Jana Zemljarič Miklavčič
Ljubljana

UDK 811.163.6'271.16:81'42

TAKSONOMIJA BESEDILNIH TIPOV ZA GRADNJO GOVORNEGA KORPUSA

Pri gradnji referenčnih korpusov si prizadevamo za čim večjo reprezentativnost in uravnoteženost, to je za nabor čim večjega števila različnih vrst besedil v razmerju, ki čim bolj ustreza jezikovni realnosti. Posebej pri gradnji govornega korpusa predstavlja zajem besedil eno najtežjih jezikovno-teoretičnih vprašanj, saj zaenkrat ne moremo objektivno izmeriti celotne govorne produkcije (ali recepcije) in določiti kvantitativnih razmerij med posameznimi tipi govorjenih besedil. Taksonomijo besedil lahko gradimo na različnih izhodiščih: glede na njihove strukturne značilnosti, vsebino, namen in/ali okoliščine, v katerih je besedilo nastalo – odvisno predvsem od namembnosti korpusa, zaradi zahtevnosti gradnje pa tudi od drugih okoliščin, npr. od razpoložljivih finančnih sredstev, velikosti ekipe sodelavcev in tehničnih možnosti.

govorjeni jezik, govorjeno besedilo, govorni korpus, zajem besedil v korpus, reprezentativnost, uravnoteženost

In building reference corpora, we strive to be as representative and balanced as possible, by selecting as many different textual genres in proportions that represent language reality. Particularly when constructing a spoken corpus, the selection of texts is one of the thorniest theoretical issues, as we cannot at any one time objectively survey the whole of spoken production (or reception) and determine the quantitative relations between different kinds of spoken text. Taxonomies of texts can be approached in different ways – with regard to their structural characteristics, content, purpose, and/or the situation in which the text arose – depending primarily on the importance of the corpus, on how demanding is its construction and on other circumstances, such as financial resources, the size of the team involved and technical capabilities.

spoken language, spoken text, corpus of spoken language, selection of corpus texts, representativeness, balance

Govorni korpusi

Govorjeni slovenščini je bilo doslej, razen v dialektologiji, posvečene relativno malo jezikoslovne pozornosti. Raziskav, ki bi temeljile na analizi in interpretaciji obsežnejšega avtentičnega gradiva – spontanega javnega in zasebnega govora – za slovenščino praktično ni bilo.[1] Slovenska teorija jezikovnih zvrsti govorjenih besedil ni posebej tipologizirala: razlikuje jih glede na okoliščine sporazumevanja in govorni položaj (socialne zvrsti), glede na namen sporočanja (funkcijske zvrsti) in

glede na prenosnik – pisni ali govorjeni jezik (Toporišič 1984: 10);[2] znotraj teh kategorij pa tipi govorjenih besedil niso bili posebej opredeljeni. Ker je govorno sporazumevanje v marsičem primarnejše od pisnega (prim. Stabej, Vitez 2000: 79), pa tudi zaradi tehnološkega razvoja, ki omogoča nove, v več pogledih revolucionarno naprednejše raziskave, se v zadnjih letih raziskovanje govorjenega jezika premika v središče ne samo jezikoslovnih, ampak tudi drugih humanističnih in družboslovnih raziskav.

Pri tem je veliko vlogo odigralo korpusno jezikoslovje, ki je poleg obsežnih zbirk pisnih besedil začelo posebno pozornost namenjati tudi govorjenim besedilom. Nastajati sta začeli dve vrsti besedilnih zbirk, *govorni korpusi* in *korpusi govora*. *Govorni korpusi* so urejene računalniške zbirke posnetkov in transkripcij spontanega govora. Posnetke govora (za najbolj dragocene veljajo posnetki, kjer vsi ali vsaj nekateri govorci ne vedo, da se njihov govor snema) se shranjuje v različnih oblikah, v zadnjem času skoraj izključno v digitalni obliki. Besedila so lahko transkribirana,[3] nato pa označena in shranjena v obliko, primerno za nadaljnje raziskave. V najsodobnejših govornih korpusih so transkribirane enote shranjene skupaj z zvočnimi posnetki, kar pomeni, da z enostavnim računalniškim ukazom hkrati z zapisano enoto besedila (členjenje enot pri govoru je poseben problem) dobimo tudi originalno zvočno podobo enote (skupaj z določeno količino levega in desnega sobesedila), kar dodatno povečuje možnosti raziskovanja govorjenega jezika. Govorni korpusi torej niso namenjeni raziskavi govora;[4] za potrebe fonetično-fonoloških raziskav in govornih tehnologij se oblikujejo posebni korpusi, t. i. *korpusi govora* (Gorjanc 2002: 7), to so najpogosteje vnaprej pripravljene in v studiu prebrane in posnete besedilne ali še pogosteje stavčne enote, torej neke vrste laboratorijski govor.

Tako zgrajen korpus je neprecenljiv vir podatkov o jeziku, namenjen analizam in interpretacijam, ki brez korpusa ne morejo nastati. Govorni korpusi služijo za preverjanje veljavnosti teoretičnih spoznanj o jeziku in za nove opise jezika – predvsem v slovaropisju in slovnici; so izrednega pomena kot referenčni vir pri učenju jezika kot tujega jezika (za učitelje in izdelovalce gradiv, pri bolj razširjeni rabi tudi za govorce), pa tudi kot vir za analize diskurza, pragmatične študije in kontrastivne analize govorjenega in pisnega jezika.

[1] Izjema so nekatere razprave v pragmalingvistiki, npr. O. Kunst Gnamuš, Govorno dejanje – družbeno dejanje, *Dialogi* 25/5–6 (1989), 83–89; M. Schlamberger Brezar, Vloga povezovalcev v diskurzu, *Jezik za danes in jutri*, ur. I. Štrukelj, Ljubljana: Društvo za uporabno jezikoslovje Slovenije: Inštitut za narodnostna vprašanja, 1998; O. Kunst Gnamuš, M. Nidorfer Šiškovič, M. Schlamberger Brezar, *Posrednost in argumentacija v govoru F(p) – T(r): zgradba stavka med informacijo, argumentacijo in konverzacijo*, Ljubljana: Pedagoški inštitut, Center za diskurzivne študije, 1997.

[2] Poleg naštetih še časovne in mernostne zvrsti (Toporišič 1984: 10).

[3] Obstajajo različne standardizirane stopnje transkripcije, ki jih izbiramo glede na namembnost korpusa: ortografska transkripcija ali transliteracija, fonemska, alofonska, akustično-fonetična in prozodična transkripcija, prim. EAGLES, 1995).

[4] Lahko pa pri teh raziskavah služijo kot izredno pomembno ali celo nujno dopolnilo.

Tako je tudi oblikovanje korpusa govorjenih besedil v slovenščini nujno zaradi več razlogov, predvsem pa zato, ker bi nam omogočil »empirični pogled na raziskovanje jezika, […] saj je analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna« (Stabej, Vitez 2000: 79). Poseben problem pri tem je velikost govornih korpusov, ki je zaenkrat še daleč od velikosti korpusov pisnih besedil, kar je glede na težavnost zbiranja gradiva razumljivo.[5]

Sicer pa fonetične in fonološke raziskave nastajajo predvsem na podlagi t. i. korpusov govora, to je laboratorijskih posnetkov govora, ki so zaradi velike redukcije šumov iz okolice za tovrstne raziskave pogosteje primernejši; znotraj jezikoslovnih tehnologij služijo korpusi govora predvsem kot osnova za razvijanje sistemov za avtomatsko razpoznavanje in sintezo govora.[6]

Z določitvijo kriterijev, s katerimi ovrednotimo korpuse, in z definiranjem vrst korpusov se je ukvarjala skupina za tipologijo korpusov pri evropski iniciativi EAGLES (Expert Advisory Group for Language Engineering, http://www.ilc.cnr.it/EAGLES96/edintro/edintro.html); določili so naslednje karakteristike korpusa:

– *velikost*, to je količina podatkov, ki jo korpus vsebuje,
– *kakovost* njegove izdelave,
– *avtentičnost* glede na kriterije, po katerih je zgrajen,
– *enostavnost zapisa* in
– *dokumentiranost*.

Pojem kakovosti korpusa v veliki meri sovpada s pojmom reprezentativnosti.[7] Če je namreč korpus vzorec, ki naj bi služil za predstavljanje jezika kot celote, mora čim bolje odslikavati realno podobo jezika. V kolikšni meri bo vzorec reprezentativen, je odvisno »najprej od obsega, v katerem predstavlja zbirko vseh besedil v realnosti; delež reprezentativnosti je torej odvisen od vnaprejšnjega definiranja celotne populacije, ki jo namerava vzorec predstavljati, in od tehnik vzorčenja«[8] (Biber 1993: 243). Pri zajemu in razvrščanju govorjenih besedil v korpus se moramo zato čim bolj približati razmerjem med tipi besedil in govorci, ki ustrezajo realni rabi jezika; zajeti se trudimo vse znane oblike besedil glede na vse znane in določljive kriterije. Po eni strani gre za razvrščanje, ki temelji na lastnostih govorca, saj pričakujemo, da se govor različnih govorcev glede na spol, starost, regionalno pripadnost, izobrazbo in družbeno-ekonomski status razlikuje. Po drugi strani pa se govorjena besedila razlikujejo tudi glede na namen in okoliščine, v katerih so nastala. Če primerjamo taksonomije besedil že zgrajenih govornih korpusov, lahko

[5] Govorna komponenta BNC, eden največjih obstoječih govornih korpusov, vsebuje pribl. 10 milijonov besed, kar predstavlje 10 % celotnega korpusa.

[6] V slovenščini se zdita izraza *govorni korpus* in *korpus govora* premalo razlikovalna; morda se bo njuna terminološka vrednost s širitvijo in ustalitvijo rabe utrdila.

[7] V določenih okoliščinah in za določene namene je lahko uporaben tudi korpus, ki ni reprezentativen glede na celotno jezikovno realnost.

[8] Prevedla J. Zemljarič Miklavčič.

ugotovimo, da se vsaka skupina raziskovalcev odloča za svojo taksonomijo tudi v primerih, ko gradijo korpus istega jezika – pač glede na namen korpusa, individualne lastnosti jezika in raziskovalni potencial (tehnične in finančne možnosti ter število usposobljenih raziskovalcev).

Tipologije govorjenih besedil v obstoječih in načrtovanih korpusih

Najstarejši govorni korpus in prva računalniška zbirka govorjenih besedil je korpus *London-Lund* (LLC). Nastal je na podlagi korpusa *Survey of English Usage* (SEU), »prikaza angleščine v rabi«, ki sodi med najstarejše korpuse sploh, še v obdobje predračunalniških besedilnih zbirk (1959). Korpus SEU sestavlja 200 približno enako dolgih besedil, od tega 100 pisnih in 100 govorjenih, skupaj 1 milijon besed. Korpus je bil namenjen za preučevanje govorjene in pisne britanske angleščine odraslih govorcev in naj bi služil kot vir za slovnični opis jezika. Tipologijo govorjenih besedil korpusa SEU predstavlja spodnja shema:

Govorjena besedila SEU:
- **dialog**
  - zasebna konverzacija
    - na štiri oči
      - posneto brez vednosti govorca
      - posneto z vednostjo govorca
    - po telefonu
  - javna diskusija
- **monolog**
  - pripravljeni
  - spontani

Slika 1: Shematski prikaz tipologije besedil v govorni komponenti korpusa SEU (http://helmer.aksis.uib.no/icame/london-lund/)

Javna diskusija je dialog, ki se odvija pred poslušalci, ti pa 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 ali komentar različnih drugih dogodkov (npr. državnih proslav), pa tudi nepripravljeni govori v parlamentarnih razpravah in podobno. *Pripravljeni monolog* je bližje pisnemu jeziku, vendar še vedno vključuje določeno mero spontanosti; to so npr. predavanja, pridige, sodni govori ipd. (Greenbaum, prav tam). Iz sheme je razvidno, da gre za razmeroma preprosto (v primerjavi s kasnejšimi), pa vendarle nazorno tipologijo besedil, vključenih v korpus; razmerje med dialogom in monologom v korpusu SEU je 76 % : 24 %. V knjižno izdajo korpusa (tudi na http://helmer.aksis.uib.no/icame/london-lund/londlundlst.html) sta

vključena dodatka, ki vsebujeta podrobnejše podatke o besedilih (datum nastanka oz. posnetka, število besed v posamezni enoti, razmerje med govorcema(-i), (ne)tajnost snemanja, tematika in pragmatična funkcija besedila) in podatke o govorcih (spol, starost in poklic/izobrazba).

Leta 1975 je Jan Svartvik na univerzi v Lundu na Švedskem začel sestrski projekt korpusa SEU; njegov namen je bil govorno komponento korpusa SEU prenesti v digitalno obliko, primerno za računalniško branje. V začetku osemdesetih let je računalniška verzija govorne komponente korpusa SEU – korpus *London-Lund* – zakrožila med zainteresiranimi znanstveniki po celem svetu. Gradivo korpusa *London-Lund* sedaj obstaja v treh verzijah: na listkovnem gradivu korpusa SEU, na CD-romu in v knjigi.[9] Na podlagi korpusa SEU je nastalo več kot 200 strokovnih in znanstvenih razprav – monografij, poglavij in člankov, najpomembnejša med njimi pa je referenčna slovnica modernega angleškega jezika.[10]

V osemdesetih letih so jezikoslovci v Veliki Britaniji zasnovali še en projekt velikih razsežnosti in v letih 1990–1994 zgradili *Britanski nacionalni korpus* (BNC). Pobuda za gradnjo jezikovnih virov za angleščino je prišla s strani britanske vlade, pri tem pa naj bi se glede na deloma prekrivne interese spodbudilo sodelovanje akademskih in industrijskih/kapitalskih sfer. K projektu so pristopile univerzitetne institucije in komercialni partnerji,[11] vlada pa je z različnih oddelkov za projekt v treh letih namenila 1,5 milijona britanskih funtov, kar naj bi zadoščalo za celotno pokritje znanstveno-raziskovalnega dela in za 50-odstotno kritje dela komercialnih partnerjev (Burnard 2000: 2).

Zahtevna gradnja korpusa BNC, t. i. *tekoči trak* (*BNC sausage machine*, Burnard 2000: 3), je tekla tako, da je bilo delo po odsekih razdeljeno med različne ustanove 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, končno generiranje oznak pa spet na univerzi v Oxfordu. Pri gradnji BNC 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 že ob svojem nastanku (prvič dostopen javnosti leta 1995) ni bil več največji britanski korpus, saj ga je v tem smislu prehitel korpus *Bank of English* (BOE), ki je prav tako vseboval tudi govorno komponento, vendar je imel tudi BNC pred BOE dve veliki prednosti: bil je skrbneje uravnotežen (razmerja med tipi besedil in med govorci) in dostopen širokemu krogu uporabnikov (BOE javnosti ni

[9] Jan Svartvik, Randolph Quirk (ur.), *A Corpus of English Conversation*, Lund: Gleerups/Liber, 1980, 893 strani.

[10] Randolph Quirk, Sidney Greenbaum, Geofrey Leech, Jan Svartvijk, *A comprehensive grammar of the English language,* London, New York: Longman, 1985.

[11] Oxford University Press (vodilni partner), Longman Group UK Ltd., Chambers Harrap, Britanska knjižnica, Univerza v Oxfordu in Univerza v Lancastru.

dostopen); v nekaterih komponentah ostaja korpus še danes nepresežen (npr. glede kakovosti jezikovnih podatkov v govorni komponenti, prim. Gorjanc 2002: 48), ne nazadnje pa je še vedno drugi največji govorni korpus.

Razmerje med govorjenimi in pisnimi besedili v BNC (10 % : 90 %) je bilo določeno na podlagi ekonomske logike, saj so izračunali, da stane 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 BNC izkazujeta zavidljivo uravnoteženost tipov besedil, vendar pa ravno zaradi tega tudi statično sliko jezika iz določenega obdobja (govorjena besedila so bila posneta v letih 1991–1994). Dandanes gredo težnje in zahteve v jezikoslovju v smeri večje dinamičnosti korpusa (priključevanja novih, časovno aktualiziranih besedil),[12] vendar pa tudi znotraj dinamičnih korpusov obstajajo uravnoteženi podkorpusi tipa BNC za posebne raziskave.

Pri gradnji govorne komponente korpusa BNC so jezikoslovci prvič uporabili metodo demografskega vzorčenja. S pomočjo statistične metode so določili reprezentativni vzorec govorcev britanske angleščine glede na spol, starost, regijsko pripadnost in socialni razred. Nadaljevanje projekta je prevzelo podjetje za raziskavo tržišča: izbrali so 124 prostovoljcev (starejših od 15 let), ki so predstavljali reprezentativni vzorec govorcev britanske angleščine;[13] med njimi je bilo približno enako število moških in žensk, ustrezno razporejenih v tri regijske in štiri starostne skupine ter v štiri socialne razrede. Tako je nastal t. i. *demografski del* govornega korpusa BNC (imenovan tudi *konverzacijski podkorpus*), obsega 4.206.058 besed, to je pribl. 40 % govorne komponente korpusa BNC.

Govorci, ki so sestavljali izbrano reprezentativno skupino, 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. Statistične obdelave celotnega konverzacijskega korpusa so pokazale, da so se razmerja uravnoteženosti, ki so bila vzpostavljena glede na izbrane govorce, precej porušila ali zameglila (*izbrani govorci* svojih sogovornikov seveda niso izbirali po načelu uravnoteženosti), kar je postalo predmet največkrat kritiziranih delov korpusa BNC. Tako je bilo npr. razmerje med moškimi in ženskami, ki je bilo v izhodišču uravnoteženo (73 moških, 75 žensk), precej neizenačeno že v konverzacijskem podkorpusu (536 moških, 561 žensk), in praktično porušeno, 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

[12] Gre za filozofijo dela korpusnega jezikoslovja, da po neki meji velikosti uravnoteženost ni več kategorija, ki bi jo bilo treba umetno dosegati, saj je implicitno dosežena; zagovornik takega pristopa je npr. J. Sinclair, ki se zavzema za čim enostavnejšo mrežo zajemanja, ob tem pa za čim večje število besedil, kvantiteta torej na prvem mestu (Sinclair 1995: 101), nadalje tudi L. Burnard, ki povzema strategijo dela ameriškega korpusnega jezikoslovja (Mitch Marcus) – »there's no data like more data« (Burnard 2001: 2).

[13] Načrtovalci korpusa so si želeli vključiti večji vzorec, npr. 1000 govorcev, vendar to zaradi praktičnih razlogov – finančnih in časovnih omejitev – ni bilo mogoče; izračunali so, da bo tudi vzorec pribl. 100 govorcev zadostil potrebam (*British National Corpus User Reference Guide*, 2000: 4).

izgovorili moški, izgovorjenih 151 besed, ki so jih izgovorile ženske. Takšno preštevanje govorcev in besed se zdi mogoče na prvi pogled brezpredmetno, vendar še zdaleč ni tako; če govorce klasificiramo na moške in ženske, je to gotovo z namenom, da bi raziskovali morebitne razlike v govorjenem jeziku enih in drugih; za take raziskave pa je treba imeti uravnoteženo razmerje skupin govorcev.[14]

Namen gradnje korpusa BNC je bil v korpus zajeti jezik, ki je statistično reprezentativen za celotno populacijo. Ob tem se je sestavljavcem že na začetku zastavilo 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 že takoj na začetku poskušali določiti kriterije zbiranja besedil tako, da bi v korpus zajeli vse znane oblike besedil. 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 demografskega nabora izpadla besedila, ki jih izbranih 153 govorcev ni produciralo, npr. predavanja, pridige, sodni govori, televizijski intervjuji itd., to je besedila, ki jih večina govorcev predvsem sprejema (posluša), producira pa jih manjšina. Ker gre v teh primerih pogosto za bolj kultivirano obliko govorjenega jezika in za govorce, ki so v določenem smislu izpostavljeni, njihov govor pogosto obvelja za reprezentativnega. Demografski del govorne komponente korpusa BNC so zato dopolnili s t. i. kontekstualnim delom, v katerem so bila besedila zbrana na podlagi besedilne tipologije; v ta del je bilo vključenih 757 besedil, skupaj 6.135.671 besed oz. pribl. 60 % govorne komponente korpusa BNC. Tipologija je bila narejena na podlagi štirih enako velikih področij, vnaprej določenih glede na namen besedila. Vsako področje je bilo navznoter razdeljeno na monologe (40 %) in dialoge (60 %), kar pomeni, da monologi znotraj posameznega področja predstavljajo 10 % besedil v celotnem kontekstualnem delu BNC, dialogi pa 15 %. To so razmerja, ki naj bi zagotavljala reprezentativnost korpusa; znotraj teh določil so bila besedila razvrščena v različne besedilne tipe, število tipov znotraj področij pa ni bilo vnaprej določeno in se je v končni obliki tudi precej razlikovalo:

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

Področje	Besedilni tipi	Št. besedil	Št. besed	%
izobraževanje in informiranje	predavanja ipd.	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	nagovori	195	1.561.167	25.44
	športni komentarji			
	govori v klubih			
	TV in radio – kontaktne oddaje			
	klubski sestanki			

Slika 2: Tematska področja, besedilni tipi in razmerja med njimi v kontekstualnem delu govorne komponente BNC (*BNC Users Reference Guide,* 2000)[15]

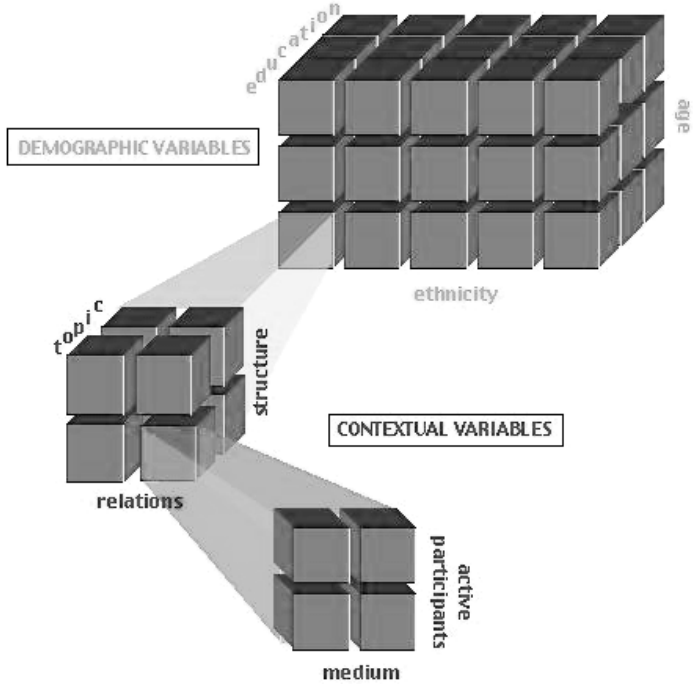
Kot je razvidno iz tabele, je bilo število besedilnih tipov znotraj področij poljubno in se je gibalo od 3 do 6. Tudi število besedil znotraj posameznega področja je bilo poljubno, od 131 do 162; kot je bilo že omenjeno, je bilo znotraj konverzacijskega dela določeno samo razmerje med monologom in dialogom ($M^k : D^k = 40$ % : 60 %). Glede na to, da je demografsko zbrani del govorne komponente BNC obsegal samo dialoge, je bilo skupno razmerje med monologi in dialogi v korpusu $M^{BNC}: D^{BNC} = 15$ % : 85 %.

Tudi v kontekstualnem delu korpusa so, čeprav je bilo izhodišče za zbiranje gradiva drugačno, poskusili ohraniti enakomerno razmerje med govorci v smislu demografske klasifikacije govorcev. Posebna pozornost je bila namenjena spolu govorcev in regijski pripadnosti. Pri spolu so se znotraj vseh kategorij trudili variirati spol govorca, kljub temu pa so naknadne analize pokazale, da je bilo razmerje med spoloma govorcev v kontekstualnem delu korpusa skrajno neuravnoteženo in nereprezentativno (3.199.812 besed so izgovorili moški, 660.071 pa ženske; Berglund 1999: 49, opomba 7). Čeprav gre le za delni prikaz celotne

[15] Podatki se deloma razlikujejo od vira, ki ga je upošteval Gorjanc 2002 (Aston, Burnard 1998).

situacije, ki je bila v resnici še mnogo kompleksnejša, je mogoče sklepati, kako težko je uravnotežiti korpus glede na strukturo govorcev in tipe besedil ter razmerja med njimi. Zato velja pri gradnji govornega korpusa postopati racionalno: zajeti čim širšo mrežo govorcev in tipov besedil ter zbrati čim večjo količino gradiva, nato pa znotraj zbranega gradiva določati uravnotežene podkorpuse.

Tretji primer govornega korpusa, ki ga želim predstaviti, je zaenkrat še nerealizirani načrt gradnje korpusa govorjene izraelske hebrejščine (CoSIH). Sestavljalci so se na podlagi preteklih izkušenj pri gradnji korpusov, predvsem BNC, odločili kombinirati demografsko in kontekstualno komponento govornega korpusa. Kot konceptualno orodje so si v ta namen zamislili večdimenzionalno celično matriko; ogrodje je matrika velikosti 5 x 3 x 3, ki temelji na demografskih komponentah, in sicer etnični pripadnosti (5 kategorij), starosti in izobrazbi (vsaka po tri kategorije). Vsaka izmed dobljenih 45 celic je navznoter sestavljena iz matrike velikosti 2 x 2 x 2, ki jo opredeljujejo kontekstualni kriteriji (odnosi med govorci: formalni/neformalni, struktura pogovora: vodeni pogovor/interakcija, tema: ± osebno); znotraj vsake tako koncipirane celice pa je nova matrika 2 x 2, ki jo določata komponenti monolog/dialog in medij: telefon/neposredno. Na ta način so



Slika 3: Matrična struktura zajema besedil v *Korpus govorjene izraelske hebrejščine*[16]

[16] *The Corpus od Spoken Israeli Hebrew*, http//:www.tau.ac.il/humanities/semitic/cosih.html; izredno natančna projekcija gradnje korpusa (2001); o rezultatih zaenkrat ni poročil.

snovalci korpusa hebrejščine sestavili fiktivno matriko s 1000 celicami, ki so jih nameravali izpolniti 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 kontekstualnih različicah. Seveda so se pri načrtovanju korpusa zavedali, da bo idealne razmere nemogoče doseči, zato so v korpus vgradili tudi varovalne mehanizme, ki naj bi prispevali k reprezentativnosti korpusa.[17]

Odločitev za kombinacijo demografske in kontekstualne komponente govornega korpusa se zdi na prvi pogled smiselna in racionalna, celo mamljiva za vse, ki se na novo lotevajo gradnje govornega korpusa, vendar pa vodi v izredno zapletena izračunavanja, za katera se v nekem trenutku zazdi, da postanejo sama sebi namen in se zaradi njihove zapletenosti raziskovalci lahko izgubijo v računanju, ne nazadnje pa se vhodna uravnoteženost demografske komponente v korpusu lahko tudi podre, če govorci sami izbirajo svoje sogovornike.

Med slovanskimi narodi se gradnjo govornega korpusa prvi realizirali Čehi. V začetku devetdesetih let se je pri načrtovanju novega slovarja češkega knjižnega jezika rodila ideja o gradnji računalniškega korpusa (Čermák 1997: 186). Leta 1994 je bil na Filozofski fakulteti v Pragi ustanovljen Oddelek za Češki nacionalni korpus (ČNK), kar je pomenilo tudi odlično osnovo za razvoj korpusne lingvistike kot posebne znanstvene discipline. Kasneje je bil ustanovljen Inštitut za ČNK, v okviru katerega je sodelovalo 9 raziskovalnih in izobraževalnih institucij iz Prage in Brna; izoblikovala se je znanstveno-raziskovalna skupina, ki se je v marsičem zgledovala po korpusnih centrih v tujini (Čermák 1997: 187), hkrati pa je razvijala lastno raziskovalno dejavnost in gradila korpus. Oteževalna okoliščina je bila, da so pridruženi partnerji pripadali izključno akademski sferi, z izjemo ene založniške hiše, kar je predstavljalo veliko oviro pri zbiranju potrebnih sredstev; večino je prispevala država, del tudi omenjena založniška hiša, sicer pa pri kapitalskih partnerjih za korpus ni bilo interesa.

Pri gradnji ČNK so se raziskovalci deloma oprli na smernice za gradnjo korpusa, ki so jih podali Atkins, Clear in Ostler (Atkins et al. 1992), deloma pa so razvili lastne kriterije, glede na naravo češkega jezika in glede na razpoložljiva sredstva. Govorna komponenta, *Praški govorni korpus* (*Pražský mluvený korpus*, ORAL-PMK), obsega okrog 800.000 besed (Čermák, brez navedbe letnice: 1).[18] Besedila so bila posneta na 304 magnetofonskih trakovih in transkribirana. Pri naboru besedil so se odločili, da bodo v začetku zbirali samo govor Prage z okolico, to je osrednjega govora, ki ga sooblikujejo tudi prišleki, govorci z različnih koncev države. Z omejenimi sredstvi Čehi seveda niso mogli zbirati besedil na enak način, kot so to počeli pri gradnji BNC; metodologija zbiranja je bila poenostavljena: pridobili so nekaj čez 100 govorcev, ki so predstavljali sociološko reprezentativni vzorec (Čermák 1997: 191); najprej so ti govorci odgovarjali na vnaprej pripravljena vpra-

[17] Bistvo varovalnega mehanizma je 5-odstotni del korpusa, ki ga predstavlja samo kontekstualna komponenta, sestavljena iz predvidoma vplivnejših besedil (mediji, parlament, sodišče).

[18] Obstaja tudi *Brnski govorni korpus*, prim. http//:ucnk.ff.cuni.cz/bmk.html.

šanja, nato pa so se morali s sogovorniki po svoji izbiri pogovarjati o poljubnih temah. Vsi pogovori so bili posneti, transkribirani in označeni. Reprezentativnost korpusa so češki raziskovalci želeli doseči z upoštevanjem štirih sociolingvističnih kategorij govorcev:

– spol,
– starost: mlajši (od 20 do 35 let) ali starejši (več kot 35 let) (Iunior/Vetus),
– izobrazba: do mature ali višja/visoka,
– govorni položaj: formalni ali neformalni.

Praški govorni korpus dokazuje, da je mogoče tudi z relativno enostavno mrežo zajema besedil (govorcev) zgraditi uporabno računalniško bazo podatkov za raziskave avtentičnega govora.[19]

Predlog za zajem govorjenih besedil v korpus govorjene slovenščine

Iz pregleda samo nekaterih znanih tipologij govorjenih besedil za zajem v govorne korpuse je razvidno, da si vsaka nova ekipa načrtovalcev korpusa določi svojo pot gradnje. Pri tem je verjetno najvplivnejša taksonomija, ki je bila izdelana za gradnjo govorne komponente BNC, pa tudi ta ni bila nikoli v celoti prevzeta, ampak na najrazličnejše načine modificirana. Pri gradnji reprezentativnega govornega korpusa se je verjetno nemogoče izogniti upoštevanju demografskega in kontekstualnega izhodišča, navznoter pa je ti dve komponenti mogoče poljubno diferencirati (glede na naravo jezika in družbeno situacijo), poleg tega pa je treba rešiti tudi vprašanje, kako ti dve komponenti kombinirati med seboj:

– kot dve komplementarni enoti (pri tem se takoj postavi vprašanje kvantitativnega razmerja med obema enotama) ali
– kot prekrivni sistem, torej v obliki večdimenzionalne matrike.

Za zajem besedil v demografsko komponento je treba statistično določiti vzorec govorcev, ki ustreza celotni populaciji. V našem primeru celotno populacijo predstavljajo (v prvi fazi predvidoma samo odrasli) govorci slovenščine.[20] Kriteriji, ki se jih pri vzorčenju govorcev navadno upošteva, so:

– spol,
– starost,
– regijska pripadnost,
– izobrazba,
– družbeno-ekonomski status,
– drugo.[21]

[19] Osnova govorne komponente ČNK je bil raziskovalni projekt Inštituta bohemističnih študij na Filozofski fakulteti Karlove univerze v Pragi, vezan na pogostnostno analizo govorjenega jezika in kasneje opisa govorjene češčine sploh (Gorjanc 2002: 58).

[20] Kasneje je govornemu korpusu vsekakor mogoče priključiti korpus otroškega in korpus mladostniškega govora; ti običajno nastajajo posebej in imajo v okviru celote status podkorpusa.

[21] Možnosti so še npr. kraj bivanja, kraj rojstva, verska pripadnost, spolna usmerjenost ipd.

Našteti kriteriji lahko v resnici služijo samo kot izhodišče, saj vsak kriterij zase zahteva poseben premislek in odpira številna vprašanja. Že pri definiranju »govorcev slovenščine« bi se bilo treba opredeliti, ali vzorčimo samo govorce slovenščine kot prvega jezika ali upoštevamo tudi delež priseljencev, kakršnega npr. izkazuje statistični popis prebivalstva.[22] Pri regijski pripadnosti je verjetno treba upoštevati tudi ustrezni delež zamejskih govorcev in morda tudi izseljencev, drugo temeljno vprašanje v zvezi s tem pa je, ali govorce opredeljujemo po kraju rojstva ali kraju bivanja (in kako dolgo morajo potemtakem že živeti v določenem kraju). Razlike v izobrazbi govorcev se predvidoma izražajo tudi v govoru posameznikov, vendar se spet postavlja vprašanje, kako določiti »izobrazbene intervale«, na katerih prihaja do razlik. Kriterij družbeno-ekonomskega statusa pa se zdi še najbolj neoprijemljiva kategorija, saj ni jasno, kako ga sploh opredelimo (življenjski standard, mesečni prihodki, položaj v družbi, položaj na delovnem mestu …), poleg tega pa je tudi vprašanje, ali tovrstni status vpliva na govor posameznika.[23]

Kontekstualno komponento najpogosteje določajo naslednji (med seboj pogosto prekrivni) kriteriji:

– spontani : pripravljeni govor,
– monolog : dialog,
– namen besedila (izobraževanje, informiranje, institucionalni govor, poslovno/službeno sporazumevanje, vsakodnevno družabno sporazumevanje, zabava in prosti čas),
– tematika,
– okoliščine (zasebno : javno),
– govorni položaj (formalno : neformalno).

Tudi tu se načrtovalcu korpusa postavljajo zahtevna vprašanja. Za raziskovanje avtentičnega govorjenega jezika so najbolj dragoceni posnetki spontanega govora, zato se jih jezikoslovci trudijo čim več zajeti v korpus. Nadalje se je treba odločiti za kvantitativno razmerje med dialogom in monologom, še težja pa je določitev tematskih ali funkcijskih (namen) področij korpusa. Vnaprej je jasno, da bodo meje med področji zabrisane, prehodne in da bo besedila pogosto težko razvrščati v posamezne kategorije. Kategoriji formalno/neformalno in zasebno/javno moramo obravnavati ločeno. Za nadaljnje raziskovanje govora, čemur korpus služi, sta obe kategoriji, govorni položaj in okoliščine, lahko zelo pomembni. Kategorija zasebnega govora bi bila v demografski komponenti korpusa gotovo zastopana v večji meri kot kategorija javno, v kontekstualni pa bi bili kategoriji kot vhodni komponenti lahko vnaprej določeni (torej zahtevano določeno število zasebnih in javnih besedil). Pri zajemu besedil v korpus je treba vsaj okvirno ločiti tudi med besedili

[22] Tudi korpus govorcev, za katere slovenščina ni prvi jezik (korpus usvajanja jezika), je za raziskovanje jezika izjemno pomemben podkorpus.

[23] Prim. članek Andreja E. Skubica *Sociolekti od izraza do pomena: kultiviranost, obrobje in eksces* v tem zborniku (str. 297–320).

množične produkcije (vsakodnevno praktično ali družabno sporazumevanje) in besedili množične recepcije, to je besedili, ki jih zaznamuje množični naslovnik in ki pogosto obveljajo za reprezentativna (medijski, pedagoški govor). Če sta demografska in kontekstualna komponenta znotraj korpusa komplementarni enoti, potem je to relativno enostavno upoštevati, seveda znotraj kontekstualne komponente.

Korpus govorjene slovenščine (KGS) lahko sestavimo iz dveh komplementarnih podkorpusov: prvega oblikuje demografska klasifikacija govorcev (*konverzacijski podkorpus*), drugega pa taksonomija besedilnih tipov (kontekstualni podkorpus). Če bi pri gradnji sledili korpusu BNC, bi med obema komponentama vzpostavili razmerje:

$$Dk : Kk = 40\ \% : 60\ \%.$$

Demografska komponenta (Dk) obsega predvsem spontani govor, saj izbrani govorci, ki predstavljajo reprezentativni vzorec glede na celotno populacijo, v določenem obdobju snemajo svoj govor ne glede na namen in tematiko. Demografska komponenta korpusa je za jezikoslovne raziskave avtentičnega govora najdragocenejše gradivo (to govori v prid čim večjega deleža Dk), a je obenem tudi izredno zahtevna za zbiranje in za transkripcijo, kar njeno velikost omejuje.

Kot vhodne kriterije za besedila demografske komponente bi bilo mogoče izbrati naslednje kategorije:

Spol:	M		Ž		
Starost:	18 do 30 let	31–55		od 56 naprej	
Izobrazba:	OŠ (8 let šolanja)	SŠ+viš+vis (9–16 let šolanja)		univ. in več (17 let šolanja oz. več)	
Regijska pripadnost:	osrednja	S in SZ	Z	SV	J in JV

Slika 4: Predlog demografske komponente KGS

Kako določiti starostne intervale, v katerih naj bi pri govorcih prihajalo do občutnejših razlik v govorjenju, je vprašanje, na katero bi bilo v resnici mogoče odgovoriti šele na podlagi analize avtentičnih govorjenih besedil, to je korpusa govorjenih besedil. Zato se je pri zajemu besedil treba za intervale odločiti na podlagi hipotez; BNC je npr. določil 6 starostnih intervalov po pribl. 10 let (in od 60 let naprej enotno), ČNK pa samo dva intervala – nad ali pod 35 let. Zgoraj je predlagana rešitev nekako v smislu mlajši, srednji in starejši, mogoče pa bi bilo zagovarjati tudi delitev na dva intervala z mejo nekje med 30 in 35 let. Podobno je tudi z izobrazbo: število intervalov je lahko poljubno, dejanska analiza stanja bo mogoča šele na podlagi zbranega gradiva, predvidevati pa je mogoče, da število let šolanja vpliva na govor posameznika, zato je predlagana delitev na 3 intervale. Pri regijski pripadnosti predlagamo delitev, ki se le deloma ujema s tipologijo slovenskih narečij: (S in SZ) skupina vključuje gorenjsko in koroško narečno skupino, (Z) primorsko in rovtarsko, (SV) štajersko in panonsko, (J in JV) dolenjsko narečno

skupino, dodana pa je kategorija »osrednja«, ki se nanaša na Ljubljano z okolico, s predpostavko, da gre v tem regijskem delu za govor, ki ne pripada nobeni narečni skupini in ki ima zaradi koncentracije in slišanosti medijev ter koncentracije šol in drugih institucij na vse govorce slovenščine zelo močan vpliv.

Reprezentativni vzorec govorcev glede na izbrane kriterije (število govorcev v posamezni skupini) bi moral biti statistično izračunan glede na celotno populacijo. Izbrani govorci bi morali nato snemati določeno količino svojih vsakodnevnih pogovorov (merjeno v urah ali dnevih), ob tem pa tudi zapisovati podatke o svojih sogovornikih.[24] Vzorec govorcev, ki bi ga dobili, ko bi zbrali vse posnetke, zagotovo ne bi bil več reprezentativen, vhodna razmerja bi se porušila, vendar bi bilo reprezentativnost mogoče spet vzpostaviti z oblikovanjem posebnega podkorpusa znotraj vseh posnetkov ali s krčenjem in dodajanjem posnetkov do uravnoteženosti, še najbolje pa bi se uravnoteženost zagotavljala s čim večjo količino podatkov oz. gradiva.

Demografska komponenta	40 %	
Kontekstualna komponenta	60 %	
Izobraževanje in informiranje	15 %	predavanja učne ure interakcija v razredu poročila komentarji okrogle mize
Javni oz. institucionalni govor	15 %	politični govor seje/sestanki vlade in drugih organov parlamentarni govor/razprava versko srečanje pridiga govor v sodnih postopkih
Poslovna oz. poklicna komunikacija	15 %	govori v podjetjih razgovori za sprejem na del. mesto govori na kongresih, konferencah prodajne demonstracije poslovni sestanki, svetovanje
Zabava in prosti čas	15 %	nagovori, zdravljice socialna interakcija športni komentarji TV in radio – kontaktne oddaje

Slika 5: Možnost zajema besedil v KGS na podlagi funkcije/namena besedila

[24] Gre za t. i. identifikacijski list govorca, vprašalnik o osebnih podatkih, ki ga izpolni govorec sam. V resnici je postopek še bolj zapleten, saj govorec v najboljšem primeru ne ve, da se pogovor snema, ob koncu pogovora je s tem seznanjen, zaprošen za sodelovanje in za podpis izjave, s katero odstopa avtorske pravice govora (posnetka in transkripcij) v znanstvene namene; šele potem izpolni identifikacijski list.

V izhodišče kontekstualne komponente je smiselno postaviti namen/funkcijo besedila, in sicer v štirih kategorijah:

– izobraževanje in informiranje,
– poslovna in poklicna komunikacija,
– javni oz. institucionalni govor,
– zabava in prosti čas.

Razdelitev na 4 osnovne kategorije je v tem primeru enaka kot pri BNC.

Eno izmed vprašanj, ki se zastavlja znotraj kontekstualne komponente, je vprašanje javnosti/zasebnosti (okoliščine) in vprašanje formalnosti/neformalnosti (govorni položaj). 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. Predvidevamo lahko, da bi bil v enotah Izobraževanje/informiranje in Institucionalni govor delež javnosti večji od deleža zasebnosti, nasprotno pa v enoti »Zabava/prosti čas« pričakujemo bistveno več govora v zasebnih okoliščinah; tudi besedilom, zbranim v demografski komponenti, predvidoma ne bi bilo težko določiti statusa javnosti oz. zasebnosti, predvidevamo pa lahko, da bo pri povprečnem vzorcu govorcev delež zasebnega govora bistveno večji od deleža javnega govora. Problem nastane pri določitvi statusa besedilom, zbranim v kontekstualni enoti »Poslovna/poklicna komunikacija«: če si kot besedilni vzorec 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), ga vseeno ne moremo imeti za zasebno besedilo.[25] Zato je morda komplementarno nasprotje, ki opisuje okoliščine govora, bolje poimenovati javno : nejavno; v tem primeru lahko predvidevamo, da bi tako zgrajen korpus vseboval približno 30 % besedil, govorjenih v javnosti.

Vnaprejšnje določanje govornega položaja (v smislu formalno/neformalno) bi bilo v nasprotju s temeljnim izhodiščem korpusne lingvistike, ki jezik opisuje šele na podlagi zbranih empiričnih podatkov. Vprašljivo pa je tudi označevanje že zbranih besedil z omenjenima oznakama, saj pogosto razen res očitnega, pa ne vedno zadostnega pogoja za določanje formalnosti (tikanje/vikanje) težko določimo druge objektivne kriterije. Tako demografska kot kontekstualna komponenta govornega korpusa lahko, če sta dovolj veliki, zagotavljata, da bodo v besedilih zastopana vsa jezikovna sredstva, ki se pojavljajo v govornih položajih z različno stopnjo formalnosti.[26]

Pri gradnji govornega korpusa za slovenščino bi bilo koristno pri zajemanju besedil v kontekstualno komponento zagotoviti dovolj besedil, govorjenih v javnosti, tako da bi ta besedila kot podkorpus omogočala študije javnega govora (med 30 in 40 % vseh besedil). Kriterijev formalno/neformalno v korpusu ne bi označevali niti

[25] Beseda *zaseben -bna -o* je v SSKJ razložena kot »nanašajoč se na posameznika kot neuradno osebo«, *zasebnost -i* pa kot »razlika med zasebnostjo in javnim delovanjem«.

[26] Razlika med kriterijema javno/nejavno in formalno/neformalno je tudi v tem, da gre v prvem primeru za dihotomijo (+/–), v drugem pa za celo paleto govornih položajev z večjo ali manjšo stopnjo (ne)formalnosti.

kot vhodni kategoriji niti v glavi, ker bi morali pri označevanju pristati na določeno mero subjektivnosti, poleg tega pa pričakujemo, da bi študije, ki bi obravnavale problematiko formalnega/neformalnega govora lahko nastale šele na podlagi že zgrajenega govornega korpusa.

Spodnja slika predstavlja eno izmed mogočih shem za zajem besedil v govorni korpus za slovenščino:

Vhodni kriteriji	**Komponenta korpusa**	**Namen besedila**	**Delež v %**	**Dialog : monolog**	**Javno : nejavno**
spol starost izobrazba regija	*demografska 40 %*	ni določen	40	prevladuje D	prevladuje Nj
namen/funkcija struktura (M/D) okoliščine (J/Nj)	*kontekstualna 60 %*	izobraževanje/ informiranje	15	prevladuje M	prevladuje J
		javni/ instituc. govor	15	predvidoma prevladuje M	prevladuje J
		poslovna/ službena kom.	15	predvidoma prevladuje D	predvidoma prevladuje Nj
		prosti čas/ zabava	15	prevladuje D	prevladuje Nj

Slika 6: Shematski prikaz predloga za zajem besedil v KGS

Kvantitativno razmerje med dialogom in monologom bi bilo mogoče opredeliti šele, ko bi bila vsa besedila zbrana, predvidevamo pa lahko, da bi bilo pri taki razdelitvi dialogov blizu 70 %. Verjetno je delež monologov v realnosti še manjši, vendar imajo izbrani govorci monologov lahko precej veliko »slišanost«, njihov govor pa lahko obvelja za reprezentativnega (pedagoški, politični govor, govor popularnih voditeljev v medijih); tudi transkribiranje monologov je seveda mnogo manj zahtevno kot transkribiranje govora, v katerem sodeluje več govorcev. Tudi razmerje med javnimi in nejavnimi besedili bi bilo znano šele po končanem zajemu besedil, predvidevamo pa lahko, da bo javnih besedil med 30 in 35 odstotki.[27]

Struktura tako sestavljenega korpusa bi poleg jezikoslovnih raziskav in opisov govorjenega jezika omogočala tudi sociološke in druge raziskave glede na obravnavane kriterije spol, starost, izobrazbo in regijsko pripadnost govorcev ter namen, strukturo besedila in okoliščine, v katerih je nastalo.

Seveda pa se postavlja vprašanje, ali je za slovenščino in za obstoječe raziskovalne potenciale in finančne možnosti res najboljša usmeritev slediti strukturi sicer izvrstnega govornega korpusa britanske angleščine. Mogoče bi bilo namreč razmišljati tudi v povsem drugi smeri, kjer demografska in kontekstualna komponenta

[27] Zadnja alinea predstavlja tudi razmerje med besedili množične recepcije (govorjena besedila, ki dosegajo velik krog poslušalcev) in besedili množične produkcije.

ne bi bili komplementarno ločeni. Celotni govor bi bilo namreč mogoče že v izhodišču razdeliti glede na okoliščine, v katerih se besedilo govori, npr. v tri osnovne kategorije: javno, nejavno (pol)uradno in zasebno:

Javna besedila	30 %	predavanje, učna ura, komentar (dogajanja), okrogla miza, politični govor, odprta seja vlade in/ali drugih organov, parlamentarni govor/razprava, pridiga, govor v sodnih postopkih, prodajna demonstracija, govor na kongresih, konferencah, športni komentar, kontaktne oddaje na TV in radiu ...
Nejavna (pol)uradna besedila	30 %	interakcija v razredu, pogovori na delovnem mestu, razgovori za sprejem na delovno mesto, govori v sodnih postopkih, parlamentarni govor/razprava, poslovni sestanki, svetovanje, pogovor med zdravnikom in pacientom, pogovor v trgovini/na trgu, na okencu katerega koli urada ali storitvene dejavnosti, pogovori z uradnimi osebami, dialog na izpitu, telefonski pogovor z uradno osebo, iskanje informacij
Zasebna besedila	30 %	nagovori, zdravljice socialna interakcija – znotraj družine, kroga prijateljev, znancev

Slika 7: Možnost zajema besedil v KGS na podlagi okoliščin nastanka besedila

Predvidevamo, da gre pri zgornjih kategorijah za troje vrst okoliščin, ki narekujejo izbiro značilnih jezikovnih sredstev: na eni strani so to javna besedila, na drugi zasebna, posebno kategorijo pa predstavljajo besedila, ki jih ne moremo umestiti v nobeno izmed teh dveh kategorij (niso niti javna niti zasebna), pa vendarle obsegajo precejšen del vsakodnevnega sporazumevanja vsakega človeka. V kategoriji javnih besedil se poleg besedil, govorjenih pred občinstvom, pojavljajo tudi besedila medijev (radio in televizija), v preostalih dveh kategorijah pa poleg prosto govorjenih besedil še besedila, govorjena po telefonu. Pri takem zajemu bi si vnaprej zagotovili veliko število različnih tipov besedil, težje pa bi bilo zagotavljati uravnoteženost po tematskih področjih in po demografskih značilnostih govorcev. Velika količina podatkov seveda zagotavlja večjo raznovrstnost besedil in govorcev, vendar pa ravno pri govornih korpusih s kvantiteto zaenkrat težko zagotavljamo uravnoteženost (kot je bilo že večkrat omenjeno – zaradi zahtevnosti zbiranja podatkov se zaenkrat ni mogoče niti približati velikostim pisnih korpusov). Demografsko raznolikost govorcev bi do neke mere lahko zagotovili tako, da bi snemali govorjena besedila v situaciji, kjer se izmenjuje veliko različnih govorcev (npr. na

informacijah o voznih redih na železniški postaji);[28] tematsko raznolikost bi bilo treba načrtovati že znotraj okvira zajema besedil, prav tako pa tudi demografsko raznolikost govorcev, npr. z izbiro samih govorcev in/ali z izbiro medijev.

Predstavljena sta bila dva izmed mogočih predlogov za zajem besedil v govorni korpus slovenščine, prvi prevzet po modelu BNC s slovenski jezikovni in družbeni situaciji prilagojenimi vhodnimi kriteriji, drugi pa izviren; slednji naj bi bil predvsem bolje prilagojen razpoložljivim tehničnim, strokovnim in finančnim možnostim.

Ne nazadnje (v resnici pa pravzaprav na začetku) se morajo načrtovalci korpusa odločiti tudi o velikosti korpusa, ki ga nameravajo zgraditi. Večkrat je bilo že omenjeno, da je gradnja govornega korpusa zaradi zbiranja zvočnih posnetkov in transkribiranja besedil izredno zahtevno in zamudno opravilo, ki zahteva dobro tehnično opremljenost ekipe, izurjeno in usklajeno ekipo sodelavcev ter veliko časa. Glede na izkušnje, zbrane ob gradnji drugih govornih korpusov, bi v primeru slovenščine verjetno lahko razmišljali o govornem korpusu velikosti med enim in petimi milijoni besed. Za velikost korpusa bo odločilnega pomena odločitev, kako bodo besedila transkribirana. Po priporočilih ekspertne grupe za gradnjo korpusov Eagles je za govorne korpuse, namenjene predvsem leksikalnim in skladenjskim raziskavam, najprimernejša oblika zapisovanja zvočnega signala v t. i. ortografski transkripciji (besede so zapisane v skladu s pravopisno tradicijo, določi pa se lista besed, oblik in morfemov, ki se pojavljajo v govorjenem jeziku, v standardnem zapisu pa jih ni).[29] Vsekakor bi bilo izrednega pomena najprej zgraditi vzorčni reprezentativni korpus velikosti med 100.000 in 200.000 besed in z njim preveriti ustreznost zajema besedil in načel transkripcije. Po morebitnih korekcijah teoretičnih predpostavk in po oceni tehnične izvedljivosti bi lahko nadaljevali z gradnjo korpusa do določene velikosti.

Pri gradnji govornega korpusa si prizadevamo najti pot, kako predstavljati in opisovati govorjeni jezik kot celoto na podlagi omejene zbirke posnetkov govora. Pri tem sledimo hipotezi, da se jezik posameznih govorcev glede na določene kriterije razlikuje (oz. jezik vseh govorcev šele tvori celotno podobo), razlikuje pa se tudi glede na govorno situacijo, v kateri nastane, in glede na namen, zaradi katerega nastane. Zato poskušamo doseči reprezentativnost vzorca tako, da izberemo govorce, ki predstavljajo reprezentativni vzorec celotne populacije, in ga dopolnimo z reprezentativnim vzorcem besedil glede na taksonomijo govorjenih besedil. Na ta način je določeno izhodišče za zajem besedil v govorni korpus in s tem za gradnjo referenčnega korpusa, prepotrebnega jezikovnega vira za raziskovanje jezika.

[28] Predlog M. Stabeja na konzultacijah aprila 2004.

[29] Zahtevnejše oblike transkribiranja so po vrsti (fonemska, alofonska, akustično-fonetična in prozodična transkripcija (EAGLES, 1995); običajno je del govornega korpusa tudi fonetično ali prozodično transkribiran.

Viri in literatura

Sue Atkins, Jeremy Clear, Nicholas Ostler, 1992: Corpus Design Criteria. *Literary and Linguistics Computing* 7/1. 1–16.

Ylva Berglund, 1999: Exploiting a Large Spoken Corpus: An End-user's Way to the BNC. *International Journal of Corpus Linguistics* 4/1.

Douglas Biber, 1993: Representativeness in Corpus Design. *Literary and Linguistics Computing* 8/4. 243–257.

Lou Burnard, 2001: Where did we go wrong? A retrospective look at the design of the BNC. *SILFI 6th International Conference, »Spoken Italian«, Congress Proceedings*. (Duisburg, 28. 6.–2. 7. 2000). Http://users.ox.ac.uk/~lou/wip/silfitalk.html.

Steve Crowdy, 1993: Spoken Corpus Design. *Literary and Linguistics Computing* 8/4. 259–265.

František Čermák, 1997: Czech National Corpus: A Case in Many Contexts. *International Journal of Corpus Linguistics* 2/2. 181–197.

František Čermák, 2001: *Pražský mluvený korpus*. Http://ucnk.ff.cuni.cz/pmk.html.

František Čermák, Věra Schmiedtová: *The Czech National Corpus Project: Its Structure and Use*. Http://ucnk.ff.cuni.cz/doc/czechnationalcorpus.doc.

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.

Tomaž Erjavec, 1996/97: Računalniške zbirke besedil. *Jezik in slovstvo* 2/3. 81–95.

Vojko Gorjanc, 2002: Jezikovna infrastruktura: kje je tu slovenščina? *38. seminar slovenskega jezika, literature in kulture. Zbornik predavanj*. Ur. B. Krakar Vogel. Ljubljana: Filozofska fakulteta. 257–270.

Vojko Gorjanc, 2002: *Jezikoslovna načela gradnje računalniških besedilnih zbirk strokovnih jezikov*. Doktorska disertacija. Ljubljana: Filozofska fakulteta.

Vojko Gorjanc, 2003: Korpusi in jezikoslovje. *Jezik in slovstvo* 48/3–4. [19]–27.

Sidney Greenbaum, Jan Svartvik, 1990: The London-Lund Corpus of Spoken English. *The London Corpus of Spoken English. Description and Research*. Ur. J. Svartvik. Lund University Press. Http://helmer.aksis.uib.no/icame/london-lund/.

Shlomo Izre'el, Benjamin Hary, Giora Rahav, 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.

Joakim Llisteri, 1996: *Preliminary Recommendations on Spoken Texts*. EAGLES (Expert Advisory Group on Language Engineering Standards).

Geoffrey Leech, Greg Myers, Jenny Thomas (ur.), 1995: *Spoken English on Computer. Transcription, Mark-up, and Applicaton*. New York: Longman.

Michael McCarthy, 1998: *Spoken Language and Applied Linguistics*. Cambridge University Press.

John, Sinclair, 1995: From theory to practice. *Spoken English on Computer. Transcription, Markup and Applications.* Ur. G. Leech, G. Myers, J. Thomas. Harlow: Longman. 99–112.

Marko Stabej, 1998: Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6. Jezikovne tehnologije, tematska številka. Ur. Z. Kačič. 96–106.

Marko Stabej, Primož Vitez, 2000: KGB (korpus govorjenih besedil) v slovenščini. *Jezikovne tehnologije*. Zbornik konference. 17.–19. oktober 2000. Ur. T. Erjavec, J. Gros. Ljubljana: Institut Jožef Stefan. 79–81.

Marko Stabej, 2003: Jezikovne tehnologije in jezikovno načrtovanje. *Jezik in slovstvo* 48/3–4. [5]–18.

Jože Toporišič, 1984: *Slovenska slovnica*. Maribor: Založba Obzorja.

Primož Vitez, 1998: Zunajjezikovne okoliščine neidealnega govora. *Jezikovne tehnologije za slovenski jezik. Zbornik konference* (Mednarodna multi-konferenca Informacijska družba – IS'98, Ljubljana, oktober 1998). Ur. T. Erjavec, J. Gros. Ljubljana: Institut Jožef Stefan. 81–83.

Primož Vitez, 1999: Od idealnih jezikovnih struktur k strategiji realnega govora. *Slavistična revija* 47/1. [23]–48.

Claire Warwick, 1997: *The Spoken Component of the BNC*. Http://www.hcu.ox.ac.uk/BNC/what/spokdesign.html.