Oblikovanje korpusa usvajanja slovenščine kot tujega jezika

Mojca Stritar*

* Center za slovenščino kot drugi/tuji jezik, Filozofska fakulteta, Univerza v Ljubljani
Kongresni trg 12, 1000 Ljubljana
mojca.stritar@ff.uni-lj.si

Povzetek

Prispevek podaja razmislek o temeljnih vprašanjih načrtovanja korpusa usvajanja slovenščine kot tujega jezika, povezanih z jezikom, prenosnikom, velikostjo, vrsto in tematiko besedil, tvorci ter kontrolnim korpusom. Ustavi se ob različnih vidikih označevanja napak, predlagane pa so tudi rešitve za pilotski korpus usvajanja slovenščine kot tujega jezika.

Slovene learner corpus design

In the article different problems considering the design of Slovene learner corpus are being discussed, such as the language, written and spoken corpora, size, text types, subjects, authors and control corpora. Various aspects of error mark-up are analyzed and finally the basic design of a pilot Slovene learner corpus is being proposed.

1. Uvod

Zaradi svoje teoretične in praktične uporabnosti vedno nujnejši del jezikovnega načrtovanja postajajo korpusi usvajanja tujega jezika, ki predstavljajo jezik, kot ga pišejo ali govorijo tisti, ki niso njegovi rojeni govorci. V pričujočem prispevku bodo pregledane in ovrednotene nekatere rešitve že obstoječih tujih tovrstnih korpusov, hkrati pa bo podan predlog za oblikovanje korpusa usvajanja slovenščine kot tujega[1] jezika (v nadaljevanju KUST).

Pregledala sem dostopne podatke o 26 obstoječih korpusih, sicer pa ves čas nastajajo še novi. Kar 23 od teh korpusov je bilo narejenih s ciljnim jezikom angleščino, katere jezikoslovna in sociolingvistična realnost sta temeljno drugačni od slovenske. Zato so neposredne vzporednice med obstoječimi in bodočim slovenskim korpusom usvajanja nemogoče in je treba njihove rešitve upoštevati kot izhodišče, ne pa kot nujen zgled.

Ključno vprašanje pri snovanju in oblikovanju vsakega korpusa je njegov namen. Korpusi usvajanja tujega jezika so uporabni tako za teoretične raziskave procesa usvajanja in opisovanje vmesnega jezika učečih se kot tudi za različne praktične aplikacije, kot so slovarji, slovnice, učbeniki ali programska orodja. Danes 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). Pri izdelavi učbenikov so korpusi vir za realne primere in nudijo gradivo za vaje prepoznavanja in odpravljanja napak (Pravec, 2002). Pregled obstoječih korpusov usvajanja pokaže, da je večinoma pomembna pedagoška uporabnost, predvsem diagnosticiranje ter odpravljanje najpogostejših napak in težav tujih govorcev (Axelsson, 2000; Lin 1999; Uzar 1998), medtem ko nekaj korpusov poleg tega deklarira tudi bolj teoretične cilje (Kennedy, 1998; Shih, 2000; Pravec, 2002; Dagneaux et al. 2001). Ker se raziskovalci slovenščine kot tujega jezika s tem praviloma ukvarjajo tako na abstraktni, teoretični kot tudi na praktični ravni, bi moral KUST nuditi relevantne in uporabniku prijazne informacije za raziskovalne in praktično-aplikativne namene.

2. Jezik

Pri korpusih usvajanja tujega jezika sta pomembna ciljni jezik, torej jezik, "ki se ga nekdo uči z namenom, da bi ga obvladal bodisi kot svoj prvi, drugi ali tuji jezik" (Pirih Svetina 2005), in izhodiščni oziroma prvi jezik, "iz katerega se nekdo uči vse druge ali tuje jezike" (navedeno delo). Velika večina obstoječih korpusov je usmerjena iz enega izhodiščnega jezika v en ciljni jezik, praviloma angleščino (Atwell et al., 2003; Axelsson, 2000; Cheng, Warren, 1999; De Cock et al., 1999; Granger, 2001; Horváth, 2003; Izumi et al., 2004; Kennedy, 1998; Lin, 1999; Pravec, 2002; Shih, 2000; Sugiura, 2000; Tenfjord et al., 2004; Tono, 2003; Uzar, 1998). Tujih govorcev slovenščine ne moremo omejiti na skupino z enim samim prvim jezikom, zato bi se bilo pri KUST-u bolje zgledovati po korpusih manj globalno razširjenih jezikov. Taka sta ASK, korpus usvajanja norveščine kot tujega jezika, ali FRIDA, korpus francoščine. V obeh je izhodiščnih jezikov več in predstavljajo največje skupine tujih govorcev oziroma priseljencev.

3. Prenosnik

Za referenčne korpuse je vprašanje prenosnika bistveno, korpusi usvajanja pa nikoli ne morejo biti reprezentativni za celotno populacijo tujih govorcev z vsemi prvimi jeziki in stopnjami znanja. Zato referenčnost ostaja utopična želja, graditelji pa zaenkrat v glavnem delajo pisne korpuse.

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.

	Pisni korpusi	**Govorni korpusi**	**Korpusi s pisnim in govornim delom**
Število besed	60.640.000	1.600.000	835.000

Tabela 1: Število besed v korpusih usvajanja tujega jezika glede na prenosnik.

[1] *Tuji jezik* se učimo "v okolju, kjer ta jezik običajno ni v uporabi" (Pirih Svetina 2005), *drugi jezik* pa je nematerni jezik, ki se v govorčevi skupnosti redno uporablja, torej prevladujoči jezik okolja, v katerem živi. Zaradi elegantnejšega poimenovanja uporabljam izraz *korpus usvajanja slovenščine kot tujega jezika* kot krovni pojem za tuji in drugi jezik.

Razlogi za tako velik delež pisnih korpusov so praktični – zajem besedil je že pri pisnih tekstih relativno zamuden, saj jih je treba pretipkati, transkripcija govorjenih tekstov pa bi bila še počasnejša. Za slovenščino še ni referenčnega govornega korpusa, ki bi reševal načelna vprašanja tega tipa, zato bi bil tudi KUST na začetku izključno pisni.

4. Velikost

Prav zaradi pretipkavanja posameznih sestavnih besedil se korpusi usvajanja po velikosti težko primerjajo z nespecializiranimi korpusi.

Korpus	Ciljni jezik	Število besed
HKUST	angleščina	25.000.000
CLC	angleščina	15.000.000
LCLE	angleščina	10.000.000
TELEC	angleščina	3.000.000
ICLE	angleščina	2.000.000
TELC	angleščina	1.300.000
SST	angleščina	1.000.000
USE	angleščina	1.000.000
CEJL	angleščina	1.000.000
CLEC	angleščina	1.000.000
Taiwanese Corpus	angleščina	730.000
HKCCE	angleščina	500.000
PELCRA	angleščina	500.000
ASK	norveščina	500.000
JPU	angleščina	400.000
POLY U	angleščina	400.000
JEFLL	angleščina	250.000
FRIDA	francoščina	200.000
LINDSEI	angleščina	100.000
MELD	angleščina	100.000
EVA	angleščina	85.000
PELE	angleščina	10.000

Tabela 2: Velikost korpusov usvajanja tujega jezika (Atwell et al., 2003; Axelsson, 2000; Cheng, Warren, 1999; De Cock et al., 1999; Granger, 2001; Horváth, 2003; Izumi et al., 2004; Kennedy, 1998; Lin, 1999; Pravec, 2002; Shih, 2000; Sugiura, 2000; Tenfjord et al., 2004; Tono, 2003; Uzar, 1998).

Zgornja tabela kaže, da je glavnina korpusov velika med pol milijona in milijonom besed. Očitno je to vsaj za angleščino srednja mera med zahtevnostjo gradnje in relevantnostjo rezultatov. Zato bi bilo to tudi smiselno izhodišče za končno verzijo KUST-a; verjetno bo že uporaba pilotske verzije pokazala, ali ne bi bila za tako fleksijski jezik, kot je slovenščina, relevantnejša kaka druga velikost.

S tem je neposredno povezana velikost sestavnih besedil, ki nikjer ne presegajo tisoč besed, sicer pa ima največ korpusov tekste z okrog petsto besedami. Japonski JEFLL izstopa z minimumom dvajset besed pri besedilih najmlajših tvorcev, starih 12 in 13 let (Pravec, 2002). Besedila z izpitov iz znanja slovenščine, ki so možen vir za KUST, imajo po dvesto besed, in tudi spisi s tečajev po izkušnjah učiteljev redko presegajo eno pisano stran. Za polmilijonski korpus bi torej potrebovali 2500 tekstov s po 200 besedami.

5. Vrsta besedil in tema

Vrste besedil in funkcijske zvrsti so v pisnih korpusih usvajanja tujega jezika zelo raznolike. Največ je spisov in esejev, pojavljajo se še pisma, dopisi, dnevniki, poročila, članki, govori, seminarske naloge in podobno. Mnogi korpusi, npr. tajski TELC, hongkonški HKUST, britanski CLC in poljska PELCRA, vključujejo tekste z jezikovnih izpitov, predvsem zaradi enostavne dosegljivosti v večjih količinah, kontroliranih pogojev tvorjenja in približno enake stopnje jezikovne zmožnosti tvorcev. Vendar se spisi, nastali v razredu, besedila z izpitov in naloge, napisane doma, razlikujejo po spodbudi pri nastanku besedil (ali so nastala spontano ali s predhodno pripravo), uporabi referenc (ali so tvorci uporabljali slovar, učne pripomočke ali se sklicevali na že napisane tekste), ter časovni omejenosti pri pisanju. Kaj verodostojneje izraža vmesni jezik? Po eni strani stres na izpitih zmanjšuje jezikovno performanco, po drugi pa smo v dejanski komunikaciji le redko brez časovnih in drugih omejitev, saj npr. sogovorniki niso pripravljeni poljubno čakati na odgovor. Zato samo redki korpusi, HKUST, britanski LCLE, hongkonški TELEC, japonski CEJL, madžarski JPU in mednarodni ICLE (Pravec, 2002; Granger, 2001), vključujejo besedila, napisana doma brez časovnih omejitev, in znotraj korpusov ti predstavljajo manjši delež besedil. Vsi ostali po dostopnih podatkih sodeč vključujejo samo tekste z različnih izpitov in testov. Tovrstni pogoji nastanka so očitno primernejši za vključevanje, zato bi bilo to smiselno upoštevati tudi v KUST-u.

Tematika sestavnih besedil je tvorcem dana vnaprej in zelo pestra, pomembna pa je, ker vpliva na izbiro besedišča. Splošne teme sprožajo uporabo drugačnega besedišča in slovničnih struktur kot bolj specifične. Ker pa še noben korpus usvajanja ni zajel reprezentativnega deleža vseh polnopomenskih besed, se raziskave bolj osredotočajo na slovnične besede in strukture, torej sama tematika ni tako ključna. Vendarle so primernejše argumentativne teme kot opisne, pripovedne, strokovne ali tehnične: priljubljeni so aktualni dogodki in družbeni problemi, služba, potovanja in konjički, razpravljanje o odnosu do ciljnega jezika ali izobraževanja, obnove prebranega in videnega.

Na rezultate vpliva tudi način zbiranja besedil. Lahko so zbrana longitudinalno, z več prispevki istega tvorca iz različnih časovnih obdobij, ali presečno, s hkratnim zajemom besedil več tvorcev. Zaradi relativne mladosti korpusov usvajanja tujega jezika in enostavnosti izvedbe so vsi pregledani korpusi presečni, tak pa bo tudi KUST.

6. Tvorci

Pogoj za uravnoteženost korpusa so tvorci besedil. Njihove notranje kriterije, kot je npr. motiviranost, je težko nadzorovati, graditelji pa pazijo na uravnoteženost in konsistentnost zunanjih dejavnikov: starosti, spola, izobrazbe, prvega jezika, učnega okolja ciljnega jezika in stopnje jezikovne zmožnosti v ciljnem jeziku. Pomembna sta tudi znanje ostalih tujih jezikov ter praktične izkušnje, ki so povezane z bivanjem učečega se v državah, kjer je ciljni jezik prvi jezik (Granger, 1998).

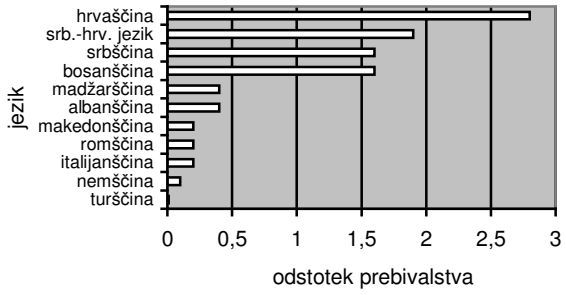
Tvorci v obstoječih korpusih so večinoma še sredi šolanja, v glavnem univerzitetni študentje ali srednješolci (Axelsson, 2000; Granger, 2001; Pravec, 2002; Tenfjord et al., 2004; Uzar, 1998). Čeprav ni nujno, da tisti, ki se še šolajo, pišejo več kot ljudje, ki so šolanje že končali, pa raziskovalci, ki so pogosto zaposleni na univerzah, lažje

pridejo do njihovih besedil. Seveda so primerni tvorci tudi starejši udeleženci tečajev tujega jezika – v vsakem primeru gre ponavadi za ljudi, ki se jezik učijo institucionalizirano, saj so le tako njihova besedila dostopna graditeljem korpusov.

Načini zbiranja podatkov o tvorcih so različni. Vir za korpusa CLC in PELCRA so izpitne pole, za korpusa ICLE in FRIDA tvorci izpolnijo vprašalnik (Pravec, 2002; Granger, 2001). Tudi načini vključevanja podatkov v korpus se razlikujejo, povsod pa so seveda anonimni. Švedski USE ima posebno bazo podatkov o tvorcih v excelovi datoteki (Axelsson, 2000), toda največkrat je to vključeno v glavo TEI vsakega besedila.

6.1. Prvi jezik

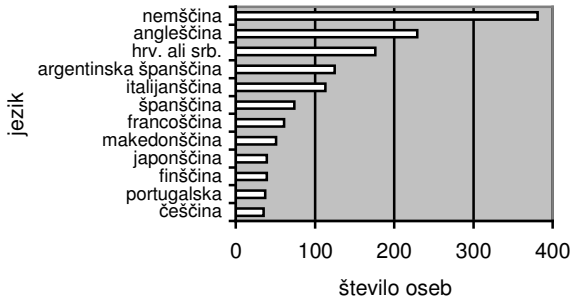
Ciljni jezik večine korpusov je angleščina, ki je tako razširjena po vsem svetu, da lahko tuji govorci angleščine z določenim prvim jezikom oblikujejo svoj korpus. Pri manjših jezikih, kjer je tudi manj konkurence, je situacija drugačna. Francoska FRIDA in norveški ASK imata več izhodiščnih jezikov, ki naj bi ustrezali jezikom največjih skupin priseljencev v tej državi. V Sloveniji je ob popisu leta 2002 87,7 odstotka prebivalstva za svoj materni jezik navedlo slovenščino (Šircelj, 2003), torej je bilo 12,3 odstotka govorcev, za katere slovenščina ni prvi jezik. Smiseln se zdi sklep, da večina govori slovenščino kot drugi jezik. Prav njihovi najpogostejši prvi jeziki bi morali biti tudi prvi jeziki v KUST-u, saj naj bi predstavljali pomemben del populacije govorcev slovenščine kot drugega jezika.



Slika 1. Materni jeziki (razen slovenščine) prebivalstva Slovenije ob popisu leta 2002 (Šircelj, 2003).

Sklepanje o govorcih slovenščine kot tujega jezika bo temeljilo na podatkih o udeležbi na tečajih slovenščine za tujce v letih od 2003 do 2005.[2] Ob tem se je treba zavedati, da bi bili za uporabnike KUST-a relevantnejši tisti tuji govorci, ki se učijo prek organiziranega učnega procesa, kar avtomatično daje prednost govorcem slovenščine kot tujega jezika.

[2] Podatki so bili pridobljeni na Centru za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani.



Slika 2. Prvi jeziki udeležencev tečajev slovenščine.[3]

Glede na obe sliki bi bilo treba v KUST-u upoštevati vsaj naslednje skupine izhodiščnih jezikov:

1. Govorci hrvaškega, srbskega oziroma bošnjaškega jezika so najštevilčnejši med govorci slovenščine kot drugega jezika (za 7,9 odstotkov prebivalstva je eden od teh jezikov materni), pa tudi med govorci slovenščine kot tujega jezika jih je 10 odstotkov. Zaradi jezikovnih razlik bi bilo najbolje za vsakega od teh treh jezikov oblikovati ločen podkorpus.

2. 0,2 odstotka prebivalcev sta leta 2002 kot svoj materni jezik navedla makedonščino, torej je za večino med njimi slovenščina drugi jezik. Tudi slabi 3 odstotki udeležencev tečajev govorijo makedonsko kot prvi jezik.

3. Nemščina kot materni jezik je med slovenskim prebivalstvom relativno redka (0,1 odstotek) in za Nemce slovenščina ni pogosto drugi jezik, zato pa je na tečajih slovenščine največ, kar 22 odstotkov udeležencev iz Nemčije, Avstrije in Švice.

4. Tudi za angleško govoreče osebe je slovenščina pomembna predvsem kot tuji jezik. Dobrih 13 odstotkov udeležencev tečajev slovenščine prihaja iz Velike Britanije, Irske, ZDA, Kanade, Avstralije in Nove Zelandije.

5. Govorci slovenščine iz Argentine, ki so večinoma potomci slovenskih izseljencev, imajo poseben status, saj je zanje slovenščina dejansko prvi jezik v družini, ne pa tudi v okolju. Njihov delež na tečajih slovenščine je 7-odstoten. Glede na jezikovno ozadje bi k njim lahko prikljućili tudi ostale govorce španščine (4 odstotke), ki prihajajo iz Španije, Mehike, Venezuele, Urugvaja, Kube, Peruja, Čila, Bolivije, Dominikanske republike in drugih držav. Korpusni podatki bodo pokazali, v kolikšni meri se njihova slovenščina razlikuje od slovenščine argentinskih izseljencev.

6. Zadnja večja skupina so govorci italijanščine (0,2 odstotka prebivalstva, dobrih 6 odstotkov na tečajih slovenščine), vendar gre tudi pri teh za precejšen delež zamejskih Slovencev, ki obiskujejo tečaje.

Druge izrazitejše skupine slovensko govorečih tujcev na tečajih so še tiste s francoščino, japonščino, finščino, portugalščino in češčino kot prvim jezikom.

Če izhajamo iz omenjene situacije, bi KUST lahko razdelili na podkorpuse, ki bi se ločili po izhodiščnih jezikih. Ti naj bi bili hrvaški, srbski in bošnjaški, makedonski, nemški, angleški, španski in italijanski. Smiselno je, da so med seboj po velikosti primerljivi, saj razmerja

[3] Hrvaški, srbski in bošnjaški jezik so združeni v isto kategorijo, ker se udeleženci tečajev slovenščine pri izpolnjevanju prijavnic pogosto ne opredelijo za enega od jezikov in podatki torej niso točni.

med populacijami niso konstantna. Tako bi imel v polmilijonskem korpusu vsak podkorpus v končni fazi 100.000 besed oziroma 400 besedil s po 250 besedami. Poleg tega bi bilo smiselno oblikovati še podkorpus ostalih izhodiščnih jezikov, ki bi bil bolj kot zanimivost, informacija ali primerjava.

6.2. Stopnja jezikovne zmožnosti

Le malo korpusov vključuje besedila začetnikov, saj ti tvorijo kratke tekste z mnogo odkloni od norme, popolnoma tujimi strukturami in malo koherentne vsebine. Precej primernejši so zato nadaljevalci ali izpopolnjevalci. Največ je korpusov, kjer so avtorji nadaljevalci (Granger, 2001; Kennedy, 1998; Pravec, 2002; Shih, 2000; Tenfjord et al., 2004). Ker je tudi na tečajih in izpitih iz slovenščine več nadaljevalcev kot izpopolnjevalcev,[4] bi se bilo v KUST-u smiselno osredotočiti na nadaljevalno stopnjo.

7. Kontrolni korpus

Uporabnost korpusa usvajanja tujega jezika poveča vzporedni kontrolni korpus jezika rojenih govorcev, ki omogoča primerjave. Tako dobimo kvantitativne podatke o pogostnosti določenih besed, besednih vrst, skladenjskih struktur in značilnostih diskurza (Tono, 2003). Nekateri za kontrolo uporabijo že obstoječe korpuse ali njihove dele, drugi pa zgradijo poseben podkorpus. Od 18 pisnih korpusov usvajanja tujega jezika, ki so relevantni ob razmišljanju o KUST-u, jih ima le šest kontrolni korpus (Granger, 2002; Tenfjord et al., 2004; Uzar, 1998; Izumi et al., 2004; Pravec, 2002). V obeh primerih si morajo biti načela gradnje približno podobna. Ker so kontrolni korpusi, ki zahtevajo precej truda graditeljev, skromno razširjeni, in ker za slovenščino obstaja referenčni korpus FIDA, bi lahko bil KUST vsaj na začetku brez posebnega kontrolnega korpusa.

8. Označevanje napak

Ker korpus usvajanja tujega jezika za razliko od običajnih korpusov nudi odklon od norme, je posebej koristno, če so v njem označene tudi napake, ki nastajajo pri jezikovni produkciji. Prav z analizo napak se je začelo raziskovanje vmesnega jezika učečih se, in čeprav danes to še zdaleč ni več edini vidik raziskav, je še vedno eden izmed najpomembnejših.

Za razliko od bolj razvitih in predvidljivih ravni označevanja, kot sta oblikoslovno označevanje in lematiziranje, je označevanje napak izjemno zamudno. Jezikovne rešitve učečih so večkrat tako nenavadne in ustvarjalne, da jih je nemogoče vnaprej predvideti. Za japonski korpus SST so skušali razviti sistem za avtomatično prepoznavanje napak. Program, ki je za uspešno delovanje potreboval dvaintrideset različnih podatkov,[5] so naučili na stopetdesetih dokumentih in ga preizkusili na šestnajstih, vendar so bili rezultati nezadovoljivi. Dodali so pravilne izjave iz podkorpusa rojenih govorcev, izjave izpraševalca in popravljene stavke iz korpusa. Rezultati so se nekoliko izboljšali, vendar so bili še vedno zgolj 43-odstotno natančni. Nazadnje so dodali umetne napake in s tem natančnost programa nekoliko zvišali, vendar bi potrebovali še več označenega gradiva za učenje, da bi bila natančnost zadovoljiva (Izumi et al., 2004). Zato označevanje poteka ročno in ni prav verjetno, da bo kmalu avtomatizirano.

Zaradi časovne potratnosti označevanja so napake zaenkrat označene le na manjšem delu korpusov. Od pregledanih korpusov usvajanja tujega jezika le v 14 označujejo napake, vendar tudi tu niso označene na vseh besedah, temveč zgolj na omejenem vzorcu, na primer na tretjini ali petini besed. Pri več kot desetmilijonskih korpusih to popolnoma zadostuje. V celoti je tako označenih 22 % vseh besed.

Za kakovostno in informativno obdelavo je treba napake klasificirati glede na določeno taksonomijo, ki naj bi omogočala opis in hkrati kvantitativno analizo. Klasifikacija je naporno, sporno in razmeroma brezplodno delo, kajti razdelimo jih lahko na različne načine glede na cilj raziskave in jezikoslovno teorijo, iz katere izhajamo. Tako celo za angleščino kot ciljni jezik ni ustaljenega načina, temveč vsak korpus deli napake po svoje. Zato so v 25-milijonskem HKUST-u označili pet milijonov besed samo s kategorijama "napaka" in "ne-napaka" (Pravec, 2002).

Možen način klasificiranja je glede na izvor napak. Delitev je več (Pirih Svetina, 2005), vendar je o izvoru brez dvosmerne komunikacije s tvorcem besedila in brez poznavanja njegovega ozadja dejansko mogoče samo ugibati. Oblikovalci korpusov tovrstnih klasifikacij zato ne uporabljajo. Edini, ki jih deli na omenjeni način, je Japanese Learners' Corpus (Sugiura, 2000).

Uporabnejše so tipologije, ki napake razvrščajo glede na spremembo predvidene ciljne oblike oziroma način odklona od nje. Tu ločimo zgrešitve ali napačne izbore, kjer je jezikovna značilnost narobe enkodirana, izpuste, kjer ni enkodirana, in dodajanja oziroma vstavitve, kjer je enkodirana odvečna jezikovna značilnost (Ragan, 2001). Včasih sta posebni kategoriji še neustrezen besedni red in sestava neobstoječe oblike iz dveh pravilnih.

V korpusih usvajanja tujega jezika se pojavlja tudi klasifikacija, ki napake razvršča znotraj posameznih jezikoslovnih ravnin in natančnejših kategorij: lahko so fonetične, oblikoslovne, oblikoskladenjske, skladenjske, pravopisne, leksikalne, povezane s samostalniki, glagoli, pridevniki in podobno. Take oznake so relativno objektivne, čeprav še vedno odvisne od interpretacije označevalca.

Najbolj priljubljene in tudi najuporabnejše so kombinacije klasifikacij. V japonskem korpusu SST ima vsaka napaka označene tri vrste podatkov: oblikoslovno umestitev, slovnično pravilo in pravilno obliko. Nabor skupaj vsebuje petinštirideset oznak. Poleg tega jih delijo tudi glede na spremembo ciljne oblike na izpuste, zamenjave in vstavitve (Izumi et al., 2004).

*I belong to two baseball **<n_num crr="teams">** team**</n_num>**.

Primer 1: Označen stavek z napako v korpusu SST: *n* označuje samostalnik, *num* pove, da gre za napako v številu, atribut *crr*= pa vsebuje pravilno obliko.

V korpusu ICLE vsaki napaki določijo krovno kategorijo, ali gre za formalno, slovnično, leksiko-slovnično, leksikalno napako, napako registra, odvečne/manjkajoče besede/napačen besedni red ali slog. Tej sledi vrsta podkategorij (Pravec, 2002). V korpusu FRIDA dobi vsaka napa-

[4] Podatki so bili pridobljeni na Centru za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani.

[5] To so ciljna beseda, dve besedi prej in dve potem, njihove besedne vrste in slovarske oblike, pet različnih kombinacij omenjenega ter prve in zadnje črke ciljne besede.

ka tri oznake: področje (slovnica, slovar, zapis itn.), kategorijo (vrsta, število itn.) in slovnično kategorijo (samostalnik, pridevnik itn.) (Granger, 2001). Norveški ASK ima zelo razčlenjeno taksonomijo na leksemske, morfološke, sintaktične, interpunkcijske in nerazvrščene napake, ki se nato še natančneje delijo (Tenfjord et al., 2004).

Zaradi težavnosti in spornosti klasificiranja napak se nekateri celo sprašujejo o smislu tega početja (Tono, 2003). Delo je pogosto intuitivno, možne so različne interpretacije, nobena analiza pa ne zajame vseh razlag (Ragan, 2001). Yukio Tono, ki je sodeloval pri več japonskih korpusih usvajanja angleščine, zaradi heterogenosti taksonomij predlaga, da bi se vsi držali splošnega nabora, kjer bi določili samo jezikovno kategorijo in način odklona od ciljne oblike, nato pa bi ga prilagajali potrebam svoje raziskave (2003). Da brez prilagajanja ne gre, je ugotovil, ko je pri ročnem označevanju shemo nehote adaptiral ciljem svoje raziskave.

Predlog klasifikacije v KUST-u je bil narejen po analizi manjšega nabora besedil tujih tvorcev[6] in napake deli na dveh ravneh.[7] Čeprav se zdi privlačna možnost, da je prvi kriterij sprememba ciljne oblike, to verjetno ni prvi podatek, ki bi zanimal profesionalne uporabnike, zato je prva raven kombinacija jezikovne ravnine in spremembe ciljne oblike. Tako so tu kategorije pravopis, besedišče, besedotvorje, oblikoslovje, besedni red, skladnja, izpust in vstavitev. K pravopisu spadajo vse besede z manjkajočimi ali odvečnimi črkami, napačna raba male oziroma velike začetnice in narobe zapisane besede. Besedišče pokriva napake, kjer so uporabljene besede, ki v slovenščini ne obstajajo ali pa so uporabljene v napačnem kontekstu. Napake zaradi napačnega tvorjenja besed so označene kot napake v besedotvorju. V oblikoslovje se uvrščajo napake zaradi nepravilnega pregibanja. Od skladenjskih napak se zaradi pogostnosti ločijo napake v besednem redu, tu gre lahko za napake znotraj samostalniških besednih zvez ali v naslonskem nizu. Izpusti in vstavitve se nanašajo na spremembo ciljne oblike in nimajo nadaljnjih atributov, ker bi bilo to glede na relativno redkost njihovega pojavljanja neekonomično. Poleg tega ugibanje o tem, kaj je v besedilu izpuščeno, zmanjšuje objektivnost in relevantnost oznak. Gotovo pa bi bilo smiselno uvesti tudi kategorijo nerazvrščenih napak za vse pojave, ki ne sodijo v nobeno drugo kategorijo.

Na drugi ravni je pomembna besedna vrsta. Strukturalistična razdelitev je pri tem nekoliko prilagojena pogostnosti pojavljanja, tako so kategorije samostalnik, pridevnik, glagol, prislov, števnik, predlog, veznik, členek in medmet. Prav veliko napak členkov in medmetov pri tujcih sicer ne pričakujemo, vprašanje pa je, kam uvrstiti samostalniške in pridevniške zaimke – ali narediti posebno kategorijo ali jih dati k samostalnikom oziroma pridevnikom? V izhodišču bi bila lahko izbrana slednja rešitev, nekoliko večji korpus učnih primerov pa bo pokazal njeno ustreznost.

1. raven	**2. raven**	**Primer**
pravopis	glagol	Ljudje so jo uzeli na piko.
besedišče	pridevnik	nisem rabil prvih luči
besedišče	členek	ta oseba ki je domneva prišla, ni res
besedotvorje	samostalnik	V ponedeljek sem poučevala francoskoščino.
oblikoslovje	glagol	V soboto in nedeljo sem šla na sprehod, sem berala
besedni red		Zjutraj zgodaj sem vstal.
skladnja		Na prvem vtisu je tema filma politična.
vstavitev[8]		
izpust		Včeraj sem zbudila zelo pozno.

Tabela 2: Primeri napak za posamezne kategorije[9]

Poskusi še natančnejšega klasificiranja, npr. zaradi napačnega spola, sklona ali števila, so pokazali, da se je pri tem nemogoče izogniti dvoumnostim in subjektivni interpretaciji. Čim bi imele tovrstne napake eno oznako, bi bila analiza otežkočena, saj bi bile druge interpretacije avtomatsko izključene.

V KUST-u poleg napak ne bodo napisane pravilne oblike. Na prvi pogled se to zdi možnost za dodatna pojasnila, posebej, kadar označevalec dobi že popravljen spis. Vendar je pogosto težko določiti, kaj je pravilna oblika očitno napačne jezikovne značilnosti. Domnevno pravilne oblike skrčijo možne interpretacije, ker označevalci od tvorcev besedil ne morejo izvedeti, kaj so v resnici hoteli povedati. Navajanje postane bolj ali manj posrečeno ugibanje, ki vsiljuje en sam vidik, zmanjšuje objektivnost označevanja in oteži priklic drugih možnosti.

Opisana klasifikacija je primerna za govorce najrazličnejših prvih jezikov, saj ni smiselno oblikovati ločenih klasifikacij za različne prve jezike. Seveda je že doživela prilagoditve in jih bo ob dejanskem označevanju učnega KUST-a nedvomno spet, saj bo treba sproti reševati probleme, kot je, kam uvrstiti uporabo napačne besedne vrste. Vendar je ne glede na (ne)popolnost klasifikacije pomembno predvsem to, da je označevanje konsistentno znotraj korpusa.

Zaradi omejene dostopnosti podatkov o obstoječih korpusih je težko podati natančen pregled tehnične plati označevanja napak, vendar jih večinoma seveda označujejo z jezikom SGML (HKUST) ali XML (LCLE, FRIDA, JEFLL, SST) (Pravec, 2002; Granger, 2001; Izumi et al., 2004). To je tudi edini smotrn označevalni jezik za KUST.

```
V soboto in nedeljo sem šla na sprehod, sem <sic
ana="TKUST.Obl.glag">berala</sic>
```

Primer 2: Označena napaka v poskusnem KUST-u (prim. op. 4); *sic* pomeni, da gre za napako, *TKUST.Obl.glag* pa, da je napaka oblikoslovna, in sicer glagola.

[6] Besedila petih tvorcev na nižji nadaljevalni stopnji s skupaj 600 besedami so nastala na tečaju slovenščine za tujce v študijskem letu 2004/2005.

[7] Klasifikacija je primerna za korpus, ki ni oblikoslovno označen. Ker sta v načrtu tudi lematizacija in oblikoslovno označevanje, bo treba klasifikacijo takrat ustrezno prilagoditi.

[8] V poskusno zbranem KUST-u (prim. op. 4) ni bilo primera vstavitve.

[9] Primeri so iz poskusno zbranega KUST-a (prim. op. 4).

9. Zaključek

Po analizi različnih že obstoječih tujih korpusov usvajanja, ki so delno lahko zgled za slovenski korpus, je nastala v članku predlagana zasnova korpusa usvajanja slovenščine kot tujega jezika, po kateri gre za pol- do enomilijonski korpus s ciljnim jezikom slovenščino in izhodiščnimi jeziki srbščino oziroma hrvaščino, makedonščino, angleščino, nemščino in italijanščino ter skupino vseh ostalih jezikov. Korpus je samo pisni. Tvorci so na nadaljevalni stopnji znanja, o njih so znani osnovni sociolingvistični podatki, podpišejo pa tudi privolitveni obrazec. Vir besedil so spisi z izpitov iz znanja slovenščine in s tečajev slovenščine, tematika je čim splošnejša. V korpusu ni besedil, napisanih doma, saj so tako pogoji nastanka relativno izenačeni za vse tekste. Kontrolnega korpusa ni, saj lahko kot to funkcionira FIDA, v njem pa so ročno označene napake po razdelani klasifikaciji, ki zajema način odklona od pričakovane oblike, jezikovno ravnino in besedno vrsto.

Naštete značilnosti in lastnosti bodo preizkušene na pilotskem korpusu usvajanja tujega jezika. Predvsem pri označevanju napak, kjer za slovenščino v ta namen še ni bilo razdelanega in preizkušenega sistema, pa je pričakovati največ sprememb in prilagoditev, ki bodo končno verzijo naredile prijaznejšo in uporabnejšo tako za znanstvene raziskovalce kot tudi za bolj praktično usmerjene pedagoge.

10. Literatura

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

Atwell, Eric, Howarth, Peter, Souter, Clive, 2003. The ISLE Corpus: Italian and German Spoken Learners' English. *ICAME Journal*. 27:5–18.

Axelsson, Margareta Westergreen, 2000. USE – The Uppsala Student English Corpus: An instrument for needs analysis. *ICAME Journal*. 24:155–157.

Cheng, Winnie, Warren, Martin, 1999. Facilitating a description of intercultural conversations: the Hong Kong Corpus of Conversational English. *ICAME Journal.* 23:5–20.

Dagneaux, Estelle, Granger, Sylviane, Meunier, Fanny, Petch-Tyson, Stephanie, 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.

De Cock, S., Granger, Sylviane, Petch-Tyson, S., 1999. *The Louvain International Database of Spoken English Interlanguage: The LINDSEI Project*. http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Lindsei/download/lindsei.pdf.

Erjavec, Tomaž, 2004. *Uvod v korpusno jezikoslovje*. http://nl.ijs.si/et/talks/korpus/korpusno.html.

Gillard, Patrick, Gadsby, Adam, 1998. Using a learners' corpus in compiling ELT dicitionaries. V S. Granger (ur.), *Learner English on Computer*. London, New York: Longman.

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

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

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.

Horváth, József, 1999. *Advanced Writing in English as a Foreign Language: A Corpus-Based Study of Processes and Products.* http://www.geocities.com/writing_site/thesis/index.html.

Izumi, Emi, Uchimoto, Kiyotaka, in Isahara, Hitoshi, 2004. SST speech corpus of Japanese Learners' English and automatic detection of learners' errors. *ICAME Journal*. 28:31–48.

Kennedy, Graeme, 1998. *An Introduction to Corpus Linguistics*. London, New York: Longman.

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.

Milton, John, 1998. Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. V S. Granger (ur.), *Learner English on Computer*. London, New York: Longman.

Pirih Svetina, Nataša, 2003. Napaka v ogledalu procesa učenja tujega jezika. *Jezik in slovstvo* 2:17–26.

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

Pravec, Norma, 2002. Survey of learner corpora. *ICAME Journal.* 26:81–114.

Ragan, Peter H., 2001. Classroom Use of a Systemic Functional Small Learner Corpus. V M. Ghadessy, A. Henry, in R. L. Roseberry (ur.), *Small Corpus Studies and ELT*. Amsterdam, Philadelphia: John Benjamins Publishing Co.

Shih, Rebecca Hsue-Hueh, 2000. Compiling Taiwanese Learner Corpus of English. *Computational Linguistic and Chinese Language Processing*. 2: 89–102.

Sugiura, Masatoshi, 2000. *On Enhancing the Writing Skill of EFL Learners in Japan: Utilizing the Insights and Technologies of Corpus Linguistics*. http://oscar.lang.nagoya-u.ac.jp/~sugiura/hawaii/680p/corpuswriting.html.

Šircelj, Milivoja, 2003. *Verska, jezikovna in narodna sestava prebivalstva Slovenije: Popisi 1921-2002.* Ljubljana: Statistični urad republike Slovenije-

Tenfjord, Kari, Meurer, Paul, 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.

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