Journal of Modern Languages
Vol. 34(2) (2024)
https://doi.org/10.22452/jml.vol34no2.7

Pre-editing rules developed for higher-quality target language texts

Kayo Tsuji
tsujikayo@omu.ac.jp
Faculty of Liberal Arts, Sciences and Global Education, Osaka Metropolitan University,
Osaka, Japan
ORCiD: https://orcid.org/0009-0005-3956-9602

Received: 25 April 2024 Accepted: 19 August 2024

Abstract

Neural machine translation (NMT/MT) has recently developed by leaps and bounds through deeply learning human brain mechanism, and its effective use has been the focus of much attention. To generate high-quality target language (TL) texts, MT users or TL learners need to edit the source language (SL) text before applying NMT. The purpose of this study was to provide pedagogical implications on pre-editing for effective NMT use with Japanese as a SL. The participants were 23 Japanese students with intermediate TL (English) proficiency, and the targeted task was a Japanese (SL) written argumentative essay and the English (TL) output thereof. Three language researchers systematically examined and analysed issues with the NMT output to detect the problematic factors in SL texts causing the issues and developed pre-editing rules. The results demonstrated the following five principles users should be aware of: (1) simplifying statements, (2) clearly stating all necessary sentence elements, (3) noting the logical placement or setting of sentence elements, (4) paying careful attention to the semantic properties of lexical items in SL texts, and (5) considering appropriacy of genre-specific SL lexical items. These rules should be applied to improve the appropriacy of syntactic and semantic aspects of TL texts.

They may also assist TL learners translating from Japanese to English independently, since the output of recent NMTs is largely considered close to human translations.

Keywords: Pre-editing Rules, Neural Machine Translation, MT Use, Academic Writing, English

1. Introduction

Neural machine translation (NMT) has emerged in recent years and has developed in accuracy compared to statistical machine translation (SMT) (Bahdanau et al., 2015; Sutskever et al., 2014). However, details of the internal structure and function of NMTs have not been revealed, resulting in it generally being considered a ‘Black Box’. In general, the translations output by recent NMTs are regarded to be largely close to human translations. For efficient use of such NMTs, recent studies (e.g., Bounaas et al., 2023; Hiraoka & Yamada, 2019; Miyata & Fujita, 2021; Zheng et al., 2022; Zhivotova et al., 2020) have focused on pre-editing – the process of editing a document before applying NMTs – and provided profound insights into its effects. Some research has reported the success of pre-editing at developing vocabulary quality (Cheng et al., 2021; Feifei et al., 2022) and improving communicative quality (Sánchez-Gijón & Kenny, 2022; Shih, 2021). Based upon the results of previous research, pre-editing contributes toward improving the quality of MT output (e.g., Kokanova et al., 2022; Simonova & Patiniotaki, 2022) and that applying pre-editing rules enables most students to produce better target-language (TL) texts (Farhana et al., 2023; Tuzcu, 2021).

2. Previous Research

2.1. Pre-editing Rules for Respective Language Pairs

One of the seminal studies on pre-editing rules for SMT is Seretan et al. (2014)[1]. Prior to this study, an online pre-editing tool was established based on rules which consider the linguistic features of English and French, with similar rules when translating in either direction. With English as a source language (SL), the rules are to “distinguish between upper- and lower-case letters, appropriately use punctuation, avoid homophone confusion, and correct grammatical

[1] Beyond establishment of an online pre-editing tool which applies said rules, this study explores the tool’s influence on MT output.

inconsistencies" (p. 1795). With French as a SL, the rules are to "correct homophones, appropriately use punctuation, and consider grammatical and style issues" (p. 1795). This study reported that an online pre-editing tool based on the said rules contributed to an improvement in the quality of MT output.

Along with recognition of NMTs advantages and effects, studies have also presented pre-editing rules for NMT use. Zhivotova et al. (2020), targeting the Russian-English language pair, developed pre-editing rules for NMT, and examined how pre-editing based on the said rules affected the quality of MT output[2]. The rules mainly focused on the following aspects: appropriate usage of words and phrases (e.g., "minimize the number of abbreviations in the text," "avoid complex words and phrases, replace with simpler synonyms where possible"), correct grammatical structures (e.g., "break down complex sentences into two or more simple ones," "start a sentence with a subject and a predicate"), and apposite text design (e.g., "check the text for punctuation errors and excessive punctuation") (p. 1784). Based on the results of evaluating the output quality using BLEU values[3], the quality of MT output with pre-editing was more than twice that of those without it, suggesting that pre-editing based on the aforesaid aspects was effective at improving the output quality.

Likewise, with an Asian-English language pair, Taufik (2020) examined English MT outputs translated from Indonesian source texts (STs) and identified the following pre-editing rules to be noted: avoiding longer sentences, clarifying the subordinate and correlative conjunctions, and paying attention to the usage of appropriate SL words. Additionally, Zheng et al. (2022), targeting Chinese (SL) product instructions, investigated how pre-editing based on Controlled Chinese Rules (CCR) affected MT outputs. The CCR required that all sentences have a clear subject, repeated expressions be removed, sentences be short, and a clear logic be established within and between sentences. The results showed statistically significant differences in adequacy, fluency, and style between MT output with and without pre-editing: The quality of the output with pre-editing was remarkably improved in terms of the said aspects.

Considering the similarities and differences of each pair, the rules for SMT developed by Seretan et al. (2014) mostly relate to the lexical, grammatical, and formatting aspects of a text. On the other hand, Zhivotova et al. (2020), Taufik (2020) and Zheng et al. (2022), studying recent

[2] It examined the written English output from a scientific and technical article written in Russian.
[3] This stands for Bilingual Evaluation Understudy values and is the most mainstream MT evaluation index.

NMTs instead, include rules not only on lexical and syntactic elements, but also on semantic aspects.

2.2. Pre-editing Rules Effective for Japanese Domains

Previous studies underline the importance of tailoring pre-editing strategies to the specific linguistic and stylistic challenges of the source language (SL) and its domain. Considering the fact that the linguistic features of each language are different, the examination of cases between Japanese and English is crucial. Inspired by the question of how pre-editing works for NMTs, Miyata and Fujita (2021) investigated the impact of manual pre-editing strategies on the accuracy of MT output for hospital medical documents, municipal office paperwork and newspapers. The results revealed that amendments such as simplifying SL sentences (SSs) and clarifying each element in a SS (for instance, by reordering phrases, adding necessary information, and using synonymous words) positively affect the syntactic and semantic qualities of NMT output. Thus, this study suggested that pre-editing rules for NMT use highlight semantic considerations in accordance with the context of source-language texts (STs).

Another influential study is by Hiraoka and Yamada (2019). They aimed to develop pre-editing rules targeting subtitle translations of TED talks[4] when translating from Japanese (SL) to English (TL) by volunteer TED Talk viewers. This study examined the effectiveness of "inserting punctuation, clarifying implicit subjects and objects in sentences, and writing proper nouns" in STs on MT output (p. 64). Beyond development of pre-editing rules, a comparative analysis was conducted between an MT output with pre-editing and one without. The results revealed that following certain rules increased the accuracy of MT output, thus concluding that pre-editing STs before applying NMTs is a promising approach for translating TED Talk subtitles. However, the study also presents cases in which the quality of MT output with pre-editing failed to improve or even deteriorated, possibly resulting from the editors' technical skills.

Apart from MT use, Tsuji (2022) investigated the educational value of students' Japanese (SL) paraphrasing on the target-language text (TT) quality. The participants initially generated their TTs on their own, then identified particular TL sentences (TSs) with low readability, and finally paraphrased SSs to re-translate. The participants paraphrased by using appropriate subjects and verbs, avoiding redundancy, and simplifying sentences, etc. Explicitly teaching paraphrasing

[4] This study involved TEDx Tokyo 2013, "a free online video service led by TED" (p. 65).

rules, with reference to the pre-editing rules known as PACE (Pym, 1990)[5], positively affected students' learning. Their final TSs demonstrated that each lexical item was ordered more logically, leading each TS to be more easily understood. This analysis revealed that the SL paraphrasing improved the quality of the TL learners' manual translation, suggesting that pre-editing rules for NMTs may be beneficial even for inexperienced translators.

3. Purpose of the Study

Recent studies exploring pre-editing rules have provided crucial pedagogical implications for MT users or TL learners to produce higher-quality TTs. Although the aforementioned studies in Japanese settings (e.g., Hiraoka & Yamada, 2019; Miyata & Fujita, 2021) contained useful suggestions, the findings might not be applicable to academic writings as they analised texts in different domains. Thus, this study investigated basic academic texts produced by Japanese learners of English. Studies conducting a systematic analysis into pre-editing rules for academic texts, i.e., argumentative or opinion essays, are few and far between. In light of the possibility that effective rules may vary by document domain, research to clarify the rules for such essays is needed in order to provide an effective pedagogical guide to MT users and TL learners alike. Accordingly, the present study focused on basic academic writing texts and attempted to investigate the following two research questions (RQs):

RQ1: What types of issue in basic SL academic essays, written by TL learners, cause problems with MT output?

RQ2: What rules should be followed during a pre-editing phase to effectively produce better TL academic essays with a Japanese-English language pair?

The findings of this study could provide MT users and TL learners with practical suggestions on how STs should be constructed before applying NMTs. They could also offer useful strategies for effective NMT use, as well as to improve students' manual translation efforts.

4. Method of the Study

This study involved 23 Japanese students with intermediate TL (English) proficiency, who were selected on the basis that all were typical Japanese university students with no experience of living

[5] PACE stands for Perkins Approved Clear English, which is the representative pre-editing rule-set for global audiences.

abroad. In terms of language proficiency, all were around B1-B2 of the Common European Framework of Reference for Languages. Explicit consent was obtained from the participants, who were given an information sheet regarding the study and the data collection, whereupon they signed their names to indicate consent.

Two argumentative essays of approximately 450 Japanese characters were assigned to each participant. The first question[6] was 'Why do you learn English? Do you think it is necessary to know more than one language? Explain why you think so' and the second, 'Do you think the world should have a universal language?' There was no randomization or blinding during the experimental process and all participants performed the same task. Participants were first asked to summarize their opinions on the assigned topics in the SL (Japanese) and then have NMT translate their writing into the TL. Among the free online NMTs, this study focused on DeepL since it has been shown to perform well on the BLEU values, that is, the most mainstream MT evaluation measure (Fujii et al., 2021). Based upon the finding that NMT tends to produce better quality texts when document-level contextual information is input (Kim et al., 2019; Miculicich et al., 2018), this study asked students to input their texts into DeepL by paragraph-unit. As all students submitted their essays, a total of 46 writing samples were collected. The MT output was examined by three linguistic analysts: a native English and two native Japanese linguists. The analysts initially identified discrepancies between the writer's intentions and the MT output, or separated awkward or mistranslated TSs, and then identified the problematic factors in the SSs causing said issues. They recorded and categorized the issues only upon mutual agreement.

5. Results and Discussion

This section firstly presented the answers to RQ1 and RQ2, and then described how the revised SSs contributed to the improvement of TSs by illustrating examples drawn from the data.

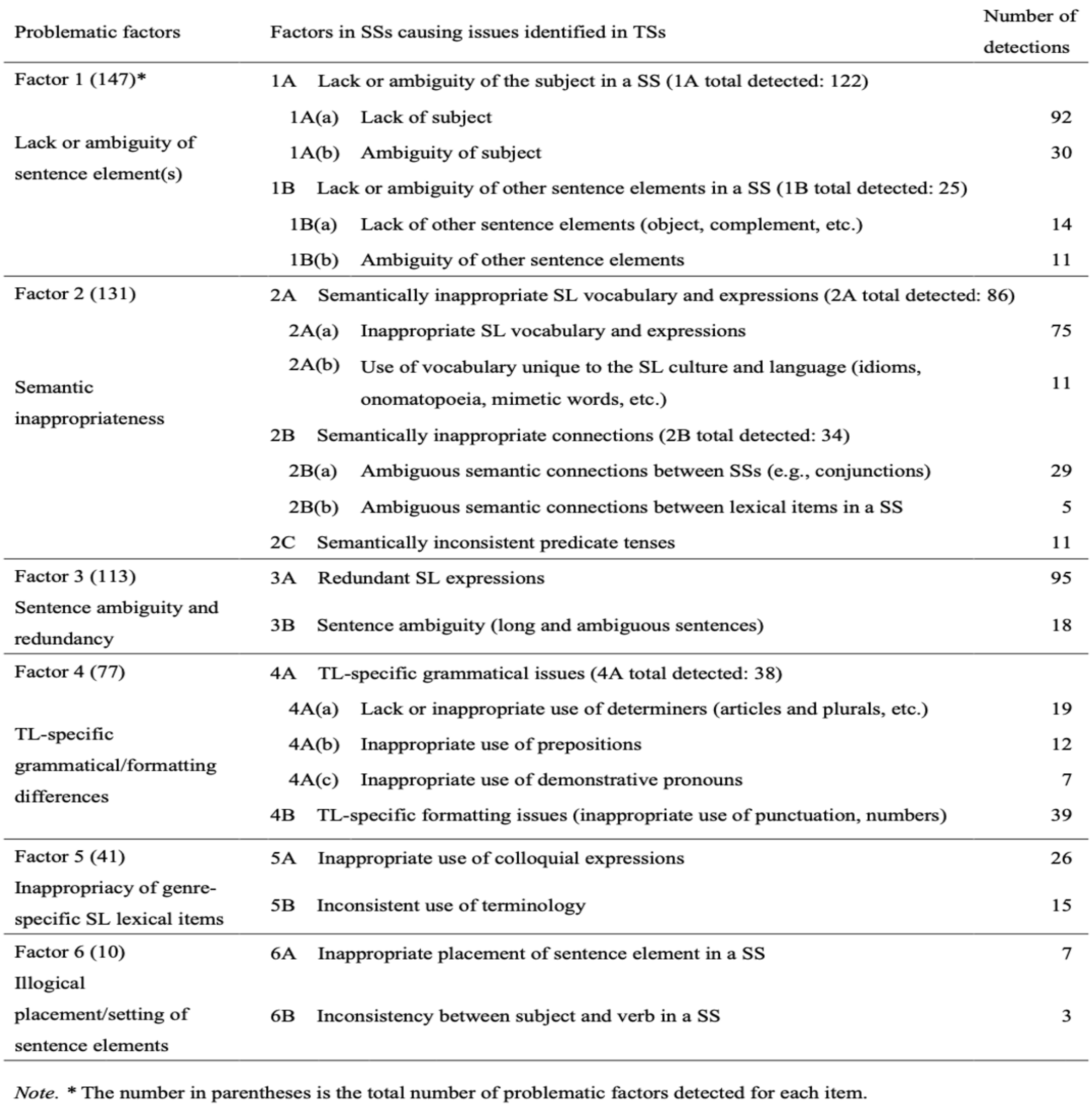
5.1. Types of Issue in SSs Causing Problems in TSs

The total number of problematic TSs was 303 out of 554. As a result of detecting the elements in SSs causing issues in TSs, 19 problems were separated into six broader categories based on commonalities in the errors. The problematic factors identified in SSs provided the answer to RQ1

[6] The textbook used in the participating classes was Weaving It Together 4 (Fourth Edition), by Milaca Broukal, published by National Geographic Learning. The questions can be found on p.13 of this textbook.

and were presented in Table 1 in order of the greatest number of issues detected. The most frequent problem related to the lack or ambiguity of sentence elements in SSs, was categorized in Factor 1. Amongst a total of 147 detections, 122 were associated with the lack or ambiguity of a subject in SSs (1A). This was followed by semantic inappropriateness (Factor 2), particularly issues relating to semantically inappropriate SL vocabulary and expressions (2A). The third was associated with sentence ambiguity and redundancy in SSs, which generated 113 detections (Factor 3). Most of the detected issues were related to redundant SL expressions (3A).

Table 1. Problematic Factors in SSs Causing Issues in TSs



Problematic factors	Factors in SSs causing issues identified in TSs	Number of detections
Factor 1 (147)*	1A Lack or ambiguity of the subject in a SS (1A total detected: 122)	
Lack or ambiguity of sentence element(s)	1A(a) Lack of subject	92
	1A(b) Ambiguity of subject	30
	1B Lack or ambiguity of other sentence elements in a SS (1B total detected: 25)	
	1B(a) Lack of other sentence elements (object, complement, etc.)	14
	1B(b) Ambiguity of other sentence elements	11
Factor 2 (131)	2A Semantically inappropriate SL vocabulary and expressions (2A total detected: 86)	
	2A(a) Inappropriate SL vocabulary and expressions	75
Semantic inappropriateness	2A(b) Use of vocabulary unique to the SL culture and language (idioms, onomatopoeia, mimetic words, etc.)	11
	2B Semantically inappropriate connections (2B total detected: 34)	
	2B(a) Ambiguous semantic connections between SSs (e.g., conjunctions)	29
	2B(b) Ambiguous semantic connections between lexical items in a SS	5
	2C Semantically inconsistent predicate tenses	11
Factor 3 (113)	3A Redundant SL expressions	95
Sentence ambiguity and redundancy	3B Sentence ambiguity (long and ambiguous sentences)	18
Factor 4 (77)	4A TL-specific grammatical issues (4A total detected: 38)	
TL-specific grammatical/formatting differences	4A(a) Lack or inappropriate use of determiners (articles and plurals, etc.)	19
	4A(b) Inappropriate use of prepositions	12
	4A(c) Inappropriate use of demonstrative pronouns	7
	4B TL-specific formatting issues (inappropriate use of punctuation, numbers)	39
Factor 5 (41)	5A Inappropriate use of colloquial expressions	26
Inappropriacy of genre-specific SL lexical items	5B Inconsistent use of terminology	15
Factor 6 (10)	6A Inappropriate placement of sentence element in a SS	7
Illogical placement/setting of sentence elements	6B Inconsistency between subject and verb in a SS	3

Note. * The number in parentheses is the total number of problematic factors detected for each item.

5.2. Rules to be Followed during a Pre-editing Phase

In response to RQ2, analysts determined that six rules, shown in Table 2, should be followed during a pre-editing phase. The rules were presented in the order of importance when constructing SSs.

Table 2. Pre-editing Rules Developed for TL Learners' MT Use

Rule No.	Rule to avoid issues in STs	Related factor
Rule 1	Simplify sentences (avoid long and redundant sentences).	3A, 3B
Rule 2	Clearly state all necessary sentence elements.	1A, 1B
Rule 3	Note logical placement/setting of sentence elements.	6A, 6B
Rule 4	Note semantically appropriate connections.	2B
Rule 5	Note semantic properties of lexical items.	2A, 2C
Rule 6	Consider appropriacy of genre-specific SL lexical items.	5A, 5B

The solution for semantic inappropriateness (Factor 2 in Table 1) included two rules: Rule 4 for the problematic factor 2B, and Rule 5 for 2A and 2C. The related factor, shown in the far-right column in Table 2, was listed as corresponding to each rule.

Comparing the rules drawn from this study to those of other studies, Rules 1, 2 and 5 can be applied to most of language pairs. Amongst studies of the Japanese-English pair, the results of this study mostly reflected those of Hiraoka and Yamada (2019) and Miyata and Fujita (2021). Especially, as a measure to improve the lack or ambiguity of sentence elements (Factor 1 in Table 1) which had the highest number of detections, Rule 2 in Table 2 (Clearly state all necessary sentence elements) is also emphasized by the aforementioned studies. This rule is consistent with the notion that, in Japan, there is a tendency to communicate by increasing the level of abstraction, and a mutual habit of 'tacit understanding' is thought to influence the 'omission of sentence elements' (Tsuji, 2021).

On the other hand, despite targeting the same language pairs, a greater number of 'ambiguous semantic connections between SSs' [2B(a) in Table 1] were detected as problematic elements in the present study. Consequently, the importance of the remedial measure 'Note semantically appropriate connections' (Rule 4 in Table 2) can be interpreted as being reasonably high, but this result may be due to the particular characteristics of the texts under analysis. While the other two studies focused on procedural documents from public institutions, newspaper articles and speeches by experts, the present study focused on learners' argumentative essays. In the realm of Japanese academia, student writing often suffers from ambiguous semantic connections between SSs [2B(a) in Table 1] and ambiguous semantic connections between lexical items in a SS [2B(b) in Table 1] (Iwasaki, 2021; Oshima, 2010). It can be inferred that the writing of the participants in this study was no exception, which may have led to certain deficiencies in the output text related to said issues.

Aside from the Japanese-English pair, the aspect relating to 2B was reflected in Zheng et al. (2022), targeting product instructions. For such instructions must establish a clear logic and clearly describe a method throughout a process.

5.3. Revising SSs to Generate Higher-quality MT Output

Focusing on the three most frequent problematic factors, the analysts explored how the MT output could be improved by revising SSs using the established pre-editing rules. These were illustrated with reference to specific examples drawn from the students' data.

5.3.1. Clearly State All Necessary Sentence Elements to Solve Factor 1 (Rule 2 in Table 2)

Regarding the lack or ambiguity of sentence elements (Factor 1 in Table 1), most observed cases were instances where MT set the subject on its own due to the lack or ambiguity of a subject in a SS. MT also frequently suggested multiple translations when it was confused with the content and context of STs. The solution for this type of issue is to guide MT in the right direction by clearly stating all necessary sentence elements. The MT translation below clearly demonstrated this problem.

> For example, Western music. If (1) you like a particular piece of Western music, (2) you will not be able to fully enjoy it if (3) you cannot understand the title of the song or the meaning of the lyrics.

This translation included structurally incomplete sentences and inappropriate personal pronouns for written documents. Specifically, from the outset it was not easy to grasp what

Western music could be an example of, as the first sentence in the MT output appeared to be just one part of a greater sentence. Upon analysing the ST content, the lack of a sentence element was noted. To solve this issue in the TS, i.e., to make the subject and verb in the TS clearer, the subject and verb in the SS should be clearly stated by rephrasing it as ‘A good example is Western music’ or ‘One example is Western music’. Likewise, the subsequent SS lacked a subject. Generally, MT cannot understand the ST’s context, or even the style required for the written documents in question, resulting in it often outputting ‘you’, ‘they’ or ‘we’ as a subject. This case was no exception. Since MT got lost while setting the subject in the TS, ‘you’ was selected [underlined parts (1), (2), and (3)]. However, the use of second-person pronouns should generally be avoided in academic writing (Hyland & Jiang, 2017). Accordingly, the SS was revised with the subject clearly stated for readers to understand who likes Western music, resulting in the following translation:

> (4) A good example of this is Western music. If (5) a person loves a particular piece of Western music, but does not understand the title or the meaning of the lyrics, (6) he or she will not be able to fully enjoy it.

The MT translation with a clear subject and verb was output in the first sentence [the underlined part (4)]. Furthermore, the subject of the following sentence was clearly stated as ‘a person’ [the underlined part (5)]. As seen in the underlined part (6) of the MT output after pre-editing the SS, by clarifying the subject, NMT automatically corrected the personal pronouns associated with the subject.

In Japanese, the subject does not exist in the same way as in English (Mikami, 1975). It is often omitted in written texts as the meaning can be understood without the subject being clearly stated (Tsuji, 2024). Thus, it is important for MT users to understand the structure and linguistic features of the TL (English) and to clearly state the subject when composing STs. Even when the subject indicates a general person or people, it should be clearly stated as ‘a person’ (for a singular subject) or ‘people’ (for plural subjects) in each SS. This example demonstrated that the subject of each sentence of the ST should be clearly written as an essential element of each sentence in order to effectively use NMT.

The second example displayed how the issues arising from the lack of subject [the factors of 1A(a) in Table 1] and the lack of other sentence elements [1B(a)] could be improved.

> Based on the above, I think (1) the answer to the second question is (2) should know.

The information of the underlined part (1) is semantically unclear and unnecessary. Unnecessary information to understand the content of the text should be removed from the ST. In the underlined part (2), the information regarding who and what were not specified. The lack of a subject and object of ‘should know’ led to semantic ambiguity. Clearly stating sentence elements other than the subject is also important for NMTs to understand a SS (Hiraoka & Yamada, 2019).

Based on the analysis of this issue type, STs should be created with clearly conveyed sentence elements, keeping in mind that MT is not yet adept at accurately reading context. To demonstrate, firstly, the subject for general people and the object of the sentence were inserted into the clause in the SS. Then, ‘the answer to the second question [the underlined part (1)]’ was deleted. Below is the MT output from the revised SS:

> Based on the above, I believe that (3) people should know (4) the universal language.

By clarifying essential elements of the SS to reduce ambiguity, the MT output was syntactically and semantically clearer [the underlined parts (3) and (4)]. Thus, the meaning of SSs should be unambiguously conveyed by paying attention to all necessary sentence elements. The clear elements in a SS would then be reflected in the MT output, leading to improvements in its readability.

Below is an example which demonstrated how a subject ought to be set to produce more simplified or higher-quality SSs:

> (1) Thanks to the existence of a common language, it is now possible to experience a variety of cultures.

To generate a more simplified or an advanced-level SS, firstly the SL expression corresponding to the above underlined part (1) should be set as the subject in the revised SS. Secondly, it is recommended to consider what sentence elements ought to be logically placed along with the subject.

> (2) The existence of a common language allows people to be exposed to a variety of cultures.

The examples above showed that appropriate subject setting [underlined part (2)] and placement of other sentence elements can create simpler TSs, displaying greater TL proficiency.

5.3.2. Note Semantic Properties of Lexical Items to Solve Factor 2 (Rule 5 in Table 2)

The second most common issue was semantic inappropriateness (Factor 2), demonstrated by the sentence below.

> Secondly, being able to speak multiple languages can be useful as a (1) weapon[7] when working in a company.

There was no syntactic discrepancy in the MT output, but rather a discrepancy related to meaning. Specifically, the underlined part (1) 'weapon' sounded strange in the TL context. In the SL, 'weapon' is commonly used metaphorically in a positive way due to its equivalent meaning as 'a useful skill', whereas it generally contains aggressive/offensive connotations in the TL. NMT could not understand the metaphorical use of the term in the particular context and translated it directly, which did not reflect the SS meaning in the TL context. Changing 'weapon' to an alternative SL expression expressed the author's true intention as follows:

> Secondly, being able to speak multiple languages can be (2) a useful skill when working in a company.

[7] Fixing the word 'weapon' would require a good understanding of the English language. It may be easier for the writer to fix the MT output in post-editing rather than trying to identify potential ambiguities in the Japanese during pre-editing. As can be seen from this case, a combination of pre- and post-editing are recommended for the most effective NMT use.

The solution for this kind of issue in TS is to refrain from using SL vocabulary metaphorically as each language has different connotations for certain words and expressions. In the event that there are no equivalent TL expressions, the vocabulary used in SSs should be carefully selected [underlined parts (2)]. It is vital to note the literal interpretations of such expressions and select appropriate SL lexical items.

Semantic inappropriateness was also displayed when expressions unique to the SL, such as onomatopoeic and mimetic words, were recognized as foreign words by NMT and output phonetically as Romanized characters. As a solution, it is recommended to insert more intelligible information such as a short SL-written description in parentheses, since there is often no TL equivalent to such words. Moreover, MT users need to ensure the correct usage of Kanji characters. Certain cases involving the wrong Kanji in a SS resulted in a TS that was semantically incongruous. The selection of SL vocabulary with more restricted meanings contributed to the elimination of translations displaying the aforesaid issues.

5.3.3. Simplify Sentences to Solve Factor 3 (Rule 1 in Table 2)

The third highest number of problems in SSs was related to sentence ambiguity and redundancy (Factor 3), illustrated by the following example:

> The second reason is that AI is evolving at such a fast pace that the accuracy of translation is gradually increasing, (1) eliminating the need for people to learn and use English on their own.

Although it had no major syntactic problems, the readability of this translation could be improved. All the ideas expressed in each part of the sentence were distinct and could be inferred from each other, however, there was potential ambiguity regarding what exactly was ‘eliminating the need for people to use English on their own’. The reason for this was that the SS had the issue of redundancy. More specifically, the content corresponding to the subject of the subsequent part overlapped with the content written in the previous part, resulting in a long and complex sentence. Such a structure was semantically ambiguous and made it difficult for the reader to grasp the meaning of the resulting TS. Vocabulary ending in -ing [underlined part (1)] should be avoided with MT use, as it leads to ambiguity in the resulting sentence (Pym, 1990)[8]. The analysts attempted to simplify the sentence without departing from the author's true intention (i.e., a literal interpretation) by: (i) separating the section that ought to be in a subsequent sentence from the previous part of the sentence, (ii) expressing the overlapping section as the indicative 'this' in the subsequent sentence [reflected in underlined part (2), below], and (iii) rephrasing the definitive expression in the subsequent sentence into 'may eliminate' [underlined part (3)] all align with the ST context. The resulting NMT output more closely reflected the author's intent:

> The second reason is that AI is evolving at such a fast pace that translation accuracy is gradually increasing. (2) This (3) may eliminate the need for people to learn and use English.

Readability improved in the MT output along with the pre-edited SS. Based on the example above, avoiding long and complex SSs, or simplifying SSs is crucial to produce better-quality TSs.

Closely related to the above is a case of sentence ambiguity [3B in Table 1], demonstrated below:

> For example, in English, the subject and verb, (1) which are the most important parts of a sentence, come first, (2) which expresses the national character of Americans, (3) who tend to express their feelings frankly.

The use of many relative pronouns made it difficult to understand the meaning of the TS. More specifically, three relative pronouns [underlined parts (1), (2), and (3)] were used in one sentence, making it difficult to grasp the relationship between what was modifying and what was being modified. This SS had a long and complex structure using punctuation marks in an attempt to convey multiple ideas in a single sentence. To reduce this ambiguity, the SS was separated based on breaks in meaning. Along with this, 'This characteristic [underlined part (4)]' was added as the subject of the subsequent sentence to ensure a smooth connection with what came before. The revised SS was input into NMT and the following TS was output:

[8] Pym (1990) developed pre-editing rules which can apply to any language pair; however, they were not created with NMT in mind.

> For example, in English, the subject and verb, which are the most important parts of a sentence, come first. (4) This characteristic expresses the national character of Americans, who tend to express their feelings frankly.

The MT output of the revised SS delivered one idea per sentence and its meaning was much clearer. Pym (1990) recommended using no more than 20 words per sentence to make up a MT translation. In the above MT output, the number of words was 19 in the first sentence and 15 in the second sentence, whereas the original MT translation contained a single sentence of 33 words. Simplifying sentences in STs is crucial to make the meaning of MT output easier to grasp. As shown in the above two examples, MT users or TL learners need to conform to the principle of 'one distinct idea per sentence' to decrease in semantic inappropriateness. Therefore, avoiding redundancy and simplifying SSs syntactically are essential during a pre-editing phase.

6. Conclusion and Limitations

This study identified 19 problematic factors in SSs written by university students which were categorized into six broad factors: These types of issue form the answer to RQ1. However, the issues related to TL linguistic elements (Factor 4 in Table 1) were judged to be difficult to note during a pre-editing phase. For "Japanese has no linguistic elements equivalent to determiners, prepositions, and singular and plural forms" (Tsuji, 2024, p. 11). Accordingly, most of these elements cannot be explicitly stated in SSs, and issues related to Factor 4 should be corrected once TSs have been output by NMTs.

In response to RQ2, six rules were presented, shown in Table 2. MT users or TL learners should be aware of the said rules to prevent the identified issues in TSs from arising. None of the rules are independent but interrelated. Pre-editing STs while keeping the rules in mind would reduce the occurrence of as many discrepancies or issues as possible in TTs resulting from NMT use in this translation context. With that being said, it should be noted that even if pre-editing is impeccably performed, MT translations may still display imperfections and issues. Post-editing is therefore also essential, especially for solving TL-specific grammatical and formatting issues (Factor 4 in Table 1), as referenced above.

While the present study offers important insights into pre-editing rules for effectively using NMT, there are some limitations in its design. The first limitation is that this study involved a limited number of participants who had a particular level of TL proficiency. Regarding the sample size, this study was conducted on a small scale during the COVID-19 pandemic, and it was not possible to extend the scope at that time. The second is that it focused on NMT output of an argumentative/opinion essay translated from Japanese to English, written by university students. Considering the above, the findings reported in this study might not be applicable to different genres of writing or TL learners with a different level of TL proficiency and SL writing aptitude in a different context. Accordingly, further research on a larger number of TL learners, adopting a longitudinal and multi-genre text design, is recommended. Furthermore, additional NMT tools need to be considered for developing more generalized rules in future studies.

Acknowledgments

This study was supported by a grant from Grant-in-Aid for Young Scientists (JSPS KAKENHI Grant number 21K13057). I would like to offer my sincerest thanks to Thomas, J. Wallestad of Marian University, and Chizuko Wallestad of Osaka Metropolitan University for the valuable analysis. Words cannot express my gratitude to Kiyo Okamoto of Osaka Metropolitan University and Benjamin Neil Smith of Kindai University for being a much-appreciated research assistant.

References

Bahdanau, D., Cho, KH., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the 3rd International Conference on Learning Representations, USA*, 1–15, https://doi.org/10.48550/arXiv.1409.0473

Bounaas, C., Zemni, B., Shehri, A. F., & Mimouna, Z. (2023). Effects of pre-editing operations on audiovisual translation using TRADOS: An experimental analysis of Saudi students' translations. *Texto Livre*, *16*(2), 1–15. http://doi.org/10.1590/1983-3652.2023.45539

Cheng, Y., Yue, S., Li, J., Deng, L., & Quan, Q. (2021). Errors of machine translation of terminology in the patent text from English into Chinese. *ASP Transactions on Computers*, *1*(1), 12–17. https://www.sciencegate.app/document/10.52810/tc.2021.100022

Farhana, B. C. D., Baharuddin, W. A. L., & Farmasari, S. (2023). Academic text quality improvement by English department students of University of Mataram: A study on pre-editing of Google neural machine translation. *Jurnal Ilmiah Profesi Pendidikan*, *8*(1), 247–254. http://doi.org/10.29303/jipp.v8i1.1186

Feifei, F., Rong, C., & Xiao, W. (2022). A study of pre-editing methods at the lexical level in the process of machine translation. *Arab World English Journal for Translation & Literary Studies*, *6*(2), 54–69. https://doi.org/10.31235/osf.io/3yrej

Fujii, R., Mita, M., Abe, K., Hanawa, K., Morishita, M., Suzuki, J., & Inui, K. (2021). Phenomenon-wise evaluation dataset towards analyzing robustness of machine translation models. *Natural Language Processing*, *28*(2), 450–478. https://doi.org/10.5715/jnlp.28.450

Hiraoka, Y., & Yamada, M. (2019). Pre-editing plus neural machine translation for subtitling: Effective pre-editing rules for subtitling of TED talks. *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, *Ireland*, *2,* 64–72. https://www.aclweb.org/anthology/W19-6710.pdf

Hyland, K., & Jiang, K. F. (2017). Is academic writing becoming more informal? *English for Specific Purposes*, *45*, 40–51.

Iwasaki, C. (2021). Analysis of issues expressed in first-year students' reports and the contribution of the writing center. *Kansai University journal of higher education*, *12*, 25–35. https://cir.nii.ac.jp/crid/1390009225309702528

Kim, Y., Tran. D. T., & Ney, H. (2019). When and why is document-level context useful in neural machine translation? *Proceedings of the Fourth Workshop on Discourse in Machine Translation*, *China*, 24–34. https://doi.org/10.18653/v1/D19-6503

Kokanova, E. S., Berendyaev, M. V., & Kulikov, N. Y. (2022). Pre-editing English news texts for machine translation into Russian. *Language Studies and Modern Humanities*, *4*(1), 25–30. https://doi.org/10.33910/2686-830X-2022-4-1-25-30

Miculicich, L., Ram, D., Pappas N., & Henderson, J. (2018). Document-level neural machine translation with hierarchical attention networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, *Belgium*, 2947–2954. http://doi.org/10.18653/v1/D18-1325

Mikami, A. (1975). *Mikami Akira ronbunsyū* [Collected academic papers written by Akira Mikami]. Kurosio Publishers.

Miyata, R., & Fujita, A. (2021). Understanding pre-editing for black-box neural machine translation. *Proceedings of the 16th Conference of the European Chapter of the Association* *for Computational Linguistics*, 1539–1550. https://doi.org/10.48550/arXiv.2102.02955

Oshima, Y. (2010). Types of problems in university students' essays: Toward a framework to promote collaborative learning to improve academic writing skills. *Kyoto University Researches in Higher Education*, *16*, 25–36. https://cir.nii.ac.jp/crid/1050001335709671040

Pym, P. J. (1990). Pre-editing and the use of simplified writing for MT: An engineer's experience of operating an MT system. In P., Mayorcas (Eds.), *Translating and the computer 10*. (pp. 80–96). Aslib.

Sánchez-Gijón, P., & Kenny, D. (2022). Selecting and preparing texts for machine translation: Pre-editing and writing for a global audience. In D. Kenny (Eds.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 81–104). Language Science Press.

Seretan, V., Bouillon, P., & Gerlach, J. (2014). A large-scale evaluation of pre-editing strategies for improving user-generated content translation. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, *Iceland*, 1793–1799. https://aclanthology.org/L14-1532/

Shih, C. (2021). How to empower machine-translation-to-web pre-editing from the perspective of Grice's cooperative maxims. *Theory and Practice in Language Studies*, *11*(12), 1554–1561. https://doi.org/10.17507/tpls.1112.07

Simonova, V., & Patiniotaki, E. (2022). Pre-editing for the translation of life-science texts from Russian into English via Google Translate. *Proceedings of New Trends in Translation and Technology 2022*, *Greece*, 259–265. https://www.researchgate.net/profile/Abdelalah-Alsolami/publication/371681915

Sutskever, I., Vinyals, O., & Le, V. Q. (2014). Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, *Canada*, *2*, 3104–3112. https://doi.org/10.48550/arXiv.1409.3215

Taufik, A. (2020). Pre-editing of Google neural machine translation. *Journal of English Language & Culture*, *10*(2), 64–74. http://doi.org/10.30813/jelc.v10i2.2137

Tsuji, K. (2021). Developing and evaluating a scoring rubric for argumentative essays: A module-based approach. *Urban Scope*, *12*, 1–13. https://urbanscope.lit.osaka-cu.ac.jp/journal/pdf/vol012/01-tsuji.pdf

Tsuji, K. (2022). Bogo parafureizingu no kyōikuteki kōka ni kansuru chōsa: Kikaihonyaku no seigengengo ni chakumoku shite [The effect of L1 paraphrasing on L2 writing: Focusing on pre-editing activity for machine translation], *Studies in the Humanities*, *73*, 33–49. https://doi.org/10.24544/ocu.20220416-008

Tsuji, K. (2024). Identifying MT errors for higher-quality target language writing. *International Journal of Translation, Interpretation, and Applied Linguistics*, *6*(1), 1–17. http://doi.org/10.4018/IJTIAL.335899

Tuzcu, A. (2021). The impact of Google Translate on creativity in writing activities. *Language Education and Technology*, *1*(1), 40–52. https://langedutech.com/letjournal/index.php/let/article/view/18/5

Zheng, Y., Peng, C., & Mu, Y. (2022). Designing controlled Chinese rules for MT pre-editing of product description text. *International Journal of Translation, Interpretation, and Applied Linguistics*, *4*(2), 1–13. http://doi.org/10.4018/IJTIAL.313919

Zhivotova, A. A., Berdonosov, V. D., & Redkolis, E. V. (2020). Improving the quality of scientific articles machine translation while writing original text. *Proceedings of 2020 International Multi-Conference on Industrial Engineering and Modern Technologies, Russia*, 1783–1786. https://doi.org/10.1109/FarEastCon50210.2020.9271