Impact of Syntactic Complexity on the Processes and Performance of Large Language Models-Leveraged Post-editing

Longhui Zou lzou4@kent.edu
Michael Carl mcarl6@kent.edu
CRITT, Kent State University, Kent, 44240, USA

Shaghayegh Momtaz momtazshaghayegh@gmail.com

Mehdi Mirzapour mehdi@proptexx.com
PropTexx, USA

Abstract

This research explores the interaction between human translators and Large Language Models (LLMs) during post-editing (PE). The study examines the impact of syntactic complexity on the PE processes and performance, specifically when working with the raw translation output generated by GPT-4. We selected four English source texts (STs) from previous American Translators Association (ATA) certification examinations. Each text is about 10 segments, with 250 words. GPT-4 was employed to translate the four STs from English into simplified Chinese. The empirical experiment simulated the authentic work environment of PE, using professional computer-assisted translation (CAT) tool, Trados. The raw translation output generated by GPT-4 was used to prepare the translation memory (TM) for the participants, and 13 words or phrases in the STs were selected to generate a term base (TB) with the English source terms and their equivalent Chinese target terms. The experiment involved 46 participants with different levels of translation expertise (30 student translators and 16 expert translators), producing altogether 2162 segments of PE versions for comparative analysis.

We implemented five syntactic complexity metrics in the context of PE, on the source text (ST) side, machine translation (MT) side, and the target text (TT) side. The metrics are chosen based on the specific syntactic difference between English and Chinese, including Incomplete Dependency Theory Metric (IDT), Dependency Locality Theory Metric (DLT), Combined IDT+DLT Metric (IDT+DLT), Left-Embeddedness (LE) and Nested Nouns Distance (NND). IDT, DLT, and IDT+DLT, are applications of linguistic complexity theories from Gibson's Incomplete Dependency Theory (IDT) and Dependency Locality Theory (DLT) (Gibson, 1998; Gibson, 2000). The metric LE is adopted and slightly modified from Coh-Metrix analysis (Graesser et al, 2011). NND is introduced in (Zou et al., 2021).

In this study, the participants' task was to post-edit the raw ChatGPT translation output, adhering to two different levels of PE guidelines (light PE [LPE] and full PE [FPE]) in Trados, according to Translation Automation User Society (TAUS) guidelines (Massardo et al., 2017). We also controlled two conditions of external search for the PE experiment: i.e., TB provided within Trados interface but no access to other external resources (TB), and access to any internet search but no TB provided within Trados interface (IS). Therefore, each participant conducted the PE of the four texts under four tasks, sequentially (i.e., LPE+TB; LPE+IS; FPE+TB; and FPE+IS).

The keystroke data during the PE sessions were recorded by both the Qualitivity plugin for Trados and Tobii Studio. The translator's eye movement data were collected with a Tobii TX 300 eye tracker. The translation process data was then converted and processed by the Trados-to-Translog II interface available at the CRITT TPR-DB (Zou et al., 2023; Zou and Carl, 2022; Yamada et al., 2022). Manual Quality assessment of the raw GPT-4 translations and the post-edited translations by human translators were conducted by ten professional translators, using an ATA-adapted error taxonomy.

Our preliminary findings demonstrate that there are significantly positive correlations between the IDT, IDT+DLT, LE and NND metrics and GPT-generated error counts. We also found that language-specific syntactic differences between English and Chinese such as directions of branching (LE) and noun modifiers (NND) can have a significantly positive influence on accuracy and minor errors in students' PE versions. Furthermore, expert translators produced significantly less fluency errors under the FPE guideline as compared to LPE guideline, whereas student translators had significantly less accuracy errors in the TB condition as compared to internet search (IS). Expert translators generally display greater mastery in understanding translation briefs and research skills compared to student translators. Process data of the student translators indicates less efficient workflows compared to experts (Hvelplund, 2016). Expert translators showed more fluent typing and less revision and refixation behavior than student translators (Tirkkonen-Condit, 2005; Carl & Schaeffer, 2017). These results suggest the need to adapt translation curricula to equip student translators with the LLMs-leveraged translation literacy, specialized research skill, and technical proficiency required for their professional advancement in generative AI-assisted translation roles.

References

Carl, M., & Schaeffer, M. J. (2017). Models of the translation process. *The handbook of translation and cognition*, 50-70.

Flanagan, M., & Christensen, T. P. (2014). Testing post-editing guidelines: how translation trainees interpret them and how to tailor them for translator training purposes. *The Interpreter and Translator Trainer*, 8(2), 257-275.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5), 223-234.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1), 1-76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*, 2000, 95-126.

Guerreiro, N. M., Rei, R., van Stigt, D., Coheur, L., Colombo, P., & Martins, A. F. (2023). xcomet: Transparent machine translation evaluation through fine-grained error detection. arXiv preprint arXiv:2310.10482.

Hvelplund, K. T. (2016). Cognitive efficiency in translation. *Reembedding Translation Process Research*. Amsterdam and Philadelphia: John Benjamins.

Massardo, I., van der Meer, J., O'Brien, S., Hollowood, F., Aranberri, N., and Drescher, K. (2017). Taus mt post-editing guidelines.

Tirkkonen-Condit, S. (2005). The monitor model revisited: Evidence from process research. *Meta*, 50(2), 405-414.

Yamada, M., Mizowaki, T., Zou, L., and Carl, M. (2022). Trados-to-translog-II: Adding gaze and qualitivity data to the CRITT TPR-DB. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 293–294, Ghent, Belgium. European Association for Machine Translation.

Yang, Z. and Li, D. (2021). Translation competence revisited: Toward a pedagogical model of translation competence. *Advances in Cognitive Translation Studies*, pages 109–138.

Zou, L., Carl, M., Mirzapour, M., Jacquenet, H., & Vieira, L. N. (2021, December). Ai-based syntactic complexity metrics and sight interpreting performance. In *International Conference on Intelligent Human Computer Interaction* (pp. 534-547). Cham: Springer International Publishing.

Zou, L. and Carl, M. (2022). Trados and the critt tpr-db: Translation process research in an ecologically valid environment. In *Model building in empirical translation studies: Proceedings of TRICKLET Conference*, May 19-20, 2022, pages 38–40.

Zou, L., Carl, M., & Gilbert, D. (2023). Integrating Trados-Qualitivity Data to the CRITT TPR-DB: Measuring Post-editing Process Data in an Ecologically Valid Setting. In *Corpora and Translation Education: Advances and Challenges* (pp. 63-86). Singapore: Springer Nature Singapore.