DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations

Ekaterina Lapshinova-Koltunski[1], Maja Popović[2], Maarit Koponen[3]

[1]Saarland University, [2]ADAPT Centre, [3]University of Eastern Finland

[1]e.lapshinova@mx.uni-saarland.de, [2]maja.popovic@adaptcentre.ie, [3]maarit.koponen@uef.fi

Setting Goals and Motivation

- A new parallel corpus containing professional and student translations of two different text registers – news and reviews
- Analyse differences in human translation
- Important for MT evaluation

Related Work

- Translationese: differences between translations and non-translations
- Studies of variation in human translation: expertise, register
- Existing corpora: RusLTC, VARTRA, KOPTE, Opacus, Opusparcus, Finnish Paraphrase Corpus

Corpus collection

news sources	articles	sent.	words
abc news	4	35	734
bbc	12	112	2,330
cbs news	5	67	1,668
chicago defender	1	6	308
cnbc	3	26	664
cnn	7	68	1,857
daily mail	3	38	889
en ndtv	1	6	169
euronews	4	30	584
fox news	1	10	300
gateway	1	7	567
independent	1	10	278
kcal	1	14	475
ny times	2	26	791
reuters	6	55	1,620
rt	1	6	142
scotsman	5	52	1,292
seattle times	2	11	294
sky	1	11	239
telegraph	5	52	1,305
upi	2	28	680
total	68	670	17,186

review domains	reviews	sent.	words
beauty	14	72	966
books	14	73	1,100
cd and vinyl	14	74	1,029
cell phones	14	65	989
grocery and food	14	69	1,045
health care	14	72	1,114
home and kitchen	14	72	1,103
movies and TV	14	77	1,168
musical instruments	14	72	1,102
patio and garden	14	73	1,162
pet supplies	14	80	1,173
sports and outdoors	14	75	1,144
toys and games	14	73	1,065
video games	14	68	1,076
total	196	1,015	15,236

Metadata

- 4 to 20 translators per group
- 0 to 37 years of experience
- MA and BA students of translation

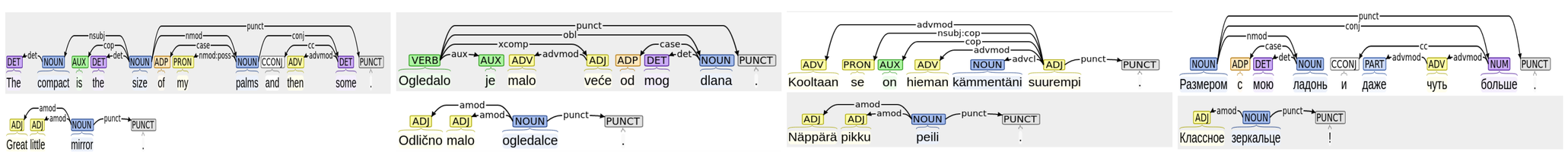
Translation task

- 4 translation variants of each EN news text: HR, RU by professionals and students
- 6 translation variants of each EN review: HR, RU, FI by professionals and students
- preserved sentence alignment, no machine translation allowed

Corpus statistics

	txt	sent	words
en	264	1,685	32,422
hr	528	3370	61,237
ru	528	3370	63,003
fi	528	3370	23,922
total	1848	11,795	180,584

Corpus Annotation



Data Analysis

Text statistics and lexical variety for full forms

text		statistics words	statistics voc	lexical variety voc/words↑	lexical variety Yule's K ↓
en news	/	17,186	4,138	**0.220**	**98.2**
en reviews	/	15,236	3,155	0.178	101.7
hr news	prof	16,662	6,009	**0.341**	86.2
	stud	16,632	5,975	0.340	**83.8**
hr reviews	prof	14,003	4,359	0.282	92.1
	stud	13,940	4,446	**0.288**	**88.2**
ru news	prof	17,469	6,079	0.340	122.9
	stud	17,054	6,076	**0.349**	**116.7**
ru reviews	prof	14,233	4,417	0.289	126.3
	stud	14,247	4,523	**0.300**	**124.1**
fi reviews	prof	11,709	4,612	**0.360**	**109.8**
	stud	12,213	4,664	0.350	112.5

Text statistics and lexical variety for lemmas

text		statistics lemmas	statistics voc	lexical variety voc/lemmas↑	lexical variety Yule's K ↓
en news	/	18,089	3,340	**0.185**	**108.1**
en reviews	/	16,342	2,350	0.143	125.2
hr news	prof	17,215	3,809	**0.222**	124.3
	stud	17,241	3,775	0.218	**122.6**
hr reviews	prof	14,785	2,785	0.188	158.2
	stud	14,809	2,838	**0.192**	**157.3**
ru news	prof	17,914	3,777	0.211	130.8
	stud	17,512	3,802	**0.217**	**126.0**
ru reviews	prof	15,163	2,667	0.176	145.0
	stud	15,116	2,771	**0.183**	**140.7**
fi reviews	prof	12,723	2,667	0.210	**158.6**
	stud	13,212	2,783	**0.211**	166.4

Overlap or distance between translators

measure	**genre**	**target language hr**	**ru**	**fi**
word overlap↓	news	58.6	55.6	/
(F1 score)	reviews	57.6	53.4	51.9
edit distance↑	news	59.5	63.7	/
(normalised)	reviews	58.2	63.6	63.4
word order	news	7.4	6.4	/
mismatch↑	reviews	5.2	5.2	4.4
inflection	news	8.7	8.5	/
mismatch↑	reviews	7.0	6.2	10.3
lexical	news	42.7	48.1	/
mismatch↑	reviews	45.8	52.2	47.9

Examples of mismatches: word order, morphology, lexical choice

News

EN: The charges will be reviewed by the Public Prosecution Service.
HR prof: državno odvjetništvo razmotrit će optužbe .
HR stud: optužbu će pregledati državno odvjetništvo .

Reviews

EN: It doesn't melt in the Florida heat and is sheer enough to be natural but have adequate coverage.
HR prof: ne topi se na vrućinama u floridi i dovoljno je prozračna da bude prirodna , ali istovremeno dobro prekrije nepravilnosti .
HR stud: ne topi se na floridskim vrućinama i dovoljno je prozračan da izgleda prirodno , ali i dovoljno prekriva .

Metrics for differences between professional and student translations

- **word unigram matching (F1 score):** different translators used same words
- **edit distance:** different translators used different words, or same words in different order
- **word order mismatch:** different translators used same words but in different positions: indicates differences in the sentence structure
- **inflection mismatch:** different translators used the same lemma but in different forms: indicates morpho-syntactic differences
- **lexical mismatch:** different translators used different words (lemmas) and/or phrases: indicates differences in lexical choice.

General observations

- news articles have a richer vocabulary than user reviews
- voc/words ratio is higher for all translations
- student translations contain a richer vocabulary than professionals (except Finnish).

Differences: professionals vs. students

- professional and student translations differ
- variation is language-/genre-dependent: lexical choice in reviews, inflectional mismatch in Finnish
- translations are more similar in news than reviews.

Corpus availability

The corpus DiHuTra is available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License and can be accessed from CLARIN http://hdl.handle.net/21.11119/0000-000A-1BA9-A or GitHub https://github.com/katjakaterina/dihutra

Acknowledgments

We were supported by EAMT, ADAPT Centre and a Kopiosto grant (SKTL). We thank the translators in Volgograd, Zagreb, Rijeka and Finland.