Proficiency and External Aides: Impact of Translation Brief and Search Conditions on Post-editing Quality

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

Masaru Yamada masaru.yamada@rikkyo.ac.jp
Takanori Mizowaki 22wv006d@rikkyo.ac.jp
Rikkyo University, Tokyo, Japan

Abstract

This study investigates the impact of translation briefs and search conditions on post-editing (PE) quality produced by participants with different levels of translation proficiency. We hired five Chinese student translators and seven Japanese professional translators to conduct full post-editing (FPE) and light post-editing (LPE), as described in the translation brief, while controlling two search conditions i.e., usage of a termbase (TB) and internet search (IS). Our results show that FPE versions of the final translations tend to have less errors than LPE versions. The FPE translation brief improves participants' performance on fluency as compared to LPE, whereas the search condition of TB helps to improve participants' performance on accuracy as compared to IS. Our findings also indicate that the occurrences of fluency errors produced by experienced translators (i.e., the Japanese participants) are more in line with the specifications addressed in translation briefs, whereas the occurrences of accuracy errors produced by inexperienced translators (i.e., our Chinese participants) depend more on the search conditions.

1 Introduction

Post-editing (PE) has become widely used in industrial translation. In some domains, more than 40% of translation practices are conducted as PE (JTF, 2020). However, PE does not mean that translators simply use machine translation (MT) to translate. In practice, MT systems must be integrated into an authentic environment, such as CAT tools, with which professional translators can perform PE operations including searches for terminology, concordance, usage of external resources on websites, etc. In addition, PE must often satisfy given quality requirements — as to whether it is full PE (FPE) or light PE (LPE) — which are normally described in the work instruction a translation brief. Thus, differences in search conditions — internet search (IS) or availability of termbase (TB) — and translation brief (i.e., LPE/FPE) may affect the translator's psychology and working style, which in turn may impact cognitive load during the translation process (effort) and the translation product (quality).

While the differences in the work environment and task conditions affect the process and product, few previous studies have taken this into account. There are possibly two reasons for it. The first reason is that an authentic translation environment capable of collecting translation process data was not previously available. For example, Translog-II, which is an experimental tool for researchers to collect translators' keyboard input and gaze data, does not provide the functionality that professional translators are used to in conventional CAT tools(Carl, 2012). The other reason has to do with translation conditions. The establishment of the international PE guidelines, ISO 18587 in 2017 has led to a certain common understanding with respect to what has to be post-edited. However, the industry definition of PE has not always been used which makes it difficult to compare the results of PE studies across the research. For example, it is often unclear whether a study focused on LPE or FPE.

Given this background, the purpose of this study is to have translators translate in an authentic translation work environment, collect translation process data, and compare and verify differences of translation performance. The term "authentic environment", in our definition, means 1) ensuring that PE is conducted in a professional CAT environment (Trados Studio in our case), as well as that translators are allowed to use IS or TB, and 2) providing appropriate work instructions that specify whether the task is FPE or LPE. The variable 1) is referred to as "search condition", and 2) is considered to be the "translation brief". In this way, this study examines differences in product-process interactions, considering these conditions (as variables) that will affect PE processes and products.

2 Literature review

There exist many comparative analyses of translation products and processes under different task conditions (from-scratch translation, PE) and work environments. However, to the best of our knowledge, no previous literature has directly examined the impact of differences in translation briefs (e.g., LPE vs. FPE) on PE quality.

A translation brief specifies the intended audience and purpose of the translation in the target language. Translation briefs are meant to bias (or prime) translators, to activate particular, but not other, "bodies of thought" (Gutt, 2004, p. 13) that answer to a specific translation expectations. Pym (2003, p. 486) points out that the notion of translation brief is "a key point in German-language Skopos theorie since 1984". For Nord (2006, p. 142) translation depends on the "conclusions the translator draws from the brief [. . .] it is no longer the source-text [alone] that guides the translator's decisions but the overall communicative purpose the target text is supposed to achieve in the target culture." Also Sturm (2017, p. 16) mentions that "Translation briefs and technical guidelines offer indications both about author intentions and the background of the target audience", and by now several translation companies offer guidelines that explain how to draft translation briefs [1].

Melby et al. (2012, p. 7) defines translation from a process-oriented perspective as follows: "Translation is the process of creating target language content that corresponds to the source content according to agreed-upon specifications". Melby's concern is the relationship of translation to the "specifications" or the required functions of the translation in the given context. This idea applies to PE tasks. Melby et al. (2014) argue that error-category-based specifications should be used to define quality in MTPE projects. The specification that Melby et al. promote is MQM (Multidimensional Quality Metrics), developed in the EU-funded QTLaunchPad project (Lommel et al., 2014).

Although the definition of FPE and LPE has been clarified with the development of ISO 18587, the international standard for PE (ISO, 2017), this definition is ambiguous for use in practice. Nunziatini and Marg (2020) provide clearer instructions and methods for post-editors, based on their own industry experience, by correlating instructions of FPE and LPE with MQM error typologies. In the same vein, (Sakamoto and Yamada, 2022) propose a risk management

[1] See https://toppandigital.com/us/blog-us/write-effective-translation-brief/, https://harryclarktranslation.co.nz/successful-translation-brief-made/.

method for which each translation task must include the client's requirements which should be broken down into issue typologies that are applied during PE with consideration of their severities.

Nitzke et al. (2019) have described this pre-production process from this risk management perspective. They claim that decision-making processes which include - among other things - the understanding of translation brief should be taught during the translation education. It is important, they say, to consider the constraints, conditions and expected translation quality for each task. This has been compiled as PE guidelines (Hu and Cadwell, 2016; Massardo et al., 2017) which stress that the translation brief has a significant impact on the translation process and product as well.

The following literature deals with interactions between different task conditions or "parameters" and differences in their performance.

Daems and Macken (2020) carried out an experiment with two groups of participants. One group were revisers who usually check/edit human translations. The other group consisted of post-editors who are used to correct MT output. In this experiment, the raw MT output and the human translations were given to both the revisers and the post-editors, but participants were not informed of the type of the texts given to them. The study compared the quality of the translations after each group's edit. The result shows that, the revisers outperformed the post-editors when they edited MT output. However, the post-editors outperformed the revisers when they edited human translations. While in every case accuracy errors remain custom underedittedcustom, this outcome suggests that different conditions influence the performance of the revisers and post-editors.

In connection with this, "search conditions" are also important. For example, is the CAT tool available for the PE, is a glossary provided, and/or is plenty of time and pay given for external searches? Whether or not sufficient working conditions are prepared is a prerequisite for fulfilling the expected requirements.

Search conditions for external resources such as IS relevant to the subject matter in the course of translation are vital to the translator, and this will also affect translators' performance. Onishi and Yamada (2020) compared search behaviors of professional and novice translators. They found that professional translators devote a higher percentage of time and operations into searches. They found a high correlation between accuracy errors and the frequency and depth of searches. These findings suggest that IS during translation will greatly affect translators' quality.

3 Experimental design

In our study, we investigate how the two controlled parameters, i.e., translation brief and external search, interact with the participants' different levels of translation proficiency.

3.1 Participants

The PE experiment was conducted with two groups of participants using the same English source texts (STs) and Google neural machine translation (GNMT) outputs. One group comprises five Chinese translation students with simplified Chinese as their L1 and English as their L2. Seven Japanese professional translators with L1 in Japanese and L2 in English make up the other group. Participants were requested to fill out a questionnaire about their basic information before attending the experiment, which included language use on a daily basis, language learning experience, language proficiency, translation experience, PE experience, etc. Chinese participants had an average of 2.4 years of professional translation experience, whereas Japanese participants had an average of 7 years, as shown in the following Table 1. The twelve participants of both groups attended the experiment individually, using the same CAT tool, Trados

Studio, version 2019, without a time limit.

Years of experience	Minimum	Maximum	Mean	Standard deviation	Variance
Participant-ZH	0	6	2.4	2.24	5.04
Participant-JA	1	17	7.0	5.24	27.43

Table 1: Translation experience of participants

3.2 Materials

We selected four English source texts (STs) with general topics from previous American Translators Association (ATA) certification examinations. These exams were intended and considered to be a general professional-level assessment for translators (Koby and Champe, 2013). Among them, two texts were for English-to-Chinese ATA exams and two texts were for English-to-Japanese ATA exams. Each text is around 250 words long and contains about 10 segments. Their readability scores (Flesch-Kincaid Grade Level) are relatively similar, as shown in Table 2.

Text	Topic	Word count	Segment count	Readability score
1	Welfare	260	11	15.5
2	Tourism	242	12	12.3
3	War	263	11	13.9
4	Racism	257	13	15.2

Table 2: General descriptions of the four STs

We used GNMT to translate the four STs into simplified Chinese and Japanese and used this material to prepare four TMs for the Chinese participants and four TMs for the Japanese participants. We chose 28 words or phrases in the STs that had terminology errors in the GNMT outputs for any of the two language pairs (i.e., English-Chinese or English-Japanese) and generated a TB with the same set of English source terms and their equivalent Chinese and Japanese target terms.

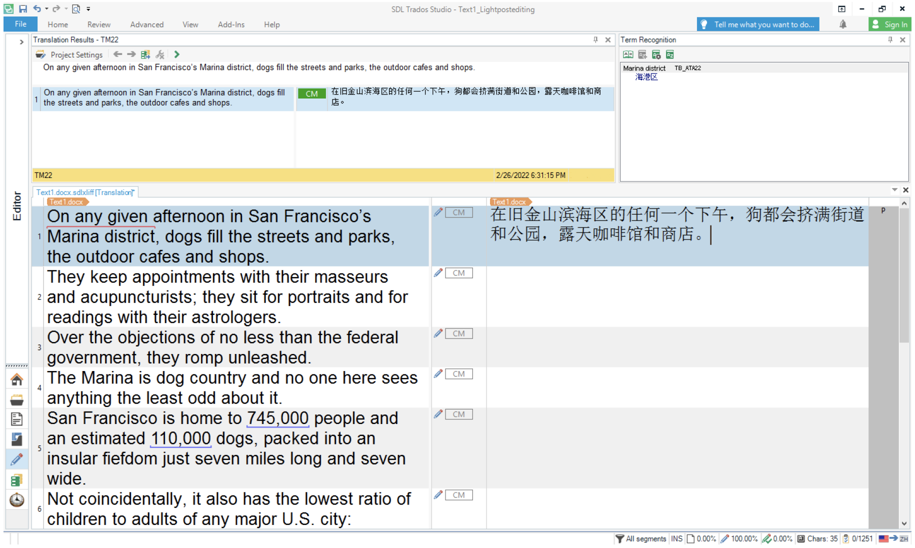
3.3 Experimental layout

We provided two kinds of translation briefs to the participants: LPE (l) and FPE (f). 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 (t), and access to any IS but no TB provided within Trados interface (s). Therefore, each participant conducted the PE of the four texts under four orthogonal tasks respectively, as illustrated in Table 3.

Brief/Condition	TB	IS
LPE	Plt	Pls
FPE	Pft	Pfs

Table 3: Four experimental tasks for each participant

For all the experimental tasks, the participants were presented with the GNMT outputs segment by segment appearing on the target text (TT) section as well as the TM section of the Trados interface in the same way as 100% matches with the TM. Under this experimental setup, we controlled that the participants of the same language pair had access to the same sets of GNMTs at a segment level. The working interface for the participants in the experiment is shown in Figure 1.



Translation Results - TM22

On any given afternoon in San Francisco's Marina district, dogs fill the streets and parks, the outdoor cafes and shops.

1	On any given afternoon in San Francisco's Marina district, dogs fill the streets and parks, the outdoor cafes and shops.	CM	在旧金山滨海区的任何一个下午，狗都会挤满街道和公园，露天咖啡馆和商店。

Term Recognition

Marina district	TB_ATA22
海港区	

1	On any given afternoon in San Francisco's Marina district, dogs fill the streets and parks, the outdoor cafes and shops.	CM	在旧金山滨海区的任何一个下午，狗都会挤满街道和公园，露天咖啡馆和商店。
2	They keep appointments with their masseurs and acupuncturists; they sit for portraits and for readings with their astrologers.	CM	
3	Over the objections of no less than the federal government, they romp unleashed.	CM	
4	The Marina is dog country and no one here sees anything the least odd about it.	CM	
5	San Francisco is home to 745,000 people and an estimated 110,000 dogs, packed into an insular fiefdom just seven miles long and seven wide.	CM	
6	Not coincidentally, it also has the lowest ratio of children to adults of any major U.S. city:	CM	

Figure 1: Working interface in Trados for the participants

Before their PE session started, each participant was given the translation briefs for the four texts with the descriptions of FPE and LPE defined by ISO 18587 (ISO, 2017), as shown in Appendix A. To ensure that the sequence of the tasks will not have an impact on our evaluation of the experiments, we randomized the experiment layout for each participant by permuting the four texts and keeping the succession of the four PE tasks in the same order.

For the TB condition, the participants were presented with the terminology appearing on the TB section of the Trados interface. Once there was one or several terms in an ST segment that they were working on match the terms in the TB that we prepared, the participants could check the TB on the upper right corner of the Trados workbench.

For the IS condition, the participants were presented with an empty TB appearing on the TB section of the Trados interface in the same way as 0% matches with TB. In this way, we assume that the participants were working in a near-authentic and familiar working environment of translation. The experimental layout is shown in Appendix B.

3.4 Data collection

The keystroke data during the PE sessions were recorded by both the Qualitivity plugin for Trados and eye tracker software (i.e., Tobii Studio 3.3.2 and Web Link). The translator's eye movement data were collected with the Tobii TX 300 eye tracker and the Eyelink 1000 plus for the EN-ZH and EN-JA experiments respectively. The translation process data (keystroke and gaze data with their production times) was then converted and processed by the newly launched research tool, Trados-Translog interface available at CRITT TPR-DB (Zou and Carl, 2022; Yamada et al., 2022). We found that the new tool can successfully be utilized to synchronize keystroke and gaze data from text production sessions into various data tables at different levels of granularity, including the text (SS), the segment (SG), the alignment group (AG), and translation unit (TU).

4 Quality assessment

Two professional translators (one Chinese and one Japanese) were hired to annotate the translation errors in the simplified Chinese and Japanese GNMT outputs as well as in the 12 PE versions. Because the STs of ATA exams are specifically designed to incorporate challenges that may result in translation errors associated with the categories and severity of errors under the grading framework of ATA, annotators in this experiment were given guidelines for error annotation based on an ATA-adapted annotation schema. [2]

Errors were divided into six types, “Mistranslation”, “Usage”, “Terminology”, “Grammar”, “Omission/Addition” and “Other”. The former four types were further annotated as “Critical” and “Minor” errors depending on the severity of errors. As a result, there were altogether ten different kinds of errors, i.e., Mistranslation_Critical, Mistranslation_Minor, Usage_Critical, Usage_Minor, Terminology_Critical, Terminology_Minor, Grammar_Critical, Grammar_Minor, Omission/Addition, and Other.

In this experiment, the annotation was conducted on the level of AG. Annotators were asked to proceed in two steps: first, they should conduct word-level alignment between the TT and their corresponding ST. Then in the second step, AGs were assigned an error as applicable. When they came across an error that they considered an omission or addition, however, they were not required to do an alignment. In other words, there are only AGs for errors excluding “Omission/Addition” in this research.

For the purpose of this study, the occurrences of “Mistranslation”, “Terminology”, and “Omission/Addition” errors were grouped under the label of “Accuracy” error, while “Usage”, “Grammar”, and “Other” errors were grouped under the label of “Fluency” error. Additionally, all kinds of translation errors were grouped under the label of “Critical” and “Minor” Errors according to their annotated severity. Therefore, we gained four subcategories of errors under study, such as ”Accuracy_Critical”, ”Accuracy_Minor”, ”Fluency_Critical”, and ”Fluency_Minor”.[3]

5 Results

5.1 Error distribution

The twelve participants produced altogether 658 segments from the output of the two GNMT systems (i.e., simplified Chinese and Japanese). Because Omission errors only occur on the ST side, and Addition errors only occur on the TT side, we count both source and target words in an AG that involve each of the aforementioned four subcategoriess of errors (i.e., ”Accuracy_Critical”, ”Accuracy_Minor”, ”Fluency_Critical”, and ”Fluency_Minor”). Since the STs for all the post-editors are identical, we can examine the total error counts for each of the four error subcategories in the raw GNMT output and the PEMT versions following the manual annotation by the two translators.

As illustrated in Figure 2, the error distribution of the GNMT has similar pattern as that of the PEMT. That is, the most frequent errors for GNMT and PEMT were, respectively, Fluency_Minor errors (50.59% and 46.86%), followed by Accuracy_Critical errors (22.93% and 30.92%), Accuracy_Minor (18.55% and 17.57%), and Fluency_Critical errors (7.93% and 4.65%). As these figures show, PEMT has lower percentages of fluency and critical accuracy errors than the GNMT. The results also show that fluency errors are usually minor errors, while accuracy errors are more often considered critical(Carl and Báez, 2019; Zou et al., 2021).

We also compare the raw total error counts for each experimental task, i.e., Pfs, Pft, Pls,

[2] See https://www.atanet.org/certification/how-the-exam-is-graded/error-categories/.

[3] In this research, “Omission/Addition” error were grouped under the label of “Critical” error, whereas “Other” error were grouped under the label of “Minor” Error.

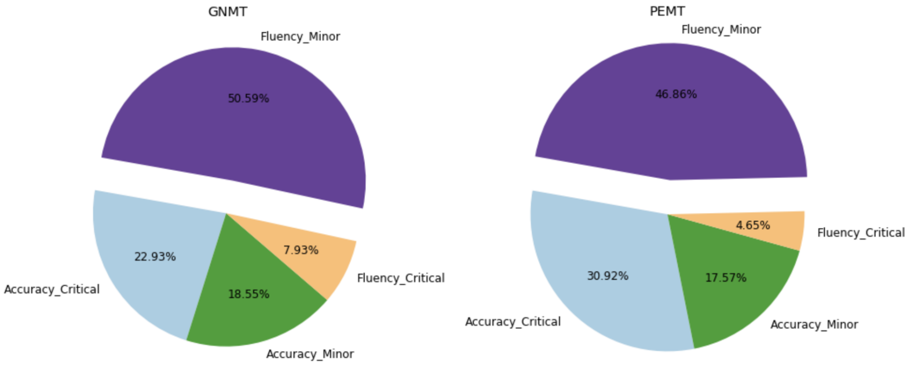


Figure 2: Error Distribution of GNMT and PEMT

Plt. The results in Figure 3 show that FPE versions have less total errors than LPE versions. Considering jointly all Japanese and Chinese versions, Pfs leads to less Fluency_Critical errors while Plt produces the most Fluency_Critical errors. Pft versions tend to have the least Fluency_Minor errors while Plt tend to have the most Fluency_Minor errors. Furthermore, Pft and Plt versions tend to have less Accuracy_Critical errors than the other two versions. Overall, compared to LPE, the translation brief of FPE improved the participants' performance on fluency, and - compared to IS - the provided TB improved the participants' performance on accuracy in our total data-set of experienced and less-experienced post-editors.

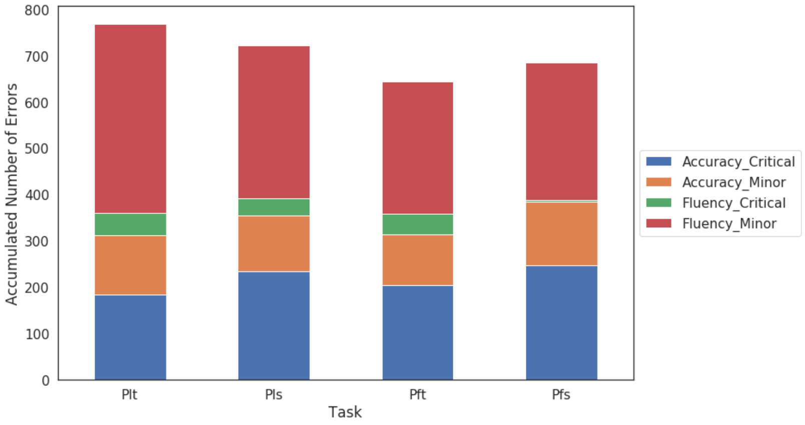


Figure 3: Error distribution across tasks

5.2 PE tasks

However, the focus of this study is to compare the effect of translation briefs (FPE or LPE) and search conditions (TB or IS) on PE quality between groups of participants with varying levels of translation proficiency. We use mixed two-way ANOVA with four dependent variables which correspond to the four error labels discussed above (i.e., Accuracy, Fluency, Critical, and Minor). The within-group independent variable consists of the four PE tasks (i.e., Plt, Pls, Pft and Pfs), and the between-group independent variable is the participant group, Chinese (zh) or Japanese (ja). Our findings indicate that across the four PE tasks, the PE versions of Chinese (novice) participants show significantly more Accuracy and Critical errors than Japanese (expert) participants. These results may be expected, as they confirm that more experienced (Japanese) translators consistently provide higher quality translations than less experienced (Chinese) ones (Shreve, 2006). While the Chinese novices produce in general more Accuracy and Critical errors, there is no significant interaction between the PE tasks and participant groups. In other words, the PE versions of Japanese participants have to the same extent less accuracy and critical errors than Chinese participants regardless of the PE tasks.

Across the four PE tasks, we also identify different tendencies of Accuracy and Fluency errors between the Chinese and Japanese groups of participants. As shown in Figure 4, there is no significant difference in Accuracy errors across all the four tasks for the Japanese participants. However, we observe that Accuracy errors fluctuate throughout the four tasks for the Chinese participants. While Chinese participants with the LPE translation brief (Plt and Pls) do not exhibit a discernible difference in Accuracy errors as compared to the FPE brief (Pft and Pfs), they tend to produce fewer Accuracy errors within the TB condition (Plt and Pft) as compared to the IS condition (Pls and Pfs). That is, they seem to be able to make better use of the TB than with free search (IS), but are largely indifferent to the translation brief.

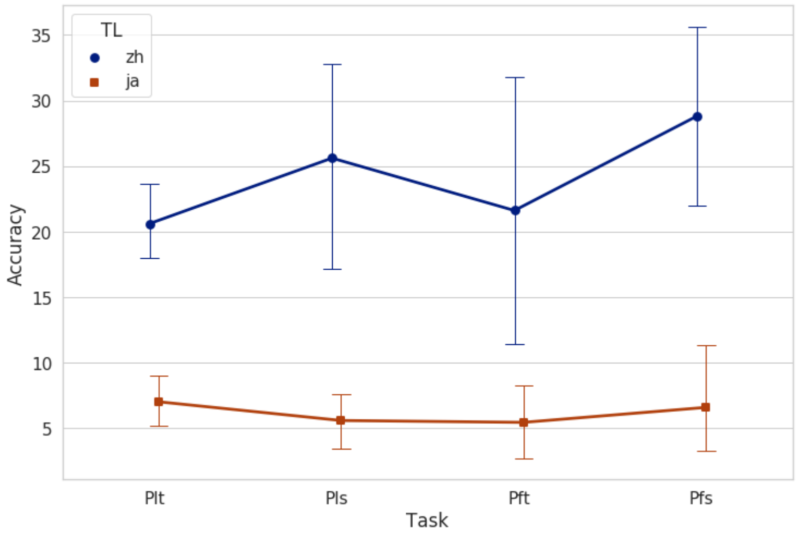


Figure 4: Comparing accuracy errors across tasks between Japanese and Chinese participants

On the other hand, we can clearly see a gradual decrease in Fluency errors for the Japanese participants when the PE task changes from LPE with TB (Plt) and LPE with IS (Pls), to FPE with TB (Pft) and FPE with IS (Pfs), as indicated in Figure 5. The PE versions of the Japanese participants have fewer Fluency errors with a FPE translation brief as opposed to LPE. Additionally, their PE versions show less Fluency errors under the IS condition as compared to TB. For the Chinese participants, however, their PE versions do not demonstrate stark differences in the occurrences of Fluency errors across the four tasks. As the LPE conditions asks to ignore Fluency issues in the MT output, this finding indicates to us that experienced (Japanese) translators are more sensitive to the translation brief than inexperienced (Chinese) translators. Surprisingly, we find that for Chinese participants, the PE versions of the FPE with TB (Pft) show more Fluency errors than the LPE with IS (Pls).

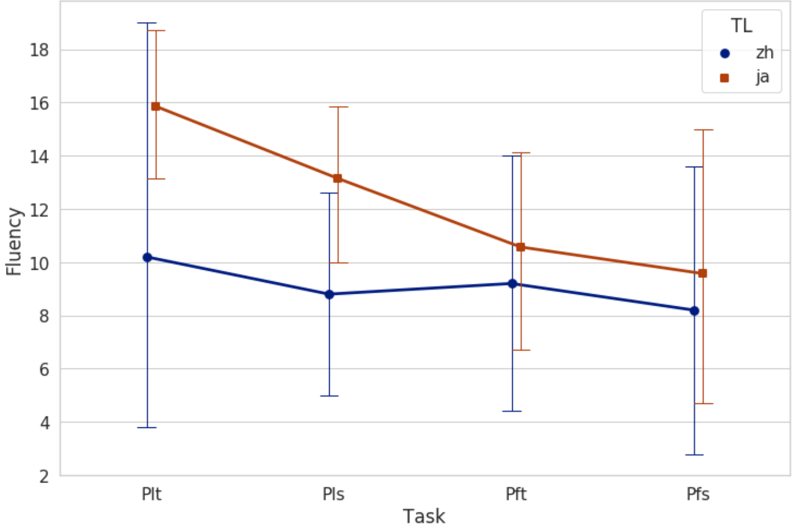


Figure 5: Comparing fluency errors across tasks between Japanese and Chinese participants

5.3 Translation proficiency

In the previous section, we investigate the difference between the two groups of participants regarding the impact of the four PE tasks. In this section, we further test if the impact of translation briefs and search conditions on PE quality are significantly different within each of the two groups of participants (Chinese and Japanese, respectively). We employ two-way ANOVA for each participant group, with four dependent variables which correspond to the four error labels (i.e., Accuracy, Fluency, Critical, and Minor). The two independent variables include translation brief (brief) and search condition (search). Our findings indicate that for Chinese participants, there are no statistically significant differences in the means of any error labels when comparing the translation briefs of FPE and LPE, but there are statistically significant differences in the mean of Accuracy errors between the TB and IS conditions ($p=.02<.05$), as shown in the following Table 4.

ANOVA Summary

	df	sum_sq	mean_sq	F	PR(>F)
Brief	1.0	7.07	7.07	0.32	0.57
Search	1.0	119.98	119.98	5.41	0.02
Brief:Search	1.0	0.71	0.71	0.03	0.86
Residual	231.0	5119.41	22.16	NaN	NaN

Table 4: ANOVA summary for Accuracy errors of Chinese participants

On the other hand, when comparing the search conditions of TB and IS, there are no statistically significant differences in the means of any error labels for Japanese participants, but there are statistically significant differences in the mean of Fluency errors between the translation briefs of FPE and LPE, as illustrated in the following Table 5. However, for both groups of participants, the interaction between translation brief and search condition has no statistically significant impact on the frequency of any error labels.

In short, the Chinese participants are more sensitive to the control of the search conditions

ANOVA Summary					
	df	sum_sq	mean_sq	F	PR(>F)
Brief	1.0	96.60	96.60	6.34	0.01
Search	1.0	40.73	40.73	2.67	0.10
Brief:Search	1.0	13.93	13.93	0.91	0.34
Residual	324.0	4938.24	15.24	NaN	NaN

Table 5: ANOVA summary for Fluency errors of Japanese participants

relating the Accuracy errors out of the four error labels. Additionally, there is no significant difference when it comes to the control of translation briefs regarding all the error labels. The Japanese participants, on the other hand, are more sensitive to the control of translation briefs relating the Fluency errors. Additionally, there is no significant difference when it comes to the control of search conditions regarding all the error labels. We suppose this results from the disparity in translation competence between inexperienced and experienced translators. Since inexperienced translators, as illustrated by the Chinese participants in this study, tend to have less profession-related competence (e.g. research skills) than experienced translators, as illustrated by the Japanese participants in this study, their PE versions have significantly less Accuracy errors when they are using the prepared set of terminology than when they are asked to search online but without proper research capabilities. Furthermore, as experienced translators tend to have a greater awareness of the differences between various translation briefs than less experienced translators do, their PE versions typically contain less Fluency errors when they are required to perform FPE rather than LPE. This is because experienced translators have more pragmatic competence than less experienced translators do (e.g., functional knowledge linked to translation briefs) (Yang and Li, 2021).

6 Conclusion

This paper aims to investigate the impact of translation briefs (full PE, FPE vs. light PE, LPE) and search conditions (provided termbase TB vs. free internet search IS) on PE quality of two groups of participants with varying levels of translation proficiency. To this purpose, four English STs from previous ATA certification exams (47 sentences, about 1,000 words) were automatically translated into simplified Chinese (zh) and Japanese (ja) by google NMT (GNMT), and were post-edited by five Chinese student translators and seven Japanese professional translators, respectively. The study was thus carried out in two language pairs (en-zh, en-ja) and the 12 post-editors produced a total of 658 segments. Keystrokes were logged and gaze data recorded, but these aspects of the experiment are not addressed in this paper.

To run the experiment under ecologically valid working conditions of professional translators, we conducted the experiment in the Trados workbench using the new Trados-Translog interface (Zou and Carl, 2022; Yamada et al., 2022). We asked participants to post-edit four texts under two types of translation briefs, i.e., FPE (f) and LPE (l), and two types of search conditions, i.e., TB (t) and IS (s). Therefore, we had four different PE tasks for each participant, i.e., Pfs, Pft, Pls, Plt.

The Chinese and Japanese GNMT outputs and the corresponding post-edited versions were annotated for translation errors based on an ATA-adapted error taxonomy. We grouped the errors under four labels, i.e., ”Accuracy”, ”Fluency”, ”Critical”, and ”Minor” errors. We calculated the error count by segment, aggregated them over the four PE tasks, and compared the error distribution in the two raw GNMT outputs (simplified Chinese and Japanese) and in the twelve post-edited versions. Our results show a similar error distribution for GNMT output and the PEMT versions. For both, GNMT and PEMT, minor fluency and critical accuracy errors were more common than other subcategories of errors. PEMT has generally lower percentages of fluency and minor errors than the GNMT, but a higher percentages of critical accuracy errors.

Looking into the error distribution for each of the tasks we see that, overall, FPE versions tend to have less errors than LPE versions. The translation brief of FPE improves in particular the participants' performance on fluency as compared to LPE, and the provided TB seems to improves the participants' performance on accuracy as compared to IS.

Across the four PE tasks, there are notable more Accuracy and Critical errors for Chinese than for Japanese participants. These results are to be expected, to the extent that the more experienced Japanese translators ought to deliver more frequently translations of higher quality than our inexperienced Chinese translators. Our findings also shows that inexperienced translators have significantly fewer accuracy errors in the TB condition as compared to searching online (IS). We assume that this is the case since less experienced translators typically possess less research skills (PACTE, 2003, 2005; Göpferich et al., 2009).

Experienced translators, on the other hand, seem to better realize implications of the translation briefs: with respect to accuracy errors, there is no significant difference across the four PE tasks. However, in the FPE condition, experienced translators produce less fluency errors as compared to LPE condition. This difference has not been observed for the less experienced translators, which suggests that experience leads to more awareness of the variations between different translation briefs.

Due to the restrictions of accessibility to the translators of our experimental language pairs, we only recruited twelve participants for this study. Therefore, there are certain limitations in the statistical results due to the relatively small sample size. However, we are currently collecting more data and intend to look into other aspects of translation process and translator behavior in future studies. The datasets are publicly available in the TPR-DB (Carl et al., 2016) and the reported results replicable.

References

Carl, M. (2012). Translog-ii: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4108–4112.

Carl, M. and Báez, M. C. T. (2019). Machine translation errors and the translation process: a study across different languages. *Journal of Specialised Translation*, 31:107–132.

Carl, M., Schaeffer, M., and Bangalore, S. (2016). The critt translation process research database. In *New directions in empirical translation process research*, pages 13–54. Springer.

Daems, J. and Macken, L. (2020). Post-editing human translations and revising machine translations: Impact on efficiency and quality. In *Translation Revision and Post-Editing*, pages 50–70. Routledge.

Göpferich, S., Jakobsen, A. L., and Mees, I. M. (2009). *Behind the mind: Methods, models and results in translation process research*, volume 37. Samfundslitteratur.

Gutt, E.-A. (2004). Applications of relevance theory to translation-a concise overview. *Retrieved March*, 4:2009.

Hu, K. and Cadwell, P. (2016). A comparative study of post-editing guidelines. *Baltic Journal of Modern Computing*, 4(2):346–353.

ISO (2017). *ISO 18587: Translation Services: Post-editing of Machine Translation Output: Requirements*. ISO.

JTF (2020). *2020 Nendo Honyaku Tsuyaku Hakusho: Dai 6 Kai Honyaku-Tsuyaku Gyokai Chosa Houkokusho [JTF Translation and Interpreting Report 2020: The 6th TI industry white paper]*. Japan Translation Federation.

Koby, G. S. and Champe, G. G. (2013). Welcome to the real world: Professional-level translator certification. *Translation & Interpreting, The*, 5(1):156–173.

Lommel, A., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.

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

Melby, A., Fields, P., Hague, D. R., Koby, G. S., and Lommel, A. (2014). Defining the landscape of translation. *Tradumàtica*, 12:0392–403.

Melby, A. K., Housley, J., Fields, P. J., and Tuioti, E. (2012). Reliably assessing the quality of post-edited translation based on formalized structured translation specifications. In *Workshop on Post-Editing Technology and Practice*.

Nitzke, J., Hansen-Schirra, S., and Canfora, C. (2019). Risk management and post-editing competence. *The Journal of Specialised Translation*, 31:239–259.

Nord, C. (2006). Translating as a purposeful activity: a prospective approach. *Teflin Journal*, 17(2):131–143.

Nunziatini, M. and Marg, L. (2020). Machine translation post-editing levels: Breaking away from the tradition and delivering a tailored service. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 309–318.

Onishi, N. and Yamada, M. (2020). Why translator competence in information searching matters: An empirical investigation into differences in searching behavior between professionals and novice translators. *Invitation to Interpreting and Translation Studies*, 22:1–22.

PACTE (2003). Building a translation competence model. *Triangulating translation: perspectives in process oriented research. Amsterdam;.*

PACTE (2005). Investigating translation competence: Conceptual and methodological issues. *Meta*, 50(2):609–619.

Pym, A. (2003). Redefining translation competence in an electronic age. in defence of a minimalist approach. *Meta: journal des traducteurs/Meta: Translators' Journal*, 48(4):481–497.

Sakamoto, A. and Yamada, M. (forthcoming, 2022). Managing clients' expectations for mtpe services through a metalanguage of translation specifications: Mppqn method. In *Metalanguages for Dissecting Translation Processes:Theoretical Development and Practical Applications*, pages 191–199. Routledge.

Shreve, G. M. (2006). The deliberate practice: translation and expertise. *Journal of translation studies*, 9(1):27–42.

Sturm, A. (2017). Metaminds: Using metarepresentation to model minds in translation. *Empirical modelling of translation and interpreting*, 7:419.

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. 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., Mirzapour, M., Jacquenet, H., and Vieira, L. N. (2021). Ai-based syntactic complexity metrics and sight interpreting performance. In *International Conference on Intelligent Human Computer Interaction*, pages 534–547. Springer.

Appendices

A Translation brief

A.1 Full post-editing

On this level of post-editing, the output shall be accurate, comprehensible and stylistically adequate, with correct syntax, grammar and punctuation. The aim of this level of post-editing is to produce an output which is indistinguishable from human translation output. Nevertheless, it is recommended that post-editors use as much of the MT output as possible. On this level of post-editing, post-editors shall focus on:

a) ensuring that no information has been added or omitted;

b) editing any inappropriate content;

c) restructuring sentences in the case of incorrect or unclear meaning;

d) producing grammatically, syntactically and semantically correct target language content;

e) applying spelling, punctuation and hyphenation rules;

f) ensuring that the style appropriate for the text type is used and that stylistic guidelines provided by the client are observed;

g) applying formatting rules.

A.2 Light post-editing

Light post-editing is normally used when the final text is not intended for publication and is mainly needed for information gisting, i.e. for rendering the main idea or point of the text. In this level of post-editing, the output shall be comprehensible and accurate but need not be stylistically adequate. At this pot-editing output level, post-editors should focus on:

a) using as much of the raw MT output as possible;

b) ensuring that no information has been added or omitted;

c) editing any inappropriate content;

d) restructuring sentences in the case of incorrect or unclear meaning.

B Experimental layout of the PE tasks

Proband	Task 1	Task 2	Task 3	Task 4
P01	Plt1	Pls2	Pft3	Pfs4
P02	Plt2	Pls3	Pft4	Pfs1
P03	Plt3	Pls4	Pft1	Pfs2
P04	Plt4	Pls1	Pft2	Pfs3
etc.				

Table 6: Experimental layout of the PE tasks

Note: The tasks for P05 are the repetition of the tasks for P01, and the tasks for P06 are the repetitions of P02, etc.