Mojca Stritar
Ljubljana
mojca.stritar@guest.arnes.si

UDK 371.3:811.163.6'243

MERILA ZA OBLIKOVANJE KORPUSOV USVAJANJA TUJEGA JEZIKA

Korpus usvajanja tujega jezika predstavlja jezik, kot ga govorijo ali pišejo tisti, ki niso njegovi rojeni govorci. V članku so predstavljeni obstoječi korpusi usvajanja za angleščino, francoščino in norveščino kot tuji jezik. Za čim večjo uravnoteženost je pri oblikovanju tovrstnih korpusov treba upoštevati različna merila, med njimi prvi in ciljni jezik, prenosnik, tehniko zajemanja besedil, velikost korpusa, vrsto besedil, temo, merila, povezana s tvorci, kontrolni korpus in uporabnost. Rešitve v obstoječih korpusih so ovrednotene z vidika načrtovanja korpusa usvajanja slovenščine kot tujega jezika, v zaključnem delu pa je podan okviren predlog za oblikovanje tega korpusa.

0 Uvod

Eden od temeljnih vidikov slovenskega jezikovnega načrtovanja je slovenščina kot tuji jezik. Njeni govorci bi morali biti vključeni »pri vseh fazah načrtovanja sodobnih jezikovnih priročnikov slovenščine, od faze oblikovanja besedilnega korpusa do jezikoslovne in leksikografske obdelave gradiva« (Stabej 2004: 12). Da bi jim omogočili čim lažje usvajanje, je treba še bolje spoznati načela, po katerih usvajanje poteka. Pri tem starejše metode vedno bolj podpirajo sodobna spoznanja korpusnega jezikoslovja. Posebej koristni so korpusi usvajanja tujega jezika, ugotovitve katerih temeljijo na dejanski jezikovni rabi.

Korpus usvajanja tujega jezika je specifičen tip korpusa, ki predstavlja jezik, kot ga pišejo ali govorijo tisti, ki niso njegovi rojeni govorci. Osnovni cilj je zbrati objektivne podatke za opis tega jezika in s tem bolje razumeti proces usvajanja, pridobljene informacije pa so uporabne tudi v pedagoškem procesu.

V pričujočem prispevku bo na kratko predstavljenih nekaj obstoječih korpusov usvajanja tujega jezika. Ker korpus usvajanja slovenščine kot tujega jezika[1] še ne

[1] Termin *tuji jezik* v tem primeru zajema tako tuji kot drugi jezik, dejansko gre torej za korpus usvajanja slovenščine kot drugega ali tujega jezika.

obstaja, bodo ključni vidiki oblikovanja obstoječih korpusov ovrednoteni glede na predvideno ustreznost pri načrtovanju slovenskega korpusa. Neposredne vzporednice med tujimi in nastajajočim slovenskim korpusom so seveda nemogoče, saj je kar 23 od 26 pregledanih korpusov[2] narejenih za angleščino, ki se od slovenščine razlikuje tako v temeljnih jezikoslovnih kot tudi sociolingvističnih vprašanjih. Ne glede na to je treba pri načrtovanju slovenskega korpusa upoštevati tuje izkušnje. Seveda je vsak korpus oblikovan specifično glede na namen in ne nazadnje tudi glede na finančna sredstva, vendar je pri vseh opazen vpliv korpusa *International Corpus of Learner English* (*ICLE*), ki je najstarejši akademski in najbolj razširjeni tovrstni projekt.

Pri korpusih usvajanja tujega jezika je posploševanje rezultatov analize na celotno populacijo tujih govorcev samo po sebi vprašljivo (Leech 1997: xix), zato pri njih o referenčnosti niti ne govorimo. Oblikovanje pa mora kljub temu upoštevati merila za čim večjo uravnoteženost. Med merili, povezanimi z jezikom, so prenosnik, vrsta besedil, slogovni postopek in tematika. Merila, povezana z načinom nastanka v korpus vključenih podatkov, obsegajo način zbiranja podatkov (longitudinalna ali presečna raziskava), spodbudo za nastanek (spontano ali po navodilih), uporabo referenc in časovno omejenost pri tvorjenju. Merila, povezana s tvorci posameznih sestavnih besedil, pa vključujejo njihove notranje kognitivne in afektivne kriterije, prvi jezik, učno okolje ciljnega jezika in trenutno jezikovno zmožnost v ciljnem jeziku (Tono 2003: 800).

1 Jezik

Pri korpusih usvajanja tujega jezika sta za razliko od nespecializiranih korpusov pomembna dva jezika: ciljni, torej jezik, »ki se ga nekdo uči z namenom, da bi ga obvladal bodisi kot svoj prvi, drugi ali tuji jezik« (Pirih Svetina 2005: 139), in izhodiščni jezik, »iz katerega se nekdo uči vse druge ali tuje jezike. Iz tega najpogosteje izvirajo napake, ki jih pripisujemo jezikovnemu prenosu (interference)« (Pirih Svetina 2005: 140).

Kot prikazuje spodnja tabela, velika večina korpusov vključuje en izhodiščni in en ciljni jezik, ki je praviloma angleščina, ker lahko glede na njeno razširjenost tuji govorci angleščine z določenim prvim jezikom oblikujejo svoj korpus. Pri manjših jezikih, kjer je tudi manj konkurence med že obstoječimi korpusi, je situacija drugačna. Francoski korpus *FRIDA* in norveški *ASK* imata več izhodiščnih jezikov, ki naj bi ustrezali jezikom največjih skupin priseljencev v tej državi. Glede na to, da slovenščino kot tuji jezik govorijo ljudje z najrazličnejšimi prvimi jeziki, se bo korpus usvajanja slovenščine verjetno moral zgledovati po njiju in imeti več izhodiščnih jezikov.

[2] Pregledani gotovo niso vsi, ker podatki o vseh niso dosegljivi, poleg tega pa bliskovito nastajajo novi, npr. korpus grščine kot tujega jezika ali korpus španščine kot tujega jezika *CEDEL2* (osebni vir M. S., april 2006).

Korpus	Izhodiščni jezik	Ciljni jezik
ASK	10 jezikov	norveški
CEJL	japonski	angleški
CLC	150 jezikov	angleški
CLEC	kitajski	angleški
ESFSLD	arabski, finski, italijanski, pundžabski, španski, turški	angleški, francoski, nemški, nizozemski, švedski
EVA	norveški	angleški
FRIDA	različni[3]	francoski
HKCCE	kantonski	angleški
HKUST	kantonski	angleški
IBLC	angleški, francoski, finski, nemški, tajski itn.	angleški
ICLE	19 jezikov	angleški
ISLE	nemški, italijanski	angleški
Japanese Learners' Corpus	japonski	angleški
JEFLL	japonski	angleški
JPU	madžarski	angleški
LCLE	160 jezikov	angleški
LINDSEI	francoski, kitajski, italijanski, japonski, bolgarski, španski, švedski	angleški
MELD	arabski, bengalski, gudžaratski, haitski kreolski, hindujski, mandarinski, poljski, španski, tajvanski, vietnamski	angleški
PELCRA	poljski	angleški
PELE	poljski	angleški
POLY U	kantonski	angleški
SST	japonski	angleški
Taiwanese Corpus	kitajski	angleški
TELC	tajski	angleški
TELEC	kantonski	angleški
USE	švedski	angleški

Tabela 1: *Izhodiščni in ciljni jeziki v korpusih usvajanja tujega jezika*.[4]

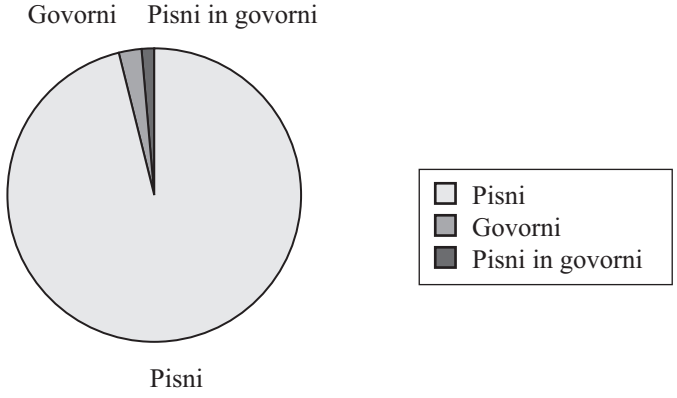
2 Prenosnik

Pri referenčnih korpusih, »ki naj bi predstavili jezik kot celoto« (Gorjanc 2005: 8), je ključno vprašanje prenosnika, v katerem so posredovana sestavna besedila. Razmerje med pisnimi in govorjenimi besedili je v tovrstnih korpusih sicer nerealno glede na resnično rabo, saj je zapisane tekste mnogo lažje pridobivati in

[3] Natančen podatek o tem v času raziskave ni bil dosegljiv.

[4] Seznam upoštevanih korpusov in njihovih kratic je naveden na koncu prispevka.

vključevati, transkripcija govorjenih pa je izjemno naporna in zamudna. Tako je v britanskem nacionalnem korpusu *BNC* le deset odstotkov govorjenih besedil.[5] Ker je referenčnost v korpusih usvajanja tujega jezika precej bolj utopična, so pri gradnji manj poudarjena govorjena besedila in se začenja s pisnimi korpusi. Od 26 pregledanih korpusov je 18 samo pisnih, pet govornih, trije pa imajo govorni in pisni del, pri čemer je pisni vedno večji od govornega.



Slika 1: *Razmerje med številom besed v pisnih, govornih ter pisnih in govornih korpusih*.

Referenčni govorni korpus za slovenščino, ki bi reševal načelna vprašanja tega tipa, šele nastaja (prim. Zemljarič Miklavčič 2004), zato bi bilo na slovenskem področju zaenkrat smiselno razmišljati predvsem o pisnem korpusu usvajanja slovenščine kot tujega jezika.

3 Tehnika zajemanja besedil

Zaenkrat se pisna besedila za korpuse usvajanja tujega jezika povsod pretipkava, saj je večina napisanih na roko. Najlažje dosegljive in obvladljive vire namreč nudijo priložnosti, kjer je skupaj veliko tujih govorcev, ki morajo pisati – v praksi to pomeni izpite iz znanja jezika ter jezikovne tečaje. Optično branje zaradi rokopisa ni izvedljivo. Vse pogosteje pa učeči se domače naloge in druga besedila napišejo v elektronski obliki, čeprav je njihovo vključevanje v korpus zaradi možnosti popravljanja in pomoči rojenih govorcev ob nastajanju ter zaradi uporabe različnih jezikovnih virov vprašljivo.

4 Velikost korpusa

Prav zaradi pretipkavanja posameznih sestavnih tekstov se korpusi usvajanja po velikosti težko primerjajo z nespecializiranimi korpusi. Zaradi tega seveda niso neuporabni, saj za mnoge raziskave zadostuje že dvesto tisoč besed (Kennedy 1998:

[5] Povzeto po spletni strani *BNC*.

73–74). Manjše korpuse je enostavneje graditi, obdelovati in interpretirati rezultate, res pa je, da je ugotovitve težje posploševati in da je verjetneje, da se določene besede ali strukture sploh ne bodo pojavile. V končni fazi je velikost odvisna od namena raziskave. Za analize tega, kateri leksemi se najpogosteje pojavljajo, zadostujejo korpusi z dvajset tisoč besedami. Ustanove, ki razvijajo komercialne pripomočke za poučevanje jezika, pa potrebujejo večmilijonske korpuse kot vire pri izdelavi slovarjev ali učbenikov. Tak je desetmilijonski *Longman Learners' Corpus* (*LCLE*), kjer so prav zaradi velikosti zunajjezikovni podatki o učečem se omejeni na najmanjšo možno mero (Granger 1998: 11).

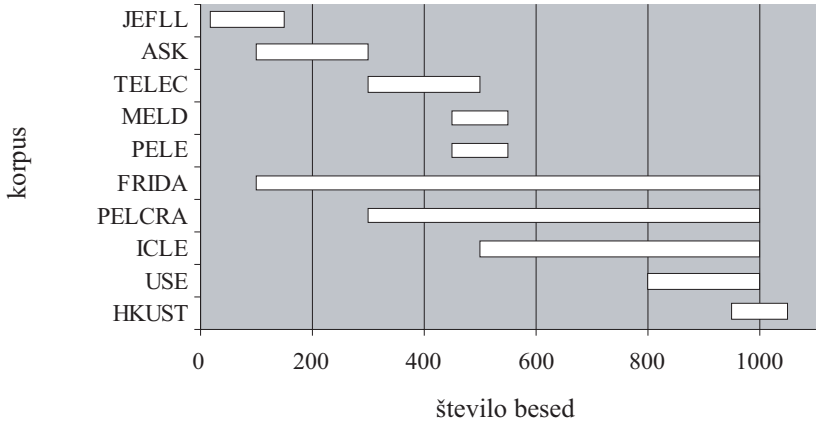
Korpus	**Število besed (samo pisni korpus)**	**Število besed (samo govorni korpus)**	**Število besed (pisni in govorni korpus)**[6]
HKUST	25.000.000		
CLC	15.000.000		
LCLE	10.000.000		
TELEC	3.000.000		
ICLE	2.000.000		
TELC	1.300.000		
SST		1.000.000	
USE	1.000.000		
CEJL	1.000.000		
CLEC	1.000.000		
Taiwanese Corpus	730.000		
HKCCE		500.000	
PELCRA			500.000
ASK	500.000		
JPU	400.000		
POLY U	400.000		
JEFLL			250.000
FRIDA	200.000		
LINDSEI		100.000	
MELD	100.000		
EVA			85.000
PELE	10.000		
ESFSLD		ni podatka	
ISLE		ni podatka	
IBLC	ni podatka		
Japanese Learners' Corpus	ni podatka	ni podatka	ni podatka
Skupaj	**60.740.000**	**1.600.000**	**835.000**

Tabela 2: *Število besed v korpusih usvajanja tujega jezika glede na prenosnik*.

[6] Dosegljivi so bili samo podatki o skupni velikosti korpusa, ne pa tudi o tem, koliko besed je v pisnem in koliko v govornem delu.

Tabela kaže, da večina korpusov obsega od pol do milijona besed. Očitno je to vsaj za angleščino kot tuji jezik tisti obseg, ki ni prezahteven za gradnjo, hkrati pa daje relevantne rezultate. Zato bi bila to tudi smiselna izhodiščna velikost korpusa usvajanja slovenščine kot tujega jezika; dejanska uporaba bo pokazala, ali niso razmerja pri tako fleksijskem jeziku drugačna in bi bila zato relevantnejša kaka druga velikost.

Z velikostjo korpusa je neposredno povezana dolžina posameznih sestavnih besedil. Učeči se v ciljnem jeziku težje producirajo tekste kot v prvem, zato so ti le redko daljši od tisoč besed, sicer pa največ korpusov vključuje besedila z okrog petsto besedami. Japonski *JEFLL* izstopa z minimalnim številom dvajset besed pri besedilih najmlajših tvorcev, starih 12 in 13 let. Predvidevamo lahko, da med dvesto ali petsto besedami ni velikih jezikovnih razlik, ki bi jih pogojevala dolžina sestavka.



Slika 2: *Velikost besedil v pisnih korpusih usvajanja tujega jezika*.

Korpusi usvajanja tujega jezika se tako kot drugi tipi korpusov ne morejo izogniti dilemi o statičnosti oziroma dinamičnosti. Ker se vmesni jezik učečih se ves čas spreminja, bi jih bilo smiselno oblikovati tako, da se nenehno dodajajo nova besedila (Gorjanc 2003: 24), s čimer korpus sledi živemu, rastočemu jeziku. Vendar ima dinamičnost svoje omejitve. Korpus zaradi velikosti sčasoma postane manj obvladljiv in neprimeren za primerjalne študije. Načelo skrbno načrtovanega izbora in uravnoteženosti zamenja oportunistično načelo količine. Ob tem ni nepomembno, da večji stroški, povezani z obdelavo večjega korpusa, povzročijo, da je dostopen manjšemu številu uporabnikov (Kennedy 1998: 61).

5 Vrsta besedil in tema

Vrste besedil in zvrsti so v korpusih usvajanja tujega jezika zelo različne. Največ je spisov ali esejev, pojavljajo se še pisma, dopisi, dnevniki, poročila, članki, govori, seminarske naloge in podobno.

Korpus	**Vrsta besedila**
ASK	besedila z izpitov
CEJL	spisi, obnove, elektronska pisma
CLC	besedila z izpitov
CLEC	besedila z izpitov, vodeni spisi, prosti spisi
FRIDA	opisni, argumentativni, pripovedna besedila, pisma, časopisni članki
HKUST	besedila z izpitov in besedila, nastala doma
IBLC	prijave na delovno mesto, poslovna komunikacija
ICLE	razpravljalni spisi
JEFLL	opisovalni in prepričevalni spisi
JPU	pripovedni in opisni eseji, razmišljanja, seminarske naloge
LCLE	odgovori na testih, pisma, poročila, dnevniki, spisi
PELCRA	besedila z izpitov
POLY U	poročila
Taiwanese Corpus	dnevniki, spisi
TELEC	osebna ali uradna pisma, uvodniki, članki, govori, spisi
TELC	besedila z izpitov, spisi
USE	spisi

Tabela 3: *Pregled vrst besedil v obstoječih pisnih korpusih*.[7]

Besedila, vključena v korpuse, so redko napisana doma, brez časovnih omejitev in z uporabo učnih pripomočkov in slovarjev. Taka besedila sicer preverjajo meje zmožnosti tvorčevega pisnega izražanja, vendar je dejanska komunikacija redko tako neomejena in brez stresa. Zato taki teksti sestavljajo le manjši delež majhnega števila korpusov, med drugim *LCLE*, *JPU*, *TELEC* in *CEJL*. Hongkonški *HKUST* poleg besedil z izpita iz angleščine vključuje besedila, ki so nastala izven razreda, kar omogoča primerjavo med teksti, ki so nastajali časovno omejeno in neomejeno, ter s tem boljše razumevanje jezikovne performance učečih se (Pravec 2002: 92). V korpusu *ICLE* je petindvajset odstotkov esejev o literaturi, ki so jih učenci pisali doma in pri tem uporabljali slovarje in druge pripomočke (Granger 2001: 5).

Veliko pogostejša so besedila z jezikovnih izpitov. Ker je gradnja korpusov ponavadi povezana z institucijami, ki izpite izvajajo, so tovrstna besedila lahko dostopna, dodatna prednost pa so kontrolirani pogoji tvorjenja in približno enaka stopnja jezikovne zmožnosti tvorcev. Izpiti, ki so vir besedil, so zelo različni. Del besedil za korpus *TELC* je pridobljen na univerzitetnih sprejemnih izpitih.[8] V *HKUST* so vključena besedila z izpita *Use of English Examination*, ki ga opravljajo učenci po končani srednji šoli in na podlagi katerega so uvrščeni v skupine pri pouku angleščine na univerzi (Pravec 2002: 86). Nastajajoči korpus grščine kot tujega jezika vključuje besedila, ki so jih v okviru testa na koncu intenzivnega tečaja napisali študentje, vključeni v študentske izmenjave.[9] V nekaterih primerih pa gre za različne zunanje, mednarodno veljavne izpite – *CLC* in *Pelcra* tako vključujeta

[7] Navedeni so le tisti korpusi, pri katerih je ta podatek dosegljiv.

[8] Podatek s spletne strani *http://iele.au.edu/corpus* julija 2004.

[9] Dimitrios Tzimokas, osebni vir M. S., april 2006.

besedila s Cambridgeovih izpitov *ESOL* (Pravec 2002: 91, 95). Natančni podatki o vseh korpusih žal niso dosegljivi.

Tematika sestavnih besedil je določena vnaprej in je zelo raznolika, pomembna pa je, ker vpliva na izbiro besedišča: splošne teme sprožajo uporabo drugačnega besedišča in slovničnih struktur kot bolj specifične, politični problemi se razlikujejo od osebnih izkušenj. Ob tem se je treba zavedati, da je s korpusom usvajanja nemogoče zajeti reprezentativen delež vseh polnopomenskih besed. Zaradi tega se raziskave bolj osredotočajo na slovnične besede in strukture, torej sama tematika niti ni tako ključna. Vendarle pa so za korpuse usvajanja bolj kot opisne, pripovedne, strokovne ali tehnične primerne argumentativne teme:[10] aktualni dogodki in problemi (*HKUST*, *ICLE*, *JPU*, *USE*), služba (*POLY U*), potovanja in konjički (*CEJL*), razpravljanje o prebranih literarnih delih (*ICLE*, *USE*), o odnosu do ciljnega jezika (*USE*) ali izobraževanja (*JPU*, *PELE*), obnove prebranega in videnega (*CEJL*).

6 Tvorci

Eden najpomembnejših pogojev za uravnoteženost korpusa usvajanja tujega jezika so tvorci besedil. Uporabni so samo podatki, zbrani z ustreznim pregledom nad učečimi se in okoliščinami nastanka besedil (Tono 2003: 801). Previdnost je posebno nujna, ker obstaja toliko učečih se in učnih situacij (Granger 1998: 7).

Notranja merila tvorcev, kot so motiviranost ali stališča do ciljnega jezika, je težko nadzirati, več pozornosti pa se namenja uravnoteženosti in konsistentnosti zunanjih dejavnikov, kot so starost, spol, izobrazba, nacionalnost, prvi jezik, učno okolje ciljnega jezika[11] in jezikovne zmožnosti v ciljnem jeziku. Pri slednji se upošteva dosežek učečega se na standardiziranih testih, če to ni mogoče, pa manj zanesljiva zunanja merila, kot je skupina na tečaju, v katero je uvrščen učeči se, ali učbenik, ki ga je nazadnje predelal. Pomembni merili sta tudi znanje drugih jezikov ter praktične izkušnje, ki so povezane z bivanjem učečega se v državah, kjer je ciljni jezik prvi jezik (Granger 1998: 9).

Tvorci se večinoma še šolajo, v glavnem gre za univerzitetne študente ali srednješolce: v korpusu *HKUST* so to dijaki, v *CLEC* in tajvanskem korpusu dijaki in univerzitetni študentje, v *ICLE*, *JPU*, *MELD*, *TELC* in *USE* samo študentje, *JEFLL* in *PELCRA* pa zajemata vse stopnje šolanja (Axelsson 2000: 155, Granger 2001: 2, 3, Pravec 2002: 86, 88, 90, 95, Tenfjord et al. 2004 in Uzar 1998). Razlogi za tako usmerjenost na tvorce, ki se še šolajo, so seveda oportunistični. Čeprav ni nujno, da šolajoči se pišejo več kot tisti, ki so šolanje že končali, graditelji korpusov in raziskovalci, ki so pogosto učitelji ali univerzitetni profesorji, lažje pridejo do njihovih besedil.

---

[10] Tipično neprimerna naslova za besedilo, vključeno v korpus, sta *Radosti angleškega podeželja* ali *Moje leto v Ameriki* (Granger 2001: 6). Manj primerne so tudi »sezonske« teme in besedišče, recimo prazniki ali doživljanje letnih časov.

[11] Npr. ali ciljni jezik ni prvi jezik učnega okolja, pomembna pa je tudi stopnja šolanja, na kateri poteka usvajanje.

Korpus	Podatki o tvorcih sestavnih besedil
CLC	narodnost, prvi jezik, stopnja jezikovne zmožnosti, starost
JEFLL	razred in vrsta šole, ki jo obiskuje tvorec
JPU	spol, leto, tečaj angleščine, ki ga obiskuje tvorec
LCLE	narodnost, stopnja jezikovne zmožnosti, država, kjer tvorec živi
MELD	starost, spol, jezikovno in izobrazbeno ozadje
PELCRA	starost, spol, znanje jezikov, obiski v angleško govorečih državah
USE	starost, spol, prvi jezik, izobrazba, čas, ki ga je tvorec preživel v angleško govorečem okolju

Tabela 4: *Pregled podatkov o tvorcih sestavnih besedil za posamezne korpuse*.[12]

Le redko so vključena besedila začetnikov, saj ti tvorijo kratka besedila z mnogo odkloni od norme, popolnoma tujimi strukturami in malo koherentne vsebine. Precej primernejši tvorci so učeči se na nadaljevalni ali izpopolnjevalni stopnji jezikovne zmožnosti.

Korpus	Jezikovna zmožnost v ciljnem jeziku[13]
ASK	B1 in B2 (nadaljevalna)
FRIDA	nadaljevalna
HKUST	nadaljevalna
ICLE	izpopolnjevalna
JEFLL	vse stopnje
LCLE	8 stopenj
MELD	izpopolnjevalna
PELCRA	vse stopnje
Taiwanese Corpus	nadaljevalna ali izpopolnjevalna

Tabela 5: *Pregled podatkov o stopnji jezikovne zmožnosti v pisnih korpusih*.

Podatki o tvorcih so pridobljeni s posebnim vprašalnikom ali s prijavnic na tečaje ali izpite. Spremlja jih privolitveni obrazec, s katerim tvorec dovoljuje uporabo besedil v raziskovalne namene. V skladu z varovanjem zasebnosti je treba zagotoviti anonimnost, zato so podatki v korpus vključeni brez imen.

7 Kontrolni korpus

Uporabnost korpusa usvajanja tujega jezika se poveča z vzporednim kontrolnim korpusom jezika rojenih govorcev, ki omogoča primerjavo z jezikom tujih govorcev. Tako dobimo kvantitativne podatke o pogostnosti določenih besed, besednih vrst, skladenjskih struktur in o značilnostih diskurza (Tono 2003: 800). V ta namen je mogoče uporabljati že obstoječe korpuse ali njihove dele ali pa zgraditi poseben podkorpus. V obeh primerih morajo biti načela gradnje podobna.

[12] Navedeni so le tisti korpusi, pri katerih je ta podatek dosegljiv.

[13] Podatki niso absolutno primerljivi, ker so pridobljeni iz različnih virov in je nemogoče ugotoviti, kakšna merila so uporabljali, vendarle pa nudijo vsaj približen vpogled v stanje.

Med 21 pisnimi korpusi usvajanja tujega jezika jih ima le šest kontrolni korpus. *ICLE* spremlja *Louvain Corpus of Native English Essays*, to je korpus esejev rojenih govorcev angleščine, ki omogoča primerjave med podobnim gradivom, ki so ga ustvarili rojeni in tuji govorci (Granger 2001: 3). Norveški *ASK* vsebuje popravljene različice besedil tujih tvorcev, s katerimi bi bilo mogoče razvozlati smisel besedil na uporabniku prijazen način, hkrati pa so primerne za pedagoške namene (Tenfjord et al. 2004). Poljskemu korpusu *PELE* je dodan vzorec besedil rojenih govorcev angleščine in besedil poljščine kot prvega jezika, ki obsega tri tisoč besed (Uzar 1998). Japonski korpus *SST* vsebuje podkorpus rojenih govorcev angleščine, ki so odgovarjali na podobna vprašanja kot Japonci, in podkorpus angleških izjav Japoncev, prevedenih nazaj v japonščino, kjer so opazne posledice jezikovnega prenosa (Izumi et al. 2004: 37). Podobno razdelano kontrolno gradivo spremlja japonski korpus *JEFLL*: japonsko pisno gradivo naj bi v prihodnosti omogočilo prepoznavanje napak v angleščini, ki so posledica jezikovnega prenosa iz japonščine, dodan pa je še korpus angleščine z 39.000 besedami. Sestavljajo ga učbeniki angleščine, ki jih uporabljajo v japonskih šolah in ki naj bi predstavljali normo, h kateri težijo njihovi učenci. Taka na prvi pogled nekoliko nenavadna norma naj bi bila za Japonce primernejši cilj kot npr. angleščina, ki jo ponuja *BNC* (Pravec 2002: 95).

8 Uporabnost

Ključno vprašanje pri oblikovanju vsakega korpusa je njegov namen. Zakaj torej potrebujemo korpuse usvajanja tujega jezika? Uporaba je usmerjena tako v smer teoretičnih raziskav kot v smer praktičnih aplikacij.

Težišče teoretičnega raziskovanja, vezanega na korpuse usvajanja tujega jezika, je primerjalna analiza, pri kateri primerjamo jezik rojenih z jezikom tujih govorcev ali različne vmesne jezike med seboj.[14] Tako je mogoče odkriti značilnosti in napake, skupne vsem učečim se, ter tiste, ki so omejene na govorce z določenim jezikovnim ozadjem. Ocenimo lahko, kako na tvorjenje vplivajo posamezne zunajjezikovne spremenljivke, kot so jezikovno ozadje, spol ali starost.

Aplikativno vrednost korpusov usvajanja jezika potrjujejo vedno številnejše izboljšave pedagoškega procesa, ki temeljijo na korpusih in vplivajo na oblikovanje treh ravni poučevanja: na učne načrte, učna gradiva in metodologijo.

V najširšem smislu korpusi spreminjajo učni načrt v celoti, torej vsebino poučevanja, vrstni red in poudarek, ki ga učitelj da določeni vsebini. Vendar je njihovo vlogo zaenkrat težko ločiti od ostalih dejavnikov. Veliko pomembnejši je vpliv korpusov na učna gradiva. Pri razvoju slovaropisja imajo že od nekdaj vodilno vlogo specializirani slovarji za učeče se, saj prvi beležijo nove rabe ali upoštevajo spreminjanje slovničnih struktur.[15] Poleg tega so prvi kot vir podatkov pričeli uporabljati korpuse. Danes

[14] Na tem področju je bila za slovenščino pionirska raziskava Nataše Pirih Svetina (prim. Pirih Svetina 2005).

je izdelava slovarjev pa tudi slovnic kompromis med podatki rojenih govorcev iz nespecializiranih korpusov, ki pokažejo tipično v ciljnem jeziku, in podatki iz korpusov usvajanja, ki povedo, katere težave so tipične za učeče se (Granger 1998: 7). Najpomembnejše značilnosti slovarjev za učeče se so tako: »uporaba omejenega besedišča v definicijah, navajanje slovničnih informacij o iztočnici, vključno s tipičnimi vezljivostnimi vzorci, načrtno izpostavljanje besednozveznih enot, od pogostih kolokacij do idiomov, radodarno vključevanje zgledov rabe, pazljivo navajanje informacij o stilni ravni iztočnice ter druge morebitne omejitve pri rabi« (Krek 2004: 4). Za večjo uporabnost so ponekod vključena še opozorila o najpogostejših napakah.

Britanski korpus *LCLE* je bil osnova za izdelavo dveh slovarjev. Pri izdelavi *Longman Active Study Dictionary* so podatki pokazali, da učenci vedno narobe uporabljajo besedo *krpe* (angl. *cloths*), ker jo zamenjujejo z besedo *obleke* (angl. *clothes*), zato so v slovar dodali posebno opozorilo. V slovarju *Longman Essential Activator*, ki vodi učence na nadaljevalni stopnji do tiste besede, ki jo potrebujejo v določenem kontekstu, pa so napačne rabe izpostavljene skupaj s predlogi, kako jih nadomestiti (Pravec 2002: 89–90).

V učbenike so na podlagi korpusov vključene najpogostejše fraze in besede namesto neživljenjskih primerov, ki so tuji učencem in nenaravni učiteljem. Korpusi so primerna osnova za vaje prepoznavanja in odpravljanja napak (Pravec 2002: 108), pa tudi za programska orodja za učenje tujega jezika. Program AutoLANG na osnovi korpusa *HKUST* omogoča interaktivno učenje prek popravljanja napak, WordPilot pa s kolokacijami pomaga širiti besedišče. TeleNex, ki temelji na korpusu *TELEC*, je spletna podatkovna baza za pomoč učiteljem angleščine, ki jim nudi slovnico ter zbirko ocenjenih učnih gradiv.

Na področju učne metodologije je precej razširjena in teoretično razčlenjena uporaba konkordanc v razredu. Učenci s seznamom konkordanc raziskujejo vzorce ciljnega jezika, pri tem pa vidijo vsak leksem ali besedno zvezo v kontekstu, kar zmanjša napačno rabo. Učijo se kolokacij in se izognejo jezikovnemu prenosu (Kennedy 1998: 288–89). V tovrstne namene so uporabni tako korpusi usvajanja tujega jezika kot tudi veliki splošni korpusi, ki so posebej primerni pri manj pogostih besedah. Ob tem ne gre zanemariti uporabnosti korpusov za računalniško razgledane učitelje, saj jim nudijo koristne informacije o jezikovnem sistemu in jih spodbujajo, da se bolj kot na lastni jezikovni čut zanašajo na dejanske jezikovne podatke (Kennedy 1998: 264).

S pregledom obstoječih korpusov usvajanja tujega jezika lahko ugotovimo, da je graditeljem pomembna pedagoška uporabnost, predvsem diagnosticiranje in odpravljanje najpogostejših napak in težav tujih govorcev. To velja za švedski korpus *USE* (Axelsson 2000: 155), poljski *PELCRA*, hongkonška *HKUST* in *TELEC*. Francoski korpus *FRIDA* želi na podlagi najpogostejših napak dolgoročno

[15] Vodilni na tem področju so slovarji iz serije Cobuild (Krek 2004).

razviti avtomatično popravljanje napak. Poljski korpus *Poly U* naj bi pomagal diplomirancem izboljšati njihovo pisanje (Lin 1999), podobno tudi poljski *PELE*, ki naj bi hkrati povečeval zanimanje za korpusno jezikoslovje. *PELE* se je že uporabljal tudi kot osnova za naloge, uporabljene pri pouku (Uzar 1998). Nekaj korpusov navaja tudi bolj teoretične cilje, med njimi *CLC*, *LCLE*, *Taiwanese Corpus*, *JPU* in *ICLE* (Kennedy 1998: 42, Shih 2000: 92, 96, Pravec 2002: 84–85 in Dagneaux et al. 2001).

Ob tem se postavlja vprašanje, kdo uporablja korpuse usvajanja tujega jezika. Čeprav to v literaturi ni posebej izpostavljeno, je jasno, da gre predvsem za profesionalne uporabnike, torej znanstvenike, sestavljavce učnih gradiv in učitelje, manj pa za tiste, ki se jezik šele učijo. Kot je razvidno, so korpusi ali popolnoma zaprti (*CLC*, *PELCRA*, *JPU*, *TELEC*) ali pa na voljo samo za akademske raziskave (*LCLE*, *ISLE*, *USE*, *ESFSLD*, *HKUST*). Le do redkih lahko brezplačno ali komercialno dostopa vsakdo (*ICLE*, *Taiwanese Corpus*). Obstaja celo mnenje, da učeči se sploh ne bi smeli dostopati do korpusnih podatkov, saj bi se s tem zmanjšala učiteljeva jezikoslovna avtoriteta (Aston 1997). Toda smisel jezikovnega pouka gotovo ni prikazati učitelja kot vsevednega, nezmotljivega boga. Učenje z odkrivanjem, kot je brskanje po seznamih konkordanc, učence dodatno motivira (Granger in Tribble 1998: 200). Seveda je treba paziti, da uporabljajo namenu ustrezne vire, iščejo na funkcionalen način in podatke pravilno interpretirajo.

9 Načrtovanje korpusa usvajanja slovenščine kot tujega jezika

Pregledana merila so ključna za načrtovanje korpusa usvajanja slovenščine kot tujega jezika. Rešitve iz že obstoječih korpusov so lahko koristno izhodišče, vendar jih zaradi specifičnosti situacije vsakega jezika ni mogoče avtomatsko prenesti na slovenski korpus.

Tako je že pri izhodiščnem jeziku. Večina korpusov, ki so narejeni za angleščino, ima samo en izhodiščni jezik, toda korpus usvajanja slovenščine bo smiseln in širše uporaben samo, če bo imel več izhodiščnih jezikov. Te bo treba poiskati med najštevilčnejšimi skupinami govorcev, za katere slovenščina ni prvi jezik, kar pa presega obseg in namen tega prispevka.

Večina korpusov usvajanja je pisnih in tak bo tudi korpus usvajanja slovenščine vsaj do takrat, ko bodo za slovenščino uveljavljena razmeroma enotna načela za oblikovanje in označevanje govornih korpusov. Sestavna besedila bo tako kot drugod po svetu treba pretipkati.

Velikost se pri obstoječih korpusih usvajanja v glavnem giblje okoli milijona besed. To bo tudi izhodišče za gradnjo slovenskega korpusa, čeprav bo končna velikost odvisna od rezultatov vmesnih analiz korpusnih podatkov in tudi od materialnih sredstev, ki bodo na voljo za gradnjo. Sestavna besedila bodo imela od dvesto do petsto besed. Napisana bodo na državno veljavnih izpitih iz znanja slovenščine, kjer

so pogoji nastanka natančno nadzorovani, in na tečajih slovenščine, določen delež pa bodo predstavljala tudi besedila, ki jih bodo učeči se napisali doma. Tematika bo splošna, vendar bodo teme tako kot v drugih korpusih usmerjene v argumentacijo.

Tvorci sestavnih besedil bodo najverjetneje udeleženci jezikovnih tečajev in izpitov, ker je veliko težje dostopati do tistih, ki se slovenščine kot tujega jezika ne učijo institucionalizirano. Njihova jezikovna zmožnost bo na nadaljevalni ali izpopolnjevalni ravni, saj začetniki tvorijo zelo kratka, manj koherentna in kohezivna besedila z veliko napakami (prim. Stritar 2005). O njih bodo znani osnovni demografski podatki, kot so spol, prvi jezik, država, v kateri živijo, izobrazba, poklic, koliko časa se učijo slovensko in podobno. Tvorci bodo podpisali privolitveni obrazec, v korpusu pa bodo anonimni in označeni le s šifro.

Nekateri korpusi usvajanja imajo poseben kontrolni korpus z besedili rojenih govorcev. Ker njihova gradnja zahteva veliko časa, bo korpus usvajanja slovenščine brez kontrolnega korpusa, primerjave z jezikom rojenih govorcev pa bodo lahko potekale na podlagi rezultatov iz referenčnega korpusa *FIDA*. Poleg tovrstnih in drugih teoretičnih raziskav bo korpus uporaben tudi za praktične namene, npr. pri izdelavi učbenikov ali slovarjev, seveda pa bosta njegova uporabnost in dostopnost javnosti tako kot še marsikatera druga značilnost precej odvisni od materialnega ozadja projekta.

10 Zaključek

Korpusi usvajanja tujega jezika vedno bolj postajajo nujno potrebna realnost vseh, ki se ukvarjajo s tem področjem. Zato sta smiselna načrtovanje in seveda tudi gradnja korpusa usvajanja slovenščine kot tujega jezika. Prvi učni poskusi so bili že narejeni,[16] vendar pa naraščajoče število tujcev, ki se učijo slovensko, kaže na nujnost izdelave sodobnega, kvantitativno utemeljenega in čim bolj uravnoteženega vira za slovenščino kot tuji jezik.

Literatura

Aston, Guy, 1997: *Small and Large Corpora in Language Learning.* <http://home.sslmit.unibo.it/~guy/wudj1.htm>.

Axelsson Westergreen, Margareta, 2000: USE – The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal* št. 24 (april 2000). 155–157. <http://helmer.aksis.uib.no/icame/ij24/use.pdf>.

Dagneaux, Estelle, Granger, Sylviane, Meunier, Fanny, Petch-Tyson, Stephanie in Vilret, Xavier, 2001: *A web interface to the International Corpus of Learner English*. <http://juppiter.fltr.ucl.ac.be/fltr/germ/etan/cecl/events/icamepr.htm#interface>.

[16] Prim. učni korpus s 600 besedami v Stritar 2005.

Gorjanc, Vojko, 2003: Korpusi in jezikoslovje. *Jezik in slovstvo* 48/3–4 (maj–avgust 2003). 19–27.

Gorjanc, Vojko, 2005: *Uvod v korpusno jezikoslovje*. Ljubljana: Izolit (Zrenja).

Granger, Sylviane, 1998: The computer learner corpus: a versatile new source of data for SLA research. Granger, Sylviane (ur.): *Learner English on Computer*. London, New York: Longman (Studies in Language and Linguistics). 3–18.

Granger, Sylviane in Tribble, Chris, 1998: Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. Granger, Sylviane (ur.): *Learner English on Computer*. London, New York: Longman (Studies in Language and Linguistics). 199–209.

Granger, Sylviane, 2001: *International Corpus of Learner English: The ICLE Project*. <http://juppiter.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl-projects/icle/download/icle.pdf>.

Izumi, Emi, Uchimoto, Kiyotaka in Isahara, Hitoshi, 2004: SST speech corpus of Japanese Learners' English and automatic detection of learners' errors. *ICAME Journal* št. 28 (april 2004). 31–48. <http://helmer.hit.uib.no/icame/ij28/Izumi.pdf>.

Kennedy, Graeme, 1998: *An Introduction to Corpus Linguistics*. London, New York: Longman (Studies in language and linguistics).

Krek, Simon, 2004: Slovarji serije COBUILD in formalizacija definicijskega jezika. *Jezik in slovstvo* 49/2 (marec–april 2004). 3–16.

Leech, Geoffrey, 1997: Introducing Corpus Annotation. Garside, Roger, Leech, Geoffrey in McEnery, Anthony (ur.): *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London, New York: Longman. 1–18.

Lin, Linda H. F., 1999: *Applying Information Technology to a corpus of student report writing to help students write better reports*. <http://elc.polyu.edu.hk/conference/papers/Lin.htm>.

Pirih Svetina, Nataša, 2005: *Slovenščina kot tuji jezik*. Ljubljana: Izolit (Zrenja).

Pravec, Norma, 2002: Survey of learner corpora. *ICAME Journal* št. 26 (april 2002). 81–114. <http://nora.hd.uib.no/icame/ij26/pravec.pdf>.

Shih, Rebecca Hsue-Hueh, 2000: Compiling Taiwanese Learner Corpus of English. *Computational Linguistic and Chinese Language Processing* 5/2 (avgust 2000). 89–102. <http://rocling.iis.sinica.edu.tw/clclp/vol5-2/paper4.pdf>.

Stabej, Marko, 2004: Slovenščina kot drugi/tuji jezik in slovensko jezikovno načrtovanje. *Jezik in slovstvo* 49/3–4 (maj–avg. 2004). 5–16.

Stritar, Mojca, 2005: *Označevanje korpusa usvajanja tujega jezika*. Seminarska naloga. Ljubljana: Filozofska fakulteta.

Tenfjord, Kari, Meurer, Paul in Hofland, Knut, 2004: *The ASK corpus – a language learner corpus of Norweigan as a second language (Poster).* <http://www.ugr.es/~talc6/talc_search/proceedings/60.html>.

Tono, Yukio, 2003: Learner corpora: design, development and applications. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster: University. 800–809.

Uzar, Rafal S., 1998: *The PELE Project: A New Perspective for Learner Language Corpora.* <http://members.fortunecity.com/pelcra/pele.htm>.

Zemljarič Miklavčič, Jana, 2004: Taksonomija besedilnih tipov za gradnjo govorjenega korpusa. Kržišnik, Erika (ur.): *Aktualizacija jezikovnozvrstne teorije na Slovenskem.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete (zbornik *Obdobja – metode in zvrsti*, 22). 503–522.

Seznam korpusov (kratice in spletni viri)

ASK: *Norsk andrespråkskorpus*. <http://www.decentius.aksis.uib.no/corpus/askdemo-home.xml> (januar 2007).

BNC: *British National Corpus*. <http://www.natcorp.ox.ac.uk/> (avgust 2006).

CLC: *Cambridge Learner Corpus*. <http://www.cambridge.org/elt/corpus/learner_corpus.htm> (avgust 2006).

CLEC: *Chinese Learner English Corpus*. <http://langbank.engl.polyu.edu.hk/corpus/clec.html> (avgust 2006).

CEDEL2: *Corpus Escrito del Español* L2. <http://www.uam.es/proyectosinv/woslac/cedel2.htm> (avgust 2006).

CEJL: *Corpus of English by Japanese Learners*. <http://www.eng.ritsumei.ac.jp/lcorpus/> (junij 2004).

ESFSLD: *European Science Foundation Second Language Databank*. <http://www.mpi.nl/world/isle/overview/overview_esfsld.html> (avgust 2006).

EVA: <http://kh.hd.uib.no/eva/> (julij 2004).

FRIDA: *French Interlanguage Database*. <http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl-projects/frida/gateway.htm> (avgust 2006).

HKCCE: *Hong Kong Corpus of Conversational English*.

HKUST: *Hong Kong University of Science and Technology Corpus*.

IBLC: *Indianapolis Business Learner Corpus*. <http://www.iupui.edu/~icic/corpusother.htm> (avgust 2006).

ICLE: *International Corpus of Learner English*. <http://juppiter.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl-projects/icle/icle.htm> (avgust 2006).

PICLE: *Polish International Corpus of Learner English* (poljski del korpusa *ICLE*). <http://www.staff.amu.edu.pl/~przemka/picle.html> (avgust 2006).

ISLE: *Interactive Spoken Language Education Corpus of non-native spoken English*. <http://nats-www.informatik.uni-hamburg.de/~isle/speech.html> (avgust 2006).

Japanese Learners' Corpus.

JEFLL: *Japanese English as a Foreign Language Learner Corpus*. <http://leo.meikai.ac.jp/~tono/jefll.html> (avgust 2006).

JPU: *Janus Pannonius University Corpus*. <http://www.geocities.com/jpu_corpus> (avgust 2006).

LCLE: *Longman Corpus of Learners' English*. <http://www.longman.com/dictionaries/corpus/learners.html> (avgust 2006).

Learner Business Letter Corpus. <http://isweb9.infoseek.co.jp/school/ysomeya/> (avgust 2006).[17]

LINDSEI: *Louvain International Database of Spoken English Interlanguage*. <http://juppiter.fltr.ucl.ac.be/fltr/germ/etan/cecl/cecl-projects/lindsei/lindsei.htm> (avgust 2006).

MELD: *Montclair Electronic Language Database*.

PELE: *Polish English Learner English*.

PELCRA: *Polish Learner English Corpus*. <http://korpus.ia.uni.lodz.pl/> (avgust 2006).

Poly U Corpus.

SST: *Standard Speaking Test Corpus*. <http://leo.meikai.ac.jp/~tono/sst/> (avgust 2006).

Taiwanese Corpus of Learner English.

TELC: *Thai English Learner Corpus*.

TELEC: TELEC *Secondary Learner Corpus*.

USE: *Uppsala Student English Corpus*. <http://hem.passagen.se/ylvaberg/useinfo1.htm> (avgust 2006).

[17] Zaradi premalo podatkov v članku ni bil obravnavan.