DOI: https://doi.org/10.22962/tnirvw.2024.14.1.003

**T&I REVIEW**
June 2024. Vol. 14. No. 1.

# Sentence Length and Translation:
## A Comparative Review of Human, NMT, and LLM Translations*

**Jin Yim****

*This paper aims to investigate the handling of long sentences in Korean-to-English translation by human translators, large language models (LLMs), and neural machine translation (NMT). Using reliable human translations in the business reports genre as a reference, the article analyzes human, NMT, and LLM translations through three analytic phases: a quantitative comparison, a qualitative analysis, and retranslation after pre-editing. The analysis found that the sentence length in the original source texts negatively correlates with translation quality in MT outputs, and that NMT's tendency to preserve the original sentence boundary often led to omissions and incomplete translations. However, retranslation after pre-editing effectively fixed both issues. The findings in this article contribute to broadening the literature on machine translation by highlighting the need to focus on the linguistic characteristics of MT, exploring different translation tendencies of the LLM and NMT models, and enhancing the representativeness of test corpora.*

**Keywords: Generative AI, Korean-English machine translation, sentence length, Google Translate, ChatGPT**

---

* This article is based on the author's manuscript included in 2023 Fall Proceedings of the Korean Association of Translation Studies.

** Ewha Womans University, Adjunct Lecturer

# 1. Introduction

Ever since machine translation (MT) came to the fore in academia, research focus has heavily been placed on evaluating MT performance by assessing MT outputs for error classification (Ragni and Nunes Vieira 2022). Evaluation is carried out automatically, manually, or using a combination of both. One of the most extensively used benchmarks for automatic evaluation is the BLEU score (Papineni et al. 2002), which compares machine-translated outcomes with human translations based on N-grams. However, there is barely any consensus among researchers about standardized evaluation criteria, as rightly argued by Lommel (2018). Different scholars use different labels to classify MT errors, often leading to confusion.

Although this is a necessary step towards overall performance improvement of MT, scholars such as S. B. Lee (2020: 88) and Lee and Choi (2023b: 78) have convincingly claimed that more MT studies should focus on specific linguistic features, which are expected to bring about more meaningful interdisciplinary discussions. In line with this argument, this article aims to investigate a specific linguistic feature of source texts (ST) and MT outcomes in a way that complements three under-represented areas in existing MT literature. First, sentence length is commonly pointed out as one of the most influential factors in translation quality, but this specific feature has rarely been the main theme of research. Second, the existing MT literature has overwhelmingly tilted towards neural machine translation, as large language models (LLMs) such as ChatGPT, only started their service in late 2022. Although scholars have begun exploring the differences in translation outputs from NMT and LLM models, this area remains widely uncharted. Third, despite the rising translation demand in the business reporting genre (Yim 2019: 139), it has been underexplored.

Against this backdrop, this article seeks to investigate how human, NMT, and LLM translators handle sentence boundaries when translating lengthy Korean ST sentences in the corporate reporting genre into English. More specifically, this paper aims to address the following research questions:

1. Does the length of ST sentences affect translation outputs from NMT and LLM models alike in the corporate reporting genre?

2. When translating long sentences in the corporate reporting genre from Korean to English, how do human, NMT, and LLM translators handle sentence boundaries? Are there differences in their tendency to preserve the structure of original sentences or split sentences?
3. If the length of ST sentences is manually reduced via pre-editing (e.g., by breaking it up into multiple sentences), does it improve the quality of both NMT and LLM outcomes?

To answer these questions, I compiled four corpora consisting of Korean ST, human translation (HT), Google Translate outcomes (MT1), and ChatGPT outcomes prompted by the author (MT2) and carried out a comparative analysis. Such an approach is expected to reveal how human translators and two different MTs handle long sentences and enhance our understanding of the impact of sentence length on MT outputs. This could eventually provide practical guidance to MT users and researchers.

## 2. Literature Review

MT performance has improved substantially since its emergence in the 1950s and the advent of artificial intelligence (AI), deep learning, and neural machine translation (NMT) (Castilho et al. 2017). Nevertheless, numerous empirical studies report unsatisfactory MT results in terms of human-parity quality, suggesting the need for human intervention for quality improvement. This presents abundant research opportunities for researchers in diverse disciplines such as computing, linguistics, and translation. This section seeks to explain why a specific linguistic feature such as sentence length deserves more attention by exploring the existing MT literature relevant to the data analyzed in this study.

### 2.1. Sentence Length in MT

An extensive body of studies has been dedicated to quality assessment via text

analyses across different language pairs, text genres, and engines. Lee and Cha (2019) pointed out that MT quality of the Korean-English language pair is still low due to linguistic differences between the two languages as well as the lack of parallel data. Hence, they argue, more empirical studies should be conducted using various text types in this language pair. This was also echoed by other scholars who tried to fill the gap by exploring diverse texts in different genres, such as legal texts, including statutes (Lee and Choi 2022, 2023a, 2023b), legal contracts (J. Lee 2022), and patents (Choi and Lee 2017); news articles (Lee and Cha 2019; C. S. Lee 2020); literary fiction (C. S. Lee 2021, 2023); non-literary texts (Park 2017, 2018); interview scripts (Lee and Cha 2023); and Korean proverbs (Kim 2018).

Empirical results from the studies mentioned above reveal a broad range of syntactic and semantic errors, which are labeled differently by various researchers. Given that this article aims to investigate the impact of a particular ST feature on MT outputs, it seems reasonable to narrow down the focus to the studies that explore ST influence in MT outputs. Several scholars mention the linkage between MT errors and ST-related elements, e.g., long and complex sentences (Park 2017, 2018; Lee and Cha 2019; J. Lee 2022; Lee and Choi 2023a); idioms (Lee and Cha 2022); dual meanings (Lee and Cha 2022); compound nouns (Park 2018), metaphors (Park 2018); text difficulty (Lee and Choi 2023b); out-of-vocabulary items (J. Lee 2022), and case markers (Kim 2018).

It is notable that long sentences are frequently mentioned as one of the problems causing translation errors in studies focusing on the evaluation of MT quality (Park 2018; Lee and Cha 2019; J. Lee 2022), but this topic has not been subject to comprehensive scholarly investigation on its own merit. Nevertheless, it is worth exploring a few works that delved into this feature more deeply (e.g., Park 2017, 2018; Lee and Cha 2019; Lee and Choi 2023a, 2023b).

Lee and Choi (2023a) compiled a corpus of 180 ST segments to compare three MT outputs in terms of overall translation quality and found a negative correlation between the two variables in all three MTs. In six groups of ST segments divided by length, translation quality decreased in longer-sentence segments. Among different NMT engines, Otran—customized NMT—handled long segments better than the other two. The negative correlation between sentence length and translation quality was

reconfirmed by the authors' subsequent study as well. In Lee and Choi (2023b), based on Korean-English Google Translate outputs of the 2,342-word ST corpus, they analyzed the correlation between ST difficulty and MT quality. To assess ST difficulty, the authors adopted two linguistic aspects—syntactic complexity (ST sentence length) and lexical and terminological difficulty. According to the analysis, translation quality did have a negative correlation with ST sentence length, while lexical and terminological difficulty had no statistically significant impact. It was the impact of ST length that made the correlation between text difficulty and translation quality significant. The findings suggest two implications. First, ST sentence length has a more significant impact on MT quality. Second, ST sentence length has more to do with text difficulty compared to lexical and terminological difficulty.

The relationship between longer Korean source sentences and English MT errors was also confirmed by Lee and Cha (2019) and Park (2017, 2018). What distinguished these three studies from others is their pre-editing approach. They identified MT errors related to long ST sentences and tried to improve the outcomes by pre-editing some of the problematic sentences to make them shorter (Park 2017: 165-166, 2018: 164; Lee and Cha 2019: 244). In these pre-editing attempts, long ST sentences were split into two or more sentences and became shorter. When the pre-edited, shorter sentences were machine-translated again, translation quality improved substantially. Although the authors effectively hinted at the impact of shorter ST sentence length on reducing translation errors, they chose not to dig deeper because their main research focus was placed on overall MT translation quality or error classification.

Arguably, the existing literature review suggests that long ST sentences are more likely prone to MT errors and thus deserve more focused research attention.

## 2.2. Translating Long Sentences and Sentence Boundaries

As argued in the previous section, longer sentences tend to cause translation errors. According to S. Lee and Y. Choi (2019: 174), sentence length is appropriate to measure text complexity because longer sentences tend to include multiple sentences or additional elements. This evidently makes text more difficult to read or understand. Sentence length

is a single factor that is used most frequently to assess text readability in various readability formulas (Choi 2013: 55). Mean sentence length is widely used in text difficulty formulas in English, e.g., the Dale-Chall readability formula (Dale and Chall 1948), the Flesch reading ease formula (Flesch 1948), the Flesch Kincaid (Kincaid et al. 1975), Lexile measures (Smith et al. 1989), and Coh-Metrix (Graesser et al. 2004). Although there is no standardized text difficulty formula yet in Korean, sentence length is included to assess text difficulty by various scholars such as Koo (2011, 2013), Chang (2012), S. Lee and Y. Choi (2019), S. H. Lee (2020) and quite a few others. Sentence length in Korean refers to the number of syllables, tokens, or word segments in each sentence. The mean sentence length is acquired by computing the average syllables and word segments included in each sentence (S. Lee and Y. Choi 2019). On the other hand, scholars such as Koo (2011, 2013) and S. H. Lee (2020) used the total number of word segments divided by the number of sentences to compute the mean sentence length.

In translation literature, sentence length is related to different translation outcomes in terms of sentence boundaries, leading to distinctive features of TT not only in MT, but also in HT. Translators who deal with a longer, difficult sentence may have two options in terms of sentence formation: preserve the original sentence boundary or alter the boundary. According to Bisiada (2013), translators tend to split long sentences rather than join multiple sentences into one. This relates to some of the universal features observed in translated texts called simplification (Blum-Kulka and Levenston 1983) and explicitation (Baker 1996). When a long and complex sentence is cut into multiple sentences, the sentence structure is simplified compared to the original one. During this process, the linkage between information sometimes disappears, so translators may compensate for the loss by adding a conjunction. This sometimes makes the translation more explicit than the original source text. Empirical evidence in some genres and language pairs support simplification and explicitation in translated texts compared to non-translated texts, although controversies exist about whether these features are "universal" or not.[1)]

It is worth noting that translation choices regarding simplification or explicitation

1) Refer to Yim (2019: 135-137) for empirical evidence of translation universals in Korean-English translation, and controversies around this concept.

usually depend on the type of texts or language pairs subject to translation. In general, some genres such as non-literary, informative texts allow translators to have more leeway, as compared to literary works and legal texts. In Korean-English translation, the mean sentence length of translated texts was found to be shorter than comparable non-translated texts in academic prose (Y. C. Lee 2019) and news articles (Goh and Lee 2016), suggesting that longer ST sentences could have been split into shorter ones. However, both studies quantitatively measured the average mean sentence length using the total number of tokens and sentences, without closely investigating whether long sentences were split or preserved.

For a more relevant look at what happens to sentence boundaries during the Korean-English translation of the corporate reporting genre, it is worth looking at Yim (2019), where the author compiled representative corpora and found longer sentences in TT than non-translated English texts presumably due to the influence of lengthy source sentences. A quantitative look did not support the hypothesis of simplification in translated texts. However, a qualitative analysis revealed that human translators frequently split ST sentences and made the length much shorter (Yim 2019: 146). Simply put, lengthy Korean sentences in this genre are often split by human translators, but it is still unknown whether MT preserves sentence boundaries or not.

The linkage between MT and human translation universals was already visited by some scholars on two fronts: One side looks at the impact of translationese in MT training data on MT performance (Graham et al. 2019; Zhang and Toral 2019), while the other investigates the trace of translation-like linguistic features from MT outcomes (Bizzoni et al. 2020; Luo and Li 2022). The former argument is relevant to this study. Zhang and Toral (2019) found that translationese inputs improve MT outcomes compared to non-translated inputs. The assumption behind this line of studies is that because all MT training data consist of parallel corpora including non-translated source texts and translated target texts, it is quite natural that MT handles translated inputs better. The assumption is partly supported by some of the previous studies mentioned in 2.1 such as Lee and Cha (2019), and Park (2017, 2018): When long ST sentences are split and thus simplified, it improved MT outcomes.

Centering on simplification and explicitation related to sentence length, this article

seeks an empirical approach to investigate how HT and two different MT outcomes deal with long sentences in this particular genre and language pair.

### 2.3. ChatGPT as a Translator

The rise of ChatGPT—a large-scale language model (LLM) developed by OpenAI—sparked instantly significant research attention in various disciplines. Researchers have shown great interest in its diverse linguistic features triggered by human prompts, and translation is one of the tasks ChatGPT is expected to improve greatly. Already, quite a few scholars such as Gao et al. (2023), Jiao et al. (2023), Lyu et al. (2023), and many others have reported a possible quality improvement of translation by ChatGPT compared to other commercial MTs such as Google Translate. However, their primary focus is placed on how to prompt ChatGPT to translate better with higher BLEU scores, rather than pointing to specific linguistic features ChatGPT can or can't handle better or worse. Particularly in the Korean and English language pair, ChatGPT's translation performance has been underexplored. A dominant proportion of MT literature has been allocated to NMT engines, which is understandable due to the LLMs' short history. Given LLMs' bright potential either as MT or an auxiliary tool assisting the work of translation, it is worth investigating its translation outcomes and relative performance to other commercial engines.

Given the research necessity described in this chapter, this article tries to broaden the existing MT literature by placing its research focus on how sentence boundaries are kept in the translation of the corporate reporting genre when translated by humans, Google Translate and ChatGPT.

## 3. Research Method

To find the answers to the research questions presented in chapter 1, this article adopted a three-phase analysis procedure. For this, a representative body of ST and HT

corpora was compiled, and MT outputs were produced as described in this chapter.

## 3.1. Corpus Compilation and Preparation

Aiming to ensure the quality of human translation as well as ST representativeness in this genre, the ST and HT corpora were collected from the forewords in Korean and English sustainability reports of the companies on the stock index of 30 Korean companies that is designed by Korea Exchange to represent the overall Korean stock market (Korea Exchange n.d.). As of June 2023, 29 out of the 30 companies had sustainability reports available both in Korean and English. They were deemed worthy of research because of their translation quality (see Appendix 1 for the list of 29 companies).

Table 1 shows the size of corpora to be analyzed in this article, based on word-level tokenization. Each corpus consists of 700 segments, each of which was manually aligned. Although each segment basically contains a single sentence, few of them may contain more than one sentence because a sentence could be split or joined during translation.

**Table 1. Size of Four Corpora**

| Corpus | No. of words | No. of segments |
|---|---|---|
| ST (source texts) | 11,865 | 700 |
| HT (human translation) | 18,342 | 700 |
| MT1(Google Translate) | 18,163 | 700 |
| MT2 (ChatGPT) | 17,434 | 700 |

Two corpora of MT outputs were collected from Google Translate (MT1) and ChatGPT (MT2). The unit of translation in this article was an individual foreword containing between five and a dozen paragraphs in a sustainability report of a single company.

With regards to Google Translate, all translations in this article were retrieved on July 6 and 7, 2023. ChatGPT outcomes require human prompts for translation. Because how

to prompt ChatGPT to translate is reported to affect the quality of its translation output (Gao et al. 2023; Jiao et al. 2023), the author referred to the prompts in Jiao et al. (2023) and modified one of them slightly for the purpose of this study into: "Translate the text into English." Given the study's research objective of comparing ChatGPT's performance with free online Google Translate, no context information was given. When the prompts were made on July 6 and 7, 2023, the author prompted ChatGPT to offer the exact model information and received the following answer: Model: GPT-3.5 (ChatGPT), Version: 2021-09 (Knowledge cutoff: September 2021).

### 3.2. Analysis Methods

The analysis consists of three phases. In the first phase, the BLEU scores of MT1 and MT2 were computed at the segment level from Tilde (n.d.). Also, the length of each of the 700 segments was collected to investigate whether sentence lengths and BLEU scores are correlated in each corpus (MT1 and MT2). To study the correlation, statistical software jamovi (The jamovi project 2023) was used.

The second phase was a qualitative analysis of the longest 100 source segments and their translations (HT, MT1, MT2), examining how the three corpora handled the boundaries of long sentences.

The final phase of the analysis was to identify the issues found in the second phase and try to modify source sentences to be machine-translated, seeking potential improvement in MT outcomes.

## 4. Research Results

According to the procedures presented in 3.3, corpora were compiled and ST sentence lengths and BLEU scores were computed for each segment as illustrated in Figure 1.

## 4.1. Quantitative Results

For a statistical outlook for the correlation between ST sentence lengths and MT output quality, a linear correlation was examined based on the Pearson correlation coefficient. Although the datasets in MT1 and MT2 were not normally distributed according to the Shapiro-Wilk test (<0.01), the sample size was large enough (N=700 for each corpus) to assume normality. Table 2 shows that the length of source sentences is negatively correlated, albeit slightly, to BLEU scores in both MT1 and MT2.

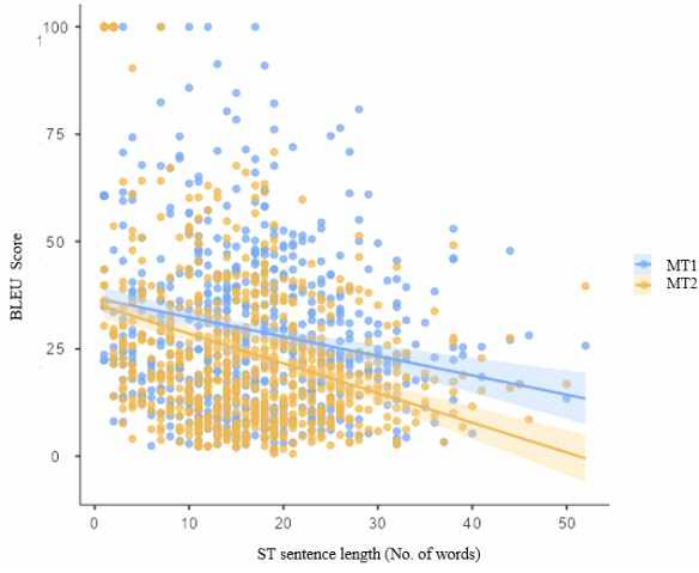


**Figure1. BLEU Scores Relative to ST Sentence Length**

**Table 2. Pearson Coefficient**

| | Corpus | | BLEU score |
|---|---|---|---|
| ST Length | MT1 | Pearson's r | -0.154*** |
| | | p-value | <.001 |
| | MT2 | Pearson's r | -0.257*** |
| | | p-value | <.001 |

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

For a detailed look at the correlation, I divided the datasets into seven groups, each containing 100 sentences depending on sentence length (Table 3) and examined the correlation again (Table 4) using the non-linear Spearman correlation matrix. As shown in Table 5, the results also confirmed a weak correlation between groups with different ST sentence lengths and their BLEU scores in MT1 and MT2 ($p < 0.05$). This indicates that longer sentence groups had lower BLEU scores in both MT outputs, which is statistically significant.

**Table 3: Mean BLEU by Sentence Length Group**

| Group | Sentence length range (No. of words) | MT1 | | MT2 | |
|---|---|---|---|---|---|
| | | Mean SL | Mean BLEU | Mean SL | Mean BLEU |
| 1 | 26-52 | 31.9 | 25.1 | 31.9 | 18.3 |
| 2 | 21-26 | 23.5 | 25.5 | 23.5 | 18.5 |
| 3 | 18-21 | 19.2 | 28.3 | 19.2 | 22.0 |
| 4 | 15-18 | 16.1 | 27.9 | 16.1 | 21.3 |
| 5 | 11-15 | 12.9 | 27.0 | 12.9 | 20.6 |
| 6 | 7-11 | 9.2 | 30.4 | 9.2 | 21.5 |
| 7 | 1-7 | 3.4 | 36.7 | 3.4 | 39.2 |

**Table 4: Spearman coefficient**

| | Corpus | | BLEU score |
|---|---|---|---|
| Sentence Length Group | MT1 | Spearman's Rho | 0.077* |
| | | p-value | 0.042 |
| | MT2 | Spearman's Rho | 0.106** |
| | | p-value | 0.005 |

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

## 4.2. Qualitative Results

For a more in-depth look at how human translators and two MT outputs handled sentence boundaries when translating long sentences, a qualitative analysis was conducted on Group 1 sentences (100 longest sentences from the ST corpus containing

26 to 52 words in Table 3) and their HT, MT1, and MT2. A manual analysis of how the three corpora translated long sentences—whether they preserved ST sentence boundaries or not—revealed differences across corpora, as shown in Table 5. It was impossible to run a *chi*-square test as the number of cells with the observation frequency of less than five exceeded 20% of the total cells. While MT1 tended to preserve the original sentence boundaries, HT and MT2 aggressively modified the sentence boundaries mainly by splitting the original sentences into multiple ones.

**Table 5: Sentence Boundary by Corpus**

| Translation choices regarding sentence boundaries by corpus | | | | |
|---|---|---|---|---|
| Corpus | Preserve | Merge | Split | Total |
| HT | 49 | 3 | 48 | 100 |
| MT1 | 77 | 1 | 22 | 100 |
| MT2 | 43 | 3 | 54 | 100 |
| Total | 169 | 7 | 124 | 300 |

There was one particularly notable point to make in this phase of analysis. By investigating the sentences that were split by HT and MT2 but preserved by MT1, I found MT1's tendency to preserve the boundary of long ST sentences occasionally led to errors, leaving incomplete sentences at the end of the segment as demonstrated below in Example 1.

Example 1: Incomplete Translation in MT1

[ST]
SK이노베이션은 현재 실행 중인 저탄소 중심의 Green Operation과 탄소 감축을 위한 투자 및 기술 개발을 더욱 가속화하겠으며, 실질적 탄소 감축 성과를 바탕으로 매년 Net Zero의 진척도를 투명하게 공개하고, 이해관계자와 지속적으로 소통해 나가겠습니다.

[HT]
SK innovation will further accelerate investment and technology development for low carbon green operation and carbon reduction that the company is currently

implementing. Based on our actual performance in carbon reduction, we will demonstrate transparency in disclosing the progress of Net Zero every year, and keep in touch with our stakeholders.

[MT1]
SK innovation will further accelerate the current low-carbon-centered Green Operation and investment and technology development for carbon reduction. I'll go out.

[MT2]
SK Innovation will further accelerate low-carbon-centered Green Operation and investments and technology development for carbon reduction. We will transparently disclose the progress of Net Zero annually based on actual carbon reduction achievements and maintain ongoing communication with stakeholders.

There were 18 cases of a similar error among 100 MT1 segments. However, this kind of error was not found in MT2, which tended to break sentences in a similar way to HT.

The analysis also revealed that MT2 sometimes omitted a large chunk of information when handling translation of long sentences as seen in Example 2. Also notable in this example was that HT split a single ST sentence into two or more sentences, while MT1 kept the original sentence boundary.

Example 2: Omissions in MT2

[ST]
기업의 이러한 지속가능경영 활동들이 보다 큰 임팩트와 성과를 창출하기 위해서는 여러 이해관계자들과 소통하고 협력하는 것이 매우 중요하기에, 삼성전자는 올해 초 글로벌 주요 이니셔티브인 유엔글로벌콤팩트(UN Global Compact)에 가입했습니다.

[HT]
To achieve meaningful progress and have material impact, we believe it is critical to communicate and work together with all stakeholders. In this context, we joined the United Nations Global Compact, an initiative to encourage businesses to adopt sustainable policies, earlier this year.

[MT1]
As effective communication and collaboration with multiple stakeholders are crucial for generating greater impact and results from these sustainable management activities, Samsung Electronics joined the UN Global Compact, a major global initiative earlier this year.

[MT2]
In order to generate greater impact and outcomes from these sustainable management activities, it is crucial to communicate and collaborate with multiple stakeholders. (omission)

Another notable aspect found in this qualitative look was MT2's problem in translating proper nouns (Example 3). When it was prompted to translate a CEO's name, it returned the name of a world-famous Korean pianist although the two share only a same family name, which is one of the most commonly used family names in Korea. The author reran this input on the date of July 25, 2023, but the result remained the same.

Example 3: Proper Noun Error in MT2

[ST]
LG전자 CEO 사장 조주완

[HT]
William (Joo-Wan) Cho, CEO & President of LG Electronics

[MT1]
Joo-Wan Cho, President and CEO of LG Electronics

[MT2]
Jo Seong-jin, CEO of LG Electronics

## 4.3. Pre-editing and Retranslation

In this section, a pre-editing process was performed, making only the minimally

necessary modifications to ST sentences for simplification and explicitation, which was then followed by retranslation. The ST segments subject to the pre-editing and retranslation were those that resulted in the two major errors in the previous section: incomplete sentence strings attached at the end of the segment in MT1 (Example 1) and major omissions in MT2 (Example 2). After choosing 10 segments for each of the two errors, the author split the original sentence into multiple sentences. Then the pre-edited segments were retranslated to be compared with the initial outcomes.

Example 4 shows how a single original sentence with a complex structure was pre-edited to be two simple-structure sentences. It was done by replacing the connective ending ("hagie") with a closing inflection ("habnida") and a period (simplification). Also, when necessary, a conjunction was added for the effect of making ST more explicit. In this case, a conjunction ("ttalaseo") was added to compensate for the meaning loss incurred from the removal of the connective ending.

Example 4: Pre-edited ST with a Connective

[ST]
기업의 이러한 지속가능경영 활동들이 보다 큰 임팩트와 성과를 창출하기 위해서는 여러 이해관계자들과 소통하고 협력하는 것이 매우 중요하기에, 삼성전자는 올해 초 글로벌 주요 이니셔티브인 유엔글로벌콤팩트 (UN Global Compact)에 가입했습니다.

[Pre-edited ST]
기업의 이러한 지속가능경영 활동들이 보다 큰 임팩트와 성과를 창출하기 위해서는 여러 이해관계자들과 소통하고 협력하는 것이 매우 중요합니다. 따라서 삼성전자는 올해 초 글로벌 주요 이니셔티브인 유엔글로벌콤팩트 (UN Global Compact)에 가입했습니다.

Pre-edited segments were retranslated by MT1 and MT2, respectively. The results were manually investigated. According to the result, there was an improvement in mean BLEU scores in both groups. The mean BLEU score for MT1 for the 10 problematic sentences was 24.0, but it only slightly rose to 25.2. For MT2, the average BLEU climbed from 8.8 to 17.0. The increase was larger in MT2 presumably because the pre-edited

input could fix major omission errors and thus substantially increased BLEU scores.

A qualitative look revealed outcome improvement in each group: Pre-edited ST successfully removed incomplete translation in MT1 and omission issues in MT2 discussed in 4.2. The findings were confirmed in the long sentences that were split (simplification) with or without a conjunction (explicitation): sentence split without a conjunction in MT1 (Example 5) and MT2 (Example 6); sentence split with an added conjunction in MT1 (Example 7).

Example 5: *Improved MT1 Outcome after Pre-editing (Simplification)*

[ST]
특히, 제조 혁신의 중심인 3D프린팅, 폐자원을 활용한 자원재순환, 그리고 Digital transformation은 이미 일부 성과가 도출되고 있는데, 두산에너빌리티의 방향성과도 부합하는 이들 신사업은 회사의 지속가능성을 담보하는 비즈니스로 계속 성장할 것입니다.

[HT] *ST split into two sentences*
We are pushing ahead with businesses such as 3D printing, which is key to achieving manufacturing innovation, resource recycling using waste resources and digital transformation. These new businesses, which are in alignment with Doosan Enerbility's future plans, are forecast to continuously grow as businesses that will guarantee the company's sustainability.

[Pre-edited ST] *ST simplified*
특히, 제조 혁신의 중심인 3D프린팅, 폐자원을 활용한 자원재순환, 그리고 Digital transformation은 이미 일부 성과가 도출되고 있습니다. 두산에너빌리티의 방향성과도 부합하는 이들 신사업은 회사의 지속가능성을 담보하는 비즈니스로 계속 성장할 것입니다.

[MT1 before pre-editing] *Preserve, incomplete translation*
In particular, 3D printing, which is the center of manufacturing innovation, resource recycling using waste resources, and digital transformation are already producing some results. will continue to grow.

[MT1 after pre-editing] *Split, improved outcome*

In particular, 3D printing, which is the center of manufacturing innovation, resource recycling using waste resources, and digital transformation are already producing some results. These new businesses, which are in line with Doosan Enerbility's direction, will continue to grow into businesses that ensure the company's sustainability.

Example 6: *Improved MT2 Outcome after Pre-editing (Simplification)*

[ST]
그러나 삼성전자는 지속가능하고도 풍요로운 환경과 사회를 만들어 나가는데 기여하는 것이 결국 기업의 경쟁력 강화와 지속적인 성장에 이르는 길임을 깊이 새기고 더 나은 미래를 향한 길을 꾸준히 만들어 나가겠습니다.

[HT] *Preserve*
However, we firmly believe that our competitiveness and sustainable growth goes hand in hand with harnessing our technological solutions to preserve the environment and contribute to building a better world.

[Pre-edited ST] *ST simplified and explicitated*
그러나 삼성전자는 지속가능하고도 풍요로운 환경과 사회를 만들어 나가는데 기여하는 것이 결국 기업의 경쟁력 강화와 지속적인 성장에 이르는 길임을 깊이 새기겠습니다. 삼성전자는 더 나은 미래를 향한 길을 꾸준히 만들어 나가겠습니다.

[MT2 before pre-editing] *Preserve, incomplete translation*
However, Samsung Electronics deeply believes that contributing to creating a sustainable and prosperous environment and society ultimately leads to strengthening the company's competitiveness and ensuring continuous growth. (omission)

[MT2 after pre-editing] *Split, improved outcome*
However, Samsung Electronics firmly believes that contributing to creating a sustainable and prosperous environment and society ultimately leads to enhancing the company's competitiveness and ensuring continuous growth. Samsung Electronics will continue to pave the way towards a better future consistently.

Example 7: *Improved MT1 Outcome after Pre-editing (Simplification and Explicitation)*

[ST]

우리는 혁신 공정 도입과 친환경 연료 · 원료 전환을 통해서 탄소 배출을 직접 감축(Reduce)하고, 재생에너지 사용을 확대해 간접 감축(Avoid)하고, 불가피한 탄소 배출은 상쇄(Compensate)하는 정교한 전략을 수립하고 실행 중입니다.

[HT] *ST split into two sentences*
We are deploying a full range of tactics to achieve our ambitious target. We reduce direct carbon emissions by introducing innovative processes and converting to green fuels and sustainable materials, avoid indirect emissions by expanding the use of renewable energy and compensate for all unavoidable emissions.

[Pre-edited ST] *ST simplified and explicitated*
우리는 혁신 공정 도입과 친환경 연료 · 원료 전환을 통해서 탄소 배출을 직접 감축(Reduce)하고 있습니다. 또한 재생에너지 사용을 확대해 간접 감축(Avoid)하고, 불가피한 탄소 배출은 상쇄(Compensate)하고 있습니다. 이렇듯 정교한 전략을 수립하고 실행 중입니다.

[MT1 before pre-editing] *Preserve, incomplete translation*
We establish sophisticated strategies to directly reduce carbon emissions through the introduction of innovative processes and conversion of eco-friendly fuels and raw materials, avoid indirect reductions by expanding the use of renewable energy, and offset unavoidable carbon emissions. Running.

[MT1 after pre-editing] *Split, improved outcome*
We are directly reducing carbon emissions through the introduction of innovative processes and the conversion of eco-friendly fuels and raw materials. In addition, we are avoiding indirect reductions (avoid) by expanding the use of renewable energy, and compensating unavoidable carbon emissions. We have formulated and implemented such a sophisticated strategy.

The results indicate that sentence length did count in both LLM and NMT outputs. Google Translate tended to preserve the original sentence boundary, while human translators and ChatGPT were more likely to split long sentences into two or more. Additionally, when a source sentence was split and thus shorter, this improved translation quality and effectively reduced omission errors in ChatGPT outcomes and incomplete translation errors in Google Translate.

## 5. Conclusion

This article aimed to shed some light on the handling of long sentences in translation, using representative, high-demand business reporting genre translations. Instead of general translation evaluation, this article placed its research focus on sentence length only. The research findings can be summarized as follows:

First, the sentence length of ST had a weak correlation with BLEU scores in MT1 and MT2. The weak correlation was also robust when each corpus was divided into seven groups depending on sentence length. This means that the negative relationship between ST sentence length and translation quality in NMT could also be valid in LLM translations.

Second, HT, MT1, and MT2 set sentence boundaries differently to translate longer sentences. HT and MT2 tended to split a single sentence into multiple sentences (simplification), while MT1 was more likely to stick to the sentence boundary of ST. Quite a few MT1 outcomes from long sentences had incomplete translation errors and odd word strings irrelevant to ST at the sentence's end. Although this issue was not observed in MT2 outcomes, it should not be interpreted as MT2's superior translation quality compared to MT1. MT2 revealed other issues, such as major omissions, when dealing with longer sentences.

Third, pre-editing of mistranslated ST sentences could effectively fix the problem and improve BLEU scores as well. Selected ST sentences were modified minimally: Long sentences were split to add the simplification feature to ST, and a conjunction was added when necessary for explicitation. Both NMT and LLM handled the pre-edited, simpler, more explicit sentences better than the original ST sentences, thereby making a noticeable improvement in reducing omission and incomplete translation.

These three findings, however, should not be generalized. Further research is needed to investigate more diverse genres, different MT engines, and different language pairs. Also, questions may be raised about whether ChatGPT and Google Translate represent overall commercial MT engines handling Korean into English or not. Also, this article used BLEU scores only because assessing the score's validity sits outside this article's research purpose. Regarding the data from ChatGPT, the author did not alter prompts for

better comparison with free online Google Translate, which might have undermined the translation quality. Offering context information on the genre and guidelines could possibly improve the quality of outcomes, which should be investigated by further research. Lastly, the focus of this article is limited to sentence length without a close look into the syntactic complexity of lengthy sentences. It is admissible that such a limitation results from the difficulties arising from classifying Korean sentences into simple, complex, and embedded sentences due to the innate nature of the agglutinative language.

Despite such shortcomings, the approach made by this article adds to the existing literature on MT outputs in three aspects. First, this article took a comparative perspective on LLM and NMT outputs and found different tendencies between the two models when handling sentence boundaries. This suggests the possibility that the findings in existing NMT studies may not be effective in translations generated by LLM models, thereby broadening the research horizon further into the testing of NMT findings in LLMs. Second, this article empirically showed the evidence that the pre-editing for adding two translation universals to ST led to improved outcomes in both NMT and LLM. This supports the quality improvement in NMT outputs when translationese is used in MT inputs (Zhang and Toral 2019). Third, this study seeks to show that the representativeness of test corpora is as important in MT as in HT research, because it is impossible to run tests under all available texts, MT models, genres, language pairs, and translation directions. It has become more important than ever in studying the rapidly evolving area of NMT and LLM translations. As rightly proposed by Luo and Li (2022: 21), a corpus-based approach to investigating linguistic features of MT eventually helps improve and develop MT systems.

## References


Baker, M. (1996). Corpus-based translation studies: the challenges that lie ahead. In Somers, H. (ed.), *Terminology, LSP and translation*. John Benjamins, 175-186.

Bisiada, M. (2013). From Hypotaxis to Parataxis: An Investigation of English-German Syntactic Convergence in Translation. Unpublished PhD dissertation, University of Manchester.

Bizzoni, Y., Juzek, T. S., España-Bonet, C., Chowdhury, K. D., Van Genabith, J. and Teich, E. (2020). How human is machine translationese? comparing human and machine translations of text and speech. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 280-290.

Blum-Kulka, S. and Levenston, E. A. (1983). Universals of lexical simplification. In Faerch, C. and G. Kasper (eds.), *Strategies in Interlanguage Communication*. London: Longman, 119-139.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J. and Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics* 108: 109-120.

Chang, M. (2012). Research on Text Difficulty Evaluation for Teaching Reading in Korean. Unpublished PhD dissertation, Korea University.

Choi, H. E. and Lee, J. (2017). A Study on the evaluation of Korean-English patent machine translation-focusing on KIPRIS K2E-PAT translation. *Interpretation and Translation* 19(1): 139-178.

Choi, M. (2013). Assessing Source Text Difficulty for Interpreter Education: With a Focus on Textual Factors of English Source Texts in English-Korean Consecutive Interpreting. Unpublished PhD dissertation, Ewha Womans University.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: instructions. *Educational Research Bulletin* 27(2): 37-54.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology* 32(3): 221-233.

Gao, Y., Wang, R. and Hou, F. (2023). How to design translation prompts for ChatGPT: an empirical study. *arXiv e-print arXiv-2304.*

Goh, G. Y. and Lee, Y. (2016). A corpus-based study of translation universals in English translations of Korean newspaper texts. *Cross-Cultural Studies* 45: 109-143.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z. (2004). Coh-Metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2): 193-202.

Graham, Y., Haddow, B. and Koehn, P. (2019). Translationese in machine translation Evaluation. *arXiv preprint arXiv:1906.09833*.

Jiao, W., Wang, W., Huang, J., Wang, X. and Tu, Z. (2023). Is ChatGPT a good translator? yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Kim, K. (2018). The influence of case markers on the machine-translation of Korean proverbs into English. *The Linguistic Association of Korea Journal* 26(3): 139-157.

Kincaid, J. P., Fishburne, Jr., Robert, P. R., Richard, L. C. and Chissom, B. S. (1975). *Derivation of*

*New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel* (RESEARCH BRANCH REPORT 8-75; Naval Technical Training). Defense Technical Information Center.

Koo, M. J. (2011). A Study of text difficulty analysis on the Korean reading materials. *Journal of Korean Language Education* 22(2): 27-48.

Koo, M. J. (2013). A Study on Measuring Text Difficulty for Korean Reading Education. Unpublished PhD dissertation, Catholic University of Korea.

Korea Exchange. (n.d.). KRX Series: KTOP30 (online) Retrieved from http://data.krx.co.kr/contents/MDC/EASY/visualController/MDCEASY500.cmd on 4 June 2024.

Lee, C. S. (2020). A study of lexical usage differences between human and machine translation in English translations of Korean newspaper editorials. *Interpretation and Translation* 22(1): 245-262.

Lee, C. S. (2021). Machine learning classification of literary translation samples by human and machine translators. *The Journal of Translation Studies* 22(1): 199-217.

Lee, C. S. (2023). A follow-up study of stylistic differences between human and machine translation with ChatGPT added in the mix. *The Journal of Translation Studies* 24(3): 539-561.

Lee, J. (2022). A Study on performance evaluation of an specialized machine translation engine in the legal domain: focusing on the Korean-to-English translation of legal contracts. *T&I Review* 12(1): 169-192.

Lee, J. H. and Cha, K. W. (2023). Human interpretation and machine translations based upon interviews with Director Joon-ho Bong. *Korean Journal of English Language and Linguistics* 23(March): 204-219.

Lee. J. and Choi, H. E. (2022). A case study of Korean-English machine translation of dual subject sentences: a comparison of statutory translations by Google Translate and Naver Papago. *T&I Review* 12(1): 211-241.

Lee, J. and Choi, H. E. (2023a). A case study on the evaluation of Korean-English legal translations by generic and custom neural machine translation engines. *Interpretation and Translation* 25(1): 75-98.

Lee, J. and Choi, H. E. (2023b). A case study on the quality of machine translation and the source text difficulty: a case of a Korean-English statutory translation by Google Translate. *Journal of Linguistic Studies* 28(1): 77-101.

Lee, J. H. and Cha, K. W. (2019). An analysis of Korean-English translation errors in Google Translate. *The Journal of Linguistics Science* 89: 221-257.

Lee, J. H. and Cha, K. W. (2022). A study of error types in Korean-English translation from Korean spoken language to Papago. *Modern English Education* 23(1): 56-65.

Lee, S. B. (2020). Review of literature on machine translation: based on arts and humanities journals covered in KCI Journal Database (from 2011 to early 2020). *Interpretation and Translation* 22(2): 75-104.

Lee, S. and Choi, Y. (2019). The development of a web-based automatic text complexity measurement program. *Journal of Korean Language Education* 30(2): 163-180.

Lee, S. H. (2020). A study on the difficulty distribution of texts for Korean education using the text difficulty measurement program. *The Studies of Korean Language and Literature* 68:

425-450.
Lee, Y. C. (2019). Spotting non-nativeness in L2 texts: a statistical approach to translationese. *Studies in English Language & Literature* 45(1): 367-388.
Lommel, A. (2018). Metrics for translation quality assessment: a case for standardising error typologies. In Moorkens, J., S. Castilho, F. Gaspari, and S. Doherty (eds.), *Translation Quality Assessment*. Springer Cham, 109-127.
Luo, J. and Li, D. (2022). Universals in machine translation? *International Journal of Corpus Linguistics* 27(1): 31-58.
Lyu, Q., Tan, J., Zapadka, M. E., Ponnatapura, J., Niu, C., Myers, K. J., Wang, G. and Whitlow, C. T. (2023). Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. *Visual Computing for Industry, Biomedicine, and Art* 6(1).
Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA, 311-318.
Park, O. (2017). Error analysis according to the typological characteristics of source text in Korean-English machine translation. *The Journal of Society for Humanities Studies in East Asia* 41: 155-183.
Park, O. (2018). Error analysis and criterion for correcting error of machine translation in terms of source language: based on the syntactic characteristics of non-literary texts. *Dongainmunhak* 44: 151-171.
Ragni, V. and Nunes Vieira, L. (2022). What has changed with neural machine translation? a critical review of human factors. *Perspectives: Studies in Translation Theory and Practice* 30(1): 137-158.
Smith, D. R., Stenner, A. J., Horabin, I. and Smith, M. (1989). The Lexile scale in theory and practice: final report for NIH grant ID-19448. International Reading Association.
The jamovi project. (2023). jamovi (Version 2.4) [Computer Software]. Retrieved from https://www.jamovi.org on 30 July 2023.
Tilde. (n.d.). Interactive BLEU [Computer Software]. Retrieved from https://www.letsmt.eu/Bleu.aspx on 10 July 2023.
Yim, J. (2019). Translation universals in translated CEO letters in sustainability reports. *The Journal of Translation Studies* 20(5): 131-162.
Yim, J. (2023). NMT vs. LLM - focused on long-sentence translation. *2023 Fall Proceedings of the Korea Association of Translation Studies*, 92-104.
Zhang, M. and Toral, A. (2019). The effect of translationese in machine translation test sets. *WMT 2019 - 4th Conference on Machine Translation, Proceedings of the Conference 1*, 73-81.

## Appendix

## List of 29 companies in corpora

AmorePacific, Doosan Enerbility, Emart, Hyundai Construction, Hyundai Mobis, Hyundai Motors, KAKAO, KB Financial Group, Kia, Korea Shipbuilding, LG Chem, LG Display, LG Electronics, Lotte Chemical, Mirae Asset Securities, NAVER, Netmarble, POSCO Holdings, Samsung C&T, Samsung Electronics, Samsung Electro-Mechanics, Samsung Fire & Marine, Samsung Life Insurance, Samsung SDI, Shinhan Financial Group, SK Hynix, SK Innovation, SK Telecom, Yuhan (29 firms that have published Korean and English sustainability reports as of July 2023)

This paper was received on 7 May 2024; revised on 5 June 2024; and accepted on 10 June 2024.

***Author's email address***

jy2812@gmail.com

***About the author***

Jin Yim is an adjunct lecturer at the Graduate School of Interpretation and Translation at Ewha Womans University. Her current research interests include sociological approaches to translation and interpreting practices, as well as corpus-based analyses of human and machine translation.