DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations

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

Abstract

The DiHuTra project aimed to design a corpus of parallel human translations of the same source texts by professionals and students. The resulting corpus consists of English news and reviews source texts, their translations into Russian and Croatian, and translations of the reviews into Finnish. The corpus will be valuable for both studying variation in translation and evaluating machine translation (MT) systems.

1 Description

Many studies have demonstrated that translated texts have different textual features than texts originally written in the given language (originals). Furthermore, some studies have shown evidence of variation between human translations generated by different translators (Rubino et al., 2016; Popović, 2020; Kunilovskaya and Lapshinova-Koltunski, 2020). Nevertheless, the number of such studies is still very small and limited to comparable corpora where different translators translated different source texts. Therefore, exact comparisons between human translations are not possible.

The DiHuTra project, formed by Saarland University, ADAPT Centre and University of Eastern Finland in 2021–2022 has aimed to design a parallel corpus to address these issues. Each source text originally written in English has been translated into three target languages: Croatian, Russian and Finnish, by two groups of translators: professionals and students. These parallel human translations will enable a better comparison of various text features as well as impact of automatic MT evaluation when used as references.

2 Data sets

The source texts consist of two sub-sets of publicly available data sets from two distinct domains:

Amazon product reviews[1] contain unique product reviews from Amazon written in English with overall ratings from 1 to 5, 1 and 2 referring to negative, 3 to neutral and 4 and 5 to positive. We selected a balanced set of reviews from 14 categories (e.g., "Sports and Outdoors", "Books", etc.) with an equal number of positive and negative reviews (14 from each of the 14 topics). In total, we included 196 reviews, containing 5.4 sentences and 93.2 words on average.

News texts were imported from the WMT (2019 and 2020) shared task[2] News test corpus. The topics vary between politics, sports, crime, health, etc. The news are longer than reviews, with 9.9 sentences and 221.7 words on average. The WMT shared tasks also contain a set of human translations of the English source texts into several languages including Russian, however, neither Croatian nor Finnish. We selected only texts which were originally written in English and had professional translations into Russian. In total, we included 68 news articles from different sources.

3 Translation process

Each English review was translated into the three target languages, Croatian, Russian and Finnish, by professionals and by students. For the news

[1] http://jmcauley.ucsd.edu/data/amazon/
[2] http://www.statmt.org/wmt20/translation-task.html

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

Table 1: Text statistics and lexical variety: (a) total number of words, (b) total number of running words, (c) ratio between vocabulary and words ↑, (d) Yule's K coefficient ↓.

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, information about age, gender, experience and the study program (for students) was collected. Translators were asked to keep the sentence alignment (not to merge or to split sentences) and not to use MT. No further restrictions were given to translators. The total number of tokens in the resulting corpus amounts to 180,584.

4 Corpus statistics

The first statistics on the shallow features in terms of running words and vocabulary in the sources and the three target languages (see Table 1). We also estimated lexical richness in terms of ratio between vocabulary 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 vocabulary/words ratio (higher value indicates richer vocabulary) and inversely proportional to Yule's K (a lower value indicates a richer vocabulary).

The corpus is valuable for studying variation in translation as it allows direct comparisons between human translations of the same source texts. Our preliminary analyses based on the shallow text statistics and matching/distance measures indicate that students used shorter sentences but richer vocabulary. To better understand these differences, we plan to carry out detailed analyses on the annotated data (we have tokenised, lemmatised, parts-of-speech tagged and parsed the data using universal dependencies). This resource is also valuable for evaluation of MT systems for the three language pairs. The Croatian (and probably Russian) part of the user reviews will be used in the WMT shared task in 2022.[3] We believe that this resource will help us to understand and improve quality issues in both human and machine translation.

The corpus is available via CLARIN[4]. The project has also a GitHub repository[5] which contains the data and some additional information. The details about the corpus can be found in (Lapshinova-Koltunski et al., 2022).

5 Acknowledgments

The creation of the corpus was supported through the EAMT sponsorship programme (2021) and by ADAPT Centre. The ADAPT Centre is funded by through the SFI Research Centres Programme and co-funded under the ERDF through Grant 13/RC/2106. The Finnish subcorpus was supported by a Kopiosto grant awarded by the Finnish Association of Translators and Interpreters. We thank the translators in Volgograd, Zagreb, Rijeka and Finland. In particular, we thank Aleksandr Besedin from VolSU for coordinating the work of the Russian translators.

References

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

Lapshinova-Koltunski, E., Popović, M., and Koponen, M. (2022). Dihutra: a parallel corpus to analyse differences between human translations. In *Proceedings of LREC 2022*, Marseille, France, June.

Popović, M. (2020). On the differences between human translations. In *Proceedings of the EAMT 2020*, pages 365–374, Lisboa, Portugal, November.

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 NAACL-HLT 2016*, pages 960–970, San Diego, California, June.

[3] https://machinetranslate.org/wmt22

[4] https://fedora.clarin-d.uni-saarland.de/dihutra/index.html

[5] https://github.com/katjakaterina/dihutra