KORPUS ŠTUDENTSKIH PREVODOV METATRANS

INES ČELIGOJ PREGELJ, MIHA POMPE, ANJA TAVČAR

UVOD

- Gradnja in analiza korpusa prevodnih napak študentov Oddelka za prevajalstvo Filozofske fakultete Univerze v Ljubljani
- Napake v študentskih prevodih poljudnoznanstvenih naravoslovnih in družboslovnih besedil iz slovenskega v angleški jezik
- Popravki, ki jih je prispeval materni govorec angleškega jezika in profesor prevajanja
- Vrsta napak (mednarodna tipologija napak v prevodih Mellange)
- Orodje WebAnno
- Analiza napak
- Analiza ujemanja med označevalci

- Slovenščina nima reprezentativnega prevajalskega korpusa (didaktičnost, pomoč študentom)
- V okviru projekta izdelan korpus -> zapolniti praznino na tem področju
- Že potekali raziskavi (Lavrič, 2009; Dobnik, 2011)
- Dodana vrednost: dvojno označevanje -> analiza razhajanja

SORODNE RAZISKAVE

- **MeLLANGE (Multilingual eLearning in LANGuage Engineering)**
- **Learner Translator Corpus (LTC):**

- vsebuje prevode v desetih jezikih
- označen z oznakami o besednih vrstah in s podatki o lemi
- vključuje popravke in oznake napak, ki dajejo podatke o vrsti napake.
- 440 študentov in profesionalnih prevajalcev
- vseh prevedenih besedil: 429, označenih: 360 (2007)
- posebnost korpusa: Mellangeva tipologija napak
- najpogostejše napake: uporaba napačnega termina, popačenje vsebine izvirnika in neskladje terminologije znotraj ciljnega besedila

SORODNE RAZISKAVE

- **Tujina:** Russian Learner Translator Corpus (Rusija), Korpus ENTRAD (Španija), Korpus LTC-UPF (Španija), KOPTE (Nemčija), PELCRA (Poljska), ...

- **Slovenija:**

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

- 122 slovenskih besedil, od tega 89 prevodov v angleški jezik
- Mellangeva tipologija napak
- Besedila so različno dolga
- različna področja
- Besedila in njihove popravke so prispevali profesorji z Oddelka za prevajalstvo Univerze v Ljubljani

SORODNE RAZISKAVE

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

- Gradnja korpusa študentskih napak
- Prevodi šestih besedil
- Študenti 2. in 3. letnika dodiplomske stopnje
- Predmeti: Prevajanje iz Francoščine v slovenščino.
- Popravki: profesorji

OZNAČEVANJE KORPUSA

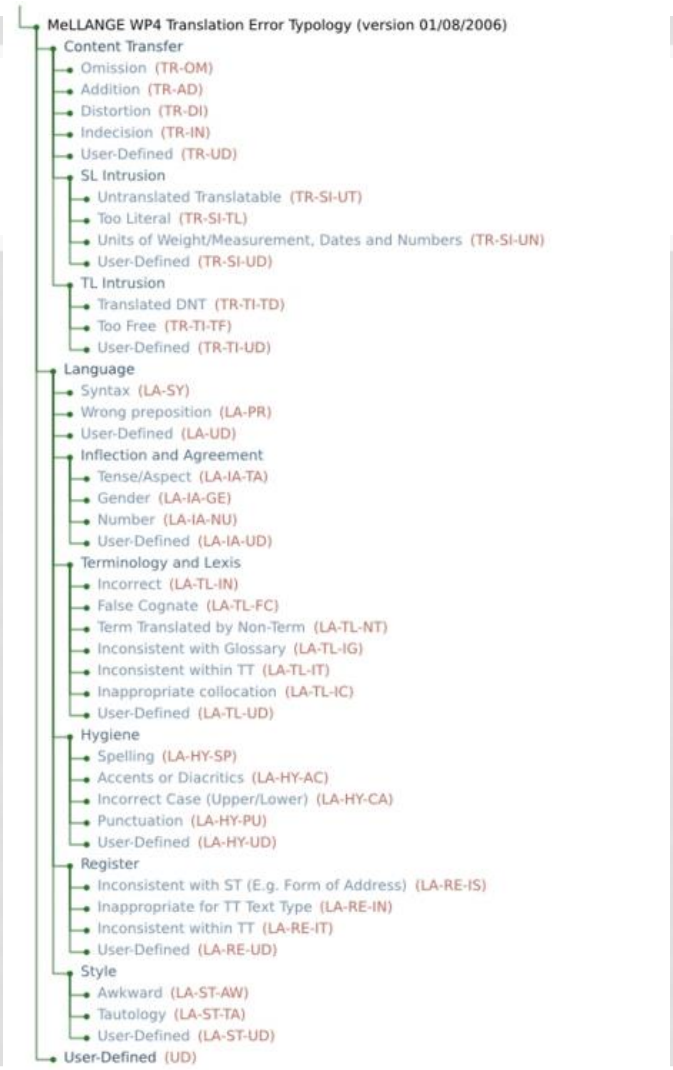
1. **Izbor besedil**
- Spletni portal Meta znanost
- 30 poljudnoznanstvenih člankov (15 s področja naravoslovnih znanosti in 15 s področja družboslovja)
- Besedila popravil izr. prof. dr. David Limon
- 2.544 povedi

2. **Označevanje popravkov**
- Popravljanje je potekalo na dveh nivojih

I. Označevanje popravkov
II. Označevanje vrste napak (Mellangeva tipologija napak); popravki

MELLANGEVA TIPOLOGIJA NAPAK

- Hierarhična shema, 3 ravni
- Temelji na razlikovanju med vsebinskimi napakami in jezikovnimi napakami
- 39 oznak napak
- Vsaka vrsta napake označena s kodo
- Celotna tipologija lokalizirana tudi v slovenski jezik
- Tipologija vključuje tudi uporabniško definirane kategorije



OZNAČEVANJE NAPAK

- 3 označevalci (3 × 10 besedil)
- WebAnno

→Orodje za jezikoslovno označevanje korpusov

→Projektno in oddaljeno delo

→Ne zahteva posebnih programerskih znanj,

→Prilagodljivost (označevanje napak na več nivojih, poljuben nabor oznak in različne načine označevanja)

→Vzporedno označevanje besedil (skladnost med označevalci)

ANALIZA BESEDIL

- Skupno število vseh zabeleženih napak → 2.484
  - 1546 oz. 63 % na področju družboslovja
  - 939 oz. 37 % na področju naravoslovja
- Družboslovje → največ jezikovnih napak iz kategorije „drugo“ (nato slogovne in skladenjske napake)
- Naravoslovje → največ slogovnih napak (nato iz kategorije jezik – „drugo“ in napake v skladnji)

ANALIZA GLEDE NA TIPOLOGIJO NAPAK

Napaka	Število napak
Jezik - neustrezen slog	389 (16 %)
Jezik - drugo	379 (15 %)
Jezik - skladnja	260 (10 %)
Jezik→terminologija→nepravilen pomen	163 (7 %)
Jezik - ločila	145 (6 %)
Prenos pomena → pomensko neustrezen	142 (6 %)
Jezik - napačna kolokacija	136 (5 %)
Jezik - neustrezen glagolski čas	125 (5 %)
Prenos pomena - preveč dobesedno	121 (5 %)
Jezik - napačen predlog	107 (4 %)
Jezik - napačna začetnica	96 (4 %)
Jezik - napačno število	72 (3 %)
Prenos pomena - dodano	39 (2 %)
Jezik - terminologija - drugo	36 (1 %)
Prenos pomena - izpust	31 (1 %)
Jezik - neprimerno za tip besedila	31 (1 %)
Jezik – istorečje	30 (1 %)
Jezik - črkovanje	26 (1 %)
Prenos pomena - preveč svobodno	25 (1 %)
Jezik - nesorodna beseda	23 (1 %)
Jezik - termin, preveden z neterminološkim izrazom	23 (1 %)
Prenos pomena - nejasno	20 (1 %)
Jezik - slog - drugo	17 (1 %)
Prenos pomena - drugo	11 (< 0,5 %)
Jezik - skloni in ujemanje - drugo	8 (< 0,5 %)
Jezik - terminologija - nekonsistentno znotraj ciljnega prevoda	7 (< 0,5 %)
Jezik - neskladno z glosarjem	7 (< 0,5 %)
Prenos pomena - prevedeni neprevedljivi izrazi (lastna imena ipd.)	4 (< 0,5 %)
Prenos pomena - poseganje v ciljni jezik - drugo	3 (< 0,5 %)
Jezik - nedosledno z izvirnim jezikom	3 (< 0,5 %)
Jezik - register - nekonsistentno znotraj ciljnega prevoda	3 (< 0,5 %)
Jezik - naglas ali diakritično znamenje	1 (< 0,5 %)
Jezik - register - drugo	1 (< 0,5 %)

ANALIZA UJEMANJA MED OZNAČEVALCEMA

- Namen: kako skladni so označevalci napak
- Rezultati vplivajo na verodostojnost rezultatov končne splošne analize
- Analiza metodologije
- Podkorpus
  - 8 dvojno označenih besedil (4 naravoslovna, 4 družboslovna)
  - 443 povedi oz. 9415 pojavnic
  - 27 % korpusa

REZULTATI ANALIZE

- Skupno število vseh napak v besedilih: 511
- Enako označenih 102
- Različno označenih 409
- Primerjava glede na prevajalca in vrsto besedil
  - označevalca sta bila pri naravoslovnih besedilih (za 20 %) bolj skladna

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**	**26**	**69**	**77**	**27**	**34**	**69**	**70**	**409**

ANALIZA NEUJEMANJ MED OZNAČEVALCEMA

- Neujemanje na nivoju fraze (25 %)
  - enako označena napaka, vendar različna označevanje v besedilu
- Neujemanje na nivoju tipa napak
  - Neujemanje na vrhnjem nivoju (prenos vsebine, jezik)(25 %)
    - razhajanje na nivoju jezika in prenosa vsebine → veliko razhajanje z največjo težo
    - diploma -> Bachelor thesis -> undergraduate dissertation jezik – pomensko neustrezno, vsebina – neskladno z izvirnim jezikom

ANALIZA NEUJEMANJ MED OZNAČEVALCEMA

- Neujemanje na srednjem nivoju (register, slog)
  - razhajanje znotraj prvih dveh večjih skupin napak (28 %)
  - Ok -> agreed -> okay

  terminologija in leksika, nepravilno; register, neprimerno za tip besedila.
- Neujemanje na spodnjem nivoju (slog, higijena)
  - razhajanje znotraj najožjih kategorij tipologije (22 %)
  - Najpogosteje *nesorodna beseda – napačna kolokacija; napačen termin – nesorodna beseda; napačen termin – napačna kolokacija.*
  - sem (biti) -> I am -> I'm

  neprimerno za tip besedila; drugo

TEŽAVE

- Neskladnost precej visoka -> negativno vpliva na uporabnost korpusa, potrebne natančne smernice, kako se označuje, tipi napak
- Definicije kategorij, osnovni in obrobni primeri (nabrani s pomočjo korpusa)
- Vsak jezik svoje vrste napak
- Manjkajoče kategorije: napačen predlog, ni pa napačen člen (določni, nedoločni)

ZAKLJUČEK

- Pilotna študija
- Cilj: izdelati zasnovo za korpus ter preizkusiti orodje Webanno, metodologijo in tipologijo za označevanje napak
- Kolikšno je razhajanje med označevalci brez predhodnega usklajevanja
- V prihodnjih raziskavah je korpus potrebno povečati in razširiti na druge tipe besedil
- Izboljšati metodologijo označevanja in izdelati robustne smernice za izboljšanje ujemanja med označevalci