Can Translation Task Difficulty Predict the Properties of Translation?

Maria Kunilovskaya, Elke Teich (University of Saarland),
Ekaterina Lapshinova-Koltunski (University of Hildesheim)
maria.kunilovskaya@uni-saarland.de

Translations are known to have lexical, morphosyntactic and semantic deviations from comparable originally-authored target language (TL). These deviations are known as translationese. Translation studies accumulated extensive evidence of translationese to raise concerns in the related research fields, such as machine translation (Artetxe et al., 2020) and contrastive studies (De Baets et al., 2020). Pinpointing the factors that contribute to translationese deviations remains a challenging task. The explanatory efforts link translationese to trends in translational behaviour (simplification, explicitation, etc), to socio-cultural factors (expertise, registers) or to the cross-linguistic nature of the translation process. It has been shown that more challenging cognitive conditions may trigger a simpler, more conventionalised (Kruger & De Sutter, 2018), more explicit (Olohan & Baker, 2000) or more implicit output (Lapshinova-Koltunski et al., 2022).

The current project aims to explore the impact of the source document complexity on the properties of translations. We rely on a range of complexity measures, including information-theoretical indices.

Previous work used average sentence surprisal to show that the amount of information in the source is positively correlated with the amount of information in the target regardless of the translation mode (Kunilovskaya et al., 2023; Przybyl et al., 2022). Other computational studies have compelling evidence that translationese deviations can be explained by the source language (SL) influence (Rabinovich, Ordan & Wintner, 2017; Evert & Neumann, 2017; Kunilovskaya & Lapshinova-Koltunski, 2020). The novelty of our approach consists (I) in the truly cross-lingual nature of the experiments where the values for the source document are used to predict the properties of the translation, (ii) in the focus on the sub-sentential source and target items (content tokens and syntactic subtrees) to represent document pairs, and (iii) in indexing the two subprocesses involved in translation: source text comprehension and SL-TL transfer. The comprehension difficulty is captured by measures of syntactic complexity such as dependency distance, hierarchical distance, tree depth, and branching factor as well as by the surprisal of the content items from GPT-2. The transfer difficulty is operationalised (a) as the entropy of translation variants for a SL item registered in a large parallel corpus and (b) as the semantic alignment score (cosine similarity between contextualised embeddings of content items). The response variable – translationese properties of target – is a document-level probability of being a translation from a classifier that can reliably distinguish translations and comparable non-translations in the TL (F1-score 80-90%) using hand-engineered delexicalised translationese predictors.

Theoretically, the more demanding SL documents can be expected to generate more deviant translations. If this is the case, translationese can be explained as a rational response to increased cognitive pressure on the assumption that producing deviant translations requires less production effort.

The preliminary regression results on a large bi-directional Europarl corpus (English-German) show that transfer difficulty indicators are more relevant to the task than syntactic complexity measures or lexical comprehension difficulty measured as surprisal of the source content items, although the correlation is very weak.

References

Artetxe, M., Labaka, G., Agirre, E., & Center, H. (2020). Translation Artifacts in Cross-lingual Transfer Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, *(EMNLP)* (pp. 7674–7684). https://github.com/pytorch/fairseq

De Baets, P., Vandevoorde, L., & De Sutter, G. (2020). On the usefulness of comparable and parallel corpora for contrastive linguistics. testing the semantic stability hypothesis. *New Approaches to Contrastive Linguistics. Empirical and Methodological Challenges*. Berlin/Boston: De Gruyter, 85–126.

Evert, S., & 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, 47–80.

Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, 1(2), 251–290.

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2020, May). Lexicogrammatic translationese across two targets and competence levels. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4102-4112).

Kunilovskaya, M., Przybyl, H., Teich, E., & Lapshinova-Koltunski, E. (2023, April). Simultaneous Interpreting as a Noisy Channel: How Much Information Gets Through. In *Proceedings of the international conference on recent advances in natural language processing* (pp. 608–618). INCOMA Ltd. https://doi.org/10.26615/978-954-452-092-2\ 066

Lapshinova-Koltunski, E., Pollkläsener, C., & Przybyl, H. (2022). Exploring explicitation and implicitation in parallel interpreting and translation corpora. *The Prague Bulletin of Mathematical Linguistics*, 119, 5–22. https://ufal.mff.cuni.cz/pbml/119/art-lapshinova-koltunski-pollklaesener-przybyl.pdf

Olohan, M., & Baker, M. (2000). Reporting that in translated english. Evidence for subconscious processes of explicitation? *Across languages and cultures*, 1(2), 141–158.

Przybyl, H., Karakanta, A., Menzel, K., & Teich, E. (2022). Exploring linguistic variation in mediated discourse: Translation vs. interpreting. In *Mediated discourse at the european parliament: Empirical investigations* (pp. 191–218). Language Science Press. https://doi.org/10.5281/zenodo.6977050

Rabinovich, E., Ordan, N., & Wintner, S. (2017, July). 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)* (pp. 530-540).