JoSTrans
The Journal of Specialised Translation

www.jostrans.org · ISSN: 1740-367X

Yamada, M. (2019). The impact of Google Neural Machine Translation on Post-editing by student translators. *The Journal of Specialised Translation, 31*, 87-106.
https://doi.org/10.26034/cm.jostrans.2019.178

The impact of Google Neural Machine Translation on Post-editing by student translators

Masaru Yamada, Kansai University

ABSTRACT

The author of this study used the design of a 2014 experiment that investigated college students' post-editing potential. The raw Google statistical machine translation (SMT) used in the 2014 experiment was replaced with the raw Google neural machine translation (NMT) of the same source text. A comparison of the results of the two studies yielded the following observations: 1) A quantitative evaluation of post-editing (PE) showed no significant difference in cognitive effort between the studies, but a significant difference in the amount of editing was observed. Overall, NMT+PE is better than SMT+PE in terms of its final product, which contains fewer errors; however, NMT+PE does not empower college students to meet professional standards of translation quality. 2) Students exhibit a poorer error correction rate in the NMT+PE condition despite similar perceived cognitive effort, which is possibly related to NMT producing human-like errors that make it more difficult for students to post-edit. 3) NMT+PE requires almost the same competence as translating a text 'from scratch' or editing human translation. Therefore, translation training is necessary for students to be able to shift their attention to the right problems (such as mistranslation) and be effective post-editors. The results of this study suggest that the more advanced, human-like translation abilities of NMT make it even more challenging for student translators to meet a professional standard of post-editing quality.

KEYWORDS

Post-editing, translation training, neural machine translation, student translator, error categories.

1. Introduction

The post-editing of machine translation (PE) has established itself as a translation service, as reflected in the International Standard (ISO 18587 2017) specifying standardised requirements for the post-editing process and post-editor competences. The requirements for PE stipulated in ISO 18587 are the same as the requirements for human translation stipulated in the ISO 17100 (2015). These requirements are also identical to those of the European Master's in Translation framework (EMT Expert Group 2017), consisting of six different skills: translation competence, linguistic and textual competence, competence in research/information mining, cultural competence, technical competence, and domain competence (ISO 18587 2017: 7). In addition, ISO 18587 requires post-editors to have certain knowledge and abilities about the post-editing task (ibid: 8), one of which is described as "a general knowledge of Machine Translation technology and a basic understanding of common errors that an MT system makes" (ibid). This seems to be a reasonable requirement for a professional who provides language services involving post-editing, but recognising errors made by an MT system is not as easy as it may seem, because different types of MT engines produce different errors. In addition, the speed of MT development, particularly neural machine

translation or NMT, is rapid and unpredictable, and it is almost impossible to predict common errors produced over time. However, if one needs to educate post-editors according to ISO 18587 professional requirements, it is important to understand common MT error types.

In 2014, the present author evaluated the quality of Google Statistical Machine Translation by investigating college language learners' post-editing (PE) performance (Yamada 2014). This study was based on an assumption that using high-quality statistical machine translation (SMT) would bolster non-professional translators' final product quality while reducing the level of effort they would need to invest in the task relative to the effort required for 'from-scratch' translation. The experiment comprehensively examined participants' perceived ease of task, the quantity of their edits, the quality of their final product, and their course grades. Although the results confirmed the initial assumption, it was concluded that the students' PE products did not meet professional standards, perhaps due to the limitations of SMT at that time.

In 2016, Google launched a neural machine translation (NMT) system with the potential to address many shortcomings of traditional SMT. Using a human side-by-side evaluation on a set of isolated simple sentences, Google NMT reduces translation errors by an average of 60% compared to Google SMT's phrase-based production system on the English-to-French and English-to-German benchmarks (Wu *et al.* 2016). Google NMT's launch attracted attention, especially in Japan's media and social networks, including the journal of the Japan Translation Federation, which described the news as "Google NMT Shock!" (Japan Translation Federation 2017). However, it is not yet known whether (Google) NMT is suitable for post-editing in terms of effort and types of error for the English-Japanese combination.

In this on-going transition, the author decided to provide a research-based assessment of Google NMT's quality in comparison with Google SMT by replicating his 2014 experiment with college students performing post-editing, but this time with NMT. The aims of this study are 1) to provide quantitative evaluations of changes in PE performance, cognitive effort, and amount of editing associated with NMT; 2) to investigate qualitatively detailed factors related to PE effort, particularly in terms of error types among SMT, NMT, and student post-editors; and 3) to identify implications for improving post-editor training.

2. Literature review

The literature on PE has grown considerably in the past decade. The general trend of PE research over this period is well organised in Koponen (2016). The growing popularity of this research area reflects the growing demand for PE. Earlier research carried out up until approximately 2010 centred on the productivity gains associated with PE in relation to time

and effort (e.g. Plitt and Masselot 2010). Although a certain degree of time reduction can be achieved by PE compared to human translation in traditional settings, temporal factors do not reflect post-editors' cognitive effort (e.g. Tatsumi 2009). Since then, a number of researchers began to investigate cognitive effort by measuring correlative variables such as self-reported perceived cognitive effort (Specia *et al*. 2010; Callison-Burch *et al*. 2012), revision amount shown by edit distance (e.g. Koponen 2012; Yamada 2014), and time spent editing different portions of text, as indicated by pauses found in keylogging data (O'Brien 2005; Lacruz *et al.* 2012) and eye tracking (e.g. Carl *et al.* 2011). Some other researchers identified difficulties during PE that contribute to increased cognitive effort and calculated correlations with effort indicators (e.g. Daems *et al*. 2015; Schaeffer and Carl 2017). In the present study, the author is interested in exploring changes in PE effort and error types emerging through the transition from SMT to NMT. Therefore, selected relevant literature on these aspects of PE is reviewed below.

Yamada (2014) investigates college language learners' PE qualifications, in terms of their overall translation grades, and how this correlates with their PE performance. The student translators' perceived effort (self-reported evaluation) was on average reduced by roughly 25% compared to the baseline effort required by a conventional human translation (HT) task. The amount of revision required to bring the MT texts to the requested level of quality was approximately 25% of the raw MT output. As for PE quality, students left about 7 errors uncorrected in their final products. The data also showed only a loose correlation between the students' general translation competence and their post-editing performance. While students who had poor grades in a traditional translation course were confirmed to be unqualified post-editors, students who received good grades were not always qualified post-editors either.

The tentative conclusion of this study was that the skill required for SMT+PE at that time was mainly the ability to correct basic linguistic errors, because the SMT engine for the English-Japanese language pair was of relatively poor quality. However, as previously stated, the MT engine has evolved into a system based on neural networks or NMT, which generates higher-quality translations and different error types. This change is expected to affect post-editing performance, including cognitive effort. Below is a summary of relevant literature on error types and cognitive effort.

Koponen (2012; Koponen *et al*. 2012) investigated the relationship between the amount of editing (evaluated with HTER) and a manual score reflecting human perceived effort. She found that sentences involving less effort, as indicated by higher manual scores or shorter post-editing times, tend to involve more edits related to word forms and simple substitutions of words of the same word class, while sentences with low scores or long

post-editing times involve more edits related to word order, edits where the word class was changed, and corrections of mistranslated idioms.

Comparin and Mendes (2017) present the results of a study involving an error annotation task of a corpus of machine translations from English into Italian. They compared error types found in raw MT and post-edited content, identified frequent and critical errors, and observed the errors' prevalence at different stages of the translation process. One pertinent finding is that 85% of the errors found in the raw MT were correctly revised through human post-editing; however, fluency errors decreased, while a relatively high number of accuracy errors were not corrected.

Daems *et al*. (2015) report on a post-editing study for general text types from English into Dutch conducted with master's students of translation. They used a fine-grained machine translation quality assessment method with error weights that correspond to severity levels and to cognitive load. They found that average MT quality (MT error weight) is a good predictor of six different post-editing effort indicators (average number of production units, average time per word, average fixation duration, average number of fixations, average pause ratio, and pause ratio), and that different types of MT errors predict different post-editing effort indicators.

3. Research question and sub-questions

As stated above, the focus of this study is to examine the impact of Google NMT on the ability of college students to be post-editors. To investigate this research question, an experiment was carried out to answer the following sub-questions, which are replicated from Yamada (2014).

Sub-question 1: What is the level of students' PE cognitive effort?
Sub-question 2: What is the level of students' PE revising effort?
Sub-question 3: What is the quality of the students' final post-edited product?

The sub-questions were respectively investigated based on the following measures: (1) Perceived ease of task, (2) Amount of editing and (3) Number and type of post-editing errors.

In contrast to the 2014 study, the relationship between a student's translation skill in terms of their translation course grades and PE performance is not explored; instead, the relationship between cognitive effort and error types will be the focus of additional detailed analysis.

3.1. Perceived ease of task

Reported perceived ease of task (sub-question 1) was used to determine PE cognitive effort. Participants were asked after completing their task to report their perceived effort by assigning a number to it as a proportion of their perceived level of ordinary translation (HT) effort (set at 100). For instance, if they felt that PE reduced their perceived effort by 20%, their rating would be 80 (100 minus 20). On the other hand, if PE was felt to increase their effort by the same ratio, then the response would be expressed as 120 (100 plus 20). Based on the results of the previous study (Yamada 2014), it is expected that NMT+PE would decrease the students' post-editing effort.

3.2. Amount of editing

The second aspect measured (sub-question 2) was the textual similarity between the raw MT output and the final target text. Out of the many automatic evaluation metrics available, GTM (General Text Manager: Turian *et al.* 2003) is utilised in this study because it is designed to evaluate relatively smaller segments or a sentence at a time. This measurement reveals the amount of text modified during the post-editing process. The values given by the metrics range from 0-1. The closer the values are to 1, the closer the final post-edited content is to the raw MT output, indicating the translators made fewer revisions. In this study, amount of post-editing is indicated by subtracting the GTM score from 1. Although the correlation between perceived effort and revision amount has not been found to be proportional (Tatsumi 2009), and a large amount of revision does not necessarily indicate high cognitive effort (Koponen 2012), the degree of textual similarity can serve as supporting evidence for perceived workload differences (Yamada 2014).

3.3. Quality

The third aspect considered is the quality of the final target text, determined by the number of translation errors made (sub-question 3). Students who participated in the experiment were instructed to perform a full post-edit, the quality of which was evaluated through MNH-TT (Mirai Hon'yaku for Translator Training) revision categories, a MeLLANGE-based (Castagnoli *et al.* 2011) error annotation system optimised exclusively for translation training purposes (details to be described below).

4. Research design, method, and participant profile

Since this is a replication of a 2014 experiment, the basic design, source text, and participant profiles are almost identical, except this time the Google NMT engine was used to pre-translate the text for the PE task.

4.1. Source text and experiment design

All the students were tasked with post-editing a 486-word English excerpt from Wikipedia into Japanese as part of their take-home examination for the course. The topic was Steve Jobs. The source text, pre-translated by Google Translate (June 2017), was prepared in a Word file format, in which the source text and raw MT output were laid out side-by-side. Students post-edited the raw MT output by overwriting it.

Participants were allowed a week to complete the entire assignment, which consisted of two passages to translate through ordinary HT in addition to the PE task. They were expected to complete the HT task first, followed by the PE task. After completing the assignment, students were asked to respond to a questionnaire about the ease of PE and provide open comments through a website prepared by the author.

4.2. Participant profile

The participant profile is nearly identical to the 2014 study except for the number of students and the university to which they were affiliated. A total of 28 students from Kansai University participated in the experiment, all of whom were majoring in English. In the previous experiment, 43 students from two universities with similar profiles took part in the study. Their English proficiency is also the same, with a TOEIC[1] score of about 800. In both studies, the students were not translation majors but participants in a practical translation course taught by the author as well other instructors at the same institute. In this hands-on practicum course students learned the basic practice of translation. Of the 15 weekly sessions, two to three were spent on PE, learning about its advantages and disadvantages, studying different types of post-editing, and practising some post-editing exercises. Given this limited training, however, students were not expected to have mastered post-editing by the end of the course.

5. Experiment results

5.1. Sub-question 1: PE effort

The results for the perceived effort required by PE are examined in this section. In response to the question "Did you find NMT+PE easier than HT?" 20 out of 28 students (71%) responded yes, while 74% answered yes in the previous experiment with SMT+PE. The students' ratings are relatively high, compared to professionals' ratings (see Yamada 2012). The next question asked students to rate their post-editing effort with a numeric ratio. A total average of 79.1% indicates that the NMT+PE task resulted roughly in a 21% reduction in PE effort when compared with HT (n=28, SD=29.7). The previous SMT+PE experiment had an average ratio of 75.1% for PE (n=43, SD=24.4). On the surface, it appears that

NMT+PE on average required greater effort than SMT+PE. However, the difference is not statistically significant (Wilcoxon signed-rank test: W = 518, p-value = 0.800). These results suggest that student post-editors who employ the NMT engine instead of SMT do not experience any additional effort during the PE process. The result is rather unexpected, given that Google NMT can offer a higher quality translation, at least in terms of fluency.

5.2. Sub-question 2: Amount of editing (GTM)

Although the relationship between cognitive effort and amount of editing is not proportional, a loose correlation between ease of PE and revision amount has been confirmed by previous studies (i.e. Yamada 2012, 2014).

The NMT and SMT groups showed an average amount of editing (1.000 minus GTM score, henceforth '1-GTM') of 0.210 (n=28, SD=0.088) and 0.247 (n=43, SD=0.064), respectively, indicating that post-editors using NMT made fewer revisions than those in the SMT group. The difference is statistically significant (Wilcoxon signed-rank test: W = 333, p-value = 0.002). With the results of sub-question 1 showing no significant difference in cognitive effort, it is inferred that NMT+PE is a case of a few edits requiring high effort. It is also noted that the amount of editing needed for NMT+PE, indicated by the 1-GTM of 0.210, is less than the amount of self-revisions (1-GTM 0.229) by students in the post-drafting phase of traditional human translation (Yamada 2012). This means that students make fewer revisions during NMT+PE than while revising their own first draft translation, which may also be related to 'effortful' editing with NMT+PE (i.e. a concentration of effort on fewer edits).

	SMT+PE	NMT+PE
Ease of PE (Effort)	75.1%	79.2%
Amount of editing (1-GTM)	0.247	0.210

Table 1. Effort and amount of editing: SMT+PE vs. NMT+PE (higher scores = higher ease of PE and higher amount of editing).

5.3. Sub-question 3: Quality

MT errors will be examined in this section to evaluate post-editing quality, as well as the reason for 'effortful' editing with NMT+PE. In this experiment, all errors were annotated using MNH-TT Revision Categories. Based on this error typology, all raw MT output errors of both NMT and SMT were annotated and categorised by severity (major or minor). Table 2 shows the total number of errors produced by the respective engines. Compared to raw NMT, raw SMT contains 1.5 times more errors, 72% of which are major errors. Major errors account for 37% of the total for raw NMT. From these results, it could be speculated that post-editing of raw SMT is more effortful than NMT because it includes a larger number of

severe errors (Daems *et al.* 2015). However, this prediction does not hold in this case, as evidenced by the sub-question 1 result where no difference in PE effort was confirmed.

	Raw SMT (2014)	Raw NMT (2017)
Major errors	31	10
Minor errors	12	17
Total	43	27

Table 2. Number of errors in raw SMT and NMT.

5.4. Post-editing quality

Post-editing quality is examined in this section based on how many errors in the raw MT were detected and correctly revised by student post-editors. For this evaluation, only major errors were considered.

	SMT+PE	NMT+PE
Average number of errors uncorrected (corrected)	6.9 (24.1)	3.20 (6.8)
Error correction rate	77.7%	68%
Variation in correction rate	41%-93%	40%-90%

Table 3. Post-editing quality: SMT+PE vs. NMT+PE.

Remarkably, the error correction rate of NMT+PE (68%) was worse than that of SMT+PE (77%). As for variance, Yamada (2014) shows that post-editors exhibit a large variance in error correction rate ranging from 93% (2 uncorrected errors) to 41% (18 uncorrected errors). The variance in the case of NMT+PE is roughly of the same range (40 – 90%). Therefore, in terms of relative quality, student post-editors using NMT do not necessarily outperform student post-editors using SMT+PE, let alone meet the professional quality standard of an 85% error correction rate (Comparin and Mendes 2017). Nevertheless, the absolute number of errors remaining in the final product after post-editing NMT is by far smaller than the SMT version, reduced nearly by half. In this respect, NMT can bolster non-professional post-editors.

5.5. Summary

There is no significant difference in cognitive effort between NMT+PE and SMT+PE (sub-question 1), despite NMT (1-GTM 0.210) requiring fewer revisions than SMT (sub-question 2). From these results, it is speculated that NMT+PE is a high-effort task. The results of sub-question 3 show a decrease in students' PE performance, with an error correction rate of 68%, compared to 77% in the SMT+PE condition (sub-question 3). By these measures, NMT focuses higher cognitive effort on fewer edits. In order to investigate the cause of this result, the following section analyses types of errors that may demand higher effort.

6. MNH-TT revision categories

This research has adopted a set of error categories (cf. revision categories) from MNH-TT (Babych *et al.* 2012) in order to compare errors before and after post-editing, and differences in errors between the types of MT and among the student translators. MNH-TT, a collaborative translation training platform, includes a menu of "revision categories," modified from MeLLANGE, that provides an error typology designed specifically for scaffolding translator competence (Toyoshima, *et al*. 2016; Yamamoto *et al*. 2016). Amongst other error categories such as Multidimensional Quality Metrics (MQM) and Dynamic Quality Framework (DQF), this set of categories is selected because it is optimised for translator training and best customised for the English-Japanese language combination (Toyoshima *et al*. 2016). MNH-TT revision categories provide a 'decision tree' diagram that guides instructors to make step-by-step error categorisation decisions. With this decision tree and the refined set of categories, multiple instructors or peer reviewers annotating the same translation error can achieve a high agreement rate (Toyoshima *et al*. 2016). Translation trainee students are taught the categories as part of the process of developing 'translator competence' (differentiated from translation competence). This serves to equip students with the ability to explain their own translation work using a set of revision categories as a meta-language that is common to all classmates when peer reviewing. One aim of the present study is to train post-editors, so the education-friendly categories are especially suited to enable learners as well as researchers and teachers to identify differences in error types between translation learners and between MT engines.

6.1. Learners' errors

We will explore what types of error translation learners make in a conventional human translation (HT) mode. An HT task was given to participants as a part of the experiment for this purpose. Students were asked to translate (not post-edit) the text, and to submit it along with the PE assignment after completion. Then, submissions were annotated by the experimenter with the MNH-TT revision categories.

The results provided in Figure 1 show an overall trend of students' translation errors where the category X3 (content distortion), also known as "mistranslation," is the most frequent one, accounting for 42% of all student errors. The X3 category is followed by X14 (target text inappropriate register), X4b (source text intrusion or translation too literal), and X2 (content addition). The overall results are roughly in line with previous studies that compare differences in errors among translation learners (Toyoshima 2016), and changes in error patterns in learners as their translation competence develops over time (Fujita 2017). It is confirmed in these studies as well as the present experiment that the error type X3 (mistranslation) is the most frequent error made by student

translators, followed by X4b, X14, X7, and X14. Translation learners consistently exhibit this signature pattern of errors where X3 occurs more frequently than the other categories. For this reason, the author carried out a separate investigation of X3 in the student translating process, using retrospective interviews to gain insight into how X3 errors occurred (Onishi *et al.* 2017). The results of this separate study will also be referred to later in this paper for discussion on post-editing errors and X3. For now, with this error distribution in mind, let us compare it to those of NMT and SMT.

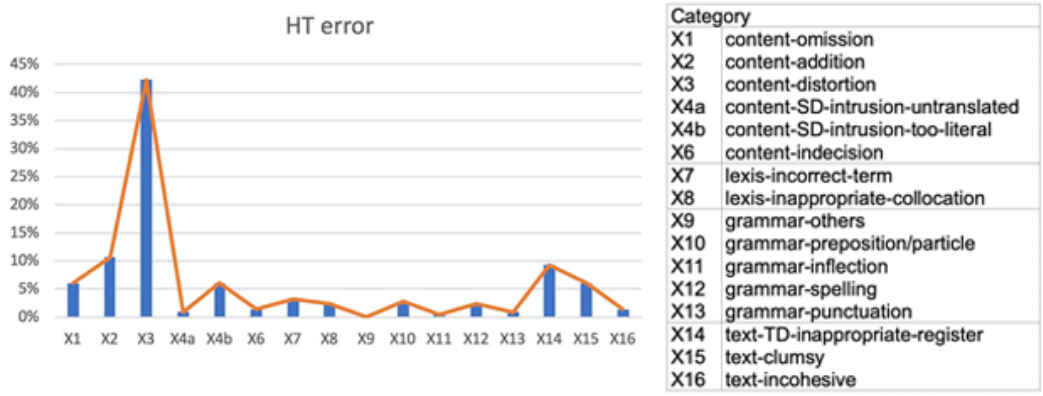


Figure 1. Error distributions of HT and MNH-TT revision categories.

6.2. SMT and NMT error distribution

Raw MT texts (SMT and NMT) before post-editing were annotated with the MNH-TT error categories. Note that error annotations are attached to errors inherent in the raw MT output, not to post-edited errors.

Comparing the error distribution of the raw MT outputs with the students' error pattern enables us to recognise interesting similarities and differences. As indicated in Figure 2, the rate of X3 is the highest for both NMT and SMT errors. However, more noticeably, the error pattern of the raw SMT is different from that of the raw NMT as well as student errors. The raw SMT presents high rates of other errors such as X4a (untranslated), X4b (too literal), X7 (incorrect terminology), X9 (syntax error), and X10 (preposition/particle error), while NMT's error distribution is similar to that of the students, with the exception of the higher frequency of X7 (incorrect terminology) in NMT.

Translation, including MT+PE, as Human-Computer interaction (HCI) must maintain a complementary relationship between human translators and machine translation, in that both need to compensate for each other's weaknesses to achieve optimal results. In this respect, SMT may be in a complementary relationship with student post-editors because its error types have a contrasting distribution when compared with error types seen in student human translation. In other words, raw SMT contains many error types, such as X4a, X4b, X7 X9, and X10, that are easy for

humans to detect and modify. In contrast, the NMT error pattern is roughly identical to that of human translators, and therefore lacks a complementary relationship. This accounts for NMT+PE requiring similar total effort as SMT+PE despite having fewer errors in the raw NMT output to address. It can be inferred that the most effort was spent on X3 errors.

Figure 2. Error distributions: HT, raw NMT, and raw SMT.

Yamada (2014) explained that one of the main causes of post-editing difficulty with SMT is improved MT fluency because human translators tend to pay the most attention to fluency factors, often neglecting accuracy (Fiederer and O'Brien 2009) and overlooking fatal meaning errors (Yamada 2014). The SMT engine with a machine-learning algorithm in 2014 had improved fluency, compared to earlier ruled-based MT systems. Thus, to some extent, it is true that improved fluency made errors harder to detect. However, taking the above findings into consideration, the author's previous explanation accounts only for error types specific to SMT such as X4a (untranslated), X4b (too literal), X7 (incorrect terminology), X9 (syntax error), and X10 (preposition/particle error).

With the advent of NMT which resolves typical SMT errors (in its raw output) before human post-editing, generating a similar error distribution to humans – NMT translates like humans, and produces errors like humans – another interest of this study became how to train students to compensate for this vulnerability, namely by improving the post-editing of X3 errors. In the following section, students' actual post-editing of X3 errors will be examined.

6.3. Post-editing errors in detail

This section examines post-editing errors. Unlike the HT and MT errors previously discussed, these are errors that are not detected or modified correctly during post-editing.

Figure 3 shows the amount of editing per segment, as indicated by the GTM score, and the percentage of post-editing errors contained in each segment for the 26 segments in the text.

Segment #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
GTM	0.156	0.258	0.288	0.265	0.166	0.119	0.241	0.160	0.131	0.096	0.165	0.245	0.202	0.150	0.154	0.271	0.069	0.127	0.199	0.035	0.187	0.228	0.170	0.221	0.206	0.145
Error rate	5%	0%	0%	0%	5%	0%	0%	0%	0%	0%	1%	14%	0%	38%	0%	0%	1%	0%	15%	0%	20%	1%	0%	0%	0%	0%

Figure 3. Amount of editing (1-GTM) and error rate per segment.

For instance, as the highest bar on the graph indicates, segment 14 contained 38% of the total errors made by all the students, meaning that, of all the errors students made during PE, 38% occurred in segment 14. This contrasts with the many segments that show 0%, in which students were able to correct the errors or the segment had no errors in the raw MT.

A closer look at segment 14 provides additional information showing the amount of editing as 1-GTM 0.150, in comparison to the average of 1-GTM 0.210, indicating less editing of this segment than average (lower 1-GTM means fewer revisions). This means the errors that occurred in segment 14 were undetected by post-editors. As another example, segment 12 has a relatively high error rate of 14% of all errors and a 1-GMT of 0.245 (higher than the average 1-GTM), indicating that post-editors spotted errors but were unable to correct them successfully.

This error-per-segment distribution was also examined in a previous experiment (Yamada 2014), exhibiting similar results. On the one hand, in SMT and NMT cases alike, segments 12 and 14 are both marked highest in error rate. On the other hand, in the case of SMT+PE, segments 2 and 4 contained high error rates, which have been resolved in the case of NMT. From this observation, it is worth noting that the types of errors in segment 12 and 14, which are found in post-editing of both SMT and NMT, represent errors that may demand high effort and be difficult for students to detect or correct. Thus, the next section will examine those two segments closely.

6.4. Quality in detail

Segment 12:
Source text
He was credited in Toy Story (1995) as an executive producer.

Raw NMT
彼は、トイストーリー（1995 年）のエグゼクティブプロデューサーとして入会しました。
[Kare-wa executive-producer to shite Toy Story (1995) ni nyuukai shimashita.]
(back-translation: He became a member as an executive producer in Toy Story (1995).)

Segment 12 includes an obvious mistranslation inherent in the raw MT output – X3, content distortion. The source word, *credited,* was rendered as *nyukai shimashita* (*became a member*), which was irrelevant to the context. Although this error seems obvious, few students could detect it, most likely due to their insufficient English language competence. Some students changed it to *shusshi shita* (invested money), which, though incorrect, makes some sense in that Steve Jobs did invest in this company. Overall, this error originates in lack of English language competence, rather than a cause specific to PE.

Segment 14:
Source text
In 1996, after Apple had failed to deliver its operating system called "Copland", Gil Amelio *turned to* NeXT Computer, and the NeXTSTEP platform became the foundation for the Mac OS X.

Raw NMT
1996 年、アップル社が "Copland"というオペレーティングシステムを提供しなかった後、Gil Amelio は NeXT Computer に転身し、NeXTSTEP プラットフォームは Mac OS X の基礎となりました。

[1996 nen ni, Appuru ga "Corpland" to iu operating system o teikyo shinakatta ato, Gil Amerlio wa NeXT Computer ni tenshin shi, NeXTSTEP platform wa Mac OSX no kiso to narimashita.]
(Back-translation: In 1996 after Apple did not deliver its operating system called "Copland", Gil Amelio moved to NeXT Computer, and the NeXTSTEP platform became the foundation for the Mac OS X.

Due to its relatively long sentence, segment 14 generates several errors related to accuracy and clarity. Apart from minor errors, the raw NMT output is very natural-sounding. It may not have been as easy for students to detect and make appropriate corrections. Particularly for the

verb phrase *turned to* (as in *Gil Amelio turned to NeXT computer*), which has been rendered as *tenshin shi* (*moved to*), it is difficult to find an appropriate translation that corresponds to the source without contextual knowledge. That is why this error was overlooked by most students. This major content-distortion is X3, and its cause may be related to students' English language competence, as seen in segment 12. The underlying issue is not specific to PE.

7. Mystery of X3 mistranslation

Toyoshima *et al.* (2016), who analysed trends in error types and improvements with respect to student learning level, found X3 to always be the most frequent error. In order to investigate the detailed causes of X3, the author of the present paper carried out an additional experiment examining the student translation process. The findings of this study are presented in this section because it includes some implications for translation training and the mystery of X3 errors.

According to Onishi *et al.* (2017), the root causes of X3 can be subcategorised into five different phases in terms of translators' attention mechanism, shown in Figure 4.

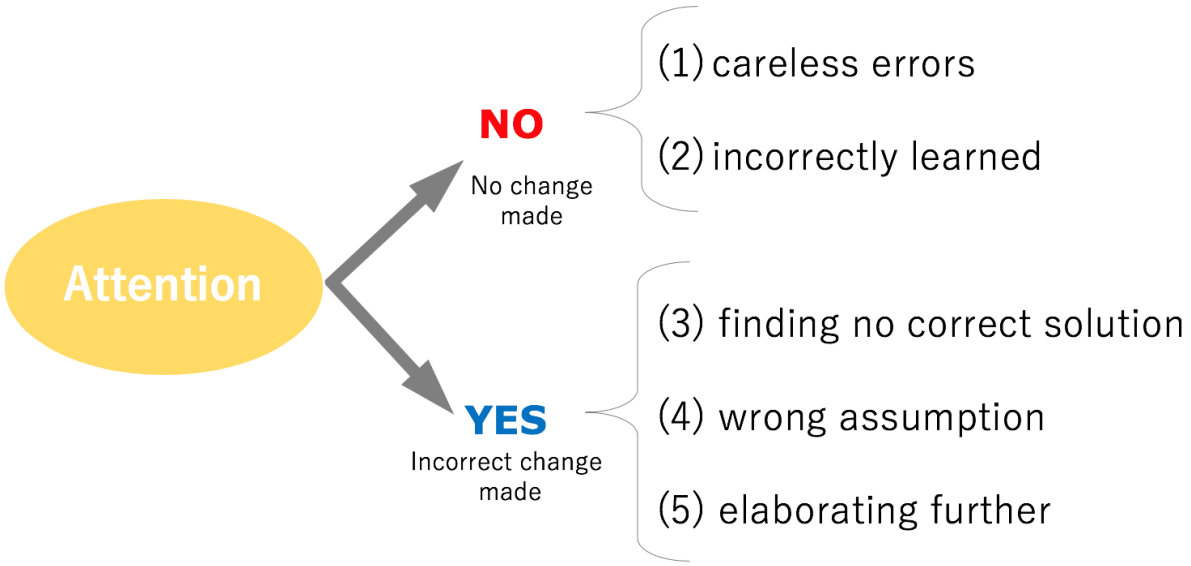


Figure 4. Attention mechanism of X3 error.

The first branching node divides the entire stage into two phases, whether or not the translator pays attention to the problematic source item. If attention is not paid, two possible scenarios can be considered: (1) *careless errors*—when a translator mistakes the meaning of the source word or text despite being capable of understanding the true intention—and (2) *incorrectly learned*—when the translator has learned and remembers the specific word or phrase incorrectly from the beginning. For (1) and (2), the translator is not aware of making the errors.

If attention is paid, three possible explanations for the X3 error can be considered. Scenario (3) *finding no correct solution* is when the translator noticed an error, but was unable to resolve it beyond an ambiguous translation (incorrect translation). Scenario (4) is similar to (3), except that the solution is based on a *wrong assumption* and results from the translator's best effort. Scenario (5) *elaborating further* means the translator understood the content and accordingly elaborated the target renditions further, but ultimately without success. Case (5) is different from (3) and (4) because the translator grasped the source text accurately, only to fail to produce a correct target rendition.

In terms of effort required for each scenario, (1) and (2) are less effortful because translators do not sustain their attention on the problem, which means errors are left uncorrected with nearly no effort. Cases (3), (4), and (5) require higher effort, although the final products are all still categorised as X3. In those scenarios, translators paying attention to the problem may attempt to revise it but do not succeed. Therefore, X3 in scenarios (3), (4), and (5) are incorrectly edited with high effort. From an educational point of view, the latter group of scenarios with attention paid to the problem are considered advanced learner behaviours, because they are evidence of conscious self-monitoring.

In this manner, two examples of X3 in segments 12 and 14 described above are examined. In segment 12, which contains a mistranslation of *credited*, most of the translators incorrectly revised the word. Although they were aware of the issue, they produced an erroneous rendition in the end. As also evidenced by the amount of editing of this segment being higher than average, this segment (this error) is indicative of effortful post-editing behaviour. In other words, students did pay attention to this particular problem during PE but could not find a correct solution, as in scenario (3), or edited it based on a wrong assumption, as in (4). In either case, this particular error X3 was an effortful error.

In the case of segment 14 with the X3 error of *turned to*, its relatively low GTM score among the majority of students suggests the error was overlooked. Therefore, scenario (2) "incorrectly learned" the meaning of *turned to* is applicable to this case, although scenario (1) "careless error" can be an option too. That is, the results suggest translators were not capable of being aware of the error and made no change to it; this X3 error was largely effortless.

As explored in this section, the causes of an X3 error can be divided into five different categories, and application of these categories to the examination of PE errors reveals that different processes were involved behind the same X3 errors. From an educational standpoint, students need to learn how to shift appropriate attention to the right problem during post-editing. This task is becoming more difficult with NMT+PE for two reasons: 1) NMT produces similar errors to humans, and 2) NMT's

language (English language) proficiency is above the level of most college students in Japan. According to one report, the current NMT engines can score 900 or better on the TOEIC (Mirai Hon'yaku 2017), exceeding this study's participants' average TOEIC score of about 800.

8. Conclusion

The author of this study used the design of a 2014 experiment (Yamada 2014) that investigated college students' potential when working as post-editors. The 2014 study's raw SMT output was replaced with raw NMT of the same source text to better understand the implications of the more advanced NMT technology. A comparison of the results of the 2014 study and the present one yielded the following observations:

(1) A quantitative evaluation showed no significant difference in cognitive effort or error correction rate, but a significant difference in the amount of editing of the raw MT output. Overall, students' NMT+PE is better than SMT+PE in terms of the final product, which contains fewer errors; however, NMT+PE does not empower college students to reduce cognitive effort or their error correction rate. It does help them to meet professional standards of translation quality.

(2) Poorer error correction in the NMT+PE condition despite similar perceived cognitive effort is related to the similarity in error distribution between NMT and humans, which renders them non-complementary. That is to say, NMT produces human-like errors which make it more difficult for students to post-edit.

(3) NMT+PE requires almost the same competence as ordinary human translation. Therefore, translation training is necessary for students to be able to shift their attention to the right problems (i.e. X3 errors) and be effective post-editors.

The present experiment has some limitations that need to be taken into consideration in future iterations. For instance, using multiple source texts would improve the generalizability of the results (Clark 1973) and controlling for time could yield different outcomes due to the effects of time pressure on cognitive function. Nevertheless, the results of this study suggest that, despite the hope that NMT could open up a new pool of post-editors in college students, the more advanced, human-like translation abilities of NMT make it even more challenging for untrained translators to meet a professional standard of post-editing quality. In this case, advances in technology have not yet eliminated the need for specialised translator education.

References

- **Babych, Bogdan, Hartley, Anthony, Kageura, Kyo, Thomas, Martin and Masao Utiyama** (2012). "MNH-TT: a collaborative platform for translator training." *Proceedings of Translation and the Computer* 34*, London, 29-30 November 2012*, 1–18. http://www.mt-archive.info/Aslib-2012-Babych.pdf (consulted 31.12.2017).

- **Callison-Burch, Chris, Koehn, Philipp, Monz, Christof, Peterson, Kay, Przybocki, Mark and Omar F. Zaidan** (2010). "Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation." Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan (eds) (2010). *ACL 2010: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden.* The Association for Computational Linguistics, 17–53. http://www.aclweb.org/anthology/W10-1700 (consulted 01.12.18).

- **Carl, Michael, Dragsted, Barbara, Elming, Jakob, Hardt, Daniel and ¬Arnt Lykke Jakobsen** (2011). "The Process of Post-Editing: A pilot study." Bernadette Sharp, Michael Zock, Michael Carl, Arnt Lykke Jakobsen (eds) (2011). *Proceedings of the 8th International NLPCS Workshop. Special Theme: Human-Machine Interaction in Translation*. Copenhagen: Samfundslitteratur, 131-142.

- **Castagnoli, Sara, Ciobanu, Dragos, Kübler, Natalie, Kunz, Kerstin and Alexandra Volanschi** (2011). "Designing a learner translator corpus for training purpose." Natalie Kübler (ed.) (2011). *Corpora, language, teaching, and resources: From theory to practice*. Peter Lang: Bern, 221-248.

- **Clark, Herbert** (1973) "The language-as-fixed-effect fallacy: A critique of language statistics in psychological research." *Journal of Verbal Learning and Verbal Behavior,* 12(4), 335-359.

- **Comparin, Lucia and Sara Mendes** (2017). "Using error annotation to evaluate machine translation and human post-editing in a business environment." *Proceedings of EAMT 2017, Prague, May 29-31.* https://ufal.mff.cuni.cz/eamt2017/user-project-product-papers/papers/user/EAMT2017_paper_76.pdf (consulted 21.12.2017).

- **Daems, Joke, Vandepitte, Sonia, Hartsuiker, Robert and Lieve Macken** (2015). "The impact of machine translation error types on post-editing effort indicators." Sharon O'Brien, Michel Simard and Joss Moorkens (eds) (2015). *Proceedings of the Fourth Workshop on Post-Editing Technology and Practice, Miami, October 30-November 3*, 31–45. https://amtaweb.org/wp-content/uploads/2015/10/MTSummitXV_WPTP4Proceedings.pdf (consulted 01.12.18).

- **EMT Expert Group** (2017) European Master's in Translation Competence Framework 2017. European Master's in Translation (EMT). https://ec.europa.eu/info/sites/info/files/emt_competence_fwk_2017_en_web.pdf (consulted 07.11.2018).

- **Fiederer, Rebecca and Sharon O'Brien** (2009). "Quality and machine translation: A realistic objective?" *The Journal of Specialised Translation* 11, 52-74.

- **Fujita, Atsushi, Tanabe, Kikuko, Toyoshima, Chiho, Yamamoto, Mayuka, Kageura, Kyo and Anthony Hartley** (2017). "Consistent classification of translation revisions: A case study of English-Japanese student translations." *Proceedings of the 11th Linguistic Annotation Workshop, Valencia, Spain, April 3*. The Association for Computational Linguistics, 57–66. http://www.aclweb.org/anthology/W17-0807 (consulted 21.12.2017).

- **ISO 17100** (2015). *Translation services – Requirements for translation services.* Geneva: International Standardization Organization.

- **ISO 18587** (2017). *Translation services – Post-editing of machine translation output – Requirements.* Geneva: International Organization for Standardization.

- **Japan Translation Federation** (2017). "Google NMT Shock." *JTF Journal* 288 (03.04.2017), 8-15.

- **Koponen, Maarit** (2012). "Comparing human perceptions of post-editing effort with post-editing operations." Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia (eds) (2012). *Proceedings of the 7th Workshop on Statistical Machine Translation, Montréal, June 7-8.* The Association for Computational Linguistics, 181–190. http://www.aclweb.org/anthology/W12-3100 (consulted 01.12.18).

- **—** (2016). "Is machine translation post-editing worth the effort? A survey of research into post-editing and effort." *The Journal of Specialised Translation* 25, 131–148.

- **Koponen, Maarit, Aziz, Wilker, Ramos, Luciana and Lucia Specia** (2012). "Post-editing time as a measure of cognitive effort." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012). *AMTA 2012 Workshop on Post-editing Technology and Practice (WPTP), San Diego, United States, 28 October*. http://157.56.13.76/AMTA2012Files/html/13/13_paper.pdf (consulted 15.10.2018).

- **Lacruz, Isabel, Shreve, Gregory M. and Erik Angelone** (2012). "Average pause ratio as an indicator of cognitive effort in post-editing: A case study." Sharon O'Brien, Michel Simard and Lucia Specia (eds) (2012). *Proceedings of the AMTA 2012 Workshop on Post-editing Technology and Practice, San Diego, 28 October*. Association for Machine Translation in the Americas, 21–30.

- **Mirai Hon'yaku** (2017). *Introducing the NMT engine with writing English Writing Ability of OverTOEIC 900 (TOEIC900 ten Ijo no Eisaku-bun Nouryoku o Motsu Shinso-Gakushu ni yoru Kikai Hon'yaku Enjine o Release)*. https://miraitranslate.com/uploads/2017/06/2d5778dcdee47e4197468bc922352179.pdf (consulted 21.12.2017).

- **O'Brien, Sharon** (2005). "Methodologies for measuring the correlations between post-editing effort and machine translatability." *Machine Translation* 19(1), 37–58.

- **Onishi, Nanami, Yamada, Masaru, Fujita, Atsushi and Kyo Kageura** (2017) "Causes of mistranslations made by student translators: Investigation into X3 in the MNH-TT revision category through retrospective Interviews." *Interpreting and Translation Studies* 18, 88-106.

- **Plitt, Mirko and François Masselot** (2010). "A Productivity test of statistical machine translation post-editing in a typical localisation context." *The Prague Bulletin of Mathematical Linguistics* 93, 7-16.

- **Schaeffer, Moritz and Michael Carl** (2017). "A minimal cognitive model for translating and post-editing." Sadao Kurohashi and Pascale Fung (2017). *Proceedings of MT Summit XVI vol. 1, Nagoya, Japan, September 18-22*, 144–155. http://aamt.info/app-def/S-102/mtsummit/2017/conference-proceedings/ (consulted 01.12.2018).

- **Specia, Lucia, Raj, Dhwaj and Marco Turchi** (2010). "Machine translation evaluation versus quality estimation." *Machine Translation* 24(1), 39–50.

- **Tatsumi, Midori** (2009). "Correlation between automatic evaluation metric scores, post-editing speed and some other factors." *MT Summit XII – The twelfth Machine Translation Summit International Association for Machine Translation hosted by the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas, 332-339. http://www.mt-archive.info/MTS-2009-Tatsumi.pdf (consulted 21.11.2018).

- **Toyoshima, Chiho, Fujita, Atsushi, Tanabe, Kikuko, Kageura, Kyo and Anthony Hartley** (2016). "Analysis of error patterns of translation students based on revision categories." *Interpreting and Translation Studies* 16, 47-65.

- **Turian Joseph P., Shen Luke and I. Dan Melamed** (2003) "Evaluation of machine translation and its evaluation." *Proceedings of the Ninth Machine Translation Summit, New Orleans, 23–27 September 2003*, 386–393. https://nlp.cs.nyu.edu/publication/papers/turian-summit03eval.pdf (consulted 01.12.18).

- **Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V. and Mohammad Norouzi** (2016) "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144*.

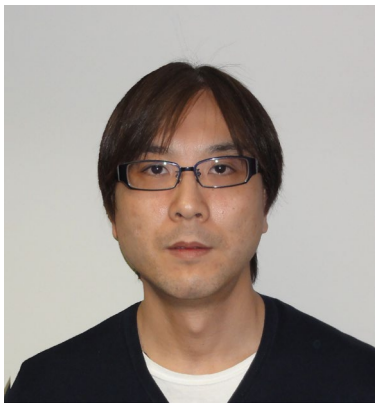
- **Yamada, Masaru** (2012) *Revising text: An empirical investigation of revision and the effects of integrating a TM and MT system into the translation process*. PhD Thesis. Rikkyo University at Tokyo.

- **-** (2014) "Can college students be post-editors? An investigation into employing language learners in machine translation plus post-editing." *Machine Translation* 29(1), 49–67.

- **Yamamoto, Mayuka, Tanabe, Kikuko and Atsushi Fujita** (2016). "Changes in translation learner's error types over time (Hon'yaku Gakushu-sha noGakushu-Katei ni okeru Error no Keikou no Henka)." *Proceedings of NLP2016* (22), 865-868. http://www.anlp.jp/proceedings/annual_meeting/2016/pdf_dir/E5-3.pdf (consulted 21.12.2017).

Biography

Masaru Yamada is a Professor in the Faculty of Foreign Language Studies at Kansai University. He specialises in Translation and Interpreting Studies with a focus on translation process research (TPR), including translation technology and post-editing, translation in language teaching (TILT), and neuroscientific approaches to TPR.



E-mail: yamada@apple-eye.com

Notes

[1] TOEIC or the Test of English for International Communication is an English language test designed specifically to measure the ability to use English in everyday workplace activities. A TOEIC score of 800 is equivalent to C1 to C2 in CEFR (The Common European Framework of Reference for Languages).

Corrigendum: Figure 1 was replaced in June 2022 by a figure containing the same content as problems were identified with the copyright of the original graphic.