Korpus študentov prevajanja MetaTrans

Anja Tavčar, Ines Čeligoj Pregelj, Miha Pompe

Oddelek za prevajalstvo, Filozofska fakulteta, Univerza v Ljubljani, Aškerčeva 2, 1000 Ljubljana
anja.tavcar09@gmail.com, ines_celigoj@hotmail.com, miha.pompe@gmail.com

Povzetek

V pričujočem prispevku sta predstavljeni gradnja in analiza korpusa prevodnih napak študentov Oddelka za prevajalstvo Filozofske fakultete Univerze v Ljubljani. V korpusu so označene napake v študentskih prevodih poljudnoznanstvenih naravoslovnih in družboslovnih besedil iz slovenskega v angleški jezik ter njihovi popravki, ki jih je prispeval materni govorec angleškega jezika in profesor prevajanja. Napakam je pripisana tudi vrsta napake, za kar je bila uporabljena mednarodna tipologija napak v prevodih Mellange. Označevanje je potekalo s spletnim anotacijskim orodjem WebAnno. Za ugotavljanje relevantnosti rezultatov je bila narejena analiza dvojnega označevanja. Najpogostejši vir napak v prevodih je neustrezen slog, analiza dvojno označenega vzorca besedil pa kaže, da je ujemanje obeh označevalcev pri označevanju napak majhno.

The building and analysis of the corpus of translation errors made by students of the Department of Translation at the Faculty of Arts in Ljubljana are presented in this paper. The corpus is comprised of labelled errors in popular natural and social science texts translated by students from Slovene to English and the corrections made by a natural speaker of English language and professor of translation. Errors are labelled by type using the Mellange international translation mistake typology. The annotation was done with the online annotation tool WebAnno. The most common mistake in translations is inappropriate style, while the analysis of dual-labelled sample texts shows that there is little accord between different annotationists.

1 Uvod

Korpusi danes predstavljajo nepogrešljiv vir jezikovnih podatkov za jezikoslovne opise, raziskave in utemeljitve. Jezikovni podatki, ki jih dajejo korpusi, omogočajo v jeziku ločevanje med tipičnim in posebnim oz. individualnim, torej prepoznavanje osrednjih in obrobnih jezikovnih pojavov (Gorjanc et al., 2005). Za slovenščino je na voljo že veliko različnih vrst korpusov, vendar pa tako kot večina drugih jezikov še vedno nimamo na voljo reprezentativnega korpusa, ki bi služil kot pomoč pri študiju prevajalstva in bil sestavljen iz izvirnikov in njihovih prevodov ter obogaten s popravki in oznakami vrst napak, ki so jih naredili študenti med prevajanjem. Takšni korpusi nedvomno pripomorejo k poučevanju prevajalcev, saj na jasen način prikazujejo tipične jezikovne težave, ki jih imajo študenti pri prevajanju v tuji jezik, kar je še posebej pomembno v okoljih, v katerih je prevajanje v tuji jezik pogosta prevodna dejavnost. Tako okolje je tudi slovensko.

To omogoča pripravo kakovostnih didaktičnih smernic in avtentičnega gradiva za izobraževanje prevajalcev. Poleg tega pa so korpusi prevajalskih napak v veliko pomoč tudi študentom prevajalstva in tujih jezikov, saj jim nudijo nepogrešljive informacije o (ne)ustreznosti prevajalskih in jezikovnih rešitev, ki so jim v pomoč pri prevajanju.

V okviru projekta »Korpus študentov prevajanja MetaTrans«,[1] katerega cilj je bil zapolniti praznino na tem področju za slovenski jezik, je bil izdelan korpus, ki vsebuje angleške prevode slovenskih poljudnoznanstvenih besedil, ki so jih prevajali študenti 2. stopnje Oddelka za prevajalstvo na Univerzi v Ljubljani, in popravke profesorja, ki je materni govorec angleškega jezika. V skladu z mednarodno tipologijo napak v prevodih Mellange je popravkom pripisana tudi vrsta napake, ki so jo naredili študentje med prevajanjem v tuji jezik.

Čeprav sta v okviru diplomske naloge in doktorske disertacije na Oddelku za prevajalstvo na Univerzi v Ljubljani že potekali podobni raziskavi (Lavrič, 2009; Dobnik, 2011), je dodana vrednost našega korpusa ta, da smo v projekt vključili analizo dvojnega označevanja napak, ki omogoča vpogled v zanesljivost označevanja korpusa. Tako prispevek vsebuje tudi analizo in diskusijo napak v korpusu ter analizo in diskusijo medsebojnega ujemanja med označevalcema.

2 Sorodne raziskave

Eden najodmevnejših tovrstnih projektov je *MeLLANGE oz. Multilingual eLearning in LANGuage Engineering* (Castagnoli et al., 2011), katerega cilj je bil ustvariti večjezični označeni in poravnan korpus prevodov, ki vključuje prevode študentov prevajalstva in strokovnjakov s tega področja. Korpus, poimenovan Learner Translator Corpus (LTC), vsebuje prevode v desetih jezikih in je poleg tega tudi označen z oznakami o besednih vrstah in s podatki o lemi, poleg tega pa vključuje popravke in oznake napak, ki dajejo podatke o vrsti napake. Posebnost korpusa je tudi ta, da je bila za potrebe označevanja napak izdelana Mellangeva tipologija napak.[2] Korpus vsebuje metapodatke o prevajalcih besedil, saj so te informacije lektorjem, ki so prispevali popravke za sestavo korpusa, pomagale pri povezovanju napak z okoliščinami, v katerih je bilo besedilo prevedeno. Pri izdelavi korpusa je sodelovalo 440 študentov in profesionalnih prevajalcev, ki so prevajali 4 vrste besedil – pravna, tehnična, poslovna in novinarska. Ob zaključku projekta je bilo vseh prevedenih besedil 429, označenih pa 360. Projekt je bil zaključen leta 2007, končna analiza pa je pokazala, da so najpogostejše

[1] Raziskava je nastala v okviru magistrskega modula Korpusi in baze podatkov v štud. l. 2015/2016 pod mentorskim vodstvom doc. dr. Darje Fišer. Članek je bil oblikovan v okviru predmeta Slovensko strokovno besedilo, prav tako v štud. l. 2015/2016, pri prof. dr. Vojku Gorjancu.

[2] http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

napake pri prevajanju besedil uporaba napačnega termina, popačenje vsebine izvirnika in neskladje terminologije znotraj ciljnega besedila.

Mellangevo tipologijo napak pa si je za osnovo pri določanju napak v svojem korpusu izbral tudi študent oddelka za prevajalstvo, Davorin Lavrič, ki je leta 2009 v svoji diplomski nalogi izdelal korpus študentskih prevodov, ki ga sestavlja 122 slovenskih besedil, od tega 89 prevodov v angleški jezik. Besedila v korpusu so različno dolga in pokrivajo različna področja, popravke pa so prispevali profesorji z Oddelka za prevajalstvo Univerze v Ljubljani (Lavrič, 2009).

V Sloveniji je leta 2011 na Oddelku za prevajalstvo na Univerzi v Ljubljani potekal še en podoben projekt. V okviru doktorske disertacije je potekala gradnja korpusa študentskih napak (Dobnik, 2011), ki vsebuje prevode šestih besedil, ki so jih prevajali študenti 2. in 3. letnika dodiplomske stopnje pri predmetu Prevajanje iz francoščine v slovenščino. Vse popravke so prispevali profesorji, žal pa korpus ni označen z istimi oznakami, kar onemogoča medsebojno primerljivost rezultatov.

Podoben projekt je potekal na desetih ruskih univerzah, kjer je bil izdelan korpus *Russian Learner Translator Corpus* oziroma *RusLTC* (Kutuzov in Kunilovskaya, 2014), ki vsebuje prevode in napake v prevodih študentov prevajalstva. Študenti so prevajali v jezikovni kombinaciji ruščina-angleščina in angleščina-ruščina. Vsi prevodi so bili narejeni kot del izpita, prevajalskih tekmovanj ali kot del domače naloge. Marca 2014 je korpus RusTLC vseboval skupno skoraj 1,2 milijona besed, 258 izvirnih besedil in 1.795 prevodov, od tega pa je bilo z napakami označenih 198 prevodov v ruski jezik in 43 prevodov v angleški jezik, označevanje pa še ni bilo zaključeno (Kutuzov in Kunilovskaya, 2014).

Korpus prevajalskih napak je leta 2005 začel nastajati še v Španiji na Univerzi v Zaragozi. Korpus ENTRAD (Serrando in Sanz, 2008) vsebuje 45 angleških besedil, ki so jih v španščino prevedli študenti pri predmetu Uvod v prevajanje angleških besedil. Materni jezik študentov je večinoma španščina in francoščina, popravljavci oziroma označevalci napak pa so bili univerzitetni profesorji. Nekatere poizvedbe v korpusu so opremljene tudi z metapodatki, kot so prevajalčeva starost, spol in materni jezik. Oznake napak niso strojno berljive in temeljijo na barvni kodi in grafičnih oznakah.

Korpus prevajalskih napak je nastal tudi na univerzi Pompeu Fabra v Barceloni. *Korpus LTC-UPF* (Espunya, 2014) vsebuje besedila, ki so jih študenti prevajali iz angleščine v katalonščino. Korpus vključuje 10 izvirnih besedil in 194 prevodov. Označevanje korpusa je potekalo ročno in samodejno. Korpus se lahko uporablja za identifikacijo pogostih napak pri prevodih učencev in za analizo njihovih vzorcev jezikovne rabe. Korpus omogoča enostaven dostop do vzorcev napak in do več različic istega izvirnega besedila, kar omogoča kvalitetnejše izobraževanje prevajalcev.

Izgradnja korpusa z označenimi prevajalskimi napakami je med letoma 2009 in 2013 potekala v nemškem izobraževalnem prostoru, in sicer na univerzi Universität des Saarlandes. Zbiranje besedil za korpus *KOPTE* (Wurm, 2013) se je začelo leta 2009 pri predmetu prevajanje iz francoščine v nemščino. Večina besedil je vzetih iz časopisnih člankov, ki so jih študenti prevajali pri pouku. Za večino besedil obstaja več prevodov, saj so študenti med poukom prevajali tudi besedila z enakim izvirnikom. Korpus vsebuje več kot 77 izvirnikov in več kot 971 prevodov v nemščino. Označevanje prevajalskih napak je potekalo ročno z orodjem *UAM Corpus Tool*. V korpusu so z orodjem *TreeTagger* pripisane tudi oblikoskladenjske oznake. Za ročno označevanje napak je pobudnik izdelave korpusa KOPTE, Andrea Wurm, izdelal lastno tipologijo napak.[3] Tudi korpus KOPTE vsebuje metapodatke, kot so podatki o jeziku, ki so se ga študenti učili, podatki o času in načinu študija, podatki o poreklu staršev idr.

Na Poljskem je leta 2001 potekal projekt *PELCRA* oziroma *Polish and English Language Corpora for Research and Applications* (Uzar, 2002), v okviru katerega je bila na univerzi Uniwersytet Łódzki narejena pilotna študija, v okviru katere je bil narejen manjši korpus (15.000 besed). Korpus vsebuje prevode 180 besedil iz poljskega jezika (L1) v angleški jezik (L2). Korpus vsebuje besedila, ki so jih prevedli študenti na Oddelku za anglistiko. Od 180 besedil je bilo glede na kvaliteto besedila, berljivost besedila in kvaliteto prevoda z napakami označenih 50 besedil. Vse povedi v korpusu pa so bile označene s +, - ali 0, kar označuje, ali je prevedena poved ustrezno, neustrezno ali dokaj ustrezno prevedena glede na izvirnik.

Nekoliko drugačno strategijo od zgoraj naštetih so na portugalskem inštitutu Instituto de Engenharia de Sistemas e Computadores uporabili pri gradnji angleško-portugalskega korpusa (Costa et al., 2014), ki vsebuje oznake napak prevodov 150 povedi, ki so bile vzete iz različnih področij in prevedene v portugalščino s strojnim prevajalnikom Google Translate ali Moses, napake v njem pa je označil materni govorec portugalščine.

Podoben projekt je potekal na oddelkih za prevajalstvo na dveh univerzah v Franciji (Universite Paris Sud, Universite Paris Diderot), kjer je 46 študentov popravljalo strojne prevode, ki so jih nato označili še glede na vrsto napak (Wisniewski et al., 2014). Število vseh povedi v korpusu je 4.854, tip napake pa je pripisan skoraj polovici. Vsi dokumenti so prevedeni iz angleščine v francoščino. Vse naloge so nato pregledali še profesorji.

Omeniti velja, da v slovenskem prostoru že imamo dva nepogrešljiva korpusa z označenimi popravki in tipi napak, in sicer korpusa *Šolar* in *Lektor*, vendar sta oba enojezična, torej namenjena učenju slovenskega jezika. *Šolar* (Rozman et al., 2010) je korpus besedil, ki so jih učenci iz slovenskih osnovnih in srednjih šol samostojno tvorili pri pouku. Korpus, ki vsebuje 967.477 besed, je prvi tovrstni korpus besedil v Sloveniji, ki je narejen po vzoru korpusov usvajanja jezikov. Vsi jezikovni popravki v korpusu so delo učiteljev. Korpus *Lektor* (Popič, 2013) pa je zbirka lektoriranih slovenskih avtorskih besedil, ki vsebuje skoraj milijon besed, celoten korpus pa vsebuje 30.258 lektorskih popravkov, ki so bili ročno označeni.

3 Označevanje korpusa

3.1 Izbor besedil

V začetni fazi nastajanja korpusa smo s slovenskega spletnega portala Meta znanost izbrali slovenska besedila in njihove angleške prevode, ki jih je v angleščino v okviru projekta Slovenska znanost gre v svet prevedlo 9 študentov magistrskega programa prevajanja na Oddelku za

[3] http://fr46.uni-saarland.de/index.php?id=3702

prevajalstvo Univerze v Ljubljani. Izvirna besedila so delo slovenskih raziskovalcev in znanstvenikov. Ker smo želeli, da bi bil korpus čim bolj reprezentativen, smo izbrali 30 poljudnoznanstvenih člankov, 15 s področja naravoslovnih znanosti in 15 s področja družboslovja. Angleški del korpusa vsebuje 2.544 povedi.

3.2 Tipologija napak

V naslednji fazi smo v prevodih študentov označili popravke, ki jih je prispeval materni govorec angleščine izr. prof. dr. David Limon z Oddelka za prevajalstvo Univerze v Ljubljani. Popravljanje je potekalo na dveh nivojih. Najprej smo v vseh besedilih označili popravke, nato pa smo popravkom pripisali tudi vrsto napake s pomočjo tipologije napak, ki je bila razvita v projektu *MeLLANGE*.[4] Za uporabo tipologije Mellange smo se odločili, ker menimo, da je za primerljivost rezultatov pomembno, da se različne korpuse z označenimi tipi prevajalskih napak da primerjati, Mellangeva tipologija pa je že bila uspešno preizkušena na drugih sorodnih projektih.

Tipologija Mellange je zasnovana kot hierarhična shema, ki v osnovi temelji na razlikovanju med vsebinskimi napakami in jezikovnimi napakami. Tipologija Mellange vsebuje 39 oznak napak, ki so razvrščene v kategorije in podkategorije. Kategorije tipologije so v celoti predstavljene v nadaljevanju članka v Tabeli 1, v razdelku 4.1.

Vse podkategorije vsebujejo različne vrste napak, kot na primer preveč dobesedno, neustrezen termin, neustrezna raba ločil, neustrezen glagolski čas, neustrezen predlog, napačna raba velike/male začetnice, prevedeno preveč svobodno, preveden neprevedljiv izraz idr. Vsaka vrsta napake je označena s kodo, ki bo v korpusu stala poleg popravka. V okviru projekta smo celotno tipologijo lokalizirali v slovenščino, da bo dostopna tudi za slovenske raziskovalce. Tipologija vključuje uporabniško definirane kategorije, ki jih uporabnik izbere v primeru, da tipologija ne vsebuje vrste napake, ki jo ta želi označiti.

3.3 Označevanje napak

Ker je bilo vseh besedil 30, smo si delo razdelili tako, da je vsak od treh označevalcev označil 10 besedil v skladu s popravki maternega govorca. Pri delu smo za označevanje napak uporabljali orodje za jezikoslovno označevanje korpusov *WebAnno* (Yimam et al., 2013), ki je eno izmed najuniverzalnejših spletnih orodij za označevanje korpusov. Orodje omogoča projektno in oddaljeno delo in ne zahteva posebnih programerskih znanj, zaradi česar je zelo uporabno za jezikoslovce, humaniste, družboslovce in študente. Ena izmed glavnih značilnosti orodja je njegova prilagodljivost, saj omogoča označevanje napak na več nivojih, poljuben nabor oznak in različne načine označevanja. Prav tako pa omogoča, da na istem projektu dela več označevalcev, ki lahko vzporedno označujejo ista besedila, na podlagi česar se lahko ugotavlja tudi skladnost med ocenjevalci (ang. *inter-annotation agreement*).

Skladnost med ocenjevalci smo določali tudi pri našem projektu, saj je to zelo pomemben podatek za ugotavljanje, kako težavna je naloga in ali so bile odločitve označevalcev pri označevanju zanesljive. Označevanje prevajalskih napak je izredno problematičen in zamuden postopek, saj se mnenja o tem, za kakšno vrsto napake gre, med označevalci pogosto razlikujejo.

Vseh besedil, ki smo jih dvojno označili, je 8 (4 naravoslovna in 4 družboslovna). Analiza ugotovitev je predstavljena v 4. razdelku, 5. razdelek pa vsebuje analizo označevanja vseh 30 besedil, ki jih korpus vsebuje, in podatke o najpogostejših/najredkejših napakah študentov, količini napak glede na področje, prevajalca in na posamezne prispevke, kjer so označevalci opazili posebnosti.

4 Analiza besedil

Korpus označenih napak zajema 30 različno dolgih besedil 9 različnih prevajalcev, in sicer 15 besedil s področja družboslovja ter 15 s področja naravoslovja. Angleški del korpusa vsebuje 2.544 povedi. Vseh označenih napak je 2.458. Od teh se jih 1.546 oziroma 63 % pojavlja v besedilih s področja družboslovja, 939 oziroma 37 % pa v besedilih s področja naravoslovja.

Besedila s področja družboslovja imajo največ jezikovnih napak, ki niso podrobneje razdelane in spadajo v kategorijo drugo (234 oz. 15 %), sledijo slogovne napake (214 oz. 14 %) in napake v skladnji (192 oz. 12 %). Pri besedilih s področja naravoslovja pa je največ slogovnih napak (175 oz. 19 %), sledijo napake v kategoriji Jezik – drugo (145 oz. 15 %) ter napake v skladnji (68 oz. 7 %).

4.1 Analiza glede na tipologijo napak

V tabeli 1 so navedene kategorije napak in njihovo število od največjega do najmanjšega.

Napaka	Število napak
Jezik - neustrezen slog	389
Jezik - drugo	379
Jezik - skladnja	260
Jezik - terminologija - nepravilen pomen	163
Jezik - ločila	145
Prenos pomena - pomensko neustrezen	142
Jezik - napačna kolokacija	136
Jezik - neustrezen glagolski čas	125
Prenos pomena - preveč dobesedno	121
Jezik - napačen predlog	107
Jezik - napačna začetnica	96
Jezik - napačno število	72
Prenos pomena - dodano	39
Jezik - terminologija - drugo	36
Prenos pomena - izpust	31
Jezik - neprimerno za tip besedila	31
Jezik - istorečje	30
Jezik - črkovanje	26

[4] http://corpus.leeds.ac.uk/mellange/images/mellange_error_typology_en.jpg

Prenos pomena - preveč svobodno	25
Jezik - nesorodna beseda	23
Jezik - termin, preveden z neterminološkim izrazom	23
Prenos pomena - nejasno	20
Jezik - slog - drugo	17
Prenos pomena - drugo	11
Jezik - skloni in ujemanje - drugo	8
Jezik - terminologija - nekonsistentno znotraj ciljnega prevoda	7
Jezik - neskladno z glosarjem	7
Prenos pomena - prevedeni neprevedljivi izrazi (lastna imena ipd.)	4
Prenos pomena - poseganje v ciljni jezik - drugo	3
Jezik - nedosledno z izvirnim jezikom	3
Jezik - register - nekonsistentno znotraj ciljnega prevoda	3
Jezik - naglas ali diakritično znamenje	1
Jezik - register - drugo	1

Tabela 1: Kategorije napak tipologije Mellange

Kot je razvidno iz tabele 1, je največ napak zaradi neustreznega sloga (389 oz. 16 %), že pri drugi kategoriji (379 oz. 15 %) pa pridemo do težave, saj Mellangeva tipologija napak ni zajemala napačne rabe določnega in nedoločnega člena, zaradi česar so vse takšne napake znotraj kategorije jezikovnih napak označene kot »drugo«, kamor so se uvrstile tudi druge jezikovne napake, zato bi bila dobrodošla nekoliko bolj razdelana tipologija. Sledijo napake v skladnji (260 oz. 10 %) ter neustrezno izbiranje terminologije (163 oz. 7 %). Študentom prevajanja nemalo napak povzročajo tudi ločila (145 oz. 6 %): v večini primerov gre tu za nepravilno postavljanje vejic. Napake, ki se tičejo prenosa pomena v ciljni jezik, so šele na šestem mestu (142 oz. 6 %).

4.2 Razlikovanje med označevalci

Pri pregledu označenih napak je opaziti, da so se označevalci različno odločali za označevanje napak, kar kaže na to, da uporabljena tipologija ni najbolj jasna. Označevalec 1 ima tako na primer 298 napak označenih kot slogovne napake, medtem ko imata v tej kategoriji druga dva označevalca skupaj takšnih napak označenih le 91. Podobno tendenco opazimo pri označevalcu 2, ki je kot Prenos pomena – dodano označil 24 od skupno 39 tako označenih napak ter 18 od skupno 31 napak v kategoriji napak Prenos pomena – izpust in 144 od skupno 260 tako označenih napak v kategoriji Jezik – skladnja.
Seveda je pri tem potrebno omeniti, da so se označevalci na začetku projektnega dela zaradi narave in namena raziskave dogovorili le katero tipologijo bodo uporabljali, pri tem pa se niso dogovarjali o natančnejših smernicah ali konkretnejših primerih vrst napak. S tem smo želeli določiti, kakšna je stopnja ujemanja brez predhodnega dogovarjanja, ki bi sicer gotovo povišal stopnjo ujemanja med vsemi označevalci.

5 Analiza ujemanja med označevalcema

Poleg splošne analize označenih napak smo izvedli tudi analizo dvojnega označevanja napak, s katero smo želeli ugotoviti, kako skladni so označevalci napak, saj ti podatki vplivajo na verodostojnost rezultatov končne splošne analize in uporabnost korpusa.
Dvojno označen podkorpus zajema 8 besedil (4 naravoslovna, 4 družboslovna) od skupno 30 besedil iz korpusa za splošno analizo, kar znaša 27 % celotnega korpusa. Podkorpus vsebuje 443 povedi oz. 9.415 pojavnic.

Analizo smo izvedli na nivoju tipologije napak, in sicer tako, da je kurator s pomočjo programa WebAnno pregledal dvojno označena besedila in v primerih, kjer se označevalca nista strinjala, izbral ustreznejšo od predlaganih možnosti oz. predlagal svojo. Med različno označevanje se je štelo tudi drugačno označevanje napake v besedilu, četudi sta ji označevalca pripisala isto vrsto napake.

Analiza je pokazala, da je skupno število analiziranih napak v 443 stavkih 511, od tega sta enako označeni 102 napaki (20 %), pri 409 primerih (80 %) v 216 (49 %) stavkih pa sta označevalca napako označila različno.

Analiza neujemanj med označevalci glede na prevajalca besedila ni pokazala nobenih posebnosti, zato v nadaljevanju predstavljamo samo rezultate analize glede na področja besedila, kjer je prihajalo do večjih razlik. Stopnja ujemanja med označevalcema pri naravoslovnih besedilih je bila 7%, pri družboslovnih pa 13%. Število različno označenih napak pri naravoslovnih besedilih je bilo 167 (33 %), pri družboslovnih pa 242 (47 %).

Besedilo		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Skupaj**
Neujemanje na nivoju fraze		7	8	16	9	7	14	23	20	104 (25 %)
Neujemanje na nivoju tipa napak	Na vrhnjem nivoju	13	2	31	20	3	8	20	7	104 (25 %)
	Na srednjem nivoju	14	10	4	30	12	7	17	20	114 (28 %)
	Na spodnjem nivoju	3	6	18	18	5	5	9	23	87 (21 %)
skupaj		37 (9 %)	26 (6 %)	69 (17 %)	77 (19 %)	27 (6 %)	34 (8 %)	69 (17 %)	70 (17 %)	409

Tabela 2: Analiza področij ujemanja med označevalcema

Razlika v številu vseh napak med besedili je bila 22 %, kar pomeni, da sta bila označevalca pri označevanju naravoslovnih besedil bolj skladna.

5.1 Analiza neujemanj med označevalcema

Podrobneje smo analizirali, do kakšnih razhajanj med označevalcema prihaja najpogosteje, izdelali tipologijo napak in rezultate ovrednotili.

Iz tabele 2 je razvidno, da gre pri 25 % napak za neujemanje na nivoju fraze, pri ostalih 75 % pa za neujemanje na nivoju tipa napak. Največ neujemanj se je pojavilo na srednjem nivoju (28 %), najmanj na spodnjem nivoju (21 %).

V nadaljevanju analize se natančneje posvetimo najpogostejšim tipom razhajanja med označevalcema.

5.2 Neujemanje na nivoju fraze

Kot smo že omenili, za neujemanje štejemo tudi napake, ki sta jim označevalca pripisala isto vrsto napake, vendar jih je program zaradi različne označitve v besedilu zaznal kot neujemanje. Kot ilustrativni primer tovrstnega neujemanja bi lahko predstavili npr. označevanje manjkajočih vejic, kjer je eden od označevalcev označil besedo pred in za manjkajočo vejico, drugi pa le besedo pred ali po mestu manjkajoče vejice. Takšnih napak je 104 (25 %), torej dobra četrtina vseh napak. Če to upoštevamo pri številu enako označenih napak, se stopnja ujemanja med označevalcema zviša za 20 %, kar je bistveno izboljšanje rezultata.

5.3 Neujemanje na nivoju tipa napak

Področje analize neujemanja smo razdelili na tri dele, kot je vidno v tabeli 2:

- **Neujemanje na vrhnjem nivoju tipologije**

Sem štejemo oznake napak, ki so se razhajale na nivoju kategorij prenosa vsebine in na nivoju jezika. Gre za največje razhajanje med označevalci in napake z največjo težo, ki najbolj negativno vplivajo na uporabno vrednost korpusa. Takšnih odstopanj pri označevanju napak se je v besedilih pojavilo 104 (25 %). Pri analizi te vrste neujemanj nismo odkrili nobenih tipičnih vzorcev razhajanj, ki bi se pogosteje pojavljali, zaradi česar tudi ni mogoče predlagati možnosti za izboljšave.

Kot primer tovrstnega neujemanja pri označevanju lahko predstavimo naslednji popravek; v enem od besedil je bila beseda *diploma* prevedena kot *Bachelor thesis* in popravljena na *undergraduate dissertation*. Omenjeni popravek je eden od označevalcev označil kot napako na nivoju jezika – *pomensko neustrezno,* drugi pa kot napako na nivoju prenosa vsebine – *nepravilno (neskladno z izvirnim jezikom)* iz podkategorije terminologija in leksika.

- **Neujemanje na srednjem nivoju tipologije**

Gre za napake, katerih oznaki se razlikujeta glede na nivo znotraj prvih dveh večjih skupin napak, npr. med kategorijo *register* in *slog*. Tovrstnih razhajanj je bilo 114 (28 %).

Opazili nismo nobenih vrst napak, ki bi se pogosteje pojavljale, vendar je očitno, da bi jih bilo mogoče zmanjšati z jasnimi smernicami za prepoznavanje tovrstnih tipov napak.

Primer tovrstnega popravka je prevod besede ok. Prevedena je bila kot agreed in popravljena na okay. Prvi označevalec je popravek označil z napako nepravilno iz podkategorije terminologija in leksika, drugi pa kot neprimerno za tip besedila iz podkategorije register. Obe podkategoriji sodita v kategorijo napak na nivoju jezika.

- **Neujemanje na spodnjem nivoju tipologije**

Pri tovrstnem neujemanju napak opažamo neujemanje znotraj najožjih kategorij tipologije, npr. znotraj kategorije *slog* ali *higiena*. Razhajanj na tem nivoju je bilo nekoliko manj, pojavilo se je 87 napak (22 %). Opazili smo, da se na tem nivoju najpogosteje pojavljajo tri kombinacije napačno označenih napak (vse znotraj kategorije *terminologija in leksika*), in sicer: *nesorodna beseda* in *napačna kolokacija* (17 napak, 19 %), *napačen termin* in *nesorodna beseda* (11 napak, 13 %) ter *napačen termin* in *napačna kolokacija* (10 napak, 11 %).

Primer tovrstnega neujemanja pri označevanju je na primer prevod besede *sem,* ki je bil iz oblike *I am* popravljen na *I'm.* Eden od popravljalcev je popravek označil kot *neprimerno za tip besedila*, drugi pa kot *drugo.* Oba popravka sodita v podkategorijo *register.*

Analiza dvojnega označevanja napak je pokazala, da je odstotek nekonsistence med označevalcema precej visok, kar negativno vpliva na uporabno vrednost korpusa, zato ne moremo trditi, da je korpus v tej različici že zanesljiv. Za izboljšanje njegove uporabne vrednosti bi potrebovali natančne smernice za označevanje napak, s katerimi bi se izognili neujemanjem zaradi različnega označevanja besedila (dogovoriti bi se bilo potrebno, da se označuje npr. le prvo besedo v popravku) in odstopanja pri izbiri tipa napake na najvišjem nivoju (potreben je dogovor, kdaj gre za napačen prevod vsebine npr. *preveč dobesedno* in kdaj za jezikovno napako npr. *nerodno* v kategoriji *slog*).

6 Zaključek

Predstavljena raziskava je pilotna; njen cilj je bil izdelati zasnovo za korpus ter preizkusiti orodje, metodologijo in tipologijo za označevanje napak. V prihodnjih raziskavah je korpus potrebno povečati ter razširiti na druge tipe besedil, na prevode besedil študentov drugih jezikovnih kombinacij in drugih nivojev študija. Predvsem pa je potrebno izboljšati metodologijo označevanja in izdelati jasne smernice, da bomo izboljšali ujemanje med označevalci, zaradi česar bo korpus kvalitetnejši in bolj uporaben. Poleg vseh teh izboljšav bi se lahko poslužili tudi rešitev, ki so bile uporabljene na primer v korpusu Šolar, kjer so se odločili, da nekatere napake označijo dvojno, interpretacije le-teh pa prepustili raziskovalcem. Kljub začetnim težavam in veliko večjim neujemanjem med označevalci napak, kot smo pričakovali, smo s pomočjo izkušenj in rezultatov pridobili veliko koristnih izsledkov, ki so nam pri nadaljnjem delu s tovrstnimi korpusi v pomoč. Vsekakor pa takšen korpus odpira veliko število možnosti raziskovanja prevodnih napak študentov in lahko služi kot orodje pri prilagajanju učnega procesa prevajanja.

7 Literatura

Ana Espunya. 2014. *The UPF learner translation corpus as a resource for translator training*. V: Language Resources and Evaluation, Volume 48, Issue 1, str. 33–43, New York. http://dl.acm.org/citation.cfm?id=2598668

Andrea Wurm. 2013. *Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE).* V: Journal of Translation and Technical Communication Research (trans-kom), zvezek 6 (2). http://www.trans-kom.eu/bd06nr02/trans-kom_06_02_06_Wurm_Eigennamen.20131212.pdf

Andrei Popescu-Belis, Margaret King in Houcine Benantar. 2002. *Towards a corpus of corrected human translations.* V: Machine translation evaluation : human evaluators meet automated metrics (LREC 2002), Third International Conference on Language Resources and Evaluation, str. 17–21, Pariz. http://www.mt-archive.info/LREC-2002-Popescu-Belis.pdf

Andrey Kutuzov in Maria Kunilovskaya. 2014. Russian Learner Translator Corpus: Design, Research Potential and Applications. V: Text, Speech and Dialogue: 17th International Conference (TSD 2014), str. 315–324, Brno. http://rus-ltc.org/references/tsd_rusltc_lnai-libre.pdf

Angela Costa, Tiago Luís in Luisa Coheur. 2014. *Translation errors from English to Portuguese: an annotated corpus.* V: Proceedings of the ninth international conference on language resources and evaluation (LREC'14), str. 1231–1234, Reykjavik. http://www.lrec conf.org/proceedings/lrec2014/pdf/199_Paper.pdf

Celia Floren Serrando in Rosa Lores Sanz. 2008. *The Application of a Parallel Coprus (English-Spanish) to the Teaching of Translation (ENTRAD PROJECT).* V: Micaela Muñoz-Calvo, Carmen Buesa-Gómez, M. Ángeles Ruiz-Moneva, ur., New Trends in Translation and Cultural Identity, str. 433–445. Cambridge Scholar Publishing, Velika Britanija.

Claudio Fantinuoli in Federico Zanettini. 2015. *New directions in corpus-based translation studies (Translation and Multilingual Natural Language Processing*. Language Science Press. Berlin.

Damjan Popič. 2013. *Je etično popravljati prevode*? V: Etika v slovenskem jeziku, literaturi in kulturi: zbornik predavanj, str. 118–123, Ljubljana.

Davorin Lavrič. 2009. *Vzporedni korpus študentskih prevodov*. Diplomska naloga. Filozofska fakulteta, Univerza v Ljubljani.

Elizaveta Kuzmenko in Andrey Kutuzov. 2014. *Russian Error-Annotated Learner English Corpus: a Tool for Computer-Assisted Language Learning.* V: Proceedings of the third workshop on NLP for computer-assisted language learning (SLTC 2014), str. 87–97, Uppsala University, Švedska. http://www.ep.liu.se/ecp/107/ecp14107.pdf

Guillaume Wisniewski, Natalie Kubler in François Yvon. 2014. *A Corpus of Machine Translation Errors Extracted from Translation Students Exercises.* V: International Conference on Language Resources and Evaluation, str. 3585–3588, Reykjavik. https://transread.limsi.fr/lrec_Wisniewskietal.pdf

Nadja Dobnik. 2011. *Analiza napak v prevodih študentov v funkciji načrtovanja in razvijanja predmetov francoskega jezika v okviru študijskega programa prevajalstva*. Doktorsko delo, Filozofska fakulteta, Univerza v Ljubljani.

Rafal S. Uzar. 2002. *A Corpus Methodology for Analysing Translation.* V: Stella Esther Ortweiler, ur., Cadernos de Tradução, str. 237–265. Lisboa.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler in Alexandra Volanschi. 2011. *Designing a Learner Translator Corpus for Training Purposes.* V: Natalie Kübler, ur., Corpora, Language, Teaching and Resources: from Theory to Practice, str. 221–248. Peter Lang. Berlin. http://www.eila.univ-paris-diderot.fr/_media/user/alexandra_volanschi/publi/castagnoli_et_al.pdf

Seid Muhie Yimam and Iryna Gurevych, Richard Eckart de Castilho in Chris Biemann. 2013. WebAnno: *A Flexible,Web-based and Visually Supported System for Distributed Annotations.* V: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations), str. 1–6, Sofija. http://www.aclweb.org/anthology/P13-4001.pdf

Tadeja Rozman, Mojca Stritar in Iztok Kosem. 2010. *Korpus šolskih pisnih izdelkov Šolar. V: Nova didaktika poučevanja slovenskega jezika*: Sporazumevanje v slovenskem jeziku, str. 4–35. Ljubljana.

Vojko Gorjanc, Simon Krek in Polona Gantar. 2005. *Slovenska leksikalna podatkovna zbirka*. Jezik in slovstvo, L/2: 3–19.

.