Translation Quality and Effort Prediction in Professional Machine Translation Post-Editing

Jennifer Vardaro
Johannes Gutenberg University
Mainz, Germany
vardaro@uni-mainz.de

Moritz Schaeffer
Johannes Gutenberg University
Mainz, Germany
mschae01@uni-mainz.de

Silvia Hansen-Schirra
Johannes Gutenberg University
Mainz, Germany
hansenss@uni-mainz.de

Abstract

The focus of this controlled eye-tracking and key-logging study is to analyze the behaviour of translation professionals at the European Commission's Directorate-General for Translation (DGT) when detecting and correcting errors in neural machine translated texts (NMT) and their post-edited versions (NMTPE). The experiment was informed by quality analyses of an authentic DGT parallel corpus (Vardaro, Schaeffer, and Hansen-Schirra 2019), consisting of English source texts and corresponding German NMT, NMTPE and revisions (REV). To identify the most characteristic error categories in NMT and NMTPE, we used the automatic error annotation tool Hjerson (Popović 2011) and the more fine-grained manual MQM framework (Lommel 2014). Results show that quality assurance measures by post-editors and revisors at the DGT are most often necessary for lexical errors. More specifically, if post-editors correct mistranslations, terminology or stylistic errors in an NMT sentence, revisors are likely to correct the same type of error in the same sentence, suggesting a certain transitivity between the NMT system and human post-editors.

In this study, carried out in Translog II (Carl 2012), participants' eye movements and typing behavior for test sentences where the error categories mistranslation, terminology, function words and stylistic errors are included will be compared to control sentences without errors. 30 language professionals from the DGT post-edited 100 English-German machine translated sentences from the DGT corpus. We examine the three error types' effect on early (first fixation durations, first pass durations) and late eye movement measures (e.g., total reading time and regression path duration) and on typing behaviour. Statistical regression analyses predict the temporal, technical, and cognitive effort during the DGT post-editing and revision process which will be corelated to the recognition and correction of said error categories. In addition, the behavioural data of the DGT translation professionals will be compared to those of a group of 30 translation students. Behavioural differences in the two groups will allow for further predictions regarding the effect of expertise on the post-editing process.in

References

Carl, Michael. 2012. 'Translog-II: A Program for Recording User Activity Data for Empirical Translation Process Research'. *International Journal of Computational Linguistics and Applications*, 2012.

Lommel, Arle. 2014. 'Multidimensional Quality Metrics Definition'. 2014. http://www.qt21.eu/mqm-definition/definition-2015-06-16.html.

Popović, Maja. 2011. 'Hjerson: An Open Source Tool for Automatic Error Classification of Machine Translation Output'. *The Prague Bulletin of Mathematical Linguistics* 96 (1). https://doi.org/10.2478/v10108-011-0011-4.

Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra. 2019. 'Comparing the Quality of Neural Machine Translation and Professional Post-Editing'. In *Proceedings of QoMEX*, 1–3. Berlin, Germany.
https://doi.org/10.1109/QoMEX.2019.8743218.