Mojca
Stritar

KORPUSI USVAJANJA TUJEGA JEZIKA

Slavistična knjižnica

Mojca Stritar

KORPUSI USVAJANJA TUJEGA JEZIKA

Zveza društev
Slavistično društvo Slovenije
Ljubljana, 2012

Mojca Stritar
KORPUSI USVAJANJA TUJEGA JEZIKA
Recenziji: red. prof. dr. Marko Stabej, izr. prof. dr. Vojko Gorjanc
Prevod povzetka: doc. dr. Darja Fišer

Zbirka Slavistična knjižnica, 17
Urednica zbirke: izr. prof. dr. Irena Novak Popov
Tehnični urednik: Matjaž Zaplotnik
Računalniški prelom in priprava na tisk: Matjaž Zaplotnik
Oblikovanje ovitka: Pšenica Kovačič

Izdala Zveza društev Slavistično društvo Slovenije
Zanjo red. prof. dr. Boža Krakar Vogel

Ljubljana, februar 2012
Tisk: CICERO, Begunje, d.o.o.
Naklada: 250 izvodov

Monografija je izšla s podporo Javne agencije za knjigo Republike Slovenije.

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

81'243

STRITAR, Mojca
Korpusi usvajanja tujega jezika / Mojca Stritar ; [prevod povzetka Darja Fišer]. – Ljubljana : Zveza društev Slavistično društvo Slovenije, 2012. – (Zbirka Slavistična knjižnica / Slavistično društvo Slovenije ; 17)

ISBN 978-961-6715-10-2

259568384

Vsebina

1 Uvod 9

2 Korpusno jezikoslovje in korpus usvajanja tujega jezika . 14
2.1 Opredelitev korpusov 14
2.2 Tipologija korpusov 15
2.3 Opredelitev korpusa usvajanja tujega jezika 18
2.3.1 Poimenovanje korpusa usvajanja tujega jezika 19
2.3.2 Vmesni jezik in usvajanje prvega in tujega jezika 20
2.3.3 Jezikovna zmožnost 22
2.4 Pregled korpusov usvajanja tujega jezika 26
2.4.1 Korpusi angleščine kot tujega jezika 27
2.4.2 Korpusi francoščine kot tujega jezika 38
2.4.3 Korpusi nemščine kot tujega jezika 40
2.4.4 Korpusi italijanščine kot tujega jezika 40
2.4.5 Korpusi španščine kot tujega jezika 41
2.4.6 Korpusa švedščine kot tujega jezika 42
2.4.7 Korpusi ostalih jezikov 42
2.4.8 Korpusi slovenščine kot tujega jezika 45

3 Oblikovanje korpusov usvajanja tujega jezika 49
3.1 Specifikacija in oblika korpusa 50
3.1.1 Prenosnik 52
3.1.2 Velikost korpusa 53
3.1.3 Jezik 58
3.1.4 Tvorci 60
3.1.5 Vrsta besedil in tema 64
3.1.6 Kontrolni korpus 69
3.2 Zajem besedil 71

3.3 Označevanje besedil 73
3.3.1 Označevanje podatkov o tvorcu 75
3.3.2 Jezikoslovno označevanje 78
3.4 Oblikovanje poskusnega korpusa usvajanja slovenščine kot tujega jezika 81
3.4.1 Specifikacija in oblika 81
3.4.2 Zajem in označevanje besedil 85

4 Načela za oblikovanje korpusa usvajanja slovenščine kot tujega jezika 92
4.1 Prenosnik 93
4.2 Vrsta besedil in tema 95
4.3 Stopnja jezikovne zmožnosti 99
4.4 Ciljni jezik 101
4.5 Izhodiščni jezik 101
4.6 Velikost 111
4.7 Učno okolje slovenščine 112
4.8 Kontrolni korpus 112
4.9 Praktični zadržki 113

5 Napake 115
5.1 Proces analize napak 118
5.1.1 Definiranje napak 118
5.1.2 Odkrivanje napak 119
5.1.3 Opisovanje napak 120
5.1.4 Razlaga napak 123
5.1.5 Vrednotenje napak 127
5.2 Označevanje in klasifikacija napak 129
5.2.1 Klasifikacija napak po površinski strukturi 133
5.2.2 Klasifikacija napak po jezikovni ravnini 134
5.2.2.1 Klasifikacije napak v tujih korpusih 135
5.2.2.2 Slovenske klasifikacije 148
5.3 Označevanje napak v korpusu slovenščine kot tujega jezika 152
5.3.1 Oblikoslovne oznake 157
5.3.2 Pravilne oblike 160
5.3.3 Povezane napake 164
5.3.4 Dvojne napake 166
5.3.5 Kategorije napak 167
5.3.5.1 Zapis 169
5.3.5.2 Besedišče 177
5.3.5.3 Oblika 185
5.3.5.4 Skladnja 189

6 Uporaba korpusov usvajanja tujega jezika 196
6.1 Teoretična uporaba podatkov 199
6.2 Pedagoška uporaba podatkov 205
6.2.1 Učna gradiva 206
6.2.2 Učna metodologija 211

7 Zaključek 216

8 Seznam korpusov usvajanja tujega jezika s spletnimi strani 219

9 Literatura 222

10 Povzetek/Summary 244

11 Imensko kazalo 246

12 Kazalo tabel 252

13 Kazalo slik 254

14 Kazalo primerov 255

15 O avtorici 258

16 Iz recenzij 259

17 Doslej izšlo 262

1 UVOD

Slovenščina kot tuji jezik je pomemben del uporabne slovenistike, njeno poučevanje pa posebna strokovna disciplina, ki se v zadnjem času dinamično razvija. Obenem postaja vse močnejše zavedanje, da bi morala biti eden od temeljnih vidikov slovenskega jezikovnega načrtovanja. Govorci slovenščine kot tujega jezika 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 ustrezno teoretično in praktično podprli njihovo usvajanje slovenščine ter ga s tem olajšali in pospešili, je treba še bolje spoznati načela, po katerih poteka. Prav to pa omogočajo korpusi usvajanja tujega jezika.

Korpusi besedil se v jezikoslovju uporabljajo že več stoletij. Slovaropisje je od nekdaj potrebovalo obsežne in relevantne vire podatkov. Delo, ki je potekalo ročno, je bilo zamudno in neizčrpno, težko ga je bilo ponoviti ali preveriti, napake pa so bile dokaj pogoste. Zato je bil za razmah besedilno usmerjenega jezikoslovja ključen razvoj računalništva. Korpusno jezikoslovje se je pojavilo v drugi polovici dvajsetega stoletja, vendar se do sredine osemdesetih let ni razširilo. Delo z računalniškimi zbirkami besedil je bilo omejeno na velike računalnike, do katerih je imelo dostop le malo raziskovalcev, ki niso mogli delati brez pomoči računalniškega strokovnjaka; v sedemdesetih letih je iskanje konkordanc – prikaza vseh korpusnih pojavitev določene besede ali besedne zveze skupaj z neposrednim kontekstom – ene besede v milijonskem

korpusu trajalo več kot eno uro (Kennedy 1998: 6). Danes je z vsesplošno prisotnostjo zmogljivih osebnih računalnikov in programske opreme korpusno jezikoslovje del lingvistike, ki se bliskovito širi po vsem svetu na najrazličnejša področja raziskovanja in uporabe.

Korpusno jezikoslovje je empirično, saj temelji na konkretnih podatkih. Tovrstne študije so prevladovale v zgodnjih petdesetih letih in v začetku šestdesetih, nato pa so jih pričenjale izpodrivati intuitivne raziskave generativne slovnice. Spor med predstavniki obeh tokov se je tako zaostril, da sta se oblikovala ločena strokovna diskurza. Skupina okrog Chomskega je menila, da »naj bi govorec določenega jezika v desetih minutah produciral za jezikoslovce več ilustrativnega gradiva, kot ga je na voljo v večmilijonskih besedilnih korpusih« (Gorjanc 2003: 20), in dvomila o relevantnosti korpusnih podatkov. Drugi tok pa se je že od začetka usmerjal v uporabnost raziskav in se potrjeval v stiku z uporabniki jezikovnega znanja (prav tam). Vanj spada korpusno jezikoslovje, ki poskuša biti z dostopnostjo izhodiščnih podatkov o besedilih, njihovem zajemu in obdelavi čim bolj objektivno, velik obseg gradiva pa zmanjšuje možnost, da bi izjemne in obrobne jezikovne pojave interpretirali kot temeljne in tipične, s čimer povečuje verodostojnost analiz (Čermák 2005: 137). Danes je uporabnost korpusov splošno priznana in razširjena, vendar korpusna lingvistika ne konkurira drugim teorijam. Z njimi je večinoma komplementarna – ne nazadnje korpusi prikazujejo jezik brez vizualnega in družbenega konteksta, zaradi česar je vsaka posplošitev zgolj sklepanje (Hunston 2002: 23). Korpusi nudijo vire podatkov za vse teorije, ki izhajajo iz besedil (Kennedy 1998: 7). Tako se je oblikovala metodologija korpusne analize in novih jezikovnih opisov, ki namesto na intuiciji temeljijo na kvantitativno preverjeni jezikovni realnosti, osmišljeni šele z ustrezno kvalitativno razlago in interpretacijo podatkov (Gorjanc 2005: 7).

Korpusna analiza se je uveljavila tudi v preučevanju usvajanja tujega jezika. Ta proces so v 20. stoletju razlagali z različnimi dejavniki od okolja in genetske zasnove do spoznavnih procesov. V 60. letih dvajsetega stoletja so razlage skušale povezati behavioristično psihologijo s strukturalnim jezikoslovjem, v 70. je bila priljubljena teza, da učeči se samostojno ustvarja jezikovni sistem, v 80. pa so iskali močnejšo povezavo z ostalimi disciplinami (Sharwood Smith 1994: 22–23). Temu ustrezno so bile raznolike metode raziskovanja. Do šestdesetih in sedemdesetih let dvajsetega stoletja je

bila poglavitna kontrastivna analiza, nato jo je nadomestila analiza napak. Analizirane zbirke besedil so vsebovale največ dva tisoč besed in so bile predvsem heterogena zbirališča napak (Nesselhauf 2004: 129). Kontekst ali podatki o pogostnosti pojavov raziskovalcev niso zanimali (Tono 2002: 8), raziskovanje je potekalo ročno. Kmalu pa so se pojavila stališča, da ne bi smeli raziskovati zgolj napak, pač pa celotni jezikovni sistem učečega se (Richards, Sampson 1974: 5). Te poglede je zares utrdilo šele korpusno jezikoslovje s korpusi usvajanja tujega jezika, ki se ne osredotočajo samo na napake, ampak jih zanima vmesni jezik v celoti (Granger 1998a: 5–6). Razmahnili so se v devetdesetih letih s korpusoma angleščine kot tujega jezika ICLE in LCLE (prim. poglavje 2.4.1). Leta 1999 je bil prvi simpozij na to temo v Hongkongu. Danes večina korpusov nastaja v Evropi in Aziji, nekaj pa tudi v Združenih državah Amerike. Po uvodnem navdušenju, ki ga je spremljalo optimistično prepričanje o širokih možnostih korpusne analize, se je vzporedno z razvojem korpusov pričelo krepiti tudi bolj prizemljeno spoznanje, da je analiza korpusnih podatkov formalna, pogosto omejena na naštevanje kvantitativnih rezultatov, zato jo je za kakovostne raziskave treba kombinirati s kvalitativnimi metodami in jo utemeljevati na novih teoretičnih predpostavkah.

Na Slovenskem so se zametki korpusnega raziskovanja začeli konec devetdesetih let. Ena prvih dosledno empirično zasnovanih raziskav je bila analiza jezikovne zmožnosti 36 udeležencev tečaja slovenščine kot tujega jezika, čeprav gradivo ni dostopno v obliki korpusa (prim. Pirih Svetina 2005). Knjiga, ki je pred vami, pa temelji na doktorski disertaciji *Korpus usvajanja slovenščine kot tujega jezika*, ki sem jo pod mentorstvom red. prof. dr. Marka Stabeja zagovarjala na Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani decembra 2009. V okviru njenega nastanka sem spomladi 2005 sestavila, ročno uredila in označila učni korpus usvajanja slovenščine kot tujega jezika. Ker je šlo za preučevanje možnosti za urejanje in označevanje korpusa, je bil miniaturen in je pomenil zgolj nabiranje izkušenj za osrednji del raziskave, poskusni korpus usvajanja slovenščine kot tujega jezika (PiKUST, prim. poglavje 3.4). Ta je bil zbran, da bi pred gradnjo pravega korpusa usvajanja slovenščine kot tujega jezika (KUST) določila in praktično preizkusila natančnejše kriterije za označevanje in zajem besedil. Tovrstni eksperimentalni korpusi so »neke vrste poligoni za raziskovanje in pridobivanje izkušenj glede samega načina tehnične ureditve

korpusa« (Stabej 1998: 97) in so bolj kot jezikoslovnim raziskavam namenjeni treningu oblikovalcev večjih, bolj uravnoteženih korpusov. Kljub temu PiKUST dovolj nazorno ilustrira možnosti, ki jih lahko nudi dejanski KUST.

Besedila učečih se z različnimi jezikovnimi ozadji in nivoji znanja, ki sem jih vključila v PiKUST, so mi bila enostavno dostopna na Centru za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani, s katerim sem sodelovala kot lektorica na tečajih slovenščine za tujce in kot organizatorka teh tečajev. Tekste sem dobila od udeležencev tečajev in v arhivih Izpitnega centra. Označila in uredila sem jih na Univerzi v Bergnu na Norveškem, na kateri sem od oktobra 2006 do januarja 2007 gostovala s štipendijo Evropske unije Marie Curie v programu BATMULT na centru Aksis. Pri vprašanjih, kot je označevanje napak, sem za modele uporabljala že obstoječe korpuse usvajanja, posebej korpus norveščine kot tujega jezika ASK (prim. poglavje 2.4.7), v katerega sem imela največ vpogleda. Vendar so bili zgledi koristni predvsem po tehnični plati. Rešitev za angleščino ali norveščino ni mogoče avtomatično prenesti na slovensko gradivo, saj ima vsak jezik svojo strukturo, posebnosti in s tem povezane specifične napake. Pri opredeljevanju kategorizacije in načel za označevanje napak za slovenščino sem torej delno upoštevala priporočila tujih graditeljev korpusov, delno slovenistično tradicijo, v veliki meri pa sem izhajala iz problemov, ki so se pokazali v zbranem gradivu PiKUST-a.

Ko je bila kategorizacija v grobem postavljena, mi je pri podrobnejšem umeščanju napak in dilemi o navajanju pravilnih oblik pomagala anketa med slovenistično-jezikoslovno populacijo. Vprašalniki so bili sicer anonimni, toda ker je izobraževalni seminar Centra za slovenščino kot drugi/tuji jezik, na katerem sem anketo izpeljala, namenjen učiteljem slovenščine kot tujega jezika, prevajalcem ter drugim strokovno usposobljenim in zainteresiranim, je upravičeno sklepanje, da so odgovarjali ljudje, ki se profesionalno ukvarjajo s slovenščino kot tujim jezikom in so tudi potencialni uporabniki KUST-a. Vrnjenih je bilo 46 vprašalnikov, v katerih so morali anketirani dvajset avtentičnih napak tujcev, večinoma iz učnega KUST-a ali iz besedil, napisanih na tečajih slovenščine, razvrstiti po posameznih kategorijah in jim pripisati pravilno obliko. Primeri so bili izbrani tako, da naj bi bila polovica razmeroma enostavnih, ostali pa so ponujali več možnosti. Vnaprej predvidenih ali pravilnih rešitev ni bilo, saj je šlo predvsem za pridobivanje mnenja strokovno

usposobljene publike. Med pregledovanjem odgovorov se je izkazalo, da v enem od primerov zaradi skopega konteksta dejansko ni bilo napake, zato je bil izključen iz analize.

Pričujoča knjiga je sestavljena iz več delov. Poglavje 2 prinaša kratko opredelitev korpusov usvajanja tujega jezika, njihovega položaja znotraj korpusnega jezikoslovja in uporabljene terminologije. Obsežen del je pregled obstoječih korpusov usvajanja tujega jezika. Opozoriti je treba, da je v tiskani obliki dosegljivih le nekaj splošnih podatkov, tako da sem se morala zanašati na bolj spremenljive spletne vire. Eden od osrednjih problemskih sklopov, to je oblikovanje korpusov usvajanja tujega jezika, je obravnavan v poglavju 3. Pregledu rešitev po posameznih parametrih v obstoječih korpusih in opisu izvedbe v PiKUST-u sledi predlog za zasnovo uravnoteženega KUST-a v poglavju 4. Poglavje 5 se osredotoča na drugo veliko vprašanje raziskave, označevanje napak. Te so pojmovane kot del kreativne produkcije učečih se, del normalnega učnega procesa (Pirih Svetina 2005: 41). Teoretičnemu razmisleku in pregledu napak v tujih korpusih sledi predlog označevanja napak za slovenščino s primeri iz PiKUST-a. Poglavje 6 pa prinaša kratek pregled področij, na katerih se uporabljajo korpusi usvajanja tujega jezika, med katerimi je treba izpostaviti integrirano kontrastivno analizo ter aplikacijo korpusnih ugotovitev na izdelavo učnih gradiv, kot so slovarji, slovnice in učbeniki.

Zaključek uvoda je primerno mesto za zahvalo: red. prof. dr. Marku Stabeju za konkretno in splošno mentorstvo, sodelavkam Centra za slovenščino kot drugi/tuji jezik in norveškega Aksisa za pomoč na različnih frontah, ekipi Trojine za »šolarske« debate in vsem mojim za vse ostalo. Najpomembnejši pa so seveda vsi anonim(izira)ni udeleženci tečajev in izpitov iz znanja slovenščine, katerih besedila so z vključitvijo v PiKUST postala del raziskave – brez njih in njihove slovenščine tega dela res ne bi bilo.

2 KORPUSNO JEZIKOSLOVJE IN KORPUS USVAJANJA TUJEGA JEZIKA

2.1 Opredelitev korpusov

Korpus je enovita, notranje strukturirana in standardno označena zbirka besedil ali njihovih delov v elektronski obliki. Besedila so avtentična, zbrana po enotnih kriterijih za namene različnih, predvsem jezikoslovnih raziskav (Gorjanc 1999: 47) in standardno označena glede na namen korpusa.

Po definiciji evropske pobude Expert Advisory Group for Language Engineering (EAGLES), ki se je največ ukvarjala s tipološkimi in drugimi vprašanji korpusnega jezikoslovja, je korpus »zbirka kosov jezika, ki so izbrani in urejeni po eksplicitnih jezikoslovnih kriterijih z namenom, da služijo kot vzorec jezika«, računalniški korpus pa je »zapisan na standarden in enovit način za poljubne iskalne naloge« (Erjavec 2004). Poleg korpusov obstajajo še drugi tipi elektronskih zbirk besedil, kot so besedilni arhiv, besedilna baza podatkov in elektronska knjižnica (Atkins idr. 1992: 1; Kennedy 1998: 4). Da taki zbirki besedil rečemo korpus, mora izpolnjevati naslednje značilnosti po tipologiji EAGLES (Erjavec 1997: 82–93; Gorjanc 2005: 11–12; Gorjanc, Fišer 2010: 14–15):

- **količina**: velikost korpusa je izražena v številu besed oziroma pojavnic (angl. *token*);
- **kakovost**: nanaša se na avtentičnost oziroma realnost besedil, ki je manjša pri pisnih korpusih in besedilih, ki so bila umetno izzvana, in na njihov zapis, torej morebitne neavtorske posege vanje in podobno;

– **dokumentiranost**: korpus je ustrezno opremljen s podatki o besedilu in samem korpusu, kar določajo standardi za zapis in označevanje elektronskih besedil;
– **enostavnost**: računalniški zapis korpusa je razumljiv, oznake ne motijo prepoznavanja samega besedila in so od njega ločene. Označeni so tudi podatki, ki bi se pri vnosu izgubili.

2.2 Tipologija korpusov

Korpusi se razlikujejo glede na velikost, žanr, prenosnik in podobno. Najosnovnejši so **splošni, nespecializirani** oziroma **referenčni korpusi**, ki predstavljajo celovito podobo jezika. Zgrajeni so po natančni metodologiji, besediloslovnih in sociolingvističnih kriterijih in naj bi bili posplošljivi. Čeprav zajamejo le vzorec populacije, so precej obsežni in pogosto sestavljeni iz podkorpusov (Gorjanc 2005: 8). Za njihovo reprezentativnost sta pomembna obseg in raznoterost besedil. Obseg zmanjšuje možnost, da bi »prevladala izjemna in obrobna raba jezikovnih enot nad osnovno in tipično; tako se zmanjša tudi problematična indukcija, ki se opira na pomanjkljivo analogijo« (Čermák 2005: 137). Raznoterost pa se nanaša na zastopanost čim večjega števila različnih ključnih zvrsti in registrov, »ustrezno vključevanje dialektalnih ali regionalnih jezikovnih zvrsti, pa tudi uravnoteženost v smislu tematskih področij« (Vintar 2008: 82). Referenčni korpus torej posreduje »vsestranske izčrpne podatke o nekem jeziku, kar naj bi dosegel s količinsko in vrstno ustreznim izborom reprezentativnih besedil v tem jeziku« (Stabej 1998: 98). Nanje se opirajo temeljne jezikoslovne raziskave, ki niso že vnaprej zamejene (Kennedy 1998: 19). Najbolj razdelani in največji so referenčni korpusi za angleščino. Prvi referenčni korpus za slovenščino je bila stomilijonska FIDA, ki je bila nadgrajena v korpus FidaPLUS[1] s 621 milijoni besed (prim. Arhar, Gorjanc 2007), v sklopu projekta Sporazumevanje v slovenskem jeziku pa nastaja referenčni korpus Gigafida v obsegu ene milijarde besed (*Korpus pisnih besedil* 2008: 10),[2] ki naj bi vključeval tudi enomilijonski govorni korpus (prim. Zemljarič Miklavčič idr. 2009).

Oportunistični korpusi so cenejša različica referenčnih, zbrani glede na dane možnosti (Erjavec 1997: 83). Slovenski

1 Prim. http://www.fidaplus.net (avgust 2011).
2 Poskusna verzija korpusa je dostopna na http://demo.gigafida.net/ (oktober 2011).

korpus Nova beseda[3] je imel leta 2005 162 milijonov besed iz različnih virov od leposlovja do časopisnega, revijalnega, strokovnega gradiva in transkripcij parlamentarnih razprav.

Govorni korpusi so del referenčnih, zaradi drugačne metodologije izdelave pa so pogosto oblikovani samostojno. Čeprav je govorjenega jezika v realnosti bistveno več kot zapisanega, so tovrstni korpusi redki, ker je njihova izdelava zaradi zamudne transkripcije draga, težka in počasna (Kennedy 1998: 20). Bolj kot prozodične jih zanimajo slovnično-leksikalne lastnosti, medtem ko so raziskavam govora namenjeni **korpusi govora** oziroma **govorne zbirke** (prim. Kačič 2007). Govorni korpus za slovenščino GOS ima milijon besed oziroma okrog 110 ur govora (*Slikovni priročnik za uporabo spletnega iskalnika po korpusu GOS* 2010: 2).[4]

Specializirani korpus ali **korpus podjezika** je »reprezentativni vzorec jezikovnih zvrsti v okviru določenega strokovnega, socialno ali geografsko opredeljenega specialnega jezika« (Vintar 2008: 79). Zbran je s točno določenim namenom oziroma za jezik v točno določeni rabi (Gorjanc 2005: 8), npr. za raziskavo otroškega ali strokovnega jezika. Za doseganje reprezentativnosti je treba upoštevati parametre uravnoteženosti, kot so »izhodiščna določitev predmetnega področja in povezav z drugimi predmetnimi področji, vprašanja besedilnovrstne zastopanosti v okviru področja in razmerij med njimi, specifičnost razmerij med različnimi tipi prenosnikov« (Gorjanc 1999: 52). Slovenski korpus Beseda[5] ima dobre tri milijone besed iz slovenskih leposlovnih del, bolj tematsko specializirani pa so korpus Vonj za raziskovanje izražanja vonja v slovenščini (prim. Arhar 2006), iKorpus za računalništvo in informatiko (prim. Vintar, Erjavec 2008), Korpus slovenskega vojaškega jezika Grizold za obramboslovje (prim. Gorjanc, Logar 2007: 643), Korpus besedil odnosov z javnostmi KoRP,[6] Večjezični korpus turističnih besedil v slovenščini, angleščini in italijanščini (prim. Mikolič idr. 2008: 60)[7] ali nastajajoči korpus novinarskih besedil (prim. Červ 2009). Podtip specializiranih korpusov so tudi **korpusi usvajanja jezika** in **korpusi usvajanja tujega jezika**. Ker jim je namenjena večina pričujoče študije, so natančneje opredeljeni v poglavju 2.3.

3 Prim. http://bos.zrc-sazu.si/s_beseda.html (avgust 2011).
4 Prim. http://www.korpus-gos.net/ (avgust 2011).
5 Prim. http://bos.zrc-sazu.si/main_si_l2.html (avgust 2011).
6 Prim. http://www.korp.fdv.uni-lj.si (junij 2010).
7 Prim. http://jt.upr.si/turisticnikorpus/ (avgust 2011).

Vzporedni korpusi so dvo- ali večjezični in zanimivi za prevodoslovje ter terminografijo (prim. Vintar 2008), saj sopostavljajo izhodiščno besedilo in prevode v druge jezike, kar omogoča vpogled v prevajalske odločitve in strategije prevajanja (Gorjanc 2005: 10). Besedila je sorazmerno težko zagotoviti, zato so pogosto vključeni samo njihovi segmenti. Pri nas so vzporedni korpusi 10-milijonski slovensko-angleški SVEZ-IJS ACQUIS,[8] 2-milijonski angleško-slovenski TRANS,[9] enomilijonski slovensko-angleški ELAN,[10] vzporedni korpus desetih jezikov TELRI[11] in 231-milijonski korpus prevodov pravnih aktov Evropske unije v uradne evropske jezike Evrokorpus.[12]

Primerljivi korpusi vključujejo tematsko, jezikovnozvrstno ali besedilnovrstno primerljiva besedila v različnih jezikih. Slovenščina je vključena v projekt MultextEast,[13] pristop vzporednega in primerljivega korpusa pa združuje Slovenski prevodoslovni korpus (SPOOK)[14] (prim. Vintar 2009).

Do začetka devetdesetih let je standardna velikost za korpuse postala milijon besed, nato pa so se pojavili **dinamični, odprti** ali **spremljevalni korpusi**, ki nenehno rastejo. Dodajanje besedil je večkrat oportunistično in neuravnoteženo, vendar se v osnovi kažeta dva pristopa: stalno dodajanje novega gradiva z ohlapnejšimi parametri za zajemanje besedil ali sprejemanje novega gradiva in izločanje starega v diahrone podkorpuse z ves čas enakimi razmerji med besedili (Gorjanc 2005: 8). Dinamični korpusi so med drugim primerni za slovaropisne raziskave; v delu gradiva bo tak nastajajoči korpus Gigafida (Logar Berginc, Šuster 2009: 61).

Vsi zgoraj omenjeni korpusi so sinhroni, **diahroni** pa kažejo jezik v spreminjanju skozi določeno časovno obdobje in s tem omogočajo preučevanje jezikovnih sprememb. Helsinški korpus angleških besedil zajema tekste od leta 700 do 1700 (Kennedy 1998: 22).

Redek tip korpusa, ki pa je pomemben za učenje jezika, je **pedagoški korpus** (prim. Tono 2004: 50). Sestavljen je iz učbenikov, zvočnih posnetkov in vsega jezika, ki so mu izpostavljeni učeči se, torej tudi učbeniškega jezika in jezika posredovanja (prim. Jurjevič 2008). Z njim preverjamo, kako

8 Prim. http://nl.ijs.si/svez/ (avgust 2011).
9 Prim. http://nl2.ijs.si/index-bi.html (avgust 2011).
10 Prim. http://nl.ijs.si/elan/ (avgust 2011).
11 Prim. http://nl.ijs.si/telri/Republic/ (avgust 2011).
12 Prim. http://evrokorpus.gov.si/ (avgust 2011).
13 Prim. http://nl.ijs.si/ME/ (avgust 2011).
14 Prim. http://lojze.lugos.si/spook/korpus.html (avgust 2011).

pogosto se učeči se srečujejo s posamezno obliko, ali pa na podlagi primerjave z referenčnim korpusom ocenjujemo, ali je jezik pri poučevanju naraven in uporaben (Hunston 2002: 16).

2.3 Opredelitev korpusa usvajanja tujega jezika

Korpus usvajanja tujega jezika je računalniška zbirka besedil v jeziku tistih, ki niso njegovi domači govorci in se jezik šele učijo oziroma ga usvajajo (prim. poglavje 2.3.1). Domači govorci, imenovani tudi rojeni govorci, so tisti, ki jim je nek jezik prvi jezik, tuji pa tisti, ki jim je isti jezik tuji (Pirih Svetina, Ferbežar 2005: 4). Pri tem seveda ne gre zgolj za govorjenje, temveč se izraz nanaša na vse sporazumevalne spretnosti. Osnovni cilj korpusa usvajanja tujega jezika je zbrati objektivne podatke, s katerimi lahko opišemo ta jezik, s tem pa bolje razumeti tudi sam proces usvajanja (Pravec 2002: 81).[15] Korpus je namenjen splošni rabi, ne le posameznim študijam. Besedila so sistematično in uravnoteženo zbrana na podlagi predvsem zunanjih kriterijev, kot sta prvi jezik ali stopnja jezikovne zmožnosti tvorcev.

Korpusi usvajanja tujega jezika so praviloma usmerjeni v ciljni jezik, to je jezik, ki se ga »nekdo uči z namenom, da bi ga obvladal bodisi kot svoj prvi, drugi ali tuji jezik« (Pirih Svetina 2005: 139). Za kakovostno analizo podatkov pa je treba poznati tudi izhodiščni jezik učečega se, iz katerega se uči vse ostale jezike in iz katerega »najpogosteje izvirajo napake, ki jih pripisujemo jezikovnemu prenosu (interference)« (navedeno delo: 140).

Korpusi usvajanja odgovarjajo na nekatera tipična vprašanja (Leech 1998: xiv–xv):

- Katere jezikovne značilnosti ciljnega jezika uporabljajo učeči se manj ali bolj pogosto kot domači govorci in kje so razlike med govorci različnih prvih jezikov?
- Kako na vedenje učečih se v ciljnem jeziku vpliva njihov prvi jezik (jezikovni transfer)?
- Kje uporabljajo strategije izogibanja, torej ne izkoriščajo vseh možnosti ciljnega jezika?

15 S tega vidika korpus usvajanja tujega jezika ni isto, kot bi bil korpus nekega jezika kot tujega jezika. V slednjem primeru bi šlo za besedila tujih govorcev, ki jezika ne usvajajo več. S stališča znanstvene radovednosti bi bili taki korpusi zanimivi, saj bi med drugim nudili podatke o fosilizaciji v jeziku. Kljub temu jih v svetovnem ali slovenskem prostoru ni, nedvomno tudi zaradi precejšnjih metodoloških težav, povezanih z izbiro govorcev.

– Kje se najbolj približajo domačim govorcem?
– Kje so najmanj podobni domačim govorcem in potrebujejo največ pomoči?

2.3.1 Poimenovanje korpusa usvajanja tujega jezika

Korpusi usvajanja tujega jezika so se najprej pojavili v angleškem govornem prostoru, pri njihovem prenosu na slovensko govorno področje pa se pojavi zadrega ob prevajanju angleškega termina *learner corpus*. Neposredni prevod *korpus učencev* ni ustrezen, kajti *učenec* se najpogosteje povezuje z otroki v osnovni šoli.[16] Boljša ustreznica bi bil *korpus učečih se*, a zaradi prostega morfema *se* zveni manj običajno. Opreti se je treba torej na angleško poimenovanje *corpus of language acquisition* in upoštevati slovensko tradicijo.

Osnovna pojma, povezana s pridobivanjem jezikovne zmožnosti v neprvem jeziku, sta učenje in usvajanje jezika. Učenje je zavestna dejavnost, ki ponavadi poteka v formalnih okoliščinah, torej vodeno, organizirano, načrtno pridobivanje drugega jezika. Usvajanje pa je lahko nezavedno in podobno procesom otrok, ki se prvi jezik učijo spontano, neorganizirano, nenačrtno (Pirih Svetina 2005: 8). Zagovorniki nativističnih teorij usvajanja tujega jezika celo trdijo, da sta rezultat učenja in usvajanja različni, ne nujno povezani vrsti znanja. Tisti, ki so se jezik formalno in zavestno učili, naj bi bolje obvladali slovnična pravila, tisti, ki so ga pridobili z nezavednim usvajanjem, pa naj bi ga bolj tekoče uporabljali v vsakdanjih situacijah. Naučeno znanje je omejeno in večinoma popravlja izhodne podatke nezavedno pridobljenega znanja (Sharwood Smith 1994: 203).

Vendar se v slovenskem strokovnem prostoru termina učenje in usvajanje prekrivata. Oba se uporabljata za zavestno ali nezavedno dejavnost in tudi rezultati so pri obeh lahko znanje o jeziku in znanje jezika. Razlikujeta pa se v kontekstu: učenje poteka v razredu, usvajanje pa neposredno iz okolja (Pirih Svetina 2005: 8). Čeprav gre pri besedilih v korpusu ponavadi za rezultat formalne situacije v razredu, je pojem učenje preozek, saj se jezikovna zmožnost govorcev razvija tudi spontano – dejansko ni učenja brez usvajanja. S tega vidika je torej ustrezen prevod *korpus usvajanja slovenščine.*

16 *Slovar slovenskega knjižnega jezika* pod geslom učenec navaja: »1. kdor se uči, se poučuje v osnovni, srednji šoli /…/ / kdor se tako uči, poučuje, glede na odnos do svojega učitelja /…/ / kdor se usposablja za kak poklic, kako obrt /…/ 2. kdor se pri kom, od koga uči in sprejema njegovo mišljenje, nazor.«

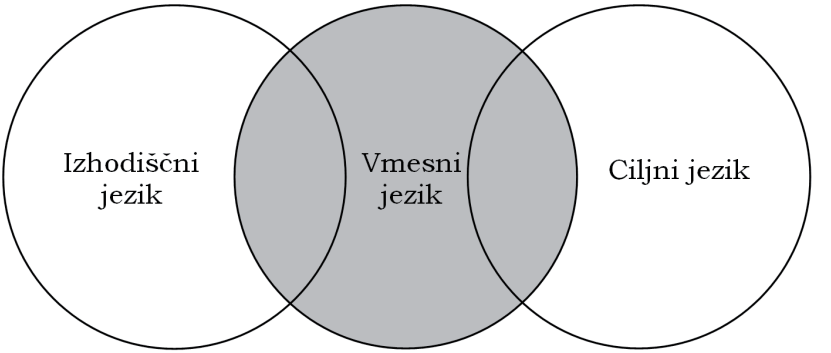
Da se poimenovanje omeji na odrasle govorce in se ne meša z otroškim usvajanjem prvega jezika, je treba dodati še razlikovanje med *drugim* in *tujim jezikom*, ki v anglosaškem okolju ni uveljavljeno. Jezik, ki se ga govorec najprej nauči, ki ga najbolje zna in v katerem je domači govorec, je prvi jezik (J1). Drugi jezik (J2, angl. *second language*, nem. *Zweitsprache*), ki se mu reče tudi jezik okolja, je jezik, ki se ga »posameznik uči/nauči poleg prvega ali za njim, lahko v šoli, predvsem pa (tudi) iz okolja; je uradni jezik in/ali jezik javnega življenja v določeni državi, človek pa se ga nauči zaradi sporazumevalnih potreb. V prvi vrsti ima torej socializacijsko funkcijo« (Ferbežar 1999: 418). Tuji jezik se nekdo uči kot svoj neprvi jezik v okolju, kjer ta jezik normalno ni v uporabi, ali za vzpostavljanje stikov zunaj lastne jezikovne skupnosti (Pirih Svetina 2005: 148). Drugi jezik je npr. italijanščina ali madžarščina za slovenske otroke na ozemljih narodnostnih manjšin (prim. Baloh 2005, Mikolič 2005), tuji pa španščina, nizozemščina ali švedščina, ki se jo Slovenci učijo v Sloveniji.

Najtočneje bi bilo zato uporabljati termin *korpus usvajanja slovenščine kot drugega ali tujega jezika*, toda v obstoječih korpusih gre v glavnem za tuji jezik in manj pogosto za drugi jezik. V slovenski situaciji se oba pojma prepletata in ju ni mogoče vedno razmejiti (Zemljarič Miklavčič 2001: 18). Nekateri tuji govorci živijo v Sloveniji, drugi ne, njihovo bivanje v slovensko govorečem okolju pa traja različno. Težko je postaviti mejo, kdaj gre za tuji in kdaj za drugi jezik. Zaradi elegantnejšega in bolj praktičnega poimenovanja je v tej raziskavi uporabljen splošni termin *korpus usvajanja tujega jezika* oziroma bolj specifični *korpus usvajanja slovenščine kot tujega jezika* (s kratico KUST), v primeru nujnega ločevanja med učenjem in usvajanjem ali med drugim in tujim jezikom pa je na to izrecno opozorjeno.

2.3.2 Vmesni jezik in usvajanje prvega in tujega jezika

Jezikovni sistem učečih je posebna oblika med izhodiščnim in ciljnim jezikom (slika 1), najpogosteje imenovana vmesni jezik (angl. *interlanguage*). Gre za relativno nestabilno fazo »v procesu učenja jezika, ki na začetku procesa učenja kaže le zelo malo značilnosti ciljnega jezika, potem pa vse več« (Pirih Svetina, Ferbežar 2005: 6). Imenuje se tudi približujoči se sistem (angl. *approximative system*, Nemser 1974: 155), saj gre za kontinuirani proces »od začetnega, skoraj neeksistira-

jočega znanja J2 proti jezikovni zmožnosti rojenega govorca« (Pirih Svetina, Ferbežar 2005: 6).



Slika 1: Vmesni jezik

Vmesni jezik je sistematično jezikovno obnašanje nedomačih govorcev določenega jezika (Sharwood Smith 1994: 200), idiosinkratično narečje (Corder 1974b: 160) z lastno slovnico in pravili, del katerih je del pravil ciljnega jezika. »Slovenščina s svojo zapleteno paradigmatsko in sintagmatsko razvejenostjo omogoča semantizacijo in sporazumevanje z zelo različnimi jezikovnimi sredstvi. Dejanskih uresničitev slovenščine, ki pa zaradi v osnovi enotnega sistema omogočajo sporazumevanje, je natanko toliko, kolikor je tudi njenih uporabnikov« (Pirih Svetina, Ferbežar 2005: 5). Za razliko od pravih narečij je vmesni jezik nestabilen in ga ne uporabljajo drugi člani širše družbene skupine (Corder 1974b: 161). Zanj je značilna fosilizacija, ki prepreči, da bi učeči se do konca razvil sporazumevalno zmožnost v ciljnem jeziku (Pirih Svetina 2005: 140). Identificirati jo je mogoče ob primerjanju vmesnega jezika z normo ciljnega jezika (Ferbežar 2007b: 195). Po sicer sporni teoriji 95 odstotkov odraslih, ki se učijo tuji jezik, nikoli ne doseže ravni domačega govorca (Selinker 1974: 34).

Ker korpus usvajanja tujega jezika nudi odklon od jezika domačih govorcev, ki naj bi predstavljal normo (prim. poglavje 4.4), je posebej koristno, če so v njem označene napake, ki nastajajo pri tvorjenju (Pravec 2002: 81). Ta dimenzija zahteva vrsto premislekov, ki jim je posvečen velik del nadaljevanja knjige (prim. poglavje 5). Z njo se srečujejo tudi oblikovalci korpusov usvajanja prvega jezika (prim. Rozman idr. 2011: 14), vendar je treba upoštevati, da sta procesa usvajanja prvega in tujega jezika podobna, ne pa identična. Razlike med njima so povzete v tabeli 1.

	Usvajanje prvega jezika	Usvajanje tujega jezika
Socialne razlike	Od zgodnjega otroštva	Kadar koli, a za prvim jezikom
	Neomejen dostop do jezika	Omejen dostop do jezika
	Neposreden in nezaveden stik z jezikom, podpora okolice	Manjša podpora okolice
	Ni pritiska	Pritisk v razredu ali okolju
	Pridobivanje družbene identitete	Spreminjanje družbene identitete
Kognitivne razlike	Vzporeden jezikovni in kognitivni razvoj	Prilagajanje že razvitih kognitivnih kategorij
	Hitro usvajanje	Individualno različna hitrost usvajanja
	Sistematičen razvoj, podoben pri vseh	Sistematičen razvoj, odvisen od posameznika in načina učenja jezika
	Prvi usvojeni jezik	Pomoč prvega in drugih usvojenih jezikov
	Mentalna slovnica je večja kot vnos	Mentalna slovnica je manjša kot vnos
	Usvajanje je neizogibno in uspešno	Usvajanje ni neizogibno in ni nujno uspešno
	Popolno obvladanje jezika	Redko popolno obvladanje jezika
	Ni fosilizacije	Pogosta fosilizacija
Afektivne razlike	Zanemarljiva vloga motivacije, osebnostnih lastnosti in stališč	Pomembna vloga motivacije, osebnostnih lastnosti in stališč

Tabela 1: Razlike med usvajanjem prvega in tujega jezika (Pirih Svetina 2005: 23–24; Corder 1982: 6)

2.3.3 Jezikovna zmožnost

Pri razpravljanju o korpusih usvajanja tujega jezika sta pomembna pojma sporazumevalna in jezikovna zmožnost. **Sporazumevalno zmožnost** sestavljajo »znanje in védenje, ki posamezniku omogoča, da lahko ustrezno in učinkovito /…/ uporablja jezik, in sposobnosti (spretnosti), da to znanje uporablja za sporazumevanje v realni situaciji – da lahko tvori in razume poljubno število besedil v različnih govornih položajih in za različne potrebe oziroma sporazumevalne namene« (Pirih Svetina 2005: 25). Gre za »znanja, sposobnosti in spretnosti, ki jih morata imeti govorec in poslušalec, da se lahko sporazumeta – razumeta in tvorita poljubno število besedil v najrazličnejših govornih položajih in za različne potrebe – sporazumevalne namene« (Ferbežar 1999: 420). Sporazumevalna zmožnost se vedno realizira v konkretnih okoliščinah, saj je sicer le »potencialna zmožnost govorca« (Ferbežar 2004b: 193).

Jezikovno zmožnost nekateri pojmujejo kot sopomenko sporazumevalne zmožnosti, drugi pa kot njen del (prim. Pi-

rih Svetina 2005; Ferbežar 1999). Pomeni obvladanje okoliščin in pravil tvorbe oz. jezikovne rabe oziroma »znanje jezika in védenje o njem, ki je brez posameznikove sposobnosti uporabe (t. i. strateške zmožnosti) v konkretni situaciji brez vrednosti« (Ferbežar 1999: 421). Sestavljata jo organizacijska in pragmatična zmožnost. Prva obsega obvladanje koda in njegovih pravil na vseh ravninah ter zmožnost oblikovati pravilna in smiselna besedila, k pragmatični pa sodita ilokucijska, povezana s prepoznavanjem funkcij besedila in sporazumevalnih namer govorca, in sociolingvistična zmožnost, ki pomeni zmožnost upoštevanja družbenih pravil jezikovnega vedenja (prav tam).

Pri opazovanju jezikovne zmožnosti učečih se so korpusi usvajanja tujega jezika postavljeni v paradoksalen položaj. Razvita sporazumevalna zmožnost naj bi bila najvišja dosežena stopnja znanja in kot taka končni cilj učenja tujega jezika (Pirih Svetina 2005: 147), vendar je mogoče o njej sklepati samo preko jezikovne performance, to je aktualne jezikovne rabe v aktualnem dogodku, oziroma preko vedênja, ki se ga da opazovati (navedeno delo: 31). Kompetenca se kaže prek performance, aktualnega razumevanja in oblikovanja besedil, in uporabniki jezika se ločujejo prav po njej (Ferbežar 1999: 423). Učeči se v določenem jezikovnem dogodku ne pokaže vse svoje jezikovne zmožnosti, saj lahko nejezikovne spremenljivke v performanci prikrijejo kompetenco (Pirih Svetina 2005: 31). Posebej je to značilno za testne situacije, v katerih lahko stres, strogi pogoji in nenaravnost situacije performanco precej zmanjšajo. Zato govorijo o **testni performanci**, jezikovni recepciji in produkciji posameznikov, ki rešujejo test (prav tam). Ta ni nujno pravi kazalec posameznikove jezikovne zmožnosti. Korpusi so za raziskovanje stopnje jezikovne zmožnosti torej uporabni zgolj ob zavedanju, da jo je mogoče izmeriti le toliko, »kolikor se (aktualno) manifestira v jezikovnem vedenju, to je performanci« (Ferbežar 1999: 423).

Ne glede na načelno neulovljivost je jezikovna zmožnost v korpusih usvajanja tujega jezika pomemben dejavnik, ki ga graditelji korpusov pogosto upoštevajo kot enega izmed ključnih parametrov za izbor besedil (prim. poglavje 4.3). Določa se ali objektivno z različnimi jezikovnimi testi (prim. Ferbežar 1999) ali subjektivno, po osebnem mnenju in opažanju pedagoških praktikov. Pogosto se uporabljajo stopnje začetniki, nadaljevalci in izpopolnjevalci, ki se lahko opredeljujejo z oznakami nižji ali višji. Različne institucije stopnje

različno definirajo in ločujejo, včasih niti ne po eksplicitnih, sorazmerno objektivnih, temveč zgolj po zunanjih kriterijih. Na tečajih slovenščine za tujce v organizaciji Centra za slovenščino kot drugi/tuji jezik so stopnje znanja denimo povezane z učbeniki, ki se uporabljajo v razredu (prim. Knez 2005). Pri tem je treba omeniti še posebno skupino, Slovane začetnike, kar pomeni udeležence tečajev iz republik nekdanje Jugoslavije, ki se slovensko še niso učili, njihov prvi jezik pa je eden od slovanskih jezikov. Pri njihovem usvajanju slovenščine je izrazit jezikovni prenos, »prenos in vgraditev značilnosti in struktur enega jezika v sistem znanja drugega jezika, ki ga želi zgraditi učeči se« (Pirih Svetina 2005: 141). Pri učenju jezikov je lahko pozitiven, saj omogoča sporazumevanje že na začetni stopnji učenja jezika, in negativen, ker pri prenašanju elementov med jeziki prihaja do napak, ki jih je zaradi podobnosti sistemov težko odpraviti (Balažic Bulc 2004: 81). »Prednosti so zlasti velike v začetnih fazah učenja/usvajanja slovenščine kot J2, saj sorodnost jezikov omogoča takojšnje sporazumevanje« (Požgaj Hadži, Ferbežar 2001: 59). Ker Slovani začetniki bistveno bolje razumejo slovenščino in se razmeroma tekoče sporazumevajo, se jih na tečajih slovenščine kot tujega jezika ne enači s t. i. popolnimi začetniki, to je s tistimi govorci, ki ne znajo nobenega od slovanskih jezikov.

Za večjo enotnost in preglednost na področju učenja tujih jezikov, posebej na področju stopenj znanja v državah članicah Evropske unije, je bil na pobudo Sveta Evrope zasnovan Skupni evropski referenčni okvir (SEJO) za jezike (*Skupni evropski jezikovni okvir* 2011). Namenjen je izboljševanju možnosti za mobilnost znotraj držav članic, olajšanju priznavanja dokazil o znanju jezika in spodbujanju sodelovanja med izobraževalnimi ustanovami v različnih državah. SEJO ločuje šest ravni jezikovnega znanja, navedenih v tabeli 2, med katerimi sta zaenkrat za slovenščino kot tuji jezik natančno opredeljeni ravni A1 (prim. Pirih Svetina, Rigler Šilc idr. 2004) in B1 (prim. Ferbežar, Knez idr. 2004).

Oznaka	Stopnja	Opis
A1	Osnovni uporabnik (vstopna raven)	Razume in uporablja pogoste vsakodnevne izraze in zelo osnovne besedne zveze za zadovoljevanje konkretnih potreb. Predstaviti zna sebe in druge ter spraševati in odgovarjati na osebna vprašanja. Obvlada preprosto interakcijo, če sogovorec govori počasi in razločno ter je pripravljen pomagati.

Oznaka	Stopnja	Opis
A2	Osnovni uporabnik (vmesna raven)	Razume stavke in pogoste izraze, ki se nanašajo na najbolj temeljna področja. Sposoben se je sporazumevati v preprostih in rutinskih opravilih, ki zahtevajo preprosto in neposredno izmenjavo informacij o znanih in rutinskih zadevah. S preprostimi besedami zna povedati nekaj o sebi in svojem neposrednem okolju ter zadovoljiti svoje trenutne potrebe.
B1	Samostojni uporabnik (raven sporazumevalnega praga)	Pri razločnem govorjenju v standardnem jeziku razume glavne točke, kadar gre za znane reči, s katerimi se redno srečuje na delu, v šoli, prostem času itn. Znajde se v večini situacij, ki se pogosto pojavljajo na potovanju po deželi, kjer se ta jezik govori. Tvoriti zna preprosta povezana besedila o temah, ki so mu znane ali ga zanimajo. Opisati zna izkušnje in dogodke, sanje, upanja in ambicije ter na kratko utemeljiti in pojasniti svoja mnenja in načrte.
B2	Samostojni uporabnik (višja raven)	Razume glavne ideje v kompleksnem besedilu tako o konkretnih kot abstraktnih temah, vključno z razpravami s svojega strokovnega področja. V medosebnih stikih govori dovolj tekoče in spontano, da njegova komunikacija z domačimi govorci ni naporna ne zanj in ne za sogovorca. Tvoriti zna jasno, poglobljeno besedilo iz širokega razpona tem in razložiti stališče o nekem aktualnem vprašanju, tako da pojasni prednosti in pomanjkljivosti različnih možnosti.
C1	Učinkoviti uporabnik (raven učinkovitosti)	Razume širok razpon zahtevnejših, daljših besedil in prepoznava implicitne pomene. Izraža se tekoče in spontano brez očitnega iskanja ustreznih izrazov. Jezik uporablja prožno in učinkovito, tako za družabne kot za akademske in poklicne namene. Tvoriti zna jasna, dobro organizirana in poglobljena besedila o kompleksnih temah. Pri tem nadzorovano uporablja organizacijske vzorce, povezovalce (konektorje) in kohezivna sredstva.
C2	Učinkoviti uporabnik (raven mojstrstva)	Z lahkoto razume tako rekoč vse, kar sliši ali prebere. Zna povzemati informacije iz različnih govornih in pisnih virov in pri tem povezati argumente in pripovedi v koherentno obliko. Izraža se spontano, zelo tekoče, natančno in tudi v kompleksnih situacijah ločuje med drobnimi pomenskimi odtenki.

Tabela 2: Globalna lestvica jezikovnega znanja SEJO (*Skupni evropski jezikovni okvir* 2011: 46)

Na Slovenskem pa se za opisovanje ravni znanja slovenščine kot tujega jezika uporablja tudi lestvica izobraževalnega programa Slovenščina za tujce (Ferbežar, Pirih Svetina 2002), ki je v veljavi na Izpitnem centru Centra za slovenščino kot drugi/tuji jezik Filozofske fakultete Univerze v Ljubljani. Ta je od leta 1994 pooblaščen za preverjanje znanja slovenskega jezika in izdajanje javno veljavne listine o znanju slovenščine kot drugega jezika, ki je med drugim pogoj za pridobitev slovenskega državljanstva, za vpis na katero od

slovenskih univerz ali visokih šol oziroma za pridobitev delovnega dovoljenja.[17]

Z izpitom iz znanja slovenščine se »meri splošna jezikovna zmožnost v slovenščini in ni bil razvit za druge, specifične namene« (Ferbežar 2007c: 8). Opravljati ga je mogoče na treh ravneh v skladu s standardi znanja, določenimi v izobraževalnem programu. Vse tri ravni upoštevajo govorca in raven njegove samostojnosti pri sporazumevanju v slovenščini (Ferbežar, Pirih Svetina 2004: 25). Osnovna raven predvideva približno 250 ur učenja, njeni govorci »se znajdejo v vseh življenjskih okoliščinah, to je ponovljivih 'rutinskih' sporazumevalnih položajih, zlasti tistih, ki vključujejo stvarne informacije« (Ferbežar, Pirih Svetina 2002: 12). Srednja raven predvideva skupaj 375 ur, govorci se »brez težav znajdejo v vsakdanjih (ponovljivih) življenjskih okoliščinah, poleg tega pa tudi dovolj dobro v nepredvidenih sporazumevalnih položajih, pogovarjati se znajo o precej širokih družbenih temah, specializirano zlasti o temah, ki so jim blizu /.../. Na tej ravni že razločujejo med knjižnim zbornim in knjižnim pogovornim jezikom in znajo glede na okoliščine in namen tudi uporabiti ustrezno zvrst« (navedeno delo: 13). Visoka raven predvideva skupaj 500 ur učenja, govorci so pri uporabi jezika povsem neodvisni in samostojni. »Sposobni so se sporazumevati v najrazličnejših okoliščinah in sporazumevalnih položajih, obvladajo tudi strokovno izrazje (predvsem lastnega poklicnega področja). Znajo razločevati med knjižnim zbornim in knjižnim pogovornim jezikom in so seznanjeni z ostalimi socialnimi zvrstmi (predvsem neknjižnim pogovornim jezikom) ter v besedilih prepoznavajo njihove značilnosti« (navedeno delo: 14). Ravni Izpitnega centra so primerljive s stopnjami po SEJO: osnovna raven s stopnjo B1, srednja z B2, visoka pa s C1 (Ferbežar, Pirih Svetina 2004: 25).

2.4 Pregled korpusov usvajanja tujega jezika

Korpusi usvajanja tujega jezika se hitro širijo tako po številu kot po obsegu in uporabnosti, tako da se v zadnjih letih uresničuje napoved ene od prvih in še danes vodilnih strokovnjakinj za korpuse usvajanja tujega jezika Sylviane Granger (1998a: xxii) o skorajšnjem izbruhu tovrstnih korpusov. Potem ko so najprej nastajali predvsem za angleščino kot tuji jezik, jih danes na vseh koncih sveta gradijo tudi za druge

17 Prim. http://www.centerslo.net/l1.asp?L1_ID=3&LANG=slo (avgust 2011).

jezike. Tudi podatki o njih so razmeroma dosegljivi, saj ima večina korpusov svoje spletne strani.

V nadaljevanju so na kratko predstavljeni najbolj znani in uveljavljeni korpusi, ki so tudi najvplivnejši, saj so vsaj deloma zgled za oblikovanje novih korpusov. Združeni so glede na ciljni jezik. Opozoriti je treba, da se nekatere njihove značilnosti, npr. velikost, hitro spreminjajo, nekatere napovedi in načrti za prihodnost, ki jih lahko zasledimo v literaturi, pa še danes ostajajo nerealizirani. Najaktualnejši podatki so običajno na spletnih straneh korpusov, ki so zbrane v poglavju 8 ob koncu knjige. Zaradi dinamičnosti spleta so sicer spletni naslovi že lahko zastareli, zato je naveden tudi datum zadnjega obiska, ko so strani delovale. Tudi sam pregled korpusov zaradi hitrega razvoja, ki področje spreminja v precej nepregleden nabor virov, ni dokončen. Manjši korpusi, ki jih pogosto za svoje trenutne potrebe izdelajo učitelji ali raziskovalci (prim. Ragan 2001: 217–218), niso omenjeni. Ker praviloma ne določajo svetovnih standardov, temveč jim kvečjemu sledijo, za pričujočo raziskavo niti niso relevantni. Za informacije o morebitnih novih ali manj znanih korpusih se je najbolje sklicevati na spletno stran Centra za angleško jezikoslovje Katoliške univerze v Louvainu v Belgiji, ki je osrednja institucija na tem področju in stalno osvežuje tako korpusno bibliografijo kot seznam obstoječih korpusov tega tipa.[18]

2.4.1 Korpusi angleščine kot tujega jezika

ICLE: International Corpus of Learner English je prvi akademski in mednarodno najbolj razširjeni korpusni projekt usvajanja tujega jezika, ki ga od leta 1990 pod vodstvom Sylviane Granger gradijo na Katoliški univerzi v belgijskem Louvainu v sodelovanju z drugimi svetovnimi univerzami. Ta pisni korpus angleščine kot tujega jezika sestavljajo spisi univerzitetnih študentov na izpopolnjevalni stopnji jezikovne zmožnosti. Podkorpusi različnih prvih jezikov vsebujejo po dvesto tisoč besed. Zaključeni so podkorpusi iz Bolgarije, Češke, Finske, Francije, Italije, Japonske, Južnoafriške republike, Kitajske, Madžarske, Nemčije, Nizozemske, Norveške, Poljske, Rusije, Španije, Švedske in Turčije, v delu pa so še podkorpusi iz Brazilije, Grčije, Litve, Maroka, Pakistana, Portugalske, Romunije in Slovenije. Skupaj ICLE trenutno vsebuje 3,7 milijona besed. Dodane ima oblikoslovne,

18 Prim. http://www.uclouvain.be/en-cecl-lcworld.html (avgust 2011).

sintaktične in semantične oznake, pa tudi oznake napak. Za analizo uporabljajo program WordSmith Tools (Granger 2001; spletna stran).

Govorna vzporednica korpusa ICLE je **LINDSEI: Louvain International Database of Spoken English Interlanguage**, ki jo na isti instituciji gradijo od leta 1995. Ima milijon besed in je sestavljen iz več podkorpusov: bolgarskega, francoskega, grškega, kitajskega, italijanskega, japonskega, nemškega, nizozemskega, poljskega, španskega in švedskega, nastajajo pa še arabski, baskovski, litvanski, norveški, portugalski (Brazilija) in turški. Vsak vsebuje prepise petdesetih petnajstminutnih intervjujev s študenti 3. ali 4. letnika, skupaj približno sto tisoč besed. Pogovori so ortografsko prepisani. Na začetku si informator izbere eno od treh splošnih tem (npr. zelo dober/slab film/predstava), na katero se pripravi in o njej govori. Čez nekaj časa se začne vključevati raziskovalec. Tik pred koncem intervjuja informator na podlagi štirih slik pove zgodbo in nazadnje izpolni obrazec s splošnimi podatki o sebi, enak tistemu za ICLE (De Cock idr. 1999: 1–3). Kot kontrolni korpus uporabljajo primerljivi korpus intervjujev z domačimi govorci angleščine LOCNEC.

Na Univerzi v Louvainu so leta 2008 začeli projekt **LONGDALE: Longitudinal Database of Learner English**, v okviru katerega želijo longitudinalno zbrati veliko bazo angleških besedil učečih se z različnimi prvimi jeziki. V njej naj bi bili raznoliki pisni in govorjeni teksti univerzitetnih študentov skozi tri leta študija od argumentativnih spisov, pripovedi in neformalnih intervjujev do povzetkov ali opisov slik. Projekt je še razmeroma na začetku, a ga želijo oblikovati podobno kot ICLE (spletna stran).

Dva izmed največjih obstoječih korpusov sta komercialna LCLE in CLC. 10-milijonski pisni **LCLE: Longman Corpus of Learners' English** je kot del skupine Longman Corpus Network v letih 1990–2002 oblikovala založba Longman. Gre za tvorce iz več kot 160 jezikovnih okolij na osmih stopnjah jezikovne zmožnosti. Korpus je uravnotežen po nacionalnostih in stopnji znanja tvorcev, po vrstah besedila, ciljni varianti (britanska ali ameriška angleščina) in državi, v kateri živijo učeči se. Na delu besedil so z jezikom XML označene napake. LCLE je dosegljiv za akademske raziskave v celoti ali le delno po podkorpusih glede na prvi jezik učečih se ali stopnjo jezikovne zmožnosti (Pravec 2002: 94; spletna stran).

Pisni korpus založbe Cambridge University Press **CLC: Cambridge Learner Corpus** je še večji, saj ima 35 milijonov

besed. Tako kot LCLE je tudi CLC del obsežnega komercialnega korpusa, tj. Cambridge International Corpusa, približno šeststomilijonske zbirke besed britanske in ameriške angleščine ter angleščine kot tujega jezika. Besedila so zbrana na Cambridgeovih izpitih ESOL:[19] po zadnjih podatkih gre za tekste 135.000 študentov iz 190 držav in 130 jezikovnih okolij. Vsak je označen s podatki o prvem jeziku, narodnosti, stopnji znanja, starosti ipd., ki jih zberejo kot del izpita, na 30 milijonih pa so označene tudi napake. Informacije so anonimne (Pravec 2002: 94, spletna stran). Korpus je dostopen samo avtorjem publikacij založbe Cambridge University Press ali sodelavcem pri izpitih ESOL.

Milijonski govorni korpus akademske angleščine **ELFA: English as a Lingua Franca in Academic Settings Corpus** so v letih 2004–2007 oblikovali na finskih univerzah v Tampereju in Helsinkih. Okoli 650 govorcev, ki so študentje, profesorji in drugi udeleženci akademske komunikacije, ima 51 različnih prvih jezikov. Korpus vključuje dogodke iz univerzitetnega življenja, izbrane na osnovi tipičnosti (npr. seminarji, zagovori doktorskih nalog, predstavitve na konferencah), vplivnosti (npr. uvodna predavanja) in prestiža (npr. vabljena ali plenarna predavanja na konferencah). 33 % besedil je monoloških, večino pa predstavljajo dialoška besedila. Okoli pet odstotkov je bilo domačih govorcev angleščine, vendar so sodelovali le v diskusijah z več udeleženci. Posnetki so transkribirani (Mauranen 2003: 522; Mauranen 2007: 38; spletna stran).

Milijonski **VOICE: Vienna-Oxford International Corpus of English**, ki ga gradijo na Univerzi na Dunaju, je namenjen angleščini kot *lingui franci*, torej kot jeziku, ki druži govorce različnih jezikov in za večino ni prvi jezik, znotraj različnih okolij, jezikovnih funkcij, udeleženjskih vlog in razmerij. Trenutno vključuje posnetke 1260 govorcev s 50 prvimi jeziki, med njimi desetino domačih govorcev, in naravnih, interaktivnih govornih dejanj: pogovorov, dialogov, razprav, predstavitev, poslovnih srečanj, delavnic, seminarjev, intervjujev, telefonskih pogovorov itn. 25 % besedil je s področja izobraževanja, 10 % iz prostega časa, 20 % iz poslovnega okolja, 35 % iz okolja poslovnih organizacij in 10 % iz znanosti. Besedila so transkribirana, ne pa tudi označena. Posebna

19 Gre za izpite Key English Test, Preliminary English Test, First Certificate in English, Certificate in Advanced English, Certificate of Proficiency in English, Business English Certificate (vse stopnje), International English Language Testing System, Certificates in English Language Skills, International Legal English Certificate, International Certificate in Financial English, Skills for Life (spletna stran CLC).

pozornost je posvečena zapisu naglaševanja, menjave koda, onomatopejskih zvokov, smeha ipd. (Breiteneder idr. 2006; spletna stran).

Pisni korpus **USE: Uppsala Student English Corpus**, ki so ga v letih 1999–2001 zgradili na Univerzi v Uppsali na Švedskem, ima 1.200.000 besed, zaželena velikost pa je dva milijona. Sestavljen je iz 1489 esejev, razdeljenih na štirinajst kategorij. Besedila s predpisano dolžino so študentje 1., 2. ali 3. letnika angleščine napisali doma v dveh do treh tednih. Uporabljali so lahko slovarje, slovnice in druge didaktične pripomočke. Pred oddajo v digitalni obliki je bilo zaželeno, da tekste še enkrat pregledajo. Oblikoslovne oznake na delu besedil so bile dodane s programom Brill, za analizo pa uporabljajo WordSmith Tools in Qwick. V bazi podatkov so anonimni podatki o 440 tvorcih besedil (Axelsson 2000: 155).

Na Univerzi v Bergnu so zgradili korpus **EVA (Evaluation of English) Corpus of Norwegian School English** z norveščino kot izhodiščnim jezikom. Ima 100.000 besed in tri podkorpuse:

- govorni podkorpus (35.000 besed): transkripcije govora 62 učencev, starih 14 ali 15 let, ki so leta 1994 sodelovali na ustnem izpitu EVA. Govor prisotnega učitelja je zaradi konteksta vključen v korpus, ni pa analiziran;
- pisni podkorpus (50.000 besed): besedila 93 učencev, ki so sodelovali na pisnem izpitu EVA, na katerem so napisali pismo osebi, ki je ne poznajo, zelo neformalno sporočilo prijatelju z že danim besediščem in spis o potovanju z danim začetkom ali koncem;
- kontrolni korpus (17.000 besed): posnetki 16 britanskih učencev iz let 1996–97 (Hasselgren 2001).

EVA se lahko uporablja vzporedno s korpusom govorjene angleščine londonskih najstnikov COLT, kar omogoča primerjave jezika mladih govorcev.

Cilj projekta PELCRA (Polish-English Language Corpus Research and Applications), ki od leta 1997 poteka na poljski Univerzi v Lodžu in britanski Univerzi v Lancastru, je zbrati korpuse sodobne poljščine in angleščine: poljski nacionalni korpus, poljski korpus angleščine kot tujega jezika **LPELC: Łódź Polish English Learner Corpus** ter vzporedni poljsko-angleški in angleško-poljski korpus. LPELC ima tri milijone besed tvorcev, ki so na vseh ravneh od učencev do izpopolnjevalnih univerzitetnih študentov angleščine. Vir pretežno pisnih besedil so bili izpiti ESOL v letih 1998–2002,

teksti pa so bili zbrani delno longitudinalno: vsak tvorec je prispeval do tri besedila, predvidoma vsakega na višji stopnji. V njih so ročno dodane oblikoslovne oznake in oznake napak. Javno dostopnih je le nekaj vzorcev iz esejev (spletna stran). Kot kontrolno gradivo se uporabljajo korpusi BNC, ANC in korpus poljščine domačih govorcev. Prvi korpusni projekt tega tipa na Poljskem pa je bil korpus **PELE: Polish English Learner English** Rafala Uzarja z Univerze v Łodzu, ki je želel z njim študentom pomagati pri pisanju in povečati zanimanje za korpusno jezikoslovje. Zasnovan je bil pilotsko in eksperimentalno, tako da obsega samo 10.000 besed študentov tretjega letnika. Označene so napake. Kot kontrolno gradivo sta vključena vzorec domačih govorcev angleščine in vzorec poljščine kot prvega jezika, vsak s po 3000 besedami (Uzar 1998).

Pisni korpus **JPU: Janus Pannonius University Corpus** je leta 1992 na madžarski Univerzi v Pecsu zgradil József Horváth. Besedila madžarskih študentov – pripovedne in opisne eseje, osebna razmišljanja in seminarske naloge – je dobil v elektronski verziji in jih ni spreminjal. Izjemo je naredil pri napakah, ki so očitno posledica napačnega tipkanja, kot je npr. *langauge* namesto *language* ali *teh* namesto *the* (Horváth 2001: 105). Skupaj ima JPU 400.000 besed, ki niso označene in so na voljo samo v tekstovni obliki (Pravec 2002: 98).

V Španiji nastaja precej anglističnih korpusnih projektov. **ENWIL: English Written Interlanguage** je korpus esejev študentov prvega letnika angleščine. V okviru projekta, ki raziskuje vpliv starosti na usvajanje anglešine, sta nastala **BAF: Barcelona Age Factor** z besedili 2063 dvojezinih, špansko in katalonsko govoreih tvorcev in **BELC: Barcelona English Language Corpus** samo z besedili tvorcev, ki jim je bilo mogoe longitudinalno slediti sedem let. Madridski **MAD Corpus** sestavljajo angleška argumentativna besedila več kot 200 študentov angleščine v prvem in četrtem letniku študija, njihova argumentativna besedila v španščini in eseji domačih govorcev angleščine (Díez Bedmar 2009: 921–923). Na Univerzi v Madridu pod vodstvom Paula Rollinsona od leta 2005 po metodologiji korpusa ICLE nastaja korpus angleščine španskih govorcev **WriCLE: Written Corpus of Learner English**, ki ima zaenkrat 750.000 besed, cilj je en milijon. Tvorci so študentje prvega in tretjega letnika angleščine in so na šestih ravneh jezikovne zmožnosti, označenih po lestvici SEJO. Spisi, shranjeni v goli tekstovni obliki, so dolgi 500–2000

besed (Mendikoetxea idr. 2009). Na isti univerzi od leta 1998 zbirajo tudi besedila za **UAM Corpus de Interlenguas escritas**. Tvorci so španski učenci angleščine na različnih nivojih od osnovne šole do podiplomskih študentov anglistike, vendar je poudarek na dodiplomskih študentih. Za primerjavo so vključena besedila domačih govorcev (Martín Úriz, brez letnice). Pisni in govorni korpus **SULEC: Santiago University Learner of English Corpus** ima pol milijona besed, zaželena velikost pa je milijon (Díez Bedmar 2009: 924).

Med mednarodnimi evropskimi korpusnimi projekti je treba omeniti **CYLIL: Corpus of Young Learner Interlanguage**, ki ima pol milijona besed. Sestavljajo ga govorjena besedila evropskih učencev, ki se učijo angleško, s prvimi jeziki nizozemščino, francoščino, grščino in italijanščino. Vključena je kontrolna skupina osmih domačih govorcev angleščine (Myles 2005: 379). **TED: Translanguage English Database** je zbirka 224 posnetkov predavanj, ki so jih imeli nedomači govorci angleščine leta 1993 na srečanju Eurospeech v Berlinu. Vsaka predstavitev traja petnajst minut, sledi ji pet minut razprave. Skupaj je petinsedemdeset ur govora. Korpus s 188 govori, zapisanimi pripravami nanje in vprašalniki za govorce je dosegljiv na šestih zgoščenkah (spletna stran). **ISLE Corpus of non-native spoken English** je zanimiv, ker ne gre za korpus usvajanja jezika, temveč za korpus govora tujega jezika z 11.484 izjavami. Nastal je kot stranski produkt projekta, katerega cilj je s pomočjo tehnologije za prepoznavanje govora izboljšati računalniško učenje angleščine za odrasle Nemce in Italijane (Atwell idr. 2003: 5). S posnetki 23 nemških in 23 italijanskih odraslih govorcev na nadaljevalni stopnji znanja angleščine, ki berejo odlomek o vzponu na Mount Everest in serijo krajših izjav, v katerih se pokažejo tipične napake pri izgovorjavi (Menzel idr. 2000: 314), so optimizirali parametre za prepoznavanje govora in ocenili sistem za prepoznavanje pričakovane napačne izgovorjave. ISLE ni splošneje uporaben, saj ima označene le napake na fonetični ravni. Za delno ročno, delno avtomatizirano označitev celotnega korpusa so potrebovali približno tristo ur (Atwell idr. 2003: 9). Ujemanje med različnimi označevalci je bilo 55-odstotno (Atwell 2001). Kontrolni korpus sestavljata posnetka dveh domačih govorcev.

V Sloveniji še ni podatkov o večjem korpusu angleščine kot tujega jezika. Januarja 2009 se je začel pilotski projekt **[kju:]** Slovenskega društva učiteljev tujega strokovnega jezika, ki ga vodi Šarolta Godnič Vičič z Univerze na Primorskem, ven-

dar o njegovem napredovanju ni podatkov.[20] Zbirali naj bi pisne izdelke študentov višjih in visokih šol, ki so opravljali test angleščine Dialang. Po zgledu kanadskega korpusa QLC je predvidena tema besedila *Kako drugačno bi bilo vaše življenje ali delo, če bi bila vaša angleščina bistveno boljša kot zdaj* (ang. *What difference would it make to your work or life if your English was significantly better than it is now?*). Besedila naj bi nastala na računalnikih v 45 minutah, dovoljena je raba slovarjev, ne pa tudi črkovalnikov ipd. Šarolta Godnič Vičič je za potrebe svoje raziskave oblikovala tudi oportunistično zbrani pisni korpus **Tourism Students' Essay (TSE)**, v katerem je dobrih 82.000 besed 47 angleških esejev slovenskih študentov prvega letnika turizma, starih 19–21 let. Vsi so se angleščino učili devet let, njihova približno ocenjena stopnja jezikovne zmožnosti pa je bila med B1 in B2. Eseji, dolgo od 1600 do 2000 besed, so bili napisani doma na osnovi vsaj treh v angleščini napisanih člankov o turizmu, študentje pa so lahko uporabljali tudi slovarje in črkovalnike (Godnič Vičič 2008: 19). Podobno je za svojo raziskavo izražanja osebnosti v znanstvenem diskurzu Martin Grad zgradil specializirani angleški korpus znanstvenih člankov s področja medicine s 140.000 besedami. Sestavljata ga podkorpus domačih govorcev angleščine in nekoliko manjši podkorpus člankov slovenskih avtorjev (Grad 2010: 77).

V Severni Ameriki je korpusov angleščine kot drugega ali tujega jezika manj kot v Evropi. Pisni **IBLC: Indianapolis Business Learner Corpus**, oblikovan v obdobju 1990–1998, vključuje besedila s tečaja poslovne komunikacije za domače in tuje univerzitetne študente, katerih prvi jeziki so bili angleščina, finščina, francoščina, nemščina in tajščina. Besedila niso označena in niso javno dostopna (Connor idr. 2002: 189; spletna stran).

Korpus **MAELC: Multimedia Adult English Learner Corpus**, zgrajen v letih 2001–2006 na ameriški univerzi v Portlandu, sestavljajo videoposnetki pouka v razredu skupaj s transkripcijami in s pisnimi besedili, ki nastajajo ob urah. V učilnici je bilo šest kamer, mikrofon pa sta nosila tudi dva študenta v vsaki skupini. Korpus ima posnetke 3600 ur pogovorov v razredu in transkripcije 150 ur govora 250 študentov na nizki stopnji jezikovne zmožnosti. Gradivo je bilo zbrano longitudinalno in ni bilo označeno, program ClassAction pa omogoča analizo in dodajanje lastnih oznak ali transkripcijskih podrobnosti (Reder idr. 2003: 547; spletna stran).

20 Prim. http://www2.arnes.si/~supsgodn/Usvajanje/pilot.html (avgust 2011).

Pisni korpus angleščine kot drugega jezika **MELD: Montclair Electronic Language Database** je nastal na univerzi Montclair v ZDA in obsega 98.000 besed. Tvorci so študentje na izpopolnjevalni ravni, tekste pa so zbrali longitudinalno. Avtomatično so bile dodane oblikoslovne oznake, slovnične napake in napake v akademskem pisanju pa so bile brez uporabe označevalnega jezika ročno označene na vzorcu 50.000 besed (Fitzpatrick, Seegmiller 2004: 223; spletna stran). Na univerzi v Michiganu pa od leta 1997 gradijo **MICASE: Michigan Corpus of Academic Corpus English**, ki ima približno 1,8 milijona besed govorjene akademske angleščine. Med korpuse usvajanja tujega jezika se uvršča le pogojno, saj je v njem 12 %, torej dobrih 200.000 besed tujih tvorcev (spletna stran).

Pisni korpus angleščine kot drugega jezika za domače govorce kanadske francoščine **QLC: Québec Learner Corpus** ima 250.000 besed in je sestavljen iz štirih podkorpusov. Na izpopolnjevalni ravni gre za besedila 400 bodočih učiteljev angleščine z univerzitetnega sprejemnega izpita, na katerem so morali leta 1997 in 1999 napisati esej z naslovom *Kako bi lahko izboljšali učenje angleščine v Quebecu*. Na nadaljevalni ravni pa gre za besedila 1500 tvorcev z naslovom *Kako bi bilo vaše življenje drugačno, če bi znali bolje angleško*, ki so nastala na uvrstitvenem testu, a zaenkrat samo podpirajo raziskave gradiva izpopolnjevalcev. Pri vnosu besedil v korpus so popravili črkovanje (Cobb 2003).

Korpusi angleščine se že dolgo živahno razvijajo tudi v Aziji. Eden najstarejših je od leta 1996 nastajajoči japonski **JEFLL: Japanese English as a Foreign Language Learner**, ki ga koordinira Yukio Tono. Cilj projekta ni splošni korpus angleščine japonskih učencev, temveč uskladitev različnih prostovoljnih projektov učiteljev angleščine. Spodbuda za nastanek je bil projekt na tokijski univerzi, v katerem so zbirali angleške spise študentov in skušali oceniti, kako učiteljeve povratne informacije vplivajo na uspešnost. Pri tem so zbrali dovolj besedil, jih označili in jim začeli dodajati nova. JEFLL ima skupaj 800.000 besed v štirih podkorpusih. V pisnem gradivu učencev angleščine (750.000 besed) so spisi z izpitov razdeljeni na sedem podkorpusov glede na letnik šolanja. Govorjeno gradivo (50.000 besed) sestavlja spontani govor, v katerem učenci opisujejo strip. Pedagoški korpus ciljnega jezika (1.200.000 besed) sestavljajo učbeniki angleščine, ki jih uporabljajo v japonskih šolah (Tono 2003b). JEFLL je v osnovi razvojen in lahko raziskuje razvoj ter spreminja-

nje vmesnega jezika (Pravec 2002: 88). Gradivo na začetni stopnji je nekoliko težje zbirati in obdelovati zaradi večjega števila napak. Težavna je transkripcija govora. Najmlajši učenci, ki so prispevali besedila, so stari dvanajst ali trinajst let, njihova pisna zmožnost je še zelo omejena. Zato so tudi besedila krajša, in sicer od dvajset do sto petdeset besed. Vsako besedilo je označeno s podatki o tvorcu. Avtomatično so bile dodane oblikoslovne, skladenjske in semantične oznake, napake pa so označili ročno v jeziku XML (Pravec 2002: 94–95).

Korpus **SST: Standard Speaking Test Corpus**[21] je od leta 1999 del večjega projekta, katerega cilj je razvoj jezikovne tehnologije, ki naj bi napake v govoru ali pisanju uporabila za pomoč pri pisanju v angleščini, strojnem prevajanju in poučevanju. Gre za govorni korpus usvajanja tujega jezika japonskih govorcev s 1200 posnetki z izpita v skupni velikosti milijon besed (Izumi idr. 2004: 31). Standard Speaking Test, na voljo za 32 različnih jezikov, ima devet stopenj jezikovne zmožnosti, ki so določene glede na vrsto besedila, natančnost, izgovorjavo, fluentnost in splošni namen (Tono 2001: 258), zato naj bi bila prednost tega korpusa, da je stopnja učenčeve jezikovne zmožnosti za vsako besedilo ocenjena strokovno in natančno, ne pa zgolj po občutku ali po letniku šolanja kot pri večini drugih korpusov. V besedilih so označene osnovne značilnosti diskurza in napake, kot kontrolno gradivo pa se uporablja podkorpus domačih govorcev angleščine, ki so odgovarjali na podobna vprašanja kot Japonci, in podkorpus, v katerem so izjave Japoncev v angleščini prevedli nazaj v japonščino ter s tem opazovali posledice jezikovnega prenosa (Tono 2001: 259; Abe, Tono 2005).

Na Japonskem obstaja še več korpusnih projektov, med drugim milijonski pisni korpus **CEJL: Corpus of English by Japanese Learners** Asaa Kojira, ki je začel nastajati leta 1995 in zajema spise, obnove in pisma tvorcev na vseh stopnjah znanja. V njem so delno označene napake, besedila pa so v goli tekstovni obliki dostopna prek interneta (spletna stran). **LBLC: Learner Business Letters Corpus** je sestavljen iz dobrih dvesto tisoč besed oziroma 1464 pisem, ki so jih napisali japonski poslovneži (Nesselhauf 2004: 129). Vključen je v večji korpus Business Letter Corpus, ki ima tudi besedila domačih govorcev angleščine: poslovna pisma (1.020.060 besed), osebna pisma (113.522 besed), osebna

21 Uporabljata se tudi imeni korpus TAO (TeleNexommunications Advancement Organization of Japan) ali korpus NICT JLE (Japanese Learner English).

pisma pomembnih oseb iz 19. stoletja, literaturo in filmske scenarije ter govore predsednikov ZDA (spletna stran). **Japanese Learners' Corpus** vsebuje 4000 povedi in ima označene napake (Sugiura 2000). **Hiroshima English Learners' Corpus** oziroma **Japanese/English Translation Corpus** Shoga Miure vsebuje besedila Japoncev v goli tekstovni obliki (Tono 2003a: 803). Delno gre tudi za vzporedni korpus, saj so učenci morali v angleščino prevesti stavke iz učbenika, prevodi in izvirniki pa so združeni v korpus.

12-milijonski pisni **PolyU Corpus** združuje različne korpusne projekte kitajskih govorcev; nekateri so nastali na univerzi Hong Kong Polytechnic University, drugi pa so del komercialnih korpusov, npr. BNC. V okviru vzporednih ali primerljivih korpusov so besedila tudi v kitajščini, japonščini in francoščini. Večinoma so jih napisali domači govorci angleščine, približno 2,5 milijona besed pa učeči se. Njihovi podkorpusi so (spletna stran; Lin 1999):

- Corpus of Business Correspondence (30.900 besed): poslovna pisma, dopisi in poročila podjetij iz Velike Britanije in Hongkonga. V korpusu ni označeno, ali gre za domače ali tuje govorce angleščine;
- Chinese Learner English Corpus (CLEC, 1.000.000 besed): pisni korpus angleščine kitajskih govorcev, in sicer srednješolcev ter študentov na petih, med seboj uravnoteženih stopnjah jezikovne zmožnosti. Vključuje pisanje iz testov, vodeno in prosto pisanje. V besedilih so z 61 tipi oznak označene napake;
- Learner Corpus of English for Business Communication (117.500 besed): poslovna besedila hongkonških študentov pri predmetu Poslovna angleščina, zbrana v letih 1999–2002;
- Learner Corpus of Essays and Reports (188.000 besed): spisi in projektna poročila o različnih temah (znanost, informacijska tehnologija, zdravstvena nega, ekonomija, družbene vede). Študentje različnih smeri so jih leta 2000 pisali pri predmetu Akademska angleščina;
- Learner Journals (165.000 besed): dnevniki, ki so jih v pisali študentje ob branju zgodbe v angleščini;
- PolyU Learner English Corpus (PLEC, 1.000.000 besed): kratki spisi, ki so jih na izpitih v angleščini napisali študentje različnih smeri.

Na Kitajskem obstaja še nekaj korpusov, o katerih je na voljo manj podatkov: govorni korpus **COLSEC: College English Learners' Spoken English Corpus**, ki ima 700.000 be-

sed, in **SWECCL: Spoken and Written English Corpus of Chinese Learners** (Quifang 2006). Del 1,3-milijonskega kitajskega korpusa Corpus of English Education je **MET Corpus** s 400.000 besedami. V njem so neoznačena angleška pisna besedila kitajskih srednješolcev (Tono 2003a: 803). **Corpus of Middle School English Education** od leta 1998 zbirajo na South China Normal University. Sestavljen je iz pisnega in govornega dela (Shih 2000: 91).

Pisni **TeleNex Student Corpus**, ki od leta 1994 nastaja na univerzi v Hongkongu pod vodstvom Quentina Allana, je del zbirke TeleCorpus, ki jo poleg tega sestavlja še TeleNex Corpus of Modern English. Korpus domačih govorcev kitajščine (kantonščine) na nadaljevalni stopnji angleščine ima tri milijone besed in vključuje besedila z izpitov in s tečajev. Zbirko, ki ni označena, učitelji centra TELEC (Teachers of English Language Education Centre) uporabljajo pri pripravi učnih gradiv. Korpus želijo dopolniti z govorno komponento in ga razširiti do obsega 10 milijonov besed (Allan 2002: 195; spletna stran). Pisni korpus angleščine **CEEAUS: The Corpus of English Essays Written by Asian University Students** pa vsebuje besedila japonskih, kitajskih in angleških študentov in ima 200.000 besed (spletna stran).

Na Univerzi v Hongkongu pod vodstvom Johna Miltona nastajata dva korpusa. 25-milijonski korpus **HKUST: Hong Kong University of Science and Technology Corpus of Learner English** je pisni. Tvorci so večinoma domači govorci kantonščine, in sicer srednješolci na nadaljevalni stopnji jezikovne zmožnosti v angleščini. Besedila so v glavnem nastala na izpitih, del pa tudi izven razreda. Označena so z oblikoslovnimi oznakami, na petih milijonih besed pa so ročno označene tudi napake (spletna stran). Govorna različica HKUST-a pa je **HKCSE: Hong Kong Corpus of Conversational English** z 900.000 besedami. Nastal je kot protiutež hongkonškemu podkorpusu ICLE, ki vsebuje transkripcije zasebnih pogovorov, vendar so zaradi omejitev pri snemanju precej nenaravni. HKCSE ima petdeset ur pogovorov, in sicer 130 spontanih pogovorov s 341 udeleženci. Vsak traja povprečno triindvajset minut. V pogovorih so sodelovali tako domači kot tuji govorci angleščine, zato je korpus primeren za medkulturne študije (Cheng, Warren 1999: 6–16).

Na Tajvanu so med letoma 1984 in 1995 izdelovali Soochow Colber Student Corpus s 227.000 besedami iz zapisanih tekstov tajvanskih študentov. Ta korpus so vključili v pisni korpus **TCLE: Taiwanese Corpus of Learner English**

s 730.000 besedami (zaželena končna velikost je milijon besed). Besedila univerzitetnih študentov angleščine na nadaljevalni in izpopolnjevalni ravni so v glavnem dnevniki in eseji, v njih pa so bile avtomatsko označene leme, oblikoslovne, skladenjske in semantične oznake. Več kot polovico tekstov so dobili v elektronski obliki, del so optično prebrali in ga naknadno pregledali, del pa so pretipkali. Pri tem so odstranili napake, ki so posledica napačnega zapisa, saj bi to oviralo označevanje (Shih 2000: 91–93). Še en tajvanski korpus je **NTOU**, ki vsebuje 53.000 besed. Besedila so v tekstovni obliki (Tono 2003a: 803).

Na Tajskem od leta 1997 gradijo pisni korpus **TELC: Thai English Learner Corpus**, ki ima 1,5 milijonov besed. Tvorci so univerzitetni študentje, domači govorci tajščine. Polovica besedil je s sprejemnih izpitov, polovica pa iz spisov v četrtem letniku. V njih so označene leme in oblikoslovne oznake (spletna stran). V Maleziji ima pisni korpus **CALES: The Corpus Archive of Learner English in Sabah/Sarawak** 400.000 besed iz esejev (Botley idr. 2007). V zadnjem času se pojavljajo tudi korpusi za arabščino, denimo pisni korpus angleščine arabskih govorcev **BALC: The BUid Arab Learner Corpus** z 287.227 besedami (Randall, Groom 2009).

2.4.2 Korpusi francoščine kot tujega jezika

Pisni korpus **FRIDA: French Interlanguage Database** je vzporednica korpusa ICLE s francoščino kot ciljnim jezikom. Od leta 1998 nastaja na belgijski univerzi v Louvainu-la-Neuve v sodelovanju s številnimi drugimi univerzami po vsem svetu. Korpus ima 450.000 besed tvorcev na nadaljevalni stopnji jezikovne zmožnosti v francoščini in tri podkorpuse: domačih govorcev angleščine, nizozemščine in ostalih prvih jezikov. V FRIDI so označeni podatki o tvorcu in napake po podobnem sistemu kot v korpusu ICLE, vendar z označevalnim jezikom XML (Granger 2003b: 466).

Pri projektu **FLLOC: French Learner Language Oral Corpora**, ki ga koordinirata Florence Myles in Rosamond Mitchell z Univerze v Southamptonu in Univerze Newcastle v Veliki Britaniji, ne gre za enoten korpus, ampak za različne govorne korpuse francoščine kot tujega jezika učencev britanskih šol:

– Young Learners Corpus vsebuje besedila najmlajših učencev, starih 5, 7 in 11 let. Za vsako starost vključuje 20 posnetkov učencev, ki prej niso znali francosko;

– Linguistic Development Corpus je bil zbran v letih 2001/2002. Vključenih je po 20 učencev, starih 13–15 let, po dveh do štirih letih učenja francoščine. Vsak je v enournem pogovoru z raziskovalcem opravil štiri naloge: opisovanje ilustracij, postavljanje vprašanj, pogovor o fotografijah, opisovanje znanih oseb;
– longitudinalni Progression Corpus je nastal v letih 1993–1996, v katerih so pri pouku francoščine dve leti spremljali 60 otrok. Gre za približno 200 ur posnetkov;
– Salford Corpus: longitudinalna raziskava 12 študentov francoščine na britanski univerzi;
– Brussels Corpus: raziskava simultanega učenja dveh tujih jezikov, francoščine in angleščine, v flamsko govorečih srednjih šolah v Belgiji. Vključenih je 150 dijakov, pri katerih so preverjali jezikovno zmožnost (govorjenje, branje, pisanje in metajezikovno znanje) v obeh jezikih, pa tudi stališča do obeh jezikov;
– longitudinalni Newcastle Corpus vsebuje gradivo 30 angleških dijakov, starih 16 in 17 let, po petih oz. šestih letih učenja francoščine;
– Reading Corpus vključuje transkripcije ustnih izpitov, ki jih opravljajo britanski dijaki pri 16 letih, in primerljiva besedila domačih govorcev francoščine;
– University of East Anglia (UEA) Corpus je bil zbran v letih 2002–2004. V simulaciji televizijske debate udeleženci, študentje francoščine oziroma angleščine kot tujega jezika, izražajo nasprotujoča si mnenja v svojem prvem in tujem jeziku (angleščini in francoščini).

Skupaj gre za 4000 datotek. Gradivo, združeno na eni spletni strani, je ortografsko prepisano po konvencijah CHILDES in oblikoslovno označeno (spletna stran, Myles 2007: 396).

Govorni korpus francoščine **InterFra** nastaja na univerzi v Stockholmu na Švedskem pod vodstvom Inge Bartning. Vključuje 98 intervjujev z dijaki, dodiplomskimi in podiplomskimi študenti, pa tudi z nekaj domačimi govorci. Za večino je francoščina tuji jezik, za nekatere drugi, vsi pa prostovoljno sodelujejo v raziskavi (spletna stran).

Pri korpusu **Lancom**, ki vključuje interakcijo v francoščini med igro vlog francosko in nizozemsko govorečih, sta vodilna Mark Debrock iz Leuvena v Belgiji in Danièle Flament-Boistrancourt iz Pariza. Trenutno ima 39 ur posnetkov. Transkribiranih in označenih je 18 ur oziroma približno 160.000 besed (spletna stran).

2.4.3 Korpusi nemščine kot tujega jezika

Pisni korpus nemščine kot tujega jezika **FALKO: Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache**, ki ima 1,2 milijona besed (Hana idr. 2010: 12), nastaja na Humboldtovi univerzi v Berlinu. Sestavljen je iz povzetkov znanstvenih člankov, ki so jih napisali univerzitetni študentje, in celotnih člankov, ki so jih povzemali, esejev in longitudinalnega dela, ki ga zbirajo v ZDA. Tvorci so na izpopolnjevalni stopnji jezikovne zmožnosti (C1 ali C2 po lestvici SEJO) in imajo različne prve jezike. V korpusu so označene leme in oblikoslovne oznake, na delu besedil pa tudi skladenjske oznake in napake. Vsi podkorpusi imajo kontrolni korpus domačih govorcev nemščine (Lüdeling idr. 2005; Siemen idr. 2006; spletna stran).

EAGLE: Error-Annotated German Learner Corpus je pisni korpus nemščine ameriških študentov na začetni stopnji, ki nastaja na Ohio State University. Sestavljen je iz dveh podkorpusov: spletnih odgovorov na različna vprašanja iz učnega gradiva (vključeni so samo odgovori, v katerih študenti uporabljajo cele povedi; skupaj okoli 33.000 besed) in spisov z izpitov (okoli 12.000 besed). V korpusu so označene napake (Boyd 2008: 1897).

ZISA: Zweitspracherwerb italienischer und spanischer Arbeiter je zbirka besedil italijanskih, španskih in portugalskih delavcev in delavk v Nemčiji. Sestavljena je bila delno presečno s 45 osebami in longitudinalno z 12 osebami, njen cilj pa je bil odkriti elemente, ki pospešujejo ali zavirajo usvajanje nemščine (Frischherz 1997: 47). Vendar ni jasno, ali gre za pravi jezikovni korpus. Podobno velja za **Saarbrücker DFG-Forschungsprojekt »Gastarbeiterkommunikation«**, ki analizira govor desetih turških in desetih italijanskih otrok (navedeno delo: 49). V Švici pa je Bruno Frischherz za svojo raziskavo nemščine priseljenih delavcev izdelal korpus z 21 intervjuji s turškimi in kurdskimi delavci. Korpus ima 43.000 besed, od tega je 26.000 besed informatorjev (Frischherz 1997: 89).

2.4.4 Korpusi italijanščine kot tujega jezika

Za italijanščino kot tuji jezik obstaja polmilijonski pisni in govorni korpus **VALICO: Varietà di Apprendimento della Lingua Italiana: Corpus Online**, ki je oblikoslovno označen in naj bi imel tudi označene napake. Med drugim naj bi po-

dal pregled besedilnih tipov, ki jih pišejo učenci italijanščine po svetu, postal pa naj bi del večjega korpusa sodobne italijanščine (spletna stran).

Na univerzi v Perugii od leta 2001 v okviru projekta Osservatorio, ki se osredotoča na italijanščino kot tuji jezik in identiteto migrantov, pod vodstvom Stefanie Spine, Silvia Pazzaglia in Mirca Perinija nastaja govorni korpus italijanščine kot tujega jezika **Corpus parlato di italiano L2**. Sestavljajo ga ortografski prepisi intervjujev, v katerih so označeni elementi govorjenega jezika in nekatere slovnične značilnosti (spletna stran).

Na univerzi v Stockholmu od leta 2001 gradijo govorni korpus usvajanja italijanščine **InterIta**. Švedski študentje odgovarjajo na vprašanja v intervjuju, kot kontrolna skupina pa so vključeni izmenjavni študentje iz Italije. Korpus je oblikoslovno označen (spletna stran).

2.4.5 Korpusi španščine kot tujega jezika

CEDEL2: Corpus Escrito del Español L2 na univerzi v Madridu gradi Cristóbal Lozano. Gre za korpus španščine domačih govorcev angleščine, ki ima trenutno 730.000 besed 2300 govorcev, od tega je 660 domačih govorcev španščine. Želena velikost korpusa je milijon besed. Tvorci so na vseh stopnjah jezikovnega znanja, besedila pa so dolga vsaj 500 besed. Vsak tvorec v zameno za sodelovanje dobi povratno informacijo o svoji jezikovni zmožnosti (spletna stran).

V Veliki Britaniji od leta 2006 gradijo projekt **SPLLOC: Spanish Learner Language Oral Corpora**, soroden projektu FLLOC (prim. poglavje 2.4.2). Zaenkrat je sestavljen iz dveh govornih korpusov. Vsak vključuje posnetke 60 učencev na začetni, nadaljevalni in izpopolnjevalni ravni, za katere je španščina tuji jezik, in 15 domačih govorcev (spletna stran).

Na Tajvanu pa od leta 2005 zbirajo besedila za **CATE: Corpus de Aprendices Taiwaneses de Español**, ki nastaja v okviru projekta Multilingual Corpora na Nacionalni univerzi Cheng Kung v Tajvanu. Korpus, dostopen prek spletnega vmesnika, vključuje tekste tajvanskih študentov s 15 univerz, za katere je angleščina prvi, španščina pa drugi tuji jezik. V prvih štirih letih so zbrali 1913 spisov oziroma 337.122 besed, ki jih spremljajo podatki o avtorju in besedilu. V podkorpusu s 181.322 besedami pa so v jeziku XML označene napake (Lu 2010: 297).

2.4.6 Korpusa švedščine kot tujega jezika

Pisni in govorni korpus švedščine **ASU: Andrespråkets strukturutveckling** je v letih 1990–1993 in 1998 oblikoval Björn Hammarberg z Univerze v Stockholmu. Gre za longitudinalni korpus s posnetki in besedili desetih tujih študentov od začetne stopnje do uspešno opravljenega izpita, potrebnega za študij na švedskih univerzah. Obsega 490.000 besed, od tega 415.000 govornih in 75.000 pisnih. Besedila so v tekstovni obliki, v pripravi pa naj bi bila tudi verzija XML. Kontrolno gradivo predstavlja 97.000 govornih in 25.000 pisnih besed sedmih domačih govorcev (spletna stran).

Pisni korpus **SVANTE: Svenska Andraspråks-texter** ali **CrossCheck Learner Corpus** ima skupaj 220.000 besed in že v zasnovi ni uravnotežen, saj vanj vključujejo ves material, ki ga dobijo:

- SSM (Svenska som Målspråk, 112.000 besed): spisi, ki so jih v letih 1972–1975 napisali odrasli z 10 različnimi prvimi jeziki;
- SFI (Swedish for Immigrants, 75.000 besed): spise je leta 1997 napisalo 275 odraslih, ki so se udeležili programa Swedish for Immigrants;
- Granska (35.000 besed): gradivo, zbrano na univerzi v Stockholmu, sestavlja 65 besedil, ki jih je napisalo 10 učečih se. Nekatera so na precej visoki ravni, npr. protokoli ali spisi;
- Argus (120.000 besed): 287 argumentativnih spisov, ki so jih napisali domači govorci švedščine na univerzi v Uppsali;
- SvSFi: kontrolni korpus desetih domačih govorcev švedščine.

Tudi pri metajezikovnih podatkih vključijo v korpus vse, kar dobijo. Korpusu naj bi priključili gradivo spisov z izpita iz znanja švedščine TISUS (Volodina 2008: 104, 205; Bigert idr. 2005).

2.4.7 Korpusi ostalih jezikov

Polmilijonski pisni korpus norveščine (bokmål) kot tujega jezika **ASK: Norsk andrespråkskorpus** je bil zgrajen v letih 2003–2006 na Univerzi v Bergnu pod vodstvom Kari Tenfjord. Besedila so zbrali na izpitu iz znanja norveščine, njihovi tvorci pa so bili priseljenci, ki so obiskovali brezplačni tečaj norveščine in opravljali izpit na nadaljevalni (B1) ali

izpopolnjevalni (B2) ravni. Teksti so uravnoteženi po prvih jezikih tvorcev: albanski, angleški, bosanski oziroma hrvaški oziroma srbski, nemški, nizozemski, somalski, španski, poljski, ruski, vietnamski. Kontrolno gradivo sestavljajo besedila domačih govorcev, ki so reševali enak izpit kot tuji govorci (sto na ravni B1 in sto na ravni B2). V ASK-u so s programom Oslo-Bergen Tagger dodali leme, oblikoslovne in skladenjske oznake. Skladnja je bila v celoti označena avtomatično, zato oznake niso nujno pravilne, medtem ko so bile oblikoslovne oznake ročno pregledane in razdvoumljene. Ročno so bile v jeziku XML označene tudi napake. Poleg prikaza konkordanc je možno vzporedno videti besedilo z napakami in popravljeno besedilo, kar naj bi olajšalo tako označevanje kot delo uporabnikom, hkrati pa se s tem preverja pravilnost označevanja (spletna stran; Tenfjord idr. 2006).

Pisni **EIC: Estonian Interlanguage Corpus** je korpus estonščine kot drugega in tujega jezika, ki ga razvijajo v okviru korpusnega projekta VAKO na Univerzi v Talinu in znotraj estonskega Nacionalnega programa za razvoj jezikovnih tehnologij.[22] EIC je dinamičen, saj besedila tvorcev na več stopnjah jezikovne zmožnosti stalno dodajajo. Trenutno ima 700.000 besed, največ je besedil rusko govorečih, povečati pa ga nameravajo na pet milijonov besed. Ima lasten konkordančnik, torej programsko orodje za prikazovanje konkordanc, in statistični modul, ki med drugim išče po lemah in napakah. Te so ročno označene na zaenkrat 500.000 besedah in natančno kategorizirane. Vsako besedilo je opremljeno z metapodatki o tvorcu (starost, spol, materni jezik, jezik, ki ga govori doma, itn.) in besedilu (število besed, tip besedila, število napak, tip napak). Spletni vmesnik ima različne možnosti dostopa za različne tipe uporabnikov, vendar je večina funkcij na voljo vsem (spletna stran).

Na Univerzi v Ouluju na Finskem v letih 2008–2011 nastaja korpus finščine **ICLFI: International Corpus of Learner Finnish**. Zaenkrat ima 600.000 besed iz raznolikih besedil, ki jih zbirajo na univerzah po svetu, kjer poučujejo finščino. Vključenih je 22 izhodiščnih jezikov: angleški, češki, estonski, islandski, italijanski, kitajski, nemški, nizozemski, poljski, ruski, španski, švedski itn. Govorci so na vseh stopnjah znanja, korpus pa bo osnova za raziskave finščine kot vmesnega jezika (Jantunen 2008, Jantunen 2010).

Projekt **JASMIN: Jongeren, Anderstaligen en Senioren in Mens-machine Interactie** je potekal v letih 2005–2007

22 Prim. http://www.keeletehnoloogia.ee/ (november 2008).

na Univerzi v Nijmegenu in Katoliški univerzi v Leuvenu. Kot razširitev 9-milijonskega govornega korpusa nizozemščine CGN (Corpus Gesproken Nederlands) je nastal tudi govorni korpus nizozemščine z dobrimi 111 urami posnetkov govorcev, izključenih iz korpusa CGN: otrok, starejših in nedomačih govorcev (s prvimi jeziki arabščino, francoščino, turščino idr.) na ravni A1 do B1 po lestvici SEJO. Posnetek vsakega govorca traja približno dvanajst minut. Od tega je polovica branja fonetično raznolikih stavkov in besedil iz učbenikov, ustreznih govorčevi jezikovni zmožnosti, polovica pa dialoga med strojem in človekom. Dialogi so oblikovani tako, da pri govorcih izzovejo določene pojave, na primer pretirano poudarjeno izgovarjanje, podaljševanje zlogov, kričanje, ponavljanje, parafraziranje. Dve tretjini materiala sta bili zbrani na nizozemskem, tretjina pa na flamskem govornem področju. Korpus je oblikoslovno označen (Cucchiarini idr. 2008).

Korpus **AKCES/CzeSL: Akvizični korpusy četiny/Czech as a second language** je dvomilijonski pisni in govorni korpus tujih govorcev češčine, ki naj bi do leta 2012 nastal na Karlovi univerzi v Pragi in Tehnični univerzi v Libercu. Sestavljajo ga štirje podkorpusi glede na prvi jezik govorcev: ruski, vietnamski, romski in preostali. Govorci so na vseh ravneh jezikovne zmožnosti po lestvici SEJO, besedila pa nastajajo v razredu ali na izpitih. Posebej pomembno je označevanje napak (Hana idr. 2010: 12–13).

Pri projektu **ESFSLD: European Science Foundation Second Language Databank** v letih 1982–1986 je šlo za longitudinalno zbiranje govornih besedil angleščine, francoščine, nemščine, nizozemščine in švedščine kot drugega jezika. Tvorci so bili priseljeni delavci, moški med 18. in 30. letom, ki se ciljnega jezika niso formalno učili in so v državo prišli pred kratkim. Vsak ciljni jezik ima dva izhodiščna jezika – arabščino, finščino, italijanščino, pundžabščino, španščino in turščino –, izbrana tako, da je njunih govorcev veliko, hkrati pa omogočata jezikoslovno najzanimivejše primerjave. Podatke 40 tvorcev so zbirali več kot dve leti in pol v štiri- do šesttedenskih intervalih. Uporabili so igro vlog, videoposnetke, obisk v spremstvu raziskovalca in samosoočenje, kjer so si informatorji ogledali svoj posnetek. Informatorji so morali poustvariti prizor, ki so si ga ogledali, dati navodila za risanje slike, pripovedovati o vsakdanjih dogodkih in o odlomku iz filma. Kontrolno gradivo sestavljajo pogovori s priseljenci, ki ne znajo jezika, priseljenci, ki že dolgo živijo v tuji državi, in domačimi govorci (Allwood idr. 1983).

V Baskiji razvijajo **korpus baskovščine kot tujega jezika**, katerega osnovni namen je razviti avtomatsko računalniško popravljanje napak. Te so razvrščene v sedem kategorij: zapis, oblikoslovje, skladnja, oblikoskladnja, semantika, slog, ločila in ostalo (Aldabe idr. 2005; Uria Garin 2006).

V Grčiji v okviru projekta Pythagoras nastaja **korpus grščine kot tujega jezika**. Sestavljen naj bi bil iz besedil izmenjavnih študentov v programu Erasmus, ki se morajo pred začetkom študija udeležiti tečaja in opraviti izpit. Na učnem korpusu z 18.000 besedami so preizkusili sistem za označevanje napak in prepoznali najpogostejše tipe (Tzimokas 2006). V Braziliji kot del projekta COMET Multilingual Corpus for Teaching and Translation nastaja korpus **USP Multilingual Learner Corpus (MLC)**. Ciljni jeziki so angleščina, nemščina in španščina. MLC naj bi bil sestavljen iz že zbranih manjših korpusov, v angleški del naj bi bila vključena brazilska komponenta korpusa ICLE. Temu naj bi dodali besedila z univerzitetnih tečajev tujih jezikov, s čimer naj bi na semester zbrali od 200 do 1000 besedil za posamezni ciljni jezik (Tagnin 2003: 941–943).

Na univerzi Montclair v ZDA je nastal pisni **pilotski korpus arabščine kot tujega jezika**. Zaenkrat ima 9000 besed, sestavlja pa ga osem besedil domačih govorcev angleščine, ki znajo arabsko na nadaljevalni in izpopolnjevalni ravni. Cilj projekta je bil poskus aplikacije sistema za označevanje napak iz korpusa FRIDA na arabščino (Abuhakema idr. 2008: 1348).

2.4.8 Korpusi slovenščine kot tujega jezika

Do nedavnega so bili vsi korpusni projekti, povezani s slovenščino kot tujim jezikom, miniaturni, *ad hoc* zbrani korpusi, v katerih so raziskovalci za potrebe specifičnih analiz na hitro združili nekaj besedil, ki so jim bila najlažje dosegljiva. Zametki korpusnega raziskovanja so se začeli konec devetdesetih let (prim. Pirih Svetina 2005). T. Balažic Bulc je za svojo raziskavo jezikovnega prenosa iz srbohrvaščine v slovenščino sestavila korpus 138 spisov z izpita iz aktivnega znanja slovenščine, ki so ga kandidati opravljali na Centru za slovenščino kot drugi/tuji jezik v letih 1998–2000. V njem so označene napake, vendar korpus ni označen po splošno veljavnih kriterijih in zato ni širše uporaben (Balažic Bulc 2004: 79). K. Rozman je v raziskavi o napakah pri učenju slovenščine zbrala korpus s 1451 besedami oziroma 12

besedili. V njem so nemško govoreči študentje slovenščine z univerze v Gradcu brez uporabe slovarjev ali drugih pripomočkov napisali krajši pisni sestavek na temo Jaz in moja družina. V drugi fazi so popravljali pretipkane in nepodpisane sestavke svojih kolegov, nazadnje pa so popravili še svoja pretipkana besedila (Rozman 2006: 6). A. Briški je v okviru raziskave o jezikovni zmožnosti v slovenščini kot drugem jeziku v osnovni šoli analizirala govorne posnetke devetih srbsko, bosansko oziroma črnogorsko govorečih učencev, ki so se predstavljali, opisovali slike in se z izpraševalcem pogovarjali o izbrani temi (Briški 2010: 32). A. Kure pa je s pomočjo jezikovnega testa ocenjevala napredek 17 otrok, ki so se udeležili 30-urnega tečaja slovenščine (prim. Kure 2010).

Sama sem učni korpus usvajanja slovenščine kot tujega jezika (**učni KUST**) oblikovala spomladi 2005. Zasnovan je bil pragmatično in oportunistično, saj je bil njegov izključni namen pridobivanje izkušenj za oblikovanje večje in uporabnejše poskusne različice KUST-a. Kot pisni korpus je vseboval zgolj 600 besed in zato ni mogel biti osnova za nobeno nadaljnjo analizo, temveč samo za preučitev možnosti urejanja korpusa in označevanja napak. Sestavljen je iz petih besedil na temo *Moj teden*. Za domačo nalogo so jih napisali udeleženci tečaja slovenščine za tujce, na katerem sem poučevala. Besedila niso bila napisana z namenom vključitve v korpus, za to sem se odločila kasneje, ker so mi bila dostopna v času raziskave. Edino navodilo, ki so ga tvorci dobili, je bilo, da v dnevniškem slogu opišejo pretekli teden. Drugih omejitev niso imeli. Tvorci so bili iz treh različnih jezikovnih okolij (francoščina, španščina, angleščina) na stopnji višjih začetnikov oziroma nižjih nadaljevalcev. Stopnja jezikovne zmožnosti je bila ocenjena glede na rezultate na razvrstitvenem testu pred začetkom tečaja in glede na učbenik za nadaljevalce *S slovenščino nimam težav*, ki so ga uporabljali pri pouku.

Učni KUST je bil ročno urejen in označen v jeziku XML. Korpus ni bil analiziran, saj to zaradi njegove majhnosti ne bi bilo smiselno. Z njim pa sem dobila dovolj izkušenj za začetek gradnje poskusnega korpusa usvajanja slovenščine kot tujega jezika (v nadaljevanju **PiKUST**). Tudi ta je bil zasnovan oportunistično, da bi pred načrtovanjem pravega KUST-a določila in preizkusila kriterije za zbiranje in označevanje besedil. Čeprav ni bil skrbno načrtovan, je uporaben za vzorce raziskav pojavov z visoko frekvenco pojavljanja ali za postavljanje hipotez, ki jih bo mogoče preverjati na večjem KUST-u. Poleg tega ilustrira možnosti, ki jih bo nudil bolj

uravnoteženi korpus usvajanja slovenščine. Besedila sem označila v okviru študijskega bivanja na Univerzi v Bergnu na Norveškem od oktobra 2006 do januarja 2007. Pomagali so mi strokovnjaki z Oddelka za kulturo, jezik in informacijsko tehnologijo,[23] ki so sodelovali pri korpusu norveščine kot tujega jezika ASK. Prav zaradi podrobnega vpogleda v ozadje ASK-a je bil norveški korpus relevanten zgled pri nastajanju PiKUST-a. Podrobnosti o oblikovanju in značilnostih PiKUST-a so razložene v poglavju 3.4, v njem pa so bile tudi ročno označene napake (prim. poglavje 5.3).

V zadnjem času sta znotraj projekta Sporazumevanje v slovenskem jeziku nastala še dva zanimiva korpusna projekta. **Korpus usvajanja slovenščine** je pisni korpus s 306 besedili oziroma 32.117 besedami. V dobrih 77 % gre za spise iz testov na tečaju slovenščine, slaba četrtina pa jih je iz uvrstitvenih testov. Skoraj polovica tvorcev, ki so bili večinoma stari med 13 in 19 let, je bila na izpopolnjevalni stopnji jezikovne zmožnosti, 37 % je nadaljevalcev, 14 % pa začetnikov. Prihajajo iz 13 držav, njihovi prvi jeziki pa so navedeni v tabeli 3 (Rozman, Stritar 2010: 35–41). Korpus ni jezikoslovno označen, v obliki XML pa je dosegljiv na spletni strani projekta.[24]

Jezik	Število besedil	Število besed	Delež
Angleščina	46	3609	11,2 %
Bosanščina	8	865	2,7 %
Francoščina	5	296	0,9 %
Hrvaščina	2	260	0,8 %
Italijanščina	76	9590	29,9 %
Madžarščina	3	152	0,5 %
Makedonščina	12	1193	3,7 %
Nemščina	19	1392	4,3 %
Nizozemščina	6	447	1,4 %
Slovenščina, italijanščina	6	570	1,8 %
Slovenščina, nemščina	4	495	1,5 %
Srbščina	25	2701	8,4 %
Španščina	90	10170	31,7 %
Ni podatka	4	377	1,2 %

Tabela 3: Prvi jeziki tvorcev v korpusu usvajanja slovenščine po državi izvora tvorcev (Rozman, Stritar 2010: 41)

23 Avdeling for kultur, språk og informasjonsteknologi (AKSIS); osrednja sta bila Kari Tenfjord in Paul Meurer.

24 Prim. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/KUS.zip (september 2011).

Korpus usvajanja slovenščine je nastal kot stranski proizvod gradnje Šolarja, skoraj milijonskega korpusa besedil, ki so jih »učenci slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku« (Rozman, Stritar, Kosem 2010: 4). Besedila pretežno domačih govorcev slovenščine, zbrana v šolskem letu 2009/2010, predstavljajo dejansko šolsko produkcijo, v njih pa so označeni jezikovni popravki, ki so jih v tekstih naredili učitelji (prim. poglavje 5.3). Tudi Šolar je v celoti dosegljiv na spletu.[25]

25 Prim. http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/SOLAR.zip (september 2011).

3 OBLIKOVANJE KORPUSOV USVAJANJA TUJEGA JEZIKA

Pri načrtovanju vsakega korpusa so koristno izhodišče primerljivi že obstoječi korpusi za druge jezike. V razvoju korpusne lingvistike vodijo angleško jezikoslovje in države z močno anglistično tradicijo, kot sta Belgija ali Švedska (Gorjanc 2003: 21). Čeprav je po mnenju nekaterih premoč angleščine škodljiva, ker v njej poteka tudi metajezikoslovno razpravljanje in zato ugledni znanstveniki niso nujno tisti, ki so najbolj nadarjeni za jezikoslovje, temveč tisti, ki so najboljši v angleščini (Uzar 1998), bi bilo nesmiselno zanemariti izkušnje in rešitve tujih, pa čeprav angleških korpusov. Zato so v nadaljevanju navedeni relevantni podatki za obstoječe korpuse usvajanja tujega jezika (podrobnosti o njih prim. v poglavju 2.4). Seveda pa neposredne vzporednice med tujimi korpusi in slovenskim gradivom niso vedno mogoče, saj se angleščina in slovenščina razlikujeta tako v temeljnih jezikoslovnih kot tudi sociolingvističnih vprašanjih.

Načrt gradnje vsakega korpusa lahko povzamemo v naslednjih fazah (Atkins idr. 1992: 2):

- specifikacija in oblika korpusa;
- strojna in programska oprema;
- zajem besedil in označevanje dokumentov;
- procesiranje zbranega gradiva;
- končna oblikovanost korpusa in povratne informacije o njem.

V tem poglavju sta predstavljeni dve fazi oblikovanja korpusov usvajanja tujega jezika, ki ju ilustrira primer oblikovanja PiKUST-a. Strojna in programska oprema, procesiranje zbranega gradiva ter končna oblikovanost korpusa pa so

manj jezikoslovna in bolj tehnološka vprašanja, zato so tu puščeni ob strani.

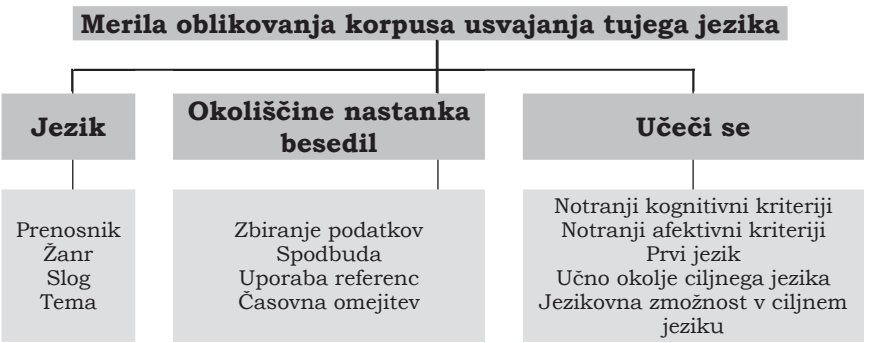
3.1 Specifikacija in oblika korpusa

Nabori besedilnih lastnosti, na katere je treba biti pozoren pri specifikaciji korpusov, so že precej izoblikovani, posebej za referenčne in splošne korpuse. Mednje spadajo (Gorjanc 2005: 26–27):

- prenosnik (pisni, govorjeni);
- avtorskost (enoavtorski, večavtorski, skupni, korporacija);
- število avtorjev;
- ustroj dokumenta (enobesedilni, večbesedilni, npr. časopis, revija, zbirka pesmi);
- besedilna predpriprava (pripravljen, na podlagi zapiskov, spontano);
- medij (natisnjen, rokopis);
- besedilna samooznaka (pesem, roman, kratka zgodba itn.);
- besedilna vrsta in govorni položaj (odprt niz – jezikovno in kulturno najbolj specifična lastnost);
- stvarnost (stvarna literatura, fikcija);
- besedilni kontekst (izobraževanje, dom, delo, prosti čas itn.);
- jezikovna funkcija (glede na vplivanjsko vlogo);
- predmetno področje (glede na dogovorjeni nabor);
- strokovnost (strokovni, nestrokovni, poljudni);
- datum izida;
- besedilni status (original, ponatis);
- jezik;
- jezikovni status (original, prevod);
- metodologija;
- ime avtorja;
- spol avtorja;
- starost avtorja;
- regijska pripadnost avtorja;
- nacionalnost avtorja;
- prvi jezik avtorja;
- avtoriteta avtorja;
- starost ciljne publike;
- velikost ciljne publike;
- družbeno razmerje avtorja do ciljne publike.

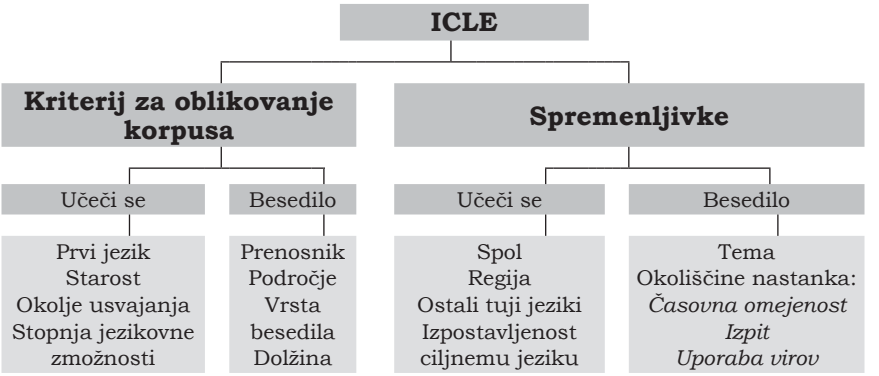
Med lastnostmi je treba določiti tiste, ki so relevantne za uravnoteženost korpusa, in tiste, ki so korpusnim dokumentom samo pripisane, ker bi lahko bile pomembne pri analizi podatkov. S tem so povezani različni pristopi k izbiri besedil. Pri pristopu proporcionalnega vzorca naj bi vzorec govorcev predstavljal celoto jezika v nekem obdobju, stratifikacijski pristop pa besedila razvrščа po različnih vidikih, npr. po jezikovni zvrsti ali po prenosniku (Gorjanc 2005: 30). Stomilijonski slovenski korpus FIDA zajema po različnih kriterijih: besedila so sinhrona, z množičnim naslovnikom in lokalno uravnotežena, izvirna ali prevedena, lektorirana ali nelektorirana ipd. (Stabej 1998: 102–105). V njegovi izpopolnjeni, referenčni različici FidaPLUS je gradivo strukturirano glede na čas izida, lektoriranost, zvrst (umetnostna in neumetnostna; pesniška, prozna in dramska; strokovna in nestrokovna besedila) in tip besedila oziroma mesto objave (Arhar, Gorjanc 2007: 98–101). V Gigafidi je na podlagi povratnih informacij iz FidePLUS taksonomija še poenostavljena. Med drugim je opuščena kategorija lektoriranosti in omejena taksonomija zvrsti (*Korpus pisnih besedil* 2008: 12–19).

Glede korpusnih lastnosti korpusi usvajanja tujega jezika nekoliko odstopajo. Da so sploh uporabni, sta pri zbiranju besedil posebej pomembna ustrezni izbor tvorcev, to je učečih se, in nadzor nad okoliščinami nastanka besedil, čeprav vseh seveda ni mogoče kontrolirati. Večino relevantnih lastnosti najdemo tudi na drugih področjih zbiranja jezikovnih informacij, na primer pri testih jezikovne zmožnosti (Tono 2003a: 801), tiste, ki so kriteriji oblikovanja korpusov usvajanja tujega jezika, pa so prikazane na sliki 2. Druge lastnosti so spremenljivke, ki ne vplivajo na vključevanje besedil v korpus. Na njihovi podlagi je pogosto mogoče iskati in delati različne primerjave (Granger 2004: 126).



Slika 2: Merila oblikovanja korpusa usvajanja tujega jezika (Tono 2003a: 800)

Pri gradnji korpusa ICLE so zbirali podatke o več kot 20 lastnostih, povezanih z učečimi se in z besedili (slika 3). V korpusu je triindvajset prvih jezikov. Tvorci so vsi približno enako stari študentje, za katere je angleščina tuji in ne drugi jezik, na nadaljevalni ali izpopolnjevalni stopnji jezikovne zmožnosti. Njihova besedila so predvidljivo dolga in vsa ali razpravljalni eseji, napisani na tečaju (po 150.000 besed na prvi jezik), ali eseji o prebranih literarnih delih, napisani doma (po 50.000 besed na prvi jezik) (Granger 2003a: 539).



Slika 3: Lastnosti besedil v korpusu ICLE (Granger 2003a: 539)

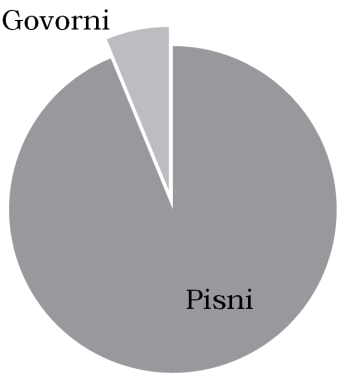
Korpusi usvajanja tujega jezika naj bi bili večinoma uravnoteženi glede na prvi jezik učečih se in/ali stopnjo njihove jezikovne zmožnosti, toda v praksi pogosto pride do odstopanj. V norveškem ASK-u je načrtovani izbor besedil temeljil na enakomernem razmerju: za vsakega od prvih jezikov naj bi zbrali 100 besedil učečih se na stopnji B1 in 80 tekstov na stopnji B2, a se je pri zbiranju pokazalo, da za somalsko in vietnamsko skupino ni dovolj informatorjev (Tenfjord idr. 2006: 1821), zato načrt ni bil izpolnjen. Nekateri korpusi, denimo švedski SVANTE (Bigert idr. 2004: 3), pa so že v izhodišču neuravnoteženi in vključujejo vsa dostopna besedila.

3.1.1 Prenosnik

Prenosnik besedil, ki sestavljajo korpus, je eno najpomembnejših vprašanj, na katerega je treba odgovoriti pri referenčnih korpusih. Prizadevanja, da bi razmerje med pisnim in govorjenim delom odslikavalo dejansko jezikovno rabo, so sicer neuresničljiva, saj je pisne tekste mnogo lažje pridobivati in vključevati. FIDA, FidaPLUS in Gigafida (prim. *Korpus pisnih besedil* 2008: 17) razen nekaj izjem, denimo transkrip-

cij parlamentarnih razprav ali zapisov branih besedil, nimajo zapisov govora. Pri govorjenih besedilih domačih govorcev sta problematična že samo snemanje in vzorčenje govorne raznolikosti, teoretično in fizično naporna ter zamudna pa je tudi transkripcija posnetkov.

Težavnost zbiranja in transkribiranja govorjenega jezika učečih se naj bi bila še podeseterjena (Granger 2004: 130), zato so korpusi usvajanja tujega jezika predvsem pisni, čeprav v zadnjih letih nastaja tudi precej govornih projektov (npr. FLLOC ali SPLLOC). Za učeče se je pisanje zelo pomembna spretnost, poleg tega pa je take korpuse bistveno lažje, hitreje in ceneje zgraditi (Leech 1998: xvii). Od korpusov, o katerih obstajajo natančnejši podatki, je dobra polovica samo pisnih, nekaj manj je govornih, nekaj pa jih ima govorni in pisni del (slika 4). Pri slednjih je švedski ASU eden redkih, kjer je govorni del večji od pisnega. Poseben je tudi korpus MAELC z govornim in pisnim delom, pri čemer govorni del sestavljajo videoposnetki interakcije v razredu.



Slika 4: Razmerje med številom besed v pisnih in govornih delih obstoječih korpusov usvajanja tujega jezika

Zaradi omejenosti na pisni prenosnik se pojavljajo pomisleki o referenčnosti korpusov usvajanja tujega jezika. Toda referenčnost, ki je problematična že v izhodišču vseh korpusov, je pri korpusih usvajanja tujega jezika še bolj utopična, saj je tu še sporneje posploševati z vzorca na celotno populacijo tujih govorcev (Leech 1998: xix).

3.1.2 Velikost korpusa

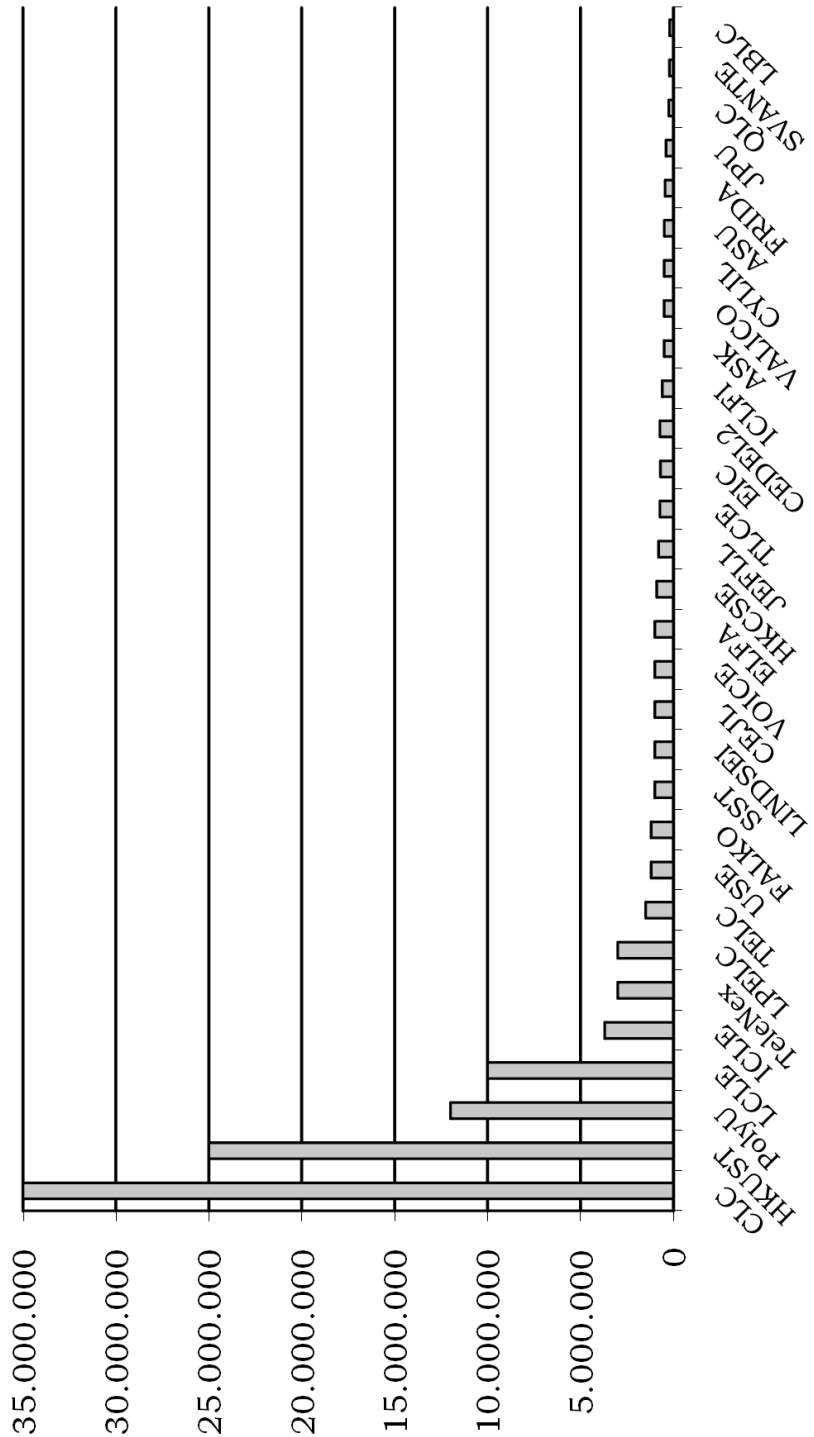
Uveljavljeno mnenje je, da korpus nikoli ne more biti prevelik (prim. Erjavec 2004). Večja količina podatkov daje zanesljи-

vejše rezultate, ki jih je ustrezneje posploševati. Veliki korpusi nudijo več primerov in omogočajo uspešnejše delovanje verjetnostnih programov, ki potrebujejo veliko količino podatkov za učenje. Verjetneje je tudi, da zajamejo obravnavano jezikoslovno vprašanje.

Korpus	**Št. besed (pisni korpus)**	**Št. besed (govorni korpus)**	**Št. besed (govorni in pisni korpus)***
CLC	35.000.000		
HKUST	25.000.000		
PolyU	12.000.000		
LCLE	10.000.000		
ICLE	3.700.000		
TeleNex	3.000.000		
LPLEC	2.800.000	200.000	
TELC	1.500.000		
USE	1.200.000		
FALKO	1.200.000		
CEJL	1.000.000		
SST		1.000.000	
LINDSEI		1.000.000	
VOICE		1.000.000	
ELFA		1.000.000	
HKCSE		900.000	
JEFLL	750.000	50.000	
WriCLE	750.000		
TCLE	730.000		
CEDEL2	730.000		
EIC	700.000		
ICLFI	600.000		
VALICO			500.000
CYLIL		500.000	
ASK	500.000		
ASU	75.000	415.000	
FRIDA	450.000		
JPU	400.000		
QLC	250.000		
SVANTE	220.000		
LBLC	200.000		
Lancom		160.000	
MELD	98.000		
EVA	50.000	35.000	
Skupaj	**100.106.000**	**6.060.000**	**500.000**

***** Pri teh korpusih niso dosegljivi podatki o velikosti posameznega dela glede na prenosnik.

Tabela 4: Število besed v korpusih usvajanja tujega jezika glede na prenosnik.

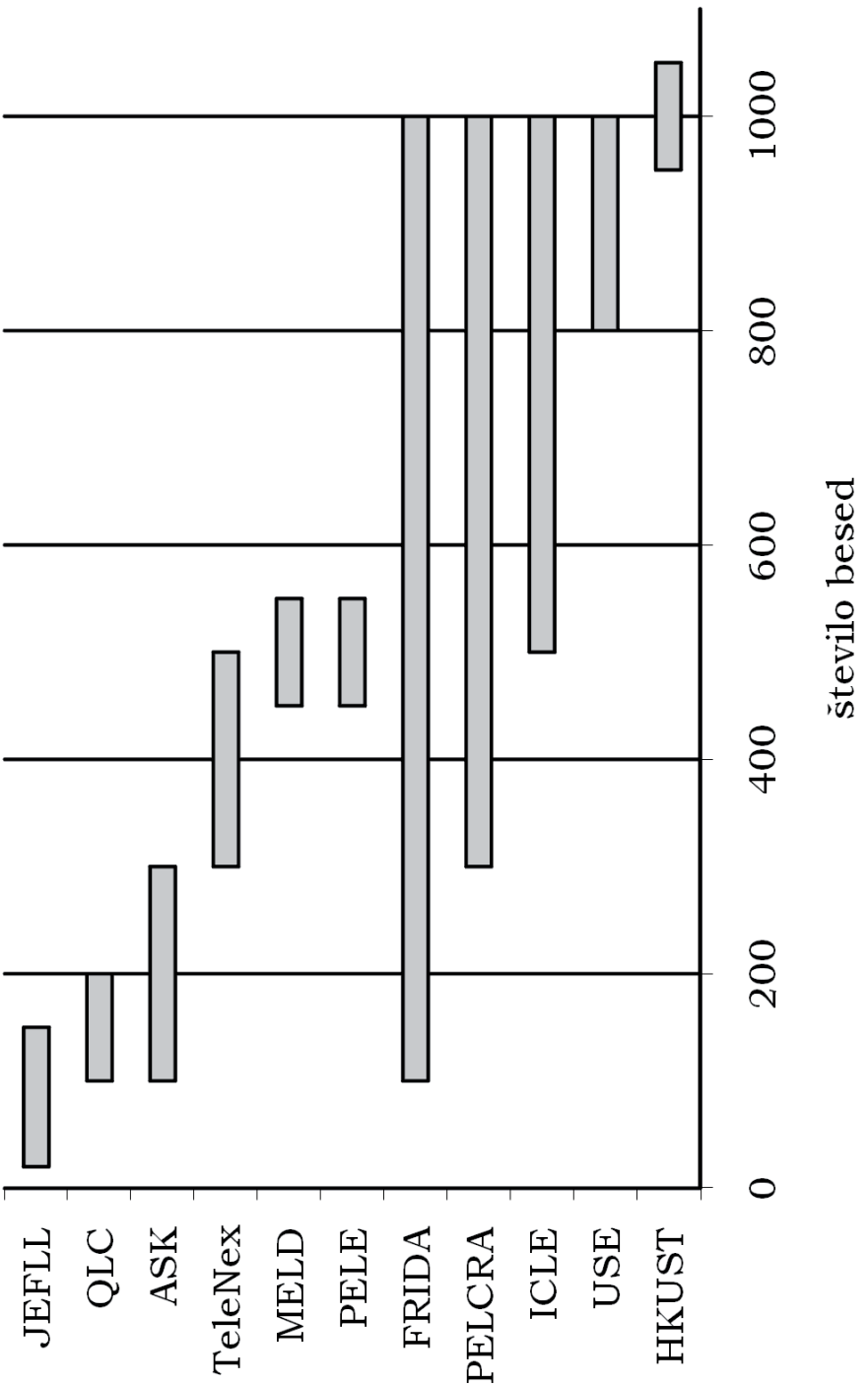


Slika 5: Velikosti obstoječih korpusov usvajanja tujega jezika (v številu besed)

V primerjavi z nespecializiranimi korpusi, ki imajo pogosto več sto milijonov besed, so korpusi usvajanja tujega jezika majhni – večina obstoječih ima od pol do enega milijona besed, kot kažeta tabela 4 in slika 5 na prejšnjih straneh. Vsaj za angleščino je to očitno obseg, ki ni prezahteven za gradnjo, a še daje relevantne rezultate. Večje vire z več milijoni besed potrebujejo ustanove, ki razvijajo komercialne pripomočke za poučevanje jezika, pri izdelavi slovarjev ali učbenikov. Založba Longman uporablja korpus LCLE, založba Cambridge University Press pa CLC. Korpusi za ostale jezike, kot sta FRIDA in ASK, so večinoma nekoliko manjši.

Relativna majhnost korpusov usvajanja tujega jezika v primerjavi z nespecializiranimi je povezana s težavnostjo zbiranja gradiva (Granger 2004: 124), ker je večinoma omejeno na organizirani jezikovni pouk, npr. tečaje na univerzah ali v jezikovnih šolah. Poleg tega je sestavna besedila večinoma treba pretipkati, označevanje pa je v veliki meri ročno. Omejenost obsega je treba upoštevati pri analizi in interpretaciji rezultatov. Ugotovitve je težje posploševati in verjetneje je, da se določene besede strukture sploh ne pojavijo. Manjša velikost pa še ne pomeni, da so korpusi usvajanja tujega jezika manj uporabni ali nujno nereprezentativni. Manjši korpusi pogosto zadostujejo za specializirane raziskave ali za uporabo v razredu (O'Keeffe, Farr 2003: 409). Kot je pokazal Biber, »je za ugotavljanje lastnosti določene jezikovne zvrsti 10 besedil že dovolj velik vzorec, od vsakega besedila pa zadošča naključno izbran vzorec v dolžini 1000 besed« (Vintar 2008: 82). Manjše korpuse je enostavneje graditi, obdelovati in interpretirati rezultate, ki so preglednejši. Uporabnik se lahko podrobno seznani z vsemi besedili in jih – ob nujnem izogibanju neutemeljenemu posploševanju – pravilneje interpretira. Ne nazadnje pa je velikost relativna. Korpus z 200.000 besedami je v splošnem korpusnem jezikoslovju miniaturen, pri raziskovanju usvajanja tujega jezika, kjer se raziskovalci tradicionalno zanašajo na precej manjše vzorce, pa je že lahko opredeljen kot velik.

Z velikostjo korpusa je povezana dolžina besedil, ki ga sestavljajo. Učeči se načeloma težje pišejo v neprvem jeziku, zato so teksti v pisnih korpusih redko daljši od tisoč besed. V pregledanih korpusih je to maksimum, medtem ko minimum praviloma ne seže pod sto. Izjema je japonski JEFLL, ki izstopa z minimalnim številom dvajset besed pri najmlajših tvorcih, starih 12 in 13 let. Največ korpusov vključuje besedila z okrog petsto besedami (slika 6).



Slika 6: Dolžina besedil v pisnih korpusih usvajanja tujega jezika

Kot pri drugih tipih korpusov se tudi pri korpusih usvajanja tujega jezika ni mogoče izogniti dilemi o statičnosti oziroma dinamičnosti. Ker se vmesni jezik učečih se spreminja, bi jih bilo smiselno oblikovati tako, da je mogoče dodajati nova besedila. Vendar dinamični korpus zaradi velikosti sčasoma postane manj obvladljiv in neprimeren za primerjalne študije, načelo skrbno načrtovanega izbora in uravnoteženosti pa v njem zamenja oportunistično načelo količine. Ob tem ni nepomembno, da je zaradi večjih stroškov, povezanih z obsežnejšo obdelavo, dostopen manjšemu številu uporabnikov (Kennedy 1998: 61). V večje korpuse namreč praviloma investirajo komercialna podjetja, ki dostop do gradiva tržijo in ne omogočajo brezplačne uporabe. Zato je edini dinamični korpus od pregledanih estonski EIC.

3.1.3 Jezik

Ena od specifičnosti, po kateri se korpusi usvajanja tujega jezika ločijo od večine drugih tipov, je, da sta zanje pomembna dva jezika: ciljni in izhodiščni. Ciljnega jezika praviloma ni treba posebej določati, problem se lahko pojavi le pri svetovno bolj razširjenih jezikih. Kaj je jezik domačih govorcev, glede na katerega se označuje odklone od norme – ali je za angleščino to britanska, ameriška, kanadska, avstralska ali katera druga različica (Leech 1998: xix)? Težava je manjša za manj razširjene jezike, kot je slovenščina, toda problematične so lahko tudi pogovorne in druge nestandardne oblike (prim. poglavje 4.4).

Za korpuse usvajanja tujega jezika sta značilni dve kombinaciji ciljnih in izhodiščnih jezikov. Za velike skupine govorcev so tipični korpusi z enim izhodiščnim in enim ciljnim jezikom (tabela 5), npr. korpus angleščine kot tujega jezika za nemške govorce.

Druga pogosta kombinacija ima več izhodiščnih in en prvi jezik (tabela 6). To je značilno za mednarodne projekte, npr. CLC, LCLE in ICLE, ali korpuse manjših jezikov, kot sta JASMIN ali ASK, v katerem naj bi izhodiščni jeziki ustrezali jezikom največjih skupin priseljencev v državi.

Korpus	Izhodiščni jezik		Ciljni jezik
CEJL	japonski		
COLSEC	kantonski		
ELFA	različni		
EVA	norveški		
HKCSE	kantonski		
HKUST	kantonski		
Japanese Learners' Corpus	japonski		
JEFLL	japonski		
JPU	madžarski		
LBLC	japonski		
MAELC	različni		
LPELC	poljski	→	angleški
PELE	poljski		
POLY U	kantonski		
QLC	francoski		
SST	japonski		
SWECCL	kantonski		
TELC	tajski		
TCLE	kitajski		
TeleNex	kantonski		
USE	švedski		
WriCLE	španski		
CEDEL2	angleški	→	španski
FLLOC	angleški	→	francoski
InterFra	švedski	→	francoski
InterIta	švedski	→	italijanski
SPLLOC	angleški	→	španski

Tabela 5: Izhodiščni in ciljni jeziki v korpusih usvajanja z enim izhodiščnim jezikom

Korpus	Izhodiščni jezik		Ciljni jezik
ASK	albanski, angleški, bosanski/hrvaški/ srbski, nemški, nizozemski, somalski, španski, poljski, ruski, vietnamski	→	norveški
CLC	130 jezikov	→	angleški
CYLIL	francoski, grški, italijanski, nizozemski	→	angleški
EIC	različni	→	estonski
ELFA	51 jezikov	→	angleški
FALKO	afrikanski, angleški, danski, italijanski, svahili, turški, uzbeški	→	nemški
FRIDA	različni	→	francoski
IBLC	angleški, francoski, finski, nemški, tajski	→	angleški
ICLE	arabski, bolgarski, cvanski, češki, finski, francoski, grški, italijanski, japonski, kitajski, litvanski, madžarski, nemški, nizozemski, norveški, pakistanski, poljski, portugalski, portugalski (braz.), romunski, ruski, slovenski, španski, švedski, turški	→	angleški

Korpus	Izhodiščni jezik		Ciljni jezik
ICLFI	22 jezikov	→	finski
ISLE	nemški, italijanski	→	angleški
JASMIN	arabski, francoski, turški in drugi	→	nizozemski
LCLE	160 jezikovnih okolij	→	angleški
LINDSEI	bolgarski, francoski, grški, kitajski, italijanski, japonski, nemški, nizozemski, poljski, španski, švedski	→	angleški
MELD	arabski, bengalski, gudžaratski, haitski kreolski, hindujski, mandarinski, poljski, španski, tajvanski, vietnamski	→	angleški
SVANTE	več jezikov	→	švedski
VOICE	50 jezikov	→	angleški

Tabela 6: Izhodiščni in ciljni jeziki v korpusih usvajanja tujega jezika z enim ciljnim jezikom

V teh korpusih so prav izhodiščni jeziki kriterij uravnoteženosti, kar pomeni, da so njihovi deleži ponavadi enakomerni. V korpusu ICLE ima vsak 200.000 besed, v ASK-u pa 180 besedil. V njem so deset jezikov izbrali oportunistično glede na to, katerih besedil je bilo dovolj v arhivih in kateri so zaradi tipološke različnosti omogočali zanimive primerjave (Tenfjord 2006).

Edini korpus, ki je sestavljen drugače od opisanih, je ESFSLD. Ker gre za mednarodni evropski projekt, ima več izhodiščnih (arabski, finski, italijanski, pundžabski, španski, turški) in več ciljnih jezikov (angleški, francoski, nemški, nizozemski, švedski). Vsak ciljni jezik je povezan z dvema izhodiščnima jezikoma, ki predstavljata veliko skupino tujih govorcev in omogočata zanimive medjezikovne primerjave (Allwood idr. 1986).

3.1.4 Tvorci

Še en pomemben pogoj za uravnoteženost korpusov usvajanja tujega jezika so podatki o tvorcu oziroma učečem se, saj so brez ustreznega pregleda nad njimi informacije iz korpusa nekoristne. Mednje spadajo starost, izobrazba, spol, prvi jezik, učno okolje ciljnega jezika, ocena stopnje jezikovne zmožnosti v ciljnem jeziku, znanje ostalih tujih jezikov in praktične izkušnje v državah, kjer je ciljni jezik prvi jezik. Po mnenju nekaterih bi bilo koristno poznati tudi abstraktnejše značilnosti, kot so notranja kognitivna (učni slog ali učne strategije) in afektivna (motiviranost in stališča do ciljnega jezika) merila tvorcev (Tono 2003a: 800), a jih je za dejansko analizo seveda težko opisati ali oceniti.

Več informacij o tvorcu imajo v glavnem akademski projekti, medtem ko so v komercialnem korpusu LCLE prav zaradi njegove izjemne velikosti omejene na najmanjšo možno mero (Granger 1998a: 11). Poleg podatkov, navedenih v tabeli 7, so vključene še informacije o izobrazbi (ASK, EIC), poklicu (ASK, EIC, HKCSE), veroizpovedi (ESFSLD), prvem jeziku staršev (CEDEL2, FRIDA, ICLE), kraju rojstva (HKCSE), letih učenja ciljnega jezika (CEDEL2, FRIDA, ICLE), tečaju (JEFLL, JPU), doseženi oceni na testu (EVA), kraju bivanja (EIC) in podobno.

Korpus	Spol	Starost	J1	Narodnost	Stopnja j. zmožnosti	Znanje ostalih jezikov	Šola	Država	Obiski v državah z J1
ASK	x	x	x	x	x	x		x	x
CEDEL2	x	x	x		x	x	x		x
CEJL	x	x	x				x		
CLC		x	x	x	x				
EIC	x	x	x		x				
ELFA	x	x	x	x					
ESFSLD	x		x						
EVA	x						x		
FRIDA	x	x	x	x		x	x		x
HKCSE	x	x	x				x		x
ICLE	x	x	x	x		x	x		x
ISLE	x		x						
LCLE				x	x			x	
MELD	x	x				x	x		
LPELC	x	x			x	x			x
USE	x	x	x				x		x

Tabela 7: Pregled podatkov o tvorcih sestavnih besedil

Ključni kriterij za izbor besedil je ocena tvorčeve jezikovne zmožnosti v ciljnem jeziku. Idealno je, če je določena na podlagi internih ali standardiziranih testov, pogosto pa temelji na vtisu ali institucionalnem statusu tvorcev. Ker včasih niti ni jasno, kako je ocenjena, podatki niso primerljivi, toda večinoma so tvorci na nadaljevalni ali izpopolnjevalni stopnji: besedila nadaljevalne stopnje so vključena v ASK (B1 in B2 po lestvici SEJO), FRIDA, HKUST, ISLE in TeleNex; besedila nadaljevalcev in izpopolnjevalcev v ICLE (od B1 do C2), QLC in TCLE; izpopolnjevalcev pa v FALKO (C1 in C2), IBLC, MELD in LPELC. Korpusi CEDEL2, CLC, JEFLL, LCLE, PolyU in SST naj bi po opredelitvah njihovih graditeljev vključevali vse stopnje in so med njimi večinoma uravnoteženi, npr. CLEC (Gui, Yang 2006). Redkeje pa

so vključena besedila začetnikov, saj so praviloma kratka z mnogo odkloni od norme, popolnoma tujimi strukturami in malo koherentne vsebine. Značilnosti njihove produkcije so izrazito prehodne, faza vmesnega jezika pa je nestabilna. Za tekste začetnikov, torej na stopnji A1 do B2 po lestvici SEJO, so se odločili v govornem korpusu JASMIN (Cucchiarini idr. 2008), od pisnih pa jih ima še največ, približno 40 %, estonski korpus EIC (spletna stran).

Glede na učno okolje ciljnega jezika ločimo med korpusi drugega in korpusi tujega jezika. ASK, ESFSLD in MELD so korpusi drugega jezika, torej jezika okolja, v katerem živijo učeči se, medtem ko so ICLE, LINDSEI in FRIDA korpusi tujega jezika, saj tvorci ne živijo v okolju ciljnega jezika. Vendar je to le redko kriterij za izbor, graditelji ponavadi dopuščajo obe možnosti: v korpusu EIC sodelujejo tako tvorci, za katere je estonščina drugi jezik, kot tisti, za katere je tuji jezik.

Kot je bilo že poudarjeno, so nekateri podatki zgolj označeni in vključeni med iskalne pogoje, medtem ko ne vplivajo na izbor besedil. To velja npr. za regijo, v kateri živi govorec, pri kateri se je v korpusu JASMIN naknadno izkazalo, da se v govoru otrok z višjimi ravnmi jezikovne zmožnosti opažajo narečni vplivi (Cucchiarini idr. 2008). Tudi starost in izobrazba sta zanimivi in zgovorni. Tvorci v obstoječih korpusih so univerzitetni študentje ali srednješolci, le redko pa otroci, pri katerih usvajanje jezika poteka drugače. V korpusu JASMIN so otroci med 7. in 16. letom ter odrasli, v korpusu HKUST dijaki, v korpusu EVA 14–15-letniki, v PolyU in TCLE dijaki in študentje, v korpusih FALKO, ICLE, JPU, MELD, TELC in USE samo študentje, JEFLL, LPELC in FLLOC pa zajemajo vse stopnje šolanja (Axelsson 2000: 155; Cucchiarini idr. 2008; Granger 2001: 2–3; Lüdeling idr. 2005; Pravec 2002: 86–95; Tenfjord idr. 2004; Uzar 1998). Razlogi za omejenost na tvorce, ki se še šolajo, so seveda pragmatični. Čeprav ni nujno, da pišejo več kot tisti, ki so šolanje že končali, graditelji korpusov in raziskovalci, ki so velikokrat učitelji ali univerzitetni profesorji, lažje pridejo do besedil oseb, vključenih v učni proces. Primerna populacija so tudi starejši udeleženci tečajev tujega jezika, kot je značilno za ASK. V vsakem primeru pa gre za ljudi v stiku z institucijami, ki poučujejo jezik, saj so le tako njihova besedila dostopna graditeljem korpusa. Pri kriteriju starosti nekoliko izstopa ESFSLD. Zaradi specifičnosti raziskave vključuje samo moške med 18. in 30. letom, ki so v državo prišli pred kratkim in so se jezika učili neformalno.

Podatki o tvorcih so največkrat pridobljeni s pomočjo posebnih obrazcev, ki pogosto vključujejo tudi dovoljenje za uporabo (prim. poglavje 3.2).

PROFIL UČENCA

Starost: Spol:

Narodnost:
Prvi jezik:
Očetov prvi jezik:
Materin prvi jezik:
Jeziki, ki jih govorite doma: (če uporabljate več kot enega, v % navedite delež rabe)

Izobrazba :
Osnovna šola (jezik): Srednja šola (jezik):
Trenutno izobraževanje: Trenutni letnik:
Ustanova:
Jezik izobraževanja: Samo angleški / Drugi jeziki (kateri) / Oboje

Leta angleščine v šoli: Leta angleščine na univerzi:

Bivanje v angleško govoreči deželi:
Kje? Kdaj? Koliko časa?

Ostali tuji jeziki (od tistega, ki ga znate najbolje, do tistega, ki ga znate najslabše):

Dovoljujem, da moj esej uporabite v raziskovalne namene.

Datum: Podpis:

Primer 1: Obrazec za korpuse ICLE, LINDSEI in FRIDA, ki ga izpolni učenec (spletna stran ICLE, prevedla M. S.)

Vsi podatki bodo anonimno uporabljeni v raziskovalne namene.

JEZIKOVNO OZADJE

Vaše začetnice:
Spol:
Starost:
Elektronski naslov:

Univerza/ustanova:
Oddelek:
Stopnja/tečaj:
Leta obiskovanja tečaja:

Vaš prvi jezik:
Očetov prvi jezik:
Materin prvi jezik:
Jeziki, ki jih govorite doma:
Starost, pri kateri ste se začeli učiti špansko (v letih):
Leta učenja španščine:

Ali ste že bili v špansko govoreči državi:
Če da, v kateri?
Kdaj?
Kako dolgo?

Prosimo, ocenite svojo zmožnost v španščini:
Govorjenje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Poslušanje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Branje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Pisanje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna

Ali govorite še katere jezike poleg angleščine in španščine?
Če da, ocenite svojo zmožnost v drugih jezikih:
Govorjenje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Poslušanje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Branje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna
Pisanje: začetna / osnovna / nižja nadaljevalna / nadaljevalna/ izpopolnjevalna / visoka izpopolnjevalna

Primer 2: Obrazec o tvorcu besedila v korpusu CEDEL2
(spletna stran CEDEL2, prevedla M. S.)

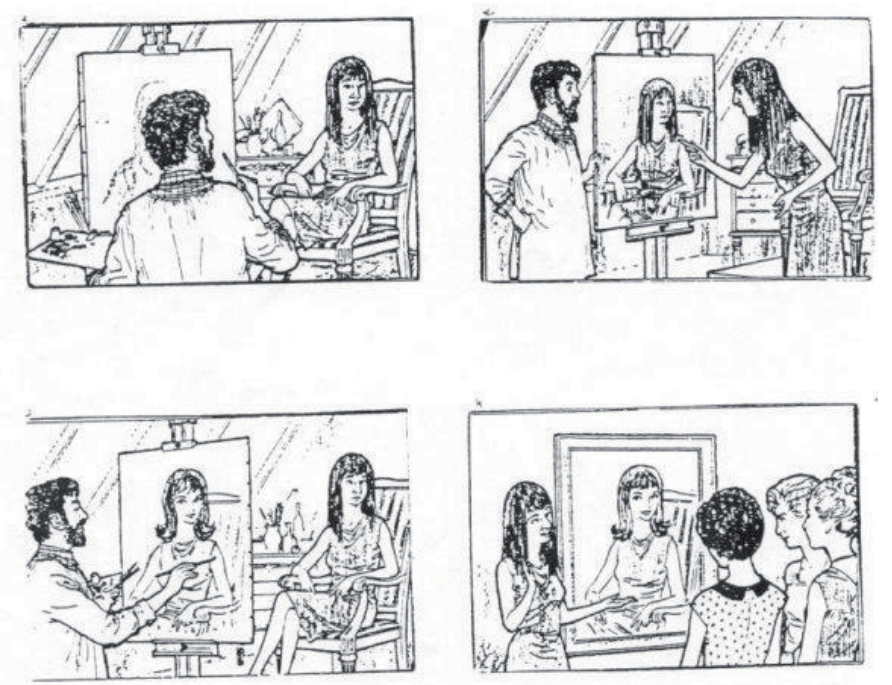
3.1.5 Vrsta besedil in tema

Uravnoteženost med vrstami in tipi sestavnih besedil pri oblikovanju referenčnih korpusov zahteva veliko premislekov. Nekateri predlagajo razlikovanje med fikcijskimi in nefikcijskimi deli, knjigami, revijami in časopisi, formalnimi in neformalnimi besedili ali izbor glede na starost, spol in poreklo avtorjev. Drugi zagovarjajo različne pristope k izboru: elitističnega na podlagi literarne ali akademske vplivnosti, naključnega, po načelu aktualnosti, glede na subjektivno presojo tipičnosti, glede na dostopnost besedil v arhivu, gle-

de na demografsko vzorčenje bralnih navad ali prilagajanje jezikoslovnim potrebam (citirano po: Kennedy 1998: 63–64).

Tudi tu je položaj korpusov usvajanja tujega jezika nekoliko drugačen. Že v izhodišču se srečujejo s paradoksom, značilnim za jezikovno testiranje: predstavljati želijo jezik nedomačih govorcev, toda do »naravne podobe« tega jezika v resnici ne morejo priti. Dostopna so zgolj besedila, ki so rezultat učnega procesa, pri njih pa je precej omejitev – produkcija je odraz testne performance, poleg vnaprej določene teme, omejenega časa in podobnega tvorec »pokaže« le področja jezikovne zmožnosti, na katerih je najsamozavestnejši, in ne celote. Zato morajo biti besedila čim bližja naravni rabi, izzvana, a napisana s čim manj kontrole (Nesselhauf 2004: 128). V govornih korpusih so to vodeni intervjuji ob pomoči slik (sliki 7 in 8) ali vnaprej pripravljenih vprašanj, v pisnih pa je največ spisov oziroma esejev. Pojavljajo se še pisma, dopisi, dnevniki, poročila, članki, seminarske naloge, obnove, povzetki, različna poslovna besedila, projektna poročila in podobno. Te vrste besedil naj bi bile najbližje integrativnemu tipu testiranja, ki na sporazumevalno zmožnost gleda celovito in ne kot na skupek nepovezanih enot (Ferbežar 1999: 432). Slog besedil je običajno utemeljevanje ali razlaganje, opisovanju in pripovedovanju se izogibajo. Za tvorce so lažje pisne naloge z jasnimi, natančnejšimi navodili, ki niso tako odvisne »od družbenega in kulturnega ozadja posameznega kandidata, njegove domišljije, nadarjenosti in ustvarjalnosti« (Ferbežar, Bajec idr. 2004: 115–116). Poleg tega so za tvorce z višjimi ravnmi jezikovne zmožnosti, kakršni največkrat sodelujejo v korpusih, primernejša besedila, ki zahtevajo kompleksnejše sposobnosti, npr. utemeljevanje ali izražanje mnenja (navedeno delo: 118). V korpusu ICLE zato odsvetujejo naslove, kot sta *Radosti angleškega podeželja* ali *Moje leto v Ameriki* (Granger 2001: 6), medtem ko LCLE, PELE, TeleNex in TCLE vključujejo tudi pripovedne in opisne tekste ali dnevniške zapiske (Pravec 2002: 92; Shih 2000: 92).

Tematika besedil je določena z naslovi in navodili (tabela 8) in je pomembna, ker vpliva na izbiro besedišča. Splošne teme spodbujajo uporabo drugačnega besedišča in slovničnih struktur kot bolj specifične ali strokovne, politični problemi se razlikujejo od osebnih izkušenj. Ker pa je s korpusi usvajanja nemogoče zajeti reprezentativen delež vseh polnopomenskih besed, se na njih temelječe raziskave bolj osredotočajo na slovnične besede in strukture, tako da tudi tematika stopi v ozadje.



Slika 7: Slike, na podlagi katerih informatorji za korpus LINDSEI pripovedujejo zgodbo (spletna stran LINDSEI)

Slika 8: Strip, ki ga je treba v zgodbi s sto besedami opisati za korpus VALICO (spletna stran VALICO)

Korpus	Teme/Naslovi
CEJL	Spisi o potovanju, vrtnarjenju ali kuhanju
	Obnove videoposnetka, japonskih ljudskih zgodb in Ezopovih basni
EVA	Pismo
	Neformalno sporočilo prijatelju z že danim besediščem
	Spis o potovanju z danim začetkom ali koncem
ICLE	Kultura, cenzura, zločin in kazen, morala
	Evropske sanje
	Človek in narava
	Zločin se ne izplača
	Feminizem lahko ženskam bolj škodi, kot koristi
	Onesnaževanje: tiha zarota
JPU	Osebne izkušnje z učenjem izven šole
	Besedila o člankih, objavljenih v časopisu na dan tvorčevega rojstva
PELE	Opisovanje oseb
	Razpravljanje o različnih šolskih sistemih
QLC	Kako bi lahko izboljšali učenje angleščine v Quebecu
	Kako bi bilo vaše življenje drugačno, če bi znali bolje angleško
USE	Ovrednotenje lastnega znanja angleščine in odnosa do nje
	Argumentiranje trditve
	Razmišljanje o nasilju v medijih na podlagi članka
	Razmišljanje o določenih elementih angleščine, o tabujih, politiki, izobrazbi in šolskih ekskurzijah
	Esej o prebranem literarnem delu ali o kulturnem problemu (suženjstvo, utopije, industrijska revolucija)

Tabela 8: Pregled tem v obstoječih pisnih korpusih

Zelo pomembna merila so povezana z okoliščinami nastanka v korpus vključenih besedil (angl. *task-related*, povezana z nalogo). Nanašajo se na longitudinalno oziroma presečno zbiranje podatkov, spodbudo pri nastanku besedil, uporabo referenc, kot so slovarji, drugi učni pripomočki ali že napisana besedila, in časovno omejenost pri pisanju. Ker so korpusi usvajanja še precej na začetku razvoja, so v glavnem presečni. Longitudinalni z več prispevki istega tvorca so LPELC, ki naj bi v končni fazi vseboval tri tekste vsakega tvorca, predvidoma vsakega na višji stopnji, ESFSLD, MAELC, MELD, LONGDALE, ASU, ki spremlja tvorce od začetka učenja do uspešno opravljenega izpita iz znanja švedščine, ter deli korpusov FALKO in FLLOC. Raziskovalci si večkrat pomagajo s kvazi longitudinalnimi korpusi, ki vsebujejo besedila, zbrana v določenem trenutku, a s tvorci na različnih stopnjah jezikovne zmožnosti (Granger 2004: 131).

Do pomembnih razlik prihaja pri kriterijih uporabe slovarjev, črkovalnikov in drugih referenčnih pripomočkov ter

časovne omejenosti pri pisanju. Naloge, napisane doma brez živčnosti in z vsemi možnimi pripomočki, se ločijo od besedil, nastalih v izpitnih situacijah na državno veljavnih izpitih iz znanja tujega jezika ali pri testih v razredu. Primerjava švedske komponente korpusa ICLE in korpusa USE je pokazala, da tvorci uporabljajo osebnejši slog, kadar je pisanje časovno omejeno, in manj osebni slog, kadar imajo dostop do drugih besedil in referenc (Ädel 2008: 35). Nesproščena situacija na izpitih, kjer so pogoji tvorjenja omejeni, vpliva na testno performanco s posebnimi okoliščinami testiranja, kot so prostor, hrup, ura dneva. Stres, trema in podobni dejavniki pa lahko še dodatno zabrišejo posameznikovo zmožnost sporazumevanja (Ferbežar 1999: 430). Toda na izpitih so pogoji nadzorovani, stopnja jezikovne zmožnosti tvorcev pa je preverjeno enaka. Po drugi strani lahko tvorci pri pisanju doma pokažejo širši razpon jezikovne zmožnosti, a jih je nemogoče nadzirati in zagotavljati, da ne prepisujejo ali da jim besedil ne pišejo drugi, uporaba pripomočkov pa lahko popači sliko o njihovi dejanski jezikovni zmožnosti. Dodatno opozorilo ob besedilih, napisanih v šolskem okolju, se nanaša na vprašljivo reprezentativnost. Entuziasti, ki so pripravljeni pokazati izdelke svojih učencev izven razreda in jih prispevati v korpus, so verjetno bolj motivirani za delo, zato je verjetno, da so tudi rezultati njihovih učencev nekoliko nadpovprečni (Tono 2003b).

Kateri tip situacije torej verodostojneje odraža vmesni jezik učečih se in ali se je v korpusih sploh smiselno omejiti samo na enega? Pisanje esejev že samo po sebi ni ravno realistična aktivnost, tipična za resnično življenje. Časovne omejitve v dejanski komunikaciji niso nerealne, saj sogovorniki večinoma niso pripravljeni poljubno čakati na odgovor. Po drugi strani tudi tvorci v korpusih prvega jezika, ki pri pisanju večinoma niso časovno omejeni, uporabljajo različne pripomočke in vire, njihovo delo pregledujejo lektorji, uredniki in drugi, uporaba referenc pa v teh korpusih pogosto sploh ni dokumentirana.

Zaradi naštetih argumentov nekateri kot najprimernejšo zagovarjajo kombinacijo obeh tipov okoliščin (Ädel 2008: 49), za kakršno so se npr. odločili v korpusu EAGLE, v katerem sta sopostavljena tudi pisanje besedil na računalniku v okviru pouka in pisanje na roko na izpitih (Boyd 2008: 1897). Toda sicer so v obstoječih korpusih pogostejša besedila z jezikovnih izpitov. Gradnja korpusov je ponavadi povezana z institucijami, ki izpite izvajajo in so jim tovrstni teksti verje-

tno enostavno dostopni. V FALKO so vključena besedila, ki jih napišejo tuji študentje v okviru izpita, na katerem dokažejo ustrezno jezikovno zmožnost v nemščini. Pri tem nimajo časa za pripravo, članka, ki ga morajo povzeti, ne poznajo vnaprej, povzetek pa napišejo v 90 minutah brez uporabe pripomočkov. Del tekstov za TELC je pridobljen na univerzitetnih sprejemnih izpitih, enako velja za QLC. 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). V korpusu JEFLL je bila časovna omejitev na izpitu dvajset minut (navedeno delo: 92), spisi v CEJL so morali biti napisani v desetih minutah. V nekaterih primerih gre za različne zunanje, mednarodno veljavne izpite – CLC in LPELC imata besedila s Cambridgeovih izpitov ESOL (Pravec 2002: 91, 95), napisana v eni uri.

Doma napisana besedila pa sestavljajo zgolj manjši delež majhnega števila korpusov, med drugim LCLE, TeleNex, MELD, CEJL in HKUST. 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). Eseji za USE so napisani doma v dveh do treh tednih z dovoljeno uporabo vseh didaktičnih pripomočkov. Pred oddajo je zaželeno, da tvorci tekste pregledajo (Axelsson 2000: 155). Tudi v seminarskih nalogah, ki so del korpusa JPU, lahko uporabljajo dodatno literaturo (Pravec 2002: 92).

3.1.6 Kontrolni korpus

Podatki iz korpusov usvajanja tujega jezika največkrat ne povedo dosti brez primerjave s podatki domačih govorcev, ki da kvantitativne informacije o pogostnosti določenih besed, besednih vrst, skladenjskih struktur in o značilnostih diskurza (Tono 2003a: 800). Večina primerjav temelji na referenčnih korpusih (prim. Tono 2003a; Tono 1998; Granger 1998b). Korpusa CLC in LCLE sta del širših korpusnih mrež Cambridge International Corpus oziroma Longman Corpus Network. Komercialno manj podprti korpusi včasih po primerljivih načelih zgradijo poseben kontrolni korpus, ki pa je praviloma premajhen za utemeljene splošne primerjave in je primernejši za reševanje raznih sprotnih vprašanj med samo analizo.

Kontrolni korpus korpusa ASK sestavljajo besedila domačih govorcev, ki so reševali enak izpit kot tuji govorci, in sicer

sto na ravni B1 in sto na ravni B2. Informatorje so iskali v raznolikih skupinah, športnih klubih, pevskih zborih, med študenti ipd., njihovo sodelovanje pa je bilo honorirano. Tudi v kontrolnem korpusu so označili napake in ugotovili, da jih je precej več, kot so pričakovali (Tenfjord 2006). ICLE spremlja korpus esejev domačih govorcev angleščine Louvain Corpus of Native English Essays (Granger 2001: 3). Poljskemu korpusu PELE je dodan miniaturni vzorec besedil domačih govorcev angleščine in besedil poljščine kot prvega jezika, ki obsega tri tisoč besed (Uzar 1998). V korpusu SVANTE so kontrolno gradivo maturitetni razpravljalni spisi domačih govorcev. Vsak podkorpus nemškega korpusa FALKO spremljajo besedila, ki so jih v primerljivih pogojih napisali domači govorci nemščine. Korpus poslovne komunikacije IBLC je sestavljen iz besedil tujih in domačih govorcev (Connor idr. 2002: 180). Kontrolno gradivo japonskega JEFLL-a je zelo razdelano: japonsko pisno gradivo naj bi v prihodnosti v angleščini omogočilo prepoznavanje napak, ki so posledica jezikovnega prenosa iz japonščine. Dodan je še pedagoški korpus angleščine z 1.200.000 besedami. Sestavljajo ga učbeniki angleščine, ki jih uporabljajo v japonskih šolah in 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 British National Corpus (Pravec 2002: 95). Zanimiv je tudi japonski English Learners' Corpus of Japanese Speakers z milijonom besed. Zasnovan je kot vzporedni korpus, kjer se kratkim spisom v angleščini pridružujejo japonski prevodi in pisanje domačih govorcev (Tono 2003a: 803; Sugiura 2000).

Kontrolno gradivo ima precej večji delež govornih korpusov. LINDSEI spremljajo intervjuji z domačimi govorci (De Cock idr. 1999: 1). V korpusu ISLE sta za primerjavo dodana posnetka dveh domačih govorcev (Atwell idr. 2003: 6), v korpusu EVA pa 26 britanskih govorcev oziroma skupaj 17.000 besed (Hasselgren 2001: 123). Švedski ASU vsebuje pogovore in besedila sedmih domačih govorcev v skupnem obsegu 123.000 besed. Japonski SST ima podkorpus domačih govorcev angleščine, ki so odgovarjali na podobna vprašanja kot Japonci, in podkorpus angleških izjav Japoncev, prevedenih nazaj v japonščino (Izumi idr. 2004: 37). HKCSE nima kontrolnega korpusa, vendar so v njem izjave tako domačih kot tujih govorcev (Cheng, Warren 1999: 6). V korpusu Salford, delu projekta FLLOC, besedila britanskih govorcev v francoščini spremljajo podobna besedila, tvorjena v angleščini. V

podkorpusu Newcastle je poleg 30 britanskih dijakov, ki se učijo francosko, iste naloge opravljalo 15 domačih govorcev, podobno pa velja za podkorpus Reading Corpus. Kontrolno gradivo imajo tudi korpusi SPLLOC, InterFra, ESFSLD in JASMIN (Allwood idr. 1983; Cucchiarini idr. 2008; Myles 2007: 396).

3.2 Zajem besedil

Faza zajema besedil za korpus bi se lahko imenovala faza presenečenja, saj se na prvi pogled zdi trivialna, posebej po dobro pripravljeni specifikaciji. Toda »gradnja korpusa (z zadostno pozornostjo do kakovosti, oblikovanja itn.) vedno traja dvakrat dlje in zahteva desetkrat več truda, kot smo mislili« (Leech 1998: xvii). Realizacija korpusa JASMIN je zahtevala mnogo več časa, kot so predvidevali, pa tudi pridobivanje govorcev se je izkazalo za še težavnejše od pričakovanj (Cucchiarini idr. 2008). Organizacija zbiranja besedil in prepričevanje tvorcev, naj jih odstopijo, terjata veliko truda, na srečo pa je prav na podlagi izkušenj obstoječih korpusov to mogoče že realneje upoštevati pri načrtovanju izdelave novih korpusov.

Z **organizacijskega stališča** je zbiranje tekstov za korpuse usvajanja zahtevno, ker jih nastane manj kot tekstov domačih govorcev, a so po drugi strani lažje dostopni. Zato ni naključje, da so graditelji korpusov tesno povezani z organizatorji jezikovnih tečajev in izpitov, kjer je skupaj več tujih govorcev. Če to dopušča projekt, se sodelujočim oddolžijo za prispevek h korpusu: v norveškem ASK-u so bili informatorji za kontrolni korpus, ki so jih iskali v športnih klubih, pevskih zborih, med študenti in podobno, honorirani (Tenfjord 2006). V nizozemskem JASMIN-u so otroške govorce našli v različnih šolah, odrasle pa v jezikovnih šolah in prek organizacij za tujce. Vsi so prejeli vstopnico za kino, bon za knjigarno ali trgovino z igračami (Cucchiarini idr. 2008). V korpusu LINDSEI pa so sodelujočim kakovostno ocenili jezikovno zmožnost, podobno ponujajo v korpusu CEDEL2. Bolj zapleteno je zbiranje besedil izven šolskega okolja. V govornem korpusu HKCSE so želeli spontano naravno gradivo, govorna dejanja zunaj institucionaliziranega okolja z vsaj dvema udeležencema, sestavljena iz več kot le ustaljenih govornih izmenjav. Snemali so na domovih, delovnih mestih, v restavracijah, parkih in na družabnih srečanjih (Cheng, Warren 1999: 8).

Resna **pravna omejitev** gradnje korpusov so avtorske pravice oziroma varovanje osebnih podatkov, zaradi katerih morajo biti tvorci anonimni ali navedeni zgolj z začetnicami, kot je v CEDEL2. V ASK-u so izključena vsa osebna, zemljepisna in druga imena, namesto njih uporabljajo kode. Podobno je v korpusu USE (spletna stran). Tudi na tem področju so sicer korpusi usvajanja manj problematični, saj zbiranje besedil ponavadi poteka v razredu ali na izpitu, kjer tvorci na pobudo raziskovalca ob njegovih pojasnilih in v njegovi prisotnosti podpišejo privolitveni obrazec oziroma dovoljenje za objavo (primer 3).

Uppsala Student English Project (USE)

Cilj projekta je korpus (računalniška zbirka besedil) gradiva študentov. Uporabljali ga bomo za jezikovne raziskave, poučevanje in izdelavo učnih materialov. Vsi prispevki so prostovoljni, z odstranitvijo imen pa bomo zagotovili anonimnost. Na vsa vprašanja lahko odgovorita koordinatorki projekta, Margareta Westergren Axelsson in Ylva Berglund.

Privolitveni obrazec

Projektu 'The Uppsala Student English Project' dovoljujem, da uporablja moje gradivo za raziskave, poučevanje, objavljanje in predstavitve (konference, delavnice itn.). Strinjam se, da lahko v celoti ali delno objavi moje gradivo v različnih oblikah, tako tiskanih kot elektronskih.

Podpis:
Ime:
Datum:

Primer 3: Privolitveni obrazec za korpus USE
(spletna stran USE, prevedla M. S.)

Posebno zamuden pa je **tehnični del** zajema besedil, torej digitalizacija. V zadnjem času vedno pogosteje to naredijo tvorci sami. Za korpusa USE in JPU so učeči se besedila oddali v elektronski obliki (Axelsson 2000: 155), toda za JPU so bile v njih kasneje popravljene napake, ki naj bi bile očitna posledica napačnega tipkanja, npr. *langauge* namesto *language* (Pravec 2002: 98). V tajvanskem TCLE je bila več kot polovica besedil že digitalizirana, del so optično prebrali in ponovno pregledali, del pa pretipkali (Shih 2000: 93). Spise za japonski CEJL so učenci napisali na računalnik med poukom, za kanadski QLC pa na sprejemnem izpitu (Cobb 2003).

Večino tekstov je torej še vedno treba pretipkati. Za govorne korpuse usvajanja sta pretipkavanje in transkribiranje edina možnost, a tudi pri pisnih načeloma ni izbire. Najlažje dosegljive in obvladljive vire namreč nudijo priložnosti, kjer je skupaj veliko tujih govorcev, ki pišejo na roko. V češkem korpusu AKCES/CzeSL se celo izogibajo besedil, oddanih v elektronski obliki, saj lahko v njih na jezik študentov vpliva uporaba različnih računalniških orodij (Hana idr. 2010: 18). Optično branje rokopisov ni izvedljivo, pretipkavanje pa je zelo počasno. Biti mora posebej previdno in pozorno, da ne pride do napak pri tipkanju, še večja pa je obratna težava. Domači govorci avtomatično popravljajo napake in nehote natipkajo pravilno obliko, čeprav je v izvirniku napaka. Pri tem lahko pomaga strategija ASK-a, kjer so študentje, ki so besedila tipkali, tekst najprej prebrali in v njem na papirju označili napake. Tako so jih ozavestili in bili pozornejši pri tipkanju. Da bi zagotovili varovanje osebnih podatkov, pa so tipkali samo na enem računalniku v uradnih prostorih (Tenfjord 2006). V korpusih večinoma niso prikazani popravki, ki so jih tvorci v svojih besedilih naredili naknadno – če je tvorec besedo prečrtal, sploh ni pretipkana. Izjema je AKCES/CzeSL, kjer so te popravke označili, saj naj bi bilo to koristno za preučevanje procesa usvajanja jezika (Hana idr. 2010: 18).

Zaradi težav s pridobivanjem dovoljenj za objavo, pa tudi zaradi drugih razlogov je dostop do korpusov pogosto omejen. CLC, LPELC, IBLC in TeleNex so popolnoma zaprti, za raziskovalne namene so na voljo LCLE, ISLE, ESFSLD, USE, ASK, HKUST, MAELC in JEFLL. Komercialno je dosegljiva ena od različic korpusa ICLE. Na internetu so bili še pred nekaj leti brezplačno dostopni samo manjši, običajno manj označeni korpusi: PELE, CEJL, JPU, TELC, MELC. Danes pa je prosto dostopnih tudi precej večjih korpusov: VOICE, EVA, PolyU, HKCSE, WriCLE, FLLOC, FALKO, ICLFI itn.

3.3 Označevanje besedil

V procesu izdelave korpusa je zelo zamudno in zahtevno tudi označevanje besedil. Izogniti se mu je mogoče, če neoznačeni korpus ostane v goli tekstovni obliki. Taki so npr. ELFA, IBLC, JPU, MAELC, TeleNex, VOICE in WriCle. Horváth je tovrstno obliko utemeljil celo kot prednost, saj naj bi se zaradi tega raziskovalec zanašal tako na empirično gradivo kot na intuicijo, ki temelji na njegovi pedagoški praksi in opazovanju (Pravec 2002: 98). Kljub vsemu pa so mnogo prijaznejši

in uporabnejši korpusi z dodanimi jezikoslovnimi in drugimi informacijami, kot so dokumentacija o korpusu, zgradba besedila, osnovni jezikoslovni podatki (npr. povedi, besede, ločila, okrajšave), oblikoslovni podatki in osnovne oblike besed, skladnja, poravnave, termini, kazalke, pragmatika, intonacija ipd. (Erjavec 2004). Označenost olajša iskanje, razširi uporabnost korpusa in omogoča večkratno uporabo ter različne obdelave (Leech 1997: 4–6). Za korpuse usvajanja tujega jezika ni enotnega seznama, kaj označevati. Poleg podatkov o tvorcih besedil (prim. poglavje 3.3.1) so v njih v glavnem označene leme, oblikoslovne in skladenjske oznake (prim. poglavje 3.3.2) ter napake. Prav te so glavna specifika in dodana vrednost korpusov usvajanja (Granger 2004: 129), zato jim je posvečeno poglavje 5. Posebej tiste oznake, ki zahtevajo ročno dodajanje, so včasih dodane zgolj manjšemu delu korpusov.

Označevanje lahko poteka ročno, avtomatsko ali polavtomatsko. Avtomatsko temelji na ročno napisanih pravilih ali avtomatsko naučenih modelih, polavtomatsko pa je ciklično: ročnemu sledi strojno označevanje, nato pa spet ročno preverjanje (Erjavec 2004). Izbira načina označevanja je odvisna od programske opreme in ravni označevanja, napake se denimo označuje samo ročno (prim. poglavje 5.2). Ne glede na način pa se je treba držati določenih smernic. Oznake morajo biti konsistentne, čeprav v dovolj velikih korpusih neizbežne manjše nekonsistentnosti ne pomenijo bistvenih razlik (Hunston 2002: 95). Vedno se mora biti možno vrniti na neobdelani, surovi korpus, izločiti oznake ali jih neodvisno shraniti (Leech 2005: 50). Uporabnik mora imeti dostop do razlage oznak ter do podatkov, kdo, kdaj in kje je označil ter kakšen del besedil je bil označen in pregledan, označevanje pa mora biti oprto na »načela pristopnosti in razumljivosti za uporabnika, ne samo za jezikoslovca« (Čermák 2005: 153). Zavedati se je treba, da oznake nikoli niso edina resnica, temveč le priročno sredstvo. »Uporabnik mora biti opozorjen, da je odločitev, če označevanje sprejme ali ne, povsem njegova« (prav tam). Oznake naj temeljijo na splošno sprejetih in ne na spornejših ali bolj specifičnih jezikovnih teorijah (Leech 1997: 6–7). Nivo označevanja naj bo »kar se le da blizu tistemu, kar je o jeziku v določenem okolju sprejeto s konsenzom, /…/ zato se je pri označevanju potrebno še posebej izogibati kompliciranim teoretičnim modelom« (Gorjanc 2005: 64). Ožje uporabni korpusi pa so lahko označeni specifično glede na svoj raziskovalni problem.

S tehničnega vidika je poleg jezikoslovnih vodil za zagotavljanje boljše dokumentiranosti, preverljive pravilnosti zapisa, enostavnejše uporabe programov za obdelavo, večje izmenljivosti in trajnosti pri označevanju treba upoštevati računalniške standarde za prenos in zapis jezikovnih podatkov (Erjavec 2003: 62). Leta 1986 je mednarodni standard za opisovanje teksta in dokumentov postal standardni posplošeni označevalni jezik SGML (angl. *standard generalized markup language*), kasneje pa ga je nadomestila njegova poenostavljena različica XML oziroma razširljivi označevalni metajezik (angl. *extensible markup language*), razvit za pošiljanje in obdelavo na svetovnem spletu. Za oba so značilni opisno označevanje namesto postopkovnega, koncept tipa dokumenta, katerega definicije določa iniciativa Text Encoding Initiative (TEI), in neodvisnost od strojne ali programske opreme, saj so dokumenti prenosljivi med različnimi strojnimi in programskimi okolji brez izgube informacij. Pomembna prednost standarda XML je, da je popolnoma odprt in dosegljiv na svetovnem spletu. Na voljo je veliko programov, ki so sposobni urejati dokumente XML, preverjati njihovo veljavnost ali jih pretvarjati iz različnih zapisov. Podpira ga večina spletnih brskalnikov, med korpusi usvajanja pa je med drugim uporabljen v korpusih ASK, FRIDA, HKUST, JEFLL, LCLE in SST (Pravec, 2002; Granger 2001; Izumi idr. 2004).

3.3.1 Označevanje podatkov o tvorcu

Podatki o tvorcu in besedilu so največkrat dokumentirani v glavi vsakega dokumenta. Za to lahko uporabljajo lastne oznake (primer 4) ali oznake TEI v jeziku XML (primer 5).

@Begin
@Participants: ELC6WABF002
@Language: Japanese
@Year in School: C1
@Sex: Female
@Age: 19
@Date: 1992-07-15
@Location: Tokai University, Hiratsuka, Kanagawa, Japan
@Writing: Hand
@Time: 10 minutes
@Type: Class Assignment
@Dictionary: No
@Task: Essay
@Correction: No

@Input: 1997-08-31 Yamada Tadao; 1997-09-25 Yamashita Keiko
@Coding:
@Distribution: Permission required from Kojiro Asao before publication.
@Warning:
*ELC6WABF002:Fun of Travel

Primer 4: Glava dokumenta v korpusu CEJL (spletna stran CEJL)

```
<!DOCTYPE pelcra>
<header>
 <filedesc>
 <titlestmt>
    <title>wui2fn08.txt</title>
 <respstmt>
     <name>Agnieska Lenko</name><resp>collected</resp>
 <name>Martin Wynne</name><resp>header</resp>
    </respstmt>
 </titlestmt>
 <publicationstmt>
    PELCRA Project, University of Lodz
 </publicationstmt>
 <authordesc>
 <author>wui2fn08</author>
 <gender>f</gender>
 <age>young adult</age>
 <institution_type>University English Philology Department</institution_type>
 <institution_name>Institute of English Studies, University of Lodz</institution_name>
 <level>2</level>
 </authordesc>
 <textdesc>
 <title>The Sweet Smell of Success (narrative or reflexive)
 </title>
 <medium>w</medium>
 <text_type>student composition</text_type>
 <timed>yes<timed>
 <reference>no</reference>
 <edited>no</edited>
 <date>Mon June 8 1998</date>
 </textdesc>
 </filedesc>
 <revisiondesc>
 <list>
 <item><date>Wed Jan 19 14:39:47 WET 2000</date>First version of header</item>
 </list>
 </revisiondesc>
```

```
</header>
<text>
<head>
'The 21st century... Doomed to failure or cut out for success...?'
</head>
<body>
<p> <s> <w AT0>The <w NN1>twentieth-century-world <c YCOM>, <w PRP>in <w AT0>the <w NN1>vanguard <w PRF>of <w AJ0>technological <c YCOM>, <w AJ0>medical <w CJC>and <w AJ0>economic <w NN1>progress <c YCOM>, <w VBZ>is <w VVN>overwhelmed <w PRP>by <w AT0>a <w AJ0>wide <w NN1>range <w PRF>of <w AV0>newly <w VVN>invented <w CJC>and <w AJ0>improved <w NN2>devils <w PRF>of <w AT0>every <w NN1>kind <c YSTP>. </s>
<s> <w CJS> If <w AV0>wisely <w VVN>used <c YCOM>, <w AT0>the <w AJ0>new <w NN2>technologies <w CJC>and <w AT0>the <w AJ0>new <w NN1>breed <w PRF>of <w AJ0>autonomous <w NN2>robots <w CJC>and <w NN2>computers <w VM0>may <w VVI>bring <w AVP>about <w AT0>a <w AJ0>successful <w NN1>cooperation <w PRP>in <w DT0>all <w NN2>spheres <w PRF>of <w DPS>our <w NN1>life <c YSTP>. </s> </p>
<p> <s> <w AT0>The <w AJ0>everlasting <w NN1>quest <w PRP>for <w NN1>money <w VVZ>makes <w NN0>people <w VVI>work <w AV0>harder <w CJC>and <w AV0>longer <c YSTP>. </s>
</p>
</body>
</text>
```

Primer 5: Glava dokumenta in označeno besedilo v korpusu LPELC (spletna stran LPELC)

Švedski USE ima podatke o tvorcih v Excelovi datoteki (Axelsson 2000: 156), ESFSLD pa ima poseben tip datotek s socio-biografskimi podatki o informatorjih. Zelo pogosto podatke o tvorcu vključuje sámo ime datoteke, ki se zaradi preglednosti in povednosti izpiše ob konkordancah (primer 6). V korpusu USE je v imenu vsake datoteke avtorjeva koda, sledi ji končnica, ki pove, za kateri letnik in vrsto besedila gre: končnica *2012.a2* razkrije, da je študent s kodo *2012* v prvem letniku (*a*) napisal besedilo na drugo temo (*2*) (Pravec 2002: 103). V TCLE oznaka <#nsysu-891-f-DES> pomeni, da gre za opis, ki ga je napisal bruc na univerzi Sun Yat-sen v prvem semestru študijskega leta 1989 (Shih 2000: 93). V korpusu TeleNex ime vsake datoteke vsebuje podatke o prenosniku, stopnji in tečaju, pogojih nastanka besedila in šoli (Allan 2002: 198), v korpusu ESFSLD pa o informatorju, prvem in ciljnem jeziku, številki cikla, katerega del je pogovor, številki sekvence cikla in tipu aktivnosti med pogovorom.

TUR, IN	pap which is in Leyden stowe to having nice
TUR, IN	e to having nice drink there. They had nice
SSP, EL	attic and loft space. The house is so nice

Primer 6: Izpis primerov v LCLE s tvorčevo nacionalnostjo in stopnjo usvajanja v stolpcu na levi (spletna stran LCLE)

3.3.2 Jezikoslovno označevanje

Računalnik zna ločevati besede, toda zaplete se pri različnih oblikah istega leksema, posebej pri jezikih z bogato morfologijo, kakršna je slovenščina. Samostalniki imajo šest različnih sklonov in tri števila, pridevniki poleg tega še tri različne stopnje, glagoli pa tri osebe, tri števila in tri čase. Zaradi tega so korpusi lematizirani. **Lematizacija** poveže odvisne oblike besed z njihovo »krovno«, slovarsko obliko oziroma lemo. Korpus je lematiziran, če ima vsaka pojavnica določeno lemo. Leme se uporabljajo »pri leksikalni analizi besedil, npr. kot iskalna funkcija konkordančnikov, avtomatski gradnji leksikonov, pri strojnem prevajanju, itd.« (Džeroski, Erjavec 2000: 14), z njimi pa je statistična analiza natančnejša in manj odvisna od konkretne oblike besede v tekstu (Atkins idr. 1992: 3). Med korpusi usvajanja tujega jezika so lematizirani ASK, EIC, FALKO in TCLE.

Avtomatski lematizatorji, ki temeljijo na ročno sestavljenih jezikoslovnih pravilih in/ali na strojnem učenju pravil iz že označenih besedil (Lönneker 2005: 196), imajo težave z enakopisnicami in z besedami, ki jih ni v leksikonu, še zahtevnejša pa je lematizacija besedil s številnimi napakami in z za knjižno normo nesprejemljivimi oblikami. To lahko ilustriramo na slovenskem gradivu, ki zaenkrat temelji na besedilih domačih govorcev slovenščine. Tudi v njih se pojavljajo neknjižne besede in pregibanje, polvikanje, zanikanje s tožilnikom, neuporaba dvojine, zamenjevanje mestnika z dajalnikom ipd. (Holozan 2006: 146–147), zato lematizatorji že vsebujejo »nekatere najpogostejše neknjižne oblike (*življenski, nebo* namesto *ne bo*, *jest* namesto *jaz*)« (navedeno delo: 146). Če iskana oblika v leksikonu besednih oblik ne obstaja, sta predvideni dve alternativi. Prva je upoštevanje tipičnih odklonov od norme v sodobnem pisnem jeziku, tako da se *stricom* lematizira v *stric*, *nevem* pa v *vedeti* (Arhar, Gorjanc 2007: 102). Druga možnost je poskus avtomatske konstrukcije leme na osnovi prepoznave besedne končnice, pri čemer se napačno konstruirane leme pojavljajo predvsem pri tujih besedah oz. lastnih imenih. Oblike, ki jih ni mogoče avtomat-

sko uganiti, ostanejo nelematizirane in se vpišejo v seznam, ki je osnova za nadgradnjo leksikona. Prav zaradi tovrstnih oblik bi bilo za morebitno avtomatsko lematizacijo slovenščine kot tujega jezika treba opisani postopek še nadgraditi.

Drugi osnovni in zelo pogosti tip korpusnega označevanja je **oblikoslovno** oziroma **oblikoskladenjsko označevanje**, imenovano tudi slovnično ali morfosintaktično označevanje oziroma označevanje delov govora (angl. *part-of-speech tagging, POS-tagging* ali *word-class syntactic tagging*). Pogosto je tudi v korpusih usvajanja tujega jezika, saj so med drugim tako označeni ASK, FALKO, FFLOC, HKUST, ICLE, InterIta, JASMIN, JEFLL, LPELC, MELD, TCLE, USE in VALICO. Vsaki pojavitvi v korpusu je določena ustrezna kontekstno odvisna morfološka oznaka (Gorjanc 2005: 66), npr. samostalnik ali pretekli deležnik, kar omogoča učinkovito iskanje, preproste skladenjske analize in nadaljnje skladenjsko ali semantično označevanje.

Avtomatski oblikoskladenjski označevalniki besedi glede na sobesedilo določijo ustrezno besedno vrsto in ponavadi delujejo v dveh fazah. V prvi slovar skupaj z oblikoslovnim analizatorjem za vsako besedo vrne njene možne oblikoslovne oznake. V drugi fazi je določena pravilna oznaka besede glede na funkcijo v besedilu, oznake pa so pripisane tudi besedam, ki jih ni v slovarju. Pravila za razdvoumljenje je mogoče pisati ročno, a zaradi kompleksnosti in zamudnosti se označevalniki ponavadi naučijo statističnega modela jezika iz predhodno ročno označenega korpusa (Erjavec 2003: 71).

Razlike med jeziki so velike predvsem zaradi pregibnosti. Za angleščino je tipično okoli petdeset, za slovenščino pa prek tisoč oznak (Erjavec 2003: 70). Prekrivanje med ročno in avtomatsko pripisanimi oznakami je pri označevanju angleščine večinoma 95-odstotno natančno (Kennedy 1998: 221), za slovenske označevalnike pa so številke 83,6–89,4 % pri oblikoslovnih oznakah in 94,4–97,6 % pri oznakah besedne vrste (Džeroski idr. 2000; Lönneker 2005: 205; Erjavec, Ledinek 2006: 172; Erjavec, Krek 2008: 52; Arhar 2009).

Podobno kot pri lematizaciji tudi pri oblikoslovnem označevanju jezika učečih se težave povzročajo napake v besedilih, saj so označevalniki naučeni na besedilih domačih govorcev. Programi za angleščino so občutljivi predvsem na oblikoskladenjske in zapisne napake, zato so razmeroma uspešni pri besedilih učečih se na izpopolnjevalni stopnji (Granger 2004: 128). Nekateri pa menijo, da bi bilo pri začetnikih včasih treba na podlagi konkretnega besedila z

netipičnimi strukturami razviti specifičen nabor oblikoslovnih oznak (Meunier 1998: 21).

Oblikoslovnim oznakam lahko dodamo še skladenjske oziroma pomenskoskladenjske. **Skladenjsko označevanje** ali parsing besedam pripisuje skladenjske vloge, loči odvisnike in druge sintaktične strukture ter besedilom določi skladenjsko strukturo. Tako označeni korpusi nudijo vpogled v to, kako deluje posamezna beseda znotraj konteksta (Kennedy 1998: 232), in omogočajo statistični pregled »distribucije skladenjskih kategorij na velikem vzorcu besedil dejanske jezikovne rabe« (Erjavec, Ledinek 2006: 162). Pogosto so ponazorjeni s skladenjskim drevesom oziroma z drevesnico, v kateri je »razločevalno opredeljen tip površinskoskladenjske odvisnosti vsake pojavnice v povedi v razmerju do njenega neposredno nadrejenega elementa« (navedeno delo: 163).

Skladenjsko označevanje poteka v treh stopnjah. Prva, morfološka, obsega oblikoslovno označevanje in lematizacijo, druga, analitična, je vezana na označevanje stavčne funkcije, zadnja, tekstogramatična, pa označuje besedilo z vidika pomenske skladnje. Danes je na zadnjem nivoju več kot štirideset različnih oznak, poleg udeleženjskih vlog, različnih tipov okoliščin in modifikacij tudi členitev po aktualnosti (Gorjanc 2005: 70).

Skladenjsko označenih korpusov ni veliko, saj po mnenju nekaterih ne odkrivajo zares novega, temveč samo potrjujejo oziroma dopolnjujejo teoretična izhodišča (navedeno delo: 71). Avtomatično skladenjsko označevanje je relativno neuspešno in precej zamudno. Slovene Dependency Treebank[26] s 45.000 ročno označenimi besedami (Ledinek, Erjavec 2009: 219) se zgleduje po češkem korpusu Prague Dependency Treebank,[27] ki je eden »najbolj ambicioznih in najbolje dokumentiranih projektov skladenjskega označevanja morfološko bogatih jezikov s prostim besednim redom« (Erjavec, Ledinek 2006: 163). Oznake v korpusih jos100k s 100.000 besedami in jos500s s 500 povedmi pa bodo osnova za avtomatsko skladenjsko označevanje nastajajočega korpusa v okviru projekta Sporazumevanje v slovenskem jeziku (Erjavec idr. 2010; Ledinek, Erjavec 2009: 223; Holozan idr. 2008: 37–65). Tudi med korpusi usvajanja je manj pogosto, tovrstne oznake imajo samo ASK, FALKO, ICLE, JEFLL in TCLE.

26 Prim. http://nl.ijs.si/sdt/, Džeroski idr. 2006.
27 Prim. http://ufal.mff.cuni.cz/pdt/.

Še redkejše je **semantično označevanje**, ki besede klasificira v semantične razrede. Označena so tudi pomenska razmerja med deli besedila, na primer med agensi in dejanji. Semantično označevanje določi smisle besedam, prepreči dvoumnosti in olajša iskanje (Bowker, Pearson 2002: 84), v korpusih usvajanja pa bi med drugim lahko pomagalo zaznati leksikalno šibkejša področja (Meunier 1998: 25). Zaenkrat so se za semantične oznake odločili v korpusih ICLE, JEFLL in TCLE. V govornih korpusih pa večkrat označujejo **značilnosti diskurza**. V SST je več kot trideset oznak razdeljenih v štiri kategorije: oznake za strukturo celotne transkripcije, oznake profila govorca v glavi dokumenta, oznake replik govorcev in oznake za mašila, ponovitve ali premore (Izumi idr. 2004: 34).

3.4 Oblikovanje poskusnega korpusa usvajanja slovenščine kot tujega jezika

3.4.1 Specifikacija in oblika

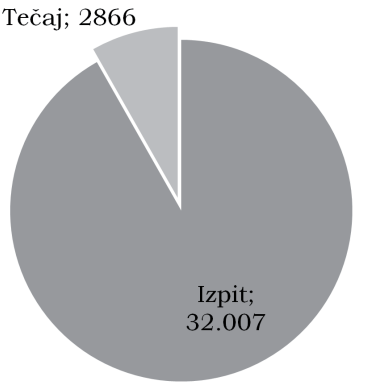
V PiKUST-u je 128 pisnih besedil 119 govorcev v skupnem obsegu 34.873 besed. Prvih jezikov je osemnajst (tabela 9). Pri tem sem upoštevala opredelitve, ki so jih navedli sami govorci, zato ločujem med hrvaščino, srbščino in bosanščino. Deleži večine jezikov imajo manj kot dva tisoč besed, torej so bolj vzorčni in na njih ne morejo temeljiti resni zaključki. Nekoliko reprezentativnejši so le za hrvaščino, srbščino in ruščino.

Jezik	**Število besedil**	**Število tvorcev**	**Število besed**
Angleščina	4	4	1517
Bolgarščina	1	1	173
Bosanščina	3	3	676
Francoščina	4	4	821
Hrvaščina	52	47	13.312
Italijanščina	1	1	591
Kitajščina	1	1	244
Jezik	**Število besedil**	**Število tvorcev**	**Število besed**
Madžarščina	3	3	1015
Makedonščina	3	3	335
Nemščina	2	2	800
Poljščina	1	1	335
Romunščina	1	1	264
Ruščina	13	11	3687
Slovaščina	3	3	1233
Srbščina	31	30	9367

Španščina	2	2	323
Tajščina	2	1	353
Ukrajinščina	1	1	180

Tabela 9: Prvi jeziki v PiKUST-u

Besedila so bila zbrana pretežno na izpitih iz znanja slovenščine kljub prvotnemu načrtu o primerljivih deležih besedil, napisanih na izpitih in na tečaju oziroma doma. Bistveno lažje je bilo namreč zbrati izpitne tekste v arhivu Izpitnega centra. Zato sem PiKUST sestavila po načelu najlažje dosegljivosti. Razmerje med besedili z izpitov in s tečajev je daleč od želenega, saj je bilo slednjih le 8,2 % (slika 9).



Slika 9: Razmerje med številom besed z izpitov in s tečajev v PiKUST-u

Največ besedil so predstavljali spisi z izpitov iz znanja slovenščine na srednji in visoki ravni (prim. poglavje 2.3.3) iz leta 2001. Dva ocenjevalca jih neodvisno ocenita od ena (najnižja ocena) do pet (najvišja ocena). Da sem jih uvrstila v korpus, je morala biti naloga ocenjena z najmanj tri, saj je bilo pri spisih z nižjo oceno napak preveč, da bi bilo označevanje še pregledno, postavlja pa se tudi vprašanje, kje pri njih je meja med slovenščino in drugimi južnoslovanskimi jeziki (prim. Požgaj Hadži, Ferbežar 2001; Pirih Svetina, Ferbežar 2005). Pridobiti sem želela besedila govorcev s čim več prvimi jeziki, da bi preizkusila možnosti za označevanje najrazličnejših napak. Izkazalo se je, da so večinoma iz hrvaškega ali srbskega govornega okolja, zato nisem vključila vseh njihovih tekstov, da ne bi bili v preveliki številčni prevladi. Pač pa sem zaradi pestrosti prvih jezikov pri govorcih, katerih prvi jezik ni bila hrvaščina oziroma srbščina, vključila tudi tekste, ki so bili ocenjeni zadovoljivo ali celo nezadostno, in na to opozorila v glavi dokumenta.

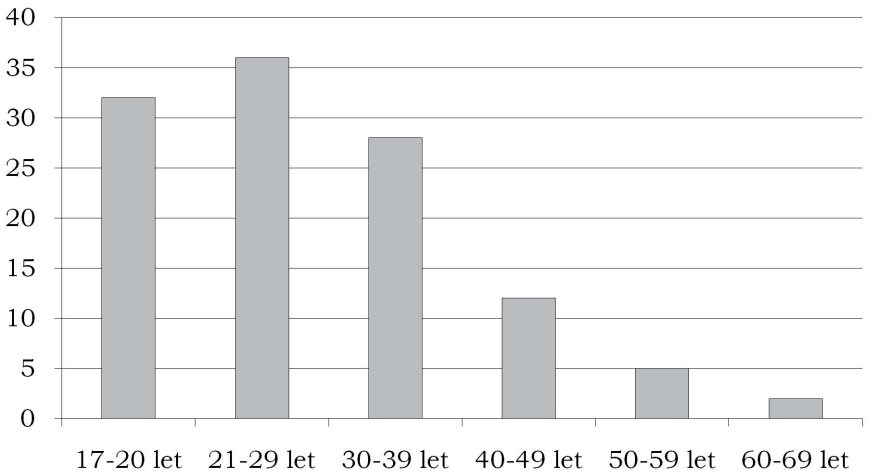
Da bi dobila besedila, napisana v manj stresnih pogojih kot na izpitu, sem po zgledu tujih korpusov oblikovala navodila za pisanje in obrazec za pridobivanje podatkov o tvorcih, po katerih je spise napisalo pet udeležencev izpopolnjevalnih skupin na poletni šoli slovenščine kot tujega jezika v Ljubljani julija 2006, in sicer doma brez tuje pomoči in brez uporabe slovarjev. Za večjo raznolikost stopenj znanja in okoliščin nastanka sem vključila tudi štiri besedila, ki so jih slovanski govorci začetniki napisali na uvrstitvenem testu pred začetkom popoldanskega tečaja slovenščine spomladi 2006, in pet besedil višjih začetnikov oziroma nižjih nadaljevalcev iz učnega KUST-a.

Ker so viri besedil tako različni, opredelitve stopenj jezikovne zmožnosti govorcev med njimi niso nujno primerljive (tabela 10). Kandidati, ki opravljajo izpit na srednji ravni, naj bi imeli nižjo jezikovno zmožnost kot tisti na visoki ravni. Na srednji naj bi se govorci brez težav znašli v vsakdanjih življenjskih okoliščinah in precej dobro v nepredvidenih govornih položajih, na visoki pa so povsem neodvisni in samostojni pri uporabi slovenskega jezika, obvladajo pisanje tudi ustvarjalnejših sestavkov in upoštevajo pravopisna pravila (Ferbežar, Pirih Svetina 2002: 13–14). Toda ker o ravni opravljanja izpita ne odloča zgolj jezikovna zmožnost kandidatov, temveč predvsem nejezikovne potrebe, kot so študij ali zaposlitev, še zdaleč ni nujno, da izpit na visoki ravni opravljajo samo tisti z najvišjo jezikovno zmožnostjo. Zato je enačenje jezikovne zmožnosti vseh govorcev, ki opravljajo izpit na določeni ravni, zgolj površinsko in ne upošteva individualnih razlik. Podobno je z besedili, ki so nastala v okviru tečajev, na katerih je jezikovna zmožnost na splošno precej nižja, individualna odstopanja pa so še večja. Kandidate, ki so sodelovali na izpitu, sem označila kot izpopolnjevalce, udeležence tečajev pa glede na raven, pripisano njihovi skupini. Uporabnik PiKUST-a se mora zavedati, da so pripisane stopnje jezikovne zmožnosti zelo splošne, pa tudi, da je meja med nadaljevalci in izpopolnjevalci lahko vprašljiva.

	Število besedil	**Število besed**
Slovani začetniki (tečaji)	3	314
Nadaljevalci (tečaji)	6	799
Izpopolnjevalci (tečaji, izpit na srednji in visoki ravni)	112	33.760

Tabela 10: Stopnje znanja v PiKUST-u

Kot zanimivost naj dodam nekaj statističnih podatkov o spremenljivkah tvorcev v PiKUST-u, ki niso bile ključne za vključevanje. 22 % besedil so napisali moški, 78 % pa ženske. Razmerje ne ustreza naravnemu, vendar je značilno tudi za druge korpuse, v korpusu ICLE je npr. 80 % žensk (Granger 2003a: 360). Tvorci so bili stari od 17 do 65 let (slika 10), povprečna starost je bila 29,8 let.



Slika 10: Starost tvorcev v PiKUST-u

Ker gre predvsem za besedila z izpitov na srednji in visoki ravni, ki so pogoj za študij ali delo v Sloveniji, je izobrazba tvorcev precej visoka. 57 je bilo študentov s srednješolsko izobrazbo, 52, največ med njimi je bilo kandidatov za izpit na visoki ravni, pa jih je imelo univerzitetno diplomo. Trije so imeli končano visoko šolo, eden magisterij, en tvorec s končano osnovno šolo je bil dijak. O izobrazbi petih tvorcev ni bilo podatka. Tudi njihovi poklici so ustrezali virom pridobivanja besedil: 42 % je bilo študentov, 21 % zdravnikov in zobozdravnikov, ostali poklici (24 %) so bili manj enotni. Šest tvorcev je bilo nezaposlenih, o devetih pa ni bilo podatka.

V PiKUST so vključeni spisi, dnevniška besedila, pritožbe in samopredstavitve (tabela 11). Na izpitih kandidati izbirajo med dvema temama, na tečajih so dobili štiri. Samopredstavitev je obvezni del uvrstitvenega testa, pisanje pritožbe je bilo obvezno v enem od izpitnih rokov.

Učno okolje ni bilo pomemben pogoj, tvorci so se učili slovenščino kot drugi ali kot tuji jezik, organizirano in sistematično ali iz okolja. Že omenjeni varovalni element, da ne bi bila vključena v celoti neslovenska besedila, je bila dose-

žena ocena pri pisnem delu izpita, ki je morala biti najmanj tri.

Tema	Število besedil
Besedila z izpitov	
Genetska revolucija	7
Internet	15
Pogovorni jezik v medijih	3
Pomanjkanje časa	27
Prehranjevanje v času bolezni norih krav	4
Pritožba zaradi hrupa iz trgovine	6
Šola za avtomate	4
Tovarna ali otroško igrišče	6
Trendi	6
Trendi in mladi	7
Zabava brez alkohola	19
Znanost – prijateljica ali sovražnica?	4
Učenje kot zabava	6
Besedila s tečajev	
Človek in narava	1
Domišljija v industrijskem svetu	2
Idealna služba	1
Nasilje med mladostniki	1
Moj teden	5
Samopredstavitev	4

Tabela 11: Vrsta in tema besedil v PiKUST-u

PiKUST je brez kontrolnega korpusa. Oblike, za katere se med označevanjem nisem mogla odločiti, ali so ustrezne ali ne, so bile preverjene v FIDI oziroma kasneje v FidiPLUS.

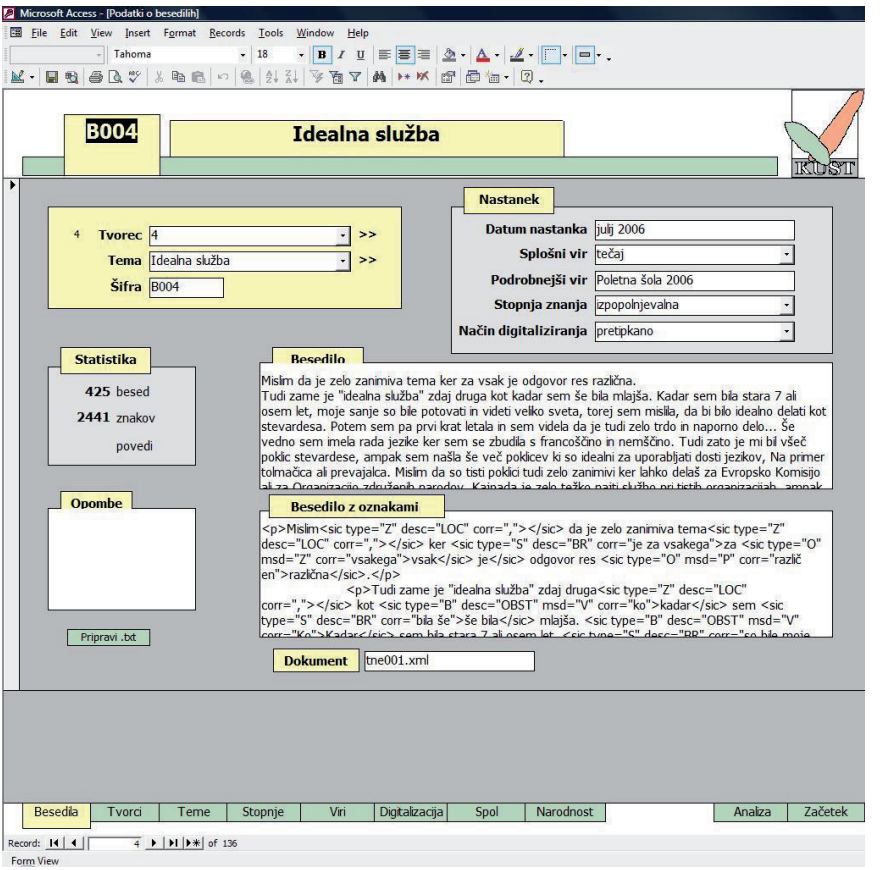
3.4.2 Zajem in označevanje besedil

Vsa besedila so bila napisana na roko, zato sem jih pretipkala. Popravkov ali vstavitev, ki so jih tvorci naredili naknadno, nisem ločevala od preostalega besedila, Če je pisec uporabil samo velike črke, je bilo to ohranjeno, saj bi bilo postavljanje velikih in malih začetnic subjektivno. Res pa je branje tako zapisanih besedil precej težje. Pretipkavanje je za en tekst s približno 200 besedami trajalo od deset do dvajset minut, za korpus z več kot 30.000 besedami torej okrog 30 ur.

Nečitljivi deli besedila so bili v PiKUST-u trije in so označeni z elementom <unclear>. Iz analize podatkov ne smejo biti izključeni, saj so lahko celo del strategije tvorcev, ki opravljajo izpit.

Teksti so zbrani v bazi podatkov v programu Microsoft Access (slika 11) skupaj z vsemi znanimi, za korpus pomembnimi podatki.

Najosnovnejši podatki o vsakem besedilu so razvidni iz imena vsakega dokumenta (tabela 12), iz katerega lahko uporabnik izve osnovne podatke: prva črka je *i* za besedila, nastala na izpitu, ali *t* za besedila s tečajev. Sledita črki, ki označujeta prvi jezik govorca, in tromestno število, ki je zaporedna številka besedila in je vsakemu tekstu pripisana naključno v vrstnem redu, kot so bila besedila označena. Oznaka *ihr037* pove, da je besedilo s številko 037 hrvaškega tvorca nastalo na izpitu, *tan001* pa, da gre za prvo besedilo angleškega govorca s tečajev slovenščine.



Slika 11: Obrazec za vnos v bazi podatkov za PiKUST

Šifra	Pomen
an	angleščina
bg	bolgarščina
bo	bosanščina
fr	francoščina
hr	hrvaščina
it	italijanščina
ki	kitajščina
ma	madžarščina
mk	makedonščina
ne	nemščina
pl	poljščina
ro	romunščina
ru	ruščina
sk	slovaščina
sr	srbščina
sp	španščina
ta	tajščina
uk	ukrajinščina

Tabela 12: Šifre prvih jezikov govorcev v imenu dokumentov v PiKUST-u

Vsako besedilo v PiKUST-u ima označene naslednje podatke:

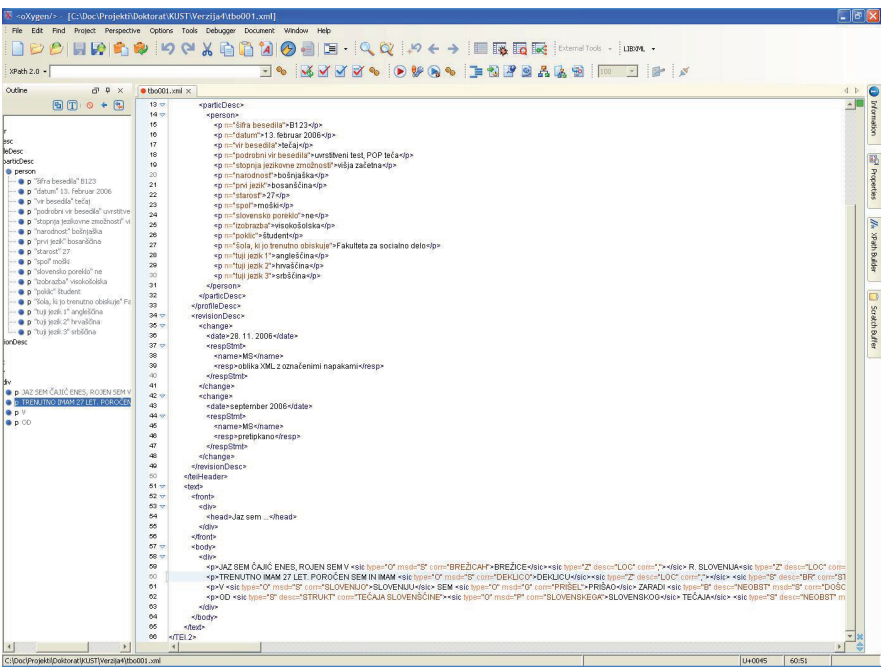
– naslov;
– šifro tvorca;
– šifro besedila;
– datum nastanka;
– vir besedila oziroma okoliščine nastanka;
– podrobnejši opis vira;
– stopnjo jezikovne zmožnosti tvorca;
– prvi jezik;
– starost;
– spol;
– morebitno slovensko poreklo;
– izobrazbo;
– poklic;
– prvi jezik staršev;
– tuje jezike, ki jih zna tvorec;
– podatke o učenju slovenščine;
– podatke o bivanju v Sloveniji.

Informacije so bile pridobljene s posebnim obrazcem ali z izpitne prijavnice, vir besedila in stopnjo jezikovne zmožnosti pa sem določila sama. Tvorci so anonimni, označeni s črko T in številko od 001 naprej. Številke sem pripisala po naključnem zaporedju. Podobno velja za besedila, ki imajo šifre od B001 naprej.

Še pomembnejše od tega je bilo označevanje napak, ki je zamudno in odvisno od njihove količine. Za eno besedilo sem povprečno porabila približno pol ure, pri tekstih z veliko napakami pa do ene ure. Ko sem označila vsa besedila, sem jih še enkrat pregledala zaradi možnih napak pri označevanju in poenotenja nekaterih rešitev. Skupaj je označevanje vsega gradiva trajalo približno osemdeset ur. Čez nekaj mesecev sem označila še nekaj besedil in s tem preverila ustreznost kriterijev ter navodil za označevanje.

Označevanje je potekalo v jeziku XML po priporočilih TEI, in sicer s pomočjo programa Oxygen XML Editor 7.2,[28] ki omogoča pregledno dodajanje XML-oznak in validacijo označenih besedil (slika 12).

Vsi znani podatki o besedilih in govorcih so v elementih <fileDesc> in <profileDesc> v glavi besedila (primer 7). V elementu <revisionDesc> so vse spremembe: kdaj je bil tekst pretipkan, ali in kdaj je bilo pridobljeno dovoljenje za objavo in kdaj so bile označene napake.



Slika 12: Okolje urejevalnika Oxygen XML Editor

28 Prim. http://www.oxygenxml.com.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="http://gandalf.hit.uib.no/ask/ASK.xsl"?>
<!DOCTYPE TEI.2 SYSTEM "http://decentius.aksis.uib.no/corpus/kust/kust.dtd">
<TEI.2>
    <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Nasilje med mladostniki</title>
        <author>T005</author>
      </titleStmt>
    </fileDesc>
    <profileDesc>
      <particDesc>
        <person>
          <p n="šifra besedila">B005</p>
          <p n="datum">julij 2006</p>
          <p n="vir besedila">tečaj</p>
          <p n="podrobni vir besedila">Poletna šola 2006</p>
          <p n="stopnja jezikovne zmožnosti">izpopolnjevalna</p>
          <p n="prvi jezik">angleščina</p>
          <p n="starost">59</p>
          <p n="spol">ženski</p>
          <p n="slovensko poreklo">ne</p>
          <p n="izobrazba">univerzitetna</p>
          <p n="poklic">tolmač</p>
          <p n="prvi jezik staršev 1">angleščina</p>
          <p n="tuji jezik 1">francoščina</p>
          <p n="tuji jezik 2">nemščina</p>
          <p n="tuji jezik 3">danščina</p>
          <p n="tuji jezik 4">nizozemščina</p>
          <p n="tuji jezik 5">irščina</p>
          <p n="koliko časa se uči slovensko">10 let</p>
          <p n="kje se uči slovensko">v Sloveniji</p>
          <p n="koliko časa živi v Sloveniji">5 let, ko ne dela drugje</p>
          <p n="kje živi v Sloveniji">Izola</p>
        </person>
      </particDesc>
    </profileDesc>
    <revisionDesc>
      <change>
        <date>13. 11. 2006</date>
        <respStmt>
          <name>MS</name>
          <resp>oblika XML z označenimi napakami</resp>
```

```
    </respStmt>
   </change>
   <change>
    <date>september 2006</date>
    <respStmt>
     <name>MS</name>
     <resp>pretipkano</resp>
    </respStmt>
   </change>
   <change>
    <date>27. 7. 2006</date>
    <respStmt>
     <name>tvorec</name>
     <resp>dovoljenje za objavo</resp>
    </respStmt>
   </change>
  </revisionDesc>
 </teiHeader>
```

Primer 7: Glava besedila v PiKUST-u

Na začetku vsakega besedila (primer 8) je tema oziroma naslov, pa tudi navodilo, ki so ga dobili tvorci. Ti namreč iz njih pogosto prepisujejo ali parafrazirajo stavke, zato je koristno oceniti, kaj je nekdo tvoril samostojno, kje pa si je pomagal z obstoječim besedilom. Stavek iz PiKUST-a *Svet mladih v sedanjosti oživlja (vrti) življenjski stil (mladost)* postane razumljivejši v kontekstu navodila *Kaj po vašem mnenju »vrti« svet mladih v sedanjosti?*. Ker so navodila del glave dokumenta, seveda niso vključena v izpis konkordanc.

```
 <text>
  <front>
   <div>
    <head>Tovarna ali igrišče?</head>
    <p>V neposredni bližini naselja, kjer živite, nameravajo zgraditi srednje veliko tovarno.
S tem bi se odprlo nekaj novih delovnih mest, ki jih zelo primanjkuje. Tovarno bi postavili na
prostor, kjer je zdaj otroško igrišče. Nekateri starši in okoljevarstveniki zbirajo podpise proti
izgradnji tovarne. Kako se boste odločili? Utemeljite razloge za tovarno in proti njej in jih
utemeljite. Na koncu se odločite, ali boste podpisali peticijo ali ne.</p>
   </div>
  </front>
  <body>
   <div>
```

Jaz sploh ne razumem<sic type="Z" desc="LOC" corr=","></sic> kako <sic type="S" desc="BR" corr="lahko v neposredni bližini naselja ljudje">v neposredni bližini naselja ljudje lahko</sic> zgradijo tovarno. Naše <sic type="S" desc="BR" corr="naselje se">se naselje</sic> nahaja v naravi<sic type="Z" desc="LOC">,</sic> in je še čisto in urejeno. Tudi je dobro za rekreacijo (narava). Če bi <sic type="S" desc="BR" corr="se sedaj nenadoma zgradila">sedaj nenadoma zgradila se</sic> tovarna<sic type="Z" desc="LOC" corr=","></sic> <sic type="S" desc="BR" corr="bi se vse">vse bi se</sic> spremenilo. Čist zrak bi <sic type="O" msd="G" corr="zamenjala">zamenjali</sic> vonj po <sic type="Z" desc="CRK" msd="S" corr="odpadkih">otpatkih</sic> in črn dim. Če <sic type="O" msd="G" corr="bomo">bi</sic> dovolili<sic type="Z" desc="LOC" corr=","></sic> da se zgradi tovarna<sic type="Z" desc="LOC" corr=","></sic> bomo <sic type="B" desc="NEOBST" msd="G" corr="vdihavali"> vdihali</sic> <sic type="O" msd="P" corr="umazani">umazan</sic> in smrdljivi zrak<sic type="Z" desc="LOC" corr=","></sic> ki ga povzroča tovarna, in potem lahko še zbolimo. Ne pišem <sic type="O" msd="Z" corr="tega">to</sic> samo zaradi sebe<sic type="Z" desc="LOC">,</sic> in svoje družine, <sic type="B" desc="OBST" msd="V" corr="ampak">vendar</sic> tudi zaradi drugih ljudi<sic type="Z" desc="LOC" corr=","></sic> ki tukaj živijo, <sic type="B" desc="OBST" msd="R" corr="večina"><sic type="Z" desc="CRK" msd="R" corr="večinoma">većinoma</sic> jih je <sic type="O" msd="P" corr="starih">staro</sic> in <sic type="O" msd="P" corr="bolnih">bolano</sic> </sic>. Naj tovarne delajo čisto na drugem koncu. Tam daleč od ljudi in narave, še manj na otroškem igrišču. Peticijo proti <sic type="O" msd="S" corr="tovarni">tovarne</sic> bom <sic type="Z" desc="CRK" msd="G" corr="podpisala">potpisala</sic><sic type="Z" desc="LOC" corr="."></sic>

 HVALA.

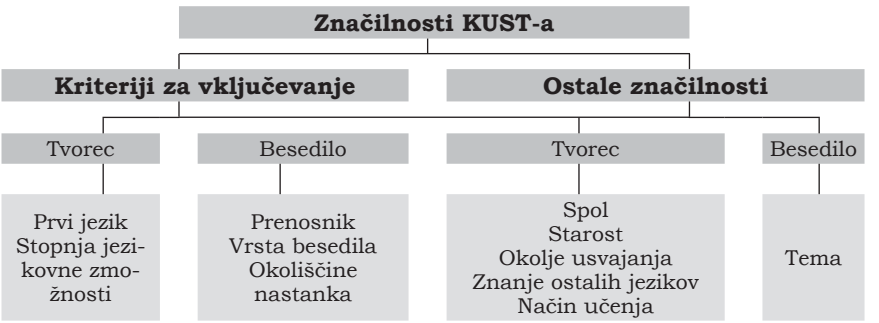
</body>

</text>

Primer 8: Besedilo z oznakami v PiKUST-u

4 NAČELA ZA OBLIKOVANJE KORPUSA USVAJANJA SLOVENŠČINE KOT TUJEGA JEZIKA

Zasnova korpusa usvajanja slovenščine kot tujega jezika temelji na rešitvah obstoječih korpusov, učnega KUST-a in predvsem PiKUST-a. Osnovni kriteriji in značilnosti KUST-a, ki bi bil čim splošneje uporaben (prim. poglavje 6), so prikazani na sliki 13.



Slika 13: Kriteriji in značilnosti KUST-a

Kriteriji za uravnoteženost, podrobneje razloženi v nadaljevanju, so pri tvorcu prvi jezik in stopnja jezikovne zmožnosti, pri besedilu pa prenosnik, vrsta besedila in okoliščine nastanka. Ostale značilnosti so zabeležene v spremni dokumentaciji v glavi dokumenta. Po njih je mogoče iskati, ne odločajo pa o vključitvi. Za tvorca so to spol, starost, okolje usvajanja, znanje ostalih tujih jezikov in način učenja slovenščine, za besedilo pa tema.

Seveda je vse podrobnosti nemogoče točno določiti vnaprej, saj korpusi zaradi nepredvidljivih okoliščin, ki se lah-

ko pojavijo med realizacijo, praktičnih zadržkov in omejitev pogosto nastajajo po metodi zaporednih približkov. Najprej skušamo oblikovati reprezentativen korpus. Ko ga pričnemo uporabljati, ugotovimo prednosti in slabosti, nato ga popravimo in po možnosti izboljšamo, postopek pa ponavljamo do želene idealnosti (Atkins idr. 1992: 4). V tem poglavju so torej dani le predlogi za KUST, ki bi lahko nastal ob idealnih pogojih, morebitne omejitve pa so opisane v poglavju 4.9.

4.1 Prenosnik

Idealni korpus usvajanja tujega jezika bi imel pisno in govorno komponento. Govorno sporazumevanje je »primarna oblika sporazumevanja tako s stališča posameznika kot s stališča družbe, poleg tega je oblika, v kateri se jezik najpogosteje udejanja« (Zemljarič Miklavčič idr. 2009). Zato je »analitična slika nekega jezika, ki elemente zajema samo iz pisnih besedil, izrazito delna in nepopolna« (Stabej, Vitez 2000: 79). Govorjeni jezik je v marsičem tako drugačen od jezika pisnih besedil, da nekateri domnevajo celo o obstoju »dveh vzporednih temeljnih slovnic (oz. jezikovnih zmožnosti) nekega jezika, govornega in pisnega« (prav tam). Vprašanja govorjenega jezika pa so bila v slovenskem jezikoslovju dolgo drugotna. »Zaradi psiholoških oz. kognitivnih in komunikacijskih razlogov, ki so danes v precejšnji meri pojasnjeni, je govor veljal za manj popolno obliko sporazumevanja, torej tudi manj prestižno, zaradi česar je bil tudi neprimeren kot vir analize oz. spoznavanja jezika« (prav tam).

Toda čeprav je govorjenje ključno in čeprav je vanj v glavnem usmerjeno tudi poučevanje tujega jezika, je večina obstoječih korpusov pisna. Razlogi so predvsem oportunistični, saj je pisanje najlažje ulovljiva jezikovna spretnost. Pisna besedila nastanejo s precejšnjim samonadzorom tvorca, lažje in ceneje jih je zajeti, pretipkati in označiti, izdelava korpusa pa je kljub temu zamudna – pri PiKUST-u je za dobrih 30.000 besed trajala približno 100 ur. Tuji govorci v vsakdanji, nešolski komunikaciji sicer na splošno še manj pišejo kot govorci prvega jezika, vendar je pisanje pri učečih se sredstvo za usvajanje in preverjanje ostalih spretnosti ter splošne jezikovne zmožnosti (Ferbežar, Pirih Svetina 2002: 34). Zbiranje avtentičnih govorjenih besedil je precej težavnejše, saj je »potrebnega precej truda, določena tehnična sredstva in (ne najenostavneje) tudi privoljenje govorcev« (Verdonik 2008: 238). Načela za zajem podatkov in transkripcijo govorjenih

besedil domačih govorcev slovenščine so precej kompleksna (prim. Zemljarič Miklavčič 2004, 2006, 2008b) in zajemajo specifične probleme, kot so deikti, besedilni aktualizatorji, diskurzni označevalci in segmentiranje govora (Zemljarič Miklavčič 2008a). Gradnja govornih korpusov je sama po sebi zahtevnejša od pisnih, kar je razvidno tudi iz transkripcij, ki vključujejo številne oznake diskurza (primer 9).

<S1> okay thank you <APPLAUSE> and then , audience has voice </S1>
<S3> of course a couple of questions erm this citizenship how much does it influence the people are they the russians allowed officially to work and everything </S3>
<S2> no yeah i can explain it's erm yeah sorry i i didn't probably didn't mention @@ it's a if you don't have a citizenship you can't vote </S2>
<S3> okay yeah </S3>
<S1> alright </S1>
<S3> i never thought it i mean [(xx)] </S3>
<S2> [no but you can't] vote and you can't actually make any kind of decisions so like for example when the after 1991 the people found okay so it was all decided by estonian side so how things should be done [not considering the people] </S2>
<S3> [yeah of course] makes sense but i mean on a more daily basis </S3>

Primer 9: Del transkribiranega besedila iz korpusa ELFA (spletna stran ELFA)

Če bi se bilo ob tem treba ukvarjati še z označevanjem napak v govorjenih besedilih, je vprašanje, ali bi uporabnost rezultatov upravičevala potrebni časovni in finančni vložek. Zato je energijo smiselno najprej usmeriti v kakovostno oblikovanje pisnega korpusa in se zadovoljiti z utemeljitvijo, da so tudi samo pisni korpusi smiselni »tako kot poligon za tehnološka vprašanja v zvezi z gradnjo in obdelavo korpusov kot tudi za določene analize in aplikacije« (Stabej, Vitez 2000: 80). Šele nato pa se lahko izdela še govorna komponenta KUST-a.

Teksti naj bi bili zbrani presečno, saj so longitudinalni korpusi omejeni na bolj specifične študije. V tehničnem smislu bi moral biti KUST zasnovan tako, da bi bilo mogoče vključevati novo, presečno ali longitudinalno zbrano gradivo.

Pri besedilih, napisanih na roko in kasneje pretipkanih za KUST, bi bilo dobro, da bi imel uporabnik poleg pretipkane različice dostop tudi do optično prebranega originalnega teksta, saj bi tako lahko preveril morebitne napake ali dvoumnosti, ki so nastale pri tipkanju manj čitljivih delov. Poleg tega morajo biti označeni nečitljivi deli (prim. Hana idr. 2010: 18; Rozman, Stritar 2010: 37).

4.2 Vrsta besedil in tema

Korpusi usvajanja tujega jezika največkrat vključujejo spise oziroma eseje, napisane na izpitih, ki naj bi bili ustrezen kompromis med naravno rabo jezika in nadzorovanostjo tvorjenja (prim. poglavje 3.1.5). Naravnost je sicer do neke mere vprašljiva, saj jih tvorci brez zunanje vzpodbude ne bi pisali, vendar je utemeljevanje, ki je v njih najpogostejše, pomemben del komunikacije v tujem jeziku že na zgodnjih stopnjah usvajanja (prim. Ferbežar, Pirih Svetina 2002).

Za KUST naj bi bila vir esejska besedila, vprašanje pa je, kje jih zbirati. Zaradi praktičnosti so smiselne priložnosti, kjer je skupaj veliko v slovenščini pišočih tujih govorcev. Možni viri so izpiti iz znanja slovenščine in tečaji za tujce. Korpus, sestavljen iz besedil s tečajev, pokaže težave tistih, ki se slovenščino učijo organizirano, kar je neposredno uporabno v pedagoški praksi. Nekateri med njimi opravljajo izpit iz znanja slovenščine, mnogi pa nikoli. Pri tekstih tega tipa pa je treba opozoriti na dva problema. Če so spisi ali seminarske naloge nastali doma, oblikovalci korpusa ne morejo biti prepričani, da so jih tvorci napisali sami, brez pomoči referenčnih virov, že obstoječih besedil ali domačih govorcev. Toda tudi če so bili napisani na tečajih, so pogoji zelo različni. Nekateri so bili pri pisanju časovno omejeni, drugi ne, nekateri so lahko uporabljali slovarje in druge vire, drugi ne. Če želimo primerljive pogoje nastanka med besedili z izpitov in s tečajev, je treba pisanje izvesti v okviru tečaja ter ga časovno in referenčno omejiti, kar se da delno izvesti tudi pri pisanju doma.

Po drugi strani se mnogi slovenščine kot tujega jezika nikoli ne učijo organizirano, temveč jo usvajajo iz okolja. Njihova besedila so najlažje dosegljiva na izpitu iz znanja slovenščine, ki ga morajo ponavadi opraviti zaradi zunanjih dejavnikov. Zaenkrat ni zadostnih dokazov, da je njihov vmesni jezik bistveno drugačen od jezika tistih, ki se učijo organizirano, in da so analize tovrstnih podatkov manj pomembne za preučevanje slovenščine kot tujega jezika ali manj relevantne za prenos v pedagoško realnost. Ugotovitve na podlagi gradiva tistih, ki slovenščino usvajajo iz okolja, so uporabne pri razvoju izpitov, pri načrtovanju posebnih tečajev, prilagojenih tej populaciji, in oblikovanju splošnih ali na specifično populacijo usmerjenih učnih gradiv.

Drugi problem je primerljivost stopenj znanja slovenščine. Pri kandidatih z izpitov je jezikovna zmožnost razmeroma

natančno ocenjena. Če so opravili npr. izpit na visoki ravni, stopnja njihove jezikovne zmožnosti ustreza visoki ravni po programu Slovenščina za tujce (prim. Ferbežar, Pirih Svetina 2002). Na tečajih so stopnje manj definirane. Določene so približno na podlagi internega uvrstitvenega testa ali učbenika, ki ga uporablja skupina pri pouku. Zato je možno, da se skupina t. i. izpopolnjevalcev na enem tečaju razlikuje od druge, posamezniki včasih izstopajo, vsi skupaj pa niso nujno primerljivi z izpopolnjevalci z izpita. Ta težava je bila opazna tudi v PiKUST-u.

Zaradi primerljivosti in večjega nadzora je torej v KUST smiselno vključevati zgolj besedila z izpitov, saj je stopnja znanja tako zanesljiveje ocenjena. Ker je vsak del izpita ocenjen posebej, je lahko kriterij za vključevanje tekstov v korpus ocena pisnega dela. Izpitna besedila so bolj primerljiva, zbrana pod enakimi pogoji. Skupna časovna omejitev za praktično in esejsko besedilo, pri katerem kandidati izbirajo med dvema temama, je 75 minut. Ne smejo uporabljati nobenih pripomočkov in naslovov ne poznajo vnaprej. Na visoki ravni kandidati ob prijavi na izpit izberejo naravoslovno ali družboslovno usmeritev. Teme razpravljalnih esejev (primeri 10–12) se omejujejo na določen problem, a ostajajo dovolj splošne, da ne zahtevajo uporabe zelo specifičnega besedišča, ki bi presegalo jezikovno zmožnost kandidatov.

Mesto ali podeželje?

Nekateri ljudje raje živijo v velikih mestih, drugi na podeželju. Navedite prednosti in slabosti življenja v mestu in na podeželju, jih utemeljite s konkretnimi primeri in o temi izrazite svoje mnenje. In h katerim se prištevate vi? Utemeljite.

Primer 10: Tema za pisanje spisa v vzorčnem testu iz slovenščine na srednji ravni

Genetska revolucija

Genetska revolucija obljublja ozdravitev za marsikaterega bolnika, s tem pa tudi bistveno podaljšanje človekovega življenja ali celo njegovo nesmrtnost. Je tak razvoj zaželen? Kaj bi za človeka pomenilo, če bi živel bistveno dlje ali tako rekoč neskončno? (po SP, 6. 1. 2001)

Kakšne bi bile posledice in kakšne prednosti bi prineslo takšno podaljšanje življenja? Bi si ga želeli? Razmišljajte o temi in svoje misli utemeljite s konkretnimi primeri.

Primer 11: Naravoslovna tema za pisanje spisa v vzorčnem testu na visoki ravni

Trendi

So osebnosti, o katerih ljudje iz množice želijo vedeti, kakšno spodnje perilo nosijo, kakšen sijaj uporabljajo na ustnicah in s čim si gladijo gube. So osebnosti, na katerih kozmetična in modna industrija preizkušata svoje izdelke, so osebnosti, ki napovedujejo trende: iz modnega kaosa znajo izločiti prav tisti posebni kos, ki se bo v sekundi prek elektronskih medijev razmnožil v očeh sveta in se nekaj tednov zatem znašel na ulicah. Trendi vrtijo zahodni svet. (prir. po SP, 6. 1. 2001)

Kaj vi menite o modnih trendih in kako se odzivate nanje? Utemeljite svoje mnenje in ga ponazorite s primeri.

Primer 12: Družboslovna tema za pisanje spisa v vzorčnem testu na visoki ravni

Eseji naj bi bili dolgi 120 besed na srednji ali 150–180 besed na visoki ravni, kar je nekoliko manj od besedil v večini drugih korpusov (prim. sliko 6). Vendar bi zaradi neenakih dolžin besedila verjetno prihajalo kvečjemu do makrobesedilnih, ne pa slovničnih in leksikalnih razlik, ki se jih s korpusi najučinkoviteje raziskuje.

Največja prednost besedil z izpitov je, da so v primerjavi s tečaji lažje dosegljiva v velikih količinah. Izpitni roki na posamezni ravni so štirikrat na leto, izpitne pole pa so shranjene v arhivih Centra za slovenščino kot drugi/tuji jezik. V zadnjih letih je izpit letno opravljalo 1300–1900 kandidatov, od tega na srednji in visoki ravni 250–350 (Ferbežar 2007a: 32; Ferbežar 2008: 38; Ferbežar 2009: 46; Ferbežar 2010: 43; Ferbežar 2011: 45). Na prijavnici, ki jo izpolni vsak kandidat, so zbrani osnovni nejezikovni podatki: ime in priimek, spol, rojstni datum in država, materni jezik, izobrazba, poklic, podatki o načinu učenja slovenščine in naslov. Iz tega se da pridobiti podatke o spolu, starosti, prvem jeziku, poklicu, izobrazbi in okoliščinah učenja oziroma usvajanja. Da bi bili za korpus relevantni podatki popolnejši, bi bilo v prijavnico treba dodati še vprašanje o ostalih tujih jezikih, ki jih zna tvorec, saj lahko vplivajo na usvajanje slovenščine. Smiselno bi bilo vključiti tudi privolitveni obrazec za sodelovanje v korpusu, tvorci pa bi bili v KUST-u anonimni in označeni zgolj s šifro.

V PiKUST-u se je pokazalo, da so tako imenovana praktična besedila, ki so prav tako del izpitov, za korpus manj primerna: so precej krajša, vključujejo veliko osebnih podatkov, ki bi jih bilo treba anonimizirati, in jezikoslovno manj zanimive praktične vsebine, kot sta poštni naslov ali datum. Največji problem pa je, da je zaradi eksplicitnosti navodil malo prave jezikovne produkcije (primera 13 in 14).

Na oglasni deski ste videli oglas za počitniško delo (npr. delo anketarja, pomoč v pisarni). Napišite prošnjo za delo, v kateri se na kratko predstavite, opišite svoje izkušnje z delom, za katerega prosite, napišite, zakaj ste ravno vi najprimernejši kandidat za to delo ipd.
Besedilo naj bo dolgo približno 60 besed. Pazite na ustrezno obliko (pošiljatelj, naslovnik …). Vse podatke v besedilu si lahko izmislite. Prosimo, pišite čitljivo. (Besedilo bo ocenjeno z oceno od 1 do 5.)

Primer 13: Navodilo za pisanje praktičnega besedila v vzorčnem testu na srednji ravni

Opazili ste spodnji razpis za delovno mesto in se nanj odzvali. Na podlagi razpisa napišite prošnjo za delo. Besedilo naj bo dolgo približno 80 besed. Pazite na ustrezno obliko (pošiljatelj, naslovnik …). Vse podatke v besedilu si lahko izmislite. Prosimo, pišite čitljivo. (Besedilo bo ocenjeno z oceno od 1 do 5).
Državna revizijska komisija objavlja javni natečaj za uradniško delovno mesto: svetovalec Državne revizijske komisije.
Pogoji za zasedbo delovnega mesta:
- univerzitetna izobrazba ekonomske, pravne ali druge ustrezne smeri,
- 2 leti delovnih izkušenj,
- opravljen državni izpit iz javne uprave oziroma če kandidat nima tega izpita, ga je dolžan opraviti v letu dni od imenovanja v naziv, ali pravniški državni izpit,
- znanje najmanj enega svetovnega jezika,
- poznavanje uporabniških programov v okolju Windows.
Kandidate vabimo, da pošljejo pisne prijave s kratkim življenjepisom v 8 dneh od dneva objave na naslov: Državna revizijska komisija za revizijo postopkov oddaje javnih naročil, Slovenska 54, Ljubljana.

Primer 14: Navodilo za pisanje praktičnega besedila v vzorčnem testu na visoki ravni

Če bi kandidati na izpitu iz znanja slovenščine sodelovali tudi pri KUST-u, bi to lahko povečalo stresnost izpitne situacije in s tem negativno vplivalo na testno performanco. Toda če bi privolitveni obrazec podpisali hkrati s prijavo na izpit, kar je najmanj deset dni pred izpitnim rokom, to na samem izpitu zaradi časovnega zamika verjetno ne bi bistveno prispevalo k stresu. Preprečiti pa je treba vsakršno pogojevanje uspeha na izpitu s sodelovanjem v korpusu. Kandidati, ki bi sodelovali v KUST-u, ne bi smeli imeti nobene prednosti na izpitu, odločitev za sodelovanje pa bi morala biti popolnoma prostovoljna.

Z opisanimi nadzorovanimi pogoji nastanka in z besedili najreprezentativnejših skupin tvorcev (prim. poglavje 4.5) bi KUST ostal odprt za več skupin uporabnikov in aplikacij: za učitelje

slovenščine kot tujega jezika in za sestavljavce učnih gradiv, za neposredno rabo v učilnici, za razvijalce in ocenjevalce izpitov in za vse, ki jih slovenščina kot tuji jezik zanima s teoretičnega stališča ali s stališča jezikovne politike (prim. Stabej 2004).

4.3 Stopnja jezikovne zmožnosti

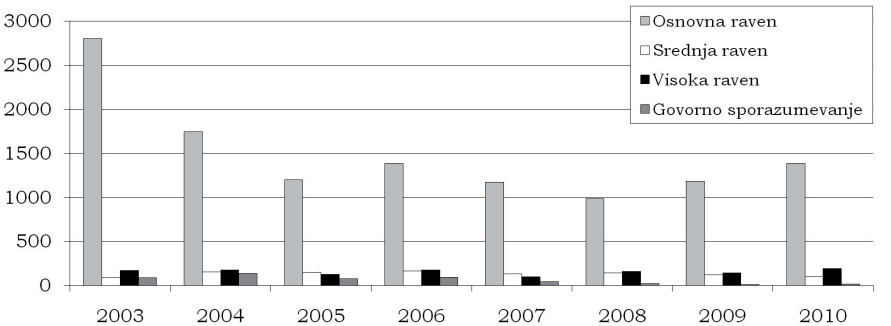
Pri stopnji jezikovne zmožnosti udeležencev naj bi KUST sledil obstoječim korpusom, v katere zaradi kratkosti, tujosti in nekoherentnosti niso vključena besedila začetnikov, temveč samo višje ravni. Teksti začetnikov in nadaljevalcev imajo nemalokrat več napak kot besed; v francoskem delu korpusa ICLE, kjer gre za besedila nadaljevalcev in izpopolnjevalcev, je ena napaka na vsakih 16 besed (Granger 2003a: 542). Označevanje je zamudno, informativnost vprašljiva, saj se je težko znajti med mnogimi neustreznimi oblikami, poleg tega pa je oblika jezikovne zmožnosti začetnikov izrazito prehodna.

Na izpitu iz znanja slovenščine na osnovni ravni, primerljivem s stopnjo B1 po lestvici SEJO (prim. poglavje 2.3.3), se ne preverja zmožnosti samostojnega tvorjenja daljših razpravljalnih besedil (Ferbežar, Pirih Svetina 2002), izpit iz govornega sporazumevanja pa je namenjen nepismenim kandidatom.[29] Zato sta ustrezna vira samo izpita na srednji in visoki ravni, primerljiva s stopnjama B2 oziroma C1 po lestvici SEJO, ki zahtevata samostojno tvorjenje esejskega besedila. Po številčnosti udeležbe na izpitu sicer izrazito prevladujejo kandidati, ki opravljajo izpit na osnovni ravni, a vseeno je bilo v zadnjih letih tudi dovolj udeležencev višjih ravni, ki bi lahko bili sodelovali v KUST-u (tabela 13, slika 14).

	Osnovna raven		**Srednja raven**		**Visoka raven**		**Govorno sporazumevanje**	
2003	2803	89 %	87	3 %	173	5 %	90	3 %
2004	1745	80 %	153	7 %	175	8 %	137	6 %
2005	1199	77 %	151	9 %	178	8 %	79	5 %
2006	1391	76 %	164	9 %	178	10 %	94	5 %
2007	1176	81 %	135	9 %	102	7 %	44	3 %
2008	989	75 %	142	12 %	161	22 %	25	2 %
2009	1185	81 %	124	8 %	145	10 %	12	1 %
2010	1390	82 %	99	6 %	194	11 %	17	1 %

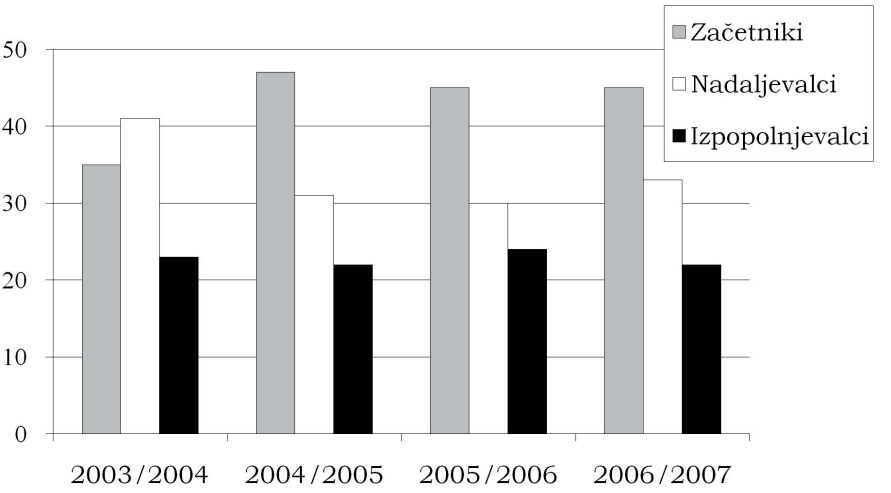
Tabela 13: Število in delež kandidatov na izpitih iz znanja slovenščine po posameznih ravneh (Ferbežar 2004a: 17; Ferbežar 2005: 22; Ferbežar 2006: 24; Ferbežar 2007a: 32; Ferbežar 2008: 38; Ferbežar 2009: 46; Ferbežar 2010: 43; Ferbežar 2011: 45)

29 Prim. http://www.centerslo.net/l2.asp?L1_ID=3&L2_ID=99&LANG=slo (september 2011).



Slika 14: Število kandidatov na izpitih iz znanja slovenščine po posameznih ravneh (Ferbežar 2004a: 17; Ferbežar 2005: 22; Ferbežar 2006: 24; Ferbežar 2007a: 32; Ferbežar 2008: 38; Ferbežar 2009: 46; Ferbežar 2010: 43; Ferbežar 2011: 45)

Tuje govorce, ki ne opravljajo izpita, a so zaradi svojega prvega jezika zanimivi za vključitev v KUST, bi lahko poiskali med udeleženci tečajev slovenščine. Čeprav so stopnje znanja na tečajih Centra za slovenščino kot drugi/tuji jezik ocenjene glede na vsakokratno populacijo tečajnikov, podatki zadovoljivo kažejo razmerja med stopnjami (slika 15). Največ, slaba polovica, je začetnikov, približno tretjina nadaljevalcev in četrtina izpopolnjevalcev.



Slika 15: Deleži začetnikov, nadaljevalcev in izpopolnjevalcev na tečajih slovenščine[30]

Govorce bi bilo najbolje izbrati iz izpopolnjevalnih skupin, njihova besedila oceniti po enakih pogojih kot na izpitu in jih na podlagi tega vključiti v KUST.

30 Podatki so bili pridobljeni v arhivu Centra za slovenščino kot drugi/tuji jezik.

4.4 Ciljni jezik

Ciljni jezik v korpusu usvajanja slovenščine kot tujega jezika je slovenščina. »Kadar gre za organizirano učenje, se vedno uči knjižno slovenščino« (Zemljarič Miklavčič 2001: 10–11). Slovenščina kot tuji jezik je »del SKJ, vsaj kar zadeva načrtovanje, gradivo, nenazadnje v precej razširjeni predstavi tudi končni cilj učenja« (prav tam). Slovenski knjižni jezik kot eksplicitna norma je »rezultat jezikovne kodifikacije in naj bi / omogočal/ (po obliki) regulirano, standardizirano jezikovno sporazumevanje. Eksplicitna norma je ponavadi povezana z uradnim in javnim sporazumevanjem, na taki ali drugačni nadregionalni ravni, in sicer predvsem v pisni (ali brani) obliki« (Stabej 2008: 89).

Načelni dvom o zadostni opisanosti in kodificiranosti slovenskega knjižnega jezika skupaj z vprašanji o implicitni in eksplicitni normi, njenem določanju, razumljivosti in sprejemljivosti (prim. Stabej 2008) zahteva posebno raziskavo. Kot normo za KUST pa lahko privzamemo slovenski knjižni jezik. Morebitne neknjižne oblike, ki so sprejemljive v pogovornem jeziku ali narečju, v korpusu niso sprejemljive in so označene kot neustrezne. Za preverjanje (ne)regularnosti pojavitev sta poleg Slovarja slovenskega knjižnega jezika in drugih jezikovnih priročnikov, ki ne temeljijo na korpusnem gradivu, na voljo referenčna korpusa FidaPLUS oz. Gigafida, ki dajeta koristne informacije o novejših oblikah. V določenih primerih pa je vprašanje o knjižnosti in s tem sprejemljivosti izrazov v korpusu prepuščeno subjektivnim odločitvam označevalcev.

4.5 Izhodiščni jezik

Po narodnosti in jezikovnem ozadju so tuji govorci slovenščine tako raznoliki, da mora KUST vključevati več izhodiščnih jezikov. Za določitev njihovega števila, nabora in deležev sem se oprla na podatke iz popisa prebivalstva, podatke o udeležencih tečajev slovenščine in o kandidatih na izpitih iz znanja slovenščine.

Podatki iz popisa dajo precej natančno število tujih govorcev, za katere je slovenščina potencialni jezik okolja. Po informacijah Statističnega urada Republike Slovenije je ob koncu leta 2005 v Sloveniji živelo 48.968, ob koncu leta 2007 68.621, ob koncu leta 2009 pa 82.316 tujcev, to je oseb s tujim državljanstvom z izdanim dovoljenjem za stalno ali

začasno prebivanje v Republiki Sloveniji, z veljavnim delovnim ali poslovnim vizumom, z začasnim zatočiščem ali s statusom begunca (Ilić idr. 2008: 75). Države, katerih državljani so, so navedene v tabeli 14.

Država oz. celina	Število oseb			
	2005	2007	2009	2011
Evropa	**47.665**	**66.874**	**79.909**	**80.258**
Albanija	35	32	59	
Avstrija	301	295	380	378
Belgija	34	35	62	
Bolgarija	72	780	770	1084
Bosna in Hercegovina	21.943	32.468	39.026	38.836
Češka republika	106	118	141	163
Črna gora		218	554	609
Danska	24	18		
Francija	182	174	184	179
Hrvaška	6955	6976	7775	7738
Italija	417	543	736	870
Kosovo			7928	9034
Madžarska	109	127	156	201
Makedonija	5122	7448	9087	8817
Moldavija			254	273
Nemčija	576	625	742	761
Nizozemska	77	94	113	111
Poljska	145	169	180	175
Romunija	136	225	195	230
Ruska federacija	373	430	530	597
Slovaška			356	451
Srbija in Črna gora	9279			
Srbija		13.782	8782	7561
Švedska	32	53		70
Švica	59	76	80	81
Ukrajina	916	1051	1135	1219
Združeno kraljestvo	157	253	352	391
Druge države	615	884	332	369
Afrika	**72**	**87**	**150**	**202**
Azija	**764**	**947**	**1428**	**1533**
Južna Amerika	**84**	**107**	**150**	**155**
Severna in Srednja Amerika	**317**	**328**	**464**	**520**
Avstralija in Oceanija	**41**	**51**	**75**	**78**
Neznana država državljanstva	**25**	**227**	**140**	
Skupaj	**48.968**	**68.621**	**82.316**	**82.746**

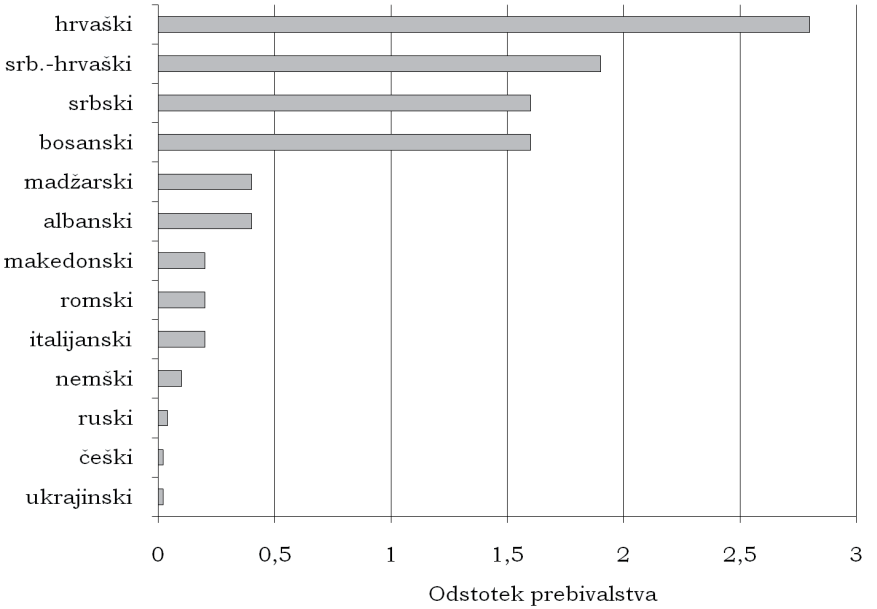
Tabela 14: Tuje prebivalstvo v Sloveniji v letih 2005–2011 (Statistični letopis 2006: 76, 96; Ilić idr. 2008: 98; Iljaš Petrovič idr. 2010; Dolenc idr. 2012)

Podatki o maternem jeziku prebivalstva (tabela 15, slika 16) so iz zadnjega popisa leta 2002, ko je 87,8 odstotka navedlo slovenščino, 12,2 odstotka pa drug materni jezik. Največje skupine govorcev med njimi so hrvaško/srbsko/bosansko, madžarsko, albansko, makedonsko, romsko, italijansko, nemško in turško govoreče, povečujejo pa se skupine z ruščino, ukrajinščino in angleščino (Šircelj 2003: 88–89). Sklepamo lahko, da večina vsaj do neke stopnje govori slovenščino kot drugi jezik ali kot drugi prvi jezik, zato lahko privzamemo, da so tudi njihovi najpogostejši prvi jeziki pomembni pri načrtovanju KUST-a.

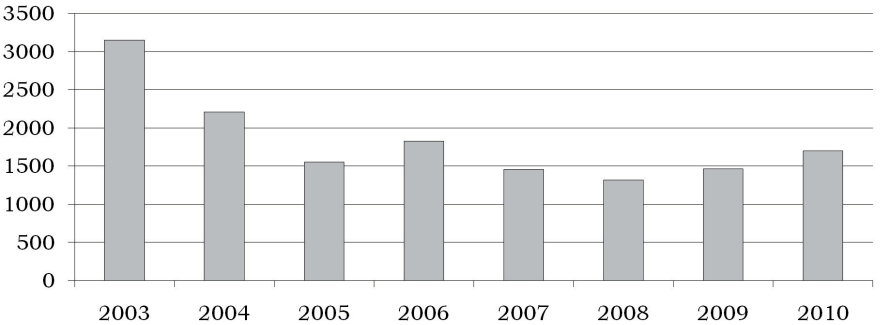
Materni jezik	Število prebivalcev
slovenski	1.723.434
italijanski	3762
madžarski	7713
romski	3834
albanski	7177
angleški	345
arabski	130
bolgarski	159
bosanski	31.499
češki	421
črnogorski	462
francoski	206
hrvaški	54.079
hrvaško-srbski	126
kitajski	216
makedonski	4760
nemški	1628
poljski	267
romunski	251
ruski	766
slovaški	294
srbski	31.329
srbsko-hrvaški	36.265
španski	129
turški	226
ukrajinski	399

Tabela 15: Materni jezik prebivalstva Slovenije po popisu leta 2002 (Šircelj 2003: 97)

Drugi vir podatkov o tujih govorcih so izpiti iz znanja slovenščine. Število kandidatov v zadnjih letih niha (slika 17) in je pogojeno z zunanjimi dejavniki, kot so zahteve za pridobitev državljanstva ali za vpis na univerzo (prim. Ferbežar 2004a: 17).



Slika 16: Najpogostejši materni jeziki (poleg slovenščine) prebivalstva Slovenije ob popisu leta 2002



Slika 17: Število kandidatov na izpitih iz znanja slovenščine (Ferbežar 2004a: 17; Ferbežar 2005: 22; Ferbežar 2006: 24; Ferbežar 2007a: 32; Ferbežar 2008: 38; Ferbežar 2009: 46; Ferbežar 2010: 43; Ferbežar 2011: 45)

Največji delež kandidatov na vseh treh ravneh govori katerega od južnoslovanskih jezikov. Izpit na srednji ravni je leta 2005 opravljalo skoraj 50 % srbsko ali hrvaško govorečih, 20 % je bilo Makedoncev. Na visoki ravni je bilo 40 % srbsko ali hrvaško govorečih in 16 % Makedoncev. Na osnovni ravni je bilo precej rusko, ukrajinsko in albansko govorečih, na srednji in visoki ravni pa so se od ostalih jezikov pojavile albanščina, romunščina, ukrajinščina, portugalščina,

španščina, francoščina, madžarščina, japonščina (Ferbežar 2006: 24). Leta 2007 so na srednji ravni prevladovali hrvaško (35 %) in makedonsko govoreči študentje (18 %), na visoki pa je bilo največ srbsko govorečih (44 %) (Ferbežar 2008: 38). Podobno je bilo leta 2008: na srednji ravni je bilo 42 % hrvaško, na visoki pa 40 % srbsko govorečih kandidatov. Nekoliko več je bilo poljsko, rusko in ukrajinsko govorečih (Ferbežar 2009: 47). Leta 2009 je bilo na srednji ravni 27 % hrvaško, 18 % makedonsko ter po 15 % srbsko in bosansko govorečih, na visoki ravni pa 34 % srbsko, 26 % makedonsko, 12 % hrvaško in 8 % rusko govorečih (Ferbežar 2010: 44). Leta 2010 je bilo na srednji ravni 42 % hrvaško, 15 % makedonsko, 13 % srbsko in 8 % bosansko govorečih, na visoki pa polovica srbsko, petina makedonsko in 9 % hrvaško govorečih (Ferbežar 2011: 45).

Podatki o udeležencih tečajev kažejo, kdo se slovenščino uči formalno in torej prihaja v stik s pedagoškimi delavci. Informacije so bile pridobljene v arhivih Centra za slovenščino kot drugi/tuji jezik. Čeprav ni edina institucija, ki organizira tečaje slovenščine za tujce, je edina specializirana za to področje, zato so njene prireditve najbolj množično obiskane (prim. Kobos, Pirih Svetina 2009).[31] V tabeli 16 so navedeni podatki o udeležencih tečajev od jeseni 2001 do jeseni 2007, kot so jih navedli sami ob prijavi. V obrazec niso vpisovali svojega prvega jezika, temveč narodnost. Imena narodnosti so prevedena glede na originalni zapis na prijavnici, niso pa prilagojena trenutnim družbenopolitičnim razmeram, zato so na seznamu tudi narodnosti, kot sta *afriška* ali *pravoslavna*.

Narodnost	**Število udeležencev**	**Delež v %**
avstrijska	370	9,55
nemška	333	8,57
argentinska	292	7,54
ameriška	247	6,38
italijanska	225	5,81
slovenska	154	3,98
srbska	149	3,85

31 Podatki drugih institucij bi bili seveda lahko drugačni; na tečaju slovenščine po programu za integracijo tujcev na Centru za permanentno izobraževanje Cene Štupar je bilo med 46 udeleženci leta 2010 15 govorcev bosanskega, 5 srbskega, 4 hrvaškega in po 3 ruskega oziroma arabskega jezika, drugi prvi jeziki učečih se pa so bili še angleščina, kitajščina, korejščina, nemščina, španščina, perzijščina, albanščina, filipinščina, francoščina, makedonščina in srbohrvaščina (Zazijal 2010: 86).

Narodnost	Število udeležencev	Delež v %
hrvaška	142	3,67
francoska	135	3,48
španska	129	3,33
britanska	118	3,05
makedonska	111	2,87
češka	107	2,76
kanadska	93	2,4
finska	89	2,3
japonska	80	2,07
belgijska	77	1,99
nizozemska	72	1,86
poljska	68	1,76
portugalska	63	1,63
bošnjaška	63	1,63
ruska	60	1,55
avstralska	50	1,29
švicarska	47	1,21
švedska	44	1,14
irska	39	1,01
madžarska	35	0,9
slovaška	28	0,72
norveška	27	0,7
ukrajinska	26	0,67
turška	25	0,65
kitajska	24	0,62
romunska	21	0,54
danska	20	0,52
brazilska	19	0,49
mehiška	16	0,41
grška	16	0,41
jugoslovanska	14	0,36
bolgarska	13	0,34
kubanska	12	0,31
črnogorska	12	0,31
albanska	11	0,28
nigerijska	11	0,28
izraelska	11	0,28
angleška, indijska, litvanska, venezuelska	po 10	po 0,26
korejska	9	0,23
iranska	7	0,18

Tabela 16: Narodnosti udeležencev tečajev slovenščine 2001–2007

Na tečajih je bilo še po največ šest predstavnikov naslednjih narodnosti: afriška, ameriško-slovenska, angleško-slovenska, beloruska, bolivijska, ciprska, čerkeška, čilenska, dominikanska, egipčanska, estonska, etiopska, filipinska,

ganska, indonezijska, iraška, islandska, južnoafriška, kamerunska, katalonska, kazaška, kenijska, kirgiška, kolumbijska, latvijska, liberijska, lihtenštajnska, madagaskarska, malezijska, maroška, moldavska, nepalska, novozelandska, osetska, pakistanska, perujska, pravoslavna, singapurska, škotska, šrilanška, švedsko-slovenska, tajska, tajvanska, tunizijska, urugvajska, vietnamska, zambijska.

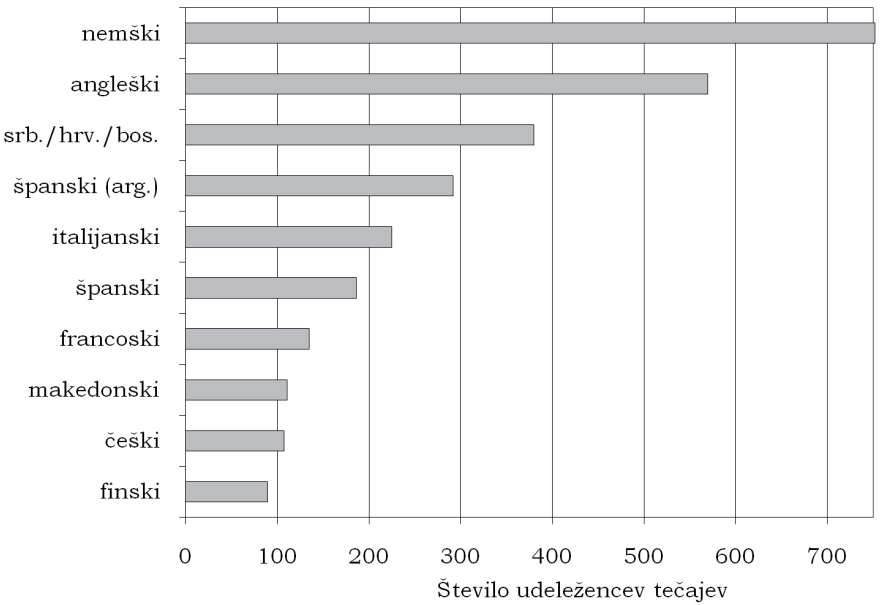
Bolj kot konkretne številke so pomembna razmerja med največjimi narodnostmi oziroma z njimi povezanimi jeziki (tabela 17, slika 18).[32]

Jezik	Število udeležencev	Narodnosti
nemški	752	avstrijska, lihtenštajnska, nemška, švicarska
angleški	570	angleška, angleško-slovenska, ameriška, ameriško-slovenska, avstralska, britanska, irska, kanadska, novozelandska, škotska
srbski/hrvaški/ bosanski	380	bošnjaška, črnogorska, hrvaška, jugoslovanska, srbska
španski (arg.)	292	argentinska
italijanski	225	
španski	186	bolivijska, čilenska, dominikanska, kolumbijska, kubanska, mehiška, perujska, španska, urugvajska, venezuelska
francoski	135	
makedonski	111	
češki	107	
finski	89	
portugalski	82	portugalska, brazilska
japonski	80	
belgijski*	77	
nizozemski	72	
poljski	68	
ruski	60	
švedski	47	švedska, švedsko-slovenska
madžarski	35	
slovaški	28	
norveški	27	
ukrajinski	26	
turški	25	
kitajski*	24	
romunski	21	
danski	20	

32 Pri jezikih, označenih z zvezdico, natančen podatek ni znan. Tisti, ki so označili slovensko narodnost, niso vključeni v tabelo. Pri kanadski narodnosti sta prva jezika angleščina ali francoščina. Točen podatek ni znan, vendar je večina udeležencev tečajev kanadskih Slovencev, ki živijo v angleško govorečem delu države, zato so vključeni k angleščini.

Jezik	Število udeležencev	Narodnosti
grški	16	
bolgarski	13	
albanski	11	
nigerijski	11	
izraelski	11	
indijski*	10	
litvanski	10	
korejski	9	
arabski	8	egipčanska, liberijska, maroška, tunizijska
iranski	7	
ostali	74	afriška, beloruska, ciprska, čerkeška, estonska, etiopska, filipinska, ganska, indonezijska, iraška, islandska, južnoafriška, kamerunska, katalonska, kazaška, kenijska, kirgiška, latvijska, madagaskarska, malezijska, moldavska, nepalska, osetska, pakistanska, pravoslavna, singapurska, šrilanška, tajska, tajvanska, vietnamska, zambijska

Tabela 17: Prvi jeziki udeležencev tečajev slovenščine



Slika 18: Deset najpogostejših prvih jezikov udeležencev tečajev slovenščine

Kot vidimo iz tabele 18 na naslednji strani, dajo različni viri različne podatke o najpogostejših prvih jezikih tujih govorcev slovenščine.

	Letopis 2005	Letopis 2007	Letopis 2009	Letopis 2011	Popis 2002	Tečaji slovenščine 2001–2007
1	hrv./srb./bos.	hrv./srb./bos./čg.	hrv./srb./bos./čg.	hrv./srb./bos./čg.	hrv./srb./bos.	nemščina
2	makedonščina	makedonščina	makedonščina	albanščina	madžarščina	angleščina
3	ukrajinščina	ukrajinščina	albanščina	makedonščina	albanščina	hrv./srb./bos.
4	azijski jeziki	azijski jeziki	angleščina	azijski jeziki	makedonščina	španščina
5	nemščina	nemščina	azijski jeziki	ukrajinščina	romščina	italijanščina
6	italijanščina	bolgarščina	ukrajinščina	bolgarščina	italijanščina	francoščina
7	ruščina	italijanščina	nemščina	angleščina	nemščina	makedonščina
8	angleščina	ruščina	bolgarščina	italijanščina	ruščina	češčina
9	francoščina	angleščina	italijanščina	nemščina	češčina	nizozemščina*
10	poljščina	romunščina	ruščina	ruščina	ukrajinščina	finščina/portugalščina/japonščina (deleži vseh treh jezikov so podobni)

* Govorcev nizozemščine je bilo 72, pripadnikov belgijske narodnosti pa 77. Pri teh je nemogoče določiti, ali govorijo nizozemsko ali francosko, a glede na prebivalstvo Belgije lahko predvidevamo, da je delež približno 60 : 40. S tem podatkom bi bila nizozemščina eden izmed desetih najpogostejših jezikov na tečajih slovenščine.

Tabela 18: Primerjava najpogostejših prvih jezikov po posameznih virih

Hrvaščina, srbščina, bosanščina, makedonščina, nemščina in italijanščina se pojavljajo v vseh virih in so najpogostejši prvi jeziki tujih govorcev slovenščine, torej morajo biti brez dvoma vključene v KUST. Govorci hrvaščine, srbščine, črnogorščine in bosanščine so med najštevilčnejšimi tako med prebivalstvom Slovenije, za katerega slovenščina ni prvi jezik, kot med udeleženci tečajev in izpitov. Makedonščina je bila materni jezik za 0,2 % prebivalcev leta 2002, v letih 2005–2011 je bila v Sloveniji dobra desetina tujcev Makedoncev. Pogosta je tudi med udeleženci tečajev in izpitov. Nemščina je kot materni jezik razmeroma redka (0,1 % tujcev leta 2002, precej več, slaba 2 %, leta 2005, 1,45 % leta 2007, 1,37 % leta 2009 in le malo več leta 2011), zato pa je na tečajih dobrih 18 % udeležencev iz nemško govorečih držav. Podobno velja za Italijane, ki jih je na tečajih slabih 6 odstotkov.

Vprašljivejši za vključevanje so jeziki, ki se pojavijo v manj virih. Če odmislimo, da je angleščina danes vseprisotni globalni jezik in bi že zato morala biti v KUST-u, je delež angleško govorečega prebivalstva od leta 2002 do 2005 narasel na skoraj en odstotek, poleg tega je skoraj najštevilčnejši jezik na tečajih, zato mora biti vključena. Število francoskih govorcev je ves čas približno enako, toda francoščina je dovolj strukturno drugačna od slovenščine, da bi jo bilo zanimivo upoštevati. Češčina ima soliden delež tako na tečajih kot v letopisu 2005. Delež ruščine med slovenskim prebivalstvom se je od leta 1991 do 2002 povečal z 0,01 % na 0,04 % (Šircelj 2003: 89), veliko rusko govorečih je tudi na tečajih, zato bi bilo ruščino smiselno vključiti. Tako so deloma upoštevani tudi ukrajinsko in belorusko govoreči, ki večinoma znajo rusko.

Madžarščina, finščina, portugalščina in nizozemščina so premalo pogoste, da bi jih bilo smiselno vključevati. Albanščina in romščina bi bili zanimivi in relevantni, toda albansko in romsko govoreči se redko formalno učijo slovensko. Ustaviti pa se je treba ob španščini. Čeprav je v okviru prebivalstva Slovenije redka, so špansko govoreči na tečajih slovenščine v samem vrhu najpogostejših jezikov. Sem poleg »pravih« špansko govorečih oseb sodijo slovenski izseljenci v Argentini, ki jih je na tečajih približno 7,5 %. Zanje je slovenščina pogosto prvi jezik v družini, ne pa tudi v okolju. Zato bi bilo nujno oblikovati španski podkorpus, v katerega bi spadali tako izseljenci kot tuji govorci iz špansko govorečih držav. Korpusni podatki bodo pokazali, v kolikšni meri se njihova slovenščina razlikuje od slovenščine argentinskih izseljencev (prim. Bonšek 2010: 89; Ban, Špelko 2006: 80–82).

V Statističnem letopisu (2006: 96) ni natančnejše razčlenitve azijskih jezikov, toda zaradi strukturne drugačnosti bi bilo zanimivo upoštevati tudi kakega od neindoevropskih jezikov. Iz popisa se ponujata kitajščina in turščina, s tečajev pa japonščina. Ta bi bila tudi najbolj praktična, saj se japonski govorci tečajev slovenščine udeležujejo pogosteje od kitajsko ali turško govorečih in so torej najlažje dosegljivi v večjem številu.

Tako bi KUST načrtovali kot strogo nadzorovan, uravnotežen korpus z desetimi izhodiščnimi jeziki, ki bi imeli enakomerno velike deleže (slika 19). Pri tem so hrvaški, srbski, bosanski in črnogorski jezik združeni v en podkorpus, saj naj med njimi ne bi bilo večjih razlik pri usvajanju slovenščine (Balažic Bulc 2004: 77).

hrvaški/ srbski/ bosanski	angleški	italijanski	španski	japonski
make-donski	nemški	francoski	ruski	češki

Slika 19: Zasnova KUST-a z enakomernimi deleži med jeziki

Prikazani izbor jezikov je določen glede na trenutno situacijo, praktični zadržki pa so podrobneje razloženi v poglavju 4.9.

4.6 Velikost

Kot pri ostalih korpusih usvajanja bi bila tudi za KUST želena velikost med 500.000 in 1.000.000 besed. Analize večje količine zbranega gradiva bodo nato pokazale, ali niso pri fleksijskem jeziku, kakršna je slovenščina, razmerja drugačna in bi bila primernejša druga velikost. Hkrati za KUST velja podobno kot za korpus govorjenih besedil (prim. Stabej, Vitez 2000: 81): dokler ne bo natančneje jasno, kako bo oblikovanje potekalo glede zbiranja besedil, tipkanja in označevanja, še nima smisla govoriti o končni velikosti.

Načrtovani izračun velikosti posameznih deležev, preračunan na približno polmilijonski korpus, je v tabeli 19. Vrednosti veljajo za besedila z izpitov na srednji ravni s 120 besedami in na visoki ravni s 150 besedami, kar je najmanjša sprejemljiva velikost. KUST bi torej imel 499.500 besed oziroma 3700 besedil.

	Izhodiščni jezik	**Stopnja**	**Število besed**	**Število besedil**
1	Angleški	B2	22.200	185
		C1	27.750	185
2	Češki	B2	22.200	185
		C1	27.750	185
3	Francoski	B2	22.200	185
		C1	27.750	185
4	Hrvaški, srbski, bosanski	B2	22.200	185
		C1	27.750	185
5	Italijanski	B2	22.200	185
		C1	27.750	185
6	Japonski	B2	22.200	185
		C1	27.750	185
7	Makedonski	B2	22.200	185
		C1	27.750	185

	Izhodiščni jezik	Stopnja	Število besed	Število besedil
8	Nemški	B2	22.200	185
		C1	27.750	185
9	Ruski	B2	22.200	185
		C1	27.750	185
10	Španski	B2	22.200	185
		C1	27.750	185
		Skupaj	**499.500 besed**	**3700 besedil**

Tabela 19: Predlagano zajemanje besedil v KUST-u

4.7 Učno okolje slovenščine

Omejevanje KUST-a na korpus slovenščine kot drugega ali kot tujega jezika se zdi nesmiselno, ker je tudi ločevanje glede na učno okolje zelo težko. Večina učečih se vsaj nekaj časa živi v Sloveniji, ima z njo tesne povezave ali živi v slovensko govorečem okolju, vsaj dokler traja tečaj. Ali je za nekoga, ki se slovensko uči v Parizu in pride dvakrat za en mesec na tečaj v Slovenijo, slovenščina tuji jezik ali v slovenskem okolju postane že drugi jezik? Ali so pripadniki slovenske manjšine v Italiji ali Argentini, ki za vsakdanjo komunikacijo izven družinskega kroga uporabljajo italijanščino ali španščino, govorci slovenščine kot drugega ali tujega jezika? Ker ni jasnega odgovora na ta vprašanja in ker zaenkrat ni dokazov, da bi bile med obema tipoma usvajanja bistvene razlike, naj tudi v KUST-u to ne bi vplivalo na vključitev ali izključitev govorca oziroma besedila.

4.8 Kontrolni korpus

Za slovenščino že obstaja precej korpusov, tudi korpus šolskih besedil Šolar, ki kaže na jezikovne težave mladih domačih govorcev (prim. Rozman, Stritar, Kosem 2010). Ni pa pisnega korpusa neprofesionalnih govorcev, ki bi nudil podatke o tem, kako pisna besedila tvorijo odrasli, ki se s pisanjem ne ukvarjajo profesionalno. Njihovi produkciji bi bila zaradi različnih razlogov (izobrazbe, starosti, nevajenosti pisanja) v nekaterih elementih lahko podobna besedila tujih govorcev (Pirih Svetina, Ferbežar 2005: 5).

Zato bi moral KUST imeti kontrolni korpus, torej podkorpus domačih govorcev slovenščine. Besedila bi bila napisana z enako časovno omejitvijo, z enakimi naslovi in ostalimi pogoji tvorjenja. Govorci bi bili primerljivi tudi demografsko, predvsem po starosti in izobrazbi. Kontrolni korpus naj bi obsegal približno enak delež kot za vsak posamezni prvi jezik, to je približno 50.000 besed ali 370 besedil.

4.9 Praktični zadržki

Opisana načela oblikovanja KUST-a veljajo ob idealnih finančnih, časovnih in kadrovskih pogojih. V realnosti pa so v veliki meri odvisna od konkretnega korpusnega projekta ter z njim povezanih finančnih in drugih zmožnosti, ki so ponavadi omejene. Zato so v nadaljevanju predstavljeni nekateri praktični problemi in predlogi za njihovo rešitev.

Prva je velikost KUST-a, ki je odvisna od materialnih in finančnih danosti. Zagotoviti je treba sodelovanje kandidatov, ki opravljajo izpit iz znanja slovenščine kot tujega jezika, in ob tem pretehtati, kako stresna bi bila zanje dodatna obremenitev. Pod čim bolj primerljivimi pogoji je treba pridobiti tudi besedila predstavnikov skupin tujejezičnih govorcev, ki obiskujejo tečaje in se slovensko organizirano učijo, izpita pa ne opravljajo. Primerno bi jih bilo povabiti na brezplačno, neobvezno opravljanje izpita ali zanje organizirati posebne izpitne roke, na katerih bi pisali pod enakimi pogoji kot redni kandidati. Povračilo za njihovo sodelovanje je spet odvisno od realnih možnosti; lahko bi dobili spričevalo o znanju slovenščine ali zgolj natančno oceno svoje pisne zmožnosti kot sodelujoči v španskem CEDEL2. To je za mnoge učeče se na višjih ravneh zadostna motivacija.

V KUST-u bi bila za vsak prvi jezik uravnotežena deleža tvorcev, ki se slovensko učijo organizirano na tečaju ali neformalno iz okolja, toda zdi se, da bi bilo to uresničljivo samo za nekatere izhodiščne jezike, denimo hrvaščino. Verjetno bi težje našli zadostno število nemško ali angleško govorečih, ki se slovensko naučijo brez udeležbe na tečajih. Zato bi bilo to v KUST-u označeno, ne bi pa vplivalo na vključitev govorca.

Razmerja med skupinami tujih govorcev slovenščine se ves čas spreminjajo in so povezana predvsem z nejezikovnimi dejavniki. Nasprotje stroge načrtovanosti deležev je popolna poljubnost glede prvih jezikov, toda neuravnoteženost gradiva oteži primerjave. Zato bi bilo KUST najbolje zasnovati v dveh vzporednih različicah, uravnoteženi in oportunistični (slika 20). V prvi bi bila besedila v natančno določenih medsebojnih razmerjih, torej z enakim številom besed, enakimi pogoji tvorjenja in enako stopnjo znanja za vse jezike. V oportunistični varianti bi bila vsa zbrana besedila v neuravnoteženih razmerjih. Uporabnik bi bil na to opozorjen in bi ob iskanju izbral različico, ki ga zanima. Med obema bi bil filter, ki bi skrbel za načrtovano uravnoteženost (prim. Stabej 1998). Podobno dvojno zasnovo bo imel milijardni korpus Gigafida,

ki bo imel natančneje strukturiran stomilijonski del, v preostanku pa so kriteriji zajemanja besedil manj fiksni (Logar Berginc, Šuster 2009: 61).

Slika 20: Dvojna zgradba KUST-a: temno osenčeni del je uravnotežen, svetli deli pa so dodani oportunistično

Za nekatere podkorpuse, npr. južnoslovanskega, se zdi tristo besedil dosegljiv cilj, za druge, npr. japonskega, je nekoliko bolj oddaljen. Prav zato je smiselna fleksibilna, a nadzorovana korpusna zasnova, ki omogoča prilagajanje realnim razmerjem. Z njo bi lahko delež katerega koli od desetih jezikov narasel, če bi bilo na voljo več gradiva, hkrati pa bi se ga dalo omejiti na primerljiv delež. Noben jezik ne bi bil vnaprej izključen. Med ostalimi jeziki bi bila npr. finska, albanska, ukrajinska, portugalska besedila, ki bi bila dostopna pri oportunističnem iskanju in bi v osnovi služila bolj kot zanimivost, za informacijo ali široko primerjavo. Če bi število govorcev in besedil tako naraslo, da bi njihov delež dosegel uravnotežene deleže drugih izhodiščnih jezikov, bi bil lahko vključen v nadzorovano različico. Pri dinamični zasnovi je sicer težje vzdrževati uravnoteženost, hkrati pa je bolj demokratična in sledi živi situaciji.

Tudi kontrolni korpus je precej odvisen od zunanjih elementov, saj njegova izdelava terja precej časa in sredstev. Vsaj na začetku je KUST lahko brez kontrolnega korpusa, primerjave z jezikom domačih govorcev pa lahko temeljijo na korpusih FidaPLUS, Gigafida in Šolar (prim. Arhar 2006; Rozman, Stritar, Kosem 2010). Čas in sredstva, ki bi jih bilo treba nameniti izdelavi kontrolnega korpusa, je bolje izkoristiti za čim kakovostnejši korpus usvajanja.

5 NAPAKE

Raziskave usvajanja tujega jezika se osredotočajo na različna področja, povezana z učečim se ali s samim procesom usvajanja (tabela 20), eno izmed prvih pa je bilo preučevanje napak. Te se sicer pojavljajo tudi v besedilih domačih govorcev; češki podatki kažejo, da je v besedilih študentov s 650 besedami mogoče najti tudi 30 resnih napak (Bušta idr. 2009: 110). V jeziku učečih se pa so napake parameter, ki kaže na to, da je »učeči se v fazi pridobivanja novega jezikovnega znanja oz. v fazi širjenja svoje jezikovne zmožnosti« (Rozman 2006: 6). Na njihovi podlagi ocenjujemo stopnjo razvitosti vmesnega jezika oziroma uspešnost pri učenju, saj jih lahko pojmujemo kot kazalce posameznikove trenutne jezikovne zmožnosti v ciljnem jeziku (Corder 1982: 7).

Osredotočenost na učenje			**Osredotočenost na učečega se**
Opisovanje	**Razlaganje**		
Značilnosti jezika učečega se	Zunanji dejavniki učečega se	Notranji dejavniki učečega se	Posamezni učeči se
Napake	Družbene okoliščine	Transfer iz prvega jezika	Splošni dejavniki, npr. motiviranost
Vrstni red usvajanja	Vnos in interakcija	Procesi učenja	Učne strategije
Variabilnost		Komunikacijske strategije	
Pragmatične značilnosti		Znanje jezikovnih univerzalij	

Tabela 20: Okvir za raziskovanje usvajanja tujega jezika (Ellis 1994: 18)

Tudi korpusi usvajanja tujega jezika so se začeli uporabljati v analizi napak, ki je kot proces iskanja pojavitev, značilnosti, vzrokov in posledic neuspešnega jezika (James 1998: 1) še vedno eden od pomembnejših vidikov njihove uporabe. Z analizo napak odkrivamo pojavljanje, naravo, vzroke in posledice »neuspešnosti« v jeziku (navedeno delo: 1). Napake raziskovalcu povedo, kako poteka proces usvajanja, učitelju pa, kako napreduje učenje. Pri njihovi analizi naj »ne bi bil zgolj sposoben odkriti in popraviti napak, pač pa tudi razumeti, zakaj je do njih prišlo, in poskusiti poiskati psihološke, didaktične ter lingvistične vzroke zanje ter jih poskusiti uzavestiti pri učečih se, da bi jih le-ti postopoma tudi odpravili« (Rozman 2006: 18). Najodmevnejši doprinos preučevanja napak v preteklosti je bilo odkritje, da večina napak v tujem jeziku ne odseva strukture prvega jezika učečih se, temveč so bolj podobne napakam, ki jih delajo otroci pri usvajanju jezika, npr. izpuščanje morfemov, dvojno označevanje semantičnih značilnosti, posploševanje pravil (Dulay, Burt, Krashen 1982: 138–139).

Napake so pomembne tudi za učeče se, saj prek njih, večinoma sicer nezavedno, preverjajo svoje hipoteze o ciljnem jeziku in se s tem učijo. Kot strategijo pri usvajanju jezika in sporazumevanju ter učenju nasploh jih uporabljajo tako otroci kot odrasli (Corder 1974a: 25; Brown 1980: 164). Ljudje tvorijo izjave, za katere vedo, da so napačne, ker je to kljub vsemu najboljša možnost, da posredujejo svoje sporočilo (Edge 1989: 8). Učeči se »prevzema posamezne elemente iz drugih jezikov, ki jih še pozna/zna, tvori analogije, postavlja hipoteze o slovničnih pravilih in eksperimentira z jezikom, da bi lahko realiziral svoje namere in potrebe po izražanju v tujem jeziku« (Rozman 2006: 16). S postopnim poskušanjem, napakami in preverjanjem hipotez se učeči se počasi približuje jezikovnemu sistemu domačih govorcev (Brown 1980: 163). Natančen opis napak mu pomaga k večji natančnosti v tujem jeziku in s tem pripomore k enemu od osrednjih ciljev poučevanja tujih jezikov (Granger 2003b: 466).

Kot veja uporabnega jezikoslovja in sodobna metoda raziskovanja usvajanja tujega jezika se je analiza napak uveljavila v šestdesetih in sedemdesetih letih dvajsetega stoletja, ko je nastala tudi večina najpomembnejših teoretičnih del o tej temi. Behavioristično stališče, da so napake neustrezni odzivi na dražljaje, ki se jim je treba izogibati, je zamenjal pogled nanje kot na sistematičen pojav, nujen v procesu usvajanja (Tono 1999). Jezik učečih se so primerjali s ciljnim jezikom, rezultat pa je bila informacija o jezikovnem znanju.

Pomembno je bilo popravljanje napak, ki lahko pomeni zgolj dajanje povratne informacije, da je v besedilu napaka, s spreminjanjem jezikovne oblike lahko vodi k popravljeni obliki ali pa z globljim ukrepom odpravi pojavljanje določenega tipa napake pri učečih se (James 1998: 237).

Analiza napak je bila pomembna v preučevanju usvajanja tujega jezika, vendar zaradi metodološke nepopolnosti ni izpolnila obetov (Dagneaux idr. 1998: 163). Opazovanje zgolj napačne produkcije ne prinaša celostnega pogleda na jezik, ker ignorira pravilni del jezika učečih se (Ellis 1994: 19). Osredotoča se na obliko in ob strani pušča funkcijo jezika (Brown 1980: 163). Ne upošteva tistega, čemur se učeči se izogibajo, in ponuja samo presek stanja v določenem trenutku, ne pa tudi razvoja znanja skozi čas (Ellis 1994: 68). Kmalu so se pokazale še druge konceptualne slabosti analize napak: heterogenost zbranega gradiva (Dagneaux idr. 1998: 164), nejasno ločevanje med opisovanjem in razlago napak, premajhna natančnost in specifičnost pri določanju kategorij napak ter poenostavljeno gledanje na vzroke za napake (Dulay, Burt, Krashen 1982: 141). Zato so jo v osemdesetih letih začeli opuščati. V devetdesetih pa se je spet uveljavila kot pomožna metoda in kot uporaben začetni korak v procesu odpravljanja napak (Pirih Svetina 2005: 41), potreben za nemoteno in uspešno komunikacijo v tujem in ne nazadnje tudi prvem jeziku. Posebej živahno je oživela računalniško podprta analiza napak (angl. *computer-aided error analysis*), ki je neposredno povezana z računalniškimi korpusi.

Vendar se je ob tem treba nenehno zavedati, da je analiza napak metoda oziroma perspektiva za opisovanje podatkov in nikakor ne jezikoslovna teorija (Hagen 2006). Dopolnjuje se s kontrastivno analizo in z analizo transfera. Kontrastivna analiza primerja prvi in ciljni jezik in na podlagi razlik skuša predvidevati, kje naj bi zaradi interferenc prihajalo do napak (prim. poglavje 6.1). Najpomembnejši pojem analize transfera je jezikovni prenos ali transfer, ki je osrednji v modelu usvajanja tujega jezika Roberta Lada. Na podlagi kontrastivne analize dveh ali več jezikov naj bi bilo mogoče napovedovati težave učečih se in določati najprimernejši vrstni red usvajanja. To se je v praksi v glavnem izkazalo kot neinformativno in netočno (Pirih Svetina 2005: 19), zato tudi analiza transfera, ki sopostavlja vmesni in prvi jezik, ne more biti samostojna metoda.

Analiza napak se osredotoča na jezikovno produkcijo učečih se, napake v recepciji so deležne precej manj teoretične pozornosti (Ellis 1994: 47). Težje jih je preučevati, saj so

podatki o recepciji dostopni samo prek sklepanja: na podlagi besednih ali nebesednih odzivov ali prek učiteljeve intuicije (Brown 1980: 163). Zato se tudi pričujoča raziskava posveča samo napakam produkcije.

5.1 Proces analize napak

Analiza napak poteka v več fazah:
- definiranje napak;
- odkrivanje napak;
- označevanje napak (hkrati je lahko tudi opis in kategorizacija);
- analiza napak (hkrati gre lahko tudi za razlago);
- vrednotenje napak.

Za korpuse usvajanja tujega jezika so ključne prve tri faze, torej definiranje, odkrivanje in označevanje napak. Razlaga in vrednotenje prideta na vrsto pri poznejši analizi korpusnih podatkov. V nadaljevanju je posebna pozornost posvečena označevanju napak (prim. poglavje 5.2), ker pa včasih prihaja do prepletanja korakov, so na kratko opisane tudi vse ostale faze.

5.1.1 Definiranje napak

Na začetku analize je treba definirati, kaj sploh je napaka. Opredelitev, da gre za »odstopanja od pričakovanega, ustreznega in pravilnega« (Pirih Svetina 2005: 35), je zelo splošna. Drugi napako definirajo kot jezikovno obliko ali kombinacijo oblik, ki je v identičnem sobesedilu in pod podobnimi pogoji domači govorec po vsej verjetnosti ne bi tvoril (Lennon 1991: 183) oziroma bi jo znal popraviti za sabo. Vprašanje pa je, s katerim jezikom primerjati vmesni jezik učečega se. Nekateri trdijo, da cilj učenja tujega jezika ni posnemanje ali nadomestek domačega govorca in da ga celo sploh ni mogoče doseči, temveč tuji govorec lahko postane le spreten uporabnik drugega jezika. Vendar ni mogoče zanikati, da je usvajanje povezano s približevanjem normi ciljnega jezika. Tujci se učijo slovensko predvsem zaradi uspešnejšega sporazumevanja s Slovenci in manj zaradi sporazumevanja z drugimi tujci v slovenščini, kar velja za angleščino ali druge globalne jezike (Pirih Svetina 2005: 36).

Napaka je razlika med dejansko produkcijo nekoga, ki se jezik uči, in »pričakovano, v idealnih razmerah pa tudi uresničeno produkcijo domačih /…/ govorcev« (Pirih Svetina 2005:

35), jezikovno »neuspel« del besedila ali pojav v besedilu, ki je nenamerno odklonski in ga avtor sam ne more popraviti (James 1998: 1). Namerno odklonske oblike, kakršne se pojavljajo v umetnostnih besedilih, niso napake, nenamerna odklonskost pa se določa s štirimi merili (navedeno delo: 65–76):

- **gramatičnost**:[33] o gramatičnosti odloča slovnica, negramatične oblike pa v nobenem kontekstu niso sprejemljive, ker v jeziku in njegovih paradigmah sploh ne obstajajo;
- **sprejemljivost**: o tem v danem kontekstu presoja naslovnik, zato je ocenjevanje sprejemljivosti vedno subjektivno;
- **pravilnost**: govorci jo presojajo glede na naučeno metajezikovno znanje;
- **tujost** in **neumestnost glede na pragmatični kontekst** (angl. *strangeness*, *infelicity*): nekatere oblike domači govorci prepoznajo kot tuje, glede na pragmatični kontekst pa je neumestno, kadar tuji govorec ne pozna vljudnostnih fraz, kadar je govorno dejanje opravljeno v neustreznih okoliščinah, kadar ima napaka pragmatične posledice, kadar je govorno dejanje prekinjeno ipd.

Mnoge oblike niso napačne same po sebi, temveč zgolj v kontekstu širše jezikovne enote (James 1998: 68), zato je treba upoštevati sobesedilo in tvorčev namen. Glede na to ločimo odkrite in prikrite napake. Odkrite že v obliki kažejo svojo odklonskost, npr. *umrel bo* namesto *umrl bo*. Pri prikritih napakah oblika ne odstopa od norme, a sporočilo ni v skladu z namero, prepoznano iz okoliščin. Odkrite napake ne ustrezajo merilu gramatičnosti in so lokalne, prikrite pa ne ustrezajo sprejemljivosti in so globalne, saj prizadenejo bistveno večji del besedila. Za njihovo identifikacijo je potrebnega precej več teksta (Pirih Svetina 2005: 38). Globalne naj bi bile najpomembnejše (Edge 1989: 2), ker resno ovirajo sporazumevanje (Dulay, Burt, Krashen 1982: 191).

5.1.2 Odkrivanje napak

Druga faza analize napak je njihovo odkrivanje v besedilu. Poteka lahko od zgoraj navzdol, od diskurza k napakam zapisa, ali od spodaj navzgor, od zapisa k diskurzu, kjer se začne pri črkovanju in ločilih, nadaljuje prek slovnice in besedišča do diskurza, ki zajema oblikovanje besedil, slog in podobno (James 1998: 268).

Na splošno je težje opaziti napake v govorjenem kot pisnem besedilu. V njem je namreč koncept napake zaradi

33 Gre za (ne)gramatičnost zgolj znotraj norm ciljnega jezika, s stališča vmesnega jezika so napačne oblike seveda gramatične (Corder 1982: 19).

narave govorjenega diskurza, ki je manj eksplicitno normiran kot pisni, še bolj odprt. Težje je opaziti napake na računalniškem zaslonu kot na papirju in pri samem sebi kot pri drugem tvorcu (James 1998: 91; Rozman 2006: 38, 92–93).

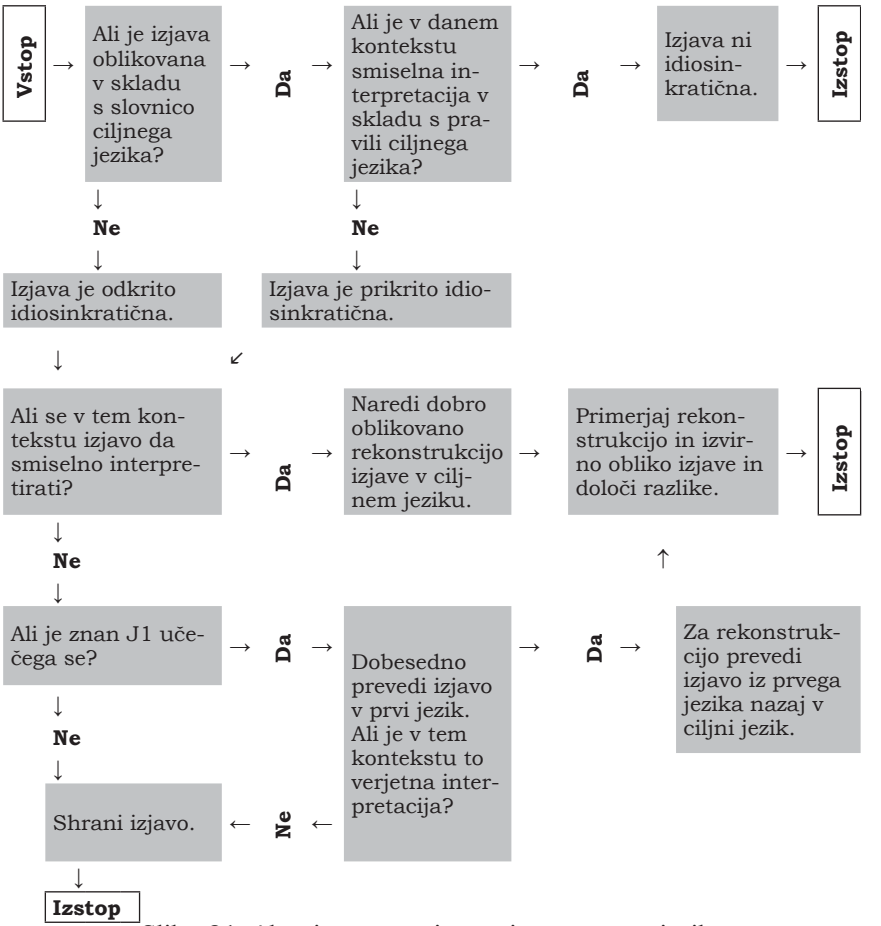
Po nekaterih izkušnjah napake najučinkoviteje odkrivajo in označujejo nedomači govorci ciljnega jezika z visoko stopnjo jezikovne zmožnosti in po možnosti s prvim jezikom, ki je tudi prvi jezik učečih se (Dagneaux idr. 1998: 165). Tuji govorci z visoko stopnjo jezikovne zmožnosti bolje interpretirajo napake, a po drugi strani ne opazijo vseh in so včasih strožji od domačih govorcev (Granger, Thewissen 2006).

5.1.3 Opisovanje napak

Opisovanje napak, ki sta mu v povezavi s konkretnim slovenskim gradivom namenjeni poglavji 5.2.2.2 in 5.3, »temelji na primerjavi oblike izreka, ki ga je sproduciral učeči se, z rekonstrukcijo tega izreka v ciljnem jeziku /.../ oziroma /.../ z rekonstrukcijo pričakovane oblike istega izreka, ki bi ga sproduciral domači govorec ciljnega jezika« (Pirih Svetina 2005: 38). Izraz *rekonstrukcija* sicer ni natančen, saj ne gre za obnavljanje in poustvarjanje obstoječega, vendar se v zvezi s popravljanjem napak pogosto uporablja kot poimenovanje za predvideno pravilno obliko neustreznega izraza. Z opisom napake izrazimo nekaj, kar je bilo samo implicitno, na njegovi osnovi lahko napake štejemo ali razvrščamo. V korpusih usvajanja tujega jezika je opisovanje napak največkrat povezano z njihovim označevanjem oziroma klasifikacijo (prim. poglavje 5.2). Sistem za opisovanje mora biti razdelan in kompleksen, hkrati pa preprost, jasen in prijazen do uporabnikov, to je tako do označevalcev napak kot do uporabnikov korpusa.

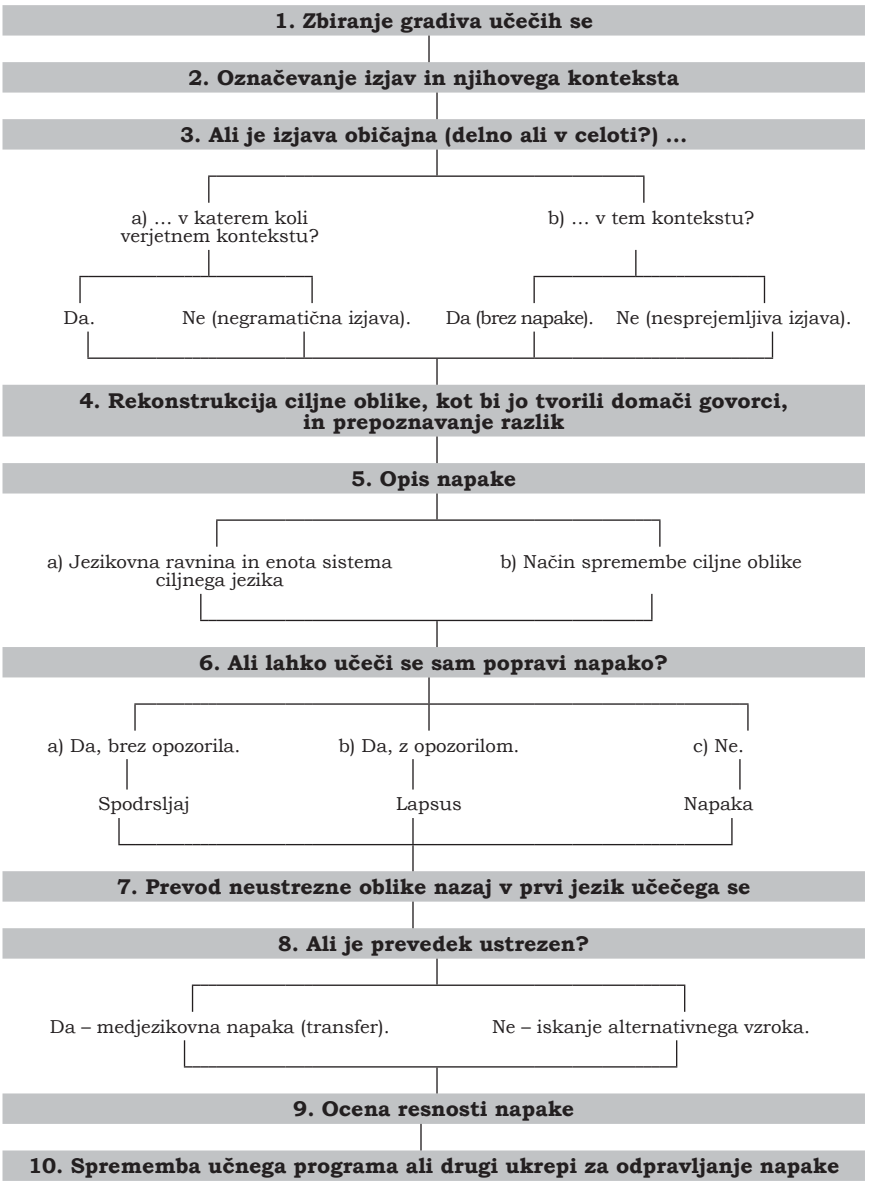
Za opisovanje napak sta Corder in James razvila vsak svoj algoritem, a zaradi abstraktnosti, zapletenosti in shematiziranja sta oba primernejša za teoretične razprave kot za praktično uporabo. Corder napačne izjave opisuje vzporedno s popravljenimi oblikami, tako da sta rezultat vsakega opisa napačna oblika in njena popravljena različica (slika 21).

Če algoritem preizkusimo z izjavo iz PiKUST-a *Menim, da časa vedno ima, če smo dobro organizirani*, ugotovimo, da je prikrito idiosinkratična: oblikovana je v skladu s slovnico ciljnega jezika, vendar v danem kontekstu nima smiselne interpretacije. Prvi jezik učečega se je znan, to je srbščina, in če izjavo prevedemo nazaj v srbščino ter nato spet v slovenščino, pridemo do sprejemljive rekonstrukcije *Menim, da čas vedno je, če smo dobro organizirani*.



Slika 21: Algoritem za opisovanje vmesnega jezika
(Corder 1974b: 167, prevedla M. S.)

Algoritem tega tipa je intuitivno verjetno res v ozadju popravljanja napak tujih govorcev (Brown 1980: 169), vendar sprejemanje besedil ni povezano s stalno primerjavo s prvim jezikom in z analiziranjem vzrokov za napake. Corderjev algoritem napake opisuje, hkrati pa jim išče vzroke, pri čemer se osredotoča samo na napake, ki so posledica transfera iz prvega jezika, in ostale tipe pušča ob strani. Corder se zaveda, da je algoritem idealiziran. Kot posebna problema izpostavlja slovnico ciljnega jezika, s katero naj bi primerjali izjavo učečega se, a ni vedno dorečena, in primere, v katerih je možnih več smiselnih interpretacij izjave. V kontekstu ni vedno dejavnikov, zaradi katerih bi bila določena interpretacija verjetnejša od drugih (Corder 1974b: 168).



Slika 22: Algoritem analize napak (James 1998: 269, prevedla M. S.)

Jamesov algoritem zajema širši proces analize napak. Začne se z zbiranjem in označevanjem izjav, opisovanje in ocenjevanje napak pa imata lahko za posledico ukrepe za popravljanje napak (slika 22).

Tudi James se osredotoča na medjezikovne napake in poudarja zmožnost učečega se, da sam popravlja napake. Prav pri tem pa gre v večini primerov za predpostavljanje, posebej,

če tisti, ki napake opisuje, nima stika z učečim se. Zato je tudi njegova delitev na spodrsljaje, lapsuse in napake (prim. poglavje 5.1.4) v glavnem prepuščena ugibanju.

5.1.4 Razlaga napak

Eden od osrednjih namenov analize napak je razlaga oziroma iskanje vzroka zanje. Mnoge razlage se opirajo na povezavo med napako in tvorčevim poznavanjem pravil ciljnega jezika. Corder loči tri tipe napak (Pirih Svetina 2005: 38):

- **predsistemska napaka**: učeči se ne pozna pravila in ne zna pojasniti, zakaj je izbral določeno obliko;
- **sistemska napaka**: učeči se uporablja in lahko pojasni pravilo, ki pa je napačno;
- **posistemska napaka**: učeči se pozna ustrezno pravilo, vendar ga uporablja nekonsistentno.

Za ugotavljanje sistemskosti napak je treba analizirati majhne jezikovne enote, ne pa jezikovnega sistema učečega se v celoti. Pri glagolskem času je lahko na predsistemski stopnji, raba določenih predlogov pa kaže sistemske napake (Brown 1980: 171).

Selinkerja zanimajo napake, ki so nastale zaradi fosilizacije in so povezane z osrednjimi procesi usvajanja. Če so rezultat prvega jezika, gre za jezikovni prenos, lahko nastanejo zaradi prenosa vadenja (angl. *transfer of training*),[34] zaradi učnih ali sporazumevalnih strategij ali posploševanja (Selinker 1974: 37).

Edge (1989: 9–11) loči tri skupine napak:

- **spodrsljaj** oz. **lapsus** (angl. *slip*) nastane zaradi nepazljivosti. Če bi bil tvorec opozorjen na napako, bi jo znal popraviti;
- **napaka** (angl. *error*): tvorec je ne bi znal popraviti, čeprav se je že učil pravilno obliko;
- **poskus** (angl. *attempt*): komaj razumljive ali nepravilne oblike, ki nastanejo, kadar tvorec še ni usvojil potrebnih pravil.

Glavni kriterij za ločevanje je govorčeva zmožnost, da napake sam popravi, zato je pri nekom določena napaka spodrsljaj, pri drugem napaka in pri tretjem poskus. Tipi napak se lahko prepletajo (navedeno delo: 11). Edge na klasifikacijo naveže tudi načine popravljanja napak pri jezikovnem pouku. Spodrsljaje lahko učenci popravljajo sami, če jim jih učitelj označi, pri napakah lahko pomagajo sošolci,

34 Gre za prenos znanja ali spretnosti, ki jih je nekdo pridobil na določenem področju, na drugo področje.

pri poskusih pa popravek predlaga učitelj (navedeno delo: 50–57).

James (1998: 83–84) Edgeovim kategorijam doda še četrti tip:

- **spodrsljaj** (angl. *slip*): neustrezna oblika pri pisanju ali govorjenju, ki jo tvorec lahko hitro opazi in popravi;
- **lapsus** (angl. *mistake*) tvorec lahko popravi, če je nanj opozorjen. Za popravek lapsusa prve stopnje zadostuje opozorilo, da je v besedilu napaka, pri drugi stopnji pa potrebuje natančnejše informacije;
- **napaka** (angl. *error*): tvorec je ne more popraviti, dokler ne usvoji novega znanja;
- **solecizem** (angl. *solecism*): kršenje naučenih pravil.

Kot je razvidno iz navedenih klasifikacij, je najbolj uveljavljeno ločevanje med lapsusi in napakami. Lapsusi nastanejo zaradi dejavnikov jezikovne performance (Dulay, Burt, Krashen 1982: 139). Delajo jih tudi domači govorci, kadar so utrujeni, neskoncentrirani, močno čustveni ipd., tvorci pa jih lahko popravijo na podlagi povratne informacije. Za razumevanje procesa učenja naj ne bi bili pomembni (Corder 1974a: 123), nekateri celo menijo, da sploh ne bi smeli biti vključeni v analizo napak (Ellis 1994: 51). Tudi N. Pirih Svetina je iz svoje raziskave izključila odstope od norme, ki so povezani z rabo diakritičnih znakov ali ki so posledica lapsusa, kot je »izpust ene črke ali zamenjava dveh črk v besedi« (2005: 123). Napake pa so posledica dejavnikov kompetence (Dulay, Burt, Krashen 1982: 139), torej premalo razvite jezikovne zmožnosti. Kot napačne jih lahko pojmujemo le ob zavedanju, da ne gre za neznanje, pač pa za dokaz trenutne, prehodne zmožnosti učečega se. Tvorci jih lahko popravijo, ko so jim razloženi vzroki zanje (James 1998: 239).

Napake in lapsuse nekateri povezujejo z načinom pridobivanja jezikovne zmožnosti, torej z usvajanjem ali učenjem jezika (tabela 21).

		Učenje	
		+	**–**
Usvajanje	**+**	Pravilna oblika ali lapsus, ki bi ga tvorec znal popraviti.	Lapsus, ki ga tvorec ne bi znal razumsko popraviti, a bi mu to lahko uspelo zaradi občutka za jezik.
	–	Napaka, ki bi jo tvorec znal popraviti, ker ima visoko zavest o jeziku.	Napaka, ki je tvorec ne bi znal popraviti.

Tabela 21: Vrste napak glede na način pridobivanja jezikovne zmožnosti (James 1998: 85)

Ne glede na različna poimenovanja, kriterije in mnenja so omenjena ločevanja predvsem teoretična, v praksi pa so manj uporabna, če ne poznamo jezikovne zmožnosti učečega se ali od njega ne dobimo povratne informacije. Ločevanje med lapsusi in napakami vedno vključuje možnost učiteljeve ali raziskovalčeve napačne predpostavke (Brown 1980: 165), ne upošteva spremenljivosti znanja učečega se (Ellis 1994: 54), v navedene kategorije pa ni mogoče uvrstiti vseh napak.

Ne glede na različna poimenovanja, kriterije in mnenja so omenjena ločevanja predvsem teoretična, v praksi pa so manj uporabna, če ne poznamo jezikovne zmožnosti učečega se ali od njega ne dobimo povratne informacije. Ločevanje med lapsusi in napakami vedno vključuje možnost učiteljeve ali raziskovalčeve napačne predpostavke (Brown 1980: 165), ne upošteva spremenljivosti znanja učečega se (Ellis 1994: 54), v navedene kategorije pa ni mogoče uvrstiti vseh napak.

Taylor zato preseže tvorčevo poznavanje jezikovnih pravil kot edini kriterij za delitev napak in pri razlagi seže širše (Pirih Svetina 2005: 39, Ellis 1994: 57–58):

- **psiholingvistične napake** se nanašajo na sistem znanja tujega jezika in na probleme, ki jih imajo učeči se pri njegovi rabi;
- **sociolingvistične napake** se nanašajo na prilagajanje jezika družbenim okoliščinam;
- **spoznavne (epistemične) napake** se nanašajo na pomanjkljivosti v védenju o svetu;
- **napake, izvirajoče iz oblike diskurza,** se nanašajo na organizacijo informacij v koherentno in kohezivno besedilo.

James (1998: 187–200) tej klasifikaciji dodaja napake zaradi uporabe komunikacijskih strategij, npr. kalke, menjavo koda, parafraziranje, in napake, povzročene zaradi situacije v razredu, to je zaradi učnega gradiva, učiteljevega govora, nalog, učiteljevih pedagoških preferenc ali uporabe slovarja. Sem spadajo tudi inducirane napake, ki jih sproži pomanjkljiva učiteljeva razlaga (Ellis 1994: 60).

Raziskave usvajanja tujega jezika se največkrat osredotočajo na psiholingvistične napake, ki so ali znotrajjezikovne ali medjezikovne kot posledica negativnega jezikovnega prenosa (tabela 22). Posebej medjezikovne naj bi bile zelo pogoste, saj naj bi učeči se ciljni jezik ves čas primerjali s sistemom jezikov, ki jih že znajo (Corder 1982: 12). Rezultati predkorpusnih raziskav pa so pokazali, da je več

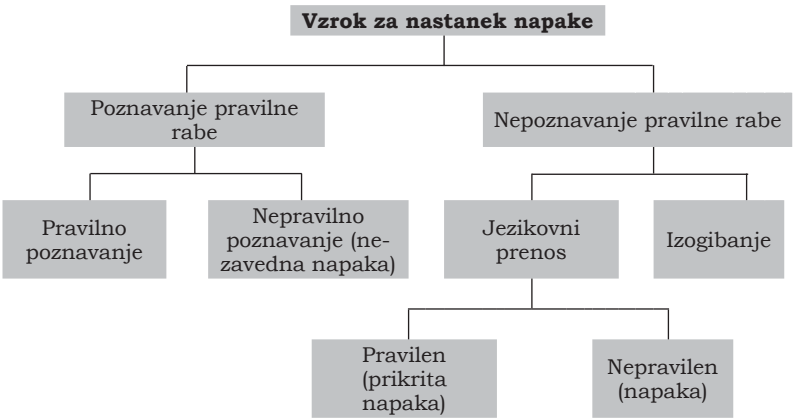
znotrajjezikovnih napak, da začetniki tvorijo več napak zaradi transfera, ki jih je več tudi pri nalogah, ki od učečih se zahtevajo prevajanje. Napake zaradi transfera so pogostejše pri odraslih kot pri otrocih in to na ravni fonologije in besedišča, manj pa na ravni slovnice. Seveda imajo napake lahko tudi več virov (Ellis 1994: 61–62).

Kategorija		**Opis/primer**
Medjezikovne napake (napake zaradi transfera)	Napake po analogiji	Lažni prijatelji, besede, ki so si v različnih jezikih podobne po obliki, a imajo različen pomen.
	Napake zaradi prenosa v strukturi	Uporaba strukture prvega jezika namesto strukture v tujem jeziku.
	Medjezikovno-znotrajjezikovne napake	Napake, ko v prvem jeziku ni razlikovanja, ki obstaja v ciljnem jeziku.
Znotrajjezikovne napake	Napake zaradi prevelikega posploševanja in napačnih analogij	Odklonska struktura na osnovi neke druge strukture ciljnega jezika.
	Napake zaradi neupoštevanja omejitev pravila	Posploševanje pravila na primere, v katerih ne velja.
	Napake kot posledica nerazumevanja struktur	Razumevanje sklonskih oblik v slovenščini kot poljubne in ne obvezne izbire.

Tabela 22: Klasifikacija napak s psiholingvističnim izvorom (Ellis 1994: 58–59, Pirih Svetina 2005: 39–40)

Brown napake klasificira v 48 tipov. Predvideva vse možne kombinacije glede na poznavanje pravil ciljnega jezika (predsistemske, sistemske in posistemske napake), obliko napake (vstavitev, izpust, zamenjava, zaporedje, prim. poglavje 6.2.1) in jezikovno ravnino (zapis/fonologija, besedišče, slovnica, diskurz). Kot izvore omenja medjezikovni in znotrajjezikovni transfer, okoliščine učenja, komunikacijske strategije, izogibanje, neustrezno naučene sporazumevalne vzorce, kognitivni in osebni slog (Brown 1980: 172–180).

Kot je že bilo omenjeno, pa iskanje vzrokov za napake in njihova razlaga ne smeta biti vključena v označevanje korpusnega gradiva, saj sta del poznejše analize. »Zavestno smo se odločili, da ne bomo uporabili ločevanj, kot je med 'napakami' in 'lapsusi' ali med 'medjezikovnimi' in 'znotrajjezikovnimi' napakami, ki jih je težko pripisovati in jih je bolje prihraniti za drugo fazo analize« (Granger 2003b: 467, prevedla M. S.). V nekaj na korpusih utemeljenih raziskavah pa so že iskali vzroke napak, npr. v korpusu Japanese Learners' Corpus (slika 23).



Slika 23: Delitev napak glede na nastanek v korpusu Japanese Learners' Corpus (Sugiura 2000)

T. Balažic Bulc (2004: 86) je iskala vzroke za jezikovni prenos iz srbohrvaščine v slovenščino (prim. poglavje 5.2.2.2). Oblikovala je korpus besedil, ga analizirala in na podlagi podatkov ugotovila, da zaradi razlik med jezikovnima sistemoma nastajajo napake pri zapisovanju vejice in glasovnih premen, pri sklanjatvenih vzorcih in dvojinskih oblikah, pri besednem redu, vezljivosti in besedotvorju, zaradi prenašanja vzorcev z dobesednim prevajanjem pa pri konektorjih in besednih zvezah.

Iz teh dveh primerov je očitno, da se v praksi tipologije prilagajajo konkretnemu gradivu. Zato je nastala splošna klasifikacija z zgolj tremi skupinami: razvojne napake so podobne tistim pri usvajanju prvega jezika, interference odsevajo strukturo prvega jezika, napake tretje skupine pa niso ne razvojne ne posledica interferenc, temveč so unikaten rezultat individualnih značilnosti učečega se ali narave naloge (Ellis 1994: 60; Pirih Svetina 2005: 40). V tej razdelitvi napak je problematično ločevanje med kategorijami (Ellis 1994: 62). Manjka tudi vpliv ostalih jezikov, ki jih učeči se zna, saj je medjezikovni transfer med prvim, drugim, tretjim in ostalimi jeziki lahko zelo močan, čeprav je odvisen od sorodnosti jezikov ali časovnega razmika med učenji (Brown 1980: 173). Predvsem pa je tudi ta klasifikacija preveč odvisna od analitikovega poznavanja učečega se, da bi bila praktična.

5.1.5 Vrednotenje napak

Zadnja faza analize napak, vrednotenje, se nanaša na njihov vpliv. Izmerljiv je posredno, prek vpliva na naslovnika (Pirih

Svetina 2005: 41), na njegovo razumevanje ali čustveno reakcijo. Vrednotenje napak odgovarja na tri glavna vprašanja (Ellis 1994: 63):

– Ali so določene napake ovrednotene kot bolj problematične od drugih?
– Ali pri vrednotenju prihaja do razlik med domačimi in tujimi govorci?
– Katere kriterije govorci uporabljajo pri vrednotenju?

Pri analizi napak je vrednotenje pogosto izpuščeno in poteka ločeno, po posebnih metodah. Tudi instrumenti so drugačni. Ponavadi gre za stavke z napakami brez širšega konteksta, ocenjevalci pa morajo oceniti razumljivost besedil, »resnost« ali »naravnost« napak ali stopnjo motečnosti (prav tam).

Vrednotenje napak je odvisno od stališča vrednotenja, zato se je v raziskavah treba odločiti, ali bodo vrednotili domači ali tuji govorci, jezikoslovci ali nejezikoslovci. Domači govorci naj bi bili najmanj strpni do tuje izgovorjave in besedišča, bolj pa do oblikoskladenjskih in semantičnih napak (Ellis 1994: 63; Pirih Svetina 2005: 41), čeprav so bile v raziskavi tolerantnosti do hrvaških govorcev slovenščine slovnične in leksikalne napake bolj moteče kot naglas (Požgaj Hadži, Balažic Bulc 2005: 252). Tuji govorci naj bi bili tolerantnejši do tujih prvin, a se njihovo razumevanje manjša z vsako napako v sporočilu (Ellis 1994: 63; Pirih Svetina 2005: 41). Slovenske domače govorce v besedilih tujcev marsikaj moti, posebej »pri ocenjevanju znanja slovenščine govorcev iz republik bivše skupne države v primerjavi z govorci drugih jezikov« (Požgaj Hadži, Balažic Bulc 2005: 252). Kljub temu pa so dokaj prizanesljivi. »Če so sporočila njihovih besedil približno razumeli, so besedila zanje postala sprejemljiva in so jih, čeprav z določenimi zadržki, prepoznali kot – slovenska« (Pirih Svetina, Ferbežar 2005: 14). Tudi na izpitih iz znanja slovenščine se napake vrednotijo različno; za nerazumljivost, ki nastane zaradi skladenjske napake ali napačno uporabljene besede, se odbije več točk kot za oblikoslovne ali nekatere pravopisne napake (Ferbežar 2007c: 32).

Pri vrednotenju se upoštevajo tudi ožji slovnični kriteriji, npr. globalnost ali lokalnost napake. Globalne napake, ki zadevajo celotno stavčno organizacijo, najresneje ovirajo komunikacijo (Dulay, Burt, Krashen 1982: 198). Lennon je predlagal dimenziji področje napake, to je velikost sobesedila, ki ga je treba analizirati za identifikacijo, in obseg oziroma velikost enote, ki jo je treba popraviti (Pirih Svetina 2005: 41). Ellis (1994: 67) omenja resnost, razumljivost in

sprejemljivost, James pa pogostnost, opaznost in motečnost. Pogostnost med drugim vključuje podatke, kako pogosto in konsistentno tvorec dela napako in ali se pojavlja tudi pri domačih govorcih. Opaznost napake je odvisna od njene odkritosti oziroma prikritosti in pogostnosti pojavljanja, na motečnost pa vplivajo napovedljivost napake, družbeno razmerje med tvorcem in naslovnikom ter stopnja kršitve družbenih norm (James 1998: 206–223).

5.2 Označevanje in klasifikacija napak

Graditelji korpusov usvajanja tujega jezika se ukvarjajo z definiranjem, odkrivanjem in predvsem označevanjem napak. Ker so te faze analize napak zelo zahtevne, mnogi pozabljajo, da označenost napak ni niti sama sebi namen ali vrednota niti že analiza vmesnega jezika, temveč zgolj priprava terena za pravo raziskovalno delo. Poleg tega ne nudi informacij o vseh vidikih vmesnega jezika, ki vplivajo na tuji zven v besedilih, kot so pretirana ali prenizka raba besed in besednih zvez (Granger 2003b: 475). Označevanje ne sme biti prejudiciranje, temveč zgolj omogočanje raziskovanja. Njegov namen ni preoblikovanje besedil v popolno obliko ciljnega jezika, pač pa le nakazovanje konteksta, v katerem se ponavljajoče pojavljajo napake (Nicholls 2003: 572, 575). Osredotoča naj se na površinske značilnosti izjav učečih se in ne na določanje virov napak (Ellis 1994: 54). Prav to je bil v preteklosti eden od glavnih očitkov analizi napak. Opisovanje se namreč nanaša na rezultat usvajanja jezika, razlaga pa na sam proces usvajanja, ki vedno vključuje dodatno sklepanje in ga je težje objektivno dokazati (Dulay, Burt, Krashen 1982: 141–142).[35]

V tem poglavju je podrobno razčlenjeno označevanje napak kot predpriprava za nadaljnjo analizo korpusnega gradiva. Sistem za označevanje mora biti dovolj informativen, da nudi koristne informacije o napakah, a obvladljiv, da se označevalec v njem ne izgubi. Biti mora splošen, da so kategorije uporabne za več jezikov, prožen, da omogoči dodajanje ali brisanje oznak, in konsistenten, pri čemer so pomembna natančna navodila za označevalce (Granger 2003b: 467). Ta naj temeljijo na jezikoslovnih načelih, a naj bodo osredotočena na praktične potrebe pri označevanju (Gorjanc 2005: 64).

35 Kako težko se je izogniti temu, kaže opis napak H. Dulay, M. Burt in S. Krashena (1982: 150). Napake, pri katerih ima površinska struktura neustrezno zaporedje, opisujejo kot dobesedni prevod struktur prvega jezika, s čimer so jim seveda že pripisali vzrok.

Za razliko od razvitejših in predvidljivejših ravni označevanja, kot sta oblikoslovno označevanje in lematizacija, je označevanje napak izjemno zamudno. Avtomatski označevalniki so neuspešni že na jeziku domačih govorcev, napake učečih se pa so še manj predvidljive in pogosto tako nenavadne, da jih ni mogoče napovedati. Primerjava treh črkovalnikov in slovničnih pregledovalnikov za francoščino je pokazala 60–80-odstotno uspešnost pri jeziku domačih in tujih govorcev, pri napakah, specifičnih za francoščino kot tuji jezik, pa je bila zgolj 25–35-odstotna (Granger 2003b: 466). Zato označevanje napak poteka ročno. Za korpus SST so sicer skušali razviti avtomatski sistem. Izbrali so trinajst najpogostejših kategorij napak in jih razdelili na tri tipe glede na to, kako se njihova struktura loči od pravilnih izjav. Program potrebuje dvaintrideset različnih podatkov: ciljne besede, dve prej in dve potem, njihove besedne vrste, njihove slovarske oblike, pet kombinacij omenjenega ter prve in zadnje črke ciljne besede. Program so naučili na stopetdesetih dokumentih in ga preizkusili na šestnajstih, vendar so bili rezultati nezadovoljivi. Še najnatančneje je zaznal napake pri členih, zato so se osredotočili nanje. Dodali so pravilne izjave iz kontrolnega korpusa domačih govorcev, izjave izpraševalca in popravljene stavke iz korpusa. Rezultati so se izboljšali, vendar so bili še vedno zgolj 43-odstotno natančni. Nazadnje so natančnost zvišali z dodajanjem umetnih napak, vendar bi potrebovali še več označenega gradiva, da bi bila natančnost zadovoljiva (Izumi idr. 2004: 37–46). V okviru projekta FreeText,[36] ki temelji na korpusu FRIDA, pa so s podatki učečih se nadgrajevali obstoječe črkovalnike in oblikoslovne označevalnike za francoščino kot prvi jezik. Sistem, ki je nastal, dobro zaznava napake zapisa, precejšen delež slovničnih napak, nekaj skladenjskih in malo napak besedišča, ne zaznava pa napak ločil, semantičnih napak besedišča ali glagolskega časa (Granger 2003b: 474).

Zaradi neuspešnosti avtomatskih poskusov ni verjetno, da bi bilo označevanje napak kmalu avtomatizirano. Brez podatkov o pogostnosti določenih napak, posebej pri skupini govorcev iz specifičnega jezikovnega okolja, ni mogoče razviti programa za njihovo identifikacijo (Meunier 1998: 26), zato bi bilo treba za razvoj uspešnih programov najprej analizirati napake v velikih korpusih učečih se (Granger 2003b: 466). Pri vnašanju besedil v korpus pa so vendarle koristni črkovalniki, ki pomagajo odpraviti lapsuse pri tipkanju – seveda zgolj tiste, ki jih v izvirnem besedilu ni.

36 Prim. http://www.latl.unige.ch/freetext/fr/index.html (avgust 2011).

Ker je ročno označevanje zamudno, so napake zaenkrat označene na manjšem delu korpusov. Zaradi včasih problematičnega ujemanja to za posamezni korpus dela le malo označevalcev, za korpus CLC denimo le dva, ki sta nadzirala drug drugega (Nicholls 2003: 572). Od pregledanih korpusov usvajanja tujega jezika, o katerih sem dobila podatke, le v 20 označujejo napake, vendar tudi v teh niso označene na vseh besedah, temveč večkrat le na omejenem vzorcu, na primer tretjini ali petini besed (tabela 23). Pri več kot desetmilijonskih korpusih to popolnoma zadostuje. V celoti je označenih približno 14 % besed vseh korpusov.

Korpus	**Število vseh besed**	**Število besed z označenimi napakami**
AKCES/CzeSL	2.000.000	2.000.000
ASK	500.000	500.000
CATE	337.122	181.322
CEJL	1.000.000	del korpusa
CLC	35.000.000	30.000.000
CLEC	1.000.000	1.000.000
EAGLE	45.000	45.000
EIC	700.000	500.000
FALKO	1.200.000	del korpusa
FRIDA	450.000	450.000
HKUST	25.000.000	5.000.000
ICLE	3.700.000	3.700.000
Japanese Learners' Corpus	ni podatka	ni podatka
JEFLL	800.000	800.000
LCLE	10.000.000	ni podatka
MELD	98.000	50.400
LPELC	3.000.000	ni podatka
PELE	10.000	10.000
SST	1.000.000	1.000.000
VALICO	500.000	500.000

Tabela 23: Korpusi usvajanja tujega jezika s številom besed z označenimi napakami

Napake so označene z različnimi označevalnimi jeziki. Nekateri imajo svoje lastne (CEJL, CLEC, Japanese Learners' Corpus), poljska PELE in LPELC uporabljata britanski označevalnik in nabor oznak CLAWS (Constituent Likelihood Automatic Word-tagging System), večinoma pa se vsi usmerjajo k jezikoma SGML (HKUST) oziroma XML (ASK, CLEC, FRIDA, JEFLL, LCLE). V korpusu ICLE so napake označene v lastnem označevalnem jeziku s programom Error Editor (primer 15), v njegovi novejši vzporednici za francoščino FRIDA pa so se že odločili za XML.

(GNN) Cigarette $Cigarettes$ is dangerous for your health.
Primer 15: Oznaka napake v korpusu ICLE

Vnašanje oznak v dokumente je različno, npr. s pomočjo programov za dodajanje oznak XML. V korpusu Chyby, v katerem so češki univerzitetni študentje sami vnašali popravke, ki so jih v njihovih besedilih naredili učitelji (Bušta idr. 2009: 114), so najprej uporabljali preprost, posebej za to narejen program WinCorr (Pala idr. 2003: 94), razumljiv tudi za manj vešče uporabnike. Kasneje so ga nadomestili s podobno aplikacijo OOCorr, ki deluje v programskem okolju OpenOffice in podpira širši nabor formatov dokumentov (Bušta idr. 2009: 114).

Označevanje napak ponavadi ni omejeno zgolj na ločevanje med neustreznimi in ustreznimi oblikami. Tako je le v korpusu HKUST, ki ima možnosti napaka in nenapaka (Pravec 2002: 100), in v korpusu MAELC, kjer nepričakovane oblike označujejo z zvezdico (primer 16). Če je dovolj zanesljiva, poleg nje navedejo pričakovano obliko (Reder idr. 2003: 553).

Actual transcript:	16:07 Chinh.1	*Mes_ *Mesico holiday?
Target:		*Mex_ *Mexican holiday?

Primer 16: Oznaka nepričakovane oblike v korpusu MAELC (Reder idr. 2003: 553)

Nasploh pa so oznake povednejše in uporabnejše, če sledijo določeni klasifikaciji napak. Korpusi usvajanja za označevanje uporabljajo klasifikacije, ki naj bi bile hierarhične, obvladljive in razširljive za potrebe specifičnih raziskav, predvsem pa dobro dokumentirane s priročnikom, v katerem so primeri, dileme in rešitve za posamezne kategorije (Granger, Thewissen 2006). Klasifikacija omogoča opis in kvantitativno analizo napak, raziskovalcem pa nudi podatke tudi o nepričakovanih napakah, ki še niso uvrščene na sezname najpogostejših. Tako so prav na podlagi označenega korpusa prišli do za domače govorce netipične in do tedaj nedokumentirane rabe konektorjev (Dagneaux idr. 1998: 172). Avtomatsko so dostopni tudi podatki o manjkajočih členih, zaimkih, veznikih itn., ki jih je v neoznačenem korpusu težko odkriti (Nicholls 2003: 573).

Interpretacija rezultatov korpusa z označenimi napakami mora upoštevati protislovno dejstvo, da kategorizacija nikoli ni enoznačna, kajti napake lahko razdelimo glede na namen raziskave in jezikoslovno kategorijo, iz katere izhajamo. Vsakršno označevanje je pristransko in vpliva tako na vnos podatkov kot na dobljene rezultate (Čermák 2009), edino pravo

orodje za označevanje, ki ga imajo označevalci, pa je njihova lastna jezikovna zmožnost (Tenfjord 2006). Označevanje in klasificiranje napak sta tako težavna, da je za marsikoga celo vprašljiv njun smisel. Delo je pogosto intuitivno, možne so različne interpretacije, nobena analiza pa ne zajame vseh razlag (Ragan 2001). Za angleščino, ki je ciljni jezik v največjem številu korpusov, ni ustaljenega enotnega načina, temveč v vsakem korpusu razvrščajo napake po svoje. Tono (2003a: 802), ki je sodeloval pri več japonskih korpusih usvajanja angleščine, je zaradi heterogenosti taksonomij predlagal, da bi se vsi držali splošnega nabora, v katerem bi določili jezikoslovno kategorijo in način odklona od ciljne oblike, nato pa bi ga prilagajali potrebam lastne raziskave. Da brez prilagajanja ne gre, je ugotovil, ko je pri označevanju shemo nehote adaptiral svojim ciljem.

Ne glede na načelno vprašljivost jezikoslovna realnost kaže, da je označevanje napak smiselno, saj bistveno olajša analizo, katere rezultati so uporabni v jezikoslovnoteoretične, jezikovnotehnološke ali pedagoške namene. Zato napake niso označene samo v korpusih usvajanja tujega jezika, temveč tudi na jeziku domačih govorcev (prim. Pala idr. 2003; Bušta idr. 2009; Rozman, Stritar, Kosem 2010), s čimer med drugim izboljšujejo delovanje črkovalnikov in programov za pregledovanje pravilnosti zapisanih besedil.

V nadaljevanju so predstavljene klasifikacije napak po površinski strukturi in po jezikovni ravnini. Za korpuse so manj relevantni nekateri drugi tipi, denimo taksonomija po komunikacijskem učinku ali komparativna taksonomija, ki primerja strukture iz jezika učečih se z drugimi strukturami, npr. iz jezika otrok (Dulay, Burt, Krashen 1982: 163, 189).

5.2.1 Klasifikacija napak po površinski strukturi

V nekaterih korpusih so napake označene formalno glede na površinsko strukturo, modifikacijo ciljne oblike (angl. *target modification*) oziroma glede na način, kako dejanska, neustrezna oblika v besedilu ne ustreza pričakovani. Seveda gre za modifikacijo ciljne oblike v prenesenem pomenu, saj je nemogoče trditi, da učeči se spremeni ciljno obliko, da bi tvoril napako (James 1998: 106). Tak način označevanja naj bi nudil koristne informacije za prepoznavo kognitivnih procesov, povezanih z usvajanjem jezika, v njem pa so bili določeni naslednji tipi napak (Dulay, Burt, Krashen 1982: 150–163; James 1998: 106–113):

– **zgrešitev** (angl. *commission*) oz. **napačni izbor** (angl. *misselection, misformation*): del besedila je neustrezen, oblika je npr. narobe tvorjena;
– **izpust** (angl. *ommission*): del besedila manjka, ponavadi so izpuščene slovnične besede;
– **vstavitev, dodajanje** (angl. *addition*): del besedila je odveč. Sem umeščajo tudi dvojno označevanje semantičnih značilnosti, regularizacijo (posploševanje pravila na izjemo) in iregularizacijo (prenos izjemne na sicer regularno obliko);
– **neustrezno zaporedje** (angl. *misordering*): neustrezen besedni red, retorično zaporedje, pa tudi oblikovanost besedil, npr. uradnih pisem;
– **sestava neobstoječe oblike iz dveh pravilnih** (angl. *blends, contamination, cross-association, hybridization*): tvorec aktivira semantično povezani strukturi, ki ustrezata namenu, in ju združi v novo, nepravilno obliko.

V korpusih se uporabljajo kategorije zgrešitev, izpust in vstavitev. V korpusu ISLE se napake, označene na fonetični ravni, ločijo na zamenjave, vstavitve in izpuste. Ker pa tako ločevanje pravzaprav ne daje dosti informacij o tem, kako učeči se usvajajo jezik (Ellis 1994: 56), ga tako raziskovalci kot oblikovalci korpusov ponavadi nadgrajujejo z jezikoslovnimi oznakami.

5.2.2 Klasifikacija napak po jezikovni ravnini

V praksi je najuporabnejše razvrščanje napak po jezikoslovnih kategorijah, ki se lahko kombinira s formalnimi oznakami po površinski strukturi (prim. Granger 2003b: 467). Ponavadi gre za klasifikacijo na več ravneh (James 1998: 105):

– po jezikovni ravnini, npr. fonologija, zapisovanje, slovnica, besedišče, besedilo, diskurz;[37]
– po posamezni sestavini oziroma enoti jezika, npr. morfem, beseda, besedna zveza, stavek, poved;
– po besedni vrsti, npr. samostalnik, glagol, pridevnik, prislov;
– po oblikoslovnih značilnostih, npr. čas, število, sklon, spol, številnost, vid, prehodnost, vezava.

37 Natančna razmejitev besedila in diskurza je za to razpravo preobsežna, toda na splošno gre pri diskurzu za besedilo v kontekstu oziroma verigo povezanih izjav in sporočil (prim. Fairclough 2001: 16–23, Titscher idr. 2005: 25–27).

Tovrstno razvrščanje se uporablja v mnogih pedagoških gradivih in je bilo pogosto tudi v predkorpusnih raziskavah jezika učečih se. Raziskovalci so denimo razdelili napake na medjezikovne in razvojne, znotraj tega pa po jezikoslovnih kategorijah (Dulay, Burt, Krashen 1982: 147). Take klasifikacije omogočajo natančen opis specifičnih napak in kvantifikacijo skupin (Ellis 1994: 54). Oznake so formalne, relativno objektivne, nedvoumne, neodvisne od označevalčevega spekuliranja in pomenijo skrajno operacionalizacijo (Pirih Svetina 2005: 127). Ker so bolj ali manj opisne, je označevanje konsistentno in ne vključuje pretirane subjektivnosti (Dagneaux idr. 1998: 166), čeprav se ji vedno ni mogoče izogniti. Manj ostre in razdelane so meje za napake v fonologiji, besedilu in diskurzu (James 1998: 105), pri slovenščini pa se pogosto prepletajo ravni zapisa, besedja in oblikoslovja (prim. poglavje 5.3.5). Tudi besednovrstno klasificiranje je lahko problematično, »kar kažejo primeri kot npr. *Kako te pišeš?*, kjer bi po eni strani lahko predvidevali kombinacijo prenosa in napačne odločitve za zaimek, po drugi pa glagolsko napako, saj tvorec ne pozna kolokacijske paradigme glagola pisati se« (Rozman 2006: 24).

Korpusi se osredotočajo na napake pri tvorjenju, toda ne na vseh ravneh. Diskurzne in besedilne napake, za prepoznavanje in označevanje katerih bi bilo treba več konteksta, pogosto ostajajo neoznačene (prim. Izumi idr. 2004: 35; Požgaj Hadži, Ferbežar 2001: 63). V nadaljevanju so na kratko opisane klasifikacije napak v nekaterih korpusih usvajanja tujega jezika in posameznih slovenskih raziskavah, saj so bile izhodišče in zgled za načela označevanja napak v korpusu usvajanja slovenščine kot tujega jezika.

5.2.2.1 Klasifikacije napak v tujih korpusih

V korpusu **ICLE** vsaki napaki ročno določijo krovno kategorijo in vrsto podkategorij; skupaj imajo 56 oznak (tabela 24). V drugi verziji so dodali razlikovanje med napako in ponesrečenim izborom (angl. *infelicity*). Medtem ko napaka krši določeno slovnično, leksikalno ali drugo jezikoslovno pravilo, se ponesrečeni izbor nanaša na nenavadno zveneč jezik (Granger, Thewissen 2006: 6) in na težave z registrom, kot so okrajšave, neformalni izrazi, sleng, klicaji, medmeti, težave s politično korektnostjo (npr. raba ženskih in moških oblik) in slogom. Ločevanje med napako in ponesrečenim izborom ni pojasnjeno.

Error	Example
Form, morphology	Watchs > watches, unpossible > impossible
Form, spelling	Their > there, independance > independence
Form, Regional, Spelling	Color > colour
Grammar, articles	The life is beautiful
Grammar, determiner, demonstrative	This elements cannot be separated > These elements
Grammar, determiner, possessive	People accept jobs according to how much they get paid but not according to his preferences > their preferences
Grammar, determiner, indefinite	He does not have some expectations > any expectations
Grammar, determiner, other	Whichever man was it you saw yesterday? > Which man
Grammar, adjectives, comparative/superlative	The more rich our culture is
Grammar, adjectives, number	The last sentences have been favourables to women. > favourable
Grammar, adjectives, order	A car red
Grammar, adverbs, order	(…) and spend there twelve months > and spend twelve months there
Grammar, nouns, case	The Berlin's wall
Grammar, nouns, number	People paid of their own life > lives
Grammar, pronoun, demonstrative	That is a principle of Grice about communication > This is a principle
Grammar, pronoun, personal	Europe is a small area, she only represents…> it only represents
Grammar, pronoun, possessive	Theirs is outdated > Ours is outdated
Grammar, pronoun, indefinite	A lower-class man is not on an equal footing with his middle- or upper-class one > upper-class counterpart
Grammar, pronoun, reflexive/reciprocal	The tobacco companies are denied permission to directly publicise themself on television > themselves on television
Grammar, pronoun, relative/interrogative	As clearly emerges from all what precedes > all that precedes
Grammar, pronoun, unclear reference	In Belgium, there are towns where they can live and study > where the students can live
Grammar, verbs, auxiliaries	We can think the reason for creating money is good > We may think
Grammar, verbs, morphology	I have swam in the river
Grammar, verbs, number	Progress have been made
Grammar, verbs, non-finite/finite	So how can people loving each other behave > So how can people who love each other behave
Grammar, verbs, tense	More and more companies become multi-national > are becoming
Grammar, verbs, voice	… are worth being underlined > are worth underlining
Grammar, word class	Because of a religion mistake > Because of a religious mistake
Lexis, conjunctions, coordinating	Neither … or

Error	Example
Lexis, connectors, logical, complex	On the other side > on the other hand
Lexis, connectors, logical, single	Indeed > ∅
Lexis, conjunctions, subordinating	Even if he … and that he decides … > and decides …
Lexical phrase	Shopping for the cooking > food shopping
Lexical phrase, false friends	Some learn a job which will never serve or anything > be of use
Lexical single	LLN is a common city > an ordinary city
Lexical single, false friends	He assisted the lecture > attended
Punctuation, confusion	He took the books, the record, the computers, he left at eight o'clock > He took the books, the record, the computers. He left at eight o'clock.
Punctuation, lexical	He went to the seaside. He spent the whole day there > He went to the seaside and spent the whole day there.
Punctuation, missing	John Lennon who was born in Liverpool, … > John Lennon, who was born in Liverpool, …
Punctuation, redundant	Women were seen conducting affairs, and bringing negotiations to satisfactory conclusions in what men always claimed to be: »a man's world« > what men always claimed to be »a man's world«
Style, incomplete	Another example.
Style, unclear	And then after all the shaking up and down between the ground and communistic ideals my head stayed empty and something else, I do not know what.
Word missing	We have a meeting Monday.
Word order	Some thirty years ago was launched the idea …
Word redundant, single	They are doing eveything possible they can
Word redundant, multiple	In short, prisons, instead of helping people to manage to get out of their plight, plunge them into despair > helping people to get out of their plight, plunge them into despair
Lexico-grammar, adjectives, complementation	Capable to go > capable of going
Lexico-grammar, adjectives, dependent preposition	Accessible for wheelchairs > accessible to wheelchairs
Lexico-grammar, conjunctions, complementations	While read a book …
Lexico-grammar, nouns, complementation	The possibility to leave
Lexico-grammar, nouns, dependent preposition	Thirst of knowledge > thirst for knowledge
Lexico-grammar, nouns, uncountable/countable	an information; evidences
Lexico-grammar, prepositions, complementation	Instead to go > instead of going
Lexico-grammar, verbs, complementation	We cannot help imagine and invent new technological systems > We cannot help imagining and inventing new technological systems

Error	Example
Lexico-grammar, verbs, dependent preposition	Move in another country > move to another country
Infelicities	Kids > children

Tabela 24: Tipi napak v korpusu ICLE (Granger, Thewissen 2006: 7–9)[38]

No	Line	Type of error	Text	Correction
1	8	Auxilliary	lot of love and care. They **do** not only expect to get it from their men **0**	**0**
2	348	Tense	correct and at least partially **answered** the problem **in question**	**answer**
3	380	Auxilliary	provides us with rules that **not necessarily have to be**	**are not necessarily**
4	798	Verb number error	the data on alcohol drinking **is** to your benefit . You	**are**
5	693	Tense	However 0 as tobacco **was getting** cheaper and	**became**
6	570	Passive-active voice	movies, they form gangs and **involve** in crimes in order to create and maintain their	**become involved**
7	348	Auxilliary	the real answer **could** be obtained only on the basis of the information	**can**
8	699	Tense	ising 0 cigarettes **has not changed since** the 50's until late 60's.	**did not change**
9	818	Tense	If it were all it **does** 0 television would by all means be	**did**
10	465	Auxilliary	rodents. How **it happens** 0 then, that it can	**does it happen**

Primer 17: Izpis napak v poljskem podkorpusu ICLE (spletna stran ICLE)

Področje napake	Kategorija napake	Primer
Oblika	Pisanje skupaj	le **portebagages** (porte-bagages) est sur le toit
	Velike/male začetnice	Mme Thatcher incarnait la peur des **anglais** (Anglais)
	Diakritična znamenja	Il existe une **ambiguité** (ambiguïté)
	Homonimija	Ils **ce** (se) déshabillent
	Druge napake zapisa	un **labirint** (labyrinthe)
Oblikoslovje	Izpeljava – predpona	le sentiment de **malcontentement** (mécontentement)
	Izpeljava – pripona	… qui continue à **évolutionner** (évoluer)
	Končnica	Elle s'occupe des enfants et des **travails** (travaux)
	Končnica – zamenjava	Il n'a pas très bien **réussit** (réussi)
	Sestavljanje	la participation de **celles-dernières** (celles-ci)

38 Ker je nabor vezan na angleško jezikoslovje, je naveden v angleškem izvirniku.

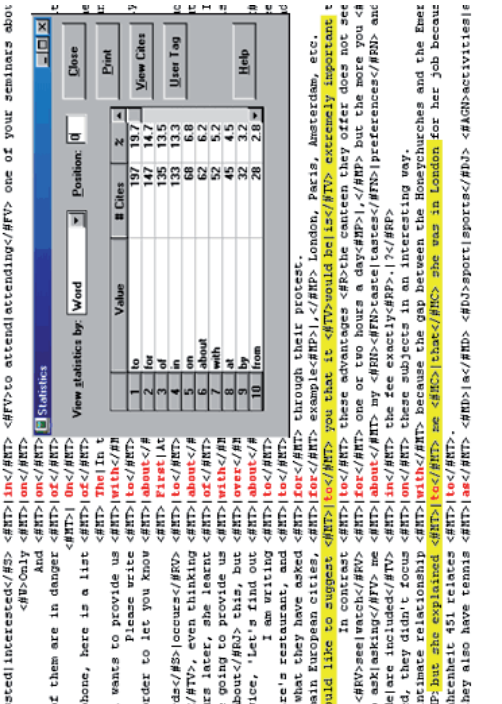
Področje napake	Kategorija napake	Primer
Slovnica	Besedna vrsta	L'unique chose **que** (qui) n'est pas bonne en Belgique...
	Pomožni glagol	Je m'**avais** (étais) très bien amusé
	Spol	La protection sociale a été **amélioré** (améliorée)
	Način	Bien qu'ils **sont** (soient) pressés, ...
	Število	Elle reprit ses **esprit** (esprits)
	Čas	J'espère **s'** (m') adapter rapidement
	Evfonija	En prenant **ce** (cet) aspect, ...
Besedišče	Pomen	Il est un peu plus *__commis à__ [engagé envers] l'idéal européen.
	Vezljivost pridevnika	... bien qu'il soit libre * **à** [de] choisir ce qu'il fait.
	Vezljivost prislova	
	Vezljivost glagola	le Canada pourrait se transformer *__dans__ [en] un type de superstructure.
	Vezava samostalnika	
	Kolokacija	leurs élèves suivent des *__cours religieux__ [cours de religion].
Skladnja	Besedni red	Je peux **m'amuser bien** (bien m'amuser).
	Manjkajoča beseda	Je crois **0** (qu') ici il y a beaucoup plus de soirées
	Odvečna beseda	le domaine social **et** (,) économique et politique
	Kohezija	A la métropole, il existe plus d'allocations et d'aide pour les chômeurs, les handicapés et les personnes âgées, **et** (tandis que) dans les îles, il n'y a pas beaucoup de soutien.
Register	Besedišče	**Quand même** (néanmoins), c'est une histoire lointaine
	Skladnja	Etant donné que **j'ai pas** (je n'ai pas) souvent présenté mes travaux ...
Stil	Nejasen	Quand la famille d'un travailleur étranger vient le rejoindre, **on** (??) est obligé d'organiser toutes les formalités avant de quitter son pays.
	Obremenjen	mais il y a des autres choses qui **m'ont donné une très grande surprise** (qui m'ont fort surpris)
Ločila	Napačno ločilo	La langue devient plus française - (:) on l'appelle maintenant le créole francisé.
	Odvečno ločilo	Ce sont surtout les différences biologiques**,** (0) qui sont présentées dans la société avec la grande force de l'organisme humaine.
	Manjkajoče ločilo	Quand j'ai du temps libre **0** (,) je veux faire des choses reposantes ...
Tipkarske napake		**qunad** (quand); **ps** (pas)

Tabela 25: Tipi napak v korpusu FRIDA
(Granger 2003b: 468–69, 477–78)

V francoskem korpusu **FRIDA** označujejo napake po zgledu korpusa ICLE, vendar v jeziku XML. Vsaka napaka ima tri oznake (tabela 25): jezikoslovno ravnino (slovnica, slovar, zapis itn.), kategorijo (vrsta, število itn.) in besedno vrsto (samostalnik, pridevnik itn.) (Granger 2001: 3).

Oblikoslovne oznake so določene le napakam, ne pa vsem besedam, ker niso mogli predvideti, do katere mere bodo napake v besedilih vplivale na natančnost avtomatskih označevalnikov. S tem so se želeli izogniti tudi delu s tekstovnimi datotekami, prepolnimi oznak (Granger 2003b: 469). Napake imajo pripisano popravljeno obliko.

Napake v korpusu **CLC** so označene po sistemu <#KODA> napačna beseda|popravljena beseda</#KODA> (slika 24). Večina kod ima dve črki, pri katerih prva predstavlja splošni tip napake glede na površinsko strukturo (neustrezno uporabljena oblika, izpust, vstavitev, beseda, ki jo je treba zamenjati, in neustrezno izpeljana beseda), druga pa besedno vrsto (Nicholls 2003: 573). Samostojne kategorije so še ločila, številnost, lažni prijatelji, ujemanje in različne posebnosti, kot so kolokacije, idiomi, register, črkovanje, ameriško črkovanje, glagolski čas, besedni red, kompleksne napake in podobno.



Slika 24: Označevanje neustreznih predlogov v korpusu CLC

V norveškem korpusu **ASK** so napake klasificirane na leksemske, morfološke, sintaktične, interpunkcijske in nerazvrščene. Temu sledijo podrobnejše kategorije (tabela 26); stilističnih napak ne označujejo. Vsaka napaka ima pripisano pravilno obliko.

Tip napake	Oznaka	Pomen	Oznaka podtipa	Pomen podtipa
Leksikalne napake	W	Napačna beseda		
	ORT	Ortografska napaka		
	PART	Besedi, napačno pisani skupaj		
	SPL	Besedi, napačno pisani narazen		
	CAP	Napačna raba velike oz. male začetnice	AGR	Napaka je nastala zaradi ujemanja s prej povzročeno napako
	FL	Nenorveška beseda		
Oblikoslovne napake	F	Napačna oblika, npr. napačen čas	AGR	Napaka je nastala zaradi ujemanja s prej povzročeno napako
	INFL	Narobe tvorjena oblika za pravilno morfosintaktično kategorijo		
Skladenjske napake	M	Manjka beseda ali fraza		
	R	Beseda ali fraza je odveč		
	O	Napačen besedni red	INV	Napaka zaradi inverzije
			OINV	Napačno posploševanje inverzijskega pravila
			MCA	Napačen položaj prislovne fraze v glavnem stavku
			SCA	Napačen položaj prislovne fraze v odvisniku
Interpunkcijske napake	PUNC	Napačno ločilo	AGR	Napaka je nastala zaradi ujemanja s prej povzročeno napako
	PUNCM	Manjka ločilo		
	PUNCR	Ločilo je odveč		
Nerazvrščene napake	X	Ni jasno, kaj je tvorec želel izraziti		

Tabela 26: Tipi napak v korpusu ASK (spletna stran ASK)

Napake je s pomočjo priročnika in izobraževalnega seminarja označevalo sedem do osem ljudi, snovalci korpusa pa bi bili zadovoljni s 70–80-odstotnim ujemanjem med njimi (Tenfjord 2006).

V nemškem korpusu **FALKO** so napake ločene na osem tipov (tabela 27).

Raven	**Primer napake iz korpusa**	**Ciljna oblika**
Pravopis	Gebeuden	Gebäuden
Besedotvorje	Analysenverfahren	Analyseverfahren
Ujemanje	dass die Bedeutung der "kleinen" Worte eine wichtige Bedeutung für den gesamten Text haben können.	dass die Bedeutungen der "kleinen" Worte eine wichtige Bedeutung für den gesamten Text haben können.
Vezava	Bei Gespräche	Bei Gesprächen
Glagolski čas	Die Realisten wollen die Wirklichkeit nicht kopieren, was mit der Erfindung der Photographie der Fall war.	Die Realisten wollten die Wirklichkeit nicht kopieren, was mit der Erfindung der Photographie der Fall war.
Glagolski način	Die Hermeneutik würde in der Neuzeit erfunden und ist keine Kompensation ...	Die Hermeneutik wurde in der Neuzeit erfunden und ist keine Kompensation ...
Besedni red	dass nur er (Josef K.) konnte durch dieses Tor eingelassen werden	dass nur er (Josef K.) durch dieses Tor eingelassen werden konnte
Izrazi	Die von F. Schlegel und F. Schleiermacher genannte Hermeneutik ist...	Die von F. Schlegel und F. Schleiermacher begründete Hermeneutik ist...

Tabela 27: Tipi napak v korpusu FALKO (Lüdeling idr. 2005)

Subjektivnim interpretacijam se izogibajo tako, da vsaki napaki določijo vse možne ciljne oblike in označijo vsako posebej (prav tam; primer 18).

izjava	*Die*	*Erklärung*	*für*	*diese*	*Phänomen*
CILJ1				dieses	
Oznaka				spol	
CILJ2					Phänomene
Oznaka					število

Primer 18: Večplastno označevanje napak v korpusu FALKO (Lüdeling idr. 2005)

V korpusu **EAGLE** so ločeno obravnavane napake črkovanja in slovnične napake. Prve so lahko oblikoslovne (*gestudiert* nam. *studiert*),[39] tipografske (*iher* nam. *hier*) ali povezane z rabo velike/male začetnice (*wetter* nam. *Wetter*). Zbrane so v

39 Zakaj so oblikoslovne napake uvrščene med napake črkovanja, v dosegljivi literaturi ni pojasnjeno.

posebnem korpusu napak črkovanja, ki ima 1697 pojavnic oziroma 1234 napak in še ni bil podrobneje analiziran (Boyd 2008: 1898). Klasifikacija slovničnih napak, ki je zaradi majhnosti korpusa precej omejena, pa vključuje pet glavnih tipov (tabela 28), od katerih ima vsak podtipe glede na določene slovnične značilnosti (prim. Boyd 2008: 1900).

Tip napake	Primer iz korpusa	Ciljna oblika
Oblika (napake znotraj besede, ki niso povezane s črkovanjem)	Ja, Ich zeige ihn ihnen.	Ja, ich zeige ihn ihnen.
Izbira	Hast du der Reiseprospekt?	Hast du den Reiseprospekt?
Ujemanje	Du arbeiten in Liechtenstein.	Du arbeitest in Liechtenstein.
Besedni red	Welcher Job diese Dinge verlangen würde?	Welcher Job würde diese Dinge verlangen?
Ločila	Gehört dir diese Jacke	Gehört dir diese Jacke?

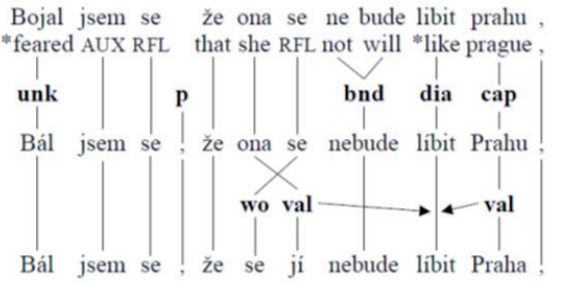
Tabela 28: Primeri napak v korpusu EAGLE (Boyd 2008: 1899)

Vsak stavek je označen neodvisno in ne glede na kontekst. Označevanje poteka v formatu korpusa FALKO, ki omogoča označevanje večbesednih, prekrivajočih se ali dvoumnih napak. Vsaka oznaka vključuje podatek o lokaciji napake, njenem opisu oz. tipu in ciljni obliki (Boyd 2008: 1898). Napake je v jeziku XML v delu korpusa označil en označevalec (navedeno delo: 1901).

Označevanje napak v korpusu **AKCES/CzeSL** naj bi ustrezalo izrazito fleksijskemu jeziku, kakršna je češčina (Hana idr. 2010: 13). Klasifikacija, ki je zaenkrat še groba in naj bi bila razširjena v nadaljnjih fazah, je nastala kot kompromis med omejitvami, ki jih prinaša proces označevanja (npr. nezmožnost označevalcev, da prepoznajo napake, ki nastanejo kot posledica interference), in raziskovalnimi zahtevami (navedeno delo: 15). Na prvi stopnji gre za transkribirani vnos oziroma originalni niz grafemov, tipi napak pa so:

– meje besed;
– ločila;
– velika/mala začetnica;
– diakritični znaki;
– črke;
– končnica;
– neznani leksem.

Na drugi stopnji gre za ortografsko in oblikoslovno popravljanje, tipi napak so:
– ujemanje;
– vezljivost;
– nanašanje zaimkov;
– kompleksne glagolske oblike;
– zanikanje;
– manjkajoči del besedila;
– odvečni del besedila;
– modalnost;
– besedni red;
– besedišče in frazeologija.



Primer 19: Stavek z označenimi napakami v korpusu AKCES: v prvi vrstici je originalno besedilo, pod njim sta angleški prevod in vrstica s tipi napak (*unk* = neznana oblika, *p* = manjkajoče ločilo, *bnd* = meje besed, *dia* = diakritični znaki, *cap* = velike/male črke, *wo* = besedni red)[40] (Hana idr. 2010: 17)

Celotni proces se začne s transkripcijo dokumenta v obliko html s pomočjo enega od urejevalnikov besedil. Na osnovi tega so generirane oznake na obeh stopnjah v formatu PML (Prague Markup Language),[41] ki temelji na jeziku XML in ga na Češkem uporabljajo za shranjevanje jezikoslovno označenih podatkov. Označevalec nato ročno popravi dokument, avtomatsko pa so dodane tudi morebitne dodatne informacije o napakah, npr. tip napake, lema, oblikoslovne oznake (Hana idr. 2010: 17–18).

Nabor napak v japonskem korpusu **SST** skuša biti čim splošnejši. Vsaka oznaka ima tri vrste podatkov (primer 20): besedno vrsto, slovnično pravilo in pričakovano pravilno obliko (Izumi idr. 2004: 35).

40 Razlaga kratice *val* v izvirnem besedilu manjka.
41 Prim. http://ufal.mff.cuni.cz/jazz/pml/index_en.html (september 2011).

*I belong to two baseball **<n_num crr=«teams«>**team**</n_num>**.
Primer 20: Stavek z označeno napako v korpusu SST: *n* = samostalnik, *num* = napaka v številu, *crr* = pravilna oblika

Napake, ki jih ni mogoče umestiti, imajo posebne oznake, tako da je v naboru skupaj petinštirideset oznak (tabela 29).

Besedna vrsta	**Kategorija napake**
Samostalnik	Končnica
	Število
	Sklon
	Številnost
	Prilastek
	Besedišče
Glagol	Končnica
	Neujemanje med osebkom in povedkom
	Oblika
	Čas
	Vid
	Glagolski glas
	Uporaba (ne)dovršnega glagola
	Zanikanje glagola
	Vprašanje
	Prilastek
	Besedišče
Modalni glagoli	Besedišče
Pridevnik	Končnica
	Uporaba stopnjevanja
	Število
	Neujemanje števila
	Številnik
	Prilastek
	Besedišče
Prislov	Končnica
	Uporaba stopnjevanja
	Besedišče
Predlog	Prilastek
	Neodvisni predlog
	Odvisni predlog
Člen	Člen
Zaimek	Končnica
	Neujemanje v spolu/številu
	Sklon
	Besedišče
Veznik	Besedišče
Oziralni zaimek	Sklon
	Besedišče

Besedna vrsta	Kategorija napake
Vprašalnica	Besedišče
Ostalo	Japonska angleščina
	Kolokacija
	Besedni red
	Neznan tip napake
	Nerazumljiva izjava

Tabela 29: Tipi napak v korpusu SST
(Izumi idr. 2004: 36–37, prevedla M. S.)

Pri poskusu avtomatskega označevanja (prim. poglavje 5.2) so napake razdelili glede na spremembo pričakovane oblike na izpuste, zamenjave in vstavitve (Izumi idr. 2004: 38).

Sodelavci kitajskega korpusa **CLEC** poudarjajo, da mora biti sistem za označevanje napak odprt. Napake so razdeljene na enajst osnovnih kategorij, ki kombinirajo jezikovno ravnino in besedno vrsto: oblika besede, glagolska besedna zveza, samostalniška besedna zveza, zaimek, pridevnik, prislov, predlog, veznik, raba besede, kolokacija in skladnja. Znotraj tega pa ločijo nadaljnjih 61 kategorij (Gui, Yang 2006).

V korpusu **Japanese Learners' Corpus** oznake vključujejo področje napake, parafrazo domačega govorca, napako, popravek in slovnično analizo napake (Sugiura 2000).

Oznaka	Pomen	Primer
< errors >	Področje napake	I <wash <in two hours>[*1]>[*2].
%ntv	Parafraza govorca angleščine kot prvega jezika	I wash it for two hours.
%err	Označena napaka in popravek	[1 in two hours = for two hours]; [2 wash for two hours = wash it for two hours];
%als	Slovnična analiza napake	[1 preposition in adverbial phrase]; [2 object of verb is missing, (pronoun)];

Tabela 30: Oznake primera napake v korpusu Japanese Learners' Corpus

*ELC: I <wash <in two hours>[*1]>[*2].
%ntv: I wash it for two hours.
%err: [1 in two hours = for two hours];
[2 wash for two hours = wash it for two hours];
%als: [1 preposition in adverbial phrase];
[2 object of verb is missing, (pronoun)];

Primer 21: Napaka z oznakami v korpusu Japanese Learners' Corpus

Češki korpus **Chyby** je za razliko od ostalih opisanih korpus domačih govorcev češčine, smiselno pa ga je omeniti, ker gre za korpus slovanskega jezika. Nastaja znotraj večjega projekta, katerega namen je prepoznati tipe napak v sponta-

nih besedilih in analizirati možnosti avtomatskega popravljanja. Ker so besedila v referenčnih korpusih v glavnem lektorirana, so razvili posebej korpus, sestavljen iz tekstov z napakami (Bušta idr. 2009: 118–119). Ima okoli 500.000 besed, v njem pa so besedila čeških univerzitetnih študentov, in sicer eseji, ocene ter uvodi v magistrske naloge (navedeno delo: 110).

Čeprav gre za domače govorce, je klasifikacija napak podobna kot v korpusih usvajanja, saj domači in tuji govorci delajo podobne napake (prim. Bigert idr. 2004). Prvotno klasifikacijo prikazuje tabela 31, kasneje pa so jo zaradi prekrivanja nekaterih kategorij ipd. popravili (tabela 32, Bušta idr.: 2009: 111). Napake so označene v jeziku XML z elementom <corr>, ki omogoča tudi gnezdenje napak (navedeno delo: 117; prim. poglavje 5.3.4).

Tip napake	**Opis**
Črkovanje	Napačen zapis besede; napako prepoznajo črkovalniki
Tipografija	Napačna raba pomišljajev, narekovajev, presledkov ipd.
Oblikoslovje in skladnja	Napačna končnica
Skladnja	Napačna vezava
Ločila	Napačna raba vejice in drugih ločil
Besedišče	Napačna raba izrazov, ki povzroči kršitev semantičnih razmerij
Stil	Neustrezna raba slenga, žargona, arhaičnih ali pogovornih besed, ponavljanje izrazov, odvečna raba kazalnih zaimkov, napačna raba trpnika, dolgi samostalniški nizi, dvoumna raba anaforičnih zaimkov

Tabela 31: Prvotna klasifikacija napak v korpusu Chyby (Pala idr. 2003: 92–93)

Tip napake	**Podtip**
Črkovanje (prepoznajo jih črkovalniki)	
Pravopis	Ločila
	i/*y* v določenih besedah
	s/*z* v določenih besedah
	Ustrezna oblika zaimka
	Raba velikih in malih črk
	Sestavljene besede
	Tuje besede
	Druge pravopisne napake
Tipografija	Vezaj in pomišljaj
	Deljenje
	Odstavki
	Presledki
	Predlogi na koncu vrstic
	Druge tipografske napake

Tip napake	Podtip
Oblikoskladnja	Oblikoslovno neustrezne oblike
	Ujemanje
	Vezljivost glagola in samostalnika
	Napake, povezane s svojino
	Druge oblikoskladenjske napake
Leksiko-semantične napake	Izraz brez pomena
	Nesmiselna ali neresnična izjava
	Druge semantične napake
Slog	Ponavljanje besed
	Redundantna raba kazalnih zaimkov
	Slengovski izrazi
	Kopičenje enakih samostalniških končnic
	Slogovno neustrezen besedni red
	Druge slogovne napake

Tabela 32: Popravljena klasifikacija napak v korpusu Chyby (Bušta idr. 2009: 115–117)

Chyby je precej bolj normativen od korpusov usvajanja tujega jezika, saj izhaja iz formulacije, da včasih besedila niso dobro napisana (Pala idr. 2003: 92). Tvorci so študentje, ki se učijo pisati, zato je največ napak stilističnih (23,25 %), sledijo jim napake ločil (21,32 %), povezane s strogimi pravili češkega pravopisa, in leksikalne napake (14,09 %) (Bušta idr.: 2009: 118).

5.2.2.2 Slovenske klasifikacije

Dosedanje slovenske klasifikacije napak so le delno korpusno utemeljene. Najbolj korpusno usmerjena je raziskava T. Balažic Bulc o najpogostejših napakah srbohrvaških govorcev slovenščine. V korpusu besedil z izpita je označila napake in tiste, ki so se pojavile več kot 52-krat oziroma petodstotno, zbrala v glavne tipe približno glede na jezikovno ravnino. Temu je pripisala več podtipov (tabela 33).

Tip napake	Podtip napake	Primer
Glasovne premene 10 % napak	Premene po zvenečnosti	kdo; nisko; opčutek; teško
	Sprememba <l> v <o> v preteklem deležniku	spremenio; učio; vidio; živeo
	Izpadanje soglasnikov	lasnik; oddaja; otujenost; rastavišče
	Vrivanje črke <j> tudi v zapisu med samoglasnika v zevi ali sklopu dveh samoglasnikov	ambijent; dijalog; pacijenti; varijante

Tip napake	Podtip napake	Primer
Sklanjatveni vzorci samostalnikov in pridevnikov 8 % napak	Moški spol	(veliko) prijatelja; s sosedima; rojstnog (kraja)
	Ženski spol	(imam) prijateljicu; (veliko) institucija; po končanoj (srednji šoli)
	Srednji spol	(posledica njihovih) dejanja; u velikim mestima
Raba konektorjev 7 % napak	Kateri namesto ki	pri prijateljici katera je nora za horoskopom
	Drugi relativni konektorji	tisto kaj jaz naredim
	Izražanje protivnosti	tri sobe bi bile za družino, a četrta za gostje
	Izražanje časa	kdaj sem prišla v novo šolo
	Izražanje namere	točno odredijo njihovo turistično zmogljivost, kako bi zmanjšali to negativno stran turizma
	Izražanje vezalnosti	zraven dobre hrane, gre in dobra pijača
Zapisovanje vejice 41 % napak		Že takrat sem vedel da bo to moj najboljši prijatelj.
Stalni besedni red 6 % napak	Mesto naslonk v stavku	v velikih mestih ceste so umazane
	Naslonski niz	zdele so mi se zelo zabavne
	Izpust pomožnega glagola	on se z našim predlogom takoj strinjal
Dvojinski vzorci 5 % napak		kot da sva stari znanci; smo se poročili

Tabela 33: Napake srbohrvaških govorcev (Balažic Bulc 2004: 81–87)

V. Požgaj Hadži in I. Ferbežar sta analizirali napake, ki jih pri ustnem delu izpita iz znanja slovenščine na osnovni ravni delajo hrvaško, srbsko in bosansko govoreči kandidati. Razdelili sta jih na fonetične, morfosintaktične in leksikalne (tabela 34) ter izpostavili neustrezne oblike, ki so posledica transfera. Jezikovne ravnine se prepletajo, posebej na začetnih ravneh znanja, težko pa je ločiti med razvojnimi napakami in interferencami (Požgaj Hadži, Ferbežar 2001: 61, 63).

Jezikovna ravnina	Napake	Primer
Fonetične napake	Izgovor srednjih namesto ozkih in širokih *e* in *o*	lepe; meso
	Nadomeščanje polglasnika z *a*	sam; osam; centar
	Trdi izgovor fonema /l/ ali izgovor /l/ v sredini in na koncu besed	žal; polnit
	Zliti izgovor fonemov /lj/ in /nj/	ljudi; stanovanje
	Mehki izgovor fonema /ć/	domaće; većinom
	Mesto naglasa kot v prvem jeziku	bíli; dánašnjem
	Tvorba preteklega deležnika na *-o* in redukcije	pogledo; reko

Jezikovna ravnina	Napake	Primer
Morfosintaktične napake	Uporaba sklonov iz prvega jezika	ímam mnogo prijatelja Slovenca
	Raba množine namesto dvojine	potem pa bomo že po telefonu dogovorile
	Raba rodilnika s številkoma tri in štiri	tri tedna; štiri meseca
	Sedanjiške končnice za tretjo osebo množine iz prvega jezika	ljudi se sprehajaju
	Napačen besedni red	lepe su bajte
Leksikalne napake	Besede in besedne zveze iz prvega jezika	kásnije; se guraju
	Dobesedni prevodi iz prvega jezika	pa nekaj fabrka je imelo tam

Tabela 34: Napake hrvaško, srbsko in bosansko govorečih kandidatov na ustnem izpitu iz znanja slovenščine na osnovni ravni (Požgaj Hadži, Ferbežar 2001: 61–63)

A. Briški je analizirala napake v govoru srbsko, bosansko oziroma črnogorsko govorečih osnovnošolcev in jih razvrstila po jezikovnih ravneh, pa tudi po besednih vrstah (tabela 35).

Jezikovna raven	Besedna vrsta	Napaka	Primer
Glasoslovna raven		Izgovor /e/ in /o/	
		Izgovor /l/ v položaju za samoglasnikom in na koncu besede	
		Napačen naglas	ósnovno; láse; hodím; jezík; tŕgovec
Oblikoslovna raven	Samostalnik	Orodnik ednine	jaz z mamico in ati
		Samostalnik *oko* v tožilniku množine	ima modre barve očesa
	Pridevnik	Tožilnik ednine	v osnovnu šolo
	Zaimek	Osebni zaimek *ona*	ime joj je Marija
	Števnik	Števnik *en* v dajalniku ednine ženskega spola	enej
	Glagol	Število	v družini so dva otroka
		Spol	imam jih rad (govori ženska)
		Povratni glagol	potem sem odločila za to
		Raba namenilnika	gremo kam ven, da se peljemo s kolesom
	Prislov	Stopnjevanje	bolj veliko družiti
		Napačni prislov	gremo nekje; gremo tam, ne gremo nikjer
	Predlog	Napačni predlog	pri družini imam
		Manjkajoči predlog	sem vpisala farmacevtsko šolo

Jezikovna raven	Besedna vrsta	Napaka	Primer
Skladnja		Prilastkov odvisnik	družina, kateri imajo dva otroka in dva starša
		Zaznamovan besedni red	Zdaj sem vpisal se na srednjo šolo ekonomosko
Besedoslovje		Interference z maternim jezikom	kao; ili; škola, strani jeziki

Tabela 35: Napake v govoru srbsko, bosansko in črnogorsko govorečih osnovnošolcev (Briški 2010: 85–87)

K. Rozman je poskušala najti najpogostejše napake nemško govorečih študentov slovenščine, vzroke zanje in predlagati nekaj strategij za njihovo odpravljanje (tabela 36). Razvrstila jih je po jezikovnih ravninah (delno tudi po besednih vrstah) in vzrokih nastanka (Rozman 2006: 39–43).

Jezikovna ravnina	Vzroki	Primer
Fonetične napake 17 % napak	Fonetični zapis	u veliki hiši; glazbo; najljupša
	Nerazlikovanje med sičniki in šumniki	druzina; maško; štarsi; na zvatbo
	Vpliv nemščine	studirati; na Šteierskem
	Narava izgovora /j/, /d/, /h/	včasi; koniček; v Gracu
	Vpliv hrvaščine*	sedamdeset; v Americi
	Vpliv angleščine	economijo
	Bližina fonemov /e/ in /i/ v izgovoru	inginer
	Nepoznavanje soglasniških premen	russčino
	Analogija izgovora črke <v> pri predlogu in v izglasju besede	v Gradcv
	Vpliv pogovorne slovenščine	pred mesci
	Neločevanje zvenečnosti glasov, onemitev izglasnega /j/, redukcija /t/	včasik; skupa; tisoč devesto
Morfološke napake 51 % napak	Neznan vzrok za napako	nimam sestra; ded in babica sta umirale
	Interference (iz nemščine, hrvaščine, pogovorne slovenščine)	v naša hiša; medve igramo; kako te pišeš; njen ime
	Neupoštevanje pravil za izpad polglasnika	enega pesa
	Nepoznavanje pregibnostnega vzorca	v soseščina; zrastal sem
	Kombiniranje dveh možnosti	rodila sem se letu 1981
	Nepoznavanje pravil za ujemanje zaimka ob samostalniku	imam ena brata
Sintaktične napake (besedni red) 5 % napak	Interference iz nemščine	v Austrii pišem se
	Prenizka stopnja jezikovne zmožnosti na začetni stopnji učenja	me blasno njenima kultura zanima

* Prvi jezik dveh študentk je bila hrvaščina, nemščina pa je bila jezik okolja (Rozman 2006: 39).

Jezikovna ravnina	Vzroki	Primer
Leksikalne napake 19 % napak	Nepoznavanje besede	živem; na universiji
	Interference iz hrvaščine	dvaindvadeset
	Neznan vzrok	dobršen interesantno
	Interference iz nemščine	mi napravimo izlet
	Interference iz pogovorne slovenščine	da dobivamo dirko
	Napačno slovenjenje	v Grazu
	Kombiniranje dveh možnosti	v bolnišci
Pravopisne napake (ločila, velika in mala začetnica, pisanje skupaj in narazen) 8 % napak	Neobvladovanje pravil za zapisovanje števil z besedo	tri in trideset
	Interference iz nemščine	Avgust
	Neločevanje mej med besedami	samojo
	Nepozornost na pravopis	Študiram zgodovina in novinarstvo; Nakupujem rad

Tabela 36: Napake nemško govorečih študentov slovenščine (Rozman 2006: 45–61)

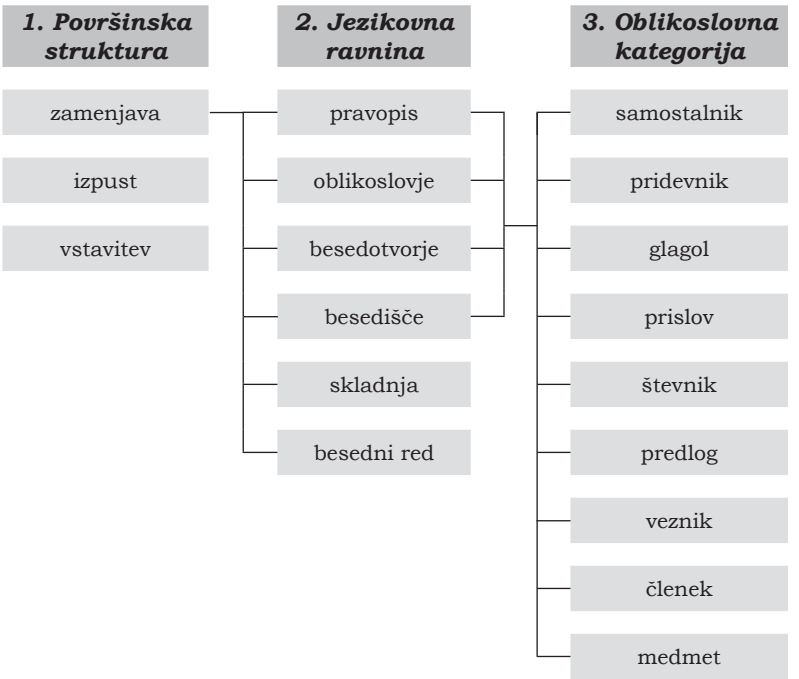
S. Starc je pregledala napake v pisanju italijansko govorečih dijakov, ki opravljajo maturo iz slovenščine kot drugega jezika. Razvrstila jih je na globalne, ki otežijo sporazumevanje, in lokalne, ki pomenijo zgolj odstopanje od norme. Globalne so po njenem mnenju besedni red, raba dialektizmov, pomensko napačno rabljene besede, glagolski vid in neustrezna raba oziralnega zaimka *ki* namesto *kar*. Lokalne so glagolska vezava, predložne zveze, napačna raba ali neuporaba dvojine, napačen sklon in pogovorni nedoločnik (Starc 2005: 164–167).

L. Pogačnik se je osredotočila na napake tujih govorcev v besednem redu. Iz različnih besedil s tečajev in izpitov slovenščine je izpisala 139 povedi z napakami v besednem redu in jih razvrstila na deset tipov z nekaj podtipi (Pogačnik 2001: 26).

5.3 Označevanje napak v korpusu slovenščine kot tujega jezika

Označevanje napak na slovenskem gradivu sem začela s preizkusno zasnovanim, miniaturnim učnim KUST-om. Vse napake so bile označene ročno v jeziku XML po tristopenjski klasifikaciji (prim. Stritar 2006): glede na površinsko strukturo, glede na jezikovno ravnino in glede na oblikoslovno kategorijo (slika 25). Na prvem nivoju so se formalno delile na zamenjave, vstavitve in izpuste (prim. poglavje 5.2.1). Pri za-

menjavah napačna oblika nadomešča pravilno, pri vstavitvah je enkodirana odvečna značilnost, pri izpustih pa značilnost ni enkodirana. Napake zaradi zamenjave so bile tako številne, da so bile zaradi preglednosti členjene naprej po jezikovni ravnini. Napakam na ravni posameznih besed – pravopisnim, oblikoslovnim, besedotvornim in napakam besedišča – pa je bila dodana še ena kategorija, besedna vrsta.



Slika 25: Klasifikacija napak v učnem KUST-u

Ustreznost tega označevanja sem preverila s pomočjo ankete med slovenisti, opisane v uvodnem poglavju, s katero naj bi bila med drugim ocenila stopnjo subjektivnosti pri razvrščanju primerov, posebej pri oblikoslovnih oznakah in navajanju popravljenih oblik. Zavedati pa se je treba, da rezultatov, prikazanih in komentiranih v nadaljevanju, ni mogoče posploševati, posebej zato, ker so anketiranci popravke predlagali brez sobesedila, ki ga vidijo označevalci korpusa.

Anketa in označevanje dodatnega korpusnega gradiva sta pokazala, da na pogled logična tristopenjska klasifikacija napak iz učnega KUST-a ni uporabna. Prva stopnja, sprememba pričakovane oblike, je striktno formalna in s tem pravzaprav nepotrebna komplikacija, saj je malo verjetno, da bi uporabniki korpusa želeli iskati vse vstavitve, temveč bi jih

bolj zanimale konkretnejše napake, npr. napake glagolskih oblik. Formalnost hitro privede do formaliziranja: primer *se trudim, da neham uživati alkohol* je označen kot kombinacija izpusta *bi* in zamenjave *neham* z *nehal*, čeprav gre za rabo napačnega glagolskega naklona.

Tako je za poskusni KUST nastala nova klasifikacija, primerna tudi za pravi KUST, čeprav lahko seveda večja količina gradiva za označevanje zahteva določene prilagoditve kategorij. Te so bile do neke mere že spremenjene, ko je bila PiKUST-ova klasifikacija sicer uspešno aplicirana na gradivo domačih govorcev v korpusu Šolar (prim. Rozman, Stritar, Kosem 2010: 14–25). Osrednja načela predlagane klasifikacije so formalnost, splošnost, objektivnost, opisnost in čim manjša interpretativnost. Označevanje napak namreč ni in ne sme biti tudi njihova analiza ali iskanje vzrokov. Natančno opisovanje napak je ločena dejavnost od sklepanja o njihovih vzrokih (Dulay, Burt, Krashen 1982: 145). Vseeno se je temu nemogoče povsem izogniti, saj razvoj in uporaba klasifikacije pokažeta in ozavestita, katere napake se sploh pojavljajo, in jih s tem – na sicer zelo površinski ravni – tudi že analizirata.

V PiKUST-u na ravni označevanja ni razlikovanja med odkritimi in prikritimi napakami, saj je pomembna tudi razumljivost sporočila (prim. Stabej 2008: 91). V primeru *Pomeriti verige z mrzlimi prsti* je glagol slovnično pravilen, vendar se ob sobesedilu pokaže, da je pomensko neustrezen, ker ne gre za ugotavljanje, ali se verige po dimenzijah prilegajo avtomobilskim gumam, temveč za njihovo nameščanje.[42] Napake so klasificirane dvostopenjsko, po jezikovni ravnini in podrobnejših značilnostih, specifičnih za posamezno ravnino. Shematično so prikazane v tabeli 37, kategorije pa so natančneje opisane v poglavju 5.3.5.

1. stopnja	**Oznaka v PiKUST-u**	**2. stopnja**	**Oznaka v PiKUST-u**
Zapis	Z	Črkovanje	CRK
		Skupaj/narazen	SN
		Mala/velika začetnica	MV
		Ločilo	LOC
		Sekundarna napaka	SEK
Beseda	B	Obstoječa beseda	OBST
		Neobstoječa beseda	NEOBST
		Sekundarna napaka	SEK

42 Vsi primeri v nadaljevanju so iz PiKUST-a, če ni navedeno drugače.

1. stopnja	Oznaka v PiKUST-u	2. stopnja	Oznaka v PiKUST-u
Oblika	O	Sekundarna napaka	SEK
Skladnja	S	Neustrezna struktura	STRUKT
		Besedni red	BR
		Izpuščeni podatki	IZP
		Odvečni podatki	VST
		Sekundarna napaka	SEK

Tabela 37: Klasifikacija napak v PiKUST-u

Napake se na prvi ravni delijo glede na jezikovno ravnino. To je najpogosteje rabljeni način, ki združuje formalnost označevanja in specifične jezikoslovne potrebe. Tudi rezultati iz ankete, komentirani v poglavju 5.3.5, so potrdili, da je v večini primerov tako razvrščanje razmeroma nesporno. Delitev na zamenjave, izpuste in vstavitve pa je vključena v ostale kategorije: zamenjave niso označene, ker so privzeta oblika za vse napake, če ni izrecno povedano drugače, izpusti in vstavitve pa so podtip napak strukture.

Na drugi ravni imajo napake podtipe. Tudi tu naj bi bilo označevanje čim manj interpretativno in čim splošnejše, podrobnosti pa prepuščene analizi usposobljenih uporabnikov. Nezapleteni označevalni sistemi so namreč konsistentnejši (Tenfjord idr. 2004) in objektivnejši. Kljub razdelanim načelom pa se je v nekaterih primerih nemogoče izogniti interpretaciji. Če se označevalec odloči, da je v kontekstu primera 22 *na to* napaka – lahko bi si razlagal tudi, da je govorec »odgovarjal« na vprašanje »Kaj rečete na to?«, kar je, sodeč po podatkih iz korpusa FidaPLUS, v slovenščini idiomatska zveza – se mora odločiti še, ali gre za napako zapisa ali besedišča. Ugibati mora, ali je tvorec želel napisati *nato* oziroma je uporabil predlog *na* namesto *za*. Pri napakah z več možnostmi oznak sta za konsistentnost označevanja zato ključna visoka usposobljenost označevalcev in priročnik za označevanje, ki opredeljuje, opisuje in prikazuje posamezne tipe (Dagneaux idr. 1998: 166). Sicer pa je načeloma označen tip napake, ki se zdi uporabnejši in povednejši (Nicholls 2003: 575).

> Na primer, če ste se odločili iti po nakupih preko interneta, najprej si morete malo pomisliti, ali je to res tako preprosto kot se piše v reklami. Najbrž ne, ker samo v nekaj minutah boste imeli prazen bančni račun, **na to** pa recite »hvala« hekerjem.

Primer 22: Napaka v PiKUST-u z več možnostmi interpretacije

V PiKUST-u ni ločevanja med lapsusi in napakami, saj bi šlo pri tem že za ugotavljanje vzrokov. Kot napake so označeni tudi primeri, ki so skoraj zagotovo nastali kot lapsusi (*današnjim svet* nam. *današnji svet*), njihova obravnava pa je prepuščena uporabnikom.

Napake so bile v PiKUST-u označene v jeziku XML po priporočilih TEI, in sicer z elementom <sic> in dodatnimi atributi. Njihove oznake, navedene v tabeli 37, so bile izbrane tako, da so za veščega uporabnika povedne same po sebi. Prva stopnja napake je označena z atributom *type*, druga z *desc*. Oblikoslovna oznaka je v atributu *msd*, rekonstrukcija pa v *corr* (primer 23).

Na tekočem smo lahko tudi z dognanji Svetovne zdravstvene organizacije o (ne) tveganosti kravjega mleka in mlečnih **<sic type="B" desc="OBST" msd="S" corr="izdelkov">predelkov</sic>**

Primer 23: Oznaka napake v PiKUST-u jeziku XML; B = besedišče, OBST= obstoječa beseda, S = samostalnik, popravljena oblika je »izdelkov«[43]

Napake so označene tik pred pojavitvijo: začetek elementa <sic> je pred samo napako in konec </sic> takoj za njo. Takega načina se držijo v vseh ostalih korpusih, izjema so le leksiko-gramatične napake v korpusu ICLE (primer 24), kjer je oznaka za začetek napake tik pred besedo, ki povzroči napako, in ne pred samo napako (Granger, Thewissen 2006). V PiKUST-u je prepoznavanje konteksta prepuščeno uporabnikom, zato so napake označene samo na besedah, ki jih je treba popraviti. Predlog *na* je v primeru 25 neustrezen, saj je po podatkih iz korpusa FidaPLUS običajna kolokacija *porabiti čas za*. Kljub temu začetek oznake napake ni pred glagolom, ki sproži napako, temveč pred besedo, ki je neustrezna, torej pred predlogom.

Students have the (XNCO) possibility to leave $possibility of leaving$

Primer 24: Napaka v korpusu ICLE z oznako pred besedo, ki sproži napako

koliko v življenju porabimo časa za vsakdanje aktivnosti, koliko **<sic type="B"desc="OBST" msd="D" corr="za">na</sic>** različna čakanja in tako naprej.

Primer 25: Mesto oznake napake v PiKUST-u

43 Možna pravilna oblika bi bil tudi samostalnik *pridelkov*, vendar korpus FidaPLUS pokaže šest zadetkov za kolokacijo *mlečni pridelek* in 5068 za *mlečni izdelek*, ki je torej precej verjetnejši.

Konkretna izvedba označevanja v jeziku XML je seveda odvisna od projekta. Oznake napak v korpusu Šolar, v katerem je bila sicer uporabljena PiKUST-ova klasifikacija, so zato nekoliko drugačne (primer 26). Za vsak učiteljev poseg v besedilo je uporabljen element <u>. Prva stopnja napake je označena z atributom *tip*, druga pa z atributom *podtip*. Za učiteljev popravek napake pa je uporabljen element , gnezden znotraj elementa <u> (Rozman, Stritar, Kosem 2010: 15).

Na **<u tip="Z" podtip="MV">slovenskem Slovenskem </u>** se je

Primer 26: Oznaka napake v Šolarju;
Z = napaka zapisa, MV = mala/velika začetnica

Pred razčlenitvijo načel je treba ponoviti, da je ne glede na natančno razdelane kategorije, omejeno število tistih, ki označujejo napake, in njihove dobre namene označevanje napak vedno subjektivno. Sama sem to izkusila pri delu s PiKUST-om tako v primerjavi z mnenji drugih jezikoslovcev kot celo z lastno tolerantnostjo. Na prvi pogled očitno neustrezne oblike so mi bile po večkratnem pregledovanju že tako domače, da sem jih bolje razumela in jih niti nisem več imela za napačne. Še bolj bi se verjetno razlikovale oznake, ki bi jih besedilom dali tuji govorci slovenščine ali domači govorci, ki ne znajo nobenega tujega jezika, a je v praksi pri označevanju večje količine gradiva to težko izvedljivo. Zato naj bo še enkrat poudarjeno, da je nemogoče določiti enoten, za vse čase in udeležence veljaven sistem označevanja napak.

5.3.1 Oblikoslovne oznake

Napake v PiKUST-u so dobile tudi ročno oblikoslovno oznako, saj korpus ni niti oblikoslovno označen niti lematiziran, ker ni bilo možnosti za edino smotrno avtomatsko označevanje. Za podoben tip oblikoslovnega označevanja so se odločili tudi v korpusu FRIDA (Granger 2003b: 469).

Anketa med slovenisti je pokazala, da je ročno pripisovanje oblikoslovnih oznak napakam razmeroma enostavno in nesporno. Tabela 38 prikazuje 19 primerov, v katerih so anketiranci napačnim pojavitvam pripisali oblikoslovno kategorijo. Nekoliko problematična sta manjkajoči prosti morfem *se* (14. primer), ki je bil v glavnem označen kot glagol, in zveza *arhitektura fakulteta* (12. primer), kjer so anketirani pri prvi besedi oklevali med samostalnikom in pridevnikom, torej med dano in pričakovano obliko.

Št.	Primer	Samostalnik	Pridevnik	Glagol	Prislov	Števnik	Predlog	Veznik	Členek	Medmet	Večbesedna enota	Drugo
1	Ljudje so jo uzeli na piko.			97							3	
2	Tam je najdla svoj mir.			100								
3	Slikal sem Grad na sončnem zahodu.	3					97					
4	Šel sem nazaj v Ljubljani.	93		2	2						2	
5	Ker je sijala polna luna, nisem rabil prvih luči.	9	80		3	6					3	
6	Med vikendom sem veliko berala.		3	94			3					
7	Na prvem vtisu je tema filma politična.	3					12				84	
8	Ob ponedeljkih poučevam nemščino.			100								
9	Plavanje, tek in aerobika so vsi športi da želim.		6	3				50			41	
11	V ponedeljek sem poučevala francoskoščino.	100										
12	Morala je iti na arhitekturo fakulteto.	23	35								42	
13	Dopoldne štiduram slovenščino.			100								
14	Včeraj sem zbudila zelo pozno in sem šla v službo.			90							5	5
15	Jutri bom praznik rojstni dan.	20		73							7	
16	Moja mama tudi je bila prevajalka.			18					18		64	
17	Zjutraj zgodaj sem vstal.			15	23				8		54	
18	Danes se zelo veliko pogovarja o zdravi prehrani.			73							27	
19	Rad grem na piknik z prijatelji.						100					
20	Nisem točno spomnim, kdaj sem se zbudil.			60							40	

Tabela 38: Odgovori ankete med slovenisti o besedni vrsti napačne oblike (v odstotkih)

V učnem KUST-u so bili samostalniki in samostalniški zaimki oziroma pridevniki in pridevniški zaimki združeni v isti kategoriji. V PiKUST-u pa je bil uporabljen nabor iz korpusov MULTEXT-East, FIDA in FidaPLUS (prim. Erjavec idr. 2001), v katerem so samostalniški in pridevniški zaimki v skupni kategoriji zaimek ločeni od samostalnikov in pridevnikov (tabela 39), saj je to koristnejše za uporabnika, hkrati pa so s tem primerljivejši rezultati iz PiKUST-a in najpomembnejših slovenskih korpusov. Kratice za oblikoslovne oznake v PiKUST-u so enake prvemu znaku kod MSD iz omenjenih korpusov. Dodana je kategorija za okrajšavo, večbesedne napake pa zaradi zapletenega gnezdenja oznak, torej umeščanja ene oznake znotraj druge, nimajo oblikoslovne oznake. Tudi anketirani slovenisti večbesednim zvezam niso določali oblikoslovja, temveč so se odločili za oznako *večbesedna enota* (7., 16., 17., 20. primer). Nekatere oznake v PiKUST-u niso uporabljene, vendar so v klasifikaciji, zato da je popolna: tuji govorci v okviru zvrsti besedil, vključenih v korpus, verjetno ne uporabljajo medmetov ali na njih ne delajo napak. Kljub temu je bolje, da je kategorija v sistemu, če bi se taka napaka pojavila, saj jo je enostavneje ignorirati pri obdelavi kot naknadno dodajati.

Oznaka	**Pomen**
S	Samostalnik
G	Glagol
P	Pridevnik
R	Prislov
Z	Zaimek
K	Števnik
V	Veznik
D	Predlog
L	Členek
M	Medmet
O	Okrajšava

Tabela 39: Oblikoslovne oznake v PiKUST-u

Da je označevanje oblikoslovja konsistentnejše, tudi pri tem velja načelo dobre prakse: oblikoslovna oznaka je pripisana izvirni in ne popravljeni obliki. Ponavadi na to ni treba opozarjati, pomembno pa postane npr. v 15. primeru iz ankete *bom praznik prvi rojstni dan*, ki so ga vprašani večinoma uvrstili h glagolu. Čeprav stoji na mestu glagola, je bil v PiKUST-u označen kot samostalnik. Ker pa je ta primer iz besedila nižjega nadaljevalca in ker bodo v KUST vključena

besedila višjih ravni jezikovne zmožnosti, ni pričakovati, da se bo pojavljalo veliko tovrstnih napak. Pogostejše je zamenjevanje besed *ni/ne*. Besedna vrsta pojavnice *ni* je v primeru *označuje ni samo sebe* označena kot glagol, *ne* v primeru *da alkohola ne treba piti* pa kot členek, ne pa glede na besedo, ki bi morala stati na njenem mestu.

Ne glede na relativno neproblematičnost oblikoslovnega označevanja se bodo pri avtomatiziranju tega procesa za KUST pojavile težave z oblikami, ki v slovenščini ne obstajajo (prim. poglavje 3.3.2). Možna rešitev za avtomatski označevalnik je kategorija neznanih oziroma neprepoznavnih oblik, kakršno ima norveški korpus ASK.

5.3.2 Pravilne oblike

Navajanje pravilnih oblik v oznakah napak vzbuja deljena mnenja zaradi problema interpretacije. Bolj kot je omejen kontekst, težje je navajati utemeljene popravke, kot je potrdila tudi anketa med slovenisti (tabela 40). Rekonstrukcije napačnih oblik v korpusu so vedno zgolj verjetne, saj ne temeljijo na poznavanju govorca, njegovega vmesnega jezika ali celo njegovem komentarju, temveč samo na obliki in sobesedilu izjave. Tudi če je pri popravljanju tvorec prisoten, se je včasih težko odločiti med možnimi rekonstrukcijami (Ellis 1994: 57). Navedene pravilne oblike pa skrčijo možne interpretacije, ker označevalci od tvorcev besedil ne morejo izvedeti, kaj so v resnici hoteli sporočiti.

Št.	Primer	Predlagani popravek	Število odgovorov	Delež odgovorov
1	**Ljudje so jo uzeli na piko.**	Ljudje so jo vzeli na piko.	41	91,1 %
		Ljudje so jo imeli na piki.	3	6,6 %
		Ljudje so jo dobili na piko.	1	2,2 %
2	**Tam je najdla svoj mir.**	Tam je našla svoj mir.	42	97,7 %
		Tam je najdla svoj mir.	1	2,3 %
3	**Slikal sem Grad na sončnem zahodu.**	ob sončnem zahodu.	31	65,9 %
		v sončnem zahodu.	13	27,6 %
		pri sončnem zahodu.	1	2,1 %
		s sončnim zahodom.	1	2,1 %
		Na Gradu sem slikal sončni zahod.	1	2,1 %
4	**Šel sem nazaj v Ljubljani.**	Šel sem nazaj v Ljubljano.	41	100 %

Št.	Primer	Predlagani popravek	Število odgovorov	Delež odgovorov
5	**Ker je sijala polna luna, nisem rabil prvih luči.**	sprednjih luči.	21	43,75 %
		dolgih luči.	12	25 %
		luči.	5	10,4 %
		prednjih luči.	4	8,3 %
		kratkih luči.	3	6,25 %
		pravih luči.	1	2,1 %
		prvih luči.	1	2,1 %
		žarometa.	1	2,1 %
6	**Med vikendom sem veliko berala.**	Med vikendom sem veliko brala.	40	100 %
7	**Na prvem vtisu je tema filma politična.**	Po prvem vtisu je	11	27,5 %
		Na prvi pogled je	11	27,5 %
		Na prvi vtis je	8	20 %
		Ob prvem vtisu je	6	15 %
		Glede na prvi vtis je	1	2,5 %
		V prvem vtisu je	1	2,5 %
		Prvi vtis je	1	2,5 %
		»pogledati v korpus«	1	2,5 %
8	**Ob ponedeljkih poučevam nemščino.**	Ob ponedeljkih poučujem nemščino.	40	100 %
9	**Plavanje, tek in aerobika so vsi športi da želim.**	so vsi športi, ki jih želim.	11	35,5 %
		so vsi športi, ki si jih želim.	8	25,8 %
		so vsi športi, ki so mi všeč.	3	9,7 %
		so moji priljubljeni športi.	3	9,7 %
		so vsi športi, ki jih imam rad.	2	6,4 %
		so vsi športi, da jih želim.	2	6,4 %
		so vsi športi, s katerimi se rad ukvarjam.	1	3,2 %
		so vsi športi, s katerimi se želim ukvarjati.	1	3,2 %
11	**V ponedeljek sem poučevala francoskoščino.**	V ponedeljek sem poučevala francoščino.	38	100 %
12	**Morala je iti na arhitekturo fakulteto.**	na fakulteto za arhitekturo.	26	81,25 %
		na arhitekturno fakulteto.	5	15,625 %
		na arhitektonsko fakulteto.	1	3,125 %
13	**Dopoldne študiram slovenščino.**	Dopoldne študiram slovenščino.	27	100 %
14	**Včeraj sem zbudila zelo pozno in sem šla v službo.**	Včeraj sem se zbudila zelo pozno in sem šla v službo.	23	100 %

Št.	Primer	Predlagani popravek	Število odgovorov	Delež odgovorov
15	**Jutri bom praznik rojstni dan.**	Jutri bom praznoval(a) rojstni dan.	22	91,6 %
		Jutri praznujem rojstni dan.	1	4,2 %
		Jutri bom praznil rojstni dan.	1	4,2 %
16	**Moja mama tudi je bila prevajalka.**	Moja mama je bila tudi prevajalka.	19	73 %
		Moja mama je tudi bila prevajalka.	7	27 %
17	**Zjutraj zgodaj sem vstal.**	Zjutraj sem zgodaj vstal.	19	95 %
		Zgodaj zjutraj sem vstal.	1	5 %
18	**Danes se zelo veliko pogovarja o zdravi prehrani.**	Danes se zelo veliko govori	13	76,5 %
		Danes se zelo veliko pogovarjamo	3	17,6 %
		Danes sem se zelo veliko pogovarjal	1	5,8 %
19	**Rad grem na piknik z prijatelji.**	Rad grem na piknik s prijatelji.	19	100 %
20	**Nisem točno spomnim, kdaj sem se zbudil.**	Ne spomnim se točno	15	83,3 %
		Nisem se spomnil	1	5,5 %
		Nisem se točno spomnila	1	5,5 %
		Ne se spomnim točno	1	5,5 %

Tabela 40: Predlagane pravilne oblike iz ankete med slovenisti

Znotraj konteksta bi bilo takoj jasno, ali se 20. primer nanaša na preteklost in ali piše moški ali ženska, zato bi bil tudi popravek enoznačnejši. V 5. primeru bi učitelj v razredu s pomočjo učenca hitro ugotovil njegov namen, v anketi pa so si predlogi popravkov celo pomensko nasprotovali.

Označevalec korpusa je manj omejen kot anketiranci, saj ima pred seboj razmeroma koherentno in kohezivno besedilo. Vseeno so problem napake govorcev z nižjo stopnjo jezikovne zmožnosti, ki so včasih tako nekoherentne in znotraj tako omejenega sobesedila, da je nemogoče navesti enoumen popravek. V povedi *Od enih do štirih sem se zjasnil med svojo hišo in to sosedo Andreja* glagol *zjasniti* ni ustrezen. Glede na kontekst besedila angleškega govorca o zimi se ponudi *skidati*. To potrdi angleško-slovenski slovar, ki *zjasniti se* navede kot prevod angleškega *clear up*, torej gre za možno napako zaradi uporabe slovarja. V stavku *Danes že vredu letališče* ni jasno, za kaj gre, tudi zaradi omejenega konteksta kratkega besedila. Popravkov pri nekaterih večjih izpustih pa sploh ni mogoče navajati, npr. v povedi *Zjutraj sem zbudila zelo pozno, nisem in zato sem morala biti hitro*.

Pri navajanju pravilnih oblik je treba upoštevati, da je tudi označevalec samo človek, ki nekaterih interpretacij ne predvidi, čeprav so popolnoma sprejemljive. V 1. primeru *Ljudje so jo uzeli na piko* je večina predlagala popravek *Ljudje so jo vzeli na piko*, slabi desetini pa se frazem *vzeti na piko* ni zdel ustrezen in so ga nadomestili z *imeti na piki* oziroma *dobiti na piko*.[44]

Spornost navedenega popravka je odvisna od tipa napake. Anketa med slovenisti je pokazala, da so manj vprašljivi pri težavah zapisa, oblikoslovja in manjkajočih besed (primeri 2, 4, 6, 8, 11, 13, 14, 19). Zapletati se začne, kjer je v jeziku več možnosti. Besedna zveza 12. primera ima *fakulteto za arhitekturo*, ki je po referenčnih virih verjetnejša, in *arhitekturno fakulteto*.[45] Besedni red oziroma stavo naslonskega niza v 16. primeru so anketiranci popravili v *Moja mama je bila tudi prevajalka* in *Moja mama je tudi bila prevajalka*. Več alternativ je pri veznikih; 9. primer so anketirani parafrazirali na različne načine od *ki jih želim* do *s katerimi se želim ukvarjati*. Podobno velja za predloge v 7. primeru.

Zaradi subjektivnosti pravilne oblike v učnem KUST-u niso bile dodane v oznake napak. Toda čeprav verjetne rekonstrukcije nikoli niso stoodstotno zanesljive, se jim ni mogoče izogniti, saj naslovnik v besedilih vedno išče namen in pomen napisanega (Tenfjord 2006). Branje je pragmatično, besedila morajo nekaj pomeniti in opisovanje napak vedno poteka glede na »prevod« napačne oblike v ciljni jezik (Corder 1982: 37). To velja tako za učitelje v razredu kot za graditelje korpusov. Zakaj torej navajati pravilne oblike, če so vedno subjektivne? Rekonstrukcije so priložnost za dodatna pojasnila, večjo razumljivost in lažje preučevanje vmesnega jezika (Reder idr. 2003: 554). S tehničnega vidika olajšajo avtomatsko lematizacijo, oblikoslovno ali skladenjsko označevanje, nudijo dostop do informacij, ki bi bile sicer dostopne le posredno (Nicholls 2003: 573), in podatke, v katerih neustreznih oblikah se lahko nekaj pojavlja – S. Granger (2003b: 470) navaja primer sedmih različnih neustreznih zapisov francoskega samostalnika *développement* v korpusu FRIDA.

V vseh pregledanih korpusih, v katerih označujejo napake, je poleg tipa napake navedena pravilna oblika. Zaželeno je,

44 *Vzeti na piko* se v SSKJ pojavi kot *jemati na piko*, v FidiPLUS ima *vzeti na piko* 136 zadetkov, *jemati na piko* pa 2. *Imeti na piki* je v SSKJ in ima v FidiPLUS 53 zadetkov. *Dobiti na piko* ni v SSKJ, v FidiPLUS je 9 zadetkov.

45 Fakulteta za arhitekturo ima v FidiPLUS 1291 zadetkov, arhitekturna fakulteta pa 10.

da je popravek čim bližje prvotni podobi izjave, ne pa idealizirana parafraza (Nicholls 2003: 573). Pri ugotavljanju, kaj je še sprejemljivo in kaj bi bil ustrezen popravek, se je najustrezneje sklicevati na referenčne korpuse, saj ti »v svojem izhodišču prisegajo na jezikovno realnost in tudi ob nepričakovanih rezultatih ne podlegajo intuiciji, vključujejo več podatkov o tipičnem besedilnem okolju ter sploh podatkov o komunikacijski realnosti« (Gorjanc 2003: 26).

Ker so koristi navajanja popravkov večje od njegovih omejitev, so bili v PiKUST-u navedeni, dvomljive oblike pa preverjene v FIDI oziroma FidiPLUS. Uporabnik mora ob iskanju in pregledovanju rezultatov upoštevati, da so pripisani popravki možna, načeloma tudi najverjetnejša, a nikakor ne ena in edina pravilna oblika (Dagneaux idr. 1998: 165). Kadar pa bi popravek zahteval preveliko ugibanje, v PiKUST-u ni zapisan in je označen z *[?]* (primer 27), saj bi označevalčevo ugibanje podatke samo popačilo (Nicholls 2003: 576). V PiKUST-u je takih le 0,7 % vseh napak. Za podobno strategijo so se med drugim odločili v korpusih ASK in AKCES/CzeSL (prim. Hana idr. 2010: 15).

Ampak to je človeški last da se spreminja in se **ugaja** s svojim časom.
Primer 27: Napaka v PiKUST-u brez navedenega popravka

5.3.3 Povezane napake

V besedilih tujih govorcev se pojavljajo oblike, ki so v danem kontekstu ustrezne, neustrezne pa postanejo, ko je popravljena napaka v sobesedilu. V zvezi *hvala na razumevanju* je neustrezen predlog *na*, toda ko je popravljen v *za*, bi moral biti zaradi vezave s popravljenim predlogom samostalnik *razumevanje* v tožilniku namesto v mestniku. Do tvorcev bi bilo krivično in neprimerno za kasnejšo analizo, če bi bilo to označeno kot prava napaka, zato se v obstoječih korpusih tega lotevajo različno. V korpusu ICLE tovrstni pojavi niso označeni (Granger, Thewissen 2006). Ker ne odražajo neznanja, pač pa slab nadzor, naj ne bi bili napake v jezikovni zmožnosti, temveč zgolj v performanci, s katerimi učeči se pokažejo konsistentnost (James 1998: 116). Tudi v učnem KUST-u jih nisem označevala. Toda v tem primeru popravljene oblike besedil ne ustrezajo normam ciljnega jezika. Zato je koristno, da so tovrstne napake posebej označene, uporabnik pa se mora ob pregledovanju korpusa zavedati, da niso prave napake.

Pri označevanju je možnih več načinov. V PiKUST-u, ki je bil zasnovan po zgledu korpusa ASK, v katerem uporabljajo posebni atribut AGR, ima vsak krovni tip napake atribut sekundarna napaka (SEK, primer 28). Ena primarna napaka lahko povzroči več sekundarnih. V povedi *Všeč mi je /.../ da so moji delovni časi zelo fleksibilni* je *čas* v množini namesto v ednini. Posledično so napačne tudi oblika glagola *biti*, svojilnega zaimka *moj* in pridevnika *fleksibilen*. Sekundarnih napak je v PiKUST-u 4,9 %; večinoma so oblikoslovne.

ker je prepovedano še slajše, najstniki posegajo <sic type="B" desc="OBST" msd="D" corr="po">za</sic> **<sic type="O" desc="SEK" msd="S" corr="alkoholu">alkoholom</sic>**.

Primer 28: Oznaka sekundarne napake v PiKUST-u

Drugačna rešitev pa je uporabljena pri označevanju napak v korpusu Šolar, kjer se tovrstne pojavitve imenujejo povezane napake. Označene so z atributom *pov* in z zaporedno številko povezanih napak znotraj odstavka (primer 29); »prve povezane napake znotraj odstavka imajo številko 1 (*pov="1"*), druge številko 2 in tako naprej; v novem odstavku se spet začne šteti od 1 naprej« (Rozman, Stritar, Kosem 2010: 24). Prednost tega načina označevanja je, da v njem ni treba opredeliti, katera od povezanih napak je dejanska napaka, katera pa zgolj njena posledica, saj se je marsikdaj o tem težko odločiti.

Izidor mu **<u tip="S" podtip="ODV" pov="1">za </u> <u tip="O" pov="1">to tega </u>** dejanje ni zameril

Primer 29: Oznaka sekundarne napake v korpusu Šolar

Poseben razmislek zahteva primer, ki se je v PiKUST-u pojavil enkrat: oblika ni bila ustrezna, toda ko je bila popravljena napaka v sobesedilu, ni bila več napačna. V stavku *Dopoldne sem šla videti nekaj hiše s moj jefica* bi namesto predloga *s* moral biti predlog *z*. Hkrati je treba zaimek *moj* nadomestiti s povratnosvojilnim *svoj*, kombinacija prvotnega predloga *s* in zaimka *svoj* pa je popolnoma ustrezna. V PiKUST-u je taka napaka, ki s popravkom sobesedila postane pravilna, ostala neoznačena. Druga, zapletenejša, a za natančneje označen korpus primernejša možnost je popraviti napačno in nato, ob popravku sobesedila, spet navesti prvotno obliko.

5.3.4 Dvojne napake

Če je določen del besedila neustrezen v več pogledih, je vsaka vrsta napake označena ločeno, s svojim elementom in atributi. Označevalni jezik XML omogoča gnezdenje oznak in s tem pripisovanje več kategorij eni sami napačni pojavitvi. Splošna tendenca označevanja dvojnih napak je, da je najprej[46] označena napaka, katere popravek jo bolj približa pričakovanim, v slovenščini obstoječim oblikam (prim. Nicholls 2003: 575).

Vrstni red dvojnih popravkov sicer ni vedno pomemben, toda nesmiselno je v oznakah popravljenih oblik navajati neslovenske besede s slovenskimi končnicami ali pravopisno neustrezne besede. V primeru 30 bi bilo nesmotrno, da bi samostalniku *uporabci* najprej dali ustrezno končnico za četrti sklon, torej *uporabce*, in ga nato popravili v *uporabnike*. Ustrezneje je najprej popraviti besedo v *uporabniki* in potem tožilnik množine *uporabnike*. Pri dvojnih napakah, ki vključujejo črkovanje, je to praviloma označeno kot manj pomembna napaka (primera 31 in 32). Včasih pa se je vendarle nemogoče izogniti navajanju popravljenih oblik, ki niso v skladu z normo, denimo pri dvojnih napakah zapisa kot v primeru 33.

življenje postane lažje za **<sic type="O" msd="S" corr="uporabnike"><sic type="B" desc="NEOBST" msd="S" corr="uporabniki">uporabci</sic></sic>**
Primer 30: Oznaka dvojne napake besede in oblike v PiKUST-u

kersemsenaučila**<sictype="B"desc="OBST"msd="P"corr="knjižnega"><sic type="Z" desc="CRK" msd="P" corr="književnega">književnega</sic></sic>** jezika
Primer 31: Oznaka dvojne napake besede in zapisa v PiKUST-u

naši dedi pod **<sic type="O" msd="S" corr="fašizmom"><sic type="Z" desc="CRK" msd="S" corr="fašizem"> fašism</sic></sic>**
Primer 32: Oznaka dvojne napake oblike in zapisa v PiKUST-u

se iz šole nosijo **<sic type="Z" desc="CRK" msd="D" corr="v glavnem"><sic type="Z" desc="SN" corr="u glavnem">uglavnem</sic></sic>** lepi spomini.
Primer 33: Oznaka dvojne napake zapisa v PiKUST-u

46 Kaj to pomeni pri konkretnem označevanju, je odvisno tudi od konkordančnika, ki podpira delovanje korpusa. Konkordančnik PiKUST-a je najprej prikazoval oznake, kot si sledijo od prve, najpomembnejše in notranje proti zunanji strani, torej od izvirne pojavitve navzven, medtem ko je pri oznakah v Šolarju ravno obratno, sledijo si od zunaj navznoter (prim. Rozman, Stritar, Kosem 2010: 24).

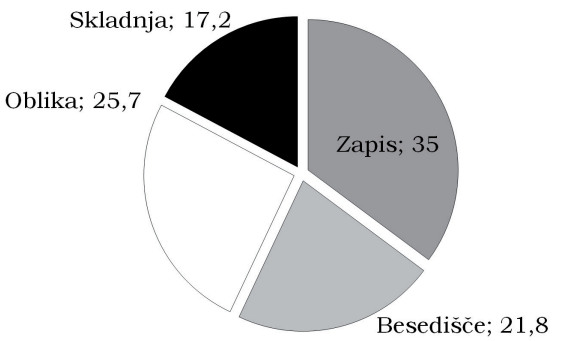
Pojavljajo se celo trojne napake. V primeru 34 gre za veliko začetnico, sklon in črkovanje. Najprej je popravljeno črkovanje; *Radiji* je spremenjen v *Radii*. Sledi sklon, ki je popravljen v mestnik. Čeprav raba napačnega sklona potencialno bolj vpliva na razumevanje kot pravopis, bi bilo nesmiselno najprej popraviti sklon in nato zapis, saj bi bila s tem kot pravilna navedena pravopisno neustrezna oblika. Najmanj pomembna napaka pa je zapis velike oziroma male začetnice.

v časopisih in <sic type="S" desc="IZP" msd="D" corr="na "></sic>**<sic type="Z" desc="MV" msd="S" corr="radiu"><sic type="O" msd="S" corr="Radiu"><sic type="Z" desc="CRK" msd="S" corr="Radii">Radiji</sic></sic></sic>**

Primer 34: Trojna napaka v PiKUST-u

5.3.5 Kategorije napak

Kategorije napak, opisane v nadaljevanju, so bile opredeljene na podlagi rešitev obstoječih tujih korpusov, učnega KUST-a, ankete med slovenisti in PiKUST-a. Kasneje so bile nekoliko prilagojeno uporabljene tudi v Šolarju, tako da so očitno dovolj univerzalne in primerne za označevanje bodočega večjega korpusa usvajanja slovenščine kot tujega jezika. Njihovo primernost potrjuje enakomerna porazdelitev iz PiKUST-a (slika 26), ki kaže, da nobena jezikovna ravnina izrazito ne odstopa, torej da npr. napak oblike ni bistveno več kot napak besedišča.[47] Porazdelitev vseh kategorij v padajočem redu pogostosti je prikazana v tabeli 41.



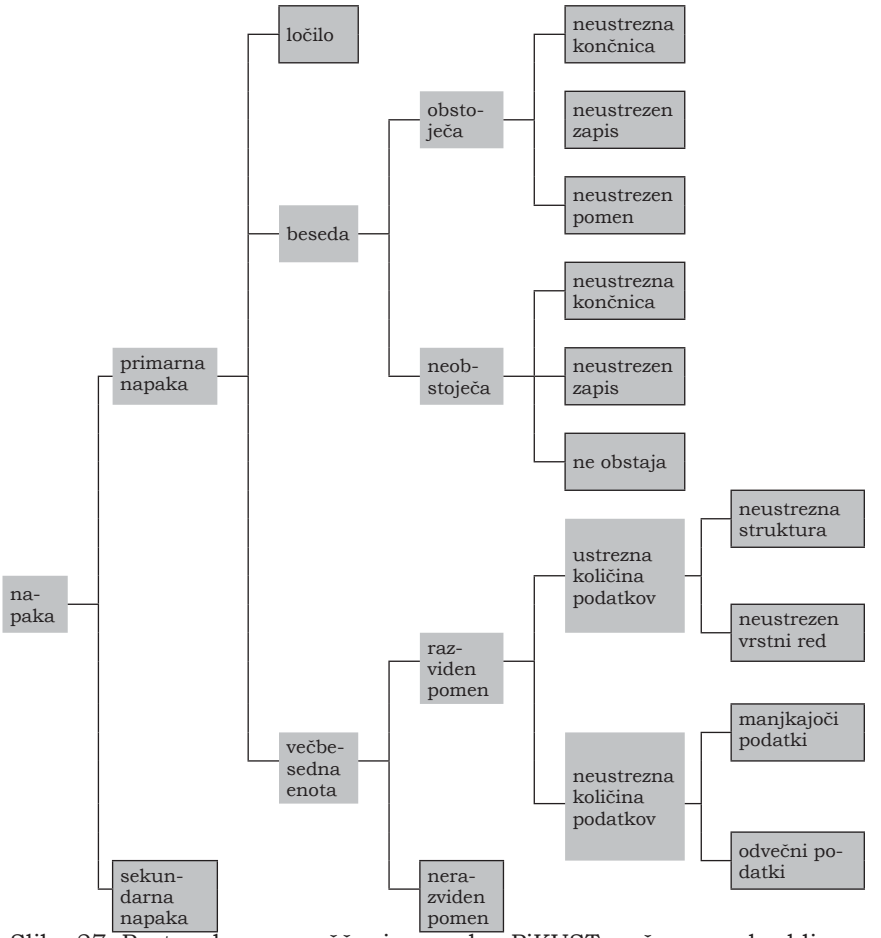
Slika 26: Deleži napak v PiKUST-u po jezikovnih ravninah

47 V izračunu so upoštevane sekundarne napake.

Kategorija napake	Število pojavitev	Delež
Oblika	1219	23,9 %
Zapis: ločilo	1102	21,7 %
Besedišče: obstoječa beseda	671	13,2 %
Skladnja: besedni red	406	7,9 %
Zapis: črkovanje	311	6,1 %
Besedišče: neobstoječa beseda	255	5 %
Sekundarna napaka	250	4,9 %
Skladnja: izpuščeni podatki	130	2,5 %
Skladnja: odvečni podatki	126	2,4 %
Zapis: skupaj/narazen	75	1,4 %
Skladnja: struktura	47	0,8 %
Zapis: mala/velika začetnica	29	0,5 %
Skupaj	**5085**	

Tabela 41: Porazdelitev kategorij napak v PiKUST-u

Napake so »največkrat kompleksne, kar pomeni, da sodijo v več kategorij hkrati, zato jih ni mogoče vedno povsem natančno lokalizirati« (Rozman 2006: 26). Zaradi tega so v nadaljevanju poglavja pri posameznih kategorijah napak, ilustriranih s primeri iz PiKUST-a, izpostavljeni problemi, kjer se je bilo pri razvrščanju treba odločiti za eno možnost, čeprav tudi ostale niso nujno neustrezne. Navedene so tudi morebitne drugačne rešitve iz Šolarja. Da bi bilo razvrščanje enotnejše in objektivnejše, sem si med označevanjem PiKUST-a pomagala s standardiziranim postopkom (slika 27). Vsaka pojavitev, zaznana kot napačna, je bila opredeljena glede na to, ali je napaka primarna ali sekundarna (prim. poglavje 5.3.3). Sledila je formalna razporeditev, ali je napaka ločila, besede ali večbesedne enote. Pri ločilu je bilo treba določiti, ali je neustrezno, odvečno ali izpuščeno. Napačne besede so bile razporejene glede na to, ali v slovenščini obstajajo ali ne. Obstoječe besede so bile razvrščene v podtip glede na neustreznost zapisa, končnice ali pomena, neobstoječe pa glede na to, ali so neustrezne zaradi končnice, zapisa ali tega, ker v slovenščini v celoti ne obstajajo. Napake večbesednih enot so se najprej razvrščale glede na ustreznost količine podatkov in glede na ustreznost vrstnega reda podatkov oziroma celotno strukturo. Sekundarne napake pa so se razvrščale po enakih tipih in oblikoslovnih oznakah, toda ker je »SEK« že oznaka podtipa, niso dobile še ene oznake podtipa.



Slika 27: Postopek za razvrščanje napak v PiKUST-u; črno so obrobljene končne kategorije napak

5.3.5.1 Zapis

Napake v zapisu – enkodiranju pisnega besedila – so omejene na pisno podobo besedila. Poleg napačnega črkovanja mednje sodijo podkategorije napak ločil, velike in male začetnice ter neustrezna pisava skupaj oziroma narazen. V klasifikaciji za korpus Šolar, v katerem je zaradi šolske produkcije besedil večji poudarek na tipično pravopisnih problemih, pa sta bili dodani še podkategoriji številka za menjavo zapisa s številko oziroma z besedo pri glavnih, vrstilnih ali drugih številkih in krajšava za neustrezno rabo okrajšav (Rozman, Stritar, Kosem 2010: 19). Napake zadnjih dveh tipov v pisanju učečih se sicer niso zelo verjetne in zato v nadaljevanju niso podrobneje razložene.

Kadar do napake zapisa pride zaradi popravka napake v sobesedilu, gre za povezane oziroma sekundarne napake. V primeru 35 je tak predlog *z*, ki je v danem kontekstu ustrezen, toda ko je popravljen svojilni zaimek v *svojo*, je treba popraviti tudi predlog v *s*. Ko pa v primeru 36 izpustimo osebni zaimek na začetku, postane odvečna tudi vejica za zvezo *v pravi trenutki*.

Dopoldne štiduram slovenščino **z** mojo mamo.
Primer 35: Povezana napaka črkovanja

Zame**,** v pravi trenutki, jaz maram kozarec dobro slovensko vino, ampak, nisem alkoholika, in nisem najstnika.
Primer 36: Povezana napaka ločila

Ker označevanje napak v korpusu ni določanje vzroka za njihov nastanek, pri napakah zapisa ni ločevanja med lapsusi in napakami. Kot take so označene tudi tiste, ki so očitno nastale pri tipkanju in najverjetneje niso posledica pomanjkljive jezikovne zmožnosti. Presoja o tem je prepuščena kasnejši analizi. Ta princip je ohranjen tudi v Šolarju (navedeno delo: 15–16).

V učnem KUST-u in v anketi med slovenisti so se napake zapisa imenovale napake pravopisa, kar pa je manj ustrezno, saj so težave zapisovanja ločil, skupaj in narazen ter z malo ali veliko začetnico pravopisne in relevantne v bolj sofisticirani pisni produkciji,[48] pri mnogih težavah s črkovanjem pa gre za specifične probleme tujih govorcev, ki ne sodijo v tradicionalno razumevanje pravopisa.[49]

Napake zapisa na razumljivost besedil ponavadi ne vplivajo tako močno kot nekateri drugi tipi, zato so ovrednotene kot manj resne. Na izpitih iz znanja slovenščine se v nalogah, ki od kandidatov zahtevajo pisanje, odbija »samo pol točke za napake zapisa, po eno pa za besedje, oblikoslovje, skladnjo« (Ferbežar 2007c: 34). Uporabnik korpusa pa naj jim sam nameni toliko pozornosti, kolikor se mu zdi potrebno in primerno.

5.3.5.1.1 *Črkovanje*

Napake črkovanja se nanašajo na neustrezen zapis glasov. T. Balažic Bulc (2004: 81) jih imenuje napake pri glasovnih pre-

48 Elektronska korespondenca z Markom Stabejem, 28. oktobra 2006.

49 V Pravilih *Slovenskega pravopisa* (2001: 69–86) so v delu o zapisovanju glasov izpostavljeni /v/, /j/, /l/, /n/, zvenečnostni pari nezvočnikov, zapis več enakih soglasnikov, zveze zobnega zapornika in zlitnika, sičnika pred šumevcem, zvez samoglasnikov in neobstojnih samoglasnikov.

menah, K. Rozman (2006: 45) pa fonetične napake, čeprav poudarja, da za pisno produkcijo ta izraz ni zares primeren. Gre torej za napake, povezane s pisno realizacijo fonemov, v glavnem soglasnikov in sklopov, ki so težavni tudi za domače govorce, kot je <nj> v *književnost*. V anketi so se slovenisti soglasno odločili, da je dilema <v> ali <u> v 1. primeru *Ljudje so jo uzeli na piko* napaka pravopisa. Tabela 42 prikazuje najpogostejše napake črkovanja v PiKUST-u. Zaradi preglednosti in lažjega razvrščanja so združene po sklopih, sicer pa so vse označene zgolj z atributom črkovanje. V stolpcu Primer iz PiKUST-a je odebeljena tista oblika, na katero se nanaša oznaka.

Napaka črkovanja	**Primer iz PiKUST-a**
Neupoštevanje premene po zvenečnosti	tovarnjaki povzročajo hrup, **ispuščajo** pline, ki onesnažujejo zraka,
Zapis <lj> in <nj>	so samozavestni, **nepotrpežlivi**, arogantni
Zapis <v>, <u> in <l>	lahko **uplivajo** na način misljenja
Zapis <j> v stiku samoglasnikov	Neprimerni **materijali**
Izpust ali vstavitev samoglasnikov	je hitrejše zato, kjer število stvari, ki moramo **narditi**, poveča.
Izpust ali vstavitev soglasnikov	kar **pretcej** najstnikov se nebi strinjalo
Zapis predlogov *z/s* in *k/h*	Ko pridem do koncerta **z** svojo punco
Zapis šumevcev in sičnikov	je starejša generacija večja, kot **mlajsa**.
Neslovenski zapis drugih glasov	tudi so informacije o **concerti**

Tabela 42: Pregled napak črkovanja v PiKUST-u

Napake zapisa velikokrat nastanejo pri fonetičnem zapisu besed, kadar tvorci ne upoštevajo premene po zvenečnosti (prim. Ferbežar 2007c: 32; Rozman 2006: 45; Balažic Bulc 2004: 81): *negdaj, glazba, iskušnje, apsolutno, ot centra*. Sklopa <lj> in <nj> sta tipični težavi črkovanja, pravopisa in tudi izgovora (Požgaj Hadži, Ferbežar 2001: 61). Pogost je izpust črke <j> zaradi otrditve v izgovoru: *škodlivo, cil življenja, v današnem življenju, spreminati*. Prav tako so napake zapisa zamenjave črk znotraj teh sklopov (*v sedajnosti*) in vstavitev črke <j>, kjer je v slovenščini ni: *revoljucija, maternjem jeziku, ljubezenj*. Težave z grafemoma <v> in <u> se poleg zamenjav pri predponah (*vpokojenci*) pojavljajo pri predlogu *v* (*u bistvu*) in v besedah tujega izvora. Posebej težaven je zapis <v> za samoglasnikom (*autoritativen*, *automobil*). Ravno tako je problematičen zapis dvoustnično izgovorjenega <l>, ki ga tuji govorci zapisujejo z <u> ali <v> (*popovnoma*). V sklop dveh samoglasnikov ali v zev tvorci vrivajo <j> (prim. Balažic Bulc 2004: 81), npr. *asocijalni*. Prihaja tudi do

drugih realizacij, denimo do izpusta <j> (*milioni*) ali zamenjave <i> z <j> (*emocjonalno stanje*).

Fonetični zapis se včasih realizira kot izpust oziroma redukcija posameznih samoglasnikov. Pri tem gre pogosto za zapis reducirane oblike, značilne za govorjeni jezik (*nardi škodo, radi majo, tko pijane, zadost, še posebi na možgane, komi čakam*; Rozman 2006: 47). Značilna je tudi zamenjava <e> in <i> zaradi bližine [é] in [i] v izgovoru (*kontroleramo*; Rozman 2006: 47). Pojavlja se še vstavljeni samoglasnik pred <r> (*obderžati*) ali v drugih soglasniških sklopih (*sočutije*). S fonetičnim zapisom besed je povezan izpust soglasnikov v soglasniških sklopih (*štiriinpedeset, kašno rastavo, šezdeset minut, smo se otujili*; Rozman 2006: 47; Balažic Bulc 2004: 81). Včasih govorci v zapisu vstavijo <t>: *bogatstvo, odvistnji*.

Ne nujno fonetično pogojena je raba neustreznega predloga *z* ali *s* (*zabava s alkoholom*), ki se v PiKUST-u pojavi 41-krat in bi jo lahko označili kot napako zapisa, besedišča, oblike ali skladnje. To je bilo preverjeno tudi v anketi med slovenisti (tabela 43). Dobra desetina anketiranih je menila, da gre za napako besedišča oziroma oblike z izbiro variante predloga, skoraj tri četrtine vprašanih pa so se odločile za pravopis. To je obveljalo v PiKUST-u in Šolarju (prim. Rozman, Stritar, Kosem 2010: 17). Enako velja za zamenjavo predlogov *k* in *h*, ki ga tudi *Slovenski pravopis* (2001: 80) omenja pri zapisu posebnih glasovnih zvez, čeprav take napake v PiKUST-u ni bilo.

Št.	Primer	Jezikovna ravnina (v %)			
		Pravopis	Besedišče	Oblikoslovje	Skladnja
19	**Rad grem na piknik z prijatelji.**	73 %	7 %	13 %	7 %

Tabela 43: Odgovori iz ankete med slovenisti o primeru napake zapisa predloga *z/s*

Čeprav naj bi bili izpust diakritičnih znakov in težave s šumevci bolj značilni za tuje kot za domače govorce, saj naj bi bila za tujce »raba diakritičnih znakov poljubna ali celo nerelevantna v vseh situacijah, še posebej pa pri zapisu tistih besed, ki jih npr. iz drugih jezikov poznajo brez diakritičnih znakov« (Pirih Svetina 2005: 124), je zanimivo, da so pogosti tudi pri pisanju domačih govorcev slovenščine (prim. Rozman, Stritar, Kosem 2010: 17). Razumevanja to niti ne oteži, kar kaže denimo izkušnja branja elektronske pošte (Pirih Svetina 2005: 124), vendar težavnost razumevanja ni edini

kriterij za določanje, kaj je napaka in kaj ni. Formalno ti primeri odstopajo od norme slovenskega knjižnega jezika, zato so v PiKUST-u označeni (prim. Rozman 2006: 47). Pojavljata se obe varianti – zapis sičnika namesto šumevca (*misljenja, zivljenje, studij*) in šumevca namesto sičnika (*ne bo najdel čaša*) – poleg tega pa še druge zamenjave (*včeš so mu*). Kasnejšim analizam je prepuščeno ugotavljanje, ali je tak zapis nastal zaradi lapsusa, pomanjkljive jezikovne zmožnosti ali prenosa iz prvega jezika.

Napake, po katerih so tuji govorci lahko prepoznavni, so neobičajni zapisi slovenskih besed, kakršnih domači govorci verjetno ne bi tvorili. Gre za neustrezen izbor grafema (*neonaziste, svoi, to be radi verjeli*), podvojene črke (*naporen process*) ali rabo neslovenskih črk (*većinoma*). Omeniti je treba tudi neustrezni zapis naglasnih znamenj (*se lahko zgodi zlórab*) in zapis tujih besed, ki ne ustreza slovenskim normam (*Mekdonalds*).

Že večkrat omenjeno osnovno pravilo klasifikacije napak za slovenščino je, da je označevanje formalno in ne temelji na ugibanju, kaj je lapsus in kaj prava napaka. V PiKUST-u je bilo kršeno v primerih očitnih lapsusov, za katere bi bilo krivično, če bi bili označeni npr. kot napake besedišča (*vska generacija, pomebno*). Podobno so se odločili slovenisti v anketi (tabela 44), kjer je slaba polovica vprašanih sicer menila, da je *štiduram* namesto *študiram* napaka besedišča, le malo manj pa, da gre za problem zapisa.

Št.	Primer	Jezikovna ravnina (v %)			
		Pravopis	Besedišče	Besedotvorje	Oblikoslovje
13	**Dopoldne štiduram slovenščino.**	32 %	41 %	4 %	23 %

Tabela 44: Odgovori iz ankete med slovenisti o primeru napake zapisa/lapsusa

Napake zapisa se včasih prekrivajo z napakami besedišča in oblike. K. Rozman (2006: 46) trdi, da bi se lahko izognili mnogim oblikoslovnim napakam, če bi v celoti odpravili napake zaradi nerazlikovanja fonemov. Načeloma so k napakam besedišča uvrščeni primeri, kjer neustrezna črka, posebej v korenu, besedo tako spremeni, da nastane nekaj čisto drugega (primer 37). To je posebej smiselno pri popačenkah (angl. *malapropism*), leksikalnih napakah, ki nastanejo zaradi zamešanja podobno zvenečih morfemov ali besed (James 1998: 134). V primeru 38 je zaradi neustrezne črke nastala

beseda z drugim pomenom. V formalnem smislu je tvorec izbral besedo *raven* namesto *reven*, če pa bi se odločili za pravopis, ker naj bi zamešal črki <a> in <e>, bi bila to že interpretacija, ker bi poskušali pokazati na fazo v mentalnem procesu, ko je prišlo do napake.[50] Ker tega ne moremo vedeti in ker je rezultat v slovenščini obstoječa beseda, ki otežuje razumevanje, je bila izbrana oznaka napake besedišča. Enako je storila K. Rozman (2006: 59) pri nerazlikovanju med grafemoma <š> in <ž> (*jušno Ameriko*). Beseda z neustreznim samoglasnikom v končnici pa je uvrščena k napakam oblike, npr. *pošiljajo brezvezne pisme*.

Po končanem pouku moram iti na postajo (**avtobusko**), kjer čakam približno petnajst minut.

Primer 37: Napaka besedišča zaradi neustrezne črke v besedi

Internet pomeni tudi denar; **ravni** ljudje, ki ne morejo si kupiti računalnik, ali ki nočejo iti preveč po internetu zaradi telefonskega računa, bojo imeli manj možnosti kot drugi.

Primer 38: Napaka zaradi uporabe podobno zveneče besede

5.3.5.1.2 *Pisanje skupaj ali narazen*

Drugi podtip napak zapisa seže v bolj sofisticirano pisno produkcijo. Gre za tipično pravopisni problem zapisovanja besed skupaj in narazen, ki se ocenjuje tudi na izpitih iz znanja slovenščine (Ferbežar 2007c: 32). Beseda, ki bi morala biti napisana skupaj, je napisana narazen, ali pa sta skupaj napisani besedi, ki bi morali biti napisani narazen. V strogo formalnem smislu bi napake prvega tipa spadale k večbesednim enotam oziroma k napakam strukture, zaradi preglednosti pa so vse združene v isti kategoriji. Tovrstne napake v okviru pravopisnih napak opisuje tudi K. Rozman. V njenem korpusu je neločevanje mej med besedami izjemno redko, saj je takih napak le 3 % (2006: 61).

Neustrezno so v PiKUST-u narazen največkrat pisane tvorjenke, pri katerih tuji govorci sestavine dojemajo kot ločene besede (*prvi krat*), skupaj pa povezave nepolnopomenskih besed, npr. členkov ali predlogov, s polnopomenskimi (*naprimer, nimaram*). Pri označevanju tega podtipa napak v PiKUST-u so bila zaradi uporabnosti kršena nekatera od prej opisanih načel. Prvo se nanaša na ročno pripisane oblikoslovne oznake. Če sta bili besedi neustrezno pisani skupaj, napaka ni dobila oblikoslovne oznake, saj pri dveh besednih

50 Elektronska korespondenca z Natašo Pirih Svetina, 2. novembra 2006.

vrstah to ne bi bilo smiselno. Drugače je bilo pri povezavi nikalnega členka in glagola (*nemorem sprostiti*), ki je bila označena kot glagol. Med napakami pisanja skupaj in narazen v PiKUST-u je namreč dobra desetina tovrstnih napačnih povezav, prav pisanje skupaj pa kaže, da se tvorci pri tem pomensko osredotočijo predvsem na glagol. Druga izjema so bile besede, neustrezno napisane narazen (*več krat*), ki so dobile oblikoslovne oznake, četudi je to kršenje načela, da je treba označevati oblike, ki se pojavljajo v besedilu, in ne rekonstrukcij. Stopnja označevalčeve interpretacije pri tovrstnih napakah je tako rekoč zanemarljiva, koristi, ki jih da oblikoslovna oznaka, pa tolikšne, da sem se odločila za izjemo. Pri avtomatskem označevanju večje količine korpusnega gradiva bi bilo to seveda drugače, prav v tovrstnih primerih pa bi bilo smiselno razmišljati o uvedbi posebne oblikoslovne kategorije.

Nekoliko poenostavljeno je med napake pisanja skupaj ali narazen vključena napačna raba oziroma izpust vezaja (*neonaziste, 17. letnega brata*). Čeprav se nanaša na ločilo, je najpogosteje povezana z zapisovanjem skupaj ali narazen, zato taka rešitev poenostavi uporabo korpusa. Podobno je vključen primer *i.t.d.* namesto *itd.*, kjer gre prav tako za odvečno ločilo. Kompleksnejšim napakam okrajšav, kot je *in p.d.* namesto *ipd.*, podtip ni določen. Težko se je odločiti za enega, za veščega uporabnika pa naj bi bila dovolj povedna zgolj kombinacija napake zapisa in oblikoslovne oznake za okrajšavo. V PiKUST-u se pojavi samo štirikrat in bo verjetno ostala obvladljiva tudi v večjem korpusu. V Šolarju pa je med napake tega tipa vključen tudi neustrezen zapis pik v datumu (Rozman, Stritar, Kosem 2010: 17).

5.3.5.1.3 *Mala in velika začetnica*

Tretji podtip napak zapisa, neustrezna raba male (primer 39) oziroma velike začetnice (primer 40), je še bolj pravopisen in razmeroma malo vpliva na razumljivost besedila. V PiKUST-u so redke, saj jih je manj kot odstotek.

Ali bo **zemlja** kot planet dovolj velika za vse?
Primer 39: Napaka zaradi neustrezne rabe male začetnice

Mislim da so tisti poklici tudi zelo zanimivi ker lahko delaš za Evropsko **Komisijo** ali za Organizacijo združenih narodov.
Primer 40: Napaka zaradi neustrezne rabe velike začetnice

Če je tvorec pisal samo z velikimi črkami, je to v PiKUST-u ohranjeno brez posebnih oznak, čeprav so izpisi konkordanc iz takih besedil zaradi navajenosti na male črke nekoliko manj pregledni.

5.3.5.1.4 *Ločilo*

Pri šolskem pouku slovenščine je raba ločil zelo poudarjena, ne nazadnje se v korpusu Šolar na to nanaša skoraj 44 % vseh učiteljskih popravkov (Arhar idr. 2001: 41). Na tečajih za tujce pa je le redko predmet posebne obravnave, tako da posebej na začetni stopnji jezikovne zmožnosti tvorci pravopisu ne posvečajo pozornosti (Rozman 2006: 81). Po programu Slovenščina za tujce morajo govorci na osnovni ravni znati uporabljati končna ločila, na srednji deloma tudi vejico, šele na visoki ravni pa na splošno velja, da »pri pisanju upoštevajo pravopisna pravila« (Ferbežar, Pirih Svetina 2002: 15). Tudi na izpitih iz znanja slovenščine to ne vpliva bistveno na doseženo oceno. Točka za ločila se odbije samo, če je besedilo povsem brez ločil (Ferbežar 2007c: 32). V učnem KUST-u napake ločil zato niso bile označene. Toda ločila so pomembna za razumljivost. Tovrstne napake so pri tujih govorcih zelo pogoste, pri srbohrvaško govorečih predstavljajo skoraj 41 % vseh napak (Balažic Bulc 2004: 80). Zaradi njihove pogostnosti, razumljivosti popravljene različice besedil in morebitnih raziskav tega področja so v PiKUST-u s posebnim podtipom označena vsa odvečna, manjkajoča ali neustrezna ločila; skupaj predstavljajo drugi najpogostejši tip napak z 21,7-odstotnim deležem v korpusu. Oblikoslovne oznake nimajo, od spremembe pričakovane oblike pa je odvisna navedba popravka. Če ločilo manjka (primer 41) ali ni ustrezno (primer 42), je naveden popravek, kadar je ločilo odvečno, pa ne (primer 43), saj že sama nenavedba popravka uporabniku pove, za kaj gre.

Nikoli ne vemo**<sic type="Z" desc="LOC" corr=","></sic>** ali so krave bile okužene ali pa ne

Primer 41: Napaka zaradi manjkajoče vejice

Kaj lahko naredimo, da <sic type="S" desc="IZP" corr="se "></sic>bodo učenci dobro počutili, ko študirajo**<sic type="Z" desc="LOC" corr="?">.</sic>**

Primer 42: Napaka zaradi neustreznega končnega ločila

ljudje, ki <sic type="S" desc="BR" corr="si ne morejo">ne morejo si</sic> kupiti <sic type="O" msd="S" corr="računalnika">računalnik</sic>**<sic type="Z" desc="LOC">,</sic>** ali ki nočejo <sic type="B" desc="OBST" corr="deskati">iti</sic> preveč po internetu zaradi telefonskega računa, bojo imeli manj možnosti kot drugi.

Primer 43: Napaka zaradi odvečnega ločila

V Šolarju, ki je zaradi gradiva domačih govorcev in beleženja učiteljskih popravkov bolj usmerjen na pravopisne probleme, so k napakam ločil uvrščeni tudi (ne)stičnost njihovega zapisa, nakazana z (ne)rabo presledka, popravki oblike narekovajev, ki so jih uporabljali učenci, in popravki deljenja besed (Rozman, Stritar, Kosem 2010: 18–19).

5.3.5.2 Besedišče

Napake besedišča so povezane s pomenom, pa tudi z obliko besed. Zajemajo uporabo besede, ki je v danem kontekstu slovnično, semantično ali pragmatično neustrezna, oziroma v slovenščini neobstoječe besede, neustrezno tvorbo besed ali uporabo neustrezne besedne vrste. Do napake pride zaradi izbora med podobnimi besedami iz spomina, iz slovarja, iz pogovornega jezika, zaradi neustreznega tvorjenja besed ali zaradi dobesednega prevajanja, pa tudi zaradi zamešanja pomenskih odnosov: uporabe nadpomenke namesto podpomenke ali obratno, uporabe manj ustrezne sopomenke ali skoraj sopomenke (James 1998: 151). V. Požgaj Hadži in I. Ferbežar (2001: 63) med leksikalne napake uvrščata besede in besedne zveze iz prvega jezika in dobesedne prevode struktur iz prvega jezika, K. Rozman (2006: 59) pa napake zaradi nenatančnega znanja ali neznanja besede, zaradi vplivov iz prvega jezika ali pogovorne slovenščine, zaradi neustrezne izbire iz slovarja itn. V anketi med slovenisti (tabela 45) so bile k tovrstnim napakam prepričljivo uvrščene raba neustreznega pridevnika oz. številnika (5. primer) in predloga (7. primer), ki so ga sicer označevali tudi kot napako skladnje (3. primer).

Št.	Primer	Jezikovna ravnina (v %)					
		Pravopis	Besedišče	Oblikoslovje	Besedni red*	Skladnja	Drugo
3	Slikal sem Grad na sončnem zahodu.	14	38	31	2	28	
5	Ker je sijala polna luna, nisem rabil prvih luči.		84			2	13
7	Na prvem vtisu je tema filma politična.		59	15		20	6

* Besedni red ni jezikovna ravnina, vendar je bil ponujen kot samostojna možnost, ker so take napake pogoste.

Tabela 45: Odgovori iz ankete med slovenisti o primerih napak besedišča

Učeči se, pa tudi domači govorci napake besedišča ocenjujejo kot resnejše ali vsaj opaznejše od npr. oblikoslovnih napak (James 1998: 145). Na razumevanje vplivajo različno

in včasih je težko oceniti, ali je beseda rabljena ustrezno ali ne. V primeru 44 ni težko razumeti, da gre pri *internetski knjigarni* za spletno oziroma internetno knjigarno, medtem ko je interpretacija glagola *ugajati se* v primeru 45 bolj prepuščena ugibanju, zato popravek v PiKUST-u ni naveden. Kjer je raba še sprejemljiva, napaka ni označena. Razumljivost in s tem povezano ustreznost izbrane besede bi bilo smiselno preverjati pri tujih govorcih, sama pa sem si pomagala s podatki iz korpusa FIDA, saj v času označevanja FidaPLUS in Gigafida še nista bili dostopni.

Na primer, navadne knjigarne so omejene s tem, da morajo imeti javni prostor za knjige na policah, medtem ko **internetska** knjigarna, kot je na primer amazon.com, ta problem nima, tako, da lahko ima na zalogi več naslovov.

Primer 44: Napaka besedišča z nesporno interpretacijo

Seveda vsa nova sredstva (televizija, mobil, internet) lahko uplivajo na način misljenja, lahko ga spreminjajo. Ampak to je človeški last da se spreminja in se **ugaja** s svojim časom.

Primer 45: Napaka besedišča s težavno interpretacijo

Napake besedišča se v PiKUST-u delijo na podtipa obstoječih in neobstoječih besed. Pri prvih gre za slovenske besede, ki so uporabljene v neustreznem kontekstu in so torej predvsem semantično ali pragmatično neustrezne, neobstoječe pa v slovenščini sploh ne obstajajo in jih govorci ustvarijo na novo ali prenesejo iz drugih jezikov. Ločevanje je sicer lahko problematično, denimo pri pogovornih besedah, vendar koristno, saj pri analizi napak besedišča zmanjša obseg potrebnega ročnega pregledovanja konkordanc. V Šolarju tega ločevanja ni bilo (prim. Rozman, Stabej, Kosem 2010: 19–20), saj bi bilo na gradivu domačih govorcev manj smiselno.

Vsaka napaka besedišča je v PiKUST-u dobila oblikoslovno oznako. Najpogostejše napake besedišča so v tabeli 46 za večjo preglednost združene glede na besedno vrsto.

Napaka besedišča	**Primer**
Neustrezen glagol	Življenje je zabavno, če **veš** živeti!
Neobstoječ glagol	Dopoldne sem **lopatil** sneg
Neustrezen glagolski vid	Potem sem pa prvi krat **letala** in sem videla da
Neustrezen samostalnik	pri plačevanju na internetu moramo dati kode bančne **karte**
Neobstoječ samostalnik	nova sredstva (televizija, **mobil**, internet)
Neustrezen pridevnik	bolj normalen **delaven** čas
Neobstoječ pridevnik	takšni četrti so vedno bili bolj ali manj **nasilijoč** in

Napaka besedišča	Primer
Neustrezen zaimek	Če **komur** pa niso všeč paralelni svetovi, naj vzame klasično
Neobstoječ zaimek	proti alkoholu in jemljenju **bilokakšnih** mamil
Neustrezen prislov	lahko jih poslišamo **domov** na radiu
Neobstoječ prislov	ker **jedino** tako bomo
Neustrezen/neobstoječ številnik	**drigi** del družbe je zelo optimističen
Neustrezen predlog	nasilje **v** televiziji, v sodobni glasbi
Neobstoječ predlog	**Črez** internet imaš možnost dobiti vse podatke
Neustrezen veznik	Vse to je dobro, **sicer** bi rada videla kako kaj danes s prijateljstvom
Neobstoječ veznik	Kako **i** kdaj takšno obnašanje preprečiti?
Neustrezna besedna vrsta	Bom **praznik** prvi rojstni dan
Neustrezno rabljen frazem	pomankanje časa, ki je **črna ovca** najinih problemov

Tabela 46: Pregled napak besedišča v PiKUST-u

Tanka meja med napakami zapisa in besedišča je bila že omenjena. Na meji med obema so tudi besede, v katerih je neustrezna le ena črka (*virtuelne, notarnjost, svojo mentoliteto, gitaro*). Lahko bi argumentirali, da gre za napake zapisa, toda kjer presega zapisovalne težave, navedene v poglavju 5.3.5.1.1, je bilo izbrano besedišče. V prvih poskusih razdelitve napak in v učnem KUST-u pa je bila poleg kategorije besedišče še kategorija besedotvorje, toda izkazalo se je, da je v praksi težko ločevati med obema. Besedotvorje je del besedoslovja, ki je tradicionalno nauk o »besedi na sploh: kaj je, katere vrste je /…/, kako se piše in izgovarja, katere pomene ima (semaziologija), katerega izvora je (etimologija), kako jo delamo (besedotvorje), kateri zvrsti pripada, kakšno stilno vrednost ima (stilistika), v kakšnih medsebojnih zvezah so posamezne besede« (Toporišič 1992: 8). Besedotvorje kaže, »kako se delajo nove besede (tvorjenke) in kako tvorjenke razumemo« (navedeno delo: 8), in je definirana morfemizacija določene besedne zveze govorne podstave, ki temelji na pretvorbeni predvidljivosti skladenjskih sestavin besedne zveze v morfeme in medmorfemska razmerja (Vidovič Muha 1983: 45). Ločiti ga je treba od oblikotvorja, ki kot veja oblikoslovja preučuje »del t. i. slovnične ravnine jezika, tj. oblike (ne pa tudi skladnje besed)« (Toporišič 1992: 152), oziroma je »tisti del slovnice, ki slovarskim enotam (besedam) določa vrstne, oblikovne in funkcijske značilnosti« (Toporišič 1991: 192). Deli se na oblikotvorje, oblikospreminjevalnost in besednovrstnost. Oblikotvorje se nanaša na tvorbo določne oblike pridevnikov, njihovo stopnjevanje, spoljenje, tvorbo glagolskih oblik in vidskih dvojic, medtem ko je

oblikospreminjevalnost povezana s tvorjenjem sklona, osebe, števila, določnosti in včasih tudi spola (Toporišič 1992: 153).

Ločevanje med napakami besedja, besedotvorja, oblikotvorja, oblikospreminjevalnosti itn. bi bilo za praktične korpusne potrebe teoretično zahtevno in prespecifično. Če oblikotvorje in oblikospreminjevalnost združimo v oblikoslovje oziroma napake oblike, ostaneta besedišče in besedotvorje. Slednje je v praksi poučevanja slovenščine za tujce omejeno, saj se v njegovem okviru večinoma obravnava tvorjenje glagolnika oziroma samostalnika iz glagola, prislova iz kakovostnega pridevnika, svojilnih in vrstnih pridevnikov, vrstilnih številnikov, deležnikov na *-n/-t, -č*. Izrazi za moške in ženske, pomanjševalnice in ljubkovalna imena, abstraktni samostalniki na *-ost*, skupna imena ter samostalniki, ki poimenujejo prostor, so še bolj na ravni opazovanja[51] – učenci se jih naučijo prepoznati in sklepati o njihovem pomenu, redko pa se od njih pričakuje, da jih bodo znali samostojno tvoriti. Še bolj omejena je zahteva po besedotvorni zmožnosti na izpitih iz znanja slovenščine za tujce, kjer je v glavnem vezana na tvorbo vrstnih pridevnikov iz samostalnikov (prim. Ferbežar idr. 2004: 110). Vprašanje, v kolikšni meri tuji govorci v navedenih primerih tvorijo oblike, ki pogosto niso popolnoma predvidljive, koliko pa gre za njihovo naučeno, slovarsko zmožnost, bi zahtevalo posebno raziskavo.

Št.	Primer	Jezikovna ravnina (v %)							
		Pravopis	Besedišče	Besedotvorje	Oblikoslovje	Besedni red	Skladnja	Izpust*	Drugo
2	Tam je najdla svoj mir.	14	26	7	50				2
6	Med vikendom sem veliko berala.	18	20	4	52		2	2	
8	Ob ponedeljkih poučevam nemščino.	2	10	19	63		5		
11	V ponedeljek sem poučevala francoskoščino.	11	43	40	5				
12	Moral je iti na arhitekturo fakulteto.	2	20	25	7	9	4	9	
18	Danes se zelo veliko pogovarja o zdravi prehrani.		82				18		

* Izpust bi težko opisali kot jezikovno ravnino, je pa specifičen tip napake, ki ga je težje vključiti med ostale ravnine.

Tabela 47: Odgovori iz ankete med slovenisti o primerih napak besedišča/besedotvorja

51 Elektronska korespondenca z Andrejo Markovič, 30. oktobra 2006.

Tudi v rezultatih ankete med slovenisti se je besedotvorje prekrivalo tako z besedišèem kot z oblikoslovjem (tabela 47). Neustrezno tvorjena pretekla deležnika glagola *najti najdla* (2. primer) in *brati berala* (6. primer) sta bila večinoma uvrščena v oblikoslovje, podobno je bilo s tvorbo sedanjika (8. primer). Besedotvorje pri nobenem od odgovorov ni bilo prepričljiva izbira, še najbolj pri zvezi *arhitektura fakulteta* (12. primer) in pri tvorjenki *francoskoščina* (11. primer), ki je bila sicer pogosteje uvrščena k napakam besedišča.

Kategorija besedotvorja se je torej pokazala kot problematična in odvečna, zato so vse napake na ravni besede združene v napakah besedišča, ki tradicionalno pomeni »besedje določenega besedila ali tvorca besedila /…/, besede danega jezika« (Toporišič 1992: 7). Vanj so anketirani nesporno uvrstili 5. in 18. primer, kjer sta rabljena neustrezni številnik oziroma pridevnik in glagol. Še vedno pa se prekrivata besedišče in oblikoslovje, zato sem se pri označevanju držala praktičnega načela, da gre za napako besedišča, kadar je problematična osnova besede, in oblikoslovja, kadar gre za končnice. Sklanjanje pridevnika po samostalniški sklanjatvi bi lahko umestili k oblikoskladnji (Pirih Svetina 2005: 79), toda v primeru 46 je težava v osnovi besede, zato je to napaka besedišča. Besedišče in oblikoslovje se srečata pri težavah glagolskega vida, ki so sicer lahko samo oblikoslovne, kadar pa gre za znotrajleksemske morfeme, ki vključujejo neustrezno tvorbo oblike, besede ali celo drugačen pomen, so že na meji oblikotvorja[52] (*vzemati* nam. *jemati*; *pričakajo* nam. *pričakujejo*; *grajevati* nam. *graditi*; *poslišati* nam. *poslušati*). Odločila sem se za kategorijo besedišča. Po analogiji so k napakam besedišča uvrščene tudi vse neustrezne glagolske in druge predpone (primer 47).

Internet je narejen za cel svet, tako da se dogajajo včasih čudovite iskušnje. Na primer, kdor želi klepetati z **neznancimi** tujci ima zdaj možnosti.

Primer 46: Napaka besedišča zaradi neustrezne osnove besede

NA NEKAKŠEN NAČIN, POMOČ KI JO DOBIVAM OD SVETOVNIH ZNAMK PRI KONSTRUIRANJU LASTNEGA IDENTITETA ME OSVOBAJA OD TEŽKEGA DELA **PROMIŠLJEVANJA** IN ZGUBLJANJA ČASA NA VPRAŠANJA KATERA V ŽIVLJENJU SPLOH NISO POMEMBNA.

Primer 47: Napaka besedišča zaradi neustrezne glagolske predpone

52 Elektronska korespondenca z Natašo Pirih Svetina, 2. novembra 2006.

Raba neustrezne vrste zaimka je označena kot napaka besedišča in ne oblike (tabela 48). V učnem KUST-u je bila neustrezna oseba svojilnega zaimka označena kot napaka oblike (*sem čistila moje stanovanje*), toda pri drugih vrstah samostalniških in pridevniških zaimkov se pomen tako spremeni, da so vse tovrstne napake dobile oznako besedišča. Uporabnik se mora zavedati, da kombinacija oznak besedišča in zaimka pomeni specifično, manj »slovarsko« in bolj »slovnično« problematiko.

Vrsta zaimka	**Primer**
Osebni zaimek	smo v enem hipu pozabili nori svet okoli **nas**. vsak otrok gre v šolo ker hoče, ker **ji** je všeč, ampak
Svojilni zaimek	Še zmeraj se spomnim **moj** prvi dan v šolo. je mislela da je alkohol **njezin** najboljši prijatelj,
Vprašalni zaimek	Beseda homo je nastala od besede humus **kaj** pomeni zemljo.
Oziralni zaimek	ali sploh lahko kdo vzame čas ali še **kar** drugega tako da vsi ljudje lahko dobijo nekaj **ki** jih zanima.
Kazalni zaimek	Kako izbrati ustrezno zdravo hrano? **Tisto** vprašanje je v zadnjem času pogosto.
Nedoločni in mnogostni zaimek	nisam moral biti dobro oblečen da se predstavim **nekateri** punci ali **marsikomu**.

Tabela 48: Primeri napak besedišča zaradi rabe neustreznega zaimka

Raba nenaslonske namesto naslonske oblike zaimka ni označena kot napaka, zato pa je označena brezpredložna raba oziralnega zaimka *kateri*, ki po teži verjetno ni primerljiva z večino drugih, saj se pojavlja tudi pri domačih govorcih slovenščine (prim. Rozman, Stritar, Kosem 2010: 20). Če popravek tovrstne napake zahteva spremembo celotne strukture, je označevanje z gnezdenjem oznak v PiKUST-u precej zapleteno (primer 48). V Šolarju je bila uporabljena preprostejša in elegantnejša rešitev brez dodatne oznake za napako strukture (primer 49).

```
del so</sic> življenja, <sic type="S" desc="STRUKT" corr="ki se mu"><sic type="B" desc="OBST" msd="Z">kateremu</sic> se</sic> težko izognemo
```

Primer 48: Označevanje neustreznih oziralnih zaimkov v PiKUST-u

```
Ljubezen ni predmet<u1 tip="Z" podtip="LOC"><p1>,</p1></u1> <u1 tip="B">katerega<p1>ki ga</p1></u1> lahko opišemo, se ga dotaknemo, je več kot to.
```

Primer 49: Označevanje neustreznih oziralnih zaimkov v Šolarju

Neustrezni števniki se v PiKUST-u niso pojavili. Izjema je bil *en* v pomenu 'neki' v primeru *s eni prijatelici*, ki je bil za-

radi neštevniške funkcije označen kot pridevnik in bi bil v pogovornem jeziku celo sprejemljiv.

Raba neustrezne besedne vrste, npr. prislova namesto predloga (primer 50), bi lahko bila napaka besedišča, oblike ali celo skladnje, ker je tvorec izbral sintagmatsko neustrezno rešitev. K. Rozman (2006: 51) jo je označila kot morfološko napako, v anketi pa so slovenisti tovrstni primer uvrščali tako k besedišču kot k besedotvorju in oblikoslovju (tabela 49). Ne glede na oznako bodo morebitne raziskave, osredotočene na ta tip napak, zahtevale ročno pregledovanje korpusnega gradiva. Sama sem sledila večinskemu mnenju iz ankete in se odločila za besedišče. Ker oznake temeljijo na izvirni in ne na pričakovani obliki, je tudi besedna vrsta v PiKUST-u določena glede na izvirno obliko.

Prej avtomobilov je bilo težko potovati

Primer 50: Napaka besedišča zaradi rabe neustrezne besedne vrste

Št.	Primer	Jezikovna ravnina (v %)			
		Besedišče	Besedotvorje	Oblikoslovje	Skladnja
15	**Jutri bom praznik rojstni dan.**	50 %	20 %	25 %	5 %

Tabela 49: Odgovori iz ankete med slovenisti o primeru napake zaradi rabe neustrezne besedne vrste

Temu je podobna raba *ni* namesto *ne* (primer 51) ali obratno (primer 52), ki bi lahko bila napaka zapisa, besedišča ali oblike. Ker imata besedi različno obliko oziroma zahtevata drugačno strukturo, sta označeni kot napaka besedišča. Besedna vrsta je označena glede na zapisano obliko kot členek *ne* ali glagol *ni*, ne pa glede na funkcijo pričakovane oblike v stavku. Podobno je z rabo *se* ali *si* kot prostega morfema (primer 53), ki ne dobi oblikoslovne oznake, medtem ko je raba *se/si* kot povratnega zaimka napaka oblike.

On je uničuje **ni** samo sebe, ampak istočasno tudi povzroča veliko težav svojim blizkim (kot so starši, otroke, mož ali žena, in tako naprej).

Primer 51: Napaka besedišča zaradi rabe *ni* namesto *ne*

Tudi ne rečem da alkohola **ne** treba piti.

Primer 52: Napaka besedišča zaradi rabe *ne* namesto *ni*

kaj **si** tiči meni, rada ispijem kakšen kozarec vina.

Primer 53: Napaka besedišča zaradi rabe *si* namesto *se*

Napake besedišča lahko zajemajo tudi večbesedne enote, in sicer neustrezno rabljene frazeme oziroma stalne besedne zveze in nekatere druge idiomatske strukture. V primeru 54 bi bilo nesmiselno zvezi *grešni kozel* dati oznako napake skladnje, ker gre za slovarski problem, še manj smotrno pa bi bilo označevati vsako besedo posebej. Podobno velja za konektorje (primer 55). Če pričakovana ustrezna zveza zahteva drugačno skladenjsko vgraditev ali če nima enobesednega popravka, v PiKUST-u povzroči sekundarne oz. povezane skladenjske napake (primer 56).

Iz tistih primerov vidimo da je človek osebno, tisti ki je **grešni kozel** tega da nima veliko prostega časa.
Primer 54: Napaka besedišča z neustrezno rabljenim frazemom

To sem menil vse **do kdaj** nisem prišel na en mednarodni tabor v Novem mestu.
Primer 55: Napaka besedišča z neustreznim konektorjem

Strokovnjaki pravijo<sic type="Z" desc="LOC" corr=","></sic> da bi <sic type="O" desc="SEK" msd="G" corr="bila">bil</sic> <sic type="O" desc="SEK" msd="P" corr="zadostna">zadosten</sic> **<sic type="B" desc="OBST" corr="desetina">deseti del</sic>** tega
Primer 56: Sekundarna napaka skladnje zaradi napake besedišča

Pogojno so napake besedišča tudi pogovorne besede ali neknjižne kolokacije, ki so v referenčnem korpusu redke (primer 57). Ker je norma PiKUST-a knjižna slovenščina, te oblike niso sprejemljive, uporabniki pa morajo oceniti, kdaj so tovrstne napake predvsem stilistične in kdaj otežujejo razumevanje oziroma kakšno težo imajo. Podobno so ravnali v norveškem ASK-u, v katerem se v nekaterih odvisnih stavkih v povezavi s prislovi pojavlja pogovorni besedni red, ki je v knjižni norveščini zaznamovan. Dilemo so rešili tako, da so označili vse tovrstne oblike, tudi če so nezaznamovane, in se zanesli na usposobljenost uporabnikov za presojo veljavnosti oznak (Tenfjord 2006).

Rabite le dobro družbo, da bi se imeli **fajn**.
Primer 57: Raba pogovornega izraza v PiKUST-u

Povezanih napak besedišča, ki bi nastale kot posledica popravka druge napake, v PiKUST-u ni bilo, kar seveda ne pomeni, da niso možne. Zato pa napake besedišča pogosto povzročijo druge povezane napake. V primeru 58 bi morali

predlog *skozi* zamenjati s *preko*, zato bi morali popraviti tudi sklon samostalnika *internet*, kar je označeno kot povezana napaka.

> Lahko pošiljamo pogodbe poslovnemu partnerju **skozi internet**, lahko se pogovarjamo **skozi internet**, cenejše kot bi ga poklicali po telefonu (na primer v Ameriki).

Primer 58: Sekundarna napaka samostalnika zaradi neustreznega predloga

5.3.5.3 Oblika

Pri tretji kategoriji napak normi ne ustreza oblika pregibnih besed (James 1998: 154). Nastanejo z napačnim pregibanjem besed oziroma neustrezno obliko besed v povedih (Pirih Svetina 2005: 82): glede na sobesedilo je samostalnik ali pridevnik v neustreznem sklonu, glagol, samostalnik ali pridevnik v napačnem številu, pridevnik je neustrezno stopnjevan, glagol ima neustrezno izražen spol ali glagolsko osebo, glagolski čas je glede na sobesedilo neustrezno izražen, napake pa lahko nastanejo tudi »zaradi spremembe glagolskega naklona ali neupoštevanja hotenjske modifikacije (ob rabi naklonskih glagolov ali naklonskih izrazov)« (navedeno delo: 65). Sem spadajo tudi neobstoječe, neparadigemske oblike besed.

V. Požgaj Hadži in I. Ferbežar (2001: 61–63) sta med morfosintaktične napake uvrstili rabo sklonov iz prvega jezika, množine namesto dvojine, rodilnika s številkoma tri in štiri, sedanjiških končnic za tretjo osebo množine iz prvega jezika in napačen besedni red. T. Balažic Bulc (2004: 81–87) je 8 % napak pripisala sklanjatvenim vzorcem samostalnikov in pridevnikov, še 5 % pa dvojinskim vzorcem. K. Rozman (2006: 48–58) je kot morfološke napake, ki jih je bilo v njenem korpusu 51 %, opredelila posledice inteferenc, neupoštevanja pravil za izpad polglasnika, nepoznavanja pregibnostnega vzorca, kombiniranja dveh možnosti, nepoznavanja pravil za ujemanje zaimka ob samostalniku in druge.

Napake oblike so brez podtipov. Na prvi pogled se jih zdi smiselno klasificirati glede na spol, sklon, število ipd., a so poskusi pokazali, da se je nemogoče izogniti dvoumnostim in subjektivnosti. V primeru 59 se zdi najverjetneje, da gre za neustrezen spol samostalnika *medij*, toda razlog se lahko skriva tudi v neustreznem sklonu ali številu. Čim bi tovrstne napake dobile eno oznako, bi bila analiza otežena, saj bi bile druge možnosti interpretacije avtomatsko izključene. Poleg tega bi bilo označevanje izjemno zamudno in nesmiselno za

razmeroma neobsežno gradivo, ki ga je mogoče ročno pregledati. Zato tudi ni ločevanja med oblikotvornimi napakami, povezanimi s tvorbo določne oblike pridevnikov, njihovim stopnjevanjem, spoljenjem, tvorbo glagolskih oblik in vidskih dvojic, in oblikospreminjevalnimi napakami, ki se nanašajo na tvorjenje sklona, osebe, števila, določnosti in včasih tudi spola (Toporišič 1992: 153). Napake oblike imajo samo oblikoslovne oznake, ki lahko zadostujejo za učinkovito analizo (prim. Rozman 2006: 48).

Veliko stvari slišimo o internetu; da je **svobodna medija**, da prinese znanost domov in da življenje postane lažje za uporabci.

Primer 59: Napaka oblike

V PiKUST-u je napak oblike 1219 oziroma slaba četrtina (tabela 50).

Napaka	**Primer**
Sklon samostalnika	Sedaj sem iz **Gorici** vozim z **avto** do Kopra dva krat
Spol samostalnika	Ampak **vprašanja** je: kako je to mogoče imeti
Število samostalnika	DRŽAVE MED SEBOJ RAZMENJAJO BLAGA, **DENARJE**
Sklon pridevnika	nimam samo **nemška** državljanstva, ampak tudi **švicarska**
Spol pridevnika	Ampak to je **človeški** last da se spreminja
Število pridevnika	Zdi se mi da tisti kritizirajo tehniške napredke so **črnogled** in tudi
Stopnjevanje pridevnika	Takšna razvada le je uvod v še **hudejšo** navado kot je droga.
(Ne)določna oblika pridevnika	Imeli so **celi** kup otrok, ker ni bilo televizije in podobnih reči.
Število glagola	Potem sem pa začnela studij in mi **so bili** tako všeč,
Oseba glagola	ko je učiteljica začela govoriti se je vse **spremenil**
Tvorba sedanjika	ter se revni ljudje **počutajo** neumni in bolj revni.
Tvorba preteklega deležnika	V soboto in nedeljo sem šla na sprehod, sem **berala**,
Glagolski čas ali naklon	Ampak se trudim, da **neham** uživati alkohol (niti kozarca).
Glagolski vid	že dolgo nekateri **predvidejo**
Nedoločnik/namenilnik	Danes evropski očeti in mami hodejo **delati** v neki podjetju,
Oblika zaimka	10 dni živiš z 37 mladim osebami in tudi z **njima** poslušaš isto glasbo,
Oblika prislova	**avtomatski** se zmanjšuje zainteresiranost mladih

Tabela 50: Pregled napak oblike v PiKUST-u

Tipična napaka oblike je neustrezna končnica samostalnika zaradi napačnega sklona (primer 60) ali končnica, ki v

slovenščini ne obstaja (primer 61). To je potrdila anketa med slovenisti, v kateri je bil neustrezni sklon precej enoglasno opredeljen kot oblikoslovna napaka (tabela 51).

Drugi stvar je da vidimo danes več **agresivnost** in s tem nasilja med **ženskih**, nekaj ki **me** zdi zelo nevarno.

Primer 60: Napaka oblike zaradi neustreznega sklona

Med časom mnogo so so postali vojaki, drugi pa borili s pestmi in **nogmi** – ali s **orožmi**.

Primer 61: Napaka oblike zaradi uporabe v slovenščini neobstoječe končnice

Št.	Primer	Jezikovna ravnina (v %)		
		Pravopis	Oblikoslovje	Skladnja
4	**Šel sem nazaj v Ljubljani.**	2 %	91 %	7 %

Tabela 51: Odgovori iz ankete med slovenisti o primeru napake zaradi rabe napačnega sklona

Napake oblike so tudi polglasnik, ki v odvisnih sklonih ne izpade ali se realizira z neslovenskim zapisom (*dedeki, brataci*), neustrezna množina samostalnika *otrok* (*otroki in mladina*) in podobne napake pri drugih samostalnikih (*strokovnjaci*). Podobno je pri pridevnikih, kjer so neustrezne končnice slovenske (*v lepim pesmah*) ali tuje (*od računalniškog »deskanja«*). Napake zaradi neustrezne rabe določne ali nedoločne oblike pridevnika so označene kot napake oblike (*Pokojnine so male*). Posamostaljeni pridevniki oziroma pridevniki v vlogi samostalnika so po zgledu korpusov FIDA, FidaPLUS oziroma Gigafida označeni kot pridevniki. Če imata pridevnik in samostalnik v besedni zvezi različni končnici, ki sta obe možni, je kot pravilen označen samostalnik, ki je jedro besedne zveze – v primeru 62 je torej popravljen pridevnik *domače* v *domačo*.

Otrok mora pisati **domače nalogo**.

Primer 62: Napaka oblike, pri kateri se končnici samostalnika in pridevnika v besedni zvezi ne ujemata

Neustrezni sedanjik pri glagolu nastane bodisi s tvorbo iz nedoločnika (*poučevam jezika, postajo agresivni*) bodisi z uporabo neustreznih končnic. V slovenščini neobstoječe oblike, kot sta *dobijem* namesto *dobim* in *jest* namesto *je*, bi lahko uvrstili k obliki, besedišču ali v nekaterih primerih celo zapisu. Kadar gre za neslovenske končnice na ustreznih osnovah, je bila izbrana oblika.

Tvorba preteklega deležnika z *-o* (*podčrtao bi*), značilna za učeče se z južnoslovanskim prvim jezikom, je bila v drugih klasifikacijah označena kot fonetična napaka (Požgaj Hadži, Ferbežar 2001: 62; Balažic Bulc 2004: 81). Ker je pri pisnih besedilih težko govoriti o fonetiki, ker se v nenormativnem zapisu jezika domačih govorcev končnica preteklega deležnika redko realizira z *-o*[53] in ker gre dejansko za oblikotvorje, sem se sama odločila za napako oblike, tudi zaradi analogije s primeri deležnikov, ki imajo neustrezno tvorjene osnove (*bo roman umrel*, *sem pa začnela*), pri katerih gre za neustrezno tvorbo oblike iz prve osebe ednine sedanjika.[54] Anketa je potrdila, da je primere, kot sta *je najdla* in *sem berala*, smiselno umestiti k obliki (prim. tabelo 48).

Tudi izbira neustreznega glagolskega časa ali naklona je napaka oblike. Čeprav prizadene strukturo več besed, gre vendarle za večbesedno enoto znotraj iste paradigme. Napake glagolskega vida v glavnem spadajo k besedišču, toda primeri, kjer je napačna samo končnica, so uvrščeni k obliki (*se moji prijatelji zavedejo*). Ravno tako je problem oblike raba kratkega, pogovornega nedoločnika (*sem morala peljat z autobusom*) – umestitev k zapisu zaradi domnevnega zapisa govorjene oblike bi bila že interpretacija vzroka za napako.

Del težav, povezanih z zaimki, je bil omenjen pri napakah besedišča, kadar pa gre samo za neustrezni sklon oziroma končnico, je to napaka oblike. Lahko je uporabljena napačna (*všeč me je*) ali v slovenščini neobstoječa oblika (*Te šole ni bilo v mom mestu*). Raba prostega morfema *se/si* spada k besedišču, kadar pa gre za povratni zaimek, je to napaka oblike pri zaimku (*da naj bi si uprašali*).

Kot pri prejšnjih kategorijah tudi pri obliki meje niso vedno ostre, zato sem se včasih oddaljila od zgolj formalističnega razvrščanja. V primeru 63 bi lahko *deklo* označili kot napako besedišča, češ da je tvorec izbral *dekla* namesto *dekle*. Ker je *dekla* v slovenščini razmeroma redek samostalnik, vezan predvsem na besedila s starejšo oziroma kmečko tematiko,[55] se ob upoštevanju konteksta – spisa o možnostih

53 Nenaglašena končnica preteklega deležnika se knjižno lahko izgovarja kot *-u* (*Slovenski pravopis* 2001: 80). V označenem delu korpusa Šolar je pretekli deležnik moškega spola ednine šestkrat zapisan s končnico *-u* in samo enkrat s končnico *-o*. Referenčni korpusi so za preverjanje nenormativnih realizacij manj primerni. Samo za ilustracijo naj bo podatek, da se v FidiPLUS zveza *sem reko* v pomenu 'sem rekel' ne pojavi, medtem ko *sem reku* nastopa 41-krat.

54 Elektronska korespondenca z Natašo Pirih Svetina, 2. novembra 2006.

55 V Gigafidi se *dekla* pojavi 5553-krat, *dekle* pa 175.531-krat. V *SSKJ* je *dekla* definirana kot »stalno najeta ženska na kmetiji za pomoč pri gospodinjskih in kmečkih delih«.

zabave brez alkohola – zdi mnogo verjetnejša razlaga, da je tvorec neustrezno sklanjal *dekle*.

Zelo je **grdo** če vidim kakšno deklo kadit.
Primer 63: Napaka oblike, ki bi bila lahko tudi napaka besedišča

Povezane napake oblike so pogoste pri ujemanju samostalnika in pridevnika; npr. v primeru 64 je samostalnik *podatek* v neustreznem sklonu, ki ga je nato treba popraviti tudi pridevniku, torej *lastne podatke*. Lahko so odvisne od predloga; v primeru 65 je treba zamenjati predlog *čez*, posledično pa še sklon samostalniške besedne zveze. Če je treba popraviti predlog, zaradi katerega se spremeni sklon samostalnika za njim, ki že v prvotnem kontekstu nima ustrezne končnice, pa to ni označeno kot povezana napaka (primer 66). Pri glagolih so pogojene predvsem z neustrezno obliko osebka; ker je v primeru 67 treba zamenjati *kdo* s *ki*, je treba spremeniti tudi glagolsko število.

To pač pomeni, da nismo omejeni, pri iskanju informacije samo na **lastnih podatkov** proizvajalcev.
Primer 64: Povezana napaka oblike pri samostalniku in pridevniku

Noro tekanje **čez vsakdanje opravke** ki jih je »nujno« treba opraviti, nas v bistvu na koncu dneva izpolnjujejo z eno votlino brezsmiselnosti.
Primer 65: Povezana napaka oblike pri samostalniku in predlogu

Pritisk za zabavati z 'dodatno sestavino' prihaja **iz sošolce, soseščine**, morda tudi **sorodnike**, in zlasti **iz prijatelje**.
Primer 66: Napaka oblike zaradi rabe neustreznega predloga in sklona

Najstnike, **kdo pije** za prvikrat, nimajo izkušnje in ne vejo, kjer so svoje meje.
Primer 67: Povezana napaka oblike pri glagolu

5.3.5.4 Skladnja

Zadnji tip so skladenjske napake, ki se nanašajo na večbesedne enote oziroma neustrezne sintagmatske povezave med besedami. Skladnja je »del nauka o besedilu, ki raziskuje naravo /…/ medsebojnih zvez besed, zvez besednih zvez, stavkov in povedi, ali ki uči, kako se poved in njene sestavine tvorijo« (Toporišič 1992: 274). Napake na nadpovedni ravni, torej napake kohezije, diskurzne, slogovne, pragmatične napake ipd., v PiKUST-u niso označene, kot niso označene v drugih korpusih usvajanja tujega jezika, saj so zaradi

odvisnosti od širšega sobesedila za korpusno označevanje manj primerne. Izjema so slogovno neustrezne besede (prim. poglavje 5.3.5.2). V Šolarju, v katerem napake temeljijo na posegih učiteljev, ki pogosto popravljajo slogovne ali pragmatične neustreznosti, pa je to večinoma označeno kot napaka strukture (Rozman, Stritar, Kosem 2010: 23).

V prvih poskusih klasifikacij za učni KUST so se pojavljale različne oznake: skladnja, besedni red, nerazvrščene napake itn. Besedni red je sicer pogost tip napake (prim. Balažic Bulc 2004), tako da ga včasih enačijo s skladenjskimi napakami (prim. Požgaj Hadži, Ferbežar 2001; Rozman 2006), vendar skladenjske napake v PiKUST-u vključujejo še neustrezne strukture ter izpuščene in manjkajoče besede ali besedne zveze. V anketi med slovenisti so bile napake skladnje razdeljene na besedni red in skladnjo. Besedni red je bil prepričljivo izbran v 16. in 17. primeru (tabela 52), skladnja pa je največ odgovorov dobila v 9. primeru. V 20. primeru so se anketiranci pogosteje odločali za oblikoslovje.

Št.	Primer	Jezikovna ravnina (v %)							
		Pravopis	Besedišče	Besedotvorje	Oblikoslovje	Besedni red	Skladnja	Izpust	Drugo
9	Plavanje, tek in aerobika so vsi športi da želim.	3	24		5		43	19	6
16	Moja mama tudi je bila prevajalka.					96	4		
17	Zjutraj zgodaj sem vstal.					90	5		5
20	Nisem točno spomnim, kdaj sem se zbudil.		18	6	29	23	18	6	

Tabela 52: Odgovori iz ankete med slovenisti o primerih napak skladnje in besednega reda

Če del besedila ni napisan v slovenščini, je v PiKUST-u označen s splošno oznako za skladenjsko napako brez podtipa napake in predlaganega popravka (primer 68).

Jaz prihajam iz Splita ki je **capital city of Dalmatia, a province in Croatia**.

Primer 68: Napaka skladnje zaradi dela besedila, napisanega v tujem jeziku

Skladenjske napake v PiKUST-u nimajo oblikoslovnih oznak, saj se nanašajo na večbesedne enote. Če bo KUST

oblikoslovno označen, te pomanjkljivosti ne bo več, saj bo vsaka beseda imela oblikoslovno oznako. Popravki so navedeni, kjer je mogoče, v primerih z nejasnim pomenom pa jih seveda ni. Pogoste so sekundarne oziroma povezane napake, posebej pri besednem redu (primer 69) ali v primerih, kjer popravek neustrezne besede povzroči spremembo celotne strukture (primer 70).

<sic type="S" desc="SEK" corr="ne pa"><sic type="B" desc="OBST" msd="V" corr="pa">**a**</sic> **ne**</sic> <sic type="O" desc="SEK" msd="P" corr="družbenega">družbeni</sic> <sic type="O" msd="S" corr="statusa">status</sic>

Primer 69: Sekundarna napaka besednega reda zaradi popravka neustreznega veznika

ker sem <sic type="S" desc="SEK">**se**</sic> <sic type="B" desc="OBST" msd="G" corr="odraščala">zbudila</sic> s francoščino in nemščino.

Primer 70: Sekundarna napaka odvečne besede zaradi popravka neustreznega glagola

5.3.5.4.1 *Besedni red*

Besedni red je bil v prvih zasnovah posebna kategorija poleg skladenjskih napak. V. Požgaj Hadži in I. Ferbežar (2001: 62) ga umeščata k morfosintaktičnim, K. Rozman (2006: 58) pa k skladenjskim napakam, zato je tudi v PiKUST-u podtip napak skladnje. Nanaša se na zaporedje besed v besedni zvezi, npr. pridevniški prilastek pred odnosnico, zaporedje izhodišča, prehoda in jedra v povedi, zaporedje naslonk v naslonskem nizu, mesto naslonskega niza v povedi (Toporišič 1992: 7). Napake besednega reda zajemajo zaporedje naslonk znotraj naslonskega niza (*mnogi, ki niso imeli priložnosti spoznati ga*), mesto naslonskega niza v povedi (*mogoče res je*), zaporedje besed znotraj samostalniške besedne zveze (*ker imam slovenščine tečaj*) in zaporedje sestavin stavčnih členov (*mednarodna politika me zanima zelo*). Povezane so predvsem s pravili, ki uravnavajo stalno stavo in vključujejo pridevniški in nepridevniški prilastek, povedkovo določilo, namenilnik ob glagolih premikanja, deležnik na *-l* glagola *biti*, glagola *iti* in *imeti* v sedanjiku, izpridevniško prislovno določilo načina, predmeta v tretjem in četrtem sklonu, če stojita skupaj, členek, predlog in veznik ter naslonski niz (Kranjc 2004: 155, Pogačnik 2001: 8–9). Predkorpusne raziskave kažejo, da naj bi tujci delali največ napak pri stavi naslonk in glagolov *imeti* in *iti* (Pogačnik 2001: 50).

Napake zaradi členitve po aktualnosti so bolj stilistične in v PiKUST-u niso označene. Pri njihovi obravnavi je treba imeti v mislih okoliščine nastanka besedila, predvsem tvorca. »Od njega je namreč odvisno, katero informacijo bo videl kot pomembnejšo in jo bo zato postavil na konec oz. v remo in katera bo z njegovega stališča manj pomembna in ji bo zato namenjeno mesto v temi, to je na začetku« (Kranjc 2004: 150). Teh informacij označevalec korpusa nima, členitev po aktualnosti pa, podobno kot organiziranost diskurza nasploh, tudi v večini učbenikov slovenščine kot tujega jezika ni posebej obravnavana (Pogačnik 2001: 35).

Začetek oznake napake besednega reda je pred besedo, ki mora prva zamenjati položaj, konec pa za zadnjo tako besedo (primer 71).

<sic type="B" desc="OBST" msd="V" corr="Ko">Kadar</sic> sem bila stara 7 ali osem let, **<sic type="S" desc="BR" corr="so bile moje sanje">moje sanje so bile</sic>** potovati in videti veliko sveta, torej sem mislila, da bi bilo idealno delati kot stevardesa.

Primer 71: Oznaka napake besednega reda v PiKUST-u

5.3.5.4.2 *Izpuščeni podatki*

Kadar v besedilu manjka beseda ali besedna zveza, je to skladenjska napaka s podtipom izpust, kar je poleg zamenjave in vstavitve ena od osnovnih kategorij v razdelitvah po površinski strukturi (prim. poglavje 5.2.1). T. Balažic Bulc (2004: 85) jo umešča k napakam besednega reda. V anketi med slovenisti je bil izrazit izpust zaznan v primeru z izpuščenim morfemom *se* (tabela 53).

Št.	Primer	Jezikovna ravnina (v %)			
		Besedišče	Oblikoslovje	Skladnja	Izpust
14	**Včeraj sem zbudila zelo pozno in sem šla v službo.**	17 %	17 %	11 %	55 %

Tabela 53: Odgovori iz ankete med slovenisti o primeru napake izpusta

Izpust je v PiKUST-u označen pred besedo, pred katero manjka oblika (primer 72). Če je izpuščena ena sama beseda, ima oblikoslovno oznako, vedno pa je naveden tudi popravek.

<sic type="S" desc="BR" corr="Nikoli nisem">Nisem nikoli</sic> mislila<sic type="Z" desc="LOC" corr=","></sic> da bi **<sic type="S" desc="IZP" msd="G"**

corr="bil"></sic> pevka poklic zame, ampak glasba mi je zelo všeč in sprejeli so me na glasbeni akademiji.

Primer 72: Oznaka napake izpusta v PiKUST-u

V določenih primerih tovrstne napake mejijo na napake oblike. V zvezi *vožnja avtobusom* namesto *vožnja z avtobusom* bi glede manjkajočega predloga lahko trdili, da gre za jezikovni prenos oziroma napačno sklepanje o brezpredložnem orodniku za izražanje sredstva dejanja in s tem za oblikoslovno napako. Ker je klasifikacija PiKUST-a formalna, taka sklepanja niso zaželena in je označen izpust. Podobno velja za izpuščeni pomožni glagol pri tvorbi preteklega ali prihodnjega časa (*vse izgledalo lepo*), ki ga N. Pirih Svetina (2005: 65) uvršča k oblikoslovnim napakam.

Večkrat so izpuščeni deležnik glagola *biti* (primer 73), prosti morfem oziroma povratni zaimek *se/si*, osebni zaimek (primer 74), členek *ne* pri dvojnem zanikanju (primer 75) ali druge besede (v primeru 76 je težko navesti zanesljiv popravek). Izpuščeni predlogi lahko povzročijo povezano napako sklona samostalniške besedne zveze, denimo v primeru *Potreba prostora in naravnih snovi je vzrok za uničevanje*, kjer bi bila popravljena oblika *Potreba po prostoru in naravnih snoveh je vzrok za uničevanje.*

Kar premislimo si kaj bi **/bilo/**,[56] če bi naša delovna doba končala pri 60 in povprečna življenska doba bi bila okoli 100 let?!

Primer 73: Napaka izpusta zaradi izpuščenega deležnika glagola *biti*

To je res, da s časom naše življenje je hitrejše zato, kjer **/se/** število stvari, ki **/jih/** moramo narditi, poveča.

Primer 74: Napaki izpusta zaradi izpuščenega osebnega zaimka in morfema *se*

Tudi res je, da marsikdo nikoli **/ne/** učijo nič, in nimajo lastnega nadzorja.

Primer 75: Napaka izpusta zaradi izpuščenega členka *ne*

Šola je najpomembnejši stvar v življenju in mislim da vsak otrok na svetu bi moral iti **/v šolo/**.

Primer 76: Napaka izpusta zaradi drugih izpuščenih besed

5.3.5.4.3 *Odvečni podatki*

Izpustu nasprotna težava, kadar je beseda ali besedna zveza v besedilu odveč, je označena s podtipom odvečni podatki oziroma vstavitev. Tudi tu gre za način odklona od pričakovane oblike in tudi tu ima napaka oblikoslovno oznako,

56 Manjkajoči deli besedila so označeni z /…/.

kadar je neustrezna samo ena beseda. Popravljena oblika ni navedena, saj je iz tipa napake jasno, da bi bilo pravilno brez označenega dela.

V anketi med slovenisti ni bilo tipičnega primera vstavitve, vendar je v zvezi s 5. primerom *nisem rabil prvih lučі* 13 % anketiranih menilo, da je pridevnik nepotreben (prim. tabelo 45).

Vstavljeni oziroma odvečni so lahko glagol *biti* (primer 77), prosti morfem oziroma zaimek *se/si* (primer 78) in različni zaimki: oziralni (primer 79), osebni (primer 80), kazalni. Ker je norma za učeče se večinoma slovenski knjižni jezik, je raba kazalnega zaimka *ta*, ki je sprejemljiva v pogovornem jeziku, označena kot neustrezna (primer 81). Pojavljajo se še vstavljeni vezniki (primer 82), členki in druge odvečne besede. Kot pri izpuščenih predlogih lahko tudi odvečni povzročijo sekundarno oziroma povezano napako samostalnika (*že trajajo poskuse po kloniranju človeka*).

Mislim da **je** denar ni tako pomemben, ker vidim da je veliko ljudi ki zasluža dosti denarja ni res veselo.

Primer 77: Odvečni glagol *biti*

Ko **se** pogledamo na to kakšen je svet, poln nasilja, agresije, statusnih simbolov.

Primer 78: Odvečni *se/si*

Nekateri ljudje pravijo da se človek pozna po oblačilih, oziroma **ki** se strinjajo s pravico »Obleka naredi človeka«.

Primer 79: Odvečni oziralni zaimek

Vsak glasbenik, posebno če **on** je koncertant, mora vaditi sam.

Primer 80: Odvečni osebni zaimek

Če si v trendu, se pravi, če si oblečen po **ta** zadnji modi, če imaš najnovejši tip avtomobila, če vsako leto zamenjaš kos pohištva misliš, da si glavni v mestu, da te vsi občudujejo, se motiš, ker s tim nisi dokazal, da nekaj veljaš.

Primer 81: Odvečni kazalni zaimek

Mislim da če imamo zelo veselo družbo, ki nam je všeč, **da** se moremo zelo lepo zabaviti.

Primer 82: Odvečni veznik

V Šolarju med tovrstne napake spadajo tudi primeri, v katerih učitelj prečrta vsebinsko neustrezen del besedila (Rozman, Stritar, Kosem 2010: 23), za KUST pa je to manj relevantno.

5.3.5.4.4 *Neustrezna struktura*

Podtip neustrezna struktura se nanaša na neustrezno sintagmatsko povezavo znotraj večbesedne enote, besedne zveze ali celega stavka. Gre za dve vrsti napak: za niz besed z razumljivim pomenom, ki je neustrezno formuliran, ali za sicer ustrezno oblikovan sklop z nejasnim pomenom.

Do napak lahko pride zaradi prenosa iz tvorčevega prvega jezika (primer 83). V. Požgaj Hadži in I. Ferbežar (2001: 62) sta dobesedne prevode besednih zvez iz prvega jezika opredelili kot leksikalne napake, vendar so te v tu predstavljeni razdelitvi načeloma omejene na enobesedne enote.

Menim, da **časa vedno ima**, če smo dobro organizirani.
Primer 83: Napaka strukture zaradi prenosa iz tvorčevega prvega jezika

Ustreznost popravkov je bila preverjena v referenčnem korpusu, kadar pa popravek zahteva popolno parafrazo ali kadar ga ni mogoče predvideti (primer 84), ni naveden. Primer 85 bi lahko razvezali npr. v *Prva značilnost interneta je, da je povsod pri nas*, kar pa je brez tvorčeve dodatne informacije že interpretacija. Podobno velja za primer 86.

Jaz menim da je dobro biti lepo oblečen in **držati do sebe**, ampak je napačna misel da je vse v oblačilih in dobro poznanih lokalih itd.
Primer 84: Napaka strukture brez jasnega popravka

Prvič z internetom je, da je povsod pri nas.
Primer 85: Napaka strukture brez jasnega popravka

Prej bi tudi spil kakšno pivo ampak vedno **v svoji meji**.
Primer 86: Napaka strukture brez jasnega popravka

Neustrezna struktura je pogosta oznaka pri dvojnih napakah, recimo pri neustrezno rabljenem naklonskem glagolu ali oziralnem zaimku (primer 87). V Šolarju so bile menjave oziralnih zaimkov *ki* in *kateri* poenostavljeno označene kot napake besedišča ne glede na širšo spremembo strukture (Rozman, Stritar, Kosem 2010: 20).

del so</sic> življenja, <sic type="S" desc="STRUKT" corr="ki se mu"><sic type="B" desc="OBST" msd="Z">**kateremu**</sic> se</sic> težko izognemo
Primer 87: Neustrezna struktura znotraj dvojne napake v PiKUST-u

6 UPORABA KORPUSOV USVAJANJA TUJEGA JEZIKA

Ključna dimenzija vsakega korpusa, ki jo je treba opredeliti še pred začetkom gradnje, je njegov namen, ki je povezan tudi s potencialnimi uporabniki. Načeloma želijo jezikoslovci s pomočjo korpusov za potrebe leksikologije, leksikografije, besediloslovja, prevodoslovja, sociolingvistike ali stilistike (Gorjanc, Fišer 2010: 16–17) čim natančneje opisati jezik. Če se omejimo na korpuse usvajanja tujega jezika, so namenjeni predvsem spoznavanju načel, po katerih poteka usvajanje. Podatki iz njih so kvantitativno utemeljeni, preverljivi, primerljivi in načeloma bližje naravni produkciji kot npr. podatki, pridobljeni z eksperimenti. Omogočajo raziskovanje različnih vidikov vmesnega jezika brez vnaprej postavljenih hipotez, preučevanje pragmatike in diskurza ter upoštevanje vpliva nejezikovnih dejavnikov na usvajanje.

Uporabni pa niso brez zadržkov; v njih na primer ni podatkov o receptivni zmožnosti učečih se ali o tem, kako sprejemljiva se jim zdi njihova lastna produkcija. Če se oblika v korpusu ne pojavi, to še ne pomeni, da je ne poznajo; »v jeziku je lahko tudi tisto, česar v korpusu ni« (Stabej 1998: 98). Domiselna uporaba korpusnih metod omogoča raznolike in zelo specifične raziskave (prim. Vintar 2003: 79), toda korpusi – nespecializirani ali usvajanja tujega jezika – niso čudežni odgovor na vsa vprašanja, pač pa zgolj v različne namene uporaben vir, ki ga lahko raziskovalci in učitelji jezika dodajo v svoj nabor jezikovnih virov (Granger 2004: 129). Nekatere pojave je še vedno treba raziskovati eksperimentalno bodisi zato, ker korpusi ne nudijo dovolj podatkov, bodisi zaradi

specifičnosti področja: elipso v jeziku je najprimerneje preučevati z introspekcijo (Čermák 2009), motiviranost ali jezikovno nadarjenost s kombinacijo korpusne in eksperimentalne metode (Nesselhauf 2004: 131–132), razvoj vmesnega jezika na individualni ravni pa z longitudinalnimi raziskavami posameznih učečih se (Granger 2004: 125). Po drugačnih metodah je treba poseči tudi pri raziskovanju recepcijske zmožnosti učečih se ali njihove strategije izogibanja, ki ga lahko nakazuje nepojavljanje določenega tipa napake (Brown 1980: 167).

Tovrstni, v teoretično raziskovanje jezika usmerjeni cilji so navedeni pri kar nekaj obstoječih korpusih usvajanja, med njimi AKCES/CzeSL, CLC, EAGLE, ICLFI, ICLE, JPU, LCLE in TCLE (Boyd 2008: 1901; Hana idr. 2010: 114; Dagneaux idr. 2001; Kennedy 1998: 42; Nicholls 2003: 576; Pravec 2002: 84–85; Shih 2000: 92, 96). Multimedijski korpus MAELC je primeren za preučevanje pedagoškega procesa (Reder idr. 2003: 547). Longitudinalni švedski ASU omogoča raziskave razvoja vmesnega jezika na individualni ravni, pa tudi med posamezniki, med tujimi in domačimi govorci, longitudinalno, med pisanjem in govorjenjem itn. (spletna stran). Namen korpusa IBLC je analizirati rabo jezika, medkulturno akomodacijo in usvajanje žanrov (Connor idr. 2002: 189), brazilski MLC pa naj bi nudil gradivo za ustreznejše poučevanje portugalsko govorečih študentov in za longitudinalno spremljanje razvoja jezikovne zmožnosti (Tagnin 2003: 941–943). Na podlagi HKCSE so raziskovali različne oblike diskurza, nedoločno izražanje in implicitnost (Cheng, Warren 1999: 6–16). Nekateri korpusi so bili zasnovani za ožje opredeljene raziskave, WriCLE je npr. namenjen raziskavam besednega reda (Mendikoetxea idr. 2009). Cilj projekta ESFSLD je bil ugotoviti, kako socialni in psihološki dejavniki vplivajo na usvajanje, opisati strukturalne in časovne značilnosti procesa ter preučiti rabo ciljnega jezika. Ugotovili so, da gre pri vseh ciljnih jezikih razvoj počasi in postopno od samostalniških prek nedoločniških do zapletenejših izjav (Allwood idr. 1983).

Enako, če ne še bolj intenzivno kot teoretičnim raziskavam so korpusi usvajanja tujega jezika namenjeni praktičnim aplikacijam, torej razvoju jezikovnih priročnikov ter metod za učenje in poučevanje jezika. Pregled literature o obstoječih korpusih kaže, da je prav pedagoška uporabnost, predvsem diagnosticiranje in odpravljanje najpogostejših napak in težav tujih govorcev, najpomembnejša graditeljem številnih korpusov, med drugimi USE, LPELC, EAGLE, SST,

HKUST in TeleNex. PolyU naj bi pomagal diplomantom izboljšati njihovo pisanje (Lin 1999), podobno PELE, ki je bil osnova nalogam, uporabljenim pri pouku (Uzar 1998), in AKCES/CzeSL (Hana idr. 2010: 114). Cilj projekta CrossCheck, katerega del je korpus SVANTE, je razviti sistem za avtomatsko prepoznavanje slovničnih napak tujih govorcev švedščine (Volodina 2008: 104, 205; Bigert idr. 2005). V slovaropisju in pri izdelavi drugih učnih gradiv naj bi se med drugim uporabljala LCLE in Lancom. Po mnenju avtorjev korpusa CLC so v učbenikih in pripravljalnem gradivu za izpit iz angleščine zaradi označenih napak bolj upoštevane dejanske težave študentov, uporaba korpusa pa naj bi povečevala konsistentnost ocenjevanja izpitov v različnih državah in obdobjih. Razvijajo tudi računalniški program za primerjavo novih spisov s starimi in s tem ustreznejše ocenjevanje. Cambridge International Corpus je sicer osnova za mnoge publikacije od učbenikov do slovarjev (Nicholls 2003: 576; Pravec 2002: 94; spletna stran).[57] Namen estonskega EIC je uporabiti obstoječe jezikovne tehnologije za polavtomatsko označevanje napak, pri čemer izkoriščajo oblikoslovni označevalnik, skladenjski razčlenjevalnik in izvirno taksonomijo napak, na njegovi osnovi pa želijo izdelati elektronski slovar in slovar učečih se (spletna stran).

Za analizo korpusov so na voljo različna programska orodja in uporabniški vmesniki. Prve generacije konkordančnikov so omogočale zgolj osnovne kvantitativne analize neoznačenih besedilnih zbirk v okoljih DOS in Unix, danes pa je eno najbolj uporabljanih orodij program WordSmith Tools,[58] prirejen za delo z obsežnimi korpusi. Podpira iskanje po označenih besedilih in izračun ključnih besed, omogoča izdelavo liste besed, izračunava besedne skupe, kolokacije, v analizo pa je možno vključiti seznam praznih besed (angl. *stoplist*).[59] Z WordSmithom analizirajo gradivo v korpusih ICLE, FRIDA, USE, LPELC, Poly U (Pravec 2002: 104; Lin 1999). V korpusu TeleNex poleg tega uporabljajo še MicroConcord in Monoconc Pro, na spletni strani je na voljo konkordančnik PatternFinder. Korpus ASK temelji na orodju Corpus Workbench. Del analize korpusa USE poteka s programom Qwick, ki ga je na univerzi v Birminghamu za analize konkordanc

57 Prim. http://www.cambridge.org/si/elt/catalogue/subject/products/item2701617/Cambridge-International-Corpus-View-all-products/?site_locale=sl_SI (avgust 2011).

58 Prim. http://www.lexically.net/wordsmith/ (september 2011).

59 Slovničnopomenske besede, ki jih pri nekaterih analizah izključimo, da ne bi motile pogostnostnih analiz (Gorjanc 2005: 88).

in kolokacij razvil Oliver Mason. Milton je HKUST analiziral z Microsoft Accessom, Excelom in programom askSam, s katerim je med drugim računal razmerje med lemami in pojavnicami, ki primerja besedišče kitajskih študentov na izpitih in v besedilih, ki niso bila napisana pod časovnim pritiskom (Pravec 2002: 105). Seveda pa mnogi korpusi – med korpusi usvajanja npr. CLC, TELC in LCLE – razvijajo svoja orodja za analizo podatkov.

Pri izbiri orodja za analizo je poleg finančnega in drugih praktičnih vidikov treba upoštevati tudi uporabnika. Čeprav v dosegljivi literaturi ni posebej izpostavljeno, je iz ciljev, ki jih navajajo graditelji korpusov usvajanja tujega jezika, jasno, da so v glavnem namenjeni profesionalnim uporabnikom, torej raziskovalcem, sestavljavcem učnih gradiv in učiteljem, manj pa tistim, ki se jezik šele učijo. Tudi KUST naj bi uporabljali predvsem učitelji slovenščine kot tujega jezika in raziskovalci na tem področju, toda učeči se in tuji govorci ne bi smeli biti vnaprej izključeni iz uporabe. Predvsem učeči se na višjih ravneh jezikovne zmožnosti lahko v korpusu usvajanja tujega jezika med drugim preverjajo standardnost svoje jezikovne produkcije, saj opisi jezika za domače govorce, ki prikazujejo le prvojezikovne nestandardne rešitve, zanje ne zadostujejo. Zato mora biti uporabniški vmesnik dovolj prijazen tudi za manj vešče uporabnike.

Ob vsem tem pa je treba opozoriti, da samo avtomatska analiza seveda ne more biti dovolj, saj zgolj podatki »ne morejo biti smiselno interpretirani brez interpretativnih zmožnosti raziskovalcev. Pri uporabi korpusnih podatkov je dragocena tako intuicija govorcev določenega jezika kot jezikoslovcev« (Gorjanc 2003: 20). Napredek v jezikoslovju je mogoč le z upoštevanjem najrazličnejših virov od besedil do introspekcije in eksperimentov. Dokazi, ki spremljajo znanstvene razprave, lahko izhajajo iz intimnega govorčevega občutka ali besedilnega korpusa, razlika pa je seveda v zaupanju dokazom, v njihovi zanesljivosti in veljavnosti (Kennedy 1998: 7–8).

6.1 Teoretična uporaba podatkov

K analizi korpusov je mogoče pristopiti na dva načina glede na razmerje med hipotezami in korpusnimi podatki. Popolni korpusni pristop (angl. *corpus-driven*) oziroma pristop od spodaj navzgor predvideva »uporabo korpusa neodvisno od

uveljavljenih jezikoslovnih interpretacij« (Gorjanc 2005: 24). Izhodišče so surovi korpusni podatki brez sklicevanja na obstoječe jezikovne modele, česar se je pri konkretnem delu zaradi obremenjenosti analiz z obstoječimi jezikoslovnimi koncepti težko strogo držati. Pri delnem korpusnem pristopu (angl. *corpus-based*) oziroma pristopu od zgoraj navzdol raziskovalec korpusne podatke aplicira na teorijo: izhaja iz teoretičnih osnov, ki jih preveri in ilustrira s korpusi. Najuspešnejša pa je kombinacija obeh pristopov oziroma nenehno prehajanje med teorijo in podatki.

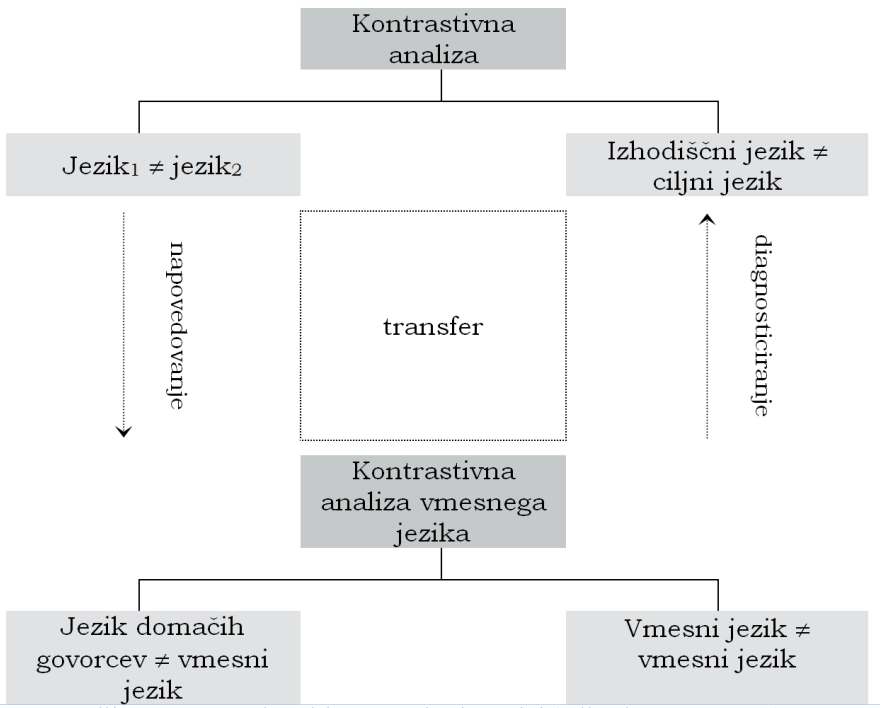
Raziskave, ki temeljijo na podatkih iz korpusov usvajanja tujega jezika, so v glavnem povezane s primerjalno ali kontrastivno analizo jezikov. Manj razvita je primerjalna analiza napak, ki napak za razliko od tradicionalne analize napak ne obravnava ločeno od njihovega konteksta (Granger 2004: 133) in je zato pogosto vključena v širše zasnovane študije vmesnega jezika.

Kontrastivna analiza jezikov izhaja iz behaviorizma in strukturalizma in primerja različne vmesne jezike ali jezik domačih govorcev z jezikom učečih se. V prvotni obliki je kot glavno oviro pri usvajanju tujega jezika pojmovala posameznikov prvi jezik, zato naj bi bilo mogoče na podlagi znanstveno ugotovljenih razlik med jezikoma napovedovati težave učečih se (Brown 1980: 148). V strogi različici, ki se osredotoča na napovedovanje problemov, kontrastivna analiza ni bila sprejeta in je pritegnila le ožji krog jezikoslovcev. Ti so ugotovili, da naj bi kot posledica transfera iz prvega jezika nastalo razmeroma malo napak – po eni od raziskav le 8–23 % slovničnih napak pri odraslih (Dulay, Burt, Krashen 1982: 103). Poleg tega so »raziskave pokazale, da učenci nimajo težav le na področju kontrastivno ugotovljenih razlik, marveč tudi pri drugih prvinah tujega jezika, in da osredotočanje na razlike ne povečuje uspešnosti in učinkovitosti pouka« (Kovačič 2003: 216). Uspešnejša je bila ohlapnejša različica kontrastivne analize, ki priznava vpliv jezikovnega prenosa in prek njega pojasnjuje napake, vendar se ne omejuje na jezik kot sistem v ožjem smislu (Brown 1980: 157), temveč v veliki meri preučuje »tudi konvencije jezikovne rabe v dveh jezikovnih skupnostih« (Kovačič 2003: 217): širše okoliščine rabe, pragmatične dejavnike, besedilne vzorce, rabo ustrezne jezikovne zvrsti glede na temo, sogovornika, okoliščine itn.

Danes je najpriljubljenejši integrirani kontrastivni model (slika 28), ki združuje kontrastivno analizo dveh jezikov s kontrastivno analizo vmesnih jezikov oziroma vmesni jezik

primerja z jezikom domačih govorcev (Gilquin 2000: 100). Vir podatkov za domače govorce so referenčni ali kontrolni korpusi, korpusi usvajanja tujega jezika pa nudijo informacije o ujemanju pravil vmesnega jezika učečih se s pravili ciljnega jezika in pomagajo odkriti tuje prvine (Granger 2004: 134). Pri primerjavah so možne različne kombinacije (Tono 2004: 54):

- primerjava med različnimi stopnjami vmesnega jezika ali med vmesnimi jeziki govorcev z različnimi prvimi jeziki;
- primerjava med korpusom usvajanja in korpusom ciljnega jezika, npr. pedagoškim korpusom;
- primerjava korpusa ciljnega in izhodiščnega jezika, s katero prepoznamo morebitne vzroke za jezikovni transfer;
- primerjava med korpusom izhodiščnega jezika in korpusom usvajanja, s katero prepoznamo s tem povezane napake ali prepogosto oz. preredko rabo;
- kombinacija primerjav vmesnega, izhodiščnega in ciljnega jezika, s katero prepoznamo kompleksna razmerja med korpusi na podlagi napak učečih se in njihove prepogoste oz. preredke rabe.



Slika 28: Integrirani kontrastivni model (Gilquin 2000: 100)

Tipična kontrastivna analiza izhaja iz empiričnih podatkov o tem, kako tuji govorci uporabljajo določeno besedo ali strukturo v primerjavi z domačimi. Ponavadi temelji na seznamih konkordanc ali listah besed, redkeje pa na spremenljivkah, kot so leksikalna raznolikost, leksikalna gostota ali leksikalna sofisticiranost. Leksikalna raznolikost je določena z razmerjem med različnicami in pojavnicami. Visoko razmerje nakazuje, da je v besedilu uporabljenih veliko različnih besed in se redko ponavljajo. Raziskave so pokazale, da pri tem ni velikih razlik med domačimi govorci in učečimi se. Z leksikalno gostoto se meri delež polnopomenskih besed, leksikalno sofisticiranost pa nekateri programi merijo tako, da primerjajo vnaprej določene sezname besed, npr. tisoč najpogostejših angleških besed, s seznami iz besedil. Drugi način deluje na podlagi dolžine besed in temelji na predpostavki, da so daljše besede bolj sofisticirane (Meunier 1998: 32–34).

Rezultate povežejo s podatki iz prvega jezika učečih se, na podlagi ugotovitev pa lahko predlagajo, kako bi v praksi učitelji in učeči se napake ozavestili ter se jim uspešneje izogibali (Tan 2005: 127). Integrirani kontrastivni model pokaže univerzalne značilnosti, skupne vsem učečim se, in tiste, ki so omejene na govorce z določenim jezikovnim ozadjem (Tono 2003a: 801). Na njegovi podlagi je mogoče predvidevati o podobi vmesnega jezika in o napakah ter oceniti vpliv zunajjezikovnih spremenljivk, kot so jezikovno ozadje, spol, starost, naloga, na tvorjenje (Granger 1998a: 13). Še vedno mu sicer očitajo, da razume vmesni jezik zgolj kot nepopolno različico ciljnega jezika (Granger 2004: 133) in da neupravičeno temelji na normi domačih govorcev kot cilju učečih se (Hunston 2002: 211). Toda zagovorniki integriranega kontrastivnega modela menijo, da je pri učečih se na izpopolnjevalni stopnji, na katere se ponavadi osredotočajo analize, smiselno ugotavljati, v čem se še razlikujejo od domačih govorcev (Granger 2004: 133), tako da naj bi bila to za učeče se ustrezna norma.

Pri kontrastivni analizi je pomembna primerljivost gradiva, tako glede prenosnika besedil, stopnje pismenosti tvorcev (besedila študentov je treba primerjati z besedili študentov in ne univerzitetnih profesorjev) ali jezikovne zvrsti. Če učeči se starostno spadajo v isti razred, to še ne pomeni, da je njihova jezikovna zmožnost primerljiva: japonski učenci angleščine so enako stari kot v korpus vključeni Evropejci, vendar je njihova jezikovna zmožnost bistveno nižja (Tono 2003a:

801). Upoštevati je treba tudi družbeno in kulturno identiteto govorcev, ki lahko botruje nekaterim razlikam. Tako je M. Tan (2005: 133) zavrnila tezo, da Tajci pri pisanju v angleščini samostalnik *teacher* uporabljajo brez člena zaradi transfera iz tajščine, ki nima členov. Ker pri drugih samostalnikih člene uporabljajo, gre verjetno za vpliv kulturnega ozadja in privzgojene uporabe nazivov, s katerimi izražajo spoštovanje, kakršen je tajski *ajarn* oziroma učitelj.

Na splošno so do sedaj izvedene kontrastivne raziskave še precej na začetku razvoja, saj so bolj prispevale k opisovanju kot interpretaciji podatkov o usvajanju tujega jezika. Razloga za to sta praktična: večinoma so jih izvajali korpusni jezikoslovci in ne strokovnjaki za usvajanje jezika, raziskana področja pa so morali najprej natančno opisati, preden so se sploh lahko lotili teoretičnega posploševanja (Granger 2004: 134–135). Zaenkrat so analize usmerjene na pojave, ki jih je enostavno preučevati z avtomatskim iskanjem in najočitneje pripomorejo k tujemu zvenu vmesnega jezika: besedišče, izražanje modalnosti, konektorji, družljivost besedišča oziroma kolokabilnost (navedeno delo: 132). Dosedanje analize za angleščino so ugotovile precej večji vpliv prvega jezika, kot so predvidevali, večjo podobnost z govorjenim jezikom domačih govorcev in manjše razlikovanje formalnosti situacij (Nesselhauf 2004: 135–136). Učeči se, predvsem tisti na nižji stopnji jezikovne zmožnosti, pogosteje od domačih govorcev uporabljajo najpogostejše besede, posebej pridevnike, in nimajo širokega nabora sopomenk (Gillard, Gadsby 1998: 161). Analiza domačih in japonskih govorcev angleščine v korpusu ICLE je pokazala, da domači govorci glagol *discuss* ('razpravljati') najpogosteje uporabljajo v zvezah *will be discussed, discussed in, as discussed*, Japonci pa v *discuss about, I will/would like to discuss* (Granger, Paquot 2006). V angleških esejih japonskih univerzitetnih študentov je tvorčeva prisotnost, ki se kaže skozi rabo osebnih ali svojilnih zaimkov v prvi osebi, precej višja kot v besedilih domačih govorcev, prav tako pa imajo več značilnosti govorjenega jezika (npr. izrazov *and so on, just, really*). Pri študentih drugega letnika je zaradi vpliva poučevanja že zaznaven očiten upad tovrstnih neustreznih struktur (McCrostie 2008: 103–105). Poljski govorci angleščine prepogosto uporabljajo prislove, ki pomenijo visoko stopnjo gotovosti (*probably, simply, definitely, actually, certainly, indeed, truly, undoubtedly*), in se izogibajo prislovu *perhaps* (Kaszubski 1998: 181). Cobb (2003) pa je poiskal skupne točke treh korpusno utemeljenih raziskav in

podatkov iz kanadskega korpusa QLC: prepogosto rabo najpogostejših besed, pogosto rabo omejenega števila sporazumevalnih vzorcev in precej manj formalnih elementov.

Na Slovenskem je Grad (2010: 86) primerjal izražanje osebnosti domačih in tujih govorcev angleščine. V diskurzu domačih govorcev je pogostejše in bolj direktno, saj izrazito pogosteje uporabljajo osebni zaimek *we*, slovenski govorci angleščine pa so raje manj izpostavljeni in se odločajo za svojilni zaimek *our*. Tudi v zvezi s slovenščino kot tujim jezikom so že bile opravljene delno korpusno utemeljene kontrastivne analize, omenjene v prejšnjih poglavjih (prim. poglavje 5.2.2.2). T. Balažic Bulc (2004: 86) je ocenila, da zaradi negativnega jezikovnega prenosa iz srbohrvaščine v slovenščino nastajajo napake pri zapisovanju vejice in glasovnih premen, pri sklanjatvenih vzorcih in dvojinskih oblikah, pri besednem redu, vezljivosti in besedotvorju, zaradi prenašanja vzorcev s pomočjo dobesednega prevajanja pa napake pri konektorjih in besednih zvezah. I. Petric (2004: 143) je iskala podobnosti in razlike v kohezivnosti in koherentnosti 23 besedil tujih slovansko govorečih in devetih primerljivih besedil domačih govorcev slovenščine. Na mikro ravni tuji govorci uporabljajo več za-oblik in junktivov, napake pa nastajajo tudi pri gradnji zahtevnejših stavčnih struktur in površinski organizaciji besedila. K. Rozman (2006: 93–95) je v besedilih dvanajstih nemško govorečih študentov slovenščine na začetni ravni našla največ morfoloških napak, ki jim sledijo leksikalne, fonetične, pravopisne in sintaktične napake. Večinoma nastanejo zaradi jezikovnega prenosa, interferenčne napake pa so pogostejše pri govorcih, katerih prvi jezik je hrvaščina.

N. Pirih Svetina (2005: 126–133) je v longitudinalni raziskavi analizirala napake 36 udeležencev tečaja slovenščine. Ugotovila je, da se delež napak z naraščanjem jezikovne zmožnosti zmanjšuje in da pogosteje prizadenejo glagolske kot samostalniške besedne zveze, kjer so tudi lokalnejše. Glede na pojavljanje v določenem oblikospreminjevalnem vzorcu so najpogostejše napake zaradi rabe napačnega sklona. Vzorec glagolov z identičnima korenoma nedoločniške in sedanjiške osnove je nezaznamovan, zato ga učeči se posplošujejo na ostale tipe, druge posplošitve pa so uvrščanje samostalnikov s končnico *-e* v sklanjatev srednjega spola, samostalnikov druge ženske sklanjatve med samostalnike moškega spola, sklanjanje pridevnikov s samostalniškimi končnicami ter raba pravila o izpadu polglasnika v neimenovalniških in

needninskih oblikah v primerih lastnostnih in vrstnih pridevnikov s končnico *-en*. Na podlagi tega je ocenila razvoj besedilne in jezikovne zmožnosti.

Ker so do sedaj opravljene kontrastivne analize usmerjene ali na zelo splošne pojave, kot so najpogostejše besede, ali pa na zelo specifične, npr. rabo posameznih besed, so predvsem površinske in kvantitativne, še vedno pa ni globinskih analiz, ki bi vključevale tudi več ročnega dela in iskanja. V korpusu z zelo natančno označenimi napakami bi bilo ročnega pregledovanja sicer manj, vendar bi se s tem zamudno delo samo preneslo z raziskovalca na označevalca (Nesselhauf 2004: 135–136).

6.2 Pedagoška uporaba podatkov

Korpuse so najprej uporabljali v slovaropisju, saj sta z njihovo pomočjo zajemanje in opisovanje besedišča enostavna in hitra. Izpis konkordanc pokaže frazeološko obnašanje iztočnice, različni sobesedilni vzorci pa njene pomene, rabo, idiomatiko in frazeologijo. Slovaropiscem so s konkordancami hitro dostopni avtentični primeri rabe (Mauranen 2003: 523). Statistični izračuni, oblikoskladenjske oznake ali podatki o avtorju in drugih okoliščinah nastanka ter objave besedila dajo dodatne informacije o iztočnici, njenem okolju in morebitni zvrstni, regionalni ali časovni specifičnosti (Vintar 2003: 78). Z razvojem korpusov pa se je njihova uporabnost širila tudi na področja, kot so besediloslovje, prevodoslovje, sociolingvistika, stilistika in didaktika poučevanja jezika. Že v dvajsetih letih prejšnjega stoletja so v ZDA in Veliki Britaniji pričeli šteti besede v besedilih, da bi našli najpogostejše in s tem pedagoško najuporabnejše. Namesto na izjeme so se osredotočili na najpogostejše lekseme, kar je postal osrednji princip usvajanja jezika na začetni stopnji, ki se je z razvojem korpusov še okrepil. Ugotavljali so tudi pogostnost posameznih glagolov ali nepravilnih glagolskih oblik v angleščini in jih aplicirali na pedagoško delo (Kennedy 1998: 264). Danes je pedagoška uporaba korpusov domačih govorcev že uveljavljena, saj pokažejo, kaj je v jeziku tipično v določeni situaciji ali zvrsti besedila. Pri poučevanju tujih govorcev pa je pomembno tudi poznavanje težav učečih se, za kar so najprimernejši vir korpusi usvajanja tujega jezika.

Korpusi so za jezikovni pouk primerni »kot referenčni vir za učitelja, učenca ali za izdelavo pedagoških referenčnih virov in učnih gradiv« (Hirci, Pisanski 2000: 30). Na poučevanje vplivajo:

– na ravni oblikovanja učnih načrtov;
– na ravni oblikovanja učnih gradiv in
– na ravni oblikovanja učne metodologije v razredu.

V najširšem smislu korpusi določajo učni proces tako, da spreminjajo celotni učni načrt: izbiro vsebine poučevanja, vrstni red in poudarek, ki ga učitelj da določeni vsebini (Kennedy 1998: 281). Snovalcem pouka se ni treba zanašati zgolj na njihovo lastno intuicijo o tem, katera slovnična vprašanja bi bilo treba obravnavati (Pisanski, Hirci 2000: 64). Vendar je vlogo korpusov zaenkrat težko ocenjevati in jo ločevati od ostalih dejavnikov, zato je več pozornosti v nadaljevanju namenjene učnim gradivom in učni metodologiji. Število konkretnih, na korpusih usvajanja tujega jezika utemeljenih rezultatov sicer še ni sorazmerno s številom publikacij, ki zagovarjajo njihovo uporabo v pedagoški praksi (Granger 2004: 136), vseeno pa so vedno številčnejše izboljšave pedagoškega procesa, ki temeljijo na podatkih iz korpusov usvajanja tujega jezika in informacijah domačih govorcev. Korpusi pozitivno vplivajo tudi na jezikovno testiranje, saj pomagajo oceniti tipičnost uporabljenih materialov, zagotavljati konsistentnost ocenjevanja ali razvijati naloge v testih (Hunston 2002: 205).

6.2.1 Učna gradiva

Sodobna učna gradiva pogosto temeljijo na referenčnih korpusih domačih govorcev, saj s tem bolj ustrezajo tipični jezikovni rabi. Korpusi usvajanja tujega jezika pa imajo pri njihovi izdelavi dve glavni funkciji. Posrednejši, zahtevnejši in zaenkrat manj izkoriščen je prenos teoretičnih ugotovitev o usvajanju tujega jezika na poučevanje, neposrednejše pa prepoznavanje težavnih mest za učeče se in aplikacija tega na gradiva, predvsem na **slovarje** (Nesselhauf 2004: 127).

Korpusno jezikoslovje je v slovaropisje vneslo številne spremembe predvsem s tem, da gradivo analizira sistematično in besed ne opisuje izolirano. Bolj kot za posamezno besedo gre »za medsebojni učinek med elementom besedila in kontekstom« (Teubert 2005: 118). Korpusno utemeljeni slovarji so bili prvi širše dostopni rezultati jezikoslovne uporabe korpusov (Gorjanc 2003: 25). Specializirani slovarji za učeče se, posebej iz serije Cobuild (prim. Krek 2004), so že od nekdaj vodilni pri razvoju slovaropisja (Gillard, Gadsby 2001: 159), saj prvi beležijo nove rabe ali upoštevajo spreminjanje slovničnih struktur. Prvi so kot vir podatkov o najpogostejših besedah pričeli uporabljati nespecializirane korpuse.

Z uporabo korpusov se je spremenilo oblikovanje razlag ali definicij gesla v slovarjih. Včasih so prevladovale razlage, »oblikovane po načelu enačbe in (skladenjske) izmenljivosti, npr. nalupiti – z lupljenjem priti do določene količine česa (SSKJ)« (Vintar 2003: 78). Na korpusih utemeljeni slovarji pa so vpeljali »razlage v obliki vezanih in lahko berljivih povedi /…/, npr. Če lupimo toliko časa, da imamo olupljenega dovolj, smo nekaj nalupili« (citirano po: prav tam). Tovrstne stavčne definicije so primernejše za šolskega uporabnika oz. učenca jezika in vsebujejo veliko več podatkov »v obliki, ki je uporabniku poznana iz splošne komunikacije. Z navajanjem tipičnih udeležencev na naraven način vključujejo tudi številne pragmatične informacije« (Gantar, Krek 2009: 153). Primeri rabe so avtentični stavki iz korpusa (Vintar 2003: 78), uporabljeno besedišče pa je omejeno: v Longmanovih slovarjih denimo na približno 2000 besed (Gillard, Gadsby 1998: 163). Poleg tega slovarji za učeče se navajajo še slovnične informacije o iztočnici vključno z vezljivostnimi vzorci, načrtno izpostavljajo besednozvezne enote, radodarno vključujejo zglede rabe in pazljivo navajajo morebitne stilne ali druge omejitve (Krek 2004: 4).

Danes je izdelava slovarjev kompromis med podatki domačih govorcev, ki povedo, kaj je tipično v ciljnem jeziku, in podatki iz korpusov usvajanja, ki pokažejo, katere težave so tipične za učeče se (Granger 1998a: 7). Na njihovi podlagi so v slovarje vključena opozorila o najpogostejših napakah in lažnih prijateljih (Gillard, Gadsby 1998: 160). Tako zasnovani so med drugim *Cambridge International Dictionary of English* (1995), ki temelji na korpusu CLC, ter *Longman Language Activator* (1993, 1997) in *Longman Dictionary of Contemporary English* (1995), ki temeljita na korpusu LCLE. Tovrstna opozorila in opombe (primer 88) učečim se pomagajo pri odpravljanju fosiliziranih napak, čeprav so v natisnjenih verzijah slovarjev precej omejeni. Elektronski slovarji lahko vključujejo več takih informacij in učečim se še bolj pomagajo pri težavah, povezanih s prenosom iz prvega jezika (Granger 2004: 137).

> **remind** […] to make someone remember a person, thing, or time from the past
> remind sb of sb/sth: *That perfume always reminded him of his mother.* | *This song reminds me of that time we all went to Brighton.*
> Don't say 'it reminds me her'. Say **it reminds me of her**.
> Don't use **remind** (= make you remember) when you mean **remember**.

Primer 88: Opozorilo v slovarju *Longman Essential Activator*, ki je nastalo na podlagi korpusa LCLE (Gillard, Gadsby 1998: 168)

Na podlagi podatkov iz korpusov usvajanja je učeče se lažje usmerjati k pravilnejši rabi v slovarju (Nicholls 2003: 576), poleg tega pa se uporabljajo še kot osnova za analizo velikosti besedišča učečih se na določeni stopnji jezikovne zmožnosti, primerjavo najpogostejših kolokacij domačih in tujih govorcev, analizo stopenj, na katerih besede vstopajo v govorčev besednjak, in konteksta, v katerem so prvič uporabljene (Gillard, Gadsby 1998: 170).

Na Slovenskem je precej poudarjena potreba po enojezičnem slovarju slovenščine, ki bi odražal aktualno jezikovno stanje in hkrati upošteval potrebe tujih govorcev. SSKJ je za takšnega govorca prezahteven in premalo razviden, obstoječi dvojezični slovarji pa so zaradi skopih oblikoslovnih in kontekstualnih podatkov primerni predvsem za receptivne dejavnosti, tako da učitelji pri tujih govorcih opažajo precej napak zaradi uporabe slovarja (Rozman 2004: 66). Dejstvo, da večina teh učiteljev svojim učencem uporabo SSKJ vendarle priporoča (navedeno delo: 65), je verjetno bolj kot dokaz njegove dejanske primernosti treba razumeti kot potrditev potrebe po enojezičnem slovarju slovenščine za tuje govorce. Vsebovati bi moral uporabnikom razumljive pomenske razlage in jim dajati »informacije o izgovorjavi, oblikoskladenjskih lastnostih leksemov, kolokacijah in ustreznosti rabe glede na zunajjezikovne okoliščine« (navedeno delo: 67). Na podlagi gradiva iz KUST-a bi bila vanj lahko vključena opozorila o tipičnih napakah. Podatki iz KUST-a bi lahko dopolnili tudi manjše slovarje, kot je prosto dostopni spletni slovensko-nemški in nemško-slovenski slovar Online SLO-DE-SLO,[60] ki vsebuje dobrih 7000 iztočnic iz treh starejših učbenikov za slovenščino kot drugi/tuji jezik[61] in z drugih tematskih področij (Lönneker, Rozman 2004: 56, 62). Ker naj bi uporabnike zanimale predvsem besede in fraze iz vsakdanje rabe jezika, bi ga bilo smiselno nadgraditi tako z najpogostejšimi lemami iz KUST-a kot tudi z opozorili o pogostih napakah pri posameznih iztočnicah.

Drugo področje, na katerem je mogoča aplikacija ugotovitev iz korpusov usvajanja tujega jezika, so **slovnice**. Nekatere že temeljijo na referenčnih korpusih domačih govorcev; *Longman Grammar of Spoken and Written English* (1999) ima

60 Prim. http://webapp.rrz.uni-hamburg.de/~slowenisch/ (julij 2010), http://www.icsi.berkeley.edu/~loenneke/Bilingual/slodeslo_gehen.html (september 2011).

61 Gre za učbenike *Odkrivajmo slovenščino* (Čuk, Mihelič, Vuga, Črnivec 1994), *Slovenščina na koncu jezika* (Pirih Svetina 1997) in *Slovenščina za tujce* (Jug-Kranjec 1978).

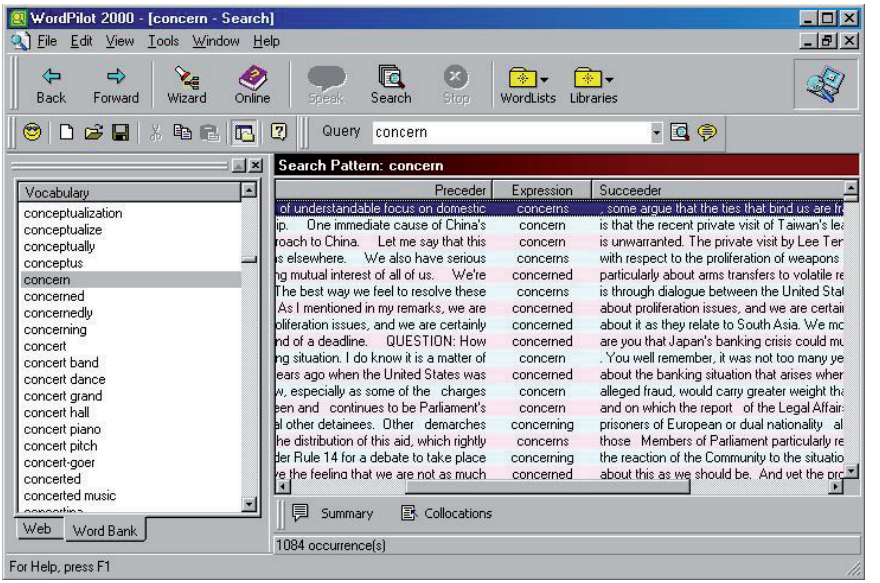
npr. za osnovo Longman Spoken and Written English Corpus. Tudi na Slovenskem obstaja potreba po »celoviti, funkcionalno zasnovani slovnici slovenskega jezika za tujce« (Pirih Svetina, Ponikvar 2006). Njen zametek predstavlja slovnica v *Sporazumevalnem pragu za slovenščino* (Ferbežar, Knez idr. 2004), idealna slovnica za učeče se pa bi morala upoštevati gradivo domačih govorcev, ki bi kazalo pogostnost pojavov, in učečih se, ki bi določalo izbiro in zaporedje tem.

Podobno kot slovnice tudi **učbeniki** zaenkrat temeljijo na korpusih domačih govorcev. Iz njih črpajo najpogostejše fraze in besede ter realne primere, tako da se izognejo zgledom rabe, ki so tuji učencem in nenaravni učiteljem (Pravec 2002: 108). Korpusi usvajanja pa na podlagi primerjave s korpusi domačih govorcev dajo podatke o prepogosti, preredki oziroma neustrezni rabi, tako da bi učbeniki lahko vključevali daljše sezname sopomenk skupaj s podatki o pogostosti, registru/stilu in opozorili o neustrezni rabi, sezname pogostih kolokacij s poudarjenimi razlikami med prvim in ciljnim jezikom ipd. (Kaszubski 1998: 183). Učeči se prek korpusov usvajanja spoznavajo tipične napake govorcev iz svojega lastnega jezikovnega okolja ali primerjajo svoj jezik z jezikom domačih govorcev (Nesselhauf 2004: 139); analiza gradiva srbohrvaških govorcev iz PiKUST-a je denimo pokazala, da bi njim prilagojene vaje za prepoznavanje in odpravljanje napak morale vključevati manj primerov neustreznih sklanjatev in glagolskega vida, več pozornosti pa bi med drugim moralo biti posvečene rabi glagolov *vedeti, znati* in *poznati* (prim. Stritar 2009: 146–149).

Eden od najobetavnejših vidikov rabe korpusov usvajanja tujega jezika, posebej tistih z označenimi napakami, so programi za **računalniško podprto učenje jezika** (angl. *CALL, computer-assisted language learning*). Kljub številnim možnostim, ki jih nudijo različne kombinacije elektronskega besedila, slikovnega, video- in avdiogradiva ter interaktivnega pristopa (Pisanski 2003: 104), je uporabnih gradiv tega tipa za jezikovni pouk malo (prim. Minn idr. 2005), posebej za slovenščino (prim. Stritar 2005, Zemljarič Miklavčič 2005, Jerman 2004).

Za interaktivno učenje angleščine je John Milton na osnovi najpogostejših napak iz korpusa HKUST in ugotovljenih leksiko-gramatičnih razlik med vmesnim in ciljnim jezikom razvil program AutoLANG. Učeči se vidijo besedilo, v katerem je v vsaki vrstici ena napaka, ki jo morajo najti in popraviti. Na voljo sta jim tudi hiperbesedilo o specifičnih napakah

in angleški slovnici nasploh ter zbirka premalo uporabljanih fraz (Milton 1998: 192). Soroden program je WordPilot (slika 29), s katerim lahko študentje med pisanjem poiščejo primere za posamezne besede ali fraze s kolokacijami v korpusih domačih govorcev (Granger 2004: 137). Nudi tudi seznam napačno rabljenih besed, nekaj malih, socialnozvrstno različnih korpusov, povezave na več slovarjev in sintetizator govora (Pravec 2002: 106; Beaugrande 2001: 23).[62]



Slika 29: Okno za iskanje konkordanc v programu WordPilot

Na francoskem korpusu FRIDA temelji projekt FreeText, katerega cilj je razvoj programa za računalniško podprto učenje francoščine na nadaljevalni in izpopolnjevalni ravni. Vključuje različne tipe nalog s področij, na katerih učeči se delajo največ napak, npr. glagolski čas in način, ujemanje, členi in predlogi. Vaje naj bi odražale tip konteksta, v katerem prihaja do napak: v nalogo za utrjevanje ujemanja preteklega deležnika z osebkom so vključili velik delež osebkov ženskega spola, ki so se v korpusu izkazali za problematične. Težave v zapisu odpravljajo z vajami, ki zahtevajo pisanje po nareku, ter posebnimi vajami za homonime in druga problematična vprašanja. Pri nalogah odprtega tipa je pomemben avtomatski sistem za prepoznavanje napak, ki na osnovi obstoječih črkovalnikov in slovničnih pregledovalnikov za fran-

62 Več informacij o obeh programih na http://home.ust.hk/~autolang/ (maj 2009).

coščino ter statistike napak iz FRIDE pregleduje uporabnikove odgovore in daje povratne informacije. Pri napakah zapisa predlaga možne rešitve, pri slovničnih napakah pa prepozna tip napake in usmerja na slovnično razlago relevantnega poglavja (Granger 2003b: 473–74). O tem, kako uspešen je v resnici ta avtomatski sistem, ni podatka.

Na osnovi korpusa TeleNex sta učiteljem angleščine na voljo pedagoška slovnica TeleGram in zbirka ocenjenih učnih gradiv TeleTeach (Pravec 2002: 87).[63] Na Tajvanu v okviru spletne strani IWiLL[64] deluje sistem za pomoč pri pisanju spisov v angleščini, uporaben za učence, učitelje in raziskovalce. Učenci lahko iščejo po spisih in odkrivajo ponavljajoče se probleme, učitelji spise ocenjujejo in pri tem izbirajo med že obstoječimi komentarji. Hkrati se avtomatsko gradi korpus angleščine kot tujega jezika, ki shranjuje oznake napak in jih uporablja pri dajanju povratnih informacij (Wible idr. 2001: 278).

6.2.2 Učna metodologija

V sodobnem poučevanju se vloga učitelja spreminja iz vira znanja v organizatorja nalog in priložnosti za usvajanje (Kennedy 1998: 281). S tega vidika na potek dela v razredu vplivajo tudi korpusi, ki omogočajo razmeroma avtonomno delo učencev, »s čimer je klasičen pristop k pouku presežen« (Pisanski 2003: 105), učenec pa prevzema del odgovornosti za posredovanje znanja (Hirci, Pisanski 2000: 27). Težišče poučevanja tujega jezika se v sodobnem času prenaša z gramatike »na leksiko, pri čemer se je zaradi drugačnega pogleda na leksikalne enote velik del gramatične problematike vključil v spoznavanje in učenje leksikalnih enot s svojo gramatično strukturo« (Gorjanc, Jurko 2004: 49). Učenje gramatike temelji na jezikovnih funkcijah, gramatičnih opisih na ravni konkretnih leksikalnih zapolnitev. Duševni leksikon je predvidljivo organiziran. Sedemdeset odstotkov vsega, kar govorec sprejme ali tvori, je v obliki fiksiranega jezikovnega elementa v spominskem slovarju, zato mu besedil ni treba vsakič znova procesirati (navedeno delo: 51). Jezik je torej organiziran predvsem okrog stalne vezljivosti besed, vendar je poučevanje to dolgo puščalo ob strani. Uporaba korpusa pa učencem omogoča, da sami odkrivajo pravila in dopolnjujejo primere iz učbenikov ali slovnic. Dejanska raba jezika

63 Prim. http://www.telenex.hku.hk/TeleNex/mainmenu.htm (maj 2009).
64 Prim. http://iwillnow.org (maj 2009).

se pogosto ne sklada s pravili iz slovnic, zato je korpus »koristen vir dodatnih primerov, ki abstraktna pravila oziroma klasične reference ilustrirajo, konkretizirajo ali pa izpostavijo njihove pomanjkljivosti« (Hirci, Pisanski 2000: 30). Korpusi sicer niso absolutno avtentični ali avtomatično realni, saj je kontekst v njih zamegljen. Čim bolj avtentične pa morajo biti vsaj dejavnosti, povezane z besedili (Mishan 2004: 219–222). V nekoliko abstraktnejšem smislu torej ni vprašanje, ali lahko korpusi predstavljajo realnost, temveč ali lahko njihova raba ustvari pogoje, v katerih so učeči se dejavni v realnem diskurzu, s čimer ga naredijo avtentičnega (navedeno delo: 226).

Nekateri menijo, da učeči se ne bi smeli dostopati do korpusnih podatkov, ker to zmanjša učiteljevo jezikoslovno avtoriteto (Aston 1997). Vendar smisel pouka ni prikazati učitelja kot vsevednega in nezmotljivega. Korpusi usvajanja učečim se nudijo boljši vpogled v njihov lastni vmesni jezik in s tem boljše zavedanje o lastni jezikovni produkciji (Pérez-Paredes 2004: 254). Učenje z odkrivanjem, kot je brskanje po seznamih konkordanc, jih dodatno motivira (Granger, Tribble 1998: 200; Nesselhauf 2004: 140), zato se uveljavlja stališče, da je raziskovanje preresno, da bi ga prepustili samo raziskovalcem (citirano po: Mishan 2004: 222).

Zaenkrat korpusov usvajanja tujega jezika pri pouku ne uporabljajo veliko, delno tudi zato, ker jih je razmeroma malo in je dostop do njih omejen. Nekateri učitelji zato sestavijo svoje male korpuse, pri čemer lahko sodelujejo tudi učenci, ki vključijo svoje tekste, besedila, ki jih že poznajo, ali celo besedila iz časopisov, televizijskih oddaj in podobno. Tako jim je vsebina bližja in bolj ustreza njihovi jezikovni zmožnosti, bolj so motivirani, naučijo pa se izdelovati in uporabljati korpus. Seveda pa je proces izdelave zamuden, označevanje v korpusu pa manj natančno (Flowerdew 2001: 364).

Delo s korpusi v razredu ponavadi poteka od spodaj navzgor, redkeje pa od zgoraj navzdol oziroma deduktivno, tako da si učenci izberejo hipotezo in jo preverijo v korpusu. Kot najbolj avtentična oblika se je uveljavilo učenje na podlagi podatkov (angl. *data-driven learning*), ki temelji na **seznamih konkordanc**. Gre za pedagoško aplikacijo raziskovalne metode, pri kateri učeči se postanejo »jezikovni detektivi« in prek avtentičnih primerov odkrivajo dejstva o ciljnem jeziku (Hunston 2002: 170). Tako se učijo kolokacij, ponavljajočih se načinov ohlapnega povezovanja besed v največkrat dvočlene besedne zveze, »ki so tipično pomensko transparentne

(pomen je razviden iz sestave pomenov komponent kolokacije) in niso pomensko motivirane« (Gorjanc, Jurko 2004: 52). Kolokacije so zanimive zaradi »besedilnih informacij, ki jih slovarji ne morejo zajeti, možnosti razvijanja samostojnega opazovanja jezika ter samokorekcije pri lastni besedilni produkciji« (navedeno delo: 54). Specifične so za posamezni jezik, učeči se pa jih zaradi napačne hipoteze o prevedljivosti celotne besedne zveze pogosto narobe uporabljajo. Za korpusno raziskovanje pri pouku so posebej primerne vezava s predlogi, napačna raba določenih besed in prepogosta ali preredka raba (Nesselhauf 2004: 140–142), pa tudi raziskave fraznih glagolov, besed s podobnim pomenom, ki jih učeči se pogosto zamenjujejo, ali izrazov za barve, hrano, dele telesa ipd., ki so pogosti v idiomatskih izrazih (Mishan 2004: 224). Manj ustrezno je raziskovanje oblik, ki se redko pojavljajo ali ki jih je težje iskati, kot sta glagolski čas in vid (Nesselhauf 2004: 140–142).

Konkretno delo v razredu se začne z izpisoma konkordanc iz korpusa domačih govorcev in korpusa usvajanja tujega jezika, ki ju učitelj da učencem, da iščejo razlike med njima. Če so manj očitne, jih usmerja z dodatnimi vprašanji. S tem ozavestijo negativne učinke jezikovnega prenosa, zaradi nevarnosti, da bi si bolj zapomnili nepravilno obliko, pa je učenje treba kombinirati s pravilnimi oblikami oziroma z nalogami za utrjevanje (Nesselhauf 2004: 140), npr. s konkordancami, v katerih manjka ključna beseda. Naloge so težje, če vključujejo različne besede in ne npr. samo enega glagola (navedeno delo: 144). Druga dejavnost v razredu, ki temelji na korpusnih podatkih in je primerna predvsem za učence na izpopolnjevalni stopnji, je sestavljanje slovarske definicije. Na podlagi seznama konkordanc sestavijo kratko definicijo besede skupaj s primeri in jo primerjajo z obstoječimi slovarji (Mishan 2004: 224).

Pri takem delu učitelj ohranja pomembno vlogo posrednika med vsebino korpusa in potrebami učencev. Na delo mora pripraviti sebe in učence ter poskrbeti, da uporabljajo ustrezne vire, iščejo na ustrezen način in podatke ustrezno interpretirajo. Včasih je konkordance bolje izpisati na papirju (Nesselhauf 2004: 143), saj se s tem prihrani čas v razredu in zmanjša strah študentov pred tehniko. Najuspešnejši pristop k delu s konkordancami naj bi se začel s papirno obliko in postopoma nadaljeval na računalniku, ker tako učenci najprej usvojijo teoretično in praktično ozadje metode (O'Keeffe, Farr 2003: 412). Treba pa je vedeti, da delo s konkordancami

ob prevelikem številu vaj postane enolično (Granger, Tribble 1998: 209). Ugotovitve je treba komentirati in ovrednotiti pragmatično sprejemljivost struktur (Flowerdew 2001: 373). Učitelji in učenci se morajo zavedati tudi, da lahko korpusi »vsebujejo nepregledana besedila, ki v nasprotju s klasičnimi referenčnimi viri niso 'očiščena napak'. /…/ V besedilih so seveda lahko netipične zveze, tipkarske napake, včasih lahko kje manjka kakšna beseda, itd.« (Hirci, Pisanski 2000: 27). Zato je tak tip učenja bolj kot za redni šolski pouk primeren za delo v manjših, ekskluzivnejših skupinah z višjo stopnjo jezikovne zmožnosti, denimo na tečajih za odrasle govorce ali na univerzi. Učenje pa je globlje, bolj celostno in posebej koristno pri oblikah, ki so ali postajajo fosilizirane (Nesselhauf 2004: 140).

Druga, manj razširjena oblika učenja na podlagi podatkov je s pomočjo pedagoško prilagojenega **modela leksikalnih funkcij**, abstraktnih zasnov, »ki imajo za različne leksikalne enote (osnovne funkcije) različne vrednosti« (Gorjanc, Jurko 2004: 56). Odnos med osnovo funkcije in njeno vrednostjo je večinoma kolokacija. Funkcija *sredstvo* bi se lahko realizirala kot *Instr(pošta) = po [pošti]; Instr(satelit) = preko [satelita]* (navedeno delo: 57). Značilnosti takega pristopa k poučevanju so imunost na jezikovno specifičnost in poljubna izbira jezikovnega para pri kontrastivnem opazovanju, »možnost uniformne obravnave novih leksikalnih enot in na ta način sprotno pletenje in utrjevanje leksikalnih mrež ob usvajanju tujega jezika, ne nazadnje pa tudi samoozaveščanje o obstoju kolokacijskih omejitev« (navedeno delo: 60).

Ob opisanih pedagoških metodah pri pouku nikakor ni smiselno opustiti že obstoječih pripomočkov in se lotevati novega izpeljevanja jezikovnih dejstev na podlagi uporabe korpusov. Mnogo učinkovitejša je kombinacija obojega (Hirci, Pisanski 2000: 30). Potreba po jezikovnem sistemu ostaja in tudi tu lahko korpusno jezikoslovje daje koristne informacije. Zaradi dela s korpusom se učitelj manj zanaša na svoj lastni jezikovni čut, ki je večkrat zavajajoč, in bolj upošteva dejanske jezikovne podatke (Kennedy 1998: 264). Korpus postane referenčni vir, korpusni podatki pa dopolnjujejo učiteljevo kompetenco, na primer pri vprašanjih vezljivosti (Pisanski, Hirci 2000: 68). Toda v praksi se kot omejitvi šolskega dela s korpusi pojavljata še dve omejitvi, tehnološka opremljenost izobraževalnih institucij in računalniško znanje. Nekateri učitelji jezikov sploh ne vedo, da korpusi obstajajo, ali jih ne znajo uporabljati. Bojijo se, »da gre za zelo

kompleksne in tehnološko zahtevne pripomočke, za katere je treba imeti ogromno računalniškega znanja, ki ga sami ne premorejo« (navedeno delo: 27). Na srečo se stalno izboljšujeta tako opremljenost šol kot jezikovnotehnološko znanje učiteljev. Šele ko bosta dosegla primerno raven, se bodo učitelji lahko posvečali usposabljanju učencev za uporabo korpusov in prek tega za učenje jezika.

7 ZAKLJUČEK

V spremenjenih družbenopolitičnih okoliščinah po slovenski osamosvojitvi in vključitvi v Evropsko unijo se slovenščina kot tuji jezik iz eksotične posebnosti, ki že s svojim obstojem pri slovensko govorečih državljanih Slovenije vzbuja blago čudenje, vztrajno spreminja v vse bolj razširjeno in samoumevno dejstvo. V jezikoslovju se to kaže v povečanem številu teoretičnih raziskav, ki preučujejo proces usvajanja in značilnosti slovenščine kot tujega oziroma vmesnega jezika učečih se, v pospešenem razvoju sodobnih učnih pripomočkov za uspešnejše učenje in usvajanje ter v zavzemanju za učinkovitejšo jezikovnopolitično podporo. Slovenščina pa ima vse višjo vrednost tudi na jezikovnem tržišču. Število jezikovnih šol, ki po vsej Sloveniji organizirajo tečaje slovenščine za tujce, v zadnjih letih med drugim narašča zaradi Uredbe o integraciji tujcev,[65] po kateri imajo državljani tretjih držav, ki imajo v Sloveniji dovoljenje za začasno ali stalno prebivanje, pravico do brezplačnega 60- ali 180-urnega tečaja slovenščine.

Pomemben jezikovni vir za preučevanje slovenščine kot tujega jezika so korpusi usvajanja tujega jezika. Od devetdesetih let dvajsetega stoletja, ko so se najprej pojavili, so vedno bolj uveljavljeni, uporabni, razširjeni in že dolgo ne več omejeni zgolj na angleščino kot tuji jezik. Med drugim obstajajo – ali nastajajo – korpusi za francoščino, nemščino, nizozemščino, norveščino, češčino, švedščino, estonščino,

65 Uradni list RS, št. 65/2008. Na to je opozarjala tudi akcija Ministrstva za notranje zadeve Učim se slovensko, da bom lahko povedal, kdo sem; prim. http://www.infotujci.si/v/24/te%C4%8Daji-slovenskega-jezika (oktober 2011).

finščino, španščino, portugalščino, arabščino, grščino in baskovščino. Kljub temu pa v splošnih premislekih o prihodnosti korpusnega jezikoslovja tovrstni korpusi še vedno ostajajo obrobni ali celo neomenjeni (prim. Leech 2005; Čermák 2009), kot da bi šlo za samostojno disciplino, ki ima s korpusnim jezikoslovjem le malo skupnega.

Na Slovenskem je bilo do sedaj izvedenih nekaj manjših korpusnih analiz, katerih gradivo je manj primerno za širšo uporabo in utemeljeno posploševanje rezultatov, saj je bilo zbrano priložnostno in oportunistično. Zato je bil namen pričujoče raziskave določiti parametre korpusa usvajanja slovenščine kot tujega jezika (KUST), ki bi bil splošno uporaben za teoretične in pedagoške potrebe. Ker slovenska družbena in jezikovna situacija ne ponuja samoumevne izbire izhodiščnih jezikov, je bila temu posvečena posebna pozornost. Prav tako zahtevno je bilo označevanje napak v besedilih, ki je osnova za analizo napak. Ta se zaradi številnih načelnih omejitev v svoji tradicionalni obliki sicer ni razširila, s pojavom jezikovnih tehnologij pa je postala pomembna pomožna metoda pri preučevanju vmesnega jezika učečih se. Predlagana klasifikacija napak v KUST-u, ki omogoča njihov opis in kvantifikacijo, je bila oblikovana na gradivu manjšega poskusnega korpusa usvajanja slovenščine kot tujega jezika in rezultatov ankete med učitelji slovenščine kot tujega jezika. Upoštevala je tako specifičnost slovenščine kot tudi korpusne zahteve po formalnem, neinterpretativnem in čim bolj enoznačnem označevanju. Seveda pa so vsi vnaprej določeni parametri za gradnjo korpusa zgolj predvidevanja ob idealnih pogojih. Poleg nekaterih omenjenih praktičnih zadržkov, povezanih predvsem z dinamiko razmerij med skupinami govorcev posameznih izhodiščnih jezikov v Sloveniji, lahko na konkretno zasnovo in podobo korpusa vplivajo še druge, zaenkrat nepredvidene omejitve.

Določeni vidiki gradnje korpusa za uspešno izvedbo zahtevajo večjo interdisciplinarnost, denimo avtomatsko lematiziranje in oblikoslovno označevanje besedil tujih govorcev, ki imajo več nepredvidljivih oblik kot teksti domačih govorcev. Pri tem bodo morali sodelovati tudi računalniški strokovnjaki, ki so se do neke mere z neustreznimi oblikami v besedilih že srečevali (prim. Holozan 2006). Največji izziv pa bo izdelava govorne komponente KUST-a. Vključevala bo zamudno snemanje, transkripcijo in označevanje govorjenih besedil. Poseben premislek bo treba nameniti samemu konceptu napake v govorjenem jeziku, ki se od pisnega zelo razlikuje na skladenjski, pa tudi na drugih ravneh.

Uporabnost korpusa usvajanja tujega jezika se zdi omejena na sorazmerno ozko področje slovenščine za tujce. Mnenja učiteljev in drugih strokovnjakov, ki so do sedaj temeljila predvsem na empiričnih izkušnjah, bodo s KUST-om lahko kvantitativno preverjena. Kateri vidiki slovenščine so za tujce v resnici najbolj problematični? Kaj je značilno za govorce različnih prvih jezikov? Ali obstajajo razlike med govorci, ki se jezik učijo formalno in organizirano, in tistimi, ki ga neformalno usvajajo iz okolja? Spoznanja in izkušnje iz KUST-a so uporabni tudi na gradivu domačih govorcev in za jezikovnotehnološke aplikacije: klasifikacija napak je že bila temelj za delno označevanje napak v korpusu usvajanja slovenščine domačih govorcev Šolar.

Kljub vedno močnejšemu uveljavljanju jezikovnih tehnologij pa bo potreba po človeški obdelavi podatkov ostala. »Niti korpusni jezikoslovci v petdesetih, ki so zavračali intuicijo, niti generativni jezikoslovci v šestdesetih, ki so zavračali korpusne podatke, niso zmogli doseči take interakcije med obsegom podatkov in uvidom, ki je značilna za mnoge uspešne korpusne analize v zadnjih letih« (Leech 2005: 37). Če jezikovne tehnologije, npr. črkovalniki ali strojni prevajalniki, prevzemajo del opravil iz dejavne jezikovne zmožnosti govorcev, jo s tem delno siromašijo, hkrati pa ima govorec »priložnost razvijati nove zmožnosti, kar se – sicer manj intenzivno – pravzaprav dogaja že skozi vso zgodovino človeške skupnosti« (Stabej 2003: 7). V kaj se bodo razvili obeti korpusov usvajanja tujega jezika, bo seveda pokazala predvsem njihova čim širša uporaba skozi čas. Morda pa bodo ugotovitve iz KUST-a vplivale na nekatere stereotipe, zakoreninjene med neprofesionalnimi domačimi govorci slovenščine – med njimi je verjetno najabstraktnejši, a hkrati najprestižnejši ta, da je slovenščina za tujce preprosto pretežka in se je ne morejo naučiti.

8 SEZNAM KORPUSOV USVAJANJA TUJEGA JEZIKA S SPLETNIMI STRANMI

ASK: Norsk andrespråkskorpus, http://ask.uib.no/ (avgust 2011)

ASU: Andrespråkets strukturutveckling, http://www.ling.su.se/staff/ham/projects.html (junij 2010)

CEDEL2: Corpus Escrito del Español L2, http://www.uam.es/proyectosinv/woslac/cedel2.htm (avgust 2011)

CEEAUS: The Corpus of English Essays Written by Asian University Students, http://language.sakura.ne.jp/s/ceeause.html (avgust 2011)

CEJL: Corpus of English by Japanese Learners, http://www.eng.ritsumei.ac.jp/lcorpus/ (junij 2010)

CLC: Cambridge Learner Corpus, http://www.cambridge.org/si/elt/catalogue/subject/custom/item3646603/Cambridge-International-Corpus-Cambridge-Learner-Corpus/ (avgust 2011)

Corpus parlato di italiano L2, http://elearning.unistrapg.it/osservatorio/Home.html (avgust 2011)

EIC: Estonian Interlanguage Corpus, http://evkk.tlu.ee (avgust 2011)

ELFA: English as a Lingua Franca in Academic Settings Corpus, http://www.helsinki.fi/englanti/elfa/elfacorpus (avgust 2011)

ESFSLD: European Science Foundation Second Language Databank, http://www.mpi.nl/world/ISLE/overview/Overview_ESFSLD.html (november 2008)

EVA Corpus of Norwegian School English, http://kh.hd.uib.no/eva/ (avgust 2011)

FALKO: Ein fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache, http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung-en/falko (junij 2010)

FLLOC: French Learner Language Oral Corpora, http://www.flloc.soton.ac.uk/ (avgust 2011)

FRIDA: French Interlanguage Database, http://www.uclouvain.be/en-cecl-frida.html (avgust 2011)

HKCSE: Hong Kong Corpus of Conversational English, http://langbank.engl.polyu.edu.hk/HKCSE/ (avgust 2011)

HKUST: Hong Kong University of Science and Technology Corpus of Learner English, http://home.ust.hk/~autolang/download_WP.htm (oktober 2006)

IBLC: Indianapolis Business Learner Corpus, http://www.liberalarts.iupui.edu/icic/research/indianapolis_business_learner_corpus (avgust 2011)

ICLE: International Corpus of Learner English, http://www.uclouvain.be/en-cecl-icle.html (avgust 2011)

ICLFI: International Corpus of Learner Finnish, http://www.oulu.fi/hutk/sutvi/oppijankieli/en/ (avgust 2011)

InterFra, http://www.fraitaklass.su.se/english/interfra/home-interfra (avgust 2011)

InterIta, http://www.fraita.su.se/interita/ (avgust 2011)

ISLE Corpus of non-native spoken English, http://nats-www.informatik.uni-hamburg.de/~isle/speech.html (avgust 2011)

JASMIN: Jongeren, Anderstaligen en Senioren in Mens-machine Interactie, http://www.esat.kuleuven.be/psi/spraak/projects/JASMIN/ (avgust 2011)

JEFLL: Japanese English as a Foreign Language Learner, http://jefll.corpuscobo.net/ (avgust 2011)

JPU: Janus Pannonius University Corpus, http://joeandco.blogspot.com/ (avgust 2011)

Lancom, http://bach.arts.kuleuven.be/elicop/ (avgust 2011)

LBLC: Learner Business Letters Corpus, http://www.someya-net.com/concordancer/ (avgust 2011)

LCLE: Longman Corpus of Learners' English, http://www.longman.com/dictionaries/corpus/learners.html (avgust 2011)

LINDSEI: Louvain International Database of Spoken English Interlanguage: http://www.uclouvain.be/en-cecl-lindsei.html (avgust 2011)

LONGDALE: Longitudinal Database of Learner English, http://cecl.fltr.ucl.ac.be/LONGDALE.html (avgust 2011)

LPELC: Łódź Polish English Learner Corpus, http://pelcra.ia.uni.lodz.pl/corpora_en.php (november 2008), http://pelcra.ia.uni.lodz.pl/corpora_en.php (avgust 2011)

MAELC: Multimedia Adult English Learner Corpus, http://www.labschool.pdx.edu/maelc_access.html (avgust 2011)

MELD: Montclair Electronic Language Database, http://chss.montclair.edu/linguistics/MELD/ (avgust 2011)

MICASE: Michigan Corpus of Academic Corpus English, http://micase.elicorpora.info/ (avgust 2011)

PolyU: http://langbank.engl.polyu.edu.hk/indexl.html (avgust 2011)

SPLLOC: Spanish Learner Language Oral Corpora, http://www.splloc.soton.ac.uk/ (avgust 2011)

SST: Standard Speaking Test Corpus, korpus TAO (Tele-Nexommunications Advancement Organization of Japan) ali korpus NICT JLE (Japanese Learner English), http://leo.meikai.ac.jp/~tono/sst/ (november 2008)

TED: Translanguage English Database, http://www. ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S04 (junij 2010)

TELC: Thai English Learner Corpus, http://ling.arts.chula.ac.th/TLE/ (avgust 2011)

TeleNex Student Corpus, http://www.telenex.hku.hk/telec/smain/sintro/intro.htm (avgust 2011)

USE: Uppsala Student English Corpus, http://hem.passagen.se/ ylvaberg/useinfo1.htm (avgust 2011)

VALICO: Varietà di Apprendimento della Lingua Italiana: Corpus Online, http://www.bmanuel.org/projects/br-HOME.html (avgust 2011)

VOICE: Vienna-Oxford International Corpus of English, http://www.univie.ac.at/voice/ (avgust 2011)

WRiCLE: Written Corpus of Learner English, http://web.uam.es/proyectosinv/woslac//Wricle/ (avgust 2011)

9 LITERATURA

Mariko Abe, Yukio Tono. Variations in L2 spoken and written English: investigating patterns of grammatical errors across proficiency levels. *Proceedings of Corpus Linguistics 2005*. Birmingham, 2005. Http://www.corpus.bham.ac.uk/PCLC/CL2005proceedings_AbeTono.doc (november 2008).

Ghazi Abuhakema, Reem Faraj, Anna Feldman, Eileen Fitzpatrick. Annotating an Arabic Learner Corpus for Error. *LREC 2008 Proceedings*. 2008: 1347–51. Http://www.lrec-conf.org/proceedings/lrec2008/pdf/343_paper.pdf (avgust 2011).

Annelie Ädel. Involvement features in writing: Do time and interaction trump register awareness? *Linking up contrastive and learner corpus research*. Ur. Gaëtanelle Gilquin, Szilvia Papp, María Belén Díez-Bedmar. Amsterdam, New York: Rodopi, 2008: 35–53.

I. Aldabe, I. Aldezabal, M. Aranazabe, B. Arrieta, A. Díaz de Ilarraza, K. Gojenola, M. Maritxalar, M. Oronoz, A. Otegi, L. Uria. *Euskarazko erroreen sailkapena: ERROREAK eta DESBIDERATZEAK datu-baseetan*. Barne-Txostena, 2005.

Quentin Grant Allan. The TeleNex secondary learner corpus: A resource for teacher development. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Ur. Sylviane Granger. Philadelphia: John Benjamins Publishing Company, 2002: 195–211.

Jens Allwood, Sven Strömqvist, Kaarlo Voionmaa. Ecology of Adult Language Acquisition: A Psycholinguistic Rese-

arch Project. *Gothenburg Papers in Theoretical Linguistics 45*. University of Göteborg, Dept of Linguistics, 1983.

Špela Arhar. Gradnja specializiranega korpusa. *Jezik in slovstvo* 51/1 (2006): 53–67.

Špela Arhar, Vojko Gorjanc. Korpus FidaPLUS: Nova generacija slovenskega referenčnega korpusa. *Jezik in slovstvo* 52/2 (2007): 95–110.

Špela Arhar. Jezikovne tehnologije in nove metode. *Slovarji, več kot le besede*. Predavanje na konferenci, 2009.

Špela Arhar Holdt, Iztok Kosem, Irena Krapš Vodopivec, Nina Ledinek, Sara Može, Mojca Stritar, Tanja Svenšek, Ana Zwitter Vitez. *Pedagoška slovnica pri projektu Sporazumevanje v slovenskem jeziku: K16 – Standard za korpusno analizo slovničnih pojavov.* Ljubljana, 2011. Http://www.slovenscina.eu/Media/Kazalniki/Kazalnik16/Kazalnik_16_Pedagoska_slovnica_SSJ.pdf (september 2011).

Guy Aston. *Small and Large Corpora in Language Learning*. 1997. Http://home.sslmit.unibo.it/~guy/wudj1.htm (junij 2004).

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

Eric Atwell. *A Corpus of German and Italian English Language Learners' (Mis)pronunciations for Project ISLE: Interactive Spoken Language Education*. 2001. Http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Events/icamepr.htm#isle (julij 2004).

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

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

Tatjana Balažic Bulc. Jezikovni prenos pri učenju sorodnih jezikov (na primeru slovenščine in srbohrvaščine). *Jezik in slovstvo* 49/3–4 (2004): 77–89.

Barbara Baloh. Slovenščina kot drugi jezik v šolah z italijanskim učnim jezikom v Slovenski Istri. *Slovenščina in njeni uporabniki v luči evropske integracije*. Ur. Vesna Mikolič, Karin Marc Bratina. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Zgodovinsko društvo za južno Primorsko, Založba Annales, 2005: 131–147.

Janja Ban, Tina Špelko. Slovenska jezikovna skupnost v Argentini. *Jezik in slovstvo* 51/3–4 (2006): 71–84.

Robert de Beaugrande. Large Corpora, Small Corpora, and the Learning of »Language«. *Small Corpus Studies and ELT*. Ur. Mohsen Ghadessy, Alex Henry, Robert L. Roseberry. Amsterdam, Philadelphia: John Benjamins Publishing Co., 2001: 3–28.

Johnny Bigert, Viggo Kann, Ola Knutsson, Ola Sjöberg. Grammar Checking for Swedish second language learners. *CALL for the Nordic Languages*. 2005. Http://www.nada.kth.se/~jsh/publications/bigkanknusjo05.pdf (december 2006).

Uroš Bonšek. *Jezikovne značilnosti Slovencev v Argentini*. Magistrsko delo. Ljubljana: Oddelek za slovenistiko Filozofske fakultete Univerze v Ljubljani, 2010.

Simon Botley, Faizal Hakim, Doreen Dillah. Investigating Spelling Errors In A Malaysian Learner Corpus. *Malaysian Journal of ELT Research* 3 (2007): 74–93. Http://www.melta.org.my/modules/tinycontent/Dos/botley_09012008.pdf (avgust 2011).

Lynne Bowker, Jennifer Pearson. *Working with Specialized Language: A practical guide to using corpora*. London in New York: Routledge, 2002.

Adriane Boyd. EAGLE: An Error-Annotated Corpus of Beginning Learner German. *LREC 2008 Proceedings*. 2008: 1897–1902. Http://www.lrec-conf.org/proceedings/lrec2010/pdf/812_Paper.pdf (september 2011).

Angelika Breiteneder, Marie-Luise Pitzl, Stefan Majewski, Theresa Klimpfinger. VOICE Recording – Methodological Challenges in the Compilation of a Corpus of Spoken ELF. *Nordic Journal of English Studies* 5/2 (2006): 161–188.

Anja Briški. *Jezikovna zmožnost v slovenščini kot drugem jeziku v osnovni šoli*. Diplomsko delo. Ljubljana: Oddelek za slavistiko in Oddelek za slovenistiko Filozofske fakultete Univerze v Ljubljani, 2010.

H. Douglas Brown. *Principles of Language Learning and Teaching*. Englewood Cliffs: Prentice-Hall, 1980.

Jan Bušta, Dana Hlaváčková, Miloš Jakubíček, Karel Pala. Classification of Errors in Text. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009*. Ur. Petr Sojka, Aleš Horák. Brno: Masaryk University, 2009: 109–119. Http://nlp.fi.muni.cz/raslan/2009/papers/6.pdf (september 2011).

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

Thomas Cobb. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review* 59/3 (2003): 393–423.

Ulla Connor, Kristen Precht, Thomas Upton. Business English: Learner data from Belgium, Finland and the U. S. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Ur. Sylviane Granger. Philadelphia: John Benjamins Publishing Company, 2002: 185–194.

S. P. Corder. The Significance of Learners' Errors. *Error Analysis: Perspectives on Second Language Acquisition*. Ur. Jack C. Richards. London: Longman, 1974a: 19–27.

S. P. Corder. Idiosyncratic Dialects and Error Analysis. *Error Analysis: Perspectives on Second Language Acquisition*. Ur. Jack C. Richards. London: Longman, 1974b: 158–171.

S. Pit Corder. *Error Analysis and Interlanguage*. London: Oxford University Press, 1982.

Catia Cucchiarini, Joris Driesen, Hugo Van hamme, E. Sanders: Recording Speech of Children, Non-Natives and Elderly People for HLT. *LREC 2008 Proceedings*. 2008. Http://www.lrec-conf.org/proceedings/lrec2008/summaries/366.html (december 2008).

František Čermák. Jezikovni korpus: sredstvo in vir spoznanj. *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina, 2005: 137–171.

František Čermák. *Some of the Current Problems of Corpus and Computational Linguistics or Fifteen Commandments and General Truths*. Predavanje na Filozofski fakulteti Univerze v Ljubljani, 2009.

Gaja Červ. Žanrski korpus novinarskih besedil. *Obdobja 24: Infrastruktura slovenščine in slovenistike*. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 89–95.

Estelle Dagneaux, Sharon Denness, Sylviane Granger. Computer-aided error analysis. *System* 26/2 (1998): 163–174.

Estelle Dagneaux, Sylviane Granger, Fanny Meunier, Stephanie Petch-Tyson, Xavier Vilret. *A web interface to the International Corpus of Learner English*. 2001. Http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Events/icamepr.htm#interface (julij 2004).

S. De Cock, Sylviane Granger, Stephanie Petch-Tyson. *The Louvain International Database of Spoken English Interlanguage: The LINDSEI Project.* 1999. Http://juppiter.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Lindsei/download/lindsei.pdf (oktober 2006).

María Belén Díez Bedmar. Struggling with English at University level: Error patterns and problematic areas of first-year students' interlanguage. *Proceedings of Corpus Linguistics 2005*. Birmingham, 2005. Http://www.corpus.bham.ac.uk/PCLC/BedmarUniEnglish.doc (november 2008).

María Belén Díez Bedmar, Szilvia Papp. The use of the English article system by Chinese and Spanish learners. *Linking up contrastive and learner corpus research*. Ur. Gaëtanelle Gilquin, Szilvia Papp, María Belén Díez Bedmar. Amsterdam, New York: Rodopi, 2008: 147–175.

María Belén Díez Bedmar. Written Learner Corpora by Spanish Students of English: An Overview. *A Survey on Corpus-based Research: Panorama de investigaciones basadas en corpus*. AELINCO, 2009: 920–933. Http://www.um.es/lacell/aelinco/contenido/pdf/62.pdf (avgust 2011).

Danilo Dolenc, Katja Kalin, Janja Povhe, Barica Razpotnik, Darja Šter, Tina Žnidaršič. Prebivalstvo. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko, 2012.

Heidi Dulay, Marina Burt, Stephen Krashen. *Language Two*. New York, Oxford: Oxford University Press, 1982.

Sašo Džeroski, Tomaž Erjavec. Strojno učenje lematizacije neznanih slovenskih besed. *Informacijska družba, Jezikovne tehnologije*. Ur. Tomaž Erjavec, Jerneja Gros. Ljubljana: Institut Jožef Stefan, 2000: 14–19.

Sašo Džeroski, Tomaž Erjavec, Jakub Zavrel. Morphosyntactic Tagging of Slovene: Evaluating PoS Taggers and Tagsets. *Second International Conference on Language Resources and Evaluation, LREC'00*. Pariz: ELRA, 2000: 1099–1104.

Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdenek Žabokrtský, Andreja Žele. Towards a Slovene

Dependency Treebank. *Fifth International Conference on Language Resources and Evaluation, LREC'06*. Genova: ELRA, 2006: 1388–1391.

Julian Edge. *Mistakes and Correction*. London, New York: Longman, 1989.

Rod Ellis. *The Study of Second Language Acquisition*. Oxford: Oxford University Press, 1994.

Tomaž Erjavec: Računalniške zbirke besedil. *Jezik in slovstvo* 42/2–3 (1997): 81–96.

Tomaž Erjavec, Vojko Gorjanc, Peter Holozan, Marko Stabej. Application to Slovene. *Specifications and Notation for MULTEXT-East Lexicon Encoding*. Ur. Tomaž Erjavec. MULTEXT-East / Concede, 2001.

Tomaž Erjavec. Označevanje korpusov. *Jezik in slovstvo* 48/3–4 (2003): 61–76.

Tomaž Erjavec. *Uvod v korpusno jezikoslovje*. 2004. Http://nl.ijs.si/et/talks/korpus/korpusno.html (november 2006).

Tomaž Erjavec, Nina Ledinek. Slovenska odvisnostna drevesnica: Prvi rezultati. *Zbornik Pete konference Jezikovne tehnologije*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2006: 162–167.

Tomaž Erjavec, Simon Krek. Oblikoskladenjske specifikacije in označeni korpusi JOS. *Zbornik Šeste konference Jezikovne tehnologije*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2008: 49–53.

Tomaž Erjavec, Darja Fišer, Simon Krek, Nina Ledinek. The JOS linguistically tagged corpus of Slovene. *LREC 2010 Proceedings*. 2010: 1806–09. Http://www.lrec-conf.org/proceedings/lrec2010/pdf/139_Paper.pdf (september 2011).

Norman Fairclough. *Language and Power*. Essex: Pearson Education, 2001.

Ina Ferbežar. Merjenje in merljivost v jeziku: Na stičišču jezikoslovja in psihologije: nekaj razmislekov. *Slavistična revija* 47/4 (1999): 417–436.

Ina Ferbežar, Nataša Pirih Svetina. *Slovenščina za tujce: Izobraževalni program, izobraževanje odraslih*. Ljubljana: Ministrstvo za šolstvo, znanost in šport, Zavod RS za šolstvo, 2002.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2003*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2004a: 17–20.

Ina Ferbežar. Jaz svakodnevno uporabljam jezik slovenščine …. *Moderno v slovenskem jeziku, literaturi in kulturi: 40. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2004b: 193–198.

Ina Ferbežar, Boštjan Bajec, Agnes Pisanski Peterlin. *Priročnik za avtorje testnih gradiv*. Ljubljana: Center za slovenščino kot drugi/tuji jezik, Krakov: TAiWPN Universitas, 2004.

Ina Ferbežar, Mihaela Knez, Andreja Markovič, Nataša Pirih Svetina, Mojca Schlamberger Brezar, Marko Stabej, Hotimir Tivadar, Jana Zemljarič Miklavčič. *Sporazumevalni prag za slovenščino*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, Ministrstvo RS za šolstvo in šport, 2004.

Ina Ferbežar, Nataša Pirih Svetina. Certificiranje slovenščine kot drugega/tujega jezika – zgodovina in perspektive. *Jezik in slovstvo* 49/3–4 (2004): 17–33.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2004*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2005: 22–24.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2005*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2006: 24–27.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2006*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2007a: 32–35.

Ina Ferbežar. Nekaj vmes. *Stereotipi v slovenskem jeziku, literaturi in kulturi: 43. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2007b: 194–197.

Ina Ferbežar. *Navodila za izvajalce izpitov iz znanja slovenščine z načeli dobre prakse*. Interno gradivo. Ljubljana: Izpitni center Centra za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete Univerze v Ljubljani, 2007c.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2007*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2008: 38–41.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2008*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2009: 46–49.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2009*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2010: 43–47.

Ina Ferbežar. Izpitni center. *Letno poročilo Centra za slovenščino kot drugi/tuji jezik 2010*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2011: 45–48.

Eileen Fitzpatrick, M. S. Seegmiller. The Montclair Electronic Language Database Project. A*pplied Corpus Linguistics: A Multidimensional Perspective*. Ur. Ulla Connor, Thomas A. Upton. Amsterdam, New York: Rodopi, 2004: 223–237.

Lynn Flowerdew. The Exploitation of Small Learner Corpora in EAP Materials Design. *Small Corpus Studies and ELT*. Ur. Mohsen Ghadessy, Alex Henry, Robert L. Roseberry. Amsterdam, Philadelphia: John Benjamins Publishing Co., 2001: 363–79.

Bruno Frischherz. *Lernen, um zu sprechen – sprechen, um zu lernen. Diskursanalytische Untersuchungen zum Zweitspracherwerb türkischer und kurdischer Asylbewerber in der Deutschschweiz*. Doktorska disertacija. Univerza v Freiburgu, 1997.

Polona Gantar, Simon Krek. Drugačen pogled na slovarske definicije: opisati, pojasniti, razložiti?. *Obdobja 24: Infrastruktura slovenščine in slovenistike*. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 151–159.

Patrick Gillard, Adam Gadsby. Using a learners' corpus in compiling ELT dicitionaries. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998: 159–171.

Gaëtanelle Gilquin. The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast* 3/1 (2000): 95–123.

Šarolta Godnič Vičič. Potentials and Challenges of ESP Learner Corpora: The Case of Modal Auxiliaries in Slovene ESP Learners' Written Interlanguage. *Jezik stroke in vloga učitelja tujega jezika stroke v slovenskem izobraževalnem prostoru: Zbornik posveta ob 10. obletnici delovanja*

Slovenskega društva učiteljev tujega strokovnega jezika. Ur. Metka Brkan. Ljubljana: Slovensko društvo učiteljev tujega strokovnega jezika. 2008: 15–30.

Vojko Gorjanc. Korpusi v jezikoslovju in korpus slovenskega jezika FIDA. *35. seminar slovenskega jezika, literature in kulture: Zbornik predavanj*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovanske jezike in književnosti Filozofske fakultete. 1999: 47–59.

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

Vojko Gorjanc, Primož Jurko. Kolokacije in učenje tujega jezika. *Jezik in slovstvo* 49/3–4 (2004): 49–62.

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

Vojko Gorjanc, Nataša Logar. Od splošnih do specializiranih korpusov: Načela gradnje glede na njihov namen. *Razvoj slovenskega strokovnega jezika. Obdobja – metode in zvrsti, 24*. Ur. Irena Orel. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2007: 637–650.

Vojko Gorjanc, Darja Fišer. *Korpusna analiza*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani. 2010.

Martin Grad. Izražanje osebnosti v akademskem diskurzu: primerjava rojenih in tujih govorcev angleščine. *Slovenske korpusne raziskave*. Ur. Špela Vintar. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2010: 72–88.

Sylviane Granger. The computer learner corpus: a versatile new source of data for SLA research. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998a: 3–18.

Sylviane Granger. Prefabricated patterns in advanced EFL writing: collocations and formulae. *Phraseology: theory, analysis and applications*. Oxford: Oxford University Press, 1998b: 145–160.

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

Sylviane Granger. *International Corpus of Learner English: The ICLE Project*. Http://juppiter.fltr.ucl.ac.be/FLTR/

GERM/ETAN/CECL/Cecl-Projects/Icle/download/icle.pdf (maj 2004). 2001.

Sylviane Granger. The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. *TESOL Quarterly* 37/3 (2003a): 538–546.

Sylviane Granger. Error-tagged Learner Corpora and CALL: A Promising Synergy. *CALICO Journal* 20/3 (2003b): 465–480.

Sylviane Granger. Computer Learner Corpus Research: Current Status and Future Prospects. *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, 2004: 123–145.

Sylviane Granger, Jennifer Thewissen. *Error Tagging Session*. Gradivo na seminarju Learner Corpus Research: Theory and Applications. Louvain-la-Neuve, 2006.

Sylviane Granger, Magali Paquot. *An L2-Oriented Approach to Phraseology*. Gradivo na seminarju Learner Corpus Research: Theory and Applications. Louvain-la-Neuve, 2006.

Shichun Gui, Huizhong Yang. Computer Analysis of Chinese Learner English. 2006. Http://lc.ust.hk/~centre/conf2001/keynote/subsect4/yang.pdf (oktober 2006).

Jon Erik Hagen. *Andrespråksteoretiske utfordringer: Er ASK en eneste stor »comparative fallacy«?*. Predavanje na seminarju ASK-workshop, Bergen, 2006.

Jirka Hana, Svatava Škodová, Alexandr Rosen, Barbora Štindlová. Error-tagged Learner Corpus of Czech. *Proceedings of the Fourth Linguistic Annotation Workshop*. Uppsala: Association for Computational Linguistics, 2010: 11–19. Http://www.aclweb.org/anthology/W/W10/W10-1802.pdf (avgust 2011).

Angela Hasselgren. The EVA Corpus of Norwegian School English. *ICAME Journal* 21 (2001): 123–124.

Nataša Hirci, Agnes Pisanski. Nove jezikovne tehnologije: vidiki uporabe računalniških korpusov. *Vestnik Društva za tuje jezike in književnosti Slovenije* 34/1–2 (2000): 27–34.

Peter Holozan. Dodatne dvoumnosti zaradi popustljivosti analizatorja pri analizi slovenskih stavkov. *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference 2006*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2006: 146–149.

Peter Holozan, Simon Krek, Matej Pivec, Simon Rigač, Simon Rozman, Aleš Velušček. *Projekt »Sporazumevanje v slovenskem jeziku«: Specifikacije za učni korpus*. 2008. Http://www.slovenscina.eu/Media/Kazalniki/Kazalnik2/SSJ_Kazalnik_2_Specifikacije-ucni-korpus_v1.pdf (september 2011).

Józef Horváth. *Advanced Writing in English as a Foreign Language*. Pécs: Lingua Franca Csoport, 2001.

Susan Hunston. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press, 2002.

Milena Ilić, Katja Kalin, Janja Povhe, Barica Razpotnik, Darja Šter, Tina Žnidaršič. Prebivalstvo. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko, 2008: 72–101.

Uršula Iljaš Petrovič, Janja Povhe, Barica Razpotnik, Darja Šter, Tina Žnidaršič. Prebivalstvo. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko, 2010.

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

Carl James. *Errors in Language Learning and Use: Exploring Error Analysis*. London, New York: Longman, 1998.

Jarmo Harri Jantunen. *Corpus driven analysis of cotextual units of meaning in learner language*. Predavanje na konferenci New trends in corpus linguistics for language teaching and translation studies. Granada, 2008.

Jarmo Harri Jantunen. *ICLFI: Overview of the project and corpus*. Predavanje na srečanju ASKeladden netwok meeting on learner corpus. 2010. Http://www.oulu.fi/hutk/sutvi/oppijankieli/tutkimus/Bergen%20ASKelladden.pdf (junij 2010).

Tanja Jerman. Slovenščina na daljavo. *Moderno v slovenskem jeziku, literaturi in kulturi: 40. seminar slovenskega jezika, literature in kulture.* Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2004: 199–201.

Ana Jurjevič. *Uporaba angleščine kot jezika posredovanja pri poučevanju slovenščine kot drugega/tujega jezika*. Diplomsko delo. Ljubljana: Oddelek za slovenistiko Filozofske fakultete, 2008.

Zdravko Kačič. *Govorne tehnologije in jezikovni viri*. Predavanje na posvetu Slovenskega društva za jezikovne teh-

nologije, Maribor, 2007. Http://www.sdjt.si/dogodki/MB2007/KACIC-jezikovni-viri-2007.pdf (avgust 2011).

Przemysław Kaszubski. Enhancing a writing textbook: a national perspective. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998: 172–185.

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

Mihaela Knez. Katere učbenike pa imate?: Predstavitev učbenikov Centra za slovenščino kot drugi/tuji jezik. *Večkulturnost v slovenskem jeziku, literaturi in kulturi: 43. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2005: 216–218.

Zuzanna Krystyna Kobos, Nataša Pirih Svetina. Učenje in poučevanje slovenščine kot neprvega jezika v Sloveniji. *Obdobja 24: Infrastruktura slovenščine in slovenistike*. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 203–209.

Brez avtorja. *Korpus pisnih besedil: Specifikacije postopkov za redno zbiranje tekstovnega gradiva za korpus*. 2008. Http://www.slovenscina.eu/Media/Kazalniki/Kazalnik1/SSJ_Kazalnik_1_Specifikacije-pisni-korpus_v1.pdf (september 2011).

Irena Kovačič. Prevajanje v učenju slovenščine kot tujega jezika. *Slovenski jezik, literatura in kultura v izobraževanju. 39. seminar slovenskega jezika literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2003: 213–220.

Simona Kranjc. Besedni red, usvajanje prvega in učenje drugega/tujega jezika. *Jezik in slovstvo* 49/3–4 (2004): 145–157.

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

Ana Kure. *Ocena napredka jezikovne zmožnosti učencev priseljencev po 30-urnem jezikovnem tečaju (sociološki in jezikoslovni vidiki)*. Diplomsko delo. Ljubljana: Oddelek za slovenistiko in Oddelek za sociologijo Filozofske fakultete Univerze v Ljubljani, 2010.

Nina Ledinek, Tomaž Erjavec. Odvisnostno površinskoskladenjsko označevanje slovenščine: Specifikacije in označeni korpusi. *Obdobja 24: Infrastruktura slovenščine in*

slovenistike. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 219–224.

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

Geoffrey Leech. Preface. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998: xiv–xx.

Geoffrey Leech. Stanje stvari v korpusnem jezikoslovju. *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina. 2005: 29–57.

Paul Lennon. Error: Some problems of definition, identification, and distinction. *Applied Linguistics* 12 (1991): 180–195.

Linda H. F. Lin. *Applying Information Technology to a corpus of student report writing to help students write better reports*. 1999. Http://elc.polyu.edu.hk/conference/papers/Lin.htm (julij 2004).

Nataša Logar Berginc, Simon Šuster. Gradnja novega korpusa slovenščine. *Jezik in slovstvo* 54/3–4 (2009): 57–68.

Birte Lönneker, Katarina Rozman. Online SLO-DE-SLO: Spletni slovensko-nemški in nemško-slovenski slovar. *Jezikovne tehnologije za slovenski jezik/Language Technologies for the Slovene Language. Zbornik konference/Proceedings of the Conference*. Ur. Tomaž Erjavec, Jerneja Gros. Ljubljana: Institut Jožef Stefan, 2004: 56–63.

Birte Lönneker. Strojno oblikoslovno označevanje slovenskih besedil: Kako daleč smo. *Slavistična revija* 53/2 (2005): 193–210.

Hui-Chuan Lu, Szu-Chia Wu. Estudio del léxico a partir de contraste de dos corpus: El CATE y el CLE. *Studies in contrastive linguistics: Proceedings of the 4th International Contrastive Linguistics Conference*. Santiago de Compostela, 2005: 391–396.

Hui-Chuan Lu. An annotated Taiwanese Learners' Corpus of Spanish, CATE. *Corpus Linguistics and Linguistic Theory* 6/2 (2010): 297–300.

Anke Lüdeling, Maik Walter, Emil Kroymann, Peter Adolphs. Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics 2005*. Birmingham, 2005. Http://www.corpus.bham.ac.uk/PCLC/Falko-CL2006.doc (junij 2010).

Ana Martín Úriz. *UAM Corpus de Interlenguas Escritas*. Http://www.uam.es/departamentos/filoyletras/filoinglesa/bin/docs/investigacion/UAM%20Corpus.pdf (avgust 2011).

Anna Mauranen. The Corpus of English as Lingua Franca in Academic Settings. *TESOL Quarterly* 37/3 (2003): 513–527.

Anna Mauranen. Refleksivnost diskurza pri mednarodnih govorcih – raba v angleščini kot lingui franci. *Jezik in slovstvo* 52/3–4 (2007): 33–51.

James McCrostie. Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal* 32 (2008): 97–114.

Amaya Mendikoetxea, Michael O'Donnell, Paul Rollinson. WriCLE: A learner corpus for Second Language Acquisition Research. *Proceedings of the Corpus Linguistics Conference CL2009*. Liverpool: Univerza v Liverpoolu, 2009. Http://ucrel.lancs.ac.uk/publications/cl2009/351_FullPaper.doc (junij 2010).

Wolfgan Menzel, Eric Atwell, Patrizia Bonaventura, Daniel Herron, Peter Howarth, Rachel Morton, Clive Souter. The ISLE corpus of non-native spoken English. *LREC 2000 2nd International Conference on Language Resources & Evaluation Proceedings*. 2000: 313–320. Http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.1687&rep=rep1&type=pdf (avgust 2011).

Fanny Meunier. Computer tools for the analysis of learner corpora. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998: 19–37.

Vesna Mikolič. Sporazumevalna zmožnost v J1 in J2 skozi prizmo časa in prostora – primer Slovenske Istre. *Slovenščina in njeni uporabniki v luči evropske integracije*. Ur. Vesna Mikolič, Karin Marc Bratina. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Zgodovinsko društvo za južno Primorsko, Založba Annales, 2005: 173–185.

Vesna Mikolič, Ana Beguš, Davorin Dukič, Miha Koderman. Vpliv namembnosti korpusa na označevanje besedilnega gradiva za »Večjezični korpus turističnih besedil«. *Zbornik Šeste konference Jezikovne tehnologije*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2008: 60–64.

John Milton. Exploiting L1 and interlanguage corpora in the design of an electronic language learning and

production environment. *Learner English on Computer*. Ur. Sylviane Granger. London, New York: Longman, 1998: 186–98.

Danny Minn, Hiroshi Sano, Marie Ino, Takahiro Nakamura. Using the BNC to create and develop educational materials and a website for learners of English. *ICAME Journal* 29 (2005): 99–113.

Freda Mishan. Authenticating corpora for language learning: a problem and its resolution. *ELT Journal* 58/3 (2004): 219–227.

Florence Myles. Interlanguage corpora and second language acquisition research: Review article. *Second Language Research* 21/4 (2005): 373–391.

Florence Myles. Using electronic corpora in SLA research. *Handbook of French Applied Linguistics*. Ur. Dalila Ayoun. Amsterdam: John Benjamins, 2007: 377–400.

William Nemser. Approximative Systems of Foreign Language Learners. *Error Analysis: Perspectives on Second Language Acquisition*. Ur. Jack C. Richards. London: Longman, 1974: 55–63.

Nadja Nesselhauf. Learner corpora and their potential for language teaching. *How to Use Corpora in Language Teaching*. Ur. John Sinclair. Amsterdam: Benjamins, 2004: 125–152.

Diane Nicholls. The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. *Proceedings of the Corpus Linguistics 2003 conference*. Ur. Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery. Lancaster: UCREL, 2003. 572–581.

Anne O'Keeffe, Fiona Farr. Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications. *TESOL Quarterly* 37/3 (2003): 389–418.

Karel Pala, Pavel Rychlý, Pavel Smrž. Text Corpus with Errors. *Text, Speech and Dialogue: 6th International Conference*. Ur. Václav Matoušek, Pavel Mautner. Berlin, Heidelberg: Springer, 2003: 90–97.

Pascual Pérez-Paredes. Learner oral corpora and network-based language teaching. *How to Use Corpora in Language Teaching*. Ur. John Sinclair. Amsterdam: Benjamins, 2004: 249–268.

Ivana Petric. Besediloslovna primerjava pisne produkcije tujih in rojenih govorcev slovenščine. *Jezik in slovstvo* 48/3–4 (2004): 133–144.

Nataša Pirih Svetina, Katarina Rigler Šilc, Marjana Lavrič, Ina Ferbežar, Tanja Jerman. *Preživetvena raven v slovenščini*. Krakov: TAiWPN Universitas, 2004.

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

Nataša Pirih Svetina, Ina Ferbežar. Slovenščine tujejezičnih govorcev. *Jezik in slovstvo* 50/6 (2005): 3–15.

Nataša Pirih Svetina, Andreja Ponikvar. Slovenska slovnica za tujce – kaj imamo, kaj pa bi. *Treći hrvatsko-slovenski slavistički skup: Zbornik povzetkov.* Opatija, 2006.

Agnes Pisanski, Nataša Hirci. Vloga besedilnih korpusov pri izboljševanju jezikovne kompetence. *Proceedings of the 2nd Slovenian and 1st International Language Technologies Conference*, 2000: 64–69.

Agnes Pisanski. Uporaba novih tehnologij pri jezikovnem pouku. *Jezik in slovstvo* 48/3–4 (2003): 102–103.

Lidija Pogačnik. *Napake v besednem redu (stalni stavi) pri učenju slovenščine kot drugega/tujega jezika*. Diplomsko delo. Ljubljana: Univerza v Ljubljani, Filozofska fakulteta, 2001.

Vesna Požgaj Hadži, Ina Ferbežar. Tudi to je slovenščina. *37. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2001: 57–68.

Vesna Požgaj Hadži, Tatjana Balažic Bulc. Meje (ne)tolerantnosti do govorcev slovenščine kot tujega jezika. *Slovenščina in njeni uporabniki v luči evropske integracije*. Ur. Vesna Mikolič, Karin Marc Bratina. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Zgodovinsko društvo za južno Primorsko, Založba Annales, 2005: 245–252.

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

Wen Quifang. *Chinese Learner Corpora and Second Language Research.* 2006. Http://call2006.fltrp.com/pdf/CALL2006plenary_Wen%20Qiufang.pdf (oktober 2006).

Peter H. Ragan. Classroom Use of a Systemic Functional Small Learner Corpus. *Small Corpus Studies and ELT*. Ur. Mohsen Ghadessy, Alex Henry, Robert L. Roseberry. Amsterdam, Philadelphia: John Benjamins Publishing Co., 2001: 208–36.

Mick Randall, Nicholas Groom. The BUiD Arab Learner Corpus: a resource for studying the acquisition of L2 English

spelling. *Proceedings of the Corpus Linguistics Conference CL2009*. 2009. Http://ucrel.lancs.ac.uk/publications/cl2009/54_FullPaper.doc (avgust 2011).

Stephen Reder, Kathryn Harris, Kristen Setzler. The Multimedia Adult ESL Learner Corpus. *TESOL Quarterly* 37/3 (2003): 546–557.

Jack C. Richards, G. P. Sampson. The study of learner English. *Error Analysis: Perspectives on Second Language Acquisition*. Ur. Jack C. Richards. London: Longman, 1974: 3–18.

Katarina Rozman. *Napaka kot izhodišče in motivacija za širjenje jezikovne zmožnosti pri učenju slovenščine kot tujega jezika*. Diplomsko delo. Ljubljana: Oddelek za slovenistiko Filozofske fakultete Univerze v Ljubljani, 2006.

Tadeja Rozman. Upoštevanje ciljnih uporabnikov pri izdelavi enojezičnega slovarja za tujce. *Jezik in slovstvo* 49/3–4 (2004): 63–75.

Tadeja Rozman, Mojca Stritar, Iztok Kosem. Korpus šolskih pisnih izdelkov Šolar. *Nova didaktika poučevanja slovenskega jezika: Sporazumevanje v slovenskem jeziku*. Ljubljana, 2010: 4–35. Http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf (avgust 2011).

Tadeja Rozman, Mojca Stritar. Korpus usvajanja slovenščine. *Nova didaktika poučevanja slovenskega jezika: Sporazumevanje v slovenskem jeziku*. Ljubljana, 2010: 36–41. Http://www.slovenscina.eu/Media/Kazalniki/Kazalnik15/Nova_didaktika_Sporazumevanje.pdf (avgust 2011).

Larry Selinker. Interlanguage. *Error Analysis: Perspectives on Second Language Acquisition*. Ur. Jack C. Richards. London: Longman, 1974: 31–54.

Michael Sharwood Smith. *Second Language Learning: Theoretical Foundations*. London, New York: Longman, 1994.

Rebecca Hsue-Hueh Shih. Compiling Taiwanese Learner Corpus of English. *Computational Linguistic and Chinese Language Processing* 5/2 (2000): 89–102.

Peter Siemen, Anke Lüdeling, Frank Henrik Müller. Falko – ein fehlerannotiertes Lernerkorpus des Deutschen. *Proceedings of Konvens 2006*. Konstanz, 2006.

Več avtorjev. *Skupni evropski jezikovni okvir: Učenje, poučevanje, ocenjevanje*. Ljubljana: Ministrstvo RS za šolstvo in šport, Urad za razvoj šolstva, 2011. Http://www.mss.gov.si/fileadmin/mss.gov.si/pageuploads/podrocje/

razvoj_solstva/Jeziki/Publikacija_SEJO_komplet.pdf (avgust 2011).

Brez avtorja. *Slikovni priročnik za uporabo spletnega iskalnika po korpusu govorjene slovenščine GOS*. Maribor, 2010. Http://www.korpus-gos.net/Content/Static/Prirocnik.pdf (avgust 2011).

Več avtorjev. *Slovar slovenskega knjižnega jezika*. Ljubljana: DZS, 2005.

Več avtorjev. *Slovenski pravopis*. Ljubljana: Slovenska akademija znanosti in umetnosti, 2001.

Marko Stabej. Besedilnovrstna sestava korpusa FIDA. *Uporabno jezikoslovje* 6 (1998): 96–106.

Marko Stabej, Primož Vitez. KGB (Korpus govorjenih besedil) v slovenščini. *Proceedings of the 2nd Slovenian and 1st International Language Technologies Conference 2000*. Ljubljana: Institut Jožef Stefan, 2000: 79–81.

Marko Stabej. Jezikovne tehnologije in jezikovno načrtovanje. *Jezik in slovstvo* 48/3–4 (2003). 5–18.

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

Marko Stabej. Kako pa govoriš!? (Spis o normi in normalnem). *Spisi o govoru*. Ur. Primož Vitez. Ljubljana: Znanstvenoraziskovalni inštitut Filozofske fakultete, 2008: 87–100.

Sonja Starc. Matura iz slovenščine kot drugega jezika v Slovenski Istri. *Slovenščina in njeni uporabniki v luči evropske integracije*. Ur. Vesna Mikolič, Karin Marc Bratina. Koper: Univerza na Primorskem, Znanstveno-raziskovalno središče, Zgodovinsko društvo za južno Primorsko, Založba Annales, 2005: 149–170.

Brez avtorja. Prebivalstvo. *Statistični letopis Republike Slovenije*. Ljubljana: Zavod Republike Slovenije za statistiko, 2006: 72–98.

Mojca Stritar. Jezik na klik: Elektronski viri za učenje slovenščine. *Večkulturnost v slovenskem jeziku, literaturi in kulturi: 43. seminar slovenskega jezika, literature in kulture*. Ljubljana: Center za slovenščino kot drugi/tuji jezik pri Oddelku za slovenistiko Filozofske fakultete, 2005: 211–215.

Mojca Stritar. Oblikovanje korpusa usvajanja slovenščine kot tujega jezika. *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference 2006*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2006: 134–139.

Mojca Stritar. Merila za oblikovanje korpusov usvajanja tujega jezika. *Jezik in slovstvo* 51/5 (2006): 59–74.

Mojca Stritar. Slovene as a Foreign Language: The Pilot Learner Corpus Perspective. *Slovenski jezik / Slovene Linguistic Studies* 7 (2009): 135–152.

Mojca Stritar. Slovenščina kot tuji jezik – od mita do korpusa. *Vloge središča: Konvergenca regij in kultur.* Ur. Irena Novak Popov. Ljubljana: Zveza društev Slavistično društvo Slovenije, 2010: 293–302.

Masatoshi Sugiura. *On Enhancing the Writing Skill of EFL Learners in Japan: Utilizing the Insights and Technologies of Corpus Linguistics*. 2000. Http://oscar.lang.nagoya-u.ac.jp/~sugiura/hawaii/680p/corpuswriting.html (julij 2004).

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

Stella E. O. Tagnin. A Multilingual Learner Corpus in Brazil. *UCREL Technical Paper: Proceedings of the Corpus Linguistics 2003 Conference*. Ur. Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery. Lancaster: University, 2003: 940–945.

Melinda Tan. Authentic language or language errors? Lessons from a learner corpus. *ELT Journal* 59/2 (2005): 126–134.

Kari Tenfjord, Paul Meurer, Knut Hofland. *The ASK corpus – a language learner corpus of Norweigan as a second language* (Poster). 2004. Http://www.ugr.es/~talc6/talc_search/proceedings/60.html (maj 2009).

Kari Tenfjord. Osebni pogovor. Bergen, 24. oktober in 19. december 2006.

Kari Tenfjord, Paul Meurer, Knut Hofland. The ASK Corpus – a Language Learner Corpus of Norwegian as a Second Language. *LREC 2006 Proceedings*. 2006: 1821–24. Http://gandalf.aksis.uib.no/lrec2006/pdf/573_pdf.pdf (avgust 2011).

Wolfgang Teubert. Korpusno jezikoslovje in leksikografija. *Študije o korpusnem jezikoslovju*. Ljubljana: Krtina, 2005: 103–136.

Stefan Titscher, Michael Meyer, Ruth Wodak, Eva Vetter. *Methods of Text and Discourse Analysis*. London: SAGE Publications, 2005.

Yukio Tono. A computer learner corpus-based analysis of the acquisition order of English grammatical morphe-

mes. *TALC 98 – Conference Proceedings*. Keble College Oxford, 1998: 183–187.

Yukio Tono. *Error Analysis: Review*. 1999. Http://leo.meikai.ac.jp/~tono/ (oktober 2006).

Yukio Tono in drugi. The Standard Speaking Test (SST) Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. *The Second Asialex International Congress*. Yonsei University, 2001: 257–262. Http://leo.meikai.ac.jp/~tono/sst/asialex2001.pdf (junij 2004).

Yukio Tono. *Learner Corpus Research State of the Applications in SLA Research.* 2000. Http://leo.meikai.ac.jp/~tono/ (oktober 2006).

Yukio Tono. Learner corpora: design, development and applications. *UCREL Technical Paper: Proceedings of the Corpus Linguistics 2003 Conference*. Ur. Dawn Archer, Paul Rayson, Andrew Wilson, Tony McEnery. Lancaster: University, 2003a: 800–809.

Yukio Tono. *JEFLL corpus: Description*. 2003b. Http://leo.meikai.ac.jp/~tono/jefll.html (oktober 2006).

Yukio Tono. Multiple comparisons of IL, L1 and TL corpora: The case of L2 acquisition of verb subcategorization patterns by Japanese learners of English. *Corpora and Language Learners*. Ur. Guy Aston, Silvia Bernardini in Dominic Stewart. Amsterdam, Philadelphia: John Benjamins, 2004: 45–66.

Jože Toporišič. *Slovenska slovnica*. 2. izdaja. Maribor: Založba Obzorja, 1991.

Jože Toporišič. *Enciklopedija slovenskega jezika*. Ljubljana: Cankarjeva založba, 1992.

Dimitrios Tzimokas. *Error tagging system for Greek learner corpus of written speech*. Prispevek na srečanju Learner Corpus Colloquium, Univerza v Louvainu, 2006.

Larraitz Uria Garin. *An Error Classification for CALL and NLP*. Prispevek na srečanju Learner Corpus Colloquium, Univerza v Louvainu, 2006.

Rafal S. Uzar. *The PELE Project: A New Perspective for Learner Language Corpora*. 1998. Http://members.fortunecity.com/pelcra/pele.htm (junij 2004).

Darinka Verdonik. Govor in jezikovne tehnologije. *Spisi o govoru*. Ur. Primož Vitez. Ljubljana: Znanstvenoraziskovalni inštitut Filozofske fakultete, 2008: 231–240.

Ada Vidovič Muha. *Zloženke v slovenskem knjižnem jeziku*. Doktorska disertacija. Ljubljana: Univerza Edvarda Kardelja v Ljubljani, Filozofska fakulteta, 1983.

Špela Vintar. Kaj izvira iz jezikovnih virov. *Jezik in slovstvo* 48/3–4 (2003): 77–88.

Špela Vintar. *Terminologija: Terminološka veda in računalniško podprta terminologija*. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, Oddelek za prevajalstvo, 2008.

Špela Vintar, Tomaž Erjavec. iKorpus in luščenje izrazja za Islovar. *Zbornik Šeste konference Jezikovne tehnologije*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2008: 65–69.

Špela Vintar. Slovenski prevodoslovni korpus. *Obdobja 24: Infrastruktura slovenščine in slovenistike*. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 385–391.

Elena Volodina. *From corpus to language classroom: Reusing Stockholm Umeå Corpus in a vocabulary exercise generator SCORVEX*. Magistrsko delo. Univerza v Gothenburgu, 2008.

David Wible, Chin-Hwa Kuo, Nai-lung Tsao, Anne Liu. An Online Writing Platform for Language Teachers. *Journal of Universal Computer Science* 7/3 (2001): 278–289.

Aleksandra Zazijal. *Tečaj slovenskega jezika po programu za integracijo tujcev kot dejavnik v procesu vključevanja priseljencev v Slovenijo*. Diplomsko delo. Ljubljana: Oddelek za slovenistiko in Oddelek za sociologijo Filozofske fakultete Univerze v Ljubljani, 2010.

Jana Zemljarič Miklavčič. *Jezikoslovni vidiki slovenščine kot tujega jezika*. Magistrsko delo. Ljubljana: Oddelek za slovanske jezike in književnosti Filozofske fakultete Univerze v Ljubljani, 2001.

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

Jana Zemljarič Miklavčič. Jezikovne tehnologije – viri in pripomočki za učenje slovenščine kot tujega jezika. *Slovenščina in njeni uporabniki v luči evropske integracije*. Ur. Vesna Mikolič, Karin Marc Bratina. Koper: Univerza na

Primorskem, Znanstveno-raziskovalno središče, Zgodovinsko društvo za južno Primorsko, Založba Annales, 2005: 253–259.

Jana Zemljarič Miklavčič. Korpus govorjene slovenščine. *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference 2006*. Ur. Tomaž Erjavec, Jerneja Žganec Gros. Ljubljana: Institut Jožef Stefan, 2006: 124–127.

Jana Zemljarič Miklavčič. Iskanje odgovorov na Vprašanja govorjenega jezika. *Jezik in slovstvo* 53/1 (2008a): 89–106.

Jana Zemljarič Miklavčič. *Govorni korpusi*. Ljubljana: Znanstvena založba Filozofske fakultete, Oddelek za prevajalstvo, 2008b.

Jana Zemljarič Miklavčič, Marko Stabej, Simon Krek, Ana Zwitter Vitez. Kaj in zakaj v referenčni govorni korpus slovenščine. *Obdobja 24: Infrastruktura slovenščine in slovenistike*. Ur. Marko Stabej. Ljubljana: Znanstvena založba Filozofske fakultete Univerze v Ljubljani, 2009: 423–428.

10 POVZETEK/SUMMARY

Korpusi usvajanja tujega jezika so računalniške zbirke besedil nedomačih govorcev jezika, ki usvajajo jezik. Uporabljajo se v teoretičnih raziskavah usvajanja tujega jezika in v praktičnih aplikacijah na učna gradiva, kot so slovarji, slovnice, učbeniki in računalniška orodja za učenje, oziroma pri delu v razredu. V knjigi so predstavljeni obstoječi tuji korpusi, med katerimi jih je največ za angleščino, čeprav v zadnjih letih intenzivno nastajajo tudi za ostale, večinoma evropske jezike. Pri oblikovanju korpusov usvajanja tujega jezika je posebna pozornost namenjena prenosniku, velikosti, ciljnemu in izhodiščnemu jeziku, tvorcem in specifikam v korpus vključenih besedil ter kontrolnemu korpusu. Nekateri vidiki procesa oblikovanja so ilustrirani z opisom oblikovanja manjšega poskusnega korpusa usvajanja slovenščine kot tujega jezika, sledi pa mu predlog zasnove uravnoteženega, splošno uporabnega korpusa usvajanja slovenščine kot tujega jezika (KUST). Poglobljen premislek v njem zahteva predvsem označevanje napak, specifično za korpuse usvajanja tujih jezikov. Potekati mora ročno in temeljiti na čim bolj objektivni klasifikaciji napak. Predlagana dvostopenjska klasifikacija izhaja iz izkušenj obstoječih korpusov in raziskav, gradiva poskusnega korpusa in ankete o označevanju napak med slovenisti, njena ustreznost pa je že bila preverjena na gradivu domačih govorcev slovenščine v korpusu Šolar.

Learner corpora are electronic collections of texts produced by non-native speakers who are learning a certain language. Such corpora are used in theoretical research into foreign language acquisition and as a valuable source of information for the preparation of learning materials, such as grammar books, textbooks and computer tools, as well as for practical use in the classroom. The book gives an overview of existing learner corpora for various languages, the majority of which have been compiled for English, although work on other, mostly European languages has increased in recent years. In the chapter on the design of learner corpora, particular attention is paid to the communication mode, corpus size, the first and the target language, as well as the authors, the specific nature of the texts included in the learner corpus and the control corpus. The key aspects of corpus compilation are addressed in the chapter describing the compilation of a pilot Slovene learner corpus, which is followed by a framework proposal for a balanced and universally applicable learner corpus of Slovene, called KUST. One of the most important features of learner corpora is error tagging, which requires careful consideration in the compilation of the KUST corpus as it has to be performed manually and should follow an error tagging scheme that is as objective as possible. The proposed two-level error classification is based on the experience of other similar projects, on the pilot corpus and on the survey of error tagging among Slovene language professionals, and has been additionally validated against material from Slovene native speakers in the Šolar corpus of student writing.

Translated by Darja Fišer

11 IMENSKO KAZALO

Abe, Mariko 35, 222
Abuhakema, Ghazi 45, 222
Ädel, Annelie 68, 222
Adolphs, Peter 234
Aldabe, I. 45, 222
Aldezabal, I. 222
Allan, Quentin Grant 37, 77, 222
Allwood, Jens 44, 60, 71, 197, 222
Aranazabe, M. 222
Arhar, Špela 15, 16, 51, 78, 79, 114, 174, 223
Arrieta, B. 222
Aston, Guy 212, 223, 241
Atkins, Sue 14, 49, 78, 93, 223
Atwell, Eric 32, 70, 223, 235
Axelsson, Margareta Westergreen 30, 62, 69, 72, 77, 223

Bajec, Boštjan 65, 228
Balažic Bulc, Tatjana 24, 45, 110, 127, 128, 148, 149, 170–172, 176, 185, 188, 190, 192, 204, 223, 237
Baloh, Barbara 20, 223
Ban, Janja 110, 223
Bartning, Inge 39
Berglund, Ylva 72
Beguš, Ana 235
Bigert, Johnny 42, 52, 147, 198, 224
Bonaventura, Patrizia 235
Bonšek, Uroš 110, 224
Botley, Simon 38, 224
Bowker, Lynne 81, 224
Boyd, Adriane 40, 68, 143, 197, 224
Breiteneder, Angelika 30, 224
Briški, Anja 46, 150, 151, 224
Brown, H. Douglas 116–118, 121, 123, 125–127, 197, 200, 224
Burt, Marina 116, 117, 119, 124, 128, 129, 133, 135, 154, 200, 226
Bušta, Jan 115, 132, 133, 147, 148, 224

Cheng, Winnie 37, 70, 71, 197, 225
Clear, Jeremy 223
Cobb, Thomas 34, 72, 203, 225
Connor, Ulla 33, 70, 197, 227, 229
Corder, S. Pit 21, 22, 115, 116, 119–121, 123–125, 163, 225
Cucchiarini, Catia 44, 62, 71, 225

Čermák, František 10, 15, 74, 132, 197, 217, 225
Červ, Gaja 16, 225

Dagneaux, Estelle 117, 120, 132, 135, 155, 164, 197, 225, 226
de Beaugrande, Robert 210, 224
Debrock, Mark 39
De Cock, Sylvie 28, 70, 226
Denness, Sharon 225
Díaz de Ilarraza, A. 222
Díez Bedmar, María Belén 31, 32, 222, 226
Dillah, Doreen 224
Dolenc, Danilo 102, 226
Driesen, Joris 225
Dukič, Davorin 235
Dulay, Heidi 116, 117, 119, 124, 128, 129, 133, 135, 154, 200, 226
Džeroski, Sašo 78–80, 226

Edge, Julian 116, 119, 123, 124, 227
Ellis, Rod 115, 117, 124–129, 134, 135, 160, 227
Erjavec, Tomaž 14–16, 53, 74, 75, 78–80, 159, 226, 227, 231, 233–235, 240, 242, 243

Fairclough, Norman 134, 227
Faraj, Reem 222
Farr, Fiona 56, 213, 236
Feldman, Anna 222
Ferbežar, Ina 18, 20–26, 65, 68, 82, 83, 93, 95–97, 99, 100, 103–105, 112, 128, 135, 149, 150, 170, 171, 174, 176, 177, 180, 185, 188, 190, 191, 195, 209, 227–229, 237
Flament-Boistrancourt, Danièle 39
Fišer, Darja 14, 196, 227, 230, 245
Fitzpatrick, Eileen 34, 222, 229
Flowerdew, Lynn 212, 214, 229
Frischherz, Bruno 40, 229

Gadsby, Adam 203, 206–208, 229
Gantar, Polona 207, 229
Gillard, Patrick 203, 206–208, 229
Gilquin, Gaëtanelle 201, 222, 226, 229
Godnič Vičič, Šarolta 32, 33, 229
Gojenola, K. 222
Gorjanc, Vojko 10, 14–17, 49–51, 74, 78–80, 129, 164, 196, 198–200, 206, 211, 213, 214, 223, 227, 230
Grad, Martin 33, 204, 230
Granger, Sylviane 11, 26–28, 38, 51–53, 56, 61, 62, 65, 67, 69, 70, 74, 75,

79, 84, 99, 116, 120, 126, 129, 130, 132, 134, 135, 138–140, 156, 157, 163, 164, 196, 197, 200–203, 206, 207, 210–212, 214, 222, 225, 226, 229, 231, 234–236
Groom, Nicholas 38, 237
Gui, Shichun 61, 146, 231

Hagen, Jon Erik 117, 231
Hakim, Faizal 224
Hammarberg, Björn 42
Hana, Jirka 40, 44, 73, 94, 143, 144, 164, 197, 198, 231
Harris, Kathryn 238
Hasselgren, Angela 30, 70, 231
Herron, Daniel 235
Hirci, Nataša 205, 206, 211, 212, 214, 231, 237
Hlaváčková, Dana 224
Hofland, Knut 240
Holozan, Peter 78, 80, 217, 227, 231, 232
Horváth, Józef 31, 73, 232
Howarth, Peter 223, 235
Hunston, Susan 10, 18, 74, 202, 206, 212, 232

Ilić, Milena 102, 232
Iljaš Petrovič, Uršula 102, 232
Ino, Marie 236
Isahara, Hitoshi 232
Izumi, Emi 35, 70, 75, 81, 130, 135, 144, 146, 232

Jakubíček, Miloš 224
James, Carl 116, 117, 119, 120, 122, 124, 125, 129, 133–135, 164, 173, 177, 185, 232
Jantunen, Jarmo Harri 43, 232
Jerman, Tanja 209, 232, 237
Jurjevič, Ana 17, 232
Jurko, Primož 211, 213, 230

Kačič, Zdravko 16, 232
Kalin, Katja 226, 232
Kann, Viggo 224
Kaszubski, Przemysław 203, 209, 233
Kennedy, Graeme 9, 10, 14–17, 58, 65, 79, 80, 197, 199, 205, 206, 211, 214, 233
Klimpfinger, Theresa 224
Knez, Mihaela 24, 209, 228, 233
Knutsson, Ola 224
Kobos, Zuzanna Krystyna 105, 233
Koderman, Miha 235
Kojiro, Asao 35
Kosem, Iztok 48, 112, 114, 133, 154, 157, 165, 166, 169, 172, 175, 177, 178, 182, 190, 194, 195, 223, 238
Kovačič, Irena 200, 233
Kranjc, Simona 191, 192, 233
Krapš Vodopivec, Irena 223
Krashen, Stephen 116, 117, 119, 124, 128, 129, 133, 135, 154, 200, 226
Krek, Simon 79, 206, 207, 227, 229, 232, 233, 243
Kroymann, Emil 234
Kuo, Chin-Hwa 242
Kure, Ana 46, 233

Lado, Robert 117
Lavrič, Marjana 237

Ledinek, Nina 79, 80, 223, 226, 227, 233
Leech, Geoffrey 18, 53, 58, 71, 74, 217, 218, 234
Lennon, Paul 118, 128, 234
Lin, Linda H. F. 36, 198, 234
Liu, Anne 242
Logar Berginc, Nataša 16, 17, 114, 230, 234
Lozano, Cristóbal 41
Lönneker, Birte 78, 79, 208, 234
Lu, Hui-Chuan 41, 234
Lüdeling, Anke 40, 62, 142, 234, 238

Majewski, Stefan 224
Maritxalar, M. 222
Markovič, Andreja 180, 228
Martín Úriz, Ana 32, 235
Mason, Oliver 199
Mauranen, Anna 29, 205, 235
McCrostie, James 203, 235
Mendikoetxea, Amaya 32, 197, 235
Menzel, Wolfgan 32, 235
Meunier, Fanny 80, 81, 130, 202, 226, 235
Meurer, Paul 47, 240
Meyer, Michael 240
Mikolič, Vesna 16, 20, 223, 235, 237, 239, 243
Milton, John 37, 199, 209, 210, 235
Minn, Danny 209, 236
Mishan, Freda 212, 213, 236
Mitchell, Rosamond 38
Miura, Shogo 36
Morton, Rachel 235
Može, Sara 223
Müller, Frank Henrik 238
Myles, Florence 32, 38, 39, 71, 236

Nakamura, Takahiro 236
Nemser, William 20, 236
Nesselhauf, Nadja 11, 35, 65, 197, 203, 205, 206, 209, 212–214, 236
Nicholls, Diane 129, 131, 132, 140, 155, 163, 164, 166, 197, 198, 208, 236

O'Donnell, Michael 235
O'Keeffe, Anne 56, 213, 236
Oronoz, M. 232
Ostler, Nicholas 223
Otegi, A. 222

Pajas, Petr 226
Pala, Karel 132, 133, 147, 148, 224, 236
Papp, Szilvia 222, 226
Paquot, Magali 203, 231
Pazzaglio, Silvio 41
Pearson, Jennifer 81, 224
Pérez-Paredes, Pascual 212, 236
Perini, Mirco 41
Petch-Tyson, Stephanie 226
Petric, Ivana 204, 236
Pirih Svetina, Nataša 11, 13, 18–26, 45, 82, 83, 93, 95, 96, 99, 105, 112, 117–120, 123–128, 135, 172, 174, 176, 181, 185, 188, 193, 204, 208, 209, 227, 228, 233, 237
Pisanski Peterlin, Agnes 205, 206, 211, 212, 214, 231, 237
Pitzl, Marie-Luise 224
Pivec, Matej 232
Pogačnik, Lidija 152, 191, 192, 237

Ponikvar, Andreja 209, 237
Povhe, Janja 226, 232
Požgaj Hadži, Vesna 24, 82, 128, 135, 149, 150, 171, 177, 185, 188, 190, 191, 195, 237
Pravec, Norma 18, 21, 28, 29, 31, 35, 62, 65, 69, 70, 72, 73, 75, 77, 132, 197–199, 209–211, 237
Precht, Kristen 225

Quifang, Wen 37, 237

Ragan, Peter H. 27, 133, 237
Randall, Mick 38, 237
Razpotnik, Barica 226, 232
Reder, Stephen 33, 132, 163, 197, 238
Richards, Jack C. 11, 225, 236, 238
Rigač, Simon 232
Rigler Šilc, Katarina 24, 237
Rollinson, Paul 31, 235
Rosen, Alexandr 231
Rozman, Katarina 45, 46, 115, 116, 120, 135, 151, 152, 168, 171–174, 176, 177, 183, 185, 186, 190, 191, 204, 208, 234, 238
Rozman, Simon 232
Rozman, Tadeja 21, 47, 48, 94, 112, 114, 133, 154, 157, 165, 166, 169, 175, 177, 178, 182, 190, 194, 195, 208, 238
Rychlý, Pavel 236

Sampson, G. P. 11, 238
Sanders, E. 225
Sano, Hiroshi 236
Schlamberger Brezar, Mojca 228
Seegmiller, M. S. 34, 229
Selinker, Larry 21, 123, 238
Setzler, Kristen 238
Sharwood Smith, Michael 10, 19, 21, 238
Shih, Rebecca Hsue-Hueh 37, 38, 65, 72, 77, 197, 238
Siemen, Peter 40, 238
Sjöberg, Ola 224
Smrž, Pavel 236
Souter, Clive 223, 235
Spina, Stefania 41
Stabej, Marko 9, 11–13, 15, 51, 93, 94, 99, 101, 111, 113, 154, 170, 178, 196, 218, 225, 227–229, 233, 234, 239, 242, 243
Starc, Sonja 152, 239
Stritar, Mojca 47, 48, 94, 112, 114, 133, 152, 154, 157, 165, 166, 169, 172, 175, 177, 182, 190, 194, 195, 209, 223, 238–240
Strömqvist, Sven 222
Sugiura, Masatoshi 36, 70, 127, 146, 240
Svenšek, Tanja 223

Šircelj, Milivoja 103, 110, 240
Škodová, Svatava 231
Špelko, Tina 110, 224
Šter, Darja 226, 232
Štindlová, Barbora 231
Šuster, Simon 17, 114, 234

Tagnin, Stella E. O. 45, 197, 240
Tan, Melinda 202, 203, 240
Tenfjord, Kari 42, 43, 47, 52, 60, 62, 70, 71, 73, 133, 142, 155, 163, 184, 240

Teubert, Wolfgang 206, 240
Thewissen, Jennifer 120, 132, 135, 138, 156, 164, 231
Titscher, Stefan 134, 240
Tivadar, Hotimir 228
Tono, Yukio 11, 17, 34–38, 51, 60, 68–70, 116, 133, 201, 202, 222, 241
Toporišič, Jože 179–181, 186, 189, 191, 241
Tribble, Chris 212, 214, 230
Tsao, Nai-lung 242
Tzimokas, Dimitrios 45, 241

Uchimoto, Kiyotaka 232
Upton, Thomas 225, 229
Uria Garin, Larraitz 45, 222, 241
Uzar, Rafal S. 31, 49, 62, 70, 198, 241

Van hamme, Hugo 225
Velušček, Aleš 232
Verdonik, Darinka 93, 241
Vetter, Eva 240
Vidovič Muha, Ada 179, 242
Vilret, Xavier 226
Vintar, Špela 15–17, 56, 196, 205, 207, 230, 242
Vitez, Primož 93, 94, 111, 239, 241
Voionmaa, Kaarlo 222
Volodina, Elena 42, 198, 242

Walter, Maik 234
Warren, Martin 37, 70, 71, 197, 225
Wible, David 211, 242
Wodak, Ruth 240
Wu, Szu-Chia 234

Yang, Huizhong 61, 146, 231

Zavrel, Jakub 226
Zazijal, Aleksandra 105, 242
Zemljarič Miklavčič, Jana 15, 20, 93, 94, 101, 209, 228, 242, 243
Zwitter Vitez, Ana 223, 243

Žabokrtský, Zdenek 226
Žele, Andreja 226
Žnidaršič, Tina 226, 232

12 KAZALO TABEL

Tabela 1: Razlike med usvajanjem prvega in tujega jezika . . . 22
Tabela 2: Globalna lestvica jezikovnega znanja SEJO 25
Tabela 3: Prvi jeziki tvorcev v korpusu usvajanja slovenščine po državi izvora tvorcev 47
Tabela 4: Število besed v korpusih usvajanja tujega jezika glede na prenosnik 54
Tabela 5: Izhodiščni in ciljni jeziki v korpusih usvajanja z enim izhodiščnim jezikom 59
Tabela 6: Izhodiščni in ciljni jeziki v korpusih usvajanja tujega jezika z enim ciljnim jezikom 60
Tabela 7: Pregled podatkov o tvorcih sestavnih besedil 61
Tabela 8: Pregled tem v obstoječih pisnih korpusih 67
Tabela 9: Prvi jeziki v PiKUST-u 82
Tabela 10: Stopnje znanja v PiKUST-u 83
Tabela 11: Vrsta in tema besedil v PiKUST-u 85
Tabela 12: Šifre prvih jezikov govorcev v imenu dokumentov v PiKUST-u 87
Tabela 13: Število in delež kandidatov na izpitih iz znanja slovenščine po posameznih ravneh 99
Tabela 14: Tuje prebivalstvo v Sloveniji v letih 2005–2011 . . .102
Tabela 15: Materni jezik prebivalstva Slovenije po popisu leta 2002 103
Tabela 16: Narodnosti udeležencev tečajev slovenščine 2001–2007 106
Tabela 17: Prvi jeziki udeležencev tečajev slovenščine 108
Tabela 18: Primerjava najpogostejših prvih jezikov po posameznih virih 109
Tabela 19: Predlagano zajemanje besedil v KUST-u 112
Tabela 20: Okvir za raziskovanje usvajanja tujega jezika 115
Tabela 21: Vrste napak glede na način pridobivanja jezikovne zmožnosti 124
Tabela 22: Klasifikacija napak s psiholingvističnim izvorom . .126

Tabela 23: Korpusi usvajanja tujega jezika s številom besed z označenimi napakami 131
Tabela 24: Tipi napak v korpusu ICLE 138
Tabela 25: Tipi napak v korpusu FRIDA 139
Tabela 26: Tipi napak v korpusu ASK 141
Tabela 27: Tipi napak v korpusu FALKO 142
Tabela 28: Primeri napak v korpusu EAGLE 143
Tabela 29: Tipi napak v korpusu SST 146
Tabela 30: Oznake primera napake v korpusu Japanese Learners' Corpus 146
Tabela 31: Prvotna klasifikacija napak v korpusu Chyby 147
Tabela 32: Popravljena klasifikacija napak v korpusu Chyby . . .148
Tabela 33: Napake srbohrvaških govorcev 149
Tabela 34: Napake hrvaško, srbsko in bosansko govorečih kandidatov na ustnem izpitu iz znanja slovenščine na osnovni ravni 150
Tabela 35: Napake v govoru srbsko, bosansko in črnogorsko govorečih osnovnošolcev 151
Tabela 36: Napake nemško govorečih študentov slovenščine . . .152
Tabela 37: Klasifikacija napak v PiKUST-u 155
Tabela 38: Odgovori ankete med slovenisti o besedni vrsti napačne oblike 158
Tabela 39: Oblikoslovne oznake v PiKUST-u 159
Tabela 40: Predlagane pravilne oblike iz ankete med slovenisti . .162
Tabela 41: Porazdelitev kategorij napak v PiKUST-u 168
Tabela 42: Pregled napak črkovanja v PiKUST-u 171
Tabela 43: Odgovori iz ankete med slovenisti o primeru napake zapisa predloga *z/s* 172
Tabela 44: Odgovori iz ankete med slovenisti o primeru napake zapisa/lapsusa 173
Tabela 45: Odgovori iz ankete med slovenisti o primerih napak besedišča 177
Tabela 46: Pregled napak besedišča v PiKUST-u 179
Tabela 47: Odgovori iz ankete med slovenisti o primerih napak besedišča/besedotvorja 180
Tabela 48: Primeri napak besedišča zaradi rabe neustreznega zaimka 182
Tabela 49: Odgovori iz ankete med slovenisti o primeru napake zaradi rabe neustrezne besedne vrste 183
Tabela 50: Pregled napak oblike v PiKUST-u 186
Tabela 51: Odgovori iz ankete med slovenisti o primeru napake zaradi rabe napačnega sklona 187
Tabela 52: Odgovori iz ankete med slovenisti o primerih napak skladnje in besednega reda 190
Tabela 53: Odgovori iz ankete med slovenisti o primeru napake izpusta 192

13 KAZALO SLIK

Slika 1: Vmesni jezik 21
Slika 2: Merila oblikovanja korpusa usvajanja tujega jezika 51
Slika 3: Lastnosti besedil v korpusu ICLE 52
Slika 4: Razmerje med številom besed v pisnih in govornih delih obstoječih korpusov usvajanja tujega jezika 53
Slika 5: Velikosti obstoječih korpusov usvajanja tujega jezika . . . 55
Slika 6: Dolžina besedil v pisnih korpusih usvajanja tujega jezika . 57
Slika 7: Slike, na podlagi katerih informatorji za korpus LINDSEI pripovedujejo zgodbo 66
Slika 8: Strip, ki ga je treba v zgodbi s sto besedami opisati za korpus VALICO 66
Slika 9: Razmerje med številom besed z izpitov in s tečajev v PiKUST-u 82
Slika 10: Starost tvorcev v PiKUST-u 84
Slika 11: Obrazec za vnos v bazi podatkov za PiKUST 86
Slika 12: Okolje urejevalnika Oxygen XML Editor 88
Slika 13: Kriteriji in značilnosti KUST-a 92
Slika 14: Število kandidatov na izpitih iz znanja slovenščine po posameznih ravneh 100
Slika 15: Deleži začetnikov, nadaljevalcev in izpopolnjevalcev na tečajih slovenščine 100
Slika 16: Najpogostejši materni jeziki (poleg slovenščine) prebivalstva Slovenije ob popisu leta 2002 104
Slika 17: Število kandidatov na izpitih iz znanja slovenščine . . .104
Slika 18: Deset najpogostejših prvih jezikov udeležencev tečajev slovenščine 108
Slika 19: Zasnova KUST-a z enakomernimi deleži med jeziki . . .111
Slika 20: Dvojna zgradba KUST-a 114
Slika 21: Algoritem za opisovanje vmesnega jezika 121
Slika 22: Algoritem analize napak 122
Slika 23: Delitev napak glede na nastanek v korpusu Japanese Learners' Corpus 127
Slika 24: Označevanje neustreznih predlogov v korpusu CLC . . .140
Slika 25: Klasifikacija napak v učnem KUST-u 153
Slika 26: Deleži napak v PiKUST-u po jezikovnih ravninah 167
Slika 27: Postopek za razvrščanje napak v PiKUST-u 169
Slika 28: Integrirani kontrastivni model 201
Slika 29: Okno za iskanje konkordanc v programu WordPilot . . .210

14 KAZALO PRIMEROV

Primer 1: Obrazec za korpuse ICLE, LINDSEI in FRIDA, ki ga izpolni učenec 63
Primer 2: Obrazec o tvorcu besedila v korpusu CEDEL2 64
Primer 3: Privolitveni obrazec za korpus USE 72
Primer 4: Glava dokumenta v korpusu CEJL 76
Primer 5: Glava dokumenta in označeno besedilo v korpusu LPELC 77
Primer 6: Izpis primerov v LCLE s tvorčevo nacionalnostjo in stopnjo usvajanja 78
Primer 7: Glava besedila v PiKUST-u 90
Primer 8: Besedilo z oznakami v PiKUST-u 91
Primer 9: Del transkribiranega besedila iz korpusa ELFA 94
Primer 10: Tema za pisanje spisa v vzorčnem testu iz slovenščine na srednji ravni 96
Primer 11: Naravoslovna tema za pisanje spisa v vzorčnem testu na visoki ravni 96
Primer 12: Družboslovna tema za pisanje spisa v vzorčnem testu na visoki ravni 97
Primer 13: Navodilo za pisanje praktičnega besedila v vzorčnem testu na srednji ravni 98
Primer 14: Navodilo za pisanje praktičnega besedila v vzorčnem testu na visoki ravni 98
Primer 15: Oznaka napake v korpusu ICLE 132
Primer 16: Oznaka nepričakovane oblike v korpusu MAELC . .132
Primer 17: Izpis napak v poljskem podkorpusu ICLE 138
Primer 18: Večplastno označevanje napak v korpusu FALKO . .142
Primer 19: Stavek z označenimi napakami v korpusu AKCES .144
Primer 20: Stavek z označeno napako v korpusu SST 145
Primer 21: Napaka z oznakami v korpusu Japanese Learners' Corpus 146
Primer 22: Napaka v PiKUST-u z več možnostmi interpretacije .155
Primer 23: Oznaka napake v PiKUST-u jeziku XML 156
Primer 24: Napaka v korpusu ICLE z oznako pred besedo, ki sproži napako 156
Primer 25: Mesto oznake napake v PiKUST-u 156
Primer 26: Oznaka napake v Šolarju 157
Primer 27: Napaka v PiKUST-u brez navedenega popravka . . .164
Primer 28: Oznaka sekundarne napake v PiKUST-u 165

Primer 29: Oznaka sekundarne napake v korpusu Šolar 165
Primer 30: Oznaka dvojne napake besede in oblike v PiKUST-u .166
Primer 31: Oznaka dvojne napake besede in zapisa v PiKUST-u 166
Primer 32: Oznaka dvojne napake oblike in zapisa v PiKUST-u .166
Primer 33: Oznaka dvojne napake zapisa v PiKUST-u 166
Primer 34: Trojna napaka v PiKUST-u 167
Primer 35: Povezana napaka črkovanja 170
Primer 36: Povezana napaka ločila 170
Primer 37: Napaka besedišča zaradi neustrezne črke v besedi . .174
Primer 38: Napaka zaradi uporabe podobno zveneče besede . . .174
Primer 39: Napaka zaradi neustrezne rabe male začetnice 175
Primer 40: Napaka zaradi neustrezne rabe velike začetnice . . .175
Primer 41: Napaka zaradi manjkajoče vejice 176
Primer 42: Napaka zaradi neustreznega končnega ločila 176
Primer 43: Napaka zaradi odvečnega ločila 176
Primer 44: Napaka besedišča z nesporno interpretacijo 178
Primer 45: Napaka besedišča s težavno interpretacijo 178
Primer 46: Napaka besedišča zaradi neustrezne osnove besede .181
Primer 47: Napaka besedišča zaradi neustrezne glagolske predpone 181
Primer 48: Označevanje neustreznih oziralnih zaimkov v PiKUST-u 182
Primer 49: Označevanje neustreznih oziralnih zaimkov v Šolarju 182
Primer 50: Napaka besedišča zaradi rabe neustrezne besedne vrste 183
Primer 51: Napaka besedišča zaradi rabe *ni* namesto *ne* 183
Primer 52: Napaka besedišča zaradi rabe *ne* namesto *ni* 183
Primer 53: Napaka besedišča zaradi rabe *si* namesto *se* 183
Primer 54: Napaka besedišča z neustrezno rabljenim frazemom 184
Primer 55: Napaka besedišča z neustreznim konektorjem 184
Primer 56: Sekundarna napaka skladnje zaradi napake besedišča 184
Primer 57: Raba pogovornega izraza v PiKUST-u 184
Primer 58: Sekundarna napaka samostalnika zaradi neustreznega predloga 185
Primer 59: Napaka oblike 186
Primer 60: Napaka oblike zaradi neustreznega sklona 187
Primer 61: Napaka oblike zaradi uporabe v slovenščini neobstoječe končnice 187
Primer 62: Napaka oblike, pri kateri se končnici samostalnika in pridevnika v besedni zvezi ne ujemata 187
Primer 63: Napaka oblike, ki bi bila lahko tudi napaka besedišča 189

Primer 64: Povezana napaka oblike pri samostalniku in pridevniku 189
Primer 65: Povezana napaka oblike pri samostalniku in predlogu 189
Primer 66: Napaka oblike zaradi rabe neustreznega predloga in sklona 189
Primer 67: Povezana napaka oblike pri glagolu 189
Primer 68: Napaka skladnje zaradi dela besedila, napisanega v tujem jeziku 190
Primer 69: Sekundarna napaka besednega reda zaradi popravka neustreznega veznika 191
Primer 70: Sekundarna napaka odvečne besede zaradi popravka neustreznega glagola 191
Primer 71: Oznaka napake besednega reda v PiKUST-u 192
Primer 72: Oznaka napake izpusta v PiKUST-u 193
Primer 73: Napaka izpusta zaradi izpuščenega deležnika glagola *biti* 193
Primer 74: Napaki izpusta zaradi izpuščenega osebnega zaimka in morfema *se* 193
Primer 75: Napaka izpusta zaradi izpuščenega členka *ne* . . . 193
Primer 76: Napaka izpusta zaradi drugih izpuščenih besed . . 193
Primer 77: Odvečni glagol *biti* 194
Primer 78: Odvečni *se/si* 194
Primer 79: Odvečni oziralni zaimek 194
Primer 80: Odvečni osebni zaimek 194
Primer 81: Odvečni kazalni zaimek 194
Primer 82: Odvečni veznik 194
Primer 83: Napaka strukture zaradi prenosa iz tvorčevega prvega jezika 195
Primer 84: Napaka strukture brez jasnega popravka 195
Primer 85: Napaka strukture brez jasnega popravka 195
Primer 86: Napaka strukture brez jasnega popravka 195
Primer 87: Neustrezna struktura znotraj dvojne napake v PiKUST-u 195
Primer 88: Opozorilo v slovarju *Longman Essential Activator*, ki je nastalo na podlagi korpusa LCLE 207

15 O AVTORICI

Mojca Stritar, rojena leta 1979, je na Filozofski fakulteti v Ljubljani diplomirala iz slovenščine in primerjalne književnosti. Takoj po diplomi je začela poučevati slovenščino kot tuji jezik na Centru za slovenščino kot drugi/tuji jezik, s čimer se še vedno ukvarja, leta 2009 pa je s tega področja doktorirala z disertacijo Korpus usvajanja slovenščine kot tujega jezika. Raziskovalno in aplikativno se ukvarja predvsem s slovenščino kot tujim jezikom in z jezikovnimi tehnologijami, pa tudi z različnimi sociolingvističnimi vidiki jezikovnega položaja na Slovenskem. Izpopolnjevala se je na univerzi v Bergnu na Norveškem. Od leta 2011 je v uredništvu znanstvene revije *Jezik in slovstvo*. Je soavtorica več učbenikov in priročnikov za slovenščino kot tuji jezik (*Slovenska beseda v živo 3a* in *3b*, *Pot do izpita iz znanja slovenščine*, *Žepna slovenščina*), dejavna pa je tudi na področju gorniške in potopisne literature.

16 IZ RECENZIJ

Avtorica je raziskovalno posegla na dve pomembni področji sodobnega slovenističnega raziskovanja, na korpusno jezikoslovje in na preučevanje slovenščine kot tujega jezika. Cilj monografije je zasnova uravnoteženega, splošneje uporabnega korpusa usvajanja slovenščine kot tujega jezika, torej računalniške zbirke besedil govorcev, za katere slovenščina ni prvi jezik. Avtorica je v ta namen po posameznih kriterijih ovrednotila lastnosti več kot 33 obstoječih tujih korpusov usvajanja, naslonila pa se je tudi na manjša lastna korpusa, ki ju je oblikovala kot poligon za preučevanje možnosti za urejanje korpusa in označevanje napak. V monografiji so zajeti vsi koraki oblikovanja korpusa, posebna pozornost pa je posvečena prenosniku jezikovne produkcije, velikosti korpusa, ciljnemu in izhodiščnemu jeziku, tvorcem besedil, vrsti besedil in temi ter kontrolnemu korpusu. Na podlagi vseh virov in nekaterih specifik položaja slovenščine kot tujega jezika avtorica predlaga priporočila za oblikovanje polmilijonskega pisnega korpusa usvajanja slovenščine kot tujega jezika, nakazane pa so tudi možnosti, prednosti in omejitve uporabe tovrstnih korpusov.

Drugi del monografije se osredotoča na ključno lastnost, po kateri se korpusi usvajanja ločijo od nespecializiranih korpusov, to je označevanje napak v jezikovni produkciji. Po splošni razpravi o razvoju metodologije določanja in pojmovanja napak avtorica predstavi proces analize napak ter predlaga klasifikacijo za korpus usvajanja slovenščine kot tujega jezika. Ta temelji na izkušnjah obstoječih korpusov,

rešitvah v slovenistični znanstveni literaturi, označevanju gradiva v avtoričinem poskusnem korpusu in ankete o napakah v jezikovni produkciji, izvedene med slovenisti. Predlagana klasifikacija za korpus usvajanja slovenščine kot tujega jezika temelji na načelih formalnosti, splošnosti, objektivnosti, opisnosti in čim manjše interpretativnosti. Sistem označevanja napak je oblikovan na dveh ravneh, to je glede na jezikovno ravnino in glede na podrobnejši opis. Z dodatnimi oblikoslovnimi oznakami in označevanjem popravljenih oblik je dovolj robusten in informativen, da zagotavlja tako izvedljivost kot vsebinsko uporabnost.

Delo ima široko potencialno bralstvo, ki sega od slovenističnih jezikoslovcev do številne publike učiteljev in učiteljev slovenščine v osnovnih in srednjih šolah (saj je npr. poglobljen razmislek o razvrstitvi in tipologiji napak pri vrednotenju pisne produkcije dragocen in inovativen tudi s stališča slovenščine kot prvega jezika), pa tudi drugih ljudi, ki se ukvarjajo s šolanjem in poučevanjem v slovenščini. Še posebej dragocena pa je monografija za specialno področje slovenščine kot drugega/tujega jezika, kjer glede na potrebe opažamo zadnja leta primanjkljaj temeljnih znanstvenih del, strokovna publika pa se iz leta v leto veča.

Čeprav se na prvi pogled zdi znanstvena monografija *Korpusi usvajanja tujega jezika* izrazito ozko specializirano delo, natančno branje komunikativnega besedila razkrije drugačno resnico. Ob zelo prepričljivem in izdatnem definiranju samega področja in pregledu mednarodne produkcije korpusov ter deloma tudi njihove uporabe avtorica ponuja uporabne rešitve ali vsaj izhodišča za vrsto nadaljnjih empiričnih raziskav slovenskega jezikovnega prostora, v poglavju o označevanju in interpretaciji t. i. napak pa temeljito nadgrajuje doslejšnja spoznanja in ponuja rešitve, ki bodo prišle prav veliko širši populaciji ljudi, ki se ukvarjajo s poučevanjem in testiranjem jezikovne zmožnosti.

red. prof. dr. Marko Stabej,
Oddelek za slovenistiko
Filozofske fakultete Univerze v Ljubljani

Mojca Stritar s svojo monografijo *Korpusi usvajanja tujega jezika* v slovenskem prostoru razvija področje korpusnega jezikoslovja, in sicer v tistem segmentu, ki je bil do sedaj slabo razvit, tj. v segmentu razvoja korpusa za raziskave usvajanja slovenščine kot tujega jezika. Jasna umestitev na korpusno jezikoslovno raziskovalno področje z namenom njegovega razvoja za slovenščino je usmeritev, ki jo lahko samo pozdravimo, še posebej zato, ker je namen njenega dela oblikovati celovita merila za oblikovanje in označevanje korpusa usvajanja slovenščine kot tujega jezika. Ker z načrti oblikovanja jezikovne infrastrukture postavlja temelje vsem gradivno usmerjenim jezikoslovnim raziskavam s področja slovenščine kot tujega jezika, pomeni njeno delo izjemno dragocen prispevek za celotno slovenistično jezikoslovje in jezikoslovje v slovenskem prostoru sploh.

Avtoričine analize posameznih parametrov, relevantnih za gradnjo korpusa usvajanja slovenščine kot tujega jezika, tvorijo celovit načrt za gradnjo tovrstnega korpusa slovenščine. Čeprav je samo oblikovanje korpusa slovenščine kot tujega jezika tehnično zahtevna naloga, pa se zdi najzahtevnejši del pri izdelavi korpusa usvajanja jezika njegovo označevanje na ravni označevanja napak. Temu je posvečeno celotno peto poglavje avtoričine monografije. Avtorica opravi analizo označevanja v tujejezičnih korpusih usvajanja jezika; kljub različnim pristopom pa je pomembno, da je avtoričino izhodišče analize in opisovanja napak v pojmovanju napake kot kreativne produkcije učečih se, torej kot normalnega dela učnega procesa. Pri kategorizaciji napak je avtorica smiselno kombinirala rešitve pri obstoječih tujejezičnih korpusih, ankete med slovenisti-jezikoslovci in diskusije s strokovnjaki s področja slovenščine kot tujega jezika. Čeprav so rešitve tudi zaradi kompleksnosti napak diskutabilne, pa je hkrati pomembno, da se avtorica zaveda potrebe po pragmatičnih odločitvah, ko gre za določanje označevalskih kategorij, saj le tak pristop omogoča tudi izgradnjo korpusa. Opisane kategorije napak so bile uporabljene v njenem poskusnem korpusu, delo na poskusnem korpusu pa je potrdilo njihovo primernost pri označevanju bodočega korpusa usvajanja slovenščine kot tujega jezika.

izr. prof. dr. Vojko Gorjanc,
Oddelek za prevajalstvo
Filozofske fakultete Univerze v Ljubljani

17 DOSLEJ IZŠLO

V *Slavistični knjižnici* so doslej izšle naslednje publikacije:

Slovenska slovnica za tretji in četrti razred srednjih in sorodnih šol. Sestavili: dr. Anton Breznik, dr. Anton Bajec, dr. Rudolf Kolarič, dr. Mirko Rupel, Anton Sovre in Jakob Šolar. Ljubljana: SD v Ljubljani, 1940. [Izven zbirke.]

Silva Trdina. *Besedna umetnost, 2: Literarna teorija*. Ljubljana: MK, 1958, 1961, 1965 (SK: učbeniki, 1).

Janko Jurančič. *Južnoslovanski jeziki*. Ljubljana: DZS, 1957 (SK). 118 str.

Vera Brnčič. *N. A. Dobroljubov*. Ljubljana: DZS, 1956 (SK). 96 str.

France Dobrovoljc in Zvone Verstovšek. *Dve imenski kazali* [Imensko kazalo k Cankarjevim *Zbranim spisom* od I. do XXI. zvezka. Imensko kazalo k Prijateljevi monografiji *Janko Kersnik, njegovo delo in doba*]. Ljubljana: DZS, 1955 (SK. Bibliografska dela, 1). 79 str.

Marja Boršnik. *Kratek bibliografski pregled slovenskega slovstva*. Ljubljana: DZS, 1955 (SK. Bibliografska dela, 2). 101 str.

Alenka Gložančev. *Imena podjetij kot jezikovnokulturno vprašanje*. Ljubljana: Rokus, 2000 (SK, 3). 167 str.

Zoltan Jan. *Poznavanje slovenske književnosti v Italiji po letu 1945*. Ljubljana: Rokus in ZDSDS, 2001 (SK, 4). 269 str.

Zoltan Jan. *Cankar, Kosovel, Zlobec in Ljubka Šorli pri Italijanih / Bibliografski dodatek*: [Slovenska književnost pri Italijanih po drugi svetovni vojni]. Ljubljana: Rokus in ZDSDS, 2001 (SK, 5). 153 str.

Marija Mercina. *Proza Cirila Kosmača: Uvod v lingvostilistično analizo*. Ljubljana: Rokus in ZDSDS, 2003 (SK, 6). 175 str.

Drago Unuk. *Zlog v slovenskem jeziku*. Ljubljana: Rokus in ZDSDS, 2003 (SK, 7). 324 str.

Peter Jurgec. *Samoglasniški nizi v slovenščini: Fonološko-fonetična analiza*. Ljubljana: Rokus, 2005 (SK, 8). 224 str.

Kozma Ahačič. *Izvirne slovenske pesmi Jovana Vesela Koseskega: Vsebinski opis, okoliščine nastanka, pregled ocen, jezik, tekstnokritični aparat ter diplomatični in kritični prepis pesmi 1818–1852*. Ljubljana: ZDSDS, 2006 (SK, 9). 195 str.

Martina Ožbot. *Prevajalske strategije in vprašanje koherence ob slovenskih prevodih Machiavellijevega* Vladarja. Ljubljana: ZDSDS, 2006 (SK, 10). 164 str.

Katarina Podbevšek. *Govorna interpretacija literarnih besedil v pedagoški in umetniški praksi*. Ljubljana: ZDSDS, 2007 (SK, 11). 305 str.

Barbara Pregelj. *Zgledno omledno: Trivialno v slovenski postmoderni književnosti*. Ljubljana: ZDSDS, 2007 (SK, 12). 210 str.

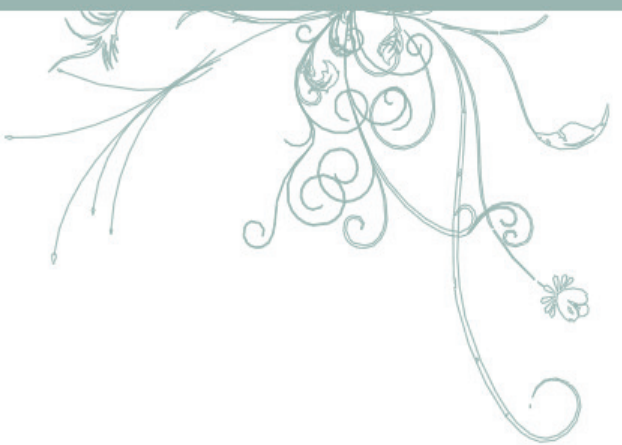
Alenka Jensterle - Doležal. *V krogu mitov: O ženski in smrti v slovenski književnosti*. Ljubljana: ZDSDS, 2008 (SK, 13). 147 str.

Petra Kodre. *Vzgoja za krepitev zavesti o narodni pripadnosti pri pouku književnosti v 20. stoletju*. Ljubljana: ZDSDS, 2008 (SK, 14). 170 str.

Mateja Pezdirc Bartol. *Najdeni pomeni: Empirične raziskave recepcije literarnega dela*. Ljubljana: ZDSDS, 2010 (SK, 15). 280 str.

Urška Perenič. *Empirično-sistemsko raziskovanje literature: Konceptualne podlage, teoretski modeli in uporabni primeri*. Ljubljana: ZDSDS, 2010 (SK, 16). 240 str.

Zveza
društev
Slavistično
društvo
Slovenije



ISBN 9789616715102

Cena: 15 EUR