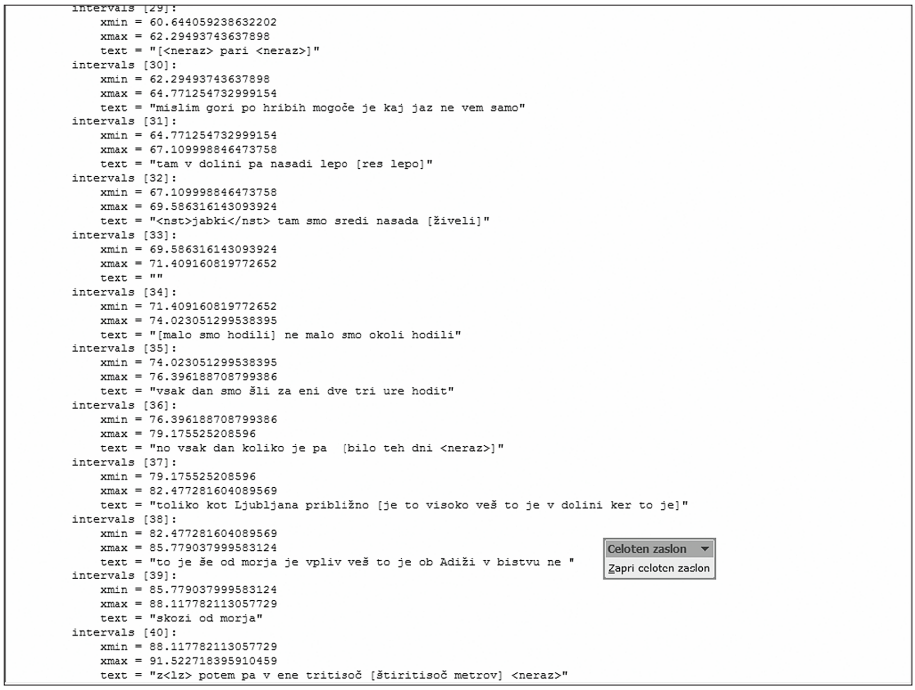
Spodnja tabela prikazuje nabor oznak, predvidenih in uporabljenih za UKGS. Tabela se nekoliko razlikuje od *Priporočenega nabora oznak za označevanje KGS* (Tabela 16); razlogi za to so bili navedeni zgoraj.

Oznaka	Pomen
<pavza>	kratka pavza (pribl. 1 sekundo)
<pavza>(5)	pavza (5 sekund)
<ime>	nadomešča lastno osebno ime
<priimek>	nadomešča priimek
<priimek><f>	nadomešča žensko obliko priimka
<naslov>	nadomešča osebni naslov
<telefonska številka>	nadomešča zasebno telefonsko številko
<neraz>	nerazumljivo
<neraz> (5)	nerazumljivi govor (5 sekund)
<?>*besedilo*</?>	nezanesljiva transkripcija
=	napačni začetek/okrnjena beseda
<repet>*besedilo* </repet>	ponovitev besede ali več besed
<nst>*beseda*</nst>	nestandardna beseda ali oblika
<tj: norv>besedilo</tj>	besedilo, izgovorjeno v tujem jeziku
<okr>*beseda*</okr>	kratica ali okrajšava
[*besedilo*]	prekrivni govor
<petje>*besedilo*</petje>	označevanje objezikovnih pojavov
<shift=vpr>*besedilo*</ shift=vpr>>	besedilo z vprašalno intonacijo
<shift=poud>besedilo</ shift=poud >	poudarjeno besedilo
<nv>smeh</nv>	neverbalni zvoki
(*opis dogodka*)	zvoki v ozadju/nekomunikacijski zvoki
 besedilo</br>	brano besedilo
<??>*besedilo*</??>	nezanesljivo, kdo je govorec

Tabela 21: Oznake UKGS

8.6 Konvertiranje

Transkripcije je v korpus s konkordančnikom in s povezavo med transkripcijami in zvočnimi posnetki konvertiral Knut Hofland na Univerzi v Bergnu. Obe transkripcijski orodji, *Praat* in *Transcriber*, omogočata funkcijo, ki vsakemu transkribiranemu segmentu (torej vsaki izjavi, ki jo omejuje menjava govorcev ali premor) na začetku in na koncu z natančnostjo 10^{-15} sekunde določi časovno oznako:

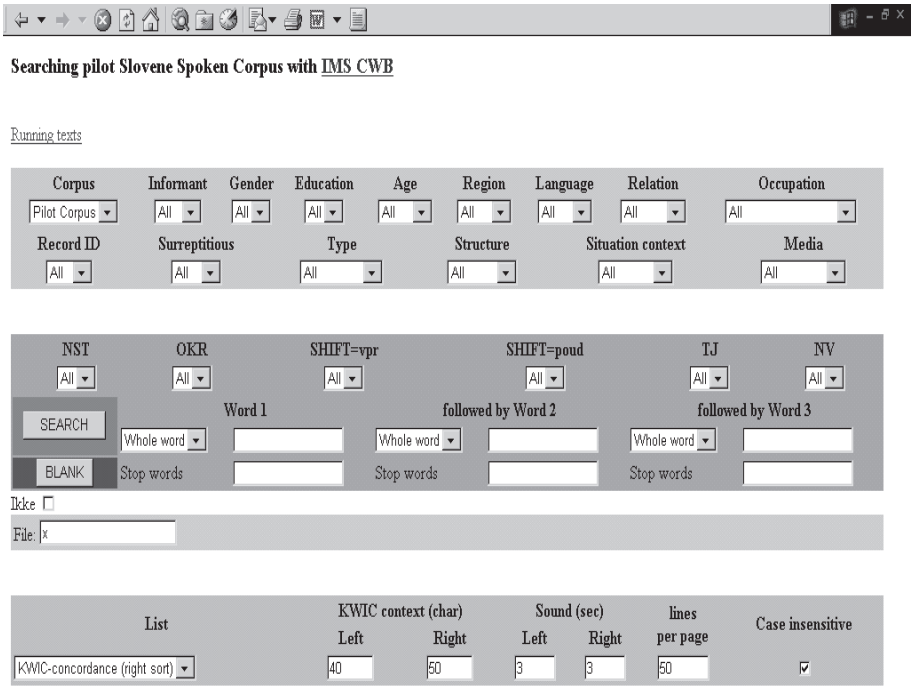


Slika 42: *WordPad* verzija transkripcije (iz *Praata*)

Program sam deli časovni odsek med začetkom in koncem izjave s številom besed in naredi interpolacijo časa za vsako besedo znotraj segmenta posebej; 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 natančno tistega, kar bi želeli.

Naslednja stopnja konvertiranja je bila narejena s programom Corpus Work Bench (CWB), ki so ga izdelali na inštitutu IMS (*Institut für Maschinelle Sprachverarbeitung*) v Stuttgartu. V tem programu so informacije o govorcih (demograf-

ski podatki) in okoliščinah snemanja (taksonomska uvrstitev besedila) dodane v posebnih stolpcih, kar v iskalnem oknu omogoča enostavno iskanje po izbranih kriterijih. Iskalno okno korpusa omogoča iskanje po različnih kriterijih, ki so bili predvideni ob načrtovanju korpusa. Kriterije, za katere so bili podatki zbrani z vpisovanjem v zbirne liste besedil in identifikacijske liste govorcev, transformirani v glavo besedila in konvertirani s CWB, lahko izbiramo in določamo v zgornjem osenčenem pasu okna, in sicer v zgornji vrstici demografske kriterije (spol govorca, izobrazbo, regijo, prvi jezik, odnos med govorci in poklic) in v spodnji vrstici besedilnovrstne kriterije (skrivaj posneta besedila, tip besedila, strukturo, okoliščine in prenosnik).



Slika 43: Iskalno okno UKGS

Besedo vtipkamo v iskalno polje v drugem osenčenem pasu, in sicer celo besedo, njen začetek, konec ali katerikoli vmesni del; okno omogoča iskanje do treh sosednjih besed ali delov besed (kolokacije).

8.7 REFLEKSIJA

Učni korpus govorjene slovenščine je bil najprej predstavljen v okviru študijskih obveznosti na Univerzi v Bergnu na Lingvističnem seminarju Oddelka za lingvistiko (3. 12. 2004) in na Oddelku za kulturo, jezik in informacijske tehnologije (10. 12. 2004); nastanek korpusa je opisan v poročilu na spletnih straneh štipendistov Marie Curie.[203]

V Ljubljani je bil UKGS najprej predstavljen na Jezikovnotehnološkem abonmaju JOTA[204] (15. 3. 2005). Sledila je predstavitev na 3. mednarodni konferenci SLOVKO 2005 v Bratislavi (*Computer treatment of Slavonic Languages*, 10.–12. november 2005), na 1. mednarodni fonetični konferenci SLOFON v Ljubljani (20.–22. april 2006) in na 5. slovenski in 1. mednarodni konferenci Jezikovne tehnologije 2006 v Ljubljani (9.–10. oktober 2006).

Predstavitve v Ljubljani so pomenile vključitev slovenske strokovne javnosti v načrtovanje govornega korpusa za slovenščino, ob spremljajočih diskusijah pa so se spremenili tudi nekateri moji pogledi na govorni korpus. Aplikacije UKGS po zaključku študijskega bivanja v Bergnu nisem mogla več spreminjati, predlogi za izboljšave pa so vključeni v poglavji *Priporočila za transkribiranje besedil v govorni korpus* in *Priporočila za označevanje besedil v govorni korpus*. Namen učnega korpusa je bil tako dosežen, saj je služil kot model za učenje gradnje govornega korpusa, pa tudi kot model za preverjanje načel za transkribiranje in označevanje. S tem je bila funkcija učnega korpusa zaključena, sledili pa so ji nadaljnji koraki v zvezi z načrtovanjem in gradnjo govornega korpusa za slovenščino.

Po mojih izkušnjah je za slovenščino realno načrtovati govorni korpus v velikosti okrog 1 milijona besed, odvisno od finančnih možnosti in razpoložljivih človeških virov. V tem okviru je nastala večina govornih korpusov po svetu, razen izjemno velikih nacionalnih projektov ali projektov, podprtih z velikim kapitalom. Največ finančnih sredstev in časa pri gradnji govornega korpusa zavzame transkribiranje in označevanje besedil. Pri transkribiranju učnega korpusa sem v povprečju potrebovala 30 minut za transkribiranje 1 minute posnetka. Tudi če bi se čas transkribiranja skrajšal, je treba upoštevati še čas za pregled in usklajevanje transkripcij ter generiranje glav besedil. Izračunala sem, da bi z dodatnim (avtomatskim) oblikoskladenjskim in skladenjskim označevanjem za gradnjo verjetno potrebovali tri leta; za finančno orientacijo pa naj služi podatek, da je gradnja desetkrat večjega nizozemskega govornega korpusa stala 440 tisoč evrov.

[203] http://helmer.aksis.uib.no/batmult/Janas_Final_Report.htm
[204] http://www2.arnes.si/~svinta/jota.html

9 Zgledi iskanja po učnem korpusu

Jezikovnih podatkov, ki jih dobimo z analiziranjem UKGS, ne moremo posploševati na govorjeno slovenščino na splošno, ker korpus ni uravnotežen. Kljub temu z njimi lahko nakažemo nekatere možnosti raziskovanja, ki jih omogoča govorni korpus. Osnovni namen primerov, ki sledijo, in sploh pričujočega poglavja ni poglobljena analiza govorjene slovenščine, ampak prikaz možnosti za raziskovanje govorjenega jezika z govornim korpusom.

9.1 MOŽNOSTI DOSTOPANJA DO BESED IN BESEDIL

Besedila v UKGS so dostopna na dva načina: kot celote ali preko konkordančnika. Dostopnost celih besedil, ki jih je mogoče kot celote ali po delih tudi poslušati, je za nekatere jezikoslovne analize pomembna. Spodnja slika prikazuje primer transkribiranega in označenega besedila, dostopnega kot dokument:

G11:	🔊	navaden {nst}stajl{/nst} to so tako recimo srednje široke hlače
	🔊	{tj}punk{/tj} {nst}stajl{/nst} so bolj take ozke
	🔊	ə pač pa {tj}rap{/tj} {nst}stajl{/nst} to so pa {tj} {?}baggypants{/?}{/tj} to so pa tako široke da ne vem
	🔊	kot da bi imel vrečke na nogah [{neraz}]
G12:	🔊	[ja take]
	🔊	{??} v {nst}iber{/nst} {nst}hudih{/nst} ne vem{/??}
G11:	🔊	potem imaš dolge majice ponavadi
	🔊	pač pri {tj}rap{/tj} {nst}stajlu{/nst} pa kapice kakšne postrani obrnjene pač pa dimije
	🔊	[to so kapice] ə zimske kapice ki se jih nosi v bistvu poleti
G13:	🔊	{??}[ja ja]{/??}
G11:	🔊	[əm] [ne vem] kaj še
G12:	🔊	[dim]
G13:	🔊	[ja]
G12:	🔊	frizure so čisto različne tako da ni nič važno v bistvu
G11:	🔊	pa potem boksarice se nujno nosi [{nv}smeh{/nv}]
G12:	🔊	[{nv}smeh{/nv}]

G11:	🔊	ne vem kaj še ja to je to
G13:	🔊	{neraz} to tudi z muziko {neraz}
	🔊	ne vem kaj poslušaš ne tako si potem ne vem te {tj} punk{/tj} si pač
	🔊	oprijet pa to ne vem srajca pa to {nst}rejparji{/nst} pa tako široko pa to
G12:	🔊	ja saj v bistvu
G11:	🔊	saj imaš tudi mislim tudi {tj}hard core{/tj} pa to
	🔊	pa eni poslušajo to bolj komercialo pa to to kar je na {okr}emtiviju{/okr} pa to
	🔊	ne vem čisto odvisno
	🔊	(govorjenje v ozadju)

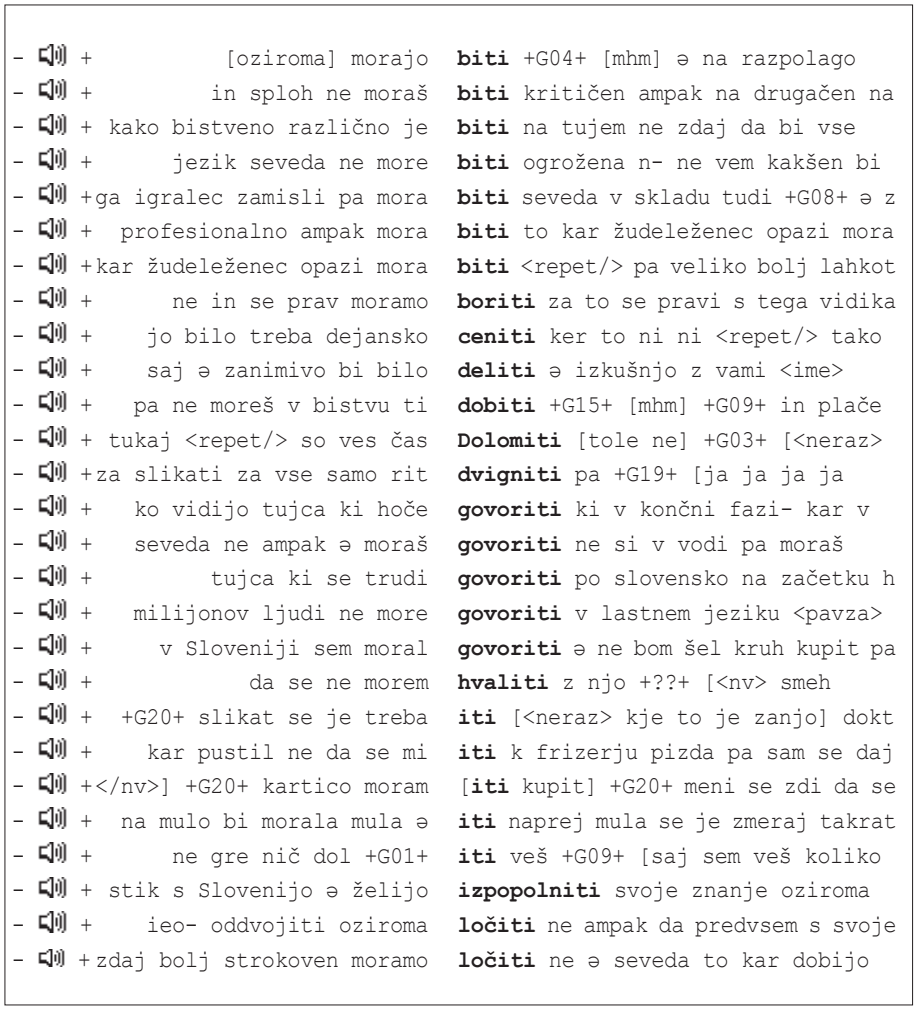
Slika 44: Del transkripcije posnetka R05, dostopne v obliki besedila

Običajnejši dostop do gradiva v korpusu je preko konkordančnika. Kot rezultat iskanja dobimo konkordančni niz. Omenjeno je že bilo, da v obstoječi aplikaciji lahko iščemo eno, dve ali tri sosednje besede; spodaj je iskalni niz po dveh začetkih besed, slovensk- jezik-:

🔊	+	slovi- Slovenije mhm ə poletna šola	**slovenskega jezika** in magistra
🔊	+	celoletne šole [celoletne šole]	**slovenskega jezika** kar pomeni
🔊	+	že pri praktični izvedbi mhm seminar	**slovenskega jezika** kulture in
🔊	+	ravno na polovici mhm ə seminar	**slovenskega jezika** literature
🔊	+	taka ocena əm ə za seminar	**slovenskega jezika** literature
🔊	+	sodi tudi že tradicionalni seminar	**slovenskega jezika** literature
🔊	+	obiskujejo ko se zberejo na seminarju	**slovenskega jezika** literature
🔊	+	<ime> <priimek> vodja poletna šole	**slovenskega jezika** magistra
🔊	+	dela ne mislim da poletna šola	**slovenskega jezika** ə magistra
🔊	+	prav zdaj poteka tudi poletna šola	**slovenskega jezika]** [o
🔊	+	nadnaslovno temo ə večkulturnost v	**slovenskem jeziku** literaturi
🔊	+	središča nenazadnje pa tako ne le	**slovenski jezik** ampak tudi
🔊	+	organizatorjem na ta način prodira	**slovenski jezik** v številna
🔊	+	bodo ali se že poklicno ukvarjajo s	**slovenskim jezikom** literaturo

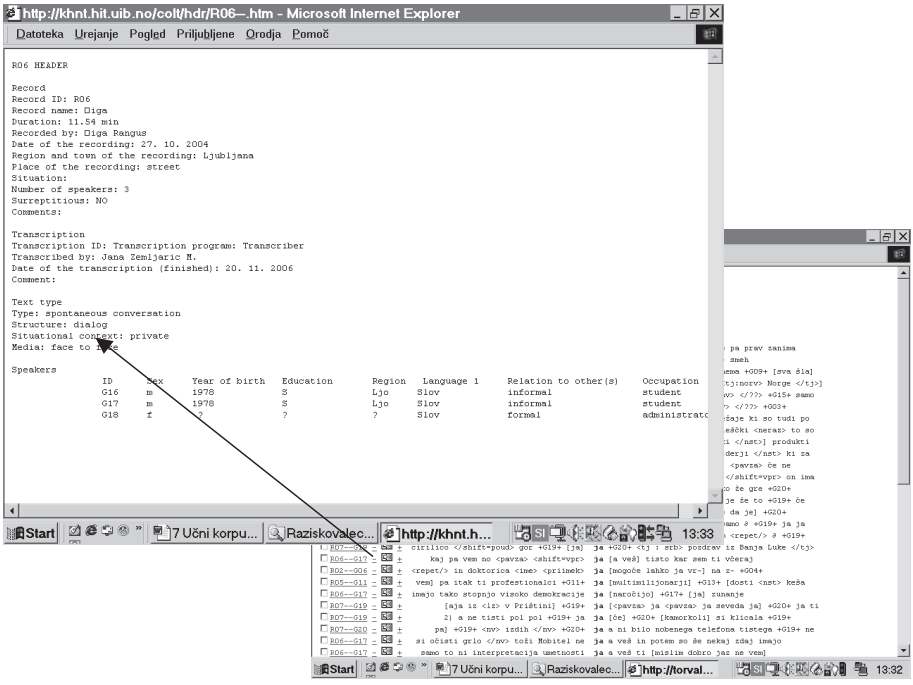
Slika 45: Iskalni niz *slovensk- jezik-*

Na obstoječi platformi je mogoče iskati tudi pa začetnem delu besede (dolžino niza lahko poljubno izberemo), končnem ali sredinskem delu besede. Spodaj je prikazano iskanje po besedah s končajem -iti, ki nam izpiše vse nedoločnike na -iti, seveda pa tudi vse ostale besede, ki se končajo na -iti:



Slika 46: Del konkordančnega niza s pripono *-iti*

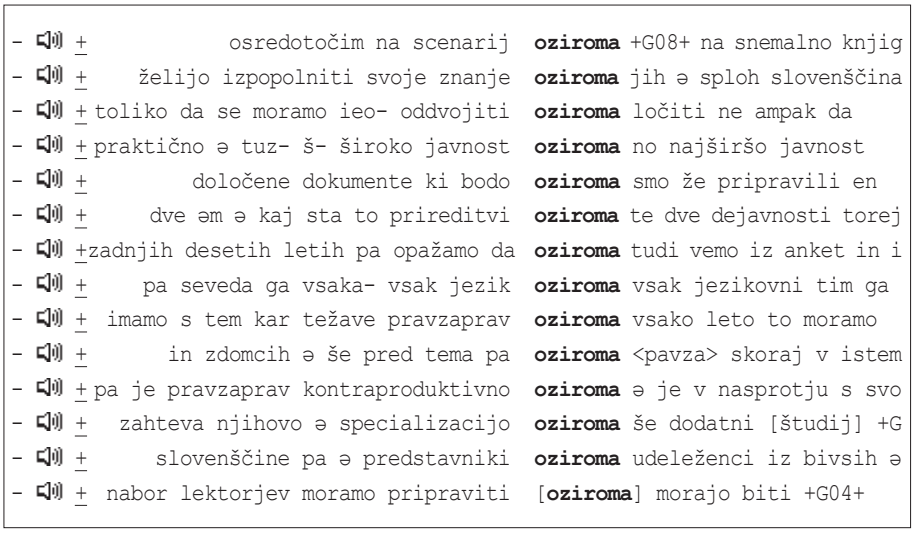
V konkordančniku je na začetku vsake vrstice šifra posnetka, kateremu pripada izjava, in šifra govorca (npr. R06 - - G17). S klikom na šifro prikličemo glavo besedila, kjer lahko preberemo podatke o posnetku, transkripciji in govorcih.



Slika 47: Povezava izjave z glavo dokumenta

9.2 ISKANJE PO KORPUSU Z OMEJEVANJEM PO DEMOGRAFSKIH IN BESEDILNOVRSTNIH KRITERIJIH

Konkordančnik UKGS omogoča iskanje po korpusu z upoštevanjem omejitev po različnih (demografskih in sobesedilnih) kriterijih, ki so bili predvideni v fazi zajemanja besedil. Tako lahko npr. po korpusu iščemo izbrano besedo, med demografskimi kriteriji pa govorce omejimo na izbrani spol, določeno starost, stopnjo izobrazbe in podobno (seveda glede na to, kaj je bilo v besedilih označeno). Enako lahko iščemo tudi po besedilnovrstnih kriterijih. Za primer si oglejmo npr. besedo *oziroma*, ki se v korpusu pojavi štirinajstkrat. Vse pojavitve pripadajo govorcem s končano univerzitetno izobrazbo, kar ugotovimo z omejevanjem iskanja po kriterijih izobrazbe. Z dodatnim upoštevanje kriterija formalni/neformalni govorni položaj pa ugotovimo, da je bila beseda vedno uporabljena v formalnem govornem položaju, da torej tudi govorci z univerzitetno izobrazbo besede *oziroma* v neformalnih govornih položajih niso uporabljali.

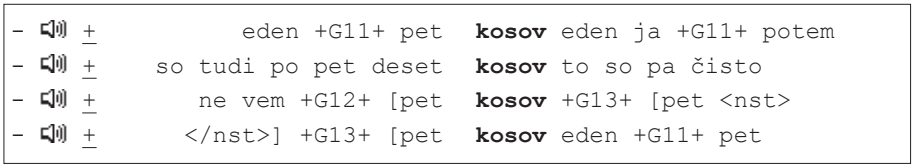


Slika 48: Iskanje po korpusu z omejitvijo na govorce z visoko izobrazbo

9.3 DRUGI ZGLEDI ISKANJA PO KORPUSU

9.3.1 Iskanje na besedni ravni

V govornem korpusu lahko besede analiziramo na različnih ravneh – lahko npr. preučujemo njihov pomen, pogostnost pojavljanja (rabo) in drugo. V UKGS najdemo besede, ki jih ni v SSKJ ali v Besedišču SJ, ni pa jih niti v pisnih korpusih Fidaplus in Nova beseda, npr. besedi *podpriročnik* in *prapriročnik*. Nekatere znane besede izkazujejo v UKGS nove pomene, kot npr. beseda *kos*, ki v določenem kontekstu UKGS pomeni denarno enoto, in sicer tisoč tolarjev:



Slika 49: Razbiranje pomena iz sobesedila v konkordancah[205]

[205] V tem primeru se iz konkordančnega zapisa ne da razbrati pomena besede; pomagamo si lahko z ustreznim izsekom iz celotnega besedila.

G11: podvozja stanejo okoli ne vem
G12: [pet kosov]
G13: [pet kosov] eden
G11: pet kosov eden ja
potem imaš pa še ležaje ki so tudi po pet deset kosov to so pa čisto različno
kolikor dobre hočeš

Čeprav je v zgornjem primeru jasno, da gre za izrazito pogovorno, celo slengovsko rabo, je to jezikovna realnost, ki lahko v določenih okoliščinah preraste v splošno rabo. Zato je govorni korpus pomemben tudi za gradnjo slovenske leksikalne podatkovne zbirke (prim. Gorjanc et al. 2005c), v kateri se zbirajo podatki o realnem jeziku, torej aktualnem leksikalnem naboru in pomenih v slovenščini. O vključitvi take ali podobne besede ali zveze v slovarsko gradivo bi morala odločati pogostnost pojavitve oz. moč besedne povezovalnosti (Gorjanc et al. 2005c, 12).

Analizo pomena si lahko ogledamo tudi na primeru besede *mhm;* v SSKJ je pomen razložen kot:

> **mhm** [mhm] medm. (m-m) *izraža obotavljanje, pomislek, dvom*: mhm, morda bo šlo // *izraža (zadržano) pritrjevanje*: mhm, je odgovoril; tisto pa, mhm

V UKGS se beseda pojavi 87-krat, polovica zadetkov je prikazanih spodaj:

+	<nv> smeh </nv>] in ka-	**[mhm]** +G07+ [ampak] ə drugače
+	v katalonščini pa ne morem	**[mhm]** +G07+ [dobesedno] sedem
+	državi kar se tega tiče ne	**[mhm]** +G07+ [s] tega vidika
+	v ozadju) +G03+ [əm] <pavza>	**[mhm]** +G09+ [hitro] eno kartico
+	kot vidite na Atlantiku ne] +G01+	**[mhm]** +G09+ [mhm] +G10+ [mhm]
+	ne moreš v bistvu ti dobiti +G15+	**[mhm]** +G09+ in plače in
+	ne] +G01+ [mhm] +G09+	**[mhm]** +G10+ [mhm] +G15+ [mhm]
+	[mhm] +G09+ [mhm] +G10+	**[mhm]** +G15+ [mhm] +G03+ obilica
+	časa za [druženje] +G04+	**[mhm]** [ampak tista prioriteta]
+	v Slovenijo preko Moskve ne	**[mhm]** [ampak ə] +G02+ [<nv>
+	vidika [ne] hə +G02+	**[mhm]** [ja <smeh> doktor
+	[bo že torej mogoče]	**[mhm]** [ja] +G02+ [ə]
+	mi zdi [da] +G02+	**[mhm]** [mhm] [ə doktorica <ime>
+	[s Slovenci] +G07+	**[mhm]** [mhm] əm kaj vam je to
+	usposobljena za slovenščino +??+	**[mhm]** [mhm] əm <ime>
+	[da] +G02+ [mhm]	**[mhm]** [ə doktorica <ime>
+	na seminarju [ane] +G03+	**[mhm]** ampak res predvsem v
+	kaj bom zdaj jaz +G04+	**[mhm]** bo mogoče <smeh> še ona
+	čez [štirideset] let +G04+	**[mhm]** bog ve kaj bo ne jaz
+	funkcijo] čeprav je res +G04+	**[mhm]** da je potreba po [nečem
+	dejansko je pa res +G04+	**[mhm]** da je- ə bo treba po
+	mi včasih smo to +G04+	**[mhm]** doživljali tako ne dosti

+	[dežela] za nas +G04+	**[mhm]**	drugače jaz sem študiral
+	[perspektive] predstaviti +G04+	**[mhm]**	in to je uspelo ne tudi
+	naredil in tako naprej +G04+	**[mhm]**	in zato pripravljamo
+	konkretno lepo ja +??+	**[mhm]**	ja hvala [<nv> smeh
+	<repet/> patetičen [ne] +G04+	**[mhm]**	ker je pa tako ne kar se
+	in [drugi] ne +G04+	**[mhm]**	ker v bistvu ə en del
+	še dodatni [študij] +G02+	**[mhm]**	mhm doktor <ime>
+	v Kataloniji [ne] +G04+	**[mhm]**	mi smo se srečali s tem
+	<ime> +G15+ je pa obilo dežja <pavza>	**[mhm]**	ne v Bergnu ne +G03+ a je
+	in slovenščini [ne] +G04+	**[mhm]**	no kot je bilo že prej
+	<repet/> naj bi bil razlog	**[mhm]**	potem ə če lahko govorim
+	[drugi] del publike +G04+	**[mhm]**	recimo ki študira
+	govo- n- [govorcev] +G04+	**[mhm]**	saj v končni fazi ne gre
+	posebej [dopovedovati] +G04+	**[mhm]**	torej seminar je zasnovan
+	[ljudi] v +G04+	**[mhm]**	tujih državah da bodo
+	od ponedeljka do petka ne ja	**mhm]**	<neraz> pa naši semenir-
+	življenje [<neraz>] +G04+	**[mhm]**	<pavza> [Slovenija] in to
+	kot za eno državo +??+ [mhm	**mhm]**	<pavza> mislim in to
+	precej na delovnem mestu +G04+	**[mhm]**	ə doktor <ime> <priimek>
+	razsežnosti ə [Slovenije] +G04+	**[mhm]**	ə je pa razlika
+	kulture +G07+ [seminarja]	**[mhm]**	ə ki prihaja iz
+	oziroma] morajo biti +G04+	**[mhm]**	ə na razpolago tako da se
+	se mi zdi] ne +G04+	**[mhm]**	ə no morda v tem
+	</nv>] [ne] +G04+	**[mhm]**	ə no saj ə zanimivo bi
+	pa isto misel ne +G04+ [mhm	**mhm]**	ə to da je na Slovenstvu
+	[da ga ima] +G04+	**[mhm]**	ə to je mogoče se slišijo

Slika 50: Beseda *mhm* v UKGS

Kot lahko vidimo, *mhm* v navedenih primerih vedno izraža pritrjevanje, (popolno) strinjanje ali pa nakazovanje, da govorec sledi pogovoru. Ta zadnji pomen oz. vloga je tudi grafično prepoznavna, saj je *mhm* največkrat izgovorjen hkrati z besedilom drugega govorca (in zato zapisan v oglatem oklepaju). V tem primeru bi lahko gradivo, izkazano v govornem korpusu, vplivalo na določitev oz. spremembo pomena besede v slovarju.

Naslednji primer nakazuje potrebo in možnost za uporabo korpusa pri jezikovni kodifikaciji. Tako npr. Verovnik v svoji razpravi o povratnem svojilnem zaimku ugotavlja, da »bi bilo treba v primeru rabe zaimka *svoj* primer-

jati tudi stanje v govorjenih in pisnih besedilih – zaradi manjše oz. nikakršne možnosti poznejšega (samo)popravljanja bi v govorjenih besedilih morda ugotovili še izrazitejše zmanjševanje rabe povratnega zaimka« (Verovnik 2005, 56). UKGS te hipoteze ne potrjuje: povratni svojilni zaimek se (v vseh sklonih in številih) pojavi dvaindvajsetkrat, od tega štirikrat v neformalnih besedilih, uporabijo pa ga skoraj vsi govorci UKGS; svojilni zaimek *moj* se v različnih oblikah pojavi desetkrat, *njegov* štirikrat in *njen* enkrat. Kot v vseh drugih zgledih pa seveda tudi v tem primeru velja velika previdnost pri posploševanju.

Med dragocene in pogosto varovane podatke v korpusih (Gorjanc 2005a, 73) sodijo *frekvenčne liste besed*, to so podatki o pogostnosti pojavljanja besed. Tudi za UKGS obstaja lista besed: v korpusu je 3118 pojavnic; ker korpus ni lematiziran, različnic ni mogoče avtomatsko prešteti. Od 3118 pojavnic se jih 2100 (dve tretjini) pojavi samo enkrat, 400 se jih pojavi po dvakrat, le okrog 600 besed pa ima tri ali več pojavitev.[206] Največkrat, skoraj 500-krat, se v UKGS pojavi beseda *je*. Spodaj je prikazana lista besed UKGS po pogostnosti pojavljanja v primerjavi s pojavitvami v korpusu Fida:

UKGS				Fida
	AF[207]	RF[208]	Pojavnica	Različnica
1	498	35.422	je	biti
2	425	30.230	ne	v
3	358	25.464	ə	in
4	313	22.263	pa	na
5	297	21.125	in	za
6	284	20.201	se	da
7	270	19.205	da	ta
8	268	19.063	to	ki
9	265	18.849	ja	pa
10	264	18.778	v	z
11	186	13.230	na	tudi
12	143	10.171	tudi	s

[206] Zdi se skoraj neverjetno, kako malo besed uporabljamo pri tvorjenju besedila. Rezultate bi bilo treba preveriti na večjem obsegu gradiva, čeprav nekatere tuje raziskave kažejo podobne rezultate, da se namreč tudi v velikih govornih korpusih več kot polovica besed pojavi samo po enkrat.

[207] AF je oznaka za absolutno frekvenco, število pojavitev v korpusu.

[208] RF je oznaka za relativno frekvenco, število pojavitev na 1000 besed v korpusu.

13	130	9.247	za	po
14	115	8.180	ki	kot
15	106	7.540	so	še
16	105	7.469	tako	ves
17	105	7.469	mhm	iz
18	98	6.971	kaj	ali
19	88	6.259	a	o
20	86	6.117	še	tako
21	84	5.975	če	
22	78	5.548	zdaj	

Slika 51: Seznam pojavnic z najvišjo frekvenco v UKGS in v Fidi[209]

Absolutna frekvenca (AF) pojavitev v korpusu kaže pričakovano stanje. Prvih deset najpogostejših pojavnic iz Fide je tudi na seznamu dvajsetih najpogostejših različnic v UKGS: *biti* (v UKGS *je* in *so*), *v, in, na, za, da, ta* (v UKGS *to*), *ki* in *pa*; poleg teh so skupne še *tudi*, *še* in *tako*. Ti podatki kažejo, da že zelo majhen korpus pri besedah z veliko frekvenco pojavljanja kaže dokaj realno podobo jezika. Značilen je tudi seznam preostalih najpogostejših besed iz UKGS, ki jih na seznamu Fide ni: *ne, ə, ja, mhm, kaj, a* so diskurzni označevalci, najznačilnejši predstavniki govorjenega jezika.

Zanimivo je primerjati tudi pogostnost osebnega zaimka *jaz* v imenovalniku: v govornem delu Fide, ki obsega 2 milijona besed, se pojavi več kot 5000-krat, torej z relativno frekvenco 2,5. V UKGS, ki ima 15.000 besed, se pojavi 38-krat, to je z relativno frekvenco 2,53. Tudi v tem primeru dobimo pri besedi s precej veliko pojavnostjo podoben rezultat v dveh govornih korpusih različnih velikosti.

V skladnji spontanega govora lahko opazujemo še druge značilnosti, npr. napačne začetke; ti so eden izmed najočitnejših razlikovalnih elementov glede na pisni jezik. Termin označuje dogodek, ki se zgodi na ravni besede; govorec besedo začne, pa je ne dokonča – je prekinjen, si premisli, se zmoti itd. Pojav si lahko ogledamo na primeru iz UKGS. V celotnem korpusu je 174 primerov napačnih začetkov, od tega 160 primerov (kar 92 %) v besedilih, kjer so udeleženci v formalnem odnosu in so besedila javna, le 14 pojavitev (8 %) pa je v neformalnih zasebnih besedilih. Razmerje je presenetljivo, kljub temu da formalna besedila v korpusu zavzemajo dve tretjini vseh besedil. Med formalnimi besedili je posnetek, dolg 55

[209] Gorjanc 2005, 73; lista ni popolnoma primerljiva z listo UKGS, ker gre tu za najpogostejše različnice (biti), pri UKGS pa za pojavnice (je, so).

minut. V njem nastopa 6 govorcev; UKGS omogoča pregled deležev izgovorjenih besed[210] posameznih govorcev:

Govorec	Št. vseh enot	Št. napačnih začetkov	Odstotek %
G07	1400	28	2
G06	892	17	1,90
G03	914	16	1,75
G02	4184	73	1,74
G04	1845	21	1,13
G05	484	5	1,03

Slika 52: Napačni začetki v formalnem govoru UKGS

Vsi govorci imajo presenetljivo podobne deleže napačnih začetkov glede na zelo različno število izgovorjenih enot. Delež je nekoliko nižji samo pri govorki, ki je govorila najmanj (G05), in pri govorki, ki je bila deloma vnaprej pripravljena, saj je vodila pogovor (G04). Vsi govorci so bili visoko izobraženi, od tega je pet slovenistov (med njimi dva doktorja in dve magistrici znanosti) in ena profesionalna novinarka; pogovor je potekal v živo na I. programu Radia Slovenija, tema je bila slovenski jezik; za enega govorca (G07) je bila slovenščina tuji jezik. Iz teh okoliščin lahko sklepamo, da je šlo za formalno obliko spontanega govora, ki bi jo lahko imeli za neke vrste standard formalnega spontanega govora v javnosti; očitno tudi v tej obliki govora obstaja določen (morda celo predvidljiv) delež napačnih začetkov, ti pa so, vsaj kolikor lahko sklepamo iz UKGS, bolj značilni za formalni kot za neformalni govor.

Ničesar še nismo povedali o vzrokih, zaradi katerih nekatere besede ostanejo neizgovorjene. Brez obširne raziskave tega tudi ni mogoče ugotoviti, lahko pa postavimo nekaj hipotez, če si ogledamo vse napačne začetke v neformalnem govoru UKGS:

```
kaj </shift=vpr> ja zdaj je prinesla  dm- samo ven [mu
               +G17+ [zdaj] me kliče  [k- ja] <shift=vpr> kaj
      </nst> do <nst> pizd </nst> pa  n- tako naprej <pavza>
             <ime> mi je rekla da so  na- naredili raziskav na
                                      fədədvəju
```

[210] Namesto besed bi bilo bolje uporabiti izraz enot, ker računalnik avtomatsko prešteva tudi polverbalne in neverbalne dogodke, npr. smeh, vzdih, tlesk ...

<pavza> itak pa ni ne petek je	**no-** je [enaintrideseti]
filozofija umetnosti ne znam ti	**po-** mislim ne da se [opredeliti]
a o umetnosti ss- ste se kaj	**pogova-** </shift=vpr>
pisana beseda nekaj nekaj <repet>	**pov-** hoče povedat ja]
+G16+ [ja] ej Rupel je car	**Ru-** Rupel je ne vem on je tak svetovljan
[cel svet je] cel> repe	**s-** cel sve je za Kerryja
je problem <shift=vpr> a o umetnosti	**ss-** ste se kaj pogova-
ne vem <repet> potem imaš ∂	[**trivial-** trivialno literaturo]
v tem ne da se potem človek	**vpra-** <shift=vpr> da je pisal to o o
da je pisal to o o vakuumu o	**člo-** v človeku ne

Slika 53: Napačni začetki v neformalnih besedilih UKGS

Splošni vtis je, da je pojava napačnih začetkov glede na količino neformalnih besedil izredno malo. Eden izmed vzrokov je morda večja interaktivnost in tekočnost neformalnega govora – izjave so krajše in se hitro izmenjujejo. V formalnem govoru govorci tvorijo daljše in zahtevnejše sintaktične strukture, zato potrebujejo več časa za načrtovanje govora; zaradi zahtevnosti formulacij so tudi napačni začetki pogostejši. Na podlagi vseh neformalnih besedil UKGS ugibam, da je do napačnih začetkov prišlo v primerih, ko so nastopile zunanje motnje pogovora in v primerih, ko so se govorci pogovarjali o vsebinsko zahtevnejših temah (konkretno o umetnosti, politiki itd.). Za empirično potrditev te teze bi bilo tudi v tem primeru potrebno analizirati večjo količino gradiva.

9.3.2 Iskanje na ravni diskurza

V korpusih iščemo tudi informacije, ki presegajo raven besede. Tako so npr. kolokacije pomemben jezikovni vir, predvsem v leksikografiji, pa tudi pri učenju jezika kot tujega jezika (prim. Gorjanc in Jurko 2004). UKGS z obstoječimi programskimi orodji sicer ne omogoča računanja statističnih vrednosti sopojavljanja, tako da moramo sopojavnice iskati sami na podlagi različnih hipotez, na nekatere pa lahko naletimo tudi po naključju. Pri analizi pojavljanja členka *a* proti *ali* sem npr. naletela na močno sopojavnost besed *a* in *veš*: *a* se v UKGS pojavi 88-krat,

od tega v sopojavitvi z *veš* 37-krat (ne preseneča, da so vsi primeri iz neformalnega govornega položaja):

+	<repet> naročili zunanje opazovalce	**a** veš	+G17+ [malo bolj]
+	nekaj] takega ne v obe smeri	**a** veš	+G17+ [no sigurno te
+	to se voziš </> pa] deset ur	**a** veš	+G17+ [<neraz> ne vem
+	to pa tako <nst> nabijejo </nst> ceno	**a** veš	+G19+ ti gre pa s
+	jezik kje pa je to v Bergnu] +G01+	**a** veš	[<neraz> <pavza> tako
+	gor ne] +back sounds+ [<neraz>] +G01+	**a** veš	cela Adiža gre +G01+
+	+G09+ a si [en čaj skuhamo] +G01+	**a** veš	da je lepo tam samo je
+	očisti grlo </nv> toži Mobitel ne ja	**a** veš	in potem so še nekaj
+	[no no saj] [to to]	**a** veš	in potem ti pušča
+	bomo videli saj ne bo dosti boljše	**a** veš	isti <nst> kurac
+	jurjev </nst> (telefon zazvoni)	**a** veš	ja <pavza> ja
+	vem če si bodo spet lahko privoščili	**a** veš	ker bodo oni spet [ja
+	+G03+ od Bergna do Nordcapa +G03+	[**a** veš	ker jaz sem mislila da
+	pa ti naši tako] dobro tam počutijo	**a** veš	ker nemško [govorijo]
+	njega enkrat bral samo sem ga tako	**a** veš	na hitrico no in
+	pa ko sem klicala +G19+ ə v Priština	**a** veš	na letališče za <nst>
+	ne </shift=vpr> [ja saj je možno	**a** veš	ne <neraz> če mene
+	to ni interpretacija umetnosti ja	**a** veš	ti [mislim dobro jaz
+	[to je] takrat že napisal	**a** veš	to se je malo
+	[a pa to je Tirolska] [pa	**a** veš	tukaj smo šli mi
+	veš zdaj ta enajsti november ima	**a** veš	<neraz> in jih s tem
+	<neraz>] +G16+ [<neraz>] <pavza>	**a** veš	<neraz> ni jasno pa
+	</shift=vpr>] +G16+ [Američan]	**a** veš	<nst> pizda </nst> ne
+	dela pa rabi posnetke ne <pavza>	**a** veš	<pavza> pa je prosila
+	malo daljše] ja <nst> čupavce </nst>	**a** veš	<pavza> tako je ne
+	[ne potem v bližini] +G01+ no in	**a** veš	<pavza> ə +G01+ tja
+	jah] <smeh> hjah </smeh> ne gre to	**a** veš	<shift=vpr> ə
+	kul pa poceni so karte v bistvu	**a** veš	ə <shift=vpr> za
+	bila ah pa] +G01+ Bolzano <neraz>	[**a** veš]	+G03+ [a pa to je
+	+G01+ [recimo da bi jo že imela ne	**a** veš]	+G09+ za poletni
+	ki kliče +G19+ [samo se nabere tega	**a** veš]	+G20+ [ja ja pol pol
+	ah pošta ali pa [<neraz>] +G19+	[**a** veš]	+G20+ no to res ja
+	pa vem no <pavza> <shift=vpr> ja	[**a** veš]	tisto kar sem ti

Slika 54: Delni izpis kolokacije *a veš* iz UKGS

Gre za izrazito govorno prvino, in sicer za diskurzni označevalec; to so izrazi (eno- ali večbesedni), ki »kažejo na povezanost diskurza s kontekstom« (Verdonik 2006, 51) oz. »opravljajo vlogo sredstva preverjanja pozornosti, hkrati pa so tudi sredstvo označevanja oz. kazanja različnih vrst udeleževanja in pritrjevanja« (Kranjc 1999, 65). Verdonik (2006) podrobneje obravnava diskurzne označevalce, vendar med njimi ni zveze »a veš«; dejansko se v njenem gradivu ta označevalec ne pojavlja, saj gre za en sam tip pogovorov – zbiranje informacij o turistični ponudbi po telefonu. Govorci se med seboj vikajo, gre torej izključno za formalni govorni položaj (čeprav ne zelo strog), v katerem zveza »a veš« sploh ni mogoča.[211] Ta ugotovitev potrjuje, kako pomembna je za celostno obravnavo nekega jezikovnega pojava reprezentativnost in uravnoteženost korpusa.

V nadaljevanju je prikazan še en primer rabe diskurznega označevalca – besede *ne* v korpusu UKGS:

mesecema se je <neraz>] per <lz>	**ne**	+G01+ ampak oni bojo ziher
ga je <lz> ko je vozil se s kolesom	**ne**	+G01+ pa nima blatnika +G01+
je vpliv veš to je ob Adiži v bistvu	**ne**	+G01+ skozi od morja +G10+
glej Oslo Bergen je v bistvu daleč	**ne**	+G03+ [tako da <pavza> ja zdaj
je v +G03+ je v <nst> penziji </nst>	**ne**	+G03+ [<shift=vpr> a to je
no glej <nst> klele </nst> je Bergen	**ne**	+G03+ [čisto tukaj a <neraz>
obilo dežja <pavza> [mhm] ne v Bergnu	**ne**	+G03+ a je <nst> scalo </nst> <n
če bi bila tukaj kakšna karta Evrope	**ne**	+G03+ da je od Ljubljane do
izhodišče za vse te taglavne fjorde	**ne**	+G09+ [kako dobro] +G09+ kako
+G01+ Adiža <pavza> a to je tole	**ne**	+G10+ <pavza> ne jaz mislim da
aja to pa ja <nst> valjda </nst>	**ne**	+G11+ na <nst> konteste </nst>
koliko jih <nst> sfuraš </nst> koliko	**ne**	+G11+ večinoma se <?> park </?>
[ali pa] karkoli je kaj takega	**ne**	+G13+ [ja] +G11+ ker je to v
pod resno misliš <neraz> literaturo	**ne**	+G16+ [<neraz>] se štejejo pač
<shift=vpr> kako lahko </shift=vpr>	**ne**	+G17+ [ja ja zastopim] ki se
sem gledal zdaj ko ta ∂m Vega	**ne**	+G17+ toži Mobitel
[zakaj ne] zato ker računalnik	**ne**	[dela ali je kaj drugega
se <?> park </?> <nst> fura </nst>	**ne**	[na kontestih] +G12+ [ja v
Američan] a veš <nst> pizda </nst>	**ne**	[oni ki imajo] <pavza> [ne

Slika 55: Delni izpis kolokacij besede *ne* iz neformalnih besedil UKGS[212]

[211] Zaenkrat lahko samo ugibamo, ali je odsotnost diskurznega povezovalca *a veš* v korpusu telefonskih pogovorov pri Verdonik (2006) morda tudi posledica regionalnih značilnosti tega korpusa.

[212] Oznake G01-G17 so identifikacijske oznake govorcev in stojijo pred njihovimi izjavami;

Prikazanih je prvih 19 vrstic kolokacij (10 %) od skupno 199, kolikor je pojavitev te besede v neformalnih besedilih korpusa (33 % vseh besedil; vseh pojavitev *ne*-ja v korpusu je 425). Vidimo lahko, da je *ne* 18-krat v funkciji diskurznega označevalca (po katerem skoraj brez izjeme sledi predaja besede), in samo enkrat v vlogi klasične nikalnice pred glagolom.[213] To pomeni precejšnje prevrednotenje pomenske napolnjenosti obravnavane besede. Predstavljajmo si tujca, ki zelo slabo razume slovensko in poskuša razumeti neformalno govorjeno besedilo, tako da »lovi« znane besede; *ne* je najverjetneje med znanimi besedami, vendar mu nepoznavanje njegove diskurzivne funkcije razumevanje besedila lahko otežuje.

Med slabo raziskana prozodična sredstva v govoru sodijo premori. Nekatere njihove funkcije so vsaj približno znane, npr. poudarjanje ali pridobivanje časa za načrtovanje govora. Hitra analiza premorov v UKGS je pokazala naslednje: v korpusu je označenih 188 premorov; od tega 68 v formalnih besedilih (66 % korpusa) in bistveno več, 120, v neformalni tretjini korpusa. V formalnem govoru so vsi premori brez izjeme dolgi največ 1 sekundo, v neformalnem govoru pa je ena četrtina premorov daljših (od 2 do 11 sekund). Gradivo torej kaže, da so premori manj moteči oz. bolj običajni (in zato sprejemljivi) v neformalnem govoru. V formalnem govoru se 93 % premorov zgodi v okviru monologa, naredi jih govorec sam od sebe, ne da bi ga kaj zmotilo; v 10 odstotkih premor podaljša še z zapolnjenim premorom (polglasnik). V neformalnem govoru so premori popolnoma drugačne narave, okoli premora se vedno »nekaj dogaja«: 24-krat po premoru spregovorita oba govorca hkrati, kot da se nista dobro razumela, komu pripada vloga govorca; 13-krat sta hkrati govorila pred premorom, očitno sta se oba ustavila v puščanju prednosti; 6-krat premoru sledi vprašanje (po Zuljan Kumar (2007) izrazito sredstvo besedilne kohezije), ki načenja novo temo; 5-krat je premor posledica motenj v okolici (ropot) ali pa glasovni zvoki iz okolice nakazujejo, da so govorci prekinili govorjenje zaradi zunanjih okoliščin (gledajo televizijo, listajo knjigo). Že ta površna analiza omejene količine gradiva je pokazala, da se premori v govoru zelo razlikujejo med seboj in da niso nekaj redundantnega in motečega, ampak imajo vedno svoj vzrok in pogosto tudi svojo funkcijo, kar je vse treba še natančneje raziskati.

</nst> ne [oni ki imajo]	<**pavza**> [ne <neraz>] potem pa oni v
[ja no] <neraz> +G01+ [<neraz>]	<**pavza**> [nič ne dela] +G15+ [ja m
fila </nst> [skozi] +G17+ <neraz>	<**pavza**> [pa] Irak <nv> pihne skozi
[ja jaz tudi nimam več] +G15+	<**pavza**> [samo na kakšno staro če se
[tukajle] +G10+ [mhm] +G03+	<**pavza**> [samo padeš dol] +G01+

[213] Če pogledamo vse pojavitve v neformalnih besedilih, je približno ena četrtina pojavitev v funkciji nikalnice.

+G10+ [<neraz> ja] +G15+	<**pavza**> [tudi letiš] +G01+ <pavza>
[<ime> ne <ime> <neraz>] +G01+	<**pavza**> [<govori s polnimi usti>
pet procentov pa [za Busha]	<**pavza**> [<neraz [ja] +G16+ [ja
tako no danes je Himalaja problem je	<**pavza**> a film ja aha <shift=vpr> a
napravo in drugi šumi) +G01+ Adiža	<**pavza**> a to je tole ne +G10+
[<neraz>] +G16+ [<neraz>]	<**pavza**> a veš <neraz> ni jasno pa
doktorat dela pa rabi posnetke ne	<**pavza**> a veš <pavza> pa je prosila
(ostro šumenje papirja) +G03+	<**pavza**> aja daj [preden] začnem ej
[kar] </smeh> +G16+ [aja]	<**pavza**> bo prišel [iskat] +G17
sigurno <shift=vpr> ane </shift=vpr>	<**pavza**> imate <shift=vpr> kako se
ja <shift=vpr> aja delaš </shift=vpr>	<**pavza**> itak pa ni ne petek je no-
Avstrija] +G10+ [Italija <pavza> med	<**pavza**> Italija ja] +G15+ [v Itali
se lotili] +G01+ [domov pojdi]	<**pavza**> ja +G10+ <??> <nv> smeh
</shift=vpr> +G11+ [ameriške firme]	<**pavza**> ja +G13+ [ti <nst> skejt
v Prištini] +G19+ ja [<pavza> ja	<**pavza**> ja seveda ja] +G20+ ja ti
je v bistvu daleč ne +G03+ [tako da	<**pavza**> ja zdaj grem jaz z z
(telefon zazvoni) a veš ja	<**pavza**> ja <shift=vpr> kaj je
spet <tj : ang> message </tj> +G20+	<**pavza**> tri mesece je mimo <nv>
bo zmagal spet malo pogoljufali a	<**pavza**> vprašanje tudi če ne ne ja
ane potem pa še tam neke formalnosti	<**pavza**> zdajle dvanajstega imam
če ne dela +G10+ [<??> no </??>]	<**pavza**> <??> aja </??> +back
v Sloveniji tako ime recimo </> +G11+	<**pavza**> <ime> <priimek> +G01+ ja
skozi usta </nv> to samo gledaš	<**pavza**> <neraz> (ropot avtomobila)
+G15+ <pavza> [tudi letiš] +G01+	<**pavza**> <neraz> <?> telemarkt </?>
ja] +G15+ [v Italiji] +G09+	<**pavza**> <nv> smeh </nv> +G09+
da je kvečjemu tukaj bila kje +G09+	<**pavza**> <nv> smeh </nv> <ime> pomoč
kje pa je to </shift=vpr> *∂* na Viču	<**pavza**> <shift=vpr> a je to
pa tudi drugače kaj pa vem no	<**pavza**> <shift=vpr> ja [a veš]
a veš <pavza> pa je prosila če lahko	<**pavza**> <shift=vpr> kaj pa dela za
</nst> a veš <pavza> tako je ne	<**pavza**> <shift=vpr> kaj sta bila
<neraz> <pavza> (5) film ja	<**pavza**> <shift=vpr> si slišal kaj
potem v bližini] +G01+ no in a veš	<**pavza**> *∂* +G01+ tja grede smo šli
do konca septembra +G15+ a svojo	<**pavza**> *∂* po mo <lz> +G15+ b <lz>
kamorkoli] si klicala +G19+ ja +G19+	<**pavza**> čakaj kako je Ie to +G19+

Slika 56: Nekateri premori v neformalnem govoru UKGS

Pri analizi diskurza s pomočjo UKGS se je izkazala pomanjkljivost, ki bi jo bilo treba pri morebitni gradnji večjega govornega korpusa odpraviti. Pri polovici po-

snetkov namreč manjka začetek sporazumevanja, ker se je oseba, ki je pogovor snemala, brez prižganega snemalnika s sogovrcem predvidoma najprej pozdravila, nato opravila vsa uvodna pojasnila, nato začela s snemanjem. Pri polovici posnetkov manjka tudi zaključek pogovora, ker je bilo snemanje zaradi različnih razlogov nenadoma prekinjeno. Tako v učnem korpusu v neformalnih besedilih ni mogoče najti primerov vzpostavljanja stika, pozdravljanja in poslavljanja. Ob morebitni gradnji KGS bi veljalo snemalce opozoriti, da poskušajo posneti govorjena besedila v čim bolj zaokroženi celoti, če je to mogoče.

9.3.3 Govorni korpus pri poučevanju in učenju jezika

Referenčni korpusi, posebej njihove govorne komponente, so pomemben jezikovni vir tudi pri poučevanju in učenju tujega jezika.[214] So namreč vir relevantnih jezikovnih podatkov, ki se lahko prenašajo v učno gradivo in v pedagoški proces. Govorni korpusi in korpusi nasploh »postopoma izrinjajo rojenega govorca s položaja jezikovnega modela ali glavnega razsodnika o jezikovni rabi« (Sinclair 2004, 5). Korpusi v tem pogledu predstavljajo velik napredek, saj rojeni govorci nekega jezika nimajo nujno uzaveščene vsestranske in izčrpne podobe o jeziku in njegovi rabi v vseh okoliščinah. Znano je tudi, da vsi rojeni govorci nimajo enake jezikovne intuicije, pisci gradiv in učitelji jezika pa se pri uporabi jezikovnih vzorcev pogosto opirajo na svojo lastno jezikovno intuicijo in rabo, ki je njim najbližja, predstavljajo kot pravilno ali najbolj sprejemljivo.

Z osredotočanjem na besede in sporazumevalne vzorce z visoko frekvenco, na splošno pa z osredotočanjem na tisto, kar je v jeziku bolj pogosto in običajno, učitelj pomaga učečemu se pri usvajanju jezika, posebej na začetni in nadaljevalni stopnji učenja. Korpusne analize omogočajo poiskati, kaj je v jeziku tipično, zato je za učitelje in pisce gradiv referenčni korpus najboljši vir realnih jezikovnih podatkov. V različnih študijah (več o tem Tsui 2004, 40–41) so na podlagi korpusnih podatkov in analiz pokazali na velike razlike med tem, kako je jezik predstavljen v učbenikih za tujce, in tem, kako je evidentiran v korpusih. Za slovenščino take analize še ne obstajajo, lahko pa problem ponazorim z opisom ene izmed jezikovnih funkcij v *Sporazumevalnem pragu za slovenščino* (Ferbežar in drugi 2004) in jo primerjam z jezikovno realnostjo, kot jo izkazuje UKGS.

[214] Za učenje jezika so sicer pomembni tudi korpusi usvajanja jezika, ki prispevajo k razumevanju procesa učenja jezika ter omogočajo analizo napak in prepoznavanje učnih potreb (prim. Stritar 2006, 134).

Funkcija »Nakazovanje, da sledimo diskurzu« (5.14*)* je v *Sporazumevalnem pragu* (str. 54) izražena z naslednjimi vzorci:

1. ***Strinjanje:*** Ja ... Razumem ... Seveda ... Točno tako ... Aha ... No vidite ...
2. ***Zadržek:*** Hm. No.
3. ***Izražanje zanimanja:*** A res? A tako? A ja?

V UKGS se v funkciji nakazovanja, da govorec sledi diskurzu, najpogosteje pojavlja *mhm* (45-krat, prim. Sliko 51); izraz v Sporazumevalnem pragu sploh ni naveden kot možni vzorec. *Seveda* se v UKGS pojavi enkrat, *aha* dvakrat, ostali izrazi iz prve točke pa se ne pojavijo. *Hm* se pojavi enkrat, *no* pa večkrat; po enkrat se pojavita vzorca *a res* in *a tako*, veliko večkrat pa se pojavi *aja*. Če bi avtorji Sporazumevalnega praga pri pisanju imeli na razpolago govorni korpus, bi morali pri izboru vzorcev za izražanje *nakazovanja, da sledimo diskurzu*, nekatere vzorce verjetno zamenjati z drugimi.

Govorni korpusni se lahko uporabljajo tudi v razredu pri pouku tujega jezika ali pri samostojnem učenju, saj ponujajo neposredni vpogled v rabo jezika. V tem primeru gre za induktivno metodo učenja – učenje iz podatkov. Z dostopom do korpusa učeči se lahko preverijo, kaj se tipično reče v določenih okoliščinah in kako se to tipično reče. Seveda ne predvidevamo, da so vsi učeči se vešči iskanja po korpusu, zato mora biti načrtovalec in koordinator takega iskanja učitelj. Možno je (in v danem trenutku tudi realno), da ima dostop do korpusa (in znanje za njegovo uporabo) samo učitelj, ki za učeče se pripravlja na korpusnih podatkih temelječe naloge. V najbolj idealnem primeru bi imeli dostop do korpusa vsi učeči se; z ustreznim znanjem za interpretacijo korpusnih podatkov bi lahko iz njega sami pridobivali jezikovne podatke.

V nadaljevanju je prikazan primer vaje, pripravljene na podlagi korpusnih podatkov; učeči se morajo iz besed z enako obliko razločiti tri različne pomene (tovrstne vaje bi bile mogoče primerna tudi za dijake, za katere slovenščina ni tuji jezik).

–	+	zdaj ena stvar na kateri trenutno	**dela** ekipa štirideset
–	+	ve da je šest ur intenzivnega	**dela** na dan [ogromno v
–	+	bistvu še z vami je kar veliko	**dela** ne mislim da
–	+	zakaj </shift=vpr> doktorat	**dela** pa rabi posnetke
–	+	<pavza> je pred nami še veliko	**dela** sicer tudi ə na
–	+	opredelitev] delite na najmanj dva	**dela** sigurno
–	+	vem] kako misliš na kakšna dva	**dela** <pavza> +G17+
–	+	+G01+ [<neraz>] <pavza> [nič ne	**dela**] +G15+ [ja mreža

Slika 57: Določanje pomena ključni besedi

Kot je bilo že nakazano, je pri zgornjem postopku pomembna zmožnost pravilne interpretacije korpusnih podatkov. Iz dobljenih informacij je namreč zlahka mogoče potegniti nesmiselne ali napačne zaključke, kar pomeni zlorabo korpusa; zato je ena izmed pomembnih nalog načrtovalcev korpusa tudi izobraževanje uporabnikov za previlno uporabo korpusnih podatkov.

9.3.4 Govorni korpus in govorne tehnologije

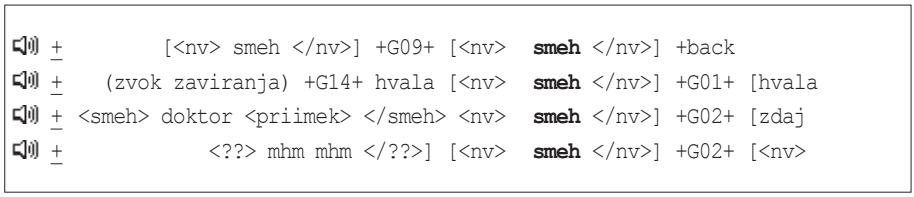
Vzporedno s prvimi raziskavami razpoznave in sinteze govora so začele nastajati tudi prve zbirke studijskih posnetkov govora, t. i. govorne zbirke, ki so služile kot izhodišče za akustične, fonetične in prozodične raziskave. Prve delujoče aplikacije, ki prepoznavajo govor v slovenščini (npr. posredovanje informacij o festivalu LENT, M-vstopnica za kino; prim. Zemljarič Miklavčič 2003, 117–118), so tako nastale izključno na podlagi govornih zbirk. Vendar se je kmalu izkazalo, da govorne zbirke ne bodo zadostovale za naprednejše aplikacije, saj v realnem govoru nenehno prihaja do dogodkov, ki v studijskih posnetkih niso predvideni, zato zavirajo delovanje aplikacij. Napake pri razpoznavanju nastajajo zaradi netekočnosti govora, obotavljanja, ponavljanja, napačnih začetkov, pa tudi zaradi kašljanja govorca, smeha, hrupa v ozadju in podobno. Nekateri slovenski govorni tehnologi so se že dokaj zgodaj zavedali pomanjkljivosti govornih zbirk, tako npr. Dobrišek opozarja na »pomanjkljivosti /govorne/ zbirke /GOPOLIS/, ki so posledica pomanjkanja splošnega znanja o slovenskem govorjenem jeziku (Dobrišek idr. 1998, 105), vendar je do bolj vsesplošne prepoznave te potrebe prišlo šele mnogo kasneje. Prvi korpus spontanega govora za potrebe govornih tehnologij je nastal v okviru že večkrat citirane disertacije D. Verdonik v sodelovanju Oddelka za slovenistiko Filozofske fakultete UL in Fakultete za elektrotehniko, računalništvo in informatiko UMB: »Temeljna teza te disertacije /je/, da se je treba pri razvoju govornih tehnologij, ki bi uspešno procesirale pogovorni govor, nasloniti na tiste veje jezikoslovja, ki preučujejo spontan govorjeni diskurz, in to v vsakdanji jezikovni rabi« (Verdonik 2006, 40). Tako je nastal korpus telefonskih poizvedovanj po turističnih informacijah TURDIS-1, ki služi kot osnova raziskav govora pri razvoju strojnega prevajanja.

Morda je bilo simbolično dokončno priznanje po potrebi za gradnjo korpusov spontanega govora s strani govornih tehnologov podano v okviru mednarodne konference Jezikovne tehnologije 2006, kjer je prvi vabljeni predavatelj Nick Campbell predstavljal zadnje dosežke pri razvoju sintetizatorja govora, namenjenega uporabi v interaktivnih dialogih med človekom in informacijskim sistemom, robotom ali govorno-prevajalno napravo (Campbell 2006, 11). Članek

opisuje več vrst neverbalnih odgovorov, ki jih je z uporabo »tradicionalnih postopkov za sintezo govora«[215] težko implementirati, in pokaže vlogo neverbalnih govornih segmentov (na eni strani smeh in mrmranje, na drugi pa tudi pogoste fraze in idiome) pri zagotavljanju povratne informacije v interaktivnem diskurzu. Segmenti, kakršne v članku raziskuje Campbell, so dostopni tudi v UKGS:

```
+           <nv> smeh </nv] +G16+ [<nv>  smeh </nv] kakšen priše
+    da ji posnamem zdaj se snema <nv>  smeh </nv> (4 sec)
+  </smeh> <nv> vzdih </nv> +G09+ <nv>  smeh </nv> +G01+ mislim
+      najprej pošlji [ko prideš] <nv>  smeh </nv> +G01+ <govo
+      [pre <lz>] +G01+ aha +G09+ <nv>  smeh </nv> +G03+ [aja v
+ne +G03+ a je <nst> scalo </nst> <nv>  smeh </nv> +G03+ no
+      Bergen] +G03+ nekako +G09+ <nv>  smeh </nv> +G03+ <pavz
+       [v Italiji] +G09+ <pavza> <nv>  smeh </nv> +G09+ <pavz
+    smeh </nv> <ime> pomoč +G09+ <nv>  smeh </nv> +G10+ <??>
+         +G01+ [mimogrede] +G09+ <nv>  smeh </nv> +G10+ <nv>
+     <priimek> +G11+ pa to +G12+ <nv>  smeh </nv> +G11+ ja saj
+        i [bez toga] </tj> +G09+ <nv>  smeh </nv> +G15+ [<okr>
+    vodne pištole [so ugotovili] <nv>  smeh </nv> +G16+ [na]
++G20+ <pavza> tri mesece je mimo <nv>  smeh </nv> +G19+ [kaj]
+        <smeh> če- če ne </smeh> <nv>  smeh </nv> bi kaj
+           smeh </nv>] ə </smeh> <nv>  smeh </nv> doktorica
+              [mhm doktor <ime>] <nv>  smeh </nv> izzvenelo
+           to čisto tako] +G02+ [<nv>  smeh </nv> ja za tem je
+       +G20+ a +G19+ halo +G19+ [<nv>  smeh </nv> je j
+          v Slovenijo +G07+ [ja] <nv>  smeh </nv> kako sem
+  na to pravi vodja poletne šole <nv>  smeh </nv> magistra
+če mogoče še zmeraj čakate +G19+ <nv>  smeh </nv> na letališču
+ <nst> jebem </nst> jim mater ej <nv>  smeh </nv> res no <nv>
+                 [pa] +G16+ [ja] <nv>  smeh </nv> to je tako
+         </smeh> več o tem +G04+ <nv>  smeh </nv> zdaj ta
+    pojdi] <pavza> ja +G10+ <??> <nv>  smeh </nv> </??> +G03+
+    tukaj bila kje +G09+ <pavza> <nv>  smeh </nv> <ime> pomoč
+        smeh </nv> res no <nv> vzdih,  smeh </nv> <nst> pička
+               [na] <pavza> [ah] <nv>  smeh </nv> <shift=vpr>
+       snemalno napravo)] +G03+ [<nv>  smeh </nv> <smeh> ja
+   po seminarju ne </smeh> +G02+ <nv>  smeh </nv> əm nekateri
+                 [mhm] ja hvala [<nv>  smeh </nv>] +??+ [<nv>
```

[215] Campbell 2006, 11; mišljene so govorne zbirke.



Slika 58: Neverbalni dogodek v govoru

9.4 ODPRTE MOŽNOSTI ZA ŠTEVILNE DRUGE RAZISKAVE

Nakazanih je bilo samo nekaj možnosti uporabe korpusa govorjene slovenščine. Predvideti je mogoče še druge raziskovalne interese specializiranih jezikoslovnih ved, npr. skladenjske analize, pragmatike, sociolingvistike, frazeologije, tudi dialektologije in drugih. Če bi bil vsaj del korpusa fonetično in prozodično označen, bi omogočal tudi fonetične in intonacijske študije. Zagotovo pa bi v govornem korpusu lahko našle zanimivo gradivo za raziskovanje tudi druge vede, npr. psihologija, sociologija, specialna pedagogika in drugi, ki jih morda še ne znamo predvideti.

Že v uvodu sem zapisala, da so največje pomanjkljivosti UKGS okrnjena demografska in besedilnovrstna sestava ter dejstvo, da korpus ni lematiziran in oblikoskladensjko označen. Kljub temu pa je bil namen gradnje UKGS dosežen, saj so bile ob gradnji preizkušene metode zbiranja in shranjevanja govorjenih besedil, preizkušena načela transkribiranja, predlagane oznake neverbalnih govornih dogodkov in nakazane nekatere možnosti korpusnih analiz. Spoznanja in izboljšave bo mogoče upoštevati pri gradnji večjega govornega korpusa.

10 Povzetek

V knjigi *Govorni korpusi* razpravljam o tem, zakaj in kako zgraditi referenčni korpus govorjene slovenščine, elektronsko zbirko transkribiranih posnetkov spontanega govora, ki bi omogočala raziskovanje govorjenega jezika.

Pri utemeljevanju odgovora na prvo vprašanje sem pregledala najodmevnejše raziskave govorjenega jezika na Slovenskem in problematiko tovrstnega raziskovanja. Izkazalo se je, da je bilo govorjenemu jeziku (predvsem spontanemu govoru) posvečene bistveno manj jezikoslovne pozornosti kot pisnemu jeziku, kar velja tudi v drugih jezikovnih okoljih. Glavne razloge za to lahko poleg drugih družbenih ter jezikoslovno teoretičnih okoliščin pripišemo dejstvu, da je bilo do nedavnega avtentično gradivo, to je posnetke govora, težje zbirati, predvsem pa težje shranjevati in pripravljati za analizo, kar pa se je z razvojem digitalne tehnologije v zadnjih desetletjih bistveno spremenilo. Tudi razvoj korpusnega jezikoslovja, ki poteka vzporedno z razvojem digitalnih tehnologij, se je na Slovenskem znašel na točki, ko je za dopolnitev jezikovne infrastrukture bistvenega pomena ravno gradnja govornega korpusa. Potencialne uporabnike referenčnega govornega korpusa lahko iščemo med jezikoslovci, specializiranimi predvsem za področja jezikovnega opisa, leksikografije, skladnje, diskurza, slovenščine kot tujega jezika in sociolingvistike, pa tudi fonetike in prozodije, če je korpus ustrezno označen. V širšem smislu je potencialnih uporabnikov govornega korpusa še več, npr. govorni tehnologi (prepoznavanje in generiranje govora), pedagogi, defektologi, sociologi, psihologi in komunikologi – vseh danes verjetno sploh ni mogoče predvideti. Z utemeljevanjem potrebe po gradnji govornega korpusa in z izkazovanjem tehnološke pripravljenosti slovenskega jezikovnega okolja na gradnjo takega korpusa utemeljujem izbrano temo in namen raziskave.

V drugem poglavju – *Govorni korpusi* – podrobneje predstavljam 12 govornih korpusov, ki so nastali v preteklem poltretjem desetletju (1980–2007); najprej predstavljam pionirsko delo na področju gradnje govornih korpusov, nato pa tiste korpuse, ki so pomembni zaradi velikosti, dobre uravnoteženosti ali tehnološkega napredka, ki so ga prinesli v razvoj gradnje govornih korpusov. Predstavljena sta korpusa britanske angleščine London-Lund in Lancaster/IBM (MARSEC), korpus govorjene ameriške angleščine Santa Barbara, mednarodni korpus angleščine ICE, govorni komponenti korpusov BNC in *Bank of English*, govorna komponenta korpusa Češkega nacionalnega korpusa, Budimpeštanski sociolingvistični intervjuji, govorni korpus britanske najstniške angleščine (COLT), švedski govorni korpus, nizozemski govorni korpus in korpusni paket francoščine, italijanščine, španščine in portugalščine (C-ORAL-ROM).

V devetdesetih letih prejšnjega stoletja je bil najvplivnejši referenčni vir med go-

vornimi korpusi Britanski nacionalni korpus, ki je s svojimi načeli gradnje ter z industrijskimi razsežnostmi produkcije postavil temelje novemu obdobju gradnje govornih korpusov; k njegovi vplivnosti je veliko pripomogla tudi dostopnost dokumentacije o gradnji korpusa. Z današnjega stališča velja za tehnološko najnaprednejšega Nizozemski govorni korpus, ki je po velikosti skoraj primerljiv z govorno komponento BNC, v tehnološkem smislu pa ponuja mnogo več, saj omogoča sinhroni dostop do zvočnih posnetkov in širši spekter jezikoslovnih analiz: gradivo je v celoti ortografsko transkribirano in oblikoskladenjsko označeno, poleg tega pa je del gradiva (1 milijon besed) transkribiran fonetično, del prozodično (250.000 besed), del (1 milijon besed) pa tudi skladenjsko označen.

Nadaljevanje je namenjeno iskanju odgovorov na vprašanje, *kako* zgraditi referenčni govorni korpus slovenskega jezika. V tem okviru sem si zastavila štiri temeljne cilje: določiti shemo za zajem besedil v korpus, določiti načela transkribiranja, izdelati priporočila za označevanje in zgraditi učni korpus za testiranje postavljenih načel.

V tretjem poglavju – *Zajem besedil v govorni korpus* – predstavljam shemo, ki zagotavlja reprezentativnost in uravnoteženost zbranega gradiva, s tem pa možnost pridobivanja relevantnih podatkov o jeziku iz korpusa. Osnova za izdelavo sheme so sociolingvistične študije in demografske analize govorcev ter nova taksonomija govorjenih besedil, prilagojena potrebam gradnje referenčnega govornega korpusa. Predlagam, da bi besedila zbirali s kombiniranjem demografske in besedilnovrstne metode. Pri tem sem upoštevala izkušnje ob gradnji drugih velikih korpusov, znotraj preizkušenih modelov pa iskala najprimernejše prilagoditve za slovenski jezik in za slovensko jezikovno situacijo. V prvem koraku predvidevam zbiranje besedil z reprezentativnim vzorcem govorcev. Načrtovani kriteriji za sestavo vzorca so spol, starost, izobrazba, regijski izvor in prvi jezik. Demografski podkorpus, ki bi nastal na ta način, bi vključeval zasebna in uradna dialoška (in morebitna monološka) besedila, govorjena v osebnem stiku ali po telefonu; zbrana besedila naj bi obsegala najmanj polovico celotnega govornega korpusa, vključevala pa bi večino zasebnih besedil celotnega korpusa. V drugem koraku predvidevam dopolnitev konverzacijskega podkorpusa z besedili, zbranimi na podlagi besedilnovrstne taksonomije; v tem delu bi zajeli predvsem javna besedila, dialoge in monologe, govorjene v osebnem stiku, po radiu, TV ali telefonu, z različno stopnjo formalnosti. S kombiniranjem demografske in besedilnovrstne metode vzorčenja dobimo bolj reprezentativen govorni korpus, kot bi ga dobili, če bi se odločili samo za eno izmed obeh metod. Postopek zajemanja pa je nekoliko olajšan s predlogom, da demografski podkorpus vključuje predvsem zasebna, besedilnovrstni pa predvsem javna besedila. Načrtovana shema mora biti dovolj

prožna, da se lahko med gradnjo korpusa ali še kasneje spreminja, če se za to pokaže potreba.

Druga temeljna odločitev, ki jo je potrebno sprejeti ob načrtovanju gradnje govornega korpusa, zadeva način zapisovanja govora. Problematika vključuje zapis govorne verige – transkribiranje, pa tudi za zapis dogodkov, ki spremljajo govor in jih lahko imamo za sestavni del govornega dogodka; izbira postopkov je tudi tu, enako kot pri zajemanju besedil, neločljivo povezana z namenom – gradnjo velikega referenčnega govornega korpusa. V četrtem poglavju – *Označevanje in transkribiranje govorjenih besedil* – tako predstavljam mednarodne transkripcijske standarde in priporočila za zapis govorjenih besedil TEI in EAGLES ter nekatere primere prozodičnega, fonetičnega in ortografskega transkribiranja obstoječih govornih korpusov. Pri predlogu priporočil za zapis govorjene slovenščine sem upoštevala zbrane mednarodne izkušnje, vendar je pri končnih odločitvah vedno prevladala specifičnost slovenske jezikovne situacije in jezika. V petem poglavju – *Predlog priporočil za transkribiranje besedil v govornem korpusu* – predlagam transkripcijsko načelo, ki bi po mojem zagotavljalo ustrezen zapis govorjene slovenščine za gradnjo referenčnega korpusa, in sicer razširjeno ortografsko transkripcijo brez ločil. Izraz »razširjena« se nanaša na prilagoditev zapisa nekaterim značilnostim govora, ki zelo odstopajo od knjižnega zapisa in ki so v govoru zelo pogoste. Predlagani način obeta, da bi zapis ohranil nekatere najbolj tipične oblikoskladenjske in leksikalne značilnosti govorjenega jezika, npr. nedoločnik brez končnega i-ja, deležnik na –l v dvojinski in množinski obliki brez končnega vokala ali z neknjižno varianto končnice (*smo delal*, *bova delale*), določne pridevnike v imenovalniku in tožilniku brez končnega i-ja, členek *una,-a,-o* v vseh oblikah, členek *en,-a,-o* v vseh oblikah, določni člen *ta* v vseh oblikah, besede *pol* (v pomenu potem), *tle, čmo* in druge (podrobnejša predstavitev je v poglavju *Končni predlog priporočil za transkribiranje govorjene slovenščine*). Predviden je tudi hkratni dostop do zvočnih posnetkov besedila, kar bi omogočalo nadgradnjo korpusa s fonetično transkripcijo, poleg tega pa transkripcijski program, ki ga predlagam za uporabo pri gradnji govornega korpusa (*Praat*), omogoča akustične analize besedil in s tem nadgradnjo s prozodičnimi oznakami.

V nadaljevanju razpravljam o oznakah v govornem korpusu. Besedilo, ki ga pripravljamo za vključitev v govorni korpus, dodatno označimo predvsem zato, da bi zapis približali avtentični obliki. Velik del označevanja poteka hkrati s transkribiranjem in o teh oznakah govorim v šestem poglavju – *Predlog priporočil za označevanje besedil v govornem korpusu*. Nabor oznak sem zasnovala po priporočilih EAGLES; vse oznake sem preizkusila pri transkribiranju učnega korpusa, v končni predlog priporočil za označevanje govornega korpusa slovenščine pa sem

vključila samo tiste, ki so se izkazale ustrezne in potrebne za slovenski jezik in za referenčni korpus, to pa so oznake za premor, napačni začetek, prekrivni govor ter za neverbalne in nekomunikacijske zvoke; v transkripciji tudi označimo, če je besedilo brano, izgovorjeno v tujem jeziku ali nerazumljivo, s posebnimi oznakami pa zakrijemo osebna imena in druge osebne podatke (Tabela 16).

V sedmem poglavju – *Oznake v glavah dokumentov* – govorim o metajezikovnih oznakah, ki so transkripcijam dodane naknadno. Najprej je predstavljena dokumentacija, kjer se ti podatki zbirajo. Na Zbirnem listu podatkov besedila (Slika 37) in Identifikacijskem listu govorca (Slika 38) so zbrani podatki, ki so pomembni za označevanje, razvrščanje in določanje posameznih vrst besedil; nabor podatkov je določen med načrtovanjem zajema besedil, podatke pa vpisuje oseba, ki posamezno besedilo snema. Na podlagi zbirnih in identifikacijskih listov uredniki izdelajo glave dokumentov. Vanje shranijo podatke o posnetku, transkripciji, taksonomski razvrstitvi besedila in govorcih. Podatki so dostopni uporabnikom korpusa in jim pomagajo pri razjasnjevanju okoliščin nastanka besedila, pri razumevanju besedila, lahko pa služijo tudi kot izhodišče za analizo. Med korpusno dokumentacijo sodi tudi *Dovoljenje za uporabo – odstop avtorskih pravic* (Slika 39). Večkrat je bilo poudarjeno, kako pomembno je imeti za vsa besedila in vse transkripcije, ki jih želimo vključiti v korpus, urejene avtorske pravice. Za učni korpus sem sama izdelala primer *Dovoljenja za uporabo*, v primeru gradnje večjega govornega korpusa pa bi morali za pravilen prenos avtorskih pravic poskrbeti pravni strokovnjaki.

Zadnji dve poglavji namenjam predstavitvi učnega korpusa govorjene slovenščine, ki je nastal na Oddelku za kulturo, jezik in jezikovne tehnologije (AKSIS) na Univerzi v Bergnu na Norveškem in je služil kot model za preverjanje ustreznosti načel za zajem besedil ter ustreznosti nabora oznak in transkripcijskih načel. V osmem poglavju – *Gradnja učnega korpusa govorjene slovenščine* – so dokumentirani posnetki in govorci korpusa, predstavljena pa je tudi njegova sestava s stališča demografskih (Tabela 19) in besedilnovrstnih kriterijev (Tabela 20). Pri demografskih kriterijih je bila dosežena manjša stopnja uravnoteženosti, večjo raznolikost besedil mi je uspelo doseči z besedilnovrstno taksonomijo; neuravnoteženo sestavo učnega korpusa je treba upoštevati pri nadaljnji uporabi korpusnih podatkov, spoznanja in izboljšave pa bi bilo mogoče upoštevati pri gradnji večjega govornega korpusa.

Kljub temu ima učni korpus tudi uporabno funkcijo in je z njim mogoče nakazati možnosti uporabe govornega korpusa. V zadnjem poglavju – *Primeri iskanja po učnem korpusu* – predstavljam možnosti dostopanja do besedil in iskanja z ome-

jevanjem po demografskih ali besedilnovrstnih kriterijih; prikazane so tudi zgledi iskanja na besedni ravni, na ravni diskurza, predstavljene so možnosti uporabe korpusa pri poučevanju in učenju tujega jezika ter nakazane povezave med govornimi korpusi in govornimi tehnologijami.

S tem je bil izpolnjen zastavljeni cilj raziskave – izoblikovana so bila teoretična in metodološka izhodišča za gradnjo reprezentativnega in uravnoteženega govornega korpusa. Naslednja naloga slovenskega korpusnega jezikoslovja je začeto delo pripeljati do konca – zgraditi referenčni korpus govorjenega jezika za slovenščino.

Summary

(prevedla/translated by Agnes Pisanski Peterlin)

Spoken Corpora is a book that discusses the whys and hows of building a reference corpus of spoken Slovene language, i.e. an electronic collection of transcribed recordings of spontaneous speech which could be used to examine spoken language.

In the Introduction I provide an overview of the most influential studies of spoken language in Slovenia and the issues involved in this type of research, in order to present the grounds for the answer to the first question. The overview shows that spoken language, above all spontaneous speech, has received considerably less attention than written language in terms of linguistic research; this has also been established for other linguistic communities. The main reason for this is the fact that until recently, authentic material, that is speech recordings, was more complicated to collect and more difficult to store and prepare for analysis: all of this, however, has changed significantly with the development of digital technology in the past few decades. Moreover, in Slovenia, the development of corpus linguistics, which is parallel to the development of digital technology, has now reached a point where the construction of a spoken corpus is essential for completing the linguistic infrastructure. Potential users of a reference corpus of spoken language include linguists focusing above all on fields such as language description, lexicography, syntax, discourse, Slovene as a foreign language and sociolinguistics, as well as phonetics and prosody, if the corpus is suitably annotated. More generally, the category of potential users may also comprise researchers from other disciplines, such as speech technology (speech analysis and synthesis), education, special education, sociology, psychology and communication studies; it is in fact probably impossible to predict all the disciplines. The subject matter and the aim of the research are further supported by establishing the need to build a spoken corpus and by ascertaining the technological possibilities available for this purpose in the Slovene linguistic community.

In Chapter Two – *Spoken Corpora* – 12 corpora of spoken language created in the past two decades and a half (1980–2007) are described: I first present the pioneer work in the field of spoken corpus collection, and then introduce the corpora important because of their size, balance or advances in technology which they have brought to the process of spoken corpus collection. The corpora presented include two corpora of British English, the London-Lund Corpus and the Lancaster/IBM Corpus (the MARSEC Corpus), a corpus of Spoken American English (the Santa Barbara Corpus), the International Corpus of English (the ICE Corpus), the spoken components of the BNC and the Bank of English Corpus, the spoken component of the Czech National Corpus, the Budapest Sociolinguistic Interview, the spoken Bergen Corpus of London Teenage Language (COLT), the Swedish Spoken Language Corpus, the Spoken Dutch Corpus and the Integrated

Reference Corpora for Spoken Romance Languages, the C-ORAL-ROM Project for French, Italian, Spanish and Portuguese.

Among the spoken corpora, the BNC was the most influential source of reference in the 1990s; with its principles of corpus design and industrial-size production, it laid the foundations for a new era of spoken corpus design; the availability of documentation on corpus design contributed to its influence. From today's point of view, the Spoken Dutch Corpus is the most advanced; in size it is almost comparable to the spoken component of the BNC, while it offers much more in terms of technology, since it enables a synchronous approach to audio recordings and offers a broader range of linguistic analyses: orthographic transcription and morphosyntactic tagging are provided for all the material; for certain segments of material, phonetic transcription (1 million words), prosodic transcription (250 000 words) and syntactic tagging (1 million words) are provided as well.

The next part of the book is dedicated to the question how to build a reference corpus of spoken Slovene. For this purpose, I proposed four main objectives: to determine corpus design criteria, to set up transcription standards, to create recommendations for tagging and to compile a pilot corpus to test the established principles.

In Chapter Three – *Corpus Design Criteria* – I present the framework designed to ensure that the materials selected are representative and balanced, which guarantees that it is possible to obtain relevant language data from the corpus. Sociolinguistic studies and demographic analyses of the speakers and a new taxonomy of spoken discourse, adapted to the needs of reference spoken corpus design, form the basis for the framework. The framework was designed on the basis of sociolinguistic studies and demographic analyses of speakers and a new taxonomy of spoken texts adjusted to the needs of spoken reference corpus building. I propose that the texts are collected by combining the demographic and the contextual method. I took into consideration the experience gained in building other large corpora and sought the most suitable adjustment for the Slovene language and the Slovene linguistic situation. As the first step, I propose collecting materials using a representative sample of speakers. The criteria for sample selection include the speaker's gender, age, education, region of origin and first language. The demographic component of the corpus collected in this way would encompass personal and official dialogue (and possible monologue) texts spoken in face-to-face and telephone conversation; the texts collected would represent at least one half of the entire spoken corpus and would include most of the personal conversation of the entire corpus. As the second step, I propose complementing the

conversation subcorpus with texts collected on the basis of text-type taxonomy; in this part, public texts, dialogues and monologues, spoken in face-to-face conversation, on the radio, television or over the telephone, with different degrees of formality would be collected. By combining the demographic and the contextual method, the spoken corpus is more representative than it would have been if a single method was chosen. The sampling procedure is made somewhat easier by the proposition that the demographic component of the corpus should include above all personal conversation, while the contextual component of the corpus should encompass above all public conversation. The proposed framework needs to be flexible enough to change in the process of corpus building or at a later point, if necessary.

The second basic decision in connection with spoken corpus design concerns the method of transcription. The problem involves the transcription of the speech chain, as well as the tagging of events accompanying the speech which may be considered part of the speech event: just as in corpus design, here, procedure selection and the purpose – large reference spoken corpus building – are inextricably bound. In Chapter Four – *Spoken text tagging and transcription* – I thus present the TEI and EAGLES international transcription standards and recommendations for spoken text transcription, as well as selected examples of prosodic, phonetic and orthographic transcription of existing spoken corpora. In my recommendations for the transcription of spoken Slovene, I took into consideration the experience from other linguistic communities, but the final decision was always based on the specific nature of the Slovene linguistic situation and the Slovene language. In Chapter 5 – *A draft of recommendations for the transcription of a spoken corpus*– I propose a transcription standard which would, in my opinion, guarantee a suitable transcription of spoken Slovene for reference corpus building, i.e. modified orthographic transcription without punctuation. The term "modified" applies to the adaptation of the transcription to certain characteristics of speech which differ substantially from the written form and are very frequent in speech. The proposed method is an attempt to retain some of the most typical morphosyntactic and lexical characteristics of spoken language, such as the infinitive without the word-final *-i*, the *-l* participle in the dual and plural forms without the final vowel, or the colloquial variant of the ending (*smo delal*, *bova delale*), the definite form of the adjective in the nominative and the accusative without the word-final *-i*, the particle *un, -a, -o* in all forms, the particle *en,-a,-o* in all forms, the definite article *ta* in all forms, words such as *pol* (meaning later), *tle, čmo* and others (for a more detailed explanations, see *The final draft of recommendations for the transcription of spoken Slovene*). The draft also proposes simultaneous access to the recordings, which would enable a phonetic

transcription of the corpus carried out at a later point, while Praat, the transcription program I recommend for building the spoken corpus, enables an acoustic text analysis, thus enabling prosody markup.

In Chapter six - *A draft of recommendations for tagging texts in a spoken corpus* - the tagging of spoken corpora is discussed. The text to be included in a spoken corpus is tagged to make the transcription as similar to the authentic form as possible. A large part of the tagging is carried out at the same time as the transcription. I based the tag set on the EAGLES recommendations and tested all the tags on the transcription of the pilot corpus; in the final draft of recommendations for tagging a spoken corpus of Slovene, I only included those tags which turned out to be useful and necessary for the Slovene language and for a reference corpus. Those tags include a pause, a false start, an overlap and non-verbal and non-communicative sounds; the transcription also includes information on whether the text was read, spoken in a foreign language or incomprehensible, while special tags are used to encode personal names and other personal data (c.f. Table 16).

In Chapter Seven – *Header tags* – I discuss metalinguistic tags which are added to transcriptions at a later point. First the documentation used for collecting this type of data is presented. In the Text information sheet (see Figure 37) and the Speaker identification sheet (Figure 38), information essential for tagging, classification and determining the text type is collected; the set of data is determined in the process of corpus design, the information is provided by the person recording the text. On the basis of the Text information sheet and the Speaker identification sheet, the editors produce document headers. These include information on the recording, transcription, taxonomic category of the text and speakers. This information is available to corpus users and may be useful for understanding the circumstances in which the text was produced, for understanding the text itself and as the starting point for analysis. Corpus documentation includes *Permission Clearance – Copyright transfer* (Figure 39). It has been stressed how important it is to obtain copyright for all the texts and transcriptions to be included in the corpus. I drafted a sample of *Permission Clearance* for the pilot corpus; however, it would be necessary to consult legal experts to ensure proper copyright transfer for a larger spoken corpus.

The last two chapters present the pilot corpus of spoken Slovene which was built at the Departement for Culture, Language and Language Technologies at the University of Bergen (Norway) and was used to test the suitability of the principles of corpus design and the suitability of the tag set and transcription principles. In Chapter Eight – *Compiling a pilot corpus of spoken Slovene* – the recordings

and the speakers included in the corpus are documented; at the same time the demographic criteria (Table 19) and text-type criteria (Table 20) of corpus composition are presented. The demographic criteria were less balanced, while the text-type taxonomy was somewhat more diverse; the fact that the composition of the pilot corpus is somewhat unbalanced needs to be taken into account in further work on corpus data, while the findings and improvements may be of use in the compilation of a larger spoken corpus.

The pilot corpus may also demonstrate the possibilities for using a spoken corpus. In the last chapter – *Examples of searching pilot corpus* – I present the possibility to access texts and to run a corpus search by limiting the demographic or text-type criteria; examples of word-level and discourse-level searching and corpus use in foreign language teaching and learning are presented as well; finally, the connections between spoken corpora and language technologies are pointed out.

The objective of this book – to establish a theoretical and methodological basis for building a representative and balanced spoken corpus – was thus fulfilled. The Slovene corpus linguists are now facing a new task: to find a way to the final realisation of the project – the building of the spoken component of a reference corpus in the near future.

Literatura

AIJMER, Karin in ALTENBERG, Bengt (ur.), 1991: *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London in New York: Longman.

ALLWOOD, Jens, GRÖNQUIST, Leif, AHLSEN, Elisabeth in GUNNARSON, Magnus, 2001: Annotations and Tools for Activity Based Spoken Language Corpus. *2nd SIGdial Workshop on Discourse and Dialogue, workshop proceedings*. Aalborg, Danska, 1.–2. september 2001.

ALLWOOD, Jens, 1998: Some Frequency Based Differences between Spoken and Written Swedish. *Proceedings from the XVIth Scandinavian Conference of Linguistics.* Department of Linguistics, University of Turku. http://www.ling.gu.se/~shirley/jenspublications/docs076-100/084.pdf

ALLWOOD, Jens, BJÖRNBERG, Maria, GRÖNQUIST, Leif, AHLSEN, Elisabeth in OTTESJÖ, Cajsa, 2001: *The Spoken Language Corpus at the Department of Linguistics.* Göteborg University. http://www.qualitative-research.net/fqs/fqs-eng.htm

ANDERSEN, Gisle, 2000: *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

ARHAR, Špela, 2007: Korpus FidaPLUS: nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52, 2. [95]–110.

ARHAR, Špela, 2004: *Gradnja specializiranega korpusa.* Diplomska naloga. Mentorja: M. Stabej in V. Gorjanc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

ASTON, Guy, BERNARDINI, Silvia in STEWART, Dominic (ur.), 2004: *Corpora and Language Learners.* Studies in Corpus Linguistics 17. Amsterdam/Philadelphia: John Benjamin's Publishing Company.

ATKINS, Sue, CLEAR, Jeremy in OSTLER, Nicholas, 1992: Corpus Design Criteria. *Literary and Linguistics Computing* 7/1. 1–16.

BALAŽIC BULC, Tatjana, 2008: *Raba in funkcija konektorjev v jezikoslovnem diskurzu na primeru slovenščine in hrvaščine kot prvega in tujega jezika*. Doktorska disertacija. Mentorica: V. Požgaj-Hadži; somentor: V. Gorjanc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slavistiko.

BERGLUND, Ylva, 1999: Exploiting a Large Spoken Corpus: An End-user's Way to the BNC. *International Journal of Corpus Linguistics* 4/1. 29–52.

BEŠTER, Marja, KRIŽAJ ORTAR, Martina, KONČINA, Marija, BAVDEK, Mojca, POZNANOVIČ, Mojca, AMBROŽ, Darinka in ŽIDAN, Stanislava, 1999: *Na pragu besedila. Učbenik za slovenski jezik v 1. letniku gimnazij, strokovnih in tehniških šol*. Ljubljana: Založba Rokus.

BEŠTER, Marja, 1996: Povojni generaciji Slovencev po svetu: narodnostna opredelitev, znanje in raba slovenščine. Vidovič Muha, A. (ur.): *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 135–151.

BEŠTER, Marja, 1994: Tip besedila kot izrazilo sporočevalčevega namena. *Uporabno jezikoslovje* 2. *Analiza diskurza* (tematska številka, ur. I. Kovačič). 44–52.

BIBER, Douglas, 1993: Representativeness in Corpus Design. *Literary and Linguistics Computing* 8/4. 243–257.

BRITISH National Corpus User Reference Guide, 2000. Ur. Lou Burnard. http://www.natcorp.ox.ac.uk/World/HTML/urg.html

BURNARD, Lou, 2000: Where Did We Go Wrong? A Retrospective Look at the Design of the BNC. *6th International Conference, »Spoken Italian«, congress proccedings*. Duisburg, 28. 6.–2. 7. 2000. http://users.ox.ac.uk/~lou/wip/silfitalk.html

BURNARD, Lou, 1995: The Text Encoding Initiative: An Overview. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 69–81.

CAMPBELL, Nick, 2006: Speech Synthesis and Discourse Information. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 11–16.

CHAFE, Wallace, DU BOIS, John in THOMPSON Sandra, 1991: Towards a New Corpus of Spoken American English. Aijmer, K. in Altenberg, B. (ur.): *English Corpus Linguistics*. London in New York: Longman. 64–82.

CINDRIČ, Alojz, 2002: Vpliv dunajske univerze na oblikovanje slovenskega izobraženstva: statistična slika študentov s Kranjske; študijske smeri, krajevni in socialni izvor. Vodopivec, P. (ur.): *Slovenci v Evropi.* Historia 5, znanstvena zbirka Oddelka za zgodovino, Filozofska fakulteta Univerze v Ljubljani. 7–4.

CROWDY, S., 1995: The BNC Spoken Corpus. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 224–234.

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

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

ČERMÁK, František, 2001: *Pražský mluvený korpus*. http://ucnk.ff.cuni.cz/pmk_bonito.html

ČERMÁK, František, 1997: Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics* 2/2. 181–197.

ČERMAK, František in SCHMIEDTOVÁ, Věra: *The Czech National Corpus Project: Its Structure and Use.* http://ucnk.ff.cuni.cz/doc/czechnationalcorpus.doc

DE SMEDT, Koenraad, GARDINER, Hazel, ORE, Espen, ORLANDI, Tito, SHORT, Harold, SOUILLOT, Jacques in VAUGHAN, William, 1999: *Computing in Humanities Education: A European Perspective*. University of Bergen,

The HIT Centre.
DOBRIŠEK, Simon in ŽGANEC GROS, Jerneja, 1998: GOPOLIS: slovenska podatkovna zbirka govorjenih poizvedovanj. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 105–108.
DOUGLAS, Fiona M., 2003: The Scottish Corpus of Texts and Speech: Problems of Corpus Design. *Literary and Linguistics Computing* 18/1. Oxford University Press. 23–37.
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
ERJAVEC, Tomaž in ŽGANEC GROS, Jerneja (ur.), 2006: *Jezikovne tehnologije*. Zbornik 5. slovenske in 1. mednarodne konference Jezikovne tehnologije. 9. mednarodna multikonferenca Informacijska družba IS 2006. 9.–10. oktober 2006. Ljubljana: Institut Jožef Stefan.
ERJAVEC, Tomaž, 2003: Označevanje korpusov. *Jezik in slovstvo* 48, 3/4. [59]–76.
ERJAVEC, Tomaž in ŽGANEC GROS, Jerneja (ur.), 2002: *Jezikovne tehnologije*. Zbornik B 5. mednarodne multikonference Informacijska družba IS'2002. 14.–15. oktober 2002. Ljubljana: Institut Jožef Stefan.
ERJAVEC, Tomaž in ŽGANEC GROS, Jerneja (ur.), 2000: *Jezikovne tehnologije za slovenski jezik*. Mednarodna multi-konferenca Informacijska družba. Zbornik konference. 17.–18. oktober 2000. Ljubljana: Institut Jožef Stefan.
ERJAVEC, Tomaž in ŽGANEC GROS, Jerneja (ur.), 1998a: *Jezikovne tehnologije za slovenski jezik*. Mednarodna multi-konferenca Informacijska družba. Zbornik konference. 6.–7. oktober 1998. Ljubljana: Institut Jožef Stefan.
ERJAVEC, Tomaž, 1998b: Oznake korpusa FIDA. *Uporabno jezikoslovje* 6. *Jezikovne tehnologije* (tematska številka, ur. Z. Kačič). 85–95.
ERJAVEC, Tomaž, 1996/97: Računalniške zbirke besedil. *Jezik in slovstvo* 42, 2/3. 81–95.
FEKONJA, Urška, 2004: *Razvoj otrokovega govora v različnih socialnih kontekstih.* Doktorska disertacija. Mentorica: L. Marjanovič Umek; somentorica: S. Kranjc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani.
FERBEŽAR, Ina, KNEZ, Mihaela, PIRIH SVETINA, Nataša, SCHLAMBERGER BREZAR, Mojca, STABEJ, Marko, TIVADAR, Hotimir in ZEMLJARIČ MIKLAVČIČ, Jana, 2004a: *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani in Ministrstvo RS za šolstvo, znanost in šport.
FERBEŽAR, Ina in PIRIH SVETINA, Nataša, 2004b: Certificiranje slovenščine

kot drugega/tujega jezika – zgodovina in perspektive. *Jezik in slovstvo* 48, 3/4. [17]–33.

GANTAR, Polona, 2007: *Stalne besedne zveze v slovenščini: korpusni pristop*. Ljubljana: ZRC SAZU.

GETTING Started on a Corpus. Cornell University. http://www.instruct1.cit.cornell.edu/courses/ling390/getting_started.htm

GORJANC, Vojko, in LOGAR, Nataša, 2007: Od splošnih do specializiranih korpusov – načela gradnje glede na njihov namen. Irena Orel (ur.): *Obdobja 24.* Ljubljana : Filozofska fakulteta, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 637–650.

GORJANC, Vojko, 2005a: *Uvod v korpusno jezikoslovje*. Zbirka Zrenja. Domžale: Založba Izolit.

GORJANC, Vojko in KREK, Simon (ur.), 2005b: *Študije o korpusnem jezikoslovju: zbornik.* Ljubljana: Krtina.

GORJANC, Vojko, KREK, Simon in GANTAR, Polona, 2005c: Slovenska leksikalna podatkovna zbirka. *Jezik in slovstvo* 50, 2. [3]–19.

GORJANC, Vojko in JURKO, Primož, 2004: Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49, 3/4. [49]–62.

GORJANC, Vojko, 2003: Korpusi in jezikoslovje. *Jezik in slovstvo* 48, 3/4. [19]–27.

GORJANC, Vojko, 2002a: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov*. Doktorska disertacija. Mentorica: A. Vidovič Muha. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

GORJANC, Vojko, 2002b: Jezikovna infrastruktura: kje je tu slovenščina? Krakar-Vogel, B. (ur.): *Zbornik predavanj 38. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti. 257–270.

GORJANC, Vojko in VINTAR, Špela, 2000: Iskanja po korpusu slovenskega jezika FIDA. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 20–26.

GORJANC, Vojko, 1999: Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. Kržišnik, Erika (ur.): *Zbornik 35. seminarja slovenskega jezika, literature in kulture.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti. 47–59.

GÖTEBORG Transcription Standard. Version 6.4. Department of Linguistics, Göteborg University. http://www.ling.gu.se:8000/magnus/papper/transcription_standard.pdf

GREENBAUM, Sidney (ur.), 1996: *Comparing English Worldwide: The Interna-*

tional Corpus of English. Oxford: Clarendon Press.
GREENBAUM, Sidney, 1991: The Development of the International Corpus of English. Aijmer, K. in Altenberg, B. (ur.): *English Corpus Linguistics*. London in New York: Longman. 83–91.
GREENBAUM, Sidney in SVARTVIK, Jan, 1990: The London-Lund Corpus of Spoken English. Svartvik, J.: *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund University Press.
http://helmer.aksis.uib.no/icame/london-lund/
GROS, Jerneja, MIHELIČ, France, DOBRIŠEK, Simon, ERJAVEC, Tomaž in ŽGANEC, Mario, 2000: A Phonetically and Prosodically Annotated Slovene Speech Corpus. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 27–30.
GUZEJ, Jožica, 1989/90: Vpliv migracij na jezik in govor posameznika. *Jezik in slovstvo* 35, 3. 52–57.
HASLERUD, Vibecke in STENSTRÖM, Anna-Brita: The Bergen Corpus of London Teenager Language (COLT). Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 99–112.
HLADKÁ, Zdeňka: *Brněnský mluvený korpus*.
http://ucnk.ff.cuni.cz/bmk.html
HRIBAR, Nataša, 2003: *Skladenjska razčlenitev sodobnega slovenskega parlamentarnega jezika.* Magistrsko delo. Mentorica: A. Vidovič Muha. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.
IZRE'EL, Shlomo, HARY, Benjamin in RAHAV, Giora, 2001: Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6/2. 171–197.
http//:www.tau.ac.il/humanities/semitic/cosih.html
JAKOPIN, Primož, 2001: Slovenski nacionalni korpus – idejni osnutek projekta. *Jezikoslovni zapiski* 7, 1–2. Ljubljana: Založba ZRC. 411–417.
JEZIKOVNE tehnologije za slovenščino 2003. *Jezik in slovstvo* 48, 3/4 (tematska številka, ur. V. Gorjanc).
JEZIKOVNE tehnologije 1998. *Uporabno jezikoslovje* 6 (tematska številka, ur. Z. Kačič). Ljubljana: FDV, Inštitut za družbene vede.
JOHANSSON, Stig, 1995: The Approach of the Text Encoding Initiative to the Encoding of Spoken Discourse. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 82–98.
JOHANSSON, Stig, 1991: Times Change, and So Do Corpora. Aijmer, K. in Altenberg, B. (ur.): *English Corpus Linguistics*. London in New York: Longman. 305–314.
KAČIČ, Zdravko, 2002: Pomen združevanja raziskovalnih potencialov pri preseganju jezikovnih pregrad v okviru jezikovnih tehnologij naslednjih generacij.

Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 111–115.

KAČIČ, Zdravko in HORVAT, Bogomir, 1998: Izgradnja infrastrukture, potrebne za razvoj govorne tehnologije za slovenski jezik. *Jezikovne tehnologije za slovenski jezik.* Ljubljana: Institut Jožef Stefan. 100–104.

KAISER, Janez in KAČIČ, Zdravko, 1998: Razvoj slovenske baze izgovorjav SpeechDat. *Uporabno jezikoslovje* 6. *Jezikovne tehnologije* (tematska številka, ur. Z. Kačič). 51–57.

KALIN GOLOB, Monika, 2004: Moderno in modno v publicističnem spletu vplivanja ter stilu slovenskih novinarskih besedil. Stabej, M. (ur.): *Zbornik predavanj 40. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 48–57.

KALIN GOLOB, Monika, 2003: Jezikovna kultura, jezikovno načrtovanje in evropsko združevanje. Vidovič Muha, A. (ur.): *Obdobja 20*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko. 255–270.

KALIN GOLOB, Monika, 2000: *Jezikovne reže*. Ljubljana: GV revije.

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.

KENNEDY, Graeme, 1998: *An Introduction to Corpus Linguistics.* Studies in Language and Linguistics. London in New York: Longman.

KNOWLES, Gerry, 1995: Converting a Corpus into a Relational Database: SEC Becomes MARSEC. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 208–223.

KOPŘIVOVÁ, Marie in WACLAWIČOVÁ, Martina, 2005: Construction of Spoken Corpus Based on the Material from the Language Area of Bohemia. *Computer Treatment of Slavic and East European Languages*. Proceedings of Third International Seminar SLOVKO. Bratislava, 10.–12. november 2005. 137–140.

KRAJNC IVIČ, Mira, 2008: *Zasebni dvogovori*. Doktorska disertacija. Mentorica: A. Vidovič Muha. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

KRANJC, Simona, 1999: *Razvoj govora predšolskih otrok*. Ljubljana: Znanstveni inštitut Filozofske fakultete.

KRANJC, Simona, 1998: Govorjena besedila in korpus slovenskega jezika. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 109–112.

KRANJC, Simona, 1996/97: Govorjeni diskurz. *Jezik in slovstvo* 42, 7. 307–320.

KREK, Simon, 2004: Slovarji serije *COBUILD* in formalizacija definicijskega jezika. *Jezik in slovstvo* 49, 2. [3]–16.

KREK, Simon, 2003: Jezikovni priročniki in novi mediji. *Jezik in slovstvo* 48, 3/4. [99]–46.

KRŽIŠNIK, Erika, 1997: Kdo govori kako. Derganc, A. (ur.): *Zbornik predavanj 33. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti.

KRŽIŠNIK, Erika, 2002: Knjižnojezikovna norma v »argentinskoslovenski« Svobodni Sloveniji. Krakar, B. (ur.): *Zbornik predavanj 38. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti.

LEECH, Geoffrey, MYERS, Greg in THOMAS, Jenny (ur.), 1995: *Spoken English on Computer: Transcription, Mark-up and Application*. New York: Longman Publishing.

LEECH, Geoffrey, 1991: The State of the Art in Corpus Linguistics. Aijmer, K., in Altenberg, B. (ur.): *English Corpus Linguistics*. London in New York: Longman. 8–29.

LLISTERI, Joakim, 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

LOGAR, Nataša, in ARHAR, Špela, 2008: *Kaj početi s korpusom strokovnih besedil KoRP*. Ljubljana : Fakulteta za družbene vede Univerze v Ljubljani.

LOGAR, Nataša, 2007: *Korpusni pristop k pridobivanju in predstavitvi jezikovnih podatkov v terminoloških slovarjih in terminoloških podatkovnih zbirkah*. Doktorska disertacija. Mentorica: M. Kalin Golob; somentor: V. Gorjanc. Ljubljana: Fakulteta za družbene vede Univerze v Ljubljani.

LOGAR, Tine, 1995: *Karta slovenskih narečij*. Ljubljana: Geodetski zavod Slovenije.

McCARTHY, Michael, 1998: *Spoken Language & Applied Linguistics*. Cambridge University Press.

McENRY, Tony in WILSON, Andrew, 1996: *Corpus Linguistics*. Edinburg University Press.

McENRY, Tony in WILSON, Andrew: *Early Corpus Linguistics and the Chomskyan Revolution*. Web-based Course on Corpus Linguistics. Part One. Department of Linguistics, Lancaster University.
http://bowland-files.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

MEYER, Charles F., 2002: *English Corpus Linguistics: An Introduction*. Oxford University Press.

OAKES, Michael P., 1998: *Statistics for Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh University Press.

PAYNE, Jonathan, 1995: The COBUILD Spoken Corpus: Transcription Conventions. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 203–207.

POGORELEC, Breda, 2004: Vase in v svet. Stabej, M. (ur.): *Zbornik predavanj 40. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 17–28.

POGORELEC, Breda, 1998: Jezikovno načrtovanje govorjenega jezika pri Slovencih. Štrukelj, I. (ur.): *Jezik za danes in jutri*. Ljubljana: Društvo za uporabno jezikoslovje. 56–64.

POGORELEC, Breda, 1989: Sociolingvistični problemi slovenske etnične skupnosti v Italiji. *Aspette metodologici e teoretici nello studio del plurilinguismo nei territori dell 'Alpe-Adria.* Videm.

POGORELEC, Breda, 1965: Vprašanja govorjenega jezika. *Jezikovni pogovori*. Ljubljana: Cankarjeva založba.

RAYSON, Paul, LEECH, Geoffrey in HODGES, Mary, 1997: Social Differentiation in the Use of English Vocabulary: Some Analyses of the Conversational Component of the British National Corpus. *International Journal of Corpus Linguistics* 2/1. 133–152.

ROZMAN, Tadeja, 2004: Upoštevanje ciljnih uporabnikov pri izdelavi enojezičnega slovarja za tujce. *Jezik in slovstvo* 49, 3/4. [63]–75.

SIGLEY, Robert, 1997: Text Categories and Where You Can Stick Them: A Crude Formality Index. *International Journal of Corpus Linguistics* 2/2. 199–238.

SINCLAIR, John McH. (ur.), 2004a: *How to Use Corpora in Language Teaching*. Studies in Corpus Linguistics 12. Amsterdam/Philadelphia: John Benjamin's Publishing Company.

SINCLAIR, John, 2004b: Corpus and Text – Basic Principles. Wynne, M. (ur.): *Developing Linguistic Corpora: A Guide to Good Practice.* http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

SINCLAIR, John, 1995: From Theory to Practice. Leech, G., Myers, G. in Thomas, J. (ur.): *Spoken English on Computer.* New York: Longman Publishing. 99–112.

SKUBIC, Andrej E., 2004: Sociolekti od izraza do pomena: kultiviranost, obrobje in eksces. Kržišnik, E. (ur.): *Obdobja 22.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 297–320.

SKUBIC, Andrej E., 1994/1995: Klasifikacija funkcijskih zvrsti in pragmatična definicija funkcije. *Jezik in slovstvo* 40, 3/4. 155–168.

SLOVAR slovenskega knjižnega jezika, 1998. Elektronska izdaja na plošči CD-ROM. Ljubljana, SAZU in ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, DZS in Amebis.

SLOVENSKI pravopis, 2003. Elektronska izdaja. Ljubljana, ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Amebis.

SMOLE, Vera, 2004: Nekaj resnic in zmot o narečjih v Sloveniji danes. Kržišnik, E. (ur.): *Obdobja 22.* Ljubljana: Filozozska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 321–330.

SMOLEJ, Mojca, 2006a: *Vpliv besedilne vrste na uresničitev skladenjskih struktur (primer narativnih besedil v vsakdanjem spontanem govoru)*. Doktorska disertacija. Mentor: M. Stabej. Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

SMOLEJ, Mojca, 2006b: Nekatere skladenjske značilnosti spontano tvorjenih besedil govorcev Ljubljane. Novak Popov, I. (ur.): *Zbornik predavanj 42. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik.

STABEJ, Marko, 2008: Kako pa govoriš!? (Spis o normi in normalnem). Vitez, P. (ur.): Spisi o govoru Ljubljana: Znanstvenoraziskovalni inštitut Filozofske fakultete. 87-100.

STABEJ, Marko, 2005: Kdo si, ki govoriš slovensko? Mikolič, V. in Marc, K. (ur.): *Slovenščina in njeni uporabniki v luči evropske integracije*. Koper: UP, ZRS, Založba Annales. 13–22.

STABEJ, Marko, 2003a: Jezikovne tehnologije in jezikovno načrtovanje. *Jezik in slovstvo* 48, 3/4. [5]–18.

STABEJ, Marko, 2003b: Med dvema stoloma: dihotomije v slovenistiki. Jesenšek, M. (ur.): *Perspektive slovenistike ob vključevanju v Evropsko zvezo*. *Zbornik Slavističnega društva Slovenije* 14. Ljubljana: Slavistično društvo Slovenije. [22]–31.

STABEJ, Marko, 2003c: Slovenščina od pet do glave. Krakar-Vogel, B. (ur.): *Zbornik predavanj 39. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko. 83–90.

STABEJ, Marko, 2003č: Ene in drugi: slovenščina in spola. *Delo*, leto 45, št. 31 (7. 2. 2003). 26.

STABEJ, Marko in VITEZ, Primož, 2000: KGB (korpus govorjenih besedil) v slovenščini. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. 79–81.

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

STABEJ, Marko, 1997: Seksizem kot jezikovnopolitični problem. Derganc, A. (ur.): *Zbornik predavanj 33. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti. 57–68.

STENSTRÖM, Anna-Brita, ANDERSEN, Gisle in HASUND, Ingrid Kristine, 2002: *Trends In Teenage Talk: Corpus Compilation, Analysis and Findings.* Studies in Corpus Linguistics 8. Amsterdam/Philadelphia: John Benjamins Publishing Company.

STRITAR, Mojca, 2006: Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije*. 134–139.

ŠABEC, Nada, 2002: Usoda slovenskega jezika med Slovenci po svetu. Krakar-Vogel, B. (ur.): *Zbornik predavanj 38. seminarja slovenskega jezika, literature in kulture*. Ljubljana: Filozozska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti.

ŠKOFIC-GUZEJ, Jožica, 1994: O oblikovanju slovenskega pogovarjalnega jezika. *Slavistična revija* 42, 4. [571]–578.

ŠPELKO, Tina in BAN, Janja, 2005: *Slovenska jezikovna skupnost v Argentini: (socio)lingvistična analiza.* Diplomsko delo. Mentor: M. Stabej. Ljubljana: Filozofska fakulteta.

TAO, Hongyn in WAUGH, Linda R., 1998: Constructing a New Corpus of Spoken American English. *Teaching and Language Corpora (TALC)*. Oxford University Press.

THOMPSON, Paul, 2004: Spoken language corpora. Wynne, M. (ur.): *Developing Linguistic Corpora: A Guide to Good Practice.* http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter5.htm

TIVADAR, Hotimir, 2008: *Kakovost in trajanje samoglasnikov v govorjenem knjižnem jeziku*. Doktorska disertacija. Mentorica: A. Vidovič Muha. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

TIVADAR, Hotimir, 2004: Podoba in funkcija govorjenega knjižnega jezika glede na neknjižne zvrsti. Kržišnik, E. (ur.): *Obdobja 22*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 437–452.

TOGNINI BONELLI, Elena, 2001: *Corpus Linguistics at Work*. Studies in Corpus Linguistics 6. Amsterdam/Philadelphia: John Benjamin's Publishing Company.

TOPORIŠIČ, Jože, 1976: *Slovenska slovnica*. Založba Obzorja Maribor.

TSUI, Amy B. M., 2004: What Teachers Have Always Wanted to Know – and How Corpora Can Help Them. Sinclair, J. (ur.): *How to Use Corpora in Lan-*

guage Teaching. Studies in Corpus Linguistics 12. Amsterdam/Philadelphia: John Benjamin's Publishing Company. 39–61.

VERDONIK, Darinka, 2007: *Jezikovni elementi spontanosti v pogovoru.* Maribor: Slavistično društvo (Zora 48).

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

VERDONIK, Darinka, 2005: Nesporazumi v komunikaciji. *Jezik in slovstvo* 50, 1. 51–64.

VEROVNIK, Tina, 2004: Govorjeni knjižni jezik v televizijskih dnevnoinformativnih oddajah: študija primera. Poler Kovačič, M. in Kalin Golob, M. (ur.): *Poti slovenskega novinarstva – danes in jutri.* Ljubljana: Fakulteta za družbene vede Univerze v Ljubljani 157–173.

VEROVNIK, Tina, 2005: *Jezikovni obronki*. Ljubljana: GV Založba.

VIDOVIČ MUHA, Ada (ur.), 2006: *Slovensko jezikoslovje danes*. Ljubljana: Slavistično društvo Slovenije.

VIDOVIČ MUHA, Ada (ur.), 2003: *Slovenski knjižni jezik – aktualna vprašanja in zgodovinske izkušnje. Obdobja 20*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Center za slovenščino kot drugi/tuji jezik, Oddelek za slovenistiko.

VIDOVIČ MUHA, Ada (ur.), 1996a: *Jezik in čas*. Razprave Filozofske fakultete. Ljubljana: Znanstveni inštitut Filozofske fakultete.

VIDOVIČ MUHA, Ada, 1996b: Razvojne prvine normativnosti slovenskega knjižnega jezika. Vidovič Muha (ur.): *Jezik in čas*. Ljubljana: Znanstveni inštitut Filozofske fakultete. 15–40.

VINTAR, Špela, 2008: *Terminologija. Terminološka veda in računalniško podprta terminografija.* Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo.

VINTAR, Špela, 2003: Kaj izvira iz jezikovnih virov. *Jezik in slovstvo* 48, 3/4. [77]–88.

VINTAR, Špela, 2003: *Uporaba vzporednih korpusov za računalniško podprto ustvarjanje dvojezičnih terminoloških virov*. Doktorska disertacija. Mentor: R. Šuštaršič. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za anglistiko in amerikanistiko.

VITEZ, Primož in ZWITTER VITEZ, Ana, 2004: Problem prozodične analize spontanega govora. *Jezik in slovstvo* 49, 6. [3]–24.

VITEZ, Primož, 1999: Od idealnih jezikovnih struktur k strategiji realnega govora. *Slavistična revija* 47, 1. [23]–48.

VITEZ, Primož, 1998: Zunajjezikovne okoliščine neidealnega govora. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 81–83.

VOGEL, Jerca, 2004: Nekateri vidiki zvrstnosti govorjenega diskurza s stališča poslušalca. Kržišnik, E. (ur.): *Obdobja 22.* Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik. 453–468.

WARWICK, Claire, 1997: *The Spoken Component of the BNC*. http://www.hcu.ox.ac.uk/BNC/what/spok_design.html

WEISS, Peter, 2001: Slovenski nacionalni korpus Maks na Inštitutu za slovenski jezika Frana Ramovša ZRC SAZU: utemeljitev. *Jezikoslovni zapiski* 7, 1–2. Ljubljana: Založba ZRC. 419–428.

WYNNE, Martin (ur.), 2005: *Developing Linguistic Corpora: A Guide to Good Practice*. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm

ZEMLJAK, Melita, KAČIČ, Zdravko, DOBRIŠEK, Simon, GROS, Jerneja in WEISS, Peter, 2002: Računalniški simbolni fonetični zapis slovenskega govora. *Slavistična revija* 50, 2. 159–169.

ZEMLJARIČ MIKLAVČIČ, Jana, 2008: Iskanje odgovorov na *Vprašanja govorjenega jezika*. *Jezik in slovstvo* 53, 2. [89]–106.

ZEMLJARIČ MIKLAVČIČ, Jana, 2007: *Načela gradnje govornega korpusa slovenščine.* Doktorska disertacija. Mentor: M. Stabej; somentor: V. Gorjanc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

ZEMLJARIČ MIKLAVČIČ, Jana, 2006: Korpus govorjene slovenščine. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije*. Ljubljana: Institut Jožef Stefan. 124–127.

ZEMLJARIČ MIKLAVČIČ, Jana in STABEJ, Marko, 2006: Zapisati nezapisljivo: transkribiranje spontanega govora za govorni korpus. *SloFon* 1. 1. slovenska mednarodna fonetična konferenca. Ljubljana, 20.–22. april 2006. 86–87.

ZEMLJARIČ MIKLAVČIČ, Jana, 2005: Jezikovne tehnologije – viri in pripomočki za učenje slovenščine kot tujega jezika. Mikolič, V. in Marc, K. (ur.): *Slovenščina in njeni uporabniki v luči evropske integracije*. Koper: UP, ZRS, Založba Annales. 253–259.

ZEMLJARIČ MIKLAVČIČ, Jana in STABEJ, Marko, 2005: Building a Pilot Spoken Corpus. *Computer Treatment of Slavic and East European Languages*. Proceedings of Third International Seminar SLOVKO. Bratislava, 10.–12. november 2005. 229–240.

ZEMLJARIČ MIKLAVČIČ, Jana, 2004: Taksonomija besedilnih tipov za gradnjo govornega korpusa. Kržišnik, E. (ur.): *Obdobja 22*. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko, Center za slovenščino kot drugi/tuji jezik.

ZEMLJARIČ MIKLAVČIČ, Jana, 2003: Jezikovne tehnologije 2002. *Jezik in slovstvo* 48, 3/4. 117–120.

ZÖGLING MARKUŠ, Aleksandra, KAČIČ, Zdravko in HORVAT, Bogomir, 2000: Razvoj slovenske baze izgovarjav POLIDAT. Erjavec, T. in Gros, J. (ur.): *Jezikovne tehnologije za slovenski jezik*. Ljubljana: Institut Jožef Stefan. 95–98.

ZULJAN KUMAR, Danila, 2005: *Govorjena briška narečna besedila z vidika besedilne skladnje.* Doktorska disertacija. Mentorica: V. Smole; somentorica: S. Kranjc. Ljubljana: Filozofska fakulteta Univerze v Ljubljani, Oddelek za slovenistiko.

ŽELEZNIKAR, Jaka, 1998: FIDA – pogoste napake pri vnosu in obdelavi besedil ter njihovo odpravljanje. *Uporabno jezikoslovje* 6. *Jezikovne tehnologije* (tematska številka, ur. Z. Kačič). 107–111.

ŽGANEC GROS, Jerneja, MIHELIČ, France in DOBRIŠEK, Simon, 2003: Govorne tehnologije: pridobivanje in pregled govornih zbirk za slovenski jezik. *Jezik in slovstvo* 48, 3/4. [47]–59.

Korpusi na internetu

I Govorni korpusi in njihova dokumentacija[216]

Survey of English Usage (Korpus London-Lund))
http://wwukw.ucl.ac./english-usage
http://helmer.aksis.uib.no/icame/london-lund/

Manual of Information to Accompany the SEC Corpus
http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM

The Bank of English Project
http://www2.lingsoft.fi/doc/engcg/Bank-of-English.html

The Bank of English (demo)
http://www.collins.co.uk/Corpus/CorpusSearch.aspx

The Bank of English User Guide
http://www.titania.bham.ac.uk/docs/svenguide.html#Introduction

The Swedish spoken language corpus at Göteborg University (opis)
http://www.ling.gu.se/projekt/tal/

COLT – The Bergen Corpus of London Teenage Language
http://helmer.aksis.uib.no/colt/

C-ORAL-ROM (Multilingvalni španski, italijanski, francoski in portugalski korpus)
http://lablita.dit.unifi.it/coralrom/index.html

International Corpus of English (ICE)
http://www.ucl.ac.uk/english-usage/ice/index.htm

Nizozemski govorni korpus (opis in demo)
http://lands.let.kun.nl/cgn/doc_English/topics/project/pro_info.htm

Britanski nacionalni korpus
http://www.natcorp.ox.ac.uk/

Češki nacionalni korpus
http://ucnk.ff.cuni.cz/

[216] Vse spletne strani, navedene v tej knjigi, so bile zadnjič preverjene 28. 6. 2009.

Madžarski nacionalni korpus
http://corpus.nytud.hu/mnsz/index_eng.html

ICAME Corpus Collection, International Computer Archive of Modern and Medieval English
http://nora.hd.uib.no/corpora.html

MICASE (*Michigan Corpus of Academic Spoken English)* demo verzija
http://www.hti.umich.edu/micase/

FIDA
http://www.fida.net/slo/index.html

FIDAPLUS
http://www.fidaplus.net/

NOVA BESEDA
http://bos.zrc-sazu.si

ELAN
http://nl.ijs.si/elan/

TRANS
http://nl2.ijs.si//index-bi.html

EVROKORPUS
http://www.sigov.si/evrokorpus

Korpus besedil odnosov z javnostmi (KoRP)
http://www.korp.fdv.uni-lj.si/

II Drugi naslovi

EAGLES: Preliminary recommendations on Spoken Texts (Expert Advisory Group on Language Engineering Standards)
http://www.ilc.cnr.it/EAGLES96/spokentx/spokentx.html

TEI
http://www.tei-c.org/

ICAME Corpus Manuals
http://khnt.hit.uib.no/icame/manuals/INDEX.HTM

ICAME (Bibliografija B. Altenberg)
http://khnt.hit.uib.no/icame/manuals/icambib3.htm

LCD (Linguistics Data Consortium)
http://www.ldc.upenn.edu/

Transcriber
http://www.etca.fr/CTA/gip/Projets/Transcriber/

Praat
http://www.fon.hum.uva.nl/praat/

WinPitch
http://www.winpitch.com/

Stvarno kazalo

Kazalo slik in tabel

KAZALO SLIK

KAZALO TABEL

Priloga: transkripcije (izbor)

Transkripcija posnetka R06 (12 min)

L	G17:	a to za faks ali kaj
L	G16:	ne ena kolegica me je prosila da ji posnamem
L	G17:	zdaj se snema {nv} smeh {/nv} (4 sec)
L		{smeh} ne moreš skriti {/smeh}
L	G16:	ne
L	G17:	ja in {shift=vpr} zakaj {/shift=vpr}
L	G16:	doktorat dela pa rabi posnetke ne {pavza} a veš {pavza}
L		pa je prosila če lahko {pavza}
L	G17:	{shift=vpr} kaj pa dela za en faks {/shift=vpr} {pavza}
L	G17:	saj ne dela več [faksa ja slovenščino ja]
L	G16:	[mislim kaj je naredila a slovenščino]
L	G17:	in {shift=vpr} ti {/shift=vpr}
L	G16:	ja nič
L	G17:	{shift=vpr} a si bil v sluľbi {/shift=vpr} {neraz}
L	G16:	ne samo tole sem šel iskat pa
L		nekaj sem šel kupit zdajle bova šla pa z {ime} še na kosilo
L		{??} {neraz} {/??}
L	G17:	{shift=vpr} kam gresta {/shift=vpr}
L	G16:	po mojem bova šla v Mirje
L	G17:	{shift=vpr} kje pa je to {/shift=vpr}
L	G16:	ə na Viču {pavza}
L	G17:	{shift=vpr} a je to gostilna ali kaj {/shift=vpr}
L	G16:	picerija pa gostilna ja
L	G17:	ja
L	G16:	əm {pavza} saj je kar dobro za jesti na bone
L	G17:	{neraz} študentski {tj} život {/tj} {pavza} (4)
L	G16:	{neraz}
L	G17:	{shift=vpr} se boš kaj postrigel {/shift=vpr}
L	G16:	ne bom kar pustil ne da se mi iti k frizerju pizda
L	G17:	pa sam se daj ne {shift=vpr} a [to {neraz} {/shift=vpr}]
L	G16:	[jah]
L	G16:	{smeh} hjah {/smeh} ne gre to a veš
L	G17:	{shift=vpr} ə {/shift=vpr} si se naveličal že
L	G16:	{shift=vpr} a [britja {/shift=vpr}] [ja]

L	G17:	[ja] mislim ə kratke [frizurce]
L	G16:	ne bom na kratko se ne bom ostrigel pa saj je boljše tako ə
L	G17:	{pavza} jah skoraj ja
L		čeprav saj ti tudi obrita glava paše
L	G16:	mah
L	G17:	za moje pojme [{neraz} kaj pa ljubice pravijo pa ne vem]
L	G16:	[meni {neraz} ja ljubice imajo raje če je malo daljše]
L	G17:	ja {nst} čupavce {/nst}
L	G16:	a veš {pavza} tako je ne {pavza}
L		{shift=vpr} kaj sta bila kaj v kinu {/shift=vpr}
L	G17:	ne nazadnje sva šla gledat
L	G17:	{shift=vpr} ə kaj sva že šla {/shift=vpr} tisto [takrat ə {?} Davidič {/?}]
L	G16:	[{neraz} ali Na robu ali]
L	G16:	{shift=vpr} in {/shift=vpr}
L	G17:	tako no
L	G16:	danes je Himalaja
L	G17:	problem je {pavza} a film
L	G16:	ja
L	G17:	aha
L	G16:	{shift=vpr} a si že gledal {/shift=vpr}
L	G17:	ne
L	G16:	po {okr} teveju {/okr} je na na {repet} Sloveniji ob osmih {pavza}
L	G17:	problem je to ker po dvajsetih minutah filma ugotoviš {neraz}
L	G17:	[za kaj] gre ne kakšen je konec vse a veš
L	G16:	[o čem je šlo]
L	G16:	ja ja {repet} ja {repet2}
L	G17:	potem je pa malo dolgčas gledati ne {pavza}
L	G17:	{shift=vpr} a boš šel gledat ti {/shift=vpr} [{neraz}]
L	G16:	[ne]
L	G16:	mislim saj ne po televiziji je
L	G17:	[aja] {neraz}
L	G16:	[ja] {neraz}
L	G16:	sem že gledal

L	G17:	{shift=vpr} in {/shift=vpr}
L	G16:	ma ni slabo mislim pokrajina je {nst} ful {/nst} lepa ane
L	G17:	čakaj {shift=vpr} to je prav film ali dokumentarec {/shift=vpr}
L	G16:	ə to je {shift=poud} film {/shift=poud}
L		{neraz} {pavza} (5)
L		film ja {pavza}
L	G17:	{shift=vpr} si slišal kaj je oni Michael {?} Moore {/?} {neraz} ali kaj imajo nekaj za študente v Ameriki {/shift=vpr}
L	G16:	ne
L	G17:	bog ve kaj ane verjetno je spet proti Bushu kaj
L	G16:	ja
L	G17:	in da so študentje začeli pištole vlačiti iz žepov in mu grozili da ga bodo
L	G16:	[{neraz}]
L	G17:	[{ns} fentali {/nst}] ja
L	G16:	{nv} phaha {/nv} saj ti Američani so {nst} prifukn-jeni {/nst} (zvok avtomobila)
L	G17:	ja (zvok avtomobila)
L	G16:	zdaj me pa prav zanima kaj bo na
L	G18:	[samo potem] so bile vodne pištole [so ugotovili] {nv} smeh {/nv}
L	G16:	[na] {pavza} [ah] {nv} smeh {/nv}
L	G17:	{shift=vpr} kaj te zanima {/shift=vpr}
L	G16:	kaj bo na volitvah
L	G17:	ah kaj Bush bo zmagal
L	G16:	spet malo pogoljufali a
L	G17:	{pavza} vprašanje tudi če ne ne
L	G16:	ja
L	G17:	tudi če ne {repeat/}
L	G16:	[po mojem pa bo]
L	G17:	[samo ima kar] veš zdaj
L	G17:	ta enajsti november ima a veš
L	G16:	{neraz} in jih s tem {nst} fila {/nst} [skozi]
L	G17:	{neraz} {pavza} [pa] Irak
L	G16:	{nv} pihne skozi nos {/nv} saj vem saj to bo ne pa malo bojo pogoljufali spet ne
L	G17:	to ne vem če si bodo spet lahko privoščili a veš ker bodo oni spet

L	G16:	[ja saj zdaj] so zdaj so {repet} naročili zunanje opazovalce a veš
L	G17:	[malo bolj]
L	G17:	{shift=vpr} kaj {/shift=vpr}
L	G16:	ja kongres je naročil in so čisto ogorčeni {nst} pizda {/nst}
L	G17:	to pa ne vem
L		ja čisto so ogorčeni kako njim ki imajo tako stopnjo visoko demokracije
L		ja
L	G16:	[naročijo]
L	G17:	[ja]
L	G16:	zunanje [opazovalce] {pavza} [mm pizda] včeraj mi je bilo všeč ej
L	G17:	[ah saj to je] čisto tipično za njih ki se skozi na demokracijo zgovoarja pa [poglej kakšna demokracija]
L	G16:	sem gledal zdaj ko ta əm Vega ne
L	G17:	{nv} si očisti grlo {/nv}
L	G16:	toži Mobitel ne
L	G17:	ja
L	G16:	a veš in potem so še nekaj zdaj imajo še {nst} mutke {/nst} dol na Kosovu ne ta
L	G17:	[ja {neraz} Mobi je izgubil zdaj]
L	G16:	[{neraz} zgubil] je to kar so že podpisali ne
L	G16:	in potem predsednik ta ə Vege [ne]
L	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 {nst} pizdi {/nst} ne kako lahko zdaj Mobitel əm se zoperstavi Združenim narodom ne
L	G17:	{shift=vpr} a [res {/shift=vpr}]
L	G16:	[Američan] a veš {nst} pizda {/nst}
L	G16:	ne [oni ki imajo] {pavza} [ne {neraz}] potem pa oni ves zgrožen {shift=vpr} kako lahko {/shift=vpr} ne
L	G17:	[ja ja zastopim] ki se itak ne {nst} šmirglajo {/nst} [Združenih narodov]
L	G17:	[{nv} smeh {/nv}]
L	G16:	[pizda] ej {nv} pihne skozi usta {/nv}
L	G16:	to samo gledaš {pavza}
L	G17:	{neraz} (ropot avtomobila)(2 sec)

L	G17:	{neraz} svojega denarja vredni ti Američani (ropot avtomobila)
L	G17:	{ime} mi je rekla da so na- naredili raziskavo na {okr} fədəvəju {/okr} ali kje že saj ne vem
L	G16:	ja
L	G17:	ə če bi Slovenci volili {pavza} Kerryja ali Busha
L	G17:	jih je petindevetdeset procentov Slovencev je za Kerryja
L	G17:	pet procentov pa [za Busha] {pavza} [{neraz}] [ja]
L	G16:	[ja pa] saj večina saj [vse] svetovne [države]
L	G17:	večina [mislim evropske {neraz}]
L	G16:	[cel svet je] cel} repet} s- cel svet {repet} je za Kerryja razen Amerike
L	G16:	ma ja saj bomo videli saj ne bo dosti boljše
L		a veš isti {nst} kurac {/nst} bo po moje
L	G16:	zdaj bo pri nas vse drugače ko bo [Janša]
L	G17:	[{nv} smeh {/nv}]
L	G17:	{nst} glih {/nst} zdajle so šli tukajle mimo
L	G16:	ja
L	G17:	pa oni {nst} kurc {/nst} {shift=vpr} kaj je že {/shift=vpr}
L		{pavza} Rupel
L	G16:	ah ta je meni meni {repet} je Rupel v redu {nst} pizda {/nst}
L	G17:	ma daj [ga {neraz}]
L	G16:	[ja] ej Rupel je car
L	G16:	Ru- Rupel je ne vem on je tak svetovljan on je brihten človek
L	G17:	{nst} kurca {/nst} je [brihten {neraz}]
L	G16:	[je je {repeat/} je je {repeat/}] Rupel je hud veš
L	G17:	takrat ko so bili Pankrti je zapisal v eno revijo
L		kako že zdaj citiral ne bom ane samo da ə
L		{pavza} piše da ə ə {shift=poud} ta je pa vredna da jo preslišite {/shift=poud}
L	G16:	{shift=vpr} a res {/shift=vpr}
L	G17:	[to je] takrat že napisal a veš
L	G16:	to se je malo zajebaval
L	G17:	ah kaj jaz vem
L	G16:	ne saj je možno da ima kakšne {nst} fore {/nst} samo meni Rupel ni tako napačen
L		{pavza} ne vem {pavza}

L	G17:	kaj jaz vem
L	G16:	ja
L	G17:	samo ne smeva preveč o politiki [{neraz}] nikoli ne ve
L	G16:	[ja {nv} smeh {/nv]
L	G17:	[{nv} smeh {/nv]
L	G16:	[{nv} smeh {/nv]
L	G17:	kakšen prišel [pa]
L	G16:	[ja]
L	G16:	{nv} smeh {/nv} to je tako {neraz}
L	G17:	ne vem no on je meni {pavza} kretenček
L	G16:	{shift=vpr} kaj bosta za novo leto ej {/shift=vpr}
L	G17:	verjetno bova kar doma ker bom delal
L	G16:	ja {shift=vpr} aja delaš {/shift=vpr}
L	G17:	{pavza} itak pa ni ne
L	G17:	petek je no- je [enaintrideseti]
L	G16:	[saj ne vem kako je] sploh
L	G17:	[petek] je enaintrideseti se mi zdi no
L	G16:	[ja]
L	G16:	{shift=vpr} a v soboto pa delaš potem ne {/ shift=vpr}
L	G17:	soboto upam da bo dal tale
L	G17:	[direktorček] ja
L	G16:	[{nst} frej {/nst}]
L	G17:	{pavza} če ne bo spet kaj zakompliciral {shift=vpr} kaj pa ti {/shift=vpr}
L	G16:	v Berlin greva
L	G17:	{shift=vpr} a {/shift=vpr} [aja saj to sta že rekla ja]
L	G16:	[sva kupila] karte
L	G17:	ə
L	G16:	kar fajn dala sva za vsako sva dala petnajst {nst} jurjev {/nst} (telefon zazvoni) a veš
L	G17:	ja {pavza} ja {shift=vpr} kaj je {/shift=vpr} {pav- za} (ropot)
L		{shift=vpr} ə {/shift=vpr} ja {pavza}
L	G18:	tole bo {ime} poklical če mu bi samo ven [dali] {pavza} [ja]
L	G17:	[zdaj] me kliče [k- ja]
L	G17:	{shift=vpr} kaj {/shift=vpr} ja zdaj je prinesla

L	G18:	dm- samo ven [mu {neraz}]
L	G17:	[zdaj je] prinesla ja ti bom prinesel
L	G17:	{pavza} (3) (ropot) ja
L		okej ajd {pavza} (3)
L	G16:	{shift=vpr} kaj pa imaš to {/shift=vpr}
L	G17:	a kurirček me kliče da
L	G17:	če mi je že prinesla pa {nst} evo {/nst} ti {smeh} pa {nst} glih {/nst} [kar] {/smeh}
L	G16:	[aja]
L	G16:	{pavza} bo prišel [iskat]
L	G17:	[ja]
L	G16:	{pavza} kul pa poceni so karte v bistvu a veš
L	G17:	ə {shift=vpr} za {/shift=vpr}
L	G17:	[aha letalske] za Berlin
L	G16:	[za Berlin]
L	G16:	je koliko je petnajst {nst} čukov {/nst} vsak ane
L	G17:	[povratna]
L	G16:	[ja] s taksami vred pa vse
L	G17:	katera linija je to
L	G16:	Easy Jet
L	G17:	aha
L	G16:	z Brnika letiš {pavza}
L	G17:	to je zastonj [skoraj]
L	G16:	[ja]
L	G16:	[{nst} pizda {/nst} to] te dražje pride če greš z [avtom če gresta] {repet} če gresta {/} dva z avtom te pride dražje
L	G17:	[{nv} pihne skozi nos {/nv}] [jah {neraz}]
L	G17:	seveda te pride [dražje]
L	G16:	[{nst} benz {/nst}] pa cestnina če grejo štirje potem ne ne samo
L	G17:	aja saj imaš vse preračunano {shift=vpr} ali kaj {/ shift=vpr}
L	G16:	mislim pač avto [te pride tam enih petintrideset] {nst} jurjev {/nst} [nekaj] takega ne v obe smeri a veš
L	G17:	[no sigurno te pride] [ja]
L	G16:	potem se pa še tam kaj voziš pa ni ne ni sploh
L	G16:	pa v eni uri si gor pizda [to se voziš {repet} to se voziš {/} pa] deset ur a veš

L	G17:	[{neraz} ne vem kako jim to uspeva ej]
L	G16:	meni tudi ni jasno
L		so pa recimo danes so pa iste karte so pa že əm
L		šestindvajset {nst} jurjev {/nst}
L	G17:	{shift=vpr} aja {/shift=vpr}
L	G16:	ja
L	G17:	nekaj sem poslušal da če naprej rezerviraš da je toliko ceneje [ne]
L	G16:	[saj je]
L	G17:	samo [{neraz}]
L	G16:	[{neraz}]
L	G17:	{pavza} a veš
L	G16:	{neraz} ni jasno pa vsak dan leti
L	G17:	m
L	G16:	vsak dan Ljubljana Berlin (hrup avtomobila 3 sec)
L		to meni ni jasno {pavza} (4)
L		veš {pavza} (2)
L	G17:	{shift=vpr} s {priimek} se kaj slišiš {/shift=vpr}
L	G16:	ja saj se kar slišiva ja {pavza} (2)
L		nič zdaj diplomo piše
L	G17:	ja
L	G16:	en izpit ima še ali kaj saj ne vem
L	G17:	[meni se zdi ja da ima ja] en izpit pa diploma ja
L	G16:	[{shift=vpr} a ima {/shift=vpr} on ima še en izpit]
L	G16:	{pavza} (2) izpit pa diploma
L	G17:	{shift=vpr} saj ti pa tudi nimaš več dosti ne {/shift=vpr}
L	G16:	jaz imam še koliko štiri izpite pa tako da se reče ane potem pa še tam neke formalnosti
L		{pavza} zdajle dvanajstega imam enega {nst} pizda {/nst} to poezijo študiram {nst} jebem {/nst} jim mater ej
L	G17:	{nv} smeh {/nv}
L	G16:	res no
L	G17:	{nv} vzdih, smeh {/nv} {nst} pička {/nst} ima šalamuna
L	G16:	ma ja da mu {nst} jebala {/nst}
L	G17:	jaz sem njega enkrat bral samo sem ga tako a veš na hitrico

L		no in zapomnil sem si samo da ima teh kletvic kolikor hočeš ne
L	G16:	ja
L	G17:	od {nst} kurcov {/nst} do {nst} pizd {/nst} pa ntako naprej {pavza}
L		pa tudi drugače kaj pa vem no {pavza}
L	G17:	{shift=vpr} ja [a veš] tisto kar sem ti včeraj bral ne {/shift=vpr}
L	G16:	[ne vem]
L	G17:	ja
L	G16:	{pavza} (2) mislim
L	G17:	sem to {nst} glih {/nst} danes zjutraj [razmišljal] mogoče je pa fora
L	G16:	[ja]
L	G17:	{nst} glih {/nst} v tem ne da se potem človek vpra- {shift=vpr} da je pisal to o
L		o vakuumu o člo- v človeku ne {/shift=vpr}
L	G16:	[ja saj je možno a veš ne {neraz} če mene vprašaš]
L	G17:	[ej ne vem o čem {neraz}]
L	G16:	[interpretacij je neskončno] a veš
L	G17:	[izgubljenost in tako naprej]
L	G17:	ja
L	G16:	vsak si [lahko]
L	G17:	[mislim] ne glede na {nst} glih {/nst} glede na interpretacijo ne če je interpretacija piše da piše o ničemer ne
L	G17:	to se [pravi]
L	G16:	[{neraz}]
L	G17:	da bi lahko to [kaj] povezal
L	G16:	[ne] {shift=poud} piše {/shift=poud} o nečem samo da mi ne vemo o čem
L	G17:	[aja]
L	G16:	[o nečem] {shift=poud} piše {/shift=poud} ne
L	G16:	to {neraz} [pesem] nekaj [pripoveduje pa] ne vemo kaj neja
L	G17:	[no no saj] [to to]
L	G16:	a veš in potem ti pušča [{nst} pizda {/nst} {neraz} ja vsaka beseda nekaj pove ne {neraz}]
L	G17:	[ja no saj vsaka pisana beseda nekaj nekaj {repet} pov- hoče povedat ja]
L	G17:	{pavza} (3) ja ja saj štekam

L	G16:	malo je težko pa veliko je no to je problem
L	G17:	{shift=vpr} a o umetnosti ss- ste se kaj pogova- {/shift=vpr} {shift=vpr} a ti znaš opredeliti umetnost {/shift=vpr}
L	G17:	kaj je [umetnost]
L	G16:	[ja imeli] lansko leto
L	G16:	za [diplomski] seminar
L	G17:	[ja]
L	G17:	{shift=vpr} in {/shift=vpr}
L	G16:	filozofija umetnosti ne znam ti po- mislim
L	G16:	ne da se [opredeliti] bolj težko so teorije samo
L	G17:	[ne da se]
L	G17:	ja
L	G16:	je težko je to opredeliti {neraz} {pavza} (2)
L		to {pavza} mislim (smeh v ozadju)
L	G17:	[{neraz} v slovenščini] v slovenščini recimo je umetnost
L	G16:	[je opredelitev]
L	G17:	delite na najmanj dva dela sigurno {shift=vpr} ane {/shift=vpr} {pavza}
L		imate {shift=vpr} kako se reče ə {/shift=vpr} pač {pavza}
L	G17:	ne vem jaz zdaj teh izrazov [ane] {shift=vpr} kako je ne vem Shakespeare {/shift=vpr}
L	G16:	[ja]
L	G17:	[{shift=vpr} Cankar pa tako naprej {/shift=vpr}] [{shift=vpr} ne {/shift=vpr}]
L	G16:	[ne ne {repet} ne {repet} gre] sploh ne gre {repet} za to {neraz}
L	G16:	fora je v tem kaj umetnost sploh je zakaj je recimo pisoar v muzeju [umetnosti]
L	G17:	[no že ampak]
L	G17:	[kolikor jaz vem se deli] v slovenščini umetnost na dva dela {shift=vpr} a se ne {/shift=vpr}
L	G16:	[stranišče pa ne]
L	G16:	samo to ni interpretacija umetnosti
L	G17:	ja
L	G16:	a veš ti [mislim dobro jaz ne vem] kako misliš na kakšna dva dela {pavza}
L	G17:	[definicija ja]

L	G17:	meni se zdi no da sem nekje prebral ali da sem slišal da
L	G17:	da {repet} je saj zdaj teh izrazov jaz ne poznam [ane] ker pač nisem
L	G16:	[ja]
L	G17:	ne vem kakšna [delitev] je to {pavza}
L	G16:	[ə]
L	G17:	əm {pavza} (2)
L		resna ali kako je že samo da je s tujko ane
L		pa potem tista ostala ane
L	G17:	zdaj pod resno misliš {neraz} literaturo ne
L	G16:	[{neraz}]
L	G17:	se štejejo pač ti Cankarji [Bevki {neraz}]
L	G16:	[glej ti imaš kanon imaš {repet} recimo] kar se literature tiče ampak ne vem
L	G16:	ne vem {repet} potem imaš ə
L	G16:	[trivial- trivialno literaturo]
L	G17:	[saj boš zdaj {neraz} pa verjetno čiste neumnosti govorim ane samo]
L	G16:	(hrup) samo to ni to zdaj {shift=vpr} kako umetnost definirati {/shift=vpr} (telefon začne zvoniti) [umetnost je] teľko definirati
L	G17:	[{neraz}]
L	G17:	ne moreš je (telefon zvoni)
L	G16:	a veš (telefon zvoni)(3)
L		ja prosim {pavza} (3)
L		zdravo živjo {pavza} (2) živjo živjo {pavza} (2)
L		aha zdravo
L	G17:	dober dan (posnetek prekinjen)

Univerza v Ljubljani
FILOZOFSKA FAKULTETA