Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1751–1760
Marseille, 20-25 June 2022

DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations

Ekaterina Lapshinova-Koltunski[1], Maja Popović[2], Maarit Koponen[3]
[1]Language Science and Technology, [2]ADAPT Centre, [3]Foreign Languages and Translation Studies
[1]Saarland University, [2]Dublin City University, [3]University of Eastern Finland
[1]e.lapshinova@mx.uni-saarland.de, [2]maja.popovic@adaptcentre.ie, [3]maarit.koponen@uef.fi

Abstract

This paper describes a new corpus of human translations which contains both professional and students translations. The data consists of English sources – texts from news and reviews – and their translations into Russian and Croatian, as well as of the subcorpus containing translations of the review texts into Finnish. All target languages represent mid-resourced and less or mid-investigated ones. The corpus will be valuable for studying variation in translation as it allows a direct comparison between human translations of the same source texts. The corpus will also be a valuable resource for evaluating machine translation systems. We believe that this resource will facilitate understanding and improvement of the quality issues in both human and machine translation. In the paper, we describe how the data was collected, provide information on translator groups and summarise the differences between the human translations at hand based on our preliminary results with shallow features.

Keywords: translation, human translation, parallel corpus, multilingual corpus, multilinguality, Russian, Croatian, Finnish, translation variation, machine translation, news translation, review translation

1. Introduction

In this paper, we describe a new parallel corpus containing professional and student translations of two different text registers – news and reviews. This corpus has been compiled to address an important problem affecting machine translation (MT), namely variation in human translation. Besides being an essential resource for MT evaluation, this corpus will also allow us to study important differences between professional and student translations across different languages (Russian, Croatian and Finnish) and text registers (news and reviews). We believe this kind of work is timely because there are studies showing evidence of variation between texts generated by different translators in terms of language patterns which may have an impact on MT evaluation too (Rubino et al., 2016; Popović, 2020). The corpus is available from the CLARIN repository[1], so it can be found though the Virtual Language Observatory[2]. Additionally, a GitHub repository[3] contains the data and some additional information.

In the MT community, we observe a growing awareness of the phenomena related to human translations, i.e. features of translated texts that make them statistically different from non-translated texts (Gellerstam, 1986; Baker, 1996; Toury, 1995). These features not only reflect the variation between text production types (translations and non-translations), but also help to detect the level of competence (professional vs. novice) in translations (Kunilovskaya and Lapshinova-Koltunski, 2020).

At the same time, we still observe a widespread lack of deep understanding of these phenomena, especially in relation to MT. Machine translation has evolved very rapidly, so that in recent years MT evaluation often also involves comparing human and machine translation quality, and even becomes a part of WMT[4] shared tasks (Barrault et al., 2020; Barrault et al., 2019). As the performance of MT systems is compared with the human performance, it is important to know more about different human translations that can be contained in the reference data. The findings on the basis of the compiled parallel corpora containing both professional and student translations of two different registers will facilitate discovery of interesting phenomena in this area, and will help us to understand and improve the quality issues in translation. Although WMT shared tasks contain a vast number of human translations involving different expertise levels and different native languages, none of these texts contain parallel human translations from the same source language text, but only comparable corpora from the same source language. Therefore it is not possible to perform an exact comparison between human translations and to analyse the details. We believe that our corpus which consists of one source text and its distinct parallel human translations into Russian, Croatian and Finnish will facilitate improvement of the quality issues in both human and machine translation.

To the best of our knowledge, there are no corpora of this kind so far, as most existing resources contain one language or one text register only. Besides that, most

[1]http://hdl.handle.net/21.11119/0000-000A-1BA9-A
[2]https://vlo.clarin.eu
[3]https://github.com/katjakaterina/dihutra
[4]https://www.statmt.org/wmt21/

of them are not publicly available.

The remainder of the paper is organised as follows. In Section 2, we describe the existing studies on variation in human translation as well as corpora that exist and that are suitable for the analysis of variation. Section 3 provides details about the corpus creation including information about English source texts, the collected translations, and translator profiles. Section 4 contains information about how the data was annotated. We describe initial analyses of the differences in translations in Section 5. In Section 6, we conclude and outline our future work.

2. Related Work

2.1. Studies of variation in human translation

There exists a substantial bulk of work that shows that (human) translations vary under impact of various factors, including language pair and text register (Cappelle and Loock, 2017; Evert and Neumann, 2017; Lapshinova-Koltunski, 2017). Most of these studies are focused on the linguistic specificity of translations which makes them different from non-translations (texts originally written in the given language) through a number of linguistic features. These features, also called translationese and translation universals, are distinctive linguistic patterns that have been extensively used in the area of corpus-based translation studies to analyse translation variation. The features facilitate automatic differentiation between translations and non-translations (Rabinovich et al., 2017; Volansky et al., 2015; Laippala et al., 2015; Baroni and Bernardini, 2006).

Only a few publications applied these features to automatically differentiate between translations produced by various translator groups. These studies showed that translation features are manifested to different degrees in translations produced by different groups of translators. Corpas Pastor et al. (2008) and Ilisei (2012) automatically distinguished between non-translations in Spanish and English-to-Spanish translations by professionals and students investigating the validity of translation universal of convergence. Convergence was defined as similarity between texts translated by translators of different proficiency levels. The authors did not find any significant differences between student and professional translations in terms of the features applied. Martínez and Teich (2017) focused on the differences in the lexical choices by professional and student translators and applied a probabilistic approach to study translation routine. Redelinghuys (2016) analysed student and professional translations of texts belonging to different registers and found a relation between translation expertise and register sensitivity – inexperienced translators in her data seemed to be more repetitive when translating informal texts, e.g. creative writing. This was also confirmed in a recent study by Bizzoni and Lapshinova-Koltunski (2021) who tested conformity of student and professional translations of various text registers to a neural model of the target language. Lapshinova-Koltunski (2020) explored linguistic features represented in the form of lexico-grammatical patterns which contributed to the automatic classification between novice and professional translations. Kunilovskaya and Lapshinova-Koltunski (2020) found correlation between the levels of professionalism and types of the translationese effects they were analysing. However, the data they were using was not entirely comparable – the sources in the English–Russian pair were not the same for student and professional translations. Popović (2020) compared shallow text features (lexical variety, lexical density, POS variety) between different translation expertise and native languages. Although the results were reported for several language pairs, the vast majority of the analysed corpora which originated from the WMT shared tasks were comparable and not parallel.

Understanding factors impacting variation in translation is also important for MT, specifically for MT evaluation, as has been indicated by Popović (2020). Moreover, there are works showing the impact of features of translated texts on machine translation evaluation (Zhang and Toral, 2019; Graham et al., 2019). Our recent studies motivate even more to the analysis of differences between various translator groups (Kunilovskaya and Lapshinova-Koltunski, 2020; Popović, 2020), as information on the differences between human translations turns to be important for MT evaluation.

We rely on the existing work mentioned above and define important issues for our analyses and corpus design: (1) the data should contain multiple translations of the *same* sources produced by translators of different levels of expertise (students and professionals); (2) the data should be heterogeneous in terms of registers (e.g. formal and informal); (3) the analysed features should be supported by studies on translated texts, both human and machine translations.

2.2. Corpora with different human translations

To our knowledge, there are not many corpora containing both professional and student translations. RusLTC (Kutuzov and Kunilovskaya, 2014) is an English-Russian parallel corpus, which contains source texts sentence-aligned with their multiple targets produced by translation students from different Russian universities. A significant part of the corpus consists of English mass-media texts translated by senior students majoring in translation studies. This corpus is much bigger than the corpus described in this paper, counting over 2.3 million word tokens in total. Each source text in the RusLTC corpus has multiple translations (8 on average). This corpus does not contain any professional translations but was used together with English-Russian professional translations of comparable mass-

media and newspaper texts from the Russian National Corpus (Plungian et al., 2005) in a number of studies (Kunilovskaya et al., 2018; Kunilovskaya and Kutuzov, 2017; Kunilovskaya and Lapshinova-Koltunski, 2019; Kunilovskaya and Lapshinova-Koltunski, 2020) to analyse variation in translation.

VARTRA (Lapshinova-Koltunski, 2013) contains professional and student translations of the same English sources. This corpus covers several text registers (e.g. fiction, instruction, popular science, etc.) but only for one language pair: English–German. Moreover, only one part of the corpus is publicly available, as the English sources and their professional translations which come from the corpus CroCo (Hansen-Schirra et al., 2012) underlie some authorship restrictions and are available for academic purposes on request only.

The corpus KOPTE (Wurm, 2016) contains multiple student translations of the same source texts for the German–French language pair. This is a valuable resource for the analysis of variation in translation by multiple translators. However, this corpus contains texts for only one language pair and does not contain professional translations.

The corpus used by Popović (2020) contains news texts in three different language pairs and five translation directions including English–Croatian, German–French and English–Finnish. The data was extracted from the publicly available WMT shared tasks. However, this corpus data has several limitations: not all translations are of the same source texts, there is a variation in translator groups, and some source texts were not written in English but translated from Czech original texts.

There are two further corpora that should be mentioned here – Opusparcus (Creutz, 2018) and the Finnish Paraphrase Corpus (Kanerva et al., 2021). Both contain alternative translations by different translators of movie and television subtitles that were extracted from Opensubtitles. Opusparcus covers six languages (German, English, Finnish, French, Russian and Swedish) and consists mainly of semi-automatically annotated alternative translations. The Finnish Paraphrase Corpus contains a smaller selection of manually annotated translations into Finnish. These corpora may contain translations by professionals and non-professionals. However, the meta data about translators is missing, therefore, there is no information on their background available.

3. Corpus Collection

Our corpus contains texts representing two distinct text registers, because existing studies show that professionals and students show different degrees of sensitivity to registers and genres (Bizzoni and Lapshinova-Koltunski, 2021; Redelinghuys, 2016) as mentioned in Section 2.1 above.

3.1. English source texts

For the English source texts, we included Amazon product reviews and news texts.

The subcorpus of the **Amazon product reviews**[5],[6] (McAuley et al., 2015) contains more than 82 million unique product reviews from Amazon written in English with overall ratings from 1 (worst) to 5 (best). The reviews are divided into 24 categories, such as “Sports and Outdoors", “Books”, “Musical Instruments”, “Movies and TV”, “Grocery and Gourmet Food”, etc. The reviews consist of 5.4 sentences and 93.2 words on average. We selected a set of reviews from fourteen categories, paying attention to the data balance: equal number of positive and negative reviews, a balanced distribution of categories (topics). In total, we included 196 reviews, fourteen from each of the fourteen topics.

The **news** texts were imported from the News test corpus of the WMT (2019 and 2020) shared task[7]. The topics of the news vary and include politics, sports, criminality, health, and others. The news are longer than reviews, with 9.9 sentences and 221.7 words on average.

Contrary to the reviews, WMT shared tasks also contain a set of human translations of the English source texts into several languages including Russian, German, Czech and some others, but no Croatian or Finnish.

Therefore, we selected only the texts which were originally written in English and translated into Russian by professional translators hired by the WMT organisers. In addition, we paid attention that the selected source texts are also translated into German by professional WMT translators, to be able to extend the resource in the future with the German student translations. In total, we included 68 news articles from different sources.

In Table 1, we provide statistics in terms of text (review/news article), sentence and word number for the different news sources and review domains. The whole set of the English source texts contains 1,685 sentences and 28,399 words (670 and 17,186 for news; 1,015 and 15,236 for reviews).

Each English review was translated into the three target languages, Croatian, Russian and Finnish, by professionals and by students. For the news corpus, Russian translations were already available from the WMT shared task and Croatian translations were produced for the purpose of this work. Finnish professional translations were not provided for the news articles. In addition to translations, the information about age, gender, experience and the study program (for students) was collected. Translators were asked too keep the sentence alignment (not to merge or to split sentences so that each English sentence corresponds to one translated sentence, which is important for current MT systems) and not to use machine translation in the pro-

[5] https://s3.amazonaws.com/amazon-reviews-pds/readme.html

[6] http://jmcauley.ucsd.edu/data/amazon/

[7] http://www.statmt.org/wmt20/translation-task.html

(a) news: different sources

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

(b) reviews: different domains (products)

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

Table 1: Distribution of sources in English news (a) and domains in English reviews (b)

cess of translation. No further restrictions were given to translators. The total number of tokens in the resulting corpus amounts to 180,584.

3.2. Croatian translations

Both professional and student translations into Croatian were produced in cooperation with the University of Zagreb and the University of Rijeka in Croatia. In total, four professional translators and twenty translation students participated, all native speakers of Croatian. Translation experience (estimated by themselves) of professional translators ranges between five and ten years, while for students the range is from 0 to five years, the majority being in the range from 2 and 4 years. The two students who indicated no experience (0 years) also indicated that they had no real professional experience yet, only work in the framework of their studies. All students were in their first or second year of master studies (MA).

3.3. Russian translations

As already mentioned above, Russian professional translations of news were imported along with the English sources from the WMT test corpus (2019 and 2020). These were extended with student translations of news. For reviews, we collected both professional and student translations. We cooperated with the Institute of Philology and Intercultural Communication of Volgograd State University (Volgograd, Russia). Overall, 10 translation trainers (professionals) and 10 translation trainees, all native speakers of Russian, participated in the data collection. Translation experience (estimated by translators themselves) ranges between five and 37 years for professional translators and one to six years for student translators. There are both advanced BA students (third or fourth year of study) and first or second year MA students amongst the student translators.

3.4. Finnish translations

For the Finnish subcorpus of review translations, we collected translations into Finnish from both professional translators and translation students. In total, seven professionals and seven students participated in the data collection. All were native Finnish speakers. The professional translations were commissioned through a Finnish translation agency. The student translations were coordinated through a students' cooperative that has members from translation training programmes in three Finnish universities. Translation experience of five of the professionals (estimated by themselves) ranged from five to 13 years, while one reported 1.5 years and one less than a year of full-time professional experience. Of the student translators, two reported no professional experience, while the others indicated that they had some experience with freelance projects alongside their studies. One had done freelance translations over approximately four years, two over approximately two years, and two for less than a year. Two had recently completed an MA degree in translation, and the other five were close to graduation (fifth or sixth year of studies).

4. Corpus Annotation

We automatically annotated the collected originals in English, as well as both professional and student translations into Croatian, Finnish and Russian with the information on tokens, lemmas, parts-of-speech, morphological information (number, tense, mood, etc.) and dependencies. To ensure the comparability across

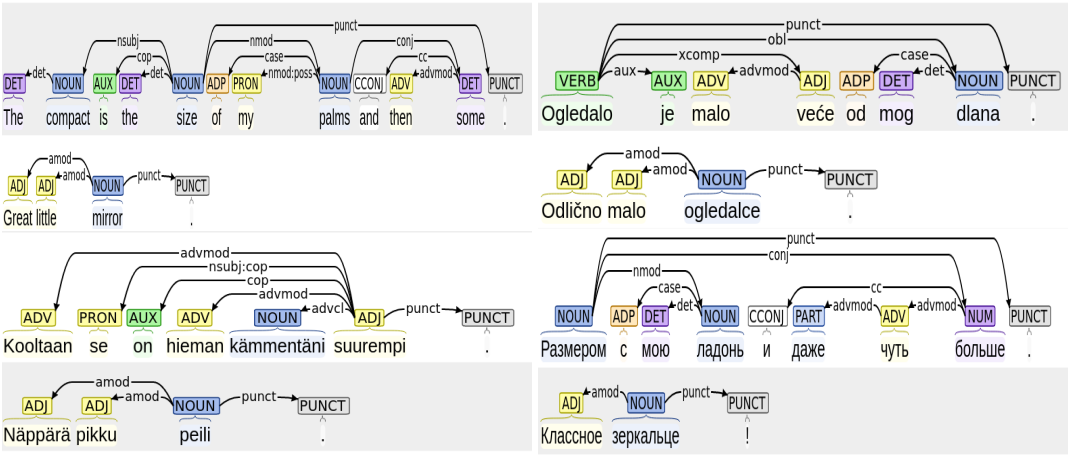

Figure 1: Example of the UD parsed data in English, Croatian, Finnish and Russian.

the four languages, we used universal parts-of-speech and universal dependencies that we obtained with the Stanford NLP Python Library Stanza (v1.2.1)[8]. Figure 1 provides an illustration of the two parsed sentences from example (1) annotated with universal part-of-speech labels and dependencies in the four language under analysis:

(1) EN: The compact is the size of my palms and then some. Great little mirror.
HR: Ogledalo je malo veće od mog dlana. Odlično malo ogledalce.
FI: Kooltaan se on hieman kämmentäni suurempi. Näppärä pikku peili.
RU: Размером с мою ладонь и даже чуть больше. Классное зеркальце!

The list of universal parts-of-speech contains 17 categories which are considered to be universal for all languages [9]. The list includes both open and close word classes, as well as punctuation. These are adjectives (ADJ), adpositions (ADP), adverbs (ADV), auxiliaries (AUX), coordinating conjunctions (CCONJ), determiners (DET), interjections (INTJ), nouns (NOUN), numerals (NUM), particles (PART), pronouns (PRON), proper nouns (PROPN), punctuation marks (PUNCT), subordinating conjunctions (SCONJ), symbols (SYM, verbs (VERB) and others (X).
Statistics on the distribution of the part-of-speech categories for each translation variant, language and register are given in Table 2.
In spite of universality of dependency relation, the annotations contain some mismatches across languages. For instance, relation `iobj` (indirect objects) is not available in the Finnish model or the relation `nsubj:pass` (passive nominal subject) is absent in the model for Croatian. Such mismatches should be considered when the universal dependencies are used to compare translations across languages (e.g. if we compare the source texts with the translations). However, they do not play any role when translation variants are compared (within every language). The list of relations contained in all the four languages include: `acl` (adnominal clause or clausal modifier of noun), `advcl` (adverbial clause modifier), `advmod` (adverbial modifier), `amod` (adjectival modifier), `appos` (appositional modifier), `aux` (auxiliary), `case` (case marking), `cc` (coordinating conjunction), `ccomp` (clausal component), `conj` (conjunct), `cop` (copula), `csubj` (clausal subject), `det` (determiner), `discourse` (discourse element, e.g. particle), `fixed` (fixed multiword expression), `mark` (marker, e.g. clausal marker), `nmod` (nominal modifier), `nsubj` (nominal subject), `nummod` (numeric modifier), `obj` (object), `obl` (oblique nominal), `parataxis`, `punct` (punctuation), `root`, `xcomp` (open clausal element)[10], see de Marneffe et al. (2021)

5. Data Analysis

We performed preliminary analysis of the data in two ways: (i) calculating shallow statistics of the texts and (ii) estimating differences between professional and student translations based on word matching and edit distance. The details are provided in the following sections.

5.1. Shallow statistics of the texts

We collected the first statistics on the shallow features in terms of sentences, running words and vocabulary in the sources and the translations in the three target languages (see Table 3). We also estimated lexical richness in the form of ratio between vocabulary (voc,

[8]Stanza is an NLP package in Python (see `https://stanfordnlp.github.io/stanza/index.html` for details) where models are all pre-trained on the Universal Dependencies v2.5 datasets.

[9]See `https://universaldependencies.org/u/pos/` for more details.

[10]A more detailed decription of the relations is provided under `https://universaldependencies.org/u/dep/`.

	news					reviews						
	en	hr		ru		en	hr		fi		ru	
		prof	stud	prof	stud		prof	stud	prof	stud	prof	stud
ADJ	1109	1798	1718	1236	1233	1234	1369	1358	959	969	1219	1252
ADP	1890	1693	1671	1969	1932	1159	999	955	130	148	1055	1119
ADV	507	613	615	520	513	890	915	973	1046	1046	858	877
AUX	834	1290	1272	180	171	983	1251	1258	1031	1083	147	162
CCONJ	445	544	534	464	442	541	656	632	506	503	539	514
DET	1456	596	639	403	366	1343	764	747	0	0	540	492
INTJ	5	4	5	2	0	16	27	25	18	20	3	1
NOUN	3152	3861	3735	4179	4242	2580	2528	2478	2465	2452	2638	2648
NUM	341	254	238	355	359	251	192	196	162	171	236	239
PART	461	84	94	205	254	483	199	234	0	0	526	549
PRON	964	457	489	823	704	1670	600	646	1070	1265	1186	1247
PROPN	1892	1285	1284	1523	1473	394	339	314	370	381	434	417
PUNCT	1937	1690	1826	2843	2744	1581	1601	1622	1823	1959	2536	2430
SCONJ	335	577	587	472	390	351	495	534	345	365	420	408
SYM	30	4	7	6	18	62	2	1	37	62	5	7
VERB	2009	1611	1692	2037	1958	1775	1766	1731	1630	1664	1775	1750
X	1	130	108	0	3	6	54	91	15	10	3	1

Table 2: Distribution of universal parts-of-speech in English, Croatian, Finnish and Russian news and reviews

which is the number of different words) and total number of words and Yule's K coefficient. Both values indicate how rich the vocabulary is in the given text, the richness being proportional to the voc/words ratio (higher value indicates richer vocabulary) and inversely proportional to Yules'K (lower value indicates richer vocabulary).

Several observations can be noted from the results: first of all, news articles in the English source texts have a notably richer vocabulary than user reviews. Furthermore, the voc/words ratio is higher for all translated texts than for the English originals, which might be surprising given that translated texts are sometimes suggested to be simpler (in terms of the vocabulary). Nevertheless, language characteristic have to be taken into account here since all three target languages are morphologically rich so that there are many different forms of one lemma. Yule's K indicates that the translations are in fact simpler than original, except for Croatian translations.

As for different translations, almost all student translations contain a richer vocabulary than professionals – the only exceptions are translations of Finnish reviews. Students seem to use less running words (shorter sentences) but more unique words, which unexpectedly points to a richer vocabulary used by students. This is somewhat surprising, since it could be expected that students are using more repeated words and generating simpler translations. An alternative view may be that professional translators often work in situations where consistent use of terminology is expected (e.g. technical texts), which may lead them to use more uniform vocabulary. On the other hand, the Finnish professional translators' higher lexical variety for full forms (compared to the students) could indicate more frequent use of the Finnish clitic particles to reflect colloquial tone. All those observations, however, require a deeper analysis for which we need further data analyses. Moreover, some differences between students and professionals seem to be small and should be proved by a significance test.

5.2. Differences between professional and human translations

We also analyse the overlap or the distance between professional and student translations within language pairs using the following metrics:

- word unigram matching (F1 score):
  different translators used same words;
- edit distance (Levenshtein, 1966):
  different translators used different words, or same words in different order (positions in the sentence).

In order to better understand the results, three additional metrics obtained by combining the two above as presented in (Popović and Ney, 2011) are used:

- 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.

(a) Statistics for full word forms

text	translators	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	professionals	16,662	6,009	**0.341**	86.2
	students	16,632	5,975	0.340	**83.8**
hr reviews	professionals	14,003	4,359	0.282	92.1
	students	13,940	4,446	**0.288**	**88.2**
ru news	professionals	17,469	6,079	0.340	122.9
	students	17,054	6,076	**0.349**	**116.7**
ru reviews	professionals	14,233	4,417	0.289	126.3
	students	14,247	4,523	**0.300**	**124.1**
fi reviews	professionals	11,709	4,612	**0.360**	**109.8**
	students	12,213	4,664	0.350	112.5

(b) Statistics for base forms (lemmas)

text	translators	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	professionals	17,215	3,809	**0.222**	124.3
	students	17,241	3,775	0.218	**122.6**
hr reviews	professionals	14,785	2,785	0.188	158.2
	students	14,809	2,838	**0.192**	**157.3**
ru news	professionals	17,914	3,777	0.211	130.8
	students	17,512	3,802	**0.217**	**126.0**
ru reviews	professionals	15,163	2,667	0.176	145.0
	students	15,116	2,771	**0.183**	**140.7**
fi reviews	professionals	12,723	2,667	0.210	**158.6**
	students	13,212	2,783	**0.211**	166.4

Table 3: Text statistics and lexical variety for (a) full forms and (b) base forms (lemmas) of the words.

Table 5 provides and overview of the overlaps and mismatches between the two variants of translations for each language pair and text register. First of all, we see that professional and student translations differ. For example, interpretation of the F1 score of 58.6 for Croatian news is that in each set of 100 words, 58.6 words are different in the two translations: professionals and students used different word/phrase or different form of a word. Edit distance of 59.5 between the same two texts should be interpreted as follows: in order to make the two translations identical, 59.5 of 100 words should be changed.

These numbers indicate notable differences between professional and student translations, which require further analysis to better understand their nature.

A very preliminary analysis in this direction, namely estimating three types of mismatches, shows that most of the differences consist of different lexical choices (45-55%), namely using different words and phrases. Different word forms are used in 6-9% cases, while same words in different order in 5-7.5% cases.

Two examples of the three types of differences between professional and student translations into Croatian are presented in Table 4: one for news and one for reviews.

Word order mismatch is presented in italic, morphological differences are underlined, and bold denotes different lexical choices (using different words and/or phrases). It is worth noting that morphological differences are often related to lexical choices, such as different prepositions requiring different cases, nouns with different genders, etc.

In addition, we can see that variation between professional and student translations is dependent on the language pair and also on the register. In reviews, there seem to be more differences in lexical choice than in news, but less differences in word order and in word forms. Also, Finnish, as the morphologically richest language from the three, exhibits the highest difference in word form choice, and the lowest one in word order. As for lexical choice mismatch, it is highest for Russian. While further analysis is certainly required, one possible reason might be different transcriptions of

en news	The charges will be reviewed by the Public Prosecution Service.
hr prof	*državno odvjetništvo* **razmotrit** *će* optužbe .
hr stud	optužbu *će* **pregledati** *državno odvjetništvo* .
en review	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 .

Table 4: Examples of mismatches between professional and student translations of news (above) and reviews (below) into Croatian: italic denotes word order, underline denotes morphology and bold denotes lexical choice.

named entities.

We also observe that translations by professionals and students are more similar in the domain of news than reviews. This could be due to the registerial differences between reviews and news: while reviews are more 'creative' and colloquial, news texts might be more standardised. Moreover, news texts are informative and objective, whereas reviews contain evaluative constructions, which may vary more in translation. However, this should also be investigated with more details in further analyses.

overlap/distance 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

Table 5: Overlap/distance between professional and student translations: word unigram matching (F1 score), edit distance, and three mismatch types obtained by combining the two above: word order mismatch, inflection mismatch and lexical mismatch.

6. Conclusion and Future Work

The presented paper describes a newly created corpus of human translations which contains texts from two different registers translated by two different groups of translators into three different target languages. The corpus represents a valuable resource to study variation in translation as it allows a direct comparison between human translations of the same source texts. Our preliminary analyses based on the shallow text statistics and matching/distance measures indicate that there are differences between professional and student translations which depend on the language pair as well as on the register. Further work is planned to better understand these differences: we plan to carry out detailed analyses using the annotated data. Besides that, we will use the information about individual variation that is available in the metadata to better understand the differences observed. In some cases, advanced student translators may have the same or similar level of proficiency as the professionals who are in the beginning of their working life.

The corpus also represents a valuable resource for evaluation of MT systems for the three language pairs. We believe that this resource will help us to understand and improve the quality issues in both human and machine translation.

As mentioned above, the corpus is available in the CLARIN. The project has additionaly a Github repositorywhich contains the data and some additional information, see the link in Section 1 above.

7. Acknowledgements

The creation of the corpus was supported by the European Association of Machine Translation (EAMT) within a sponsorship of activities (2021) and by ADAPT Centre. The ADAPT SFI Centre for Digital Media Technology is funded by Science Foundation Ireland through the SFI Research Centres Programme and is co-funded under the European Regional Development Fund (ERDF) through Grant 13/RC/2106. The Finnish subcorpus was supported by a Kopiosto grant awarded by the Finnish Association of Translators and Interpreters (SKTL). We also thank the teams of translators in Volgograd, Zagreb, Rijeka and Finland. In particular, we thank Aleksandr Besedin from Volgograd State University (Institute of Philology and Intercultural Communication) for coordinating the work of the Russian translators.

8. Bibliographical References

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H.L. Somers, editor, *LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, page 175–186. John Benjamins Publishing Company, Amsterdam.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (wmt20). In *Proceedings of the Fifth Conference on Machine Translation (WMT 2020)*, pages 1–55, Online, November.

Bizzoni, Y. and Lapshinova-Koltunski, E. (2021). Measuring translationese across levels of expertise: Are professionals more surprising than students? In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 53–63, Reykjavik, Iceland (Online), May 31 - June 02. Linköping University Electronic Press, Sweden.

Cappelle, B. and Loock, R. (2017). Typological differences shining through : The case of phrasal verbs in translated English. In Isabelle Delaere, et al., editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, pages 235–264. De Gruyter Mouton, Berlin.

Corpas Pastor, G., Mitkov, R., Afzal, N., and Garcia-Moya, L. (2008). Translation universals: do they exist? a corpus-based and nlp approach to convergence. In *Proceedings of the LREC-2008 Workshop on Building and Using Comparable Corpora*, pages 1–7.

de Marneffe, M.-C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal dependencies. *Computational Linguistics*, 47:255–308.

Evert, S. and Neumann, S. (2017). The impact of translation direction on characteristics of translated texts : A multivariate analysis for English and German. *Empirical Translation Studies: New Methodological and Theoretical Traditions*, 300:47.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin et al., editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund.

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in machine translation evaluation. *CoRR*, abs/1906.09833.

Ilisei, I. (2012). *A machine learning approach to the identification of translational language: an inquiry into translationese*. Doctoral thesis, University of Wolverhampton.

Kunilovskaya, M. and Kutuzov, A. (2017). Universal Dependencies-based syntactic features in detecting human translation varieties. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 27–36, Prague, Czech Republic.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019). Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2020). Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4102–4112, Marseille, France, May.

Kunilovskaya, M., Morgoun, N., and Pariy, A. (2018). Learner vs. professional translations into Russian: Lexical profiles. *The International Journal for Translation and Interpreting Research*, 10.

Laippala, V., Kanerva, J., Missilä, A., Pyysalo, S., Salakoski, T., and Ginter, F. (2015). Towards the Classification of the Finnish Internet Parsebank : Detecting Translations and Informality. In *Nodalida*. Linköping University Electronic Press, Sweden.

Lapshinova-Koltunski, E. (2017). Exploratory analysis of dimensions influencing variation in translation: The case of text register and translation method. In Gert De Sutter, et al., editors, *Empirical Translation Studies: New Methodological and Theoretical Traditions*, volume 300 of *TILSM series*, pages 207–234. Mouton de Gruyter. TILSM series.

Lapshinova-Koltunski, E. (2020). Exploring linguistic differences between novice and professional translators with text classification methods. In Lore Vandevoorde, et al., editors, *New Empirical Perspectives on Translation and Interpreting*, Routledge Advances in Translation and Interpreting Studies, pages 215–238. Routledge.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707–710, February.

Martínez, J. M. M. and Teich, E. (2017). Modeling routine in translation with entropy and surprisal: A comparison of learner and professional translations. In Larissa Cercel, et al., editors, *Kreativität und Hermeneutik in der Translation*. Narr Francke Attempto Verlag.

McAuley, J., Targett, C., Shi, Q., and van den Hengel, A. (2015). Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, pages 43–52, Santiago, Chile.

Popović, M. and Ney, H. (2011). Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4), December.

Popović, M. (2020). On the differences between human translations. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020)*, pages 365–374, Lisboa, Portugal, November.

Rabinovich, E., Ordan, N., and Wintner, S. (2017). Found in translation: Reconstructing phylogenetic language trees from translations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 530–540, Vancouver, Canada, July. Association for Computational Linguistics.

Redelinghuys, K. (2016). Levelling-out and register variation in the translations of experienced and inexperienced translators: a corpus-based study. *Stellenbosch Papers in Linguistics*, 45(0):189–220.

Rubino, R., Lapshinova-Koltunski, E., and van Genabith, J. (2016). Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016)*, pages 960–970, San Diego, California, June.

Toury, G. (1995). *Descriptive Translation Studies - and Beyond*. John Benjamins Publishing Company, benjamins edition.

Volansky, V., Ordan, N., and Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Zhang, M. and Toral, A. (2019). The effect of translationese in machine translation test sets. *CoRR*, abs/1906.08069.

9. Language Resource References

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Hansen-Schirra, S., Neumann, S., and Steiner, E. (2012). *Cross-Linguistic Corpora for the Study of Translations: Insights from the Language Pair English-German*, volume 11 of *Text, Translation, Computational Processing [TTCP]*. Walter de Gruyter GmbH, Berlin/Boston.

Kanerva, J., Ginter, F., Chang, L.-H., Rastas, I., Skantsi, V., Kilpeläinen, J., Kupari, H.-M., Saarni, J., Sevón, M., and Tarkka, O. (2021). Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online), May 31–2 June. Linköping University Electronic Press, Sweden.

Kutuzov, A. and Kunilovskaya, M. (2014). Russian learner translator corpus. In Petr Sojka, et al., editors, *Text, Speech and Dialogue*, volume 8655 of *Lecture Notes in Computer Science*, pages 315–323. Springer International Publishing.

Lapshinova-Koltunski, E. (2013). VARTRA: A comparable corpus for analysis of translation variation. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 77–86, Sofia, Bulgaria, August. Association for Computational Linguistics.

Plungian, V., Reznikova, T., and Sitchinava, D. (2005). Russian National Corpus: General description [Nacional'nyj korpus russkogo jazyka: obshhaja harakteristika]. *Scientific and technical information. Series 2: Information processes and systems*, 3:9–13.

Wurm, Andrea. (2016). *Presentation of the KOPTE Corpus – version 2*.