Error Annotation in Post-Editing Machine Translation: Investigating the Impact of Text-to-Speech Technology

Justus Brockmann
Centre for Translation Studies
University of Vienna
justus.brockmann@univie.ac.at

Claudia Wiesinger
Centre for Translation Studies
University of Vienna
claudia.wiesinger@univie.ac.at

Dragoș Ciobanu
Centre for Translation Studies
University of Vienna
dragos.ioan.ciobanu@univie.ac.at

Abstract

As post-editing of machine translation (PEMT) is becoming one of the most dominant services offered by the language services industry (LSI), efforts are being made to support the provision of this service with additional technologies. We present text-to-speech (T2S) as a potential attention-raising technology for post-editors. Our study was conducted with university students and included both PEMT and MT error annotation of a creative text with and without T2S. Focusing on the error annotation data, our analysis finds that participants under-annotated fewer MT errors in the T2S condition compared to the silent condition. At the same time, more over-annotation was recorded. Finally, annotation performance corresponded to participants' attitudes towards using T2S.

1 Introduction

With machine translation (MT) adoption and the provision of post-editing machine translation (PEMT) services on the rise, Translation Process Research (TPR) has been questioning whether the ways in which PEMT is currently being carried out (in dedicated PEMT tools, in simple word processing software, or in computer-assisted translation (CAT) tools/translation environment tools (TEnT), and with or without the use of additional technologies) optimally support post-editors, both from a process- and a product-oriented point of view (Moorkens and O'Brien, 2017). As the technological possibilities are growing, there is an uptake of speech tools such as automatic speech recognition (ASR; speech-to-text) by professional translators (ELIA et al., 2022), and this practice has become one of the focal points of TPR (Dragsted, Mees and Hansen, 2011; Ciobanu, 2014, 2016; Mesa-Lao, 2014; Zapata, Castilho and Moorkens, 2017; Liyanapathirana, 2021).

While the use of automatic speech synthesis (text-to-speech; T2S) has received comparatively little attention both from the language services industry (LSI) and the research community, translators and revisers are known to read aloud translations during (self-)revision (Allain, 2010; Ciobanu, 2016; Scocchera, 2017). This intuitively perceived benefit of aurally processing a text points to the potential of T2S as an attention-raising technology that may also help post-editors identify subtle neural machine translation (NMT) errors.

The practice of PEMT remains a particular challenge for Translation Studies students, despite the transition from statistical MT (SMT) to NMT which has reduced the absolute number of errors to be corrected in the raw MT output (Yamada, 2019). Moreover, the phenomena of over- and under-editing continue to preoccupy both academia and the LSI (Nitzke and Gros, 2020). We share the view that students need to be exposed to a variety of translation tools early and often, and we believe that introducing them to additional technologies such as T2S will prove beneficial for honing the skills needed to succeed as future post-editors. Rather than segregating tools and technologies to separate courses, we support integrative tasks which combine error annotation using standardised typologies, PEMT,

and T2S as ideal opportunities to build confidence, competence, and speed when performing PEMT.

This paper describes the results of a small-scale study investigating the impact of T2S on PEMT error annotation, alongside participant attitudes towards using T2S for PEMT.

To that end, we first present previous research on the use of speech tools for PEMT, as well as the use of error typologies in the LSI and in translator training. This is followed by our research questions and methodology. In the last two sections we present the results of our study and discuss the implications of teaching PEMT by introducing T2S and error annotation into the mix.

2 Previous Research

PEMT has been identified as the service with the highest growth potential in the LSI (ELIA et al., 2022). The widespread adoption of data-driven MT since the 2000s (Kenny, 2020) has brought considerable change to the industry, and professional translators are increasingly being asked to carry out PEMT tasks. While claims of MT achieving near or full human parity in terms of translation quality (Wu et al., 2016; Hassan et al., 2018) should be taken with a grain of salt (Läubli, Sennrich and Volk, 2018), MT has been shown to enable productivity and quality gains in translation tasks (e.g. Guerberof Arenas, 2014; Sánchez-Gijón, Moorkens and Way, 2019).

However, despite MT quality improvements and the clear industry need for qualified post-editors (most recently embodied by the GALA MTPE Training Special Interest Group[1]), European Translation Studies programmes have been found to lack hands-on PEMT training both at undergraduate and postgraduate levels (Ginovart Cid and Colominas Ventura, 2020). While interest in MT literacy is growing in the research community (cf. Bowker and Ciro, 2019), many translators are still reluctant to embrace MT as a tool (ELIA et al., 2022). Limited knowledge and experience regarding MT use in university-trained translators is likely to be a contributing factor to this reticence.

In parallel to the lack of hands-on PEMT training in Translation Studies syllabi, previous work has also highlighted a lack of familiarity of translation educators and students with translation quality assessment (TQA) practices (Doherty et al., 2018). This training blind spot may come as a surprise since TQA practices, which include the use of error typologies and scorecards, are common in the LSI (Lommel, 2018).

Quality management, which frequently involves TQA processes, has been identified as a key competence for professional translators (European Master's in Translation, 2017), and in the context of MT, the ability to perform TQA in the form of error annotation with predefined typologies is a useful skill for engine evaluation and PEMT research, among others (Popović, 2018). Moreover, the active reflection on error types may help improve the current issue of over- and under-editing, which is common in PEMT (Nitzke and Gros, 2020). There is therefore a competence gap between academia and industry in relation to both PEMT and TQA practices.

Interest in MT is growing in the LSI; however, it has been shown that translation tools do not optimally support post-editors, which leads to dissatisfaction among users (Moorkens and O'Brien, 2017). In parallel, dictating with automatic speech recognition (ASR) tools instead of, or in addition to, typing has been recognised as an alternative, more ergonomic working mode, and has attracted the interest of several scholars (Dragsted, Mees and Hansen, 2011; Ciobanu, 2014, 2016; Mesa-Lao, 2014; Zapata, Castilho and Moorkens, 2017; Liyanapathirana, 2021). ASR is also seeing an uptake among professional translators (ELIA et al., 2022). Consequently, new applications offering multi-modal forms of translator-computer interaction (TCI) have been developed (Teixeira et al., 2019; Herbig et al., 2020). In these examples, multimodal features include the use of ASR for translation and PEMT. In Interpreting Studies, the integration of ASR into computer-aided interpreting tools is also being investigated to support the work of interpreters (Fantinuoli, 2017; Defrancq and Fantinuoli, 2021).

Comparatively little attention has so far been given to potential applications of text-to-speech (T2S) technology in translation, revision, and PEMT tasks, which allow translators/post-editors to listen to an artificial computer voice 'reading out' the text they are working on. While the tools currently used in the LSI do not support T2S by default and only Trados Studio offers a T2S plug-in[2] to date, there is evidence of translators seeking

[1] https://www.gala-global.org/knowledge-center/professional-development/sigs

[2] https://community.rws.com/product-groups/trados-portfolio/

other ways of aurally processing text in their work (Allain, 2010; Ciobanu, 2016; Scocchera, 2017). A study that introduced T2S in the translation revision process (Ciobanu, Ragni and Secară, 2019) yielded encouraging results regarding revisers' error correction performance; however, further research on the effects of T2S on translators' work is certainly needed (Ciobanu and Secară, 2020).

The study by Ciobanu, Ragni and Secară (2019) found revision with T2S to be conducive to correcting more errors – above all Accuracy errors – compared to revision in silence. This has promising implications for the integration of T2S in PEMT since Accuracy has been identified as one of the major challenges for NMT (Vardaro, Schaeffer and Hansen-Schirra, 2019). Given that the use of T2S seemed to have an attention-raising effect in the revision study, we contend that this technology may also be beneficial for PEMT and error annotation – especially for translation students.

To our knowledge, there is a lack of empirical evidence on: (i) the impact of T2S technology on PEMT performance, productivity, and post-editors' attitudes towards this mode of working; and (ii) the impact of T2S on error annotation performance in the context of translator training. We aimed to fill these research gaps with a small-scale study conducted with 17 university students.

3 Methodology

3.1 Study design

The study involved 16 undergraduate students of Transcultural Communication and 1 postgraduate student of Translation. Participants were quasi-randomly allocated to two groups, G1 and G2 based on their responses to a pre-experiment questionnaire. The groups were roughly balanced regarding the participants' language skills and translation experience. Most participants were German native speakers with an English language level of C1 and very limited translation experience. Due to constraints imposed by the COVID-19 pandemic, the study was carried out fully online. In order to control the experiment conditions in this online setting, the participants were asked to work in front of their active webcams and to observe strict time limits.

The source text we used in our study was a 1,800-word excerpt from the 2019 stage adaptation of Hanif Kureishi's 1985 screenplay *My Beautiful Laundrette*. In an exploratory preparation stage, this English text was translated into German with the freely available MT engines DeepL[3], Microsoft Translator[4], and Google Translate[5]. The resulting raw MT output was evaluated by a member of the research team through error annotation according to the DQF subset of the harmonised DQF-MQM error typology[6]. We decided on using the output from Google Translate in our experiment because it contained fewer errors than the output from Microsoft Translator, and more errors than the output from DeepL, thus qualifying as a moderate PEMT challenge for our participants. We then split the source text into four equal parts of roughly 450 source words each to obtain texts of comparable length for our four experiment conditions.

Participants carried out the error annotation and PEMT tasks in Microsoft Word 365. The built-in Read Aloud function in Microsoft Word was used for synthetic voices, allowing participants to access both source and target text speech synthesis seamlessly during the final condition regardless of their computers' operating systems and without making major changes to their previous working environment. The source and target texts were displayed in a three-column table format. Each table cell represented one segment from the stage play script. The first column contained the English source text, the second and third columns contained identical copies of the German output from Google Translate. This way, participants could annotate the MT errors in the second column and post-edit the output in the third column, thus providing a more convenient way of working than combining annotations and post-edits in a single cell.

Prior to the experiment, our participants attended an introductory workshop in which they practised PEMT and error annotation on a 124-word excerpt from the play. This was done in preparation for the actual experiment tasks, which required the participants to both annotate and post-

[3] https://www.deepl.com/; translation retrieval date: 7/04/2021

[4] https://www.bing.com/translator; translation retrieval date: 7/04/2021

[5] https://translate.google.com/; translation retrieval date: 7/04/2021

[6] https://www.taus.net/qt21-project#harmonized-error-typology

edit the four target text parts during two separate experiment sessions with two parts each. Our participants were provided with instructions both during the introductory workshop and in writing on how to use the T2S functionality in the PEMT task, as well as how to change the synthetic voices if desired. The written instructions also included relevant keyboard shortcuts for Windows and MacOS that the participants could use to increase productivity when using T2S: play/pause/skip-back/skip-forward/increase or decrease reading speed.

The two experiment sessions were carried out on two separate days within a two-week interval. Each session was split up into two 45-minute parts, and in each part the participants carried out their task in a different working condition: 1. in silence for both groups; 2. with source text sound (STS), or with target text sound (TTS), depending on the group; 3. with TTS, or STS, again depending on the group; and 4. with both STS and TTS. For the working conditions that included T2S, the students were instructed to use the speech functionality for each segment they worked on at least once. We reversed the order in which the two groups were confronted with the first sound condition to counteract the potential influence of growing familiarity with the text (**Table 1**).

	Part 1	**Part 2**	**Part 3**	**Part 4**
G1 (n=9) P2, P3, P5, P6, P10, P14, P15, P18, P21	Silence	STS	TTS	STS+ TTS
G2 (n=8) P1, P4, P7, P8, P9, P16, P19, P23	Silence	TTS	STS	STS+ TTS

Table 1: Distribution of experimental groups, parts, and sound conditions

This paper focuses on the annotations made in Part 1 (silence) and Part 4 (STS+TTS) for two main reasons: firstly, as reported in Wiesinger et al. (forthcoming), our participants' average PEMT performance was highest in Part 4, both in terms of error correction rate and productivity. Secondly, these were the two parts where all participants worked both on the same content, and in the same conditions – i.e., in silence in Part 1 and with both types of sound in Part 4.

The experiment sessions were followed by a feedback meeting, allowing the participants to ask questions and to compare their performance. Moreover, a total of six questionnaires were answered by the participants throughout the experiment: one during recruitment, one after each part, and one after the feedback meeting, allowing us to collect data on their prior experience, perceived performance, and evolving attitudes.

For the error annotation task, the participants were introduced to the DQF subset of the harmonised DQF-MQM error typology, which contains eight high-level error types and 33 granular error types. The typology also features four severity levels to add a weight to errors, complemented by a 'Kudos' option to praise exceptional performance. Participants were instructed to use the numerical identifier assigned to each high-level and granular error type, as well as severity level when making annotations. This way, a 'Mistranslation' error with 'Major' severity, for instance, would be annotated via the MS Word comment function with the label: 1–13–2 (i.e., Accuracy–Mistranslation–Major).

3.2 The Gold Standard

In order to establish a reference against which the participants' submissions could be compared, two members of the research team annotated and post-edited the MT output, and merged their annotations by mutual agreement into a gold standard version. For the purpose of our study, this gold standard was assumed to contain all of the errors that needed to be corrected in the MT output: 91 errors in Part 1, 75 errors in Part 2, 62 errors in Part 3, and 45 errors in Part 4.

3.3 Complementary work

Complementary work in Wiesinger et al. (forthcoming) has involved an analysis of the study data regarding the effect of T2S on post-editing performance and productivity. The final experiment condition (STS+TTS) resulted in the highest proportion of MT errors corrected in line with our gold standard. Although productivity grew on average, we saw that the highest improvement in PEMT quality came with the lowest improvement in productivity.

In the present analysis we re-visit the data collected in the study, focusing in more detail on the impact of T2S on the high-level error types annotated by the participants, as well as the

relationship between the participants' attitudes and their annotation performance.

3.4 Research questions

Our research questions were:

- RQ1: Which of the two conditions (silence, or STS+TTS) is more conducive to over-annotation?

- RQ2: Which of the two conditions (silence, or STS+TTS) is more conducive to under-annotation?

- RQ3: What is the relationship between the participants' attitudes and their error annotation performance?

4 Results

4.1 Error annotation

We measured our participants' annotation performance in Part 1 and Part 4 by comparing each participant's annotations against our gold standard (GS) annotated version which contained 91 errors in Part 1, and 45 in Part 4.

'Over-annotation' refers to cases where the participant annotated an error not present in the GS. On average, 21% of the total annotations made by our participants were labelled as over-annotations in the silence condition (Part 1). For the STS+TTS condition (Part 4), the average percentage was 34%.

'Under-annotation' refers to cases where the participant did not annotate an error present in the GS. On average, 52% of the errors present in the Part 1 MT output were not annotated. For Part 4, the figure was 46%.

Since the participants were asked to observe strict time limits for the experiment parts, the amount of text they managed to annotate varied depending on individual productivity. We took this into account in our calculations. Over-annotations were calculated as percentages of the total number of annotations each participant made in the respective part. Under-annotations were calculated as the percentage of GS errors present but left un-annotated in the portion of text they worked on in each part.

However, averages only tell part of the story. Predictably, we observed that not all participants annotated the same number of errors in the two parts. There was, in fact, considerable variation among participants.

Over-annotation went up for all but two participants (**Figure 1**): P8, who registered a slight decrease, and P10 (no change). This is not surprising, given that the total number of annotations made by all participants remained almost the same (492 in Part 1, 487 in Part 4), but the number of errors present in the GS halved from Part 1 to Part 4 (91 in Part 1, 45 in Part 4). Possible reasons for the increase in over-annotation include that some participants might have been trying to annotate errors at a similar or higher rate than in the previous parts, or that their approach to translation defects was more critical in the sound conditions. However, these speculations could only be confirmed by obtaining more qualitative data on the process from participants.

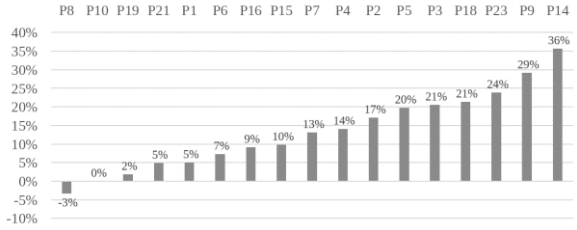

Figure 1: Increases/decreases in **over-annotations** made by participants in Part 4 compared to Part 1

On the other hand, under-annotation went down in Part 4 (STS + TTS) for 12 of the 17 participants, with decreases ranging from 2 to 27 percentage points (**Figure 2)**.

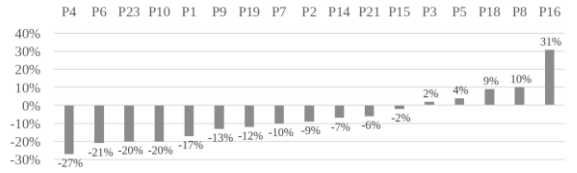

Figure 2: Increases/decreases in **under-annotations** made by participants in Part 4 compared to Part 1

4.2 Attitudes

When looking at the responses to a pre-experiment questionnaire item that asked whether the participants see any major advantages or disadvantages in using T2S, we can broadly classify the answers given by the participants as indicating a positive, neutral, or negative attitude. A positive answer is one where the participant expects advantages from the use of T2S. In a neutral answer, the participant indicates that they are unsure about any advantages or disadvantages. In a negative answer, the participant would state that they expect disadvantages from using T2S or prefer working without it. Generally, our

participants' answers indicated a largely positive attitude towards using T2S.

Of the 17 participants, there were only 6 who indicated a negative attitude towards T2S in the pre-experiment questionnaire. Three of them changed their minds over the course of the experiment, indicating positive attitudes in the final questionnaire after the experiment. This leaves three participants (P9, P14, P16) who kept their negative attitudes towards T2S even after using the technology.

It should also be noted that none of the participants changed their attitude towards using T2S to negative after the experiment.

Moreover, the attitudes towards annotating errors during PEMT were also largely positive. In the questionnaire answered after completing Part 1, only 5 out of the 17 participants indicated that they did not see any advantages in PEMT with error annotation.

5 Discussion

In an ideal world, introducing this new mode of working would enable post-editors to reduce both their over-annotation and under-annotation scores.

In response to RQ1, we observed that STS+TTS was the condition in which all participants except two annotated more errors which were not actually there – so their over-annotation scores went up, in some cases by over 20% (**Figure 1**). This is not necessarily detrimental to the target text, although it lowers the post-editor's productivity.

At the same time (and in response to RQ2), the STS+TTS condition was also the condition in which fewer actual GS errors were missed by all but 5 participants (**Figure 2**). While there is an outlier here with an increase in under-annotation of 31 percentage points (P16), qualitative data revealed that this participant experienced technical difficulties in using the Read Aloud feature – thus offering an example of the detrimental impact on performance posed by user-specific technical challenges.

Overall, missing fewer real errors is extremely valuable and can improve target text quality if corrected well, provided that the over-annotations and their corresponding corrections do not introduce new errors.

Our data suggests that, when performing PEMT with STS+TTS, participants made more preferential annotations, but also missed fewer genuine errors. In the words of P18: "By listening to the segments in the target language that were translated only by a machine, I can detect errors more easily as the translation sounds unnatural to me." Although not ideal – the ideal would be for post-editors to only make necessary annotations –, identifying more genuine errors while also making what could be classed as 'preferential annotations' could be considered an acceptable compromise.

In any case, what these figures show is that limited practice without personalised feedback does not result in ideal performance improvements for an entire group, although encouraging signs could already be seen. For example, at the end of the experiment, for 5 students the percentage by which they over-annotated was actually below the one by which they decreased their under-annotation performance. This is a move in the right direction. 5 different students, though, were at the other end of the spectrum, with both higher over-annotations (which is tolerable) **and** higher under-annotations (which is not ideal).

With sufficient practice, though, annotating errors and subsequently correcting them can reach a level of quality which makes this task useful not just for an individual – "You have a clearer picture of what kind of errors you have to correct" (P6) – but also for a group collaborating on a PEMT project – "It is helpful if you work with others; in that case you don't have to explain to them your decision every time. And if the person you are doing the post-editing for wants to know why you corrected something, it is easier to explain." (P10)

Furthermore, the qualitative data obtained from the questionnaires (RQ3) suggest that the perception of T2S as a useful tool for error annotation and PEMT will depend on personal preferences and attitudes.

The three participants who did not change their negative attitudes towards T2S were also among those whose error annotation performance changed for the worse between Part 1 and Part 4. P16 had the largest increase in under-annotation (31 percentage points), while P9 and P14 had the largest increases in over-annotation (29 and 36 percentage points, respectively).

Conversely, those whose error annotation performance changed for the better between Part 1 and Part 4 generally indicated positive attitudes towards the use of T2S. Of the three students with the highest decrease in under-annotation (P4, P6, P23), the first two indicated a positive attitude

before and after the experiment, while P23 changed their attitude from negative to positive in the final questionnaire. P23 shares third place in reducing under-annotation with P10 who kept a neutral attitude throughout the experiment. The only participant to reduce over-annotation (P8) had a positive attitude throughout.

Other participants perceived speech synthesis as beneficial for text comprehension more generally: "Speech synthesis made understanding sentences with slang words much easier. I could understand the spoken words in the context of the sentence better, even though I had never heard them before." (P21)

6 Conclusions

Despite rapid advances in technologies such as machine translation and speech synthesis, the professional environments in which translators, revisers, and post-editors work have remained largely unchanged.

Post-editors are expected to identify and correct at an ever faster rate the unpredictable and often subtle errors produced by neural machine translation engines, but their attention is not yet enhanced and stimulated by multi-modal input. Our experiment shows that integrating S2T into PEMT workflows can be easily done with existing tools and has practical benefits – similar to how integrating T2S into *revision* workflows improved revisers' performance in a previous experiment.

Moreover, although both the task and the technologies used in the experiment were unfamiliar to the participants, progress was recorded to different degrees concerning performance and attitudes. Continued practice supplemented by regular, personalised feedback is likely to accelerate such progress.

A more seamless integration of T2S into current CAT tools would enable further studies to be conducted in more authentic environments, and more natural-sounding artificial voices would improve the user experience. Even at this stage, though, we see T2S as having perceived benefits for content comprehension and error identification, alongside measurable benefits for reducing error under-annotation.

Future work could thus include investigating the impact of T2S on error annotation and PEMT carried out by professional post-editors, with other text types and language pairs than the ones used in this study.

References

Allain, Jean-François. 2010. Repenser la révision. *Traduire. Revue française de la traduction*, (223):114–120.

Bowker, Lynne and Jairo B. Ciro. 2019. *Machine Translation and Global Research: Towards Improved Machine Translation Literacy in the Scholarly Community*. Emerald Publishing Limited.

Ciobanu, Dragoș. 2014. Of Dragons and Speech Recognition Wizards and Apprentices. *Tradumàtica: tecnologies de la traducció*, 12:524–538.

Ciobanu, Dragoș. 2016. Automatic Speech Recognition in the professional translation process. *Translation Spaces*, 5(1):124–144.

Ciobanu, Dragoș, Valentina Ragni, and Alina Secară. 2019. Speech Synthesis in the Translation Revision Process: Evidence from Error Analysis, Questionnaire, and Eye-Tracking. *Informatics*, 6(4)(51).

Ciobanu, Dragoș and Alina Secară. 2020. Speech recognition and synthesis technologies in the translation workflow. In Minako O'Hagan (ed.). *The Routledge Handbook of Translation and Technology*. Routledge, 91–106.

Defrancq, Bart and Claudio Fantinuoli. 2021. Automatic speech recognition in the booth: Assessment of system performance, interpreters' performances and interactions in the context of numbers. *Target-International Journal of Translation Studies*, 33(1):73–102.

Doherty, Stephen, Joss Moorkens, Federico Gaspari, and Sheila Castilho. 2018. On Education and Training in Translation Quality Assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (eds.). *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing, 95–106.

Dragsted, Barbara, Inger M. Mees, and Inge G. Hansen. 2011. Speaking your translation: students' first encounter with speech recognition technology. *The International Journal for Translation & Interpreting Research*, 3(1):10–43.

European Language Industry Association (ELIA), EMT, EUATC, FIT Europe, GALA, LIND, Women in Localization. 2022. *2022 European Language Industry Survey. Trends, expectations and concerns of the European language industry*. Available at: https://fit-europe-rc.org/wp-content/uploads/2022/03/ELIS-2022_survey_results_final_report.pdf?x77803 (Accessed: 23 March 2022).

European Master's in Translation. 2017. *EMT Competence Framework*. Available at: https://ec.europa.eu/info/sites/info/files/emt_competence_fwk_2017_en_web.pdf. (Accessed: 11 May 2022)

Fantinuoli, Claudio. 2017. Speech Recognition in the Interpreter Workstation. *Proceedings of Translating and the Computer 39*, 25–34.

Ginovart Cid, Clara and Carme Colominas Ventura. 2020. The MT Post-Editing Skill Set. Course descriptions and educators' thoughts. In Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera (eds.). *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. Routledge.

Guerberof Arenas, Ana. 2014. Correlations between productivity and quality when post-editing in a professional context. *Machine Translation*, 28:165–186.

Hassan, Hany, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Reqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving Human Parity on Automatic Chinese to English News Translation. *CoRR*, abs/1803.05567. Available at: http://arxiv.org/abs/1803.05567.

Herbig, Nico, Tim Düwel, Santanu Pal, Kalliopi Meladaki, Mahsa Monshizadeh, Antonio Krüger, and Josef van Genabith. 2020. MMPE: A Multi-Modal Interface for Post-Editing Machine Translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. *ACL 2020*, 1691–1702.

Kenny, Dorothy. 2020. Machine Translation. In Mona Baker and Gabriela Saldanha (eds.). *Routledge Encyclopedia of Translation Studies*. 3rd edn. Routledge, 305–310.

Läubli, Samuel, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4791–4796.

Liyanapathirana, Jeevanthi. 2021. Integrating post-editing with Dragon speech recognizer: a use case at international organizations. *43rd Translating and the Computer conference*. Available at: https://www.asling.org/tc43/videos/Liyanapathirana.mp4.

Lommel, Arle. 2018. Metrics for Translation Quality Assessment: A Case for Standardising Error Typologies. In Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (eds.). *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing, 109–127.

Mesa-Lao, Bartholomé. 2014. Speech-Enabled Computer-Aided Translation: A Satisfaction Survey with Post-Editor Trainees. *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, 99–103.

Moorkens, Joss and Sharon O'Brien. 2017. Assessing user interface needs of post-editors of machine translation. In Dorothy Kenny (ed.). *Human Issues in Translation Technology*. London: Routledge, 109–130.

Nitzke, Jean and Anne-Kathrin Gros. 2020. Preferential Changes in Revision and Post-Editing. In Maarit Koponen, Brian Mossop, Isabelle S. Robert and Giovanna Scocchera (eds.). *Translation Revision and Post-editing: Industry Practices and Cognitive Processes*. Routledge, 21–34.

Popović, Maja. 2018. Error Classification and Analysis for Machine Translation Quality Assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari and Stephen Doherty (eds.). *Translation Quality Assessment: From Principles to Practice*. Cham: Springer International Publishing, 129–158.

Sánchez-Gijón, Pilar, Joss Moorkens, and Andy Way. 2019. Post-editing neural machine translation versus translation memory segments. *Machine Translation*, (33):31–59.

Scocchera, Giovanna. 2017. Translation Revision as Rereading: Different Aspects of the Translator's and Reviser's Approach to the Revision Process. *Mémoires du livre / Studies in Book Culture*, 9(1).

Teixeira, Carlos S. C., Joss Moorkens, Daniel Turner, Joris Vreeke, and Andy Way. 2019. Creating a Multimodal Translation Tool and Testing Machine Translation Integration Using Touch and Voice. *Informatics*, 6(1)(13).

Vardaro, Jennifer, Moritz Schaeffer, and Silvia Hansen-Schirra. 2019. Translation Quality and Error Recognition in Professional Neural Machine Translation Post-Editing. *Informatics*, 6(3).

Wiesinger, Claudia, Justus Brockmann, Alina Secară, and Dragoș Ciobanu. Forthcoming. Speech-enabled machine translation post-editing in the context of translator training. *Peter Lang: Łódź Studies in Language*. The Łódz-ZHAW Duo Colloquium, 2-3 September, 2021.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144. Available at: http://arxiv.org/abs/1609.08144 (Accessed 11 May, 2022).

Yamada, Masaru. 2019. The impact of Google Neural Machine Translation on Post-editing by student translators. *JosTrans. The Journal of Specialised Translation* [Preprint], (31). Available at: https://www.jostrans.org/issue31/art_yamada.php (Accessed: 3 January 2021).

Zapata, Julian, Sheila Castilho, and Joss Moorkens. 2017. Translation Dictation vs. Post-editing with Cloud-based Voice Recognition: A Pilot Experiment. *Proceedings of MT Summit XVI. Vol.2 Commercial MT Users and Translators Track*, 173–186