The impact of machine translation on the translation quality of undergraduate translation students

Jia Zhang jia.zhang2@unsw.edu.au
School of Humanities and Languages
University of New South Wales, Sydney, 2052, Australia
Hong Qian hongqian@uic.edu.cn
Department of Languages and Cultures
BNU-HKBU United International College, Zhuhai, 519087, China

Abstract

The importance of machine translation (MT) and post-editing (PE), as well as the importance of MT and PE training, has been widely acknowledged, and specialised courses have recently been introduced at universities worldwide. However, MT courses are usually offered to students at the postgraduate level or in the last year of an undergraduate programme. In addition, existing empirical studies have mainly investigated the impact of MT on postgraduate students or undergraduate students in the last year of their studies. The present paper reports on a study that aimed to determine the possible effects of MT and PE on the translation quality of undergraduate students in the early stages of translator training. Methodologically, an experiment was conducted to compare the students' (n = 10) post-editing machine translation (PEMT)-based translations and from-scratch translations. Several methods of translation quality assessment were adopted, including rubric-based scoring and error analysis. It was found that the quality of students' PE translations was compromised in comparison to the quality of their from-scratch translations. In addition, errors were more homogenised in the PEMT-based translations. It is hoped that this study can shed light on the role of PEMT in translator training and contribute to the curricula and course designs of PE for translator education.

1. Machine translation and translator training

Following several decades of development, machine translation (MT) systems can now translate more accurately than ever before. However, due to the complexity of human languages, MT cannot yet truly or fully convey the meaning of a text in the target language; thus, post-editing machine translation (PEMT) has become necessary. Post-editing (PE) refers to the process of improving machine-generated translations. House (2017, p. 20) pointed out that, in the future, translators would 'have to devote considerably more time to pre- and post-editing of texts'. Therefore, PE should be an essential skill for all translators.

In recent years, PEMT has been introduced at universities with the aim of training would-be translators with this skill. Since translators are destined to become post-editors (Pym, 2013), training programmes for translators should be redesigned. Some universities offer specialised courses, while others incorporate PEMT as an essential module in courses on translation technology. For example, a PEMT course was introduced for students in the Localisation Master's programme at Universitat Autònoma de Barcelona in Spain in 2009 and in 2017 (Arenas & Moorkens, 2019), while the University of Helsinki in Finland provided a course on PE for

Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track, pages 99–108
September 4–8, 2023, Macau SAR, China.

undergraduate students and postgraduate students (Koponen, 2015); furthermore, the University of Exeter in Britain offers a course in machine-assisted translation at the final-year undergraduate level, including a PE workshop (Belam, 2003). Trainers appear to have reached a consensus that these specialised courses should be offered at the postgraduate level or towards the end of an undergraduate programme and that undergraduate students in the early stages of translator training should not be introduced to the knowledge or skills pertaining to MT and PE.

There have always been concerns about whether teachers should allow novice translation students to use and post-edit MT because novice translators do not have the confidence or experience to critically evaluate the output of a technology (Bowker, 2015) nor the linguistic competence to identify errors in machine-suggested translations in the early stages of translator training; thus, the quality of their translations may be affected. One may even suspect that their reliance on machine-suggested translations might affect the development of their translation competencies, such as critical thinking and creativity. In addition, some technologies, such as MT, are regarded as being more complex tools and are thus difficult for undergraduate students to master, which is why they are often integrated into translation curricula later in the programme. As a result, translation programmes may forbid undergraduate students from resorting to MT before a specialised course is offered.

However, it has been observed that undergraduate translation students may use MT as a reference in their translation assignments even without having received any appropriate training in PEMT.

Empirical evidence suggesting the negative impact of MT and PE on undergraduate translation students' translation performances in previous studies is insufficient. Empirical studies often recruit postgraduate or undergraduate students in the last year of their programmes (e.g., Jia et al., 2019; Wang et al., 2021; Zaretskaya et al., 2016). Less attention has been paid to the possible effects of PEMT on the quality of students' translations if they are introduced to PEMT at an early stage in their translator training. Even when attempts to compare undergraduate students' PE results and from-scratch translations are made, such comparisons are usually based on an overall quality assessment with a score being assigned to each translation product. There is a lack of detailed and closer examinations of the quality of students' translation products with or without MT assistance.

If MT and PE are proven to be beneficial for novice translation students to a certain extent, teachers might consider ways of integrating PEMT as a course component in translator training for students in an earlier stage at the undergraduate level. Nonetheless, novice translation students should be informed about the possible negative impacts of PEMT in order to interact with it more effectively.

2. Research questions

In light of the above discussion, the current research aimed to explore the impact of MT and PE on the quality of undergraduate translation students' translations in the early stages of their translator training by comparing their from-scratch translations and their PEMT-assisted translations. The research questions (RQs) for the study were as follows:

1) How do undergraduate translation students' from-scratch translations differ from their PEMT-based translations?

2) What are these students' perceptions of the use of PEMT in translation?

3) What are the pedagogical implications of the use of PEMT in translator training at the undergraduate level?

An experiment was designed to compare novice translation students' PEMT and from-scratch translations in an attempt to answer the first RQ. Students' perceptions of MT and PE were also solicited to understand the analysis of the experimental results. It is hoped that this

study can shed light on the role of PEMT in translator training and contribute to the curricula and course design of PE for undergraduate education.

3. Methodology

3.1. Experimental design

An experiment was conducted amongst novice translation students to compare PEMT-assisted translations and from-scratch translations. Before the implementation of the experiment, ethical approval was obtained from the ethics committee of the university. Once the project had been approved, the participants were recruited quickly. In the experiment, the participants

(1) were briefed about the tasks,

(2) signed the informed consent form,

(4) translated the first text from scratch,

(5) post-edited the machine-generated translation of the second text (the order of the two translation tasks was randomised), and

(6) completed a survey.

The participants were second-year undergraduates who were enrolled in a translation programme at a university based in Zhuhai, China, at the time of the experiment. Their educational backgrounds were comparable, as the students were native Chinese speakers who had been learning English for more than ten years from primary school onwards. At the time of the experiment, all the students have taken at least three fundamental translation courses in which they obtained grades higher than B+; according to the university's grading system, a B+ indicates good competence in the performance in a course. In addition, the students had not taken any other translation courses on or off campus prior to the experiment. They had little translation experience and no knowledge of PEMT.

The participants were asked to translate two texts of around 300 words each from English into Chinese. One of the texts was translated from scratch, while the other was translated by post-editing machine-generated output. The direction of English-to-Chinese translation was chosen in our study in consideration of the commonly accepted belief that translation into the native tongue is easier than translation into a non-native language.

Several methods were adopted to guarantee that the textual difficulty of the two texts was similar. Firstly, the sources of the two texts were controlled. The texts were selected from the Accreditation Test for Translators and Interpreters in China (CATTI). As the national qualification test for translators in China, the level III exam questions must be controlled to maintain consistent levels of difficulty over the years. Two texts with a similar topic were chosen from the level III exam of December 2017 and were adapted by the researchers. The texts, adapted from two news reports, involved few or no professional terms from a specific field. No professional knowledge was needed to understand or translate the texts.

Secondly, some linguistic features were referred to as markers of text difficulty. The type-token ratio, the number of sentences, the average sentence length, the number of different sentence types and the level of the words in each text were calculated using the corpus tools AntConc and AntWordProfiler. The texts were rewritten to ensure that the markers were comparable.

Thirdly, the number of problem triggers, which were annotated by three raters, was comparable in both texts: Problem triggers were defined as words, phrases or sentences in the source texts that might cause translation errors. According to the three raters, there were 15, 21 and 20 problem triggers in text 1 and 17, 21 and 21 in text 2. The author then further edited the texts based on the results of the annotations.

The texts were edited to control the overall difficulty at the lexical, syntactic, semantic and pragmatic levels. The researchers rewrote the texts, and a native speaker was invited to

assist with the editing and proofreading once all the preparatory work mentioned above was complete. The texts were comparable in terms of difficulty and were also clear and accurate.

The environment in which the experiment was conducted was a laboratory for translator training in which the students had attended classes for one semester. The lighting, room temperature and noise level were maintained at the same level. The participants could choose to sit in the same seat in which they sat in the class and adjust the height of their chairs and the positions of their computer monitors. They could use the computers in the room or bring their personal computers. All of the above requirements guaranteed that the students completed the translations in a safe, quiet and comfortable setting. Each participant went through all the steps individually in the laboratory at a time slot that they had chosen to avoid possible stress created by the presence of peer participants.

The participants were provided with the same hard-copy dictionary and had no access to an internet connection. As this research intended to explore the quality of the students' translations, the students' decisions in the translation process should be based on their knowledge of and thoughts about translation instead of drawing on online resources.

The participants were also briefed about the translation standards of achieving accuracy and fluency, which are two basic requirements for novice translation students. These requirements were essentially the same as those in the students' translation assignments in the previous year of learning to translate.

The students recorded their translation time on the document for each text and were told to submit their translations when they had decided that their translations had met the quality standards. The expected time for the translation of a 300-word text is approximately 45 to 60 minutes, but no specific time limit was set for the translation tasks.

Immediately after they had completed both translation tasks, the participants were instructed to complete a questionnaire survey inquiring about their attitudes to and perceptions of MT and PE. The survey included several open-ended questions. The students could answer the questions in either Chinese or English according to their preferences. The researcher, who is also an experienced translator with over 15 years of experience, translated the answers that were in Chinese for further analysis.

3.2. Data collection

The translation products, including the from-scratch translations and the PEMT-assisted translations, were first rated by three raters who assessed the translations and scored each translation based on the rubrics provided by the researchers. The three raters were experienced translator trainers based in China who had taught foundational translation courses for at least three years.

The rubrics that were used were divided into translation accuracy and language quality. Therefore, each translation product was given a total score, an accuracy sub-score and a fluency sub-score. Inter-rater agreement was tested before the scores for each translation were finalised by averaging the scores given by the three raters. Paired-sample *t*-tests were conducted to explore the relationships between the scores for the from-scratch translations and those for the PEMT-assisted translations.

A paired-sample *t*-test was also conducted to reveal the relationship between the time spent translating from scratch and the time for the PEMT to reveal the students' translation efficiency in both tasks.

The error analysis of the translation products was implemented in two ways. Firstly, with reference to the TAUS Harmonised DQF-MQM Error Typology, errors in the translations were first annotated independently by the two researchers; the results revealed that both researchers agreed about most of the errors. When a discrepancy occurred, the two researchers engaged in discussions to reach an agreement. The errors were divided into four types: The count for each type of error and the total error count in each translation product were then calculated. The

paired-sample *t*-tests revealed the relationship between error counts in the from-scratch translations and those in the PEMT-assisted translations. Secondly, the errors in the translation products were observed and examined further in a qualitative manner to identify other issues related to translations with or without MT.

The survey of the students' attitudes to and perceptions of MT and PE was analysed manually with the assistance of NVivo to identify significant arguments, which would help to understand the results of the experiment in more depth.

4. Analysis and discussion

4.1. Rubric-based scoring

After the three raters had completed the grading, each translation product was given an overall score, which was divided into an accuracy sub-score and a fluency sub-score. Inter-rater agreement was verified by calculating the intraclass correlation coefficient (ICC). ICC estimates and 95% confidence intervals were calculated using the SPSS statistical package version 27 based on a mean-rating (k = 3), absolute-agreement, 2-way mixed-effects model. The value was .69. The scores were thus deemed to be reliable.

The score for each translation was then obtained by averaging the three scores given by the three raters. Paired-sample *t*-tests were conducted to understand the relationship between the scores for the from-scratch translations and those for the PEMT-assisted translations. Table 1 presents the results of the *t*-tests, and Figure 1 shows the corresponding boxplots. The from-scratch translations are indicated by "H", while the PEMT-assisted translations are marked as "M".

	MEAN	STD. DEVIATION	STD. ERROR MEAN	95% CONFIDENCE INTERVAL OF THE DIFFERENCE		T	DF	SIG. (2-TAILED)
				Lower	Upper			
SCORE H – SCORE M	5.93	6.23	1.97	1.48	10.39	3.01	9.00	0.01
ACCURACY H – ACCURACY M	2.97	2.82	0.89	0.95	4.98	3.33	9.00	0.01
FLUENCY H – FLUENCY M	2.97	3.77	1.19	0.27	5.66	2.49	9.00	0.03

Table 1 *T*-test results for the rubric-based scores

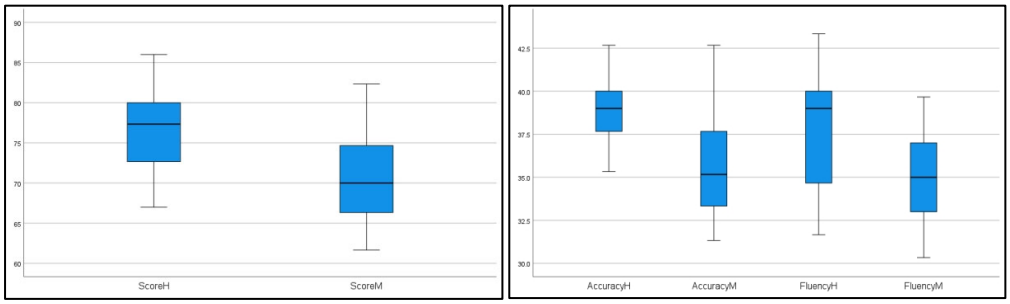

Figure 1 Boxplots of the rubric-based scores

The scores for the from-scratch translations (M = 76.7, SD = 5.34) were significantly higher than those for the PEMT-assisted translations (M = 70.77, SD = 5.88), with $t(9) = 3.01$, $p = .01$.

Similarly, the accuracy scores (M = 38.8, SD = 1.98) for the from-scratch translations were significantly higher than those (M = 35.83, SD = 3.19) for the PEMT-assisted translations, with $t(9) = 3.33$, $p = .01$.

The fluency scores (M = 37.9, SD = 3.47) for the from-scratch translations were significantly higher than those (M = 34.93, SD = 2.77) for the PEMT-assisted translations, with $t(9) = 2.49$, $p = .03$.

This result shows that the students' performances in the PEMT-assisted translations were not as good as they were in the from-scratch translations. The quality of the PE results decreased.

4.2. Translation time

The students were instructed to record the start time and the end time for the two tasks. The researchers then calculated the translation time needed for each task. Again, a paired-sample *t*-test revealed the relationship between the time spent translating from scratch and the PE time. In Table 2 and Figure 2, TIME H refers to the time needed for from-scratch translations, while TIME M indicates the time required for PE.

				95% CONFIDENCE INTERVAL OF THE DIFFERENCE				
	MEAN	STD. DEVIATION	STD. ERROR MEAN	Lower	Upper	T	DF	SIG. (2-TAILED)
TIME H – TIME M	7.20	18.17	5.75	-5.80	20.20	1.25	9.00	0.24

Table 2 *T*-test result for translation time

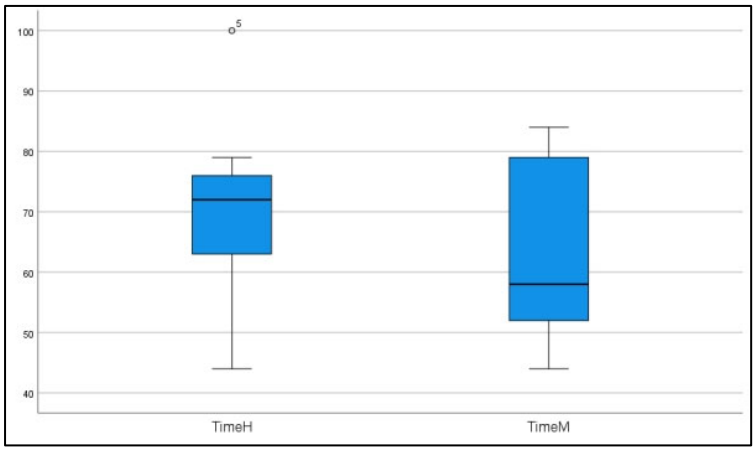

Figure 2 Boxplot of the translation time

The translation time for the from-scratch translations (M = 70.50, SD = 14.79) was not significantly different from that for the PEMT-assisted translations (M = 63.30, SD = 14.59), with $t(9) = 1.25$, $p = .25$. The PEMT output did not increase the translation efficiency of the participants in general.

When each student's translation time was examined closely, it was found that only half of the participants reported a subtle decrease in the time needed for the PE task, while three other participants spent the same amount of time on both tasks. It is worth noting that two participants spent significantly more time on the PE task.

4.3. Error counts

TAUS Harmonised DQF-MQM Error Typology is an internationally recognised framework for the assessment of translation quality. It is not only used to assess automated translations but also to assess post-edited machine translation and human translations. As the texts chosen for

the translations only involved certain types of translation errors, the typology was adapted to suit the purposes of the annotations, as displayed in Table 3.

ID	Error type	Definition
1	Accuracy	The target text does not accurately reflect the source text, allowing for any differences authorised by the specifications.
2	Fluency	Issues related to the form or content of a text, irrespective of whether it is a translation or not.
4	Style	The text has stylistic problems.
7	Verity	The text makes statements that contradict the world of the text.

Table 3 Adapted TAUS Harmonised DQF-MQM Error Typology

Both researchers identified the errors in the students' translations by referring to the error types. The annotation results were compared, and the two researchers engaged in discussions when they had different opinions about some translation errors. The errors in each translation product were thus confirmed. The errors in each type were counted, and a total error count was calculated. Paired-sample *t*-tests revealed the relationships amongst the error counts, as shown in Table 4 and Figure 3.

				95% CONFIDENCE INTERVAL OF THE DIFFERENCE				
	MEAN	STD. DEVIATION	STD. ERROR MEAN	Lower	Upper	T	DF	SIG. (2-TAILED)
ERROR H - ERROR M	-2.50	4.17	1.32	-5.48	0.48	-1.90	9.00	0.09
ERROR TYPE I H - ERROR TYPE I M	-3.50	4.09	1.29	-6.43	-0.57	-2.71	9.00	0.02
ERROR TYPE II H - ERROR TYPE II M	0.20	0.79	0.25	-0.36	0.76	0.80	9.00	0.44
ERROR TYPE III H - ERROR TYPE III M	0.50	3.63	1.15	-2.10	3.10	0.44	9.00	0.67
ERROR TYPE IV H - ERROR TYPE IV M	0.30	0.48	0.15	-0.05	0.65	1.96	9.00	0.08

Table 4 *T*-test results for the error counts

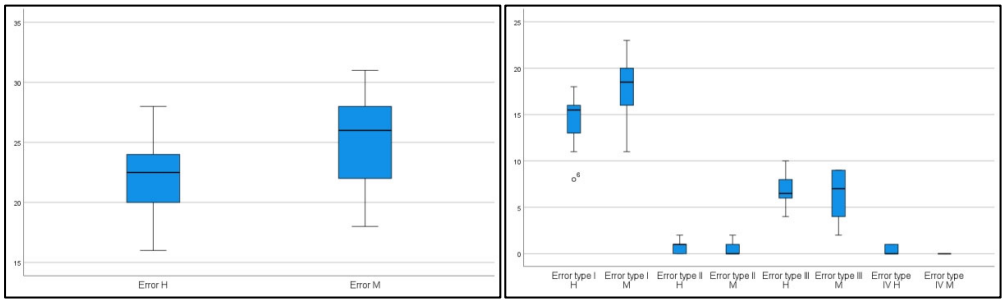

Figure 3 Boxplots of the error counts

The total error count in the from-scratch translations (M = 22.30, SD = 3.368) was lower than it was in the PEMT-assisted translations (M = 24.80, SD = 4.341), but not significantly so, with $t(9) = -1.90$, $p = .09$. Of the four types of errors, only type I accuracy errors in the from-scratch translations (M = 14.4, SD = 3.169) were significantly lower than they were in the PEMT-assisted translations (M = 17.90, SD = 3.725), with $t(9) = -2.71$, $p = .02$.

The number of errors might have increased when the students were engaged in PE, but a definite conclusion based on our data could not be drawn. This result may have been affected by the small number of participants who were recruited and the small number of error counts for each error type given that the length of each text was only 300 words. Further studies must be conducted to determine the impact of MT and PE on students' error counts.

4.4. Error observation

The errors in the students' translation products were observed closely. It could clearly be seen that, in the from-scratch translations, the errors triggered by the same point were very different. However, when the students engaged in PEMT, the errors tended to be homogenised. The error types and words used to translate a certain point were exactly the same.

An example is the phrase "native English speaker" in the text translated from scratch, which was translated using different renderings. Some students made errors when translating this phrase, but the error types differed. These students intended to be more creative in their translations but still made various errors, as shown in Table 5.

Source text	Target text	Error
native English speaker	土生土长的英语母语者[back translation: English native speaker born and raised in an English-speaking country]	Over-translation
	那些英语母语者（英语可能并不是他们所在国的唯一语言）[back translation: English native speaker (English might not be the only language in their country)]	Addition
	以英语为主要语言的人[back translation: People who use English as the main language]	Mistranslation
	天生就讲英语的人[back translation: people who speak English after they were born]	Unidiomatic

Table 5 Examples of errors in the from-scratch translations

The translation errors in the PEMT-based translations were identical. The two phrases in Table 6 were translated as identical Chinese versions in the PE task by most of the participants. Seven out of 10 students mistranslated "English speakers with no other language" as "英语使用者" (English users). Similarly, seven students translated the phrase "simple but standard grammar" word for word, which resulted in awkwardness in the target text. It can be inferred that the students could not improve on the unidiomatic or awkward expressions provided by the MT. The students may not have been able to identify all the errors in the machine-generated output, and their critical thinking was also impacted.

Source text	Target text	Error	Frequency
English speakers with no other language	英语使用者[back translation: English users]	Mistranslation	7/10
Globish -- a distilled form of English, stripped down to 1,500 words and simple but standard grammar	1500 个单词和简单但标准的语法[1500 words and simple but standard grammar]	Awkward	7/10

Table 6 Examples of errors in PEMT-based translations

4.5. Students' perceptions

Contrary to the experimental results, eight out of 10 participants stated in the survey that they felt more confident when post-editing the MT output. As a result, they also felt more confident about the quality of their PEMT-assisted translations. Even though some of the students doubted the quality of the MT, they did trust MT to a certain extent, as they clearly expressed that MT helped them to understand the source text better, particularly with regard to the text structure and complicated sentences. In addition, they believed that the MTs provided good references that decreased their efforts to find the correct words.

Of note, all ten students also said that they invested more effort in the PEMT tasks and made more judgement about the quality of and adjustments to the MT output. This echoed the analysis of the time spent on the translation tasks to some extent. According to the survey, such efforts were mainly aimed at improving awkwardness in the machine-generated translations.

One point worth noting is that none of the students mentioned that such an increase in effort was the result of their insufficient translation competence or language proficiency. Instead, they believed that the main reasons were their unfamiliarity with MT and their lack of PE training.

5. Concluding remarks

This research required students to perform from-scratch translations and PEMT-assisted translations and compared the quality of the products with the aim of exploring the impact of MT and PE on the translation performances of undergraduate students in the early stages of translator training.

The quality of the students' PEMT-assisted translations was compromised in comparison to that of their from-scratch translations. The overall score, the accuracy sub-score and the fluency sub-score for the PEMT-assisted translations were significantly lower than those for the from-scratch translations. The total error counts and accuracy error counts in the PEMT tasks were higher than those in the from-scratch translation tasks.

The students' perceptions of translation quality were the opposite. The students felt more confident when having a pre-translated version to hand and thus had more confidence in the outcomes of the translations.

The students' translation efficiency was not improved via MT assistance. The time spent on PE was reduced, but not significantly from that spent on the from-scratch translations. Two students obviously spent more time on the PEMT-assisted translation. This result is consistent with the students' perceptions. Most students expressed feeling annoyed and burdened because correcting "weird" expressions took them more time. The students attributed the increase in effort to a lack of MT knowledge and PE training rather than to their translation competence or language proficiency.

The students' translation errors in the PEMT tasks were homogenous. A closer examination of their translation products revealed that they could not identify an error made by a machine and tended to retain these errors in their final translation products.

It is thus probably concluded that MT may not benefit undergraduate students in the early stages of translator training in the absence of specialised MT and PE training. Without any training, MT impacted negatively on their translation quality and possibly on their critical thinking. If students rely too extensively on MT too early in their translator training, one might be concerned that MT might have a negative impact on the development of their translation competence. However, since the students raised the issue of training, the next step could be to test the effectiveness of MT and PE training on the translation performances of undergraduate translation students.

Even if PE training is not to be incorporated at an early stage in translator training, it is strongly suggested that trainers and teachers should provide lectures about basic MT knowledge. The students obviously trusted MT to a certain degree, particularly with regard to understanding the source text. Although they were not engaged in PEMT directly, they were willing to use MT as a helpful reference in their translations. As many MT systems are freely available online, preventing students from accessing them would be difficult. Therefore, specific guidance should be provided to students to make them aware of the best practices when interacting with MT at this stage. For example, students should be aware that, although technology is overwhelming the industry and education sectors, the fundamental factor for translation learners, language proficiency, must still be prioritised.

As this was a small-scale experiment with only ten participants and two texts of 300 words each, we understand that there is a limitation in terms of generalising the results of the experiment. Our findings prove that this topic requires further exploration. Experiments involving more participants and longer texts could be conducted to provide more empirical evidence in this regard.

Acknowledgement

This work is supported by the FHSS Teaching and Learning Grant (FHSS TLSE Committee) of BNU-HKBU United International College.

References

Arenas, A. G., & Moorkens, J. (2019). Machine translation and post-editing training as part of a master's programme. *Journal of Specialised Translation*, 31, 217–238.

Belam, J. (2003). Buying up to falling down: A deductive approach to teaching post-editing. *Proceedings of MT Summit IX Workshop on Teaching Translation Technologies and Tools*, 1–10.

Bowker, L. (2015). Computer-aided translation: Translator training. In S. Chan (Ed.), *Routledge encyclopedia of translation technology* (pp. 126–142). Routledge.

House, J. (2017). *Translation: The basics*. Routledge.

Jia, Y. F., Carl, M., & Wang, X. L. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *Journal of Specialised Translation*, 31, 60–86.

Koponen, M. (2015). How to teach machine translation post-editing? Experiences from a post-editing course. *Proceedings of the 4th Workshop on Post-editing Technology and Practice*.

Pym, A. (2013). Translation Skill-Sets in a Machine-Translation Age. *Meta*, 58(3), 487–503. https://doi.org/10.7202/1025047ar

Wang, X., Wang, T., Muñoz Martín, R., & Jia, Y. (2021). Investigating usability in postediting neural machine translation: Evidence from translation trainees' self-perception and performance. *Across Languages and Cultures*, 22(1), 100–123. https://doi.org/10.1556/084.2021.00006

Zaretskaya, A., Vela, M., Pastor, G. C., & Seghiri, M. (2016). Measuring post-editing time and effort for different types of machine translation errors. *New Voices in Translation Studies*, 15, 63–92.