ReCALL (2021), 1–14
doi:10.1017/S0958344021000124

RESEARCH ARTICLE

An investigation of machine translation output quality and the influencing factors of source texts

Sangmin-Michelle Lee
Kyung Hee University, Republic of Korea (sangminlee@khu.ac.kr)

Abstract

The use of machine translation (MT) in the academic context has increased in recent years. Hence, language teachers have found it difficult to ignore MT, which has led to some concerns. Among the concerns, its accuracy has become a major factor that shapes language teachers' pedagogical decision to use MT in their language classrooms. Despite the urgency of the issue, studies on MT output quality in foreign language education remain scarce. Moreover, as MT is advancing every year, updated studies are imperative. Therefore, the present study investigated the quality of MT outputs (Google Translate) from Korean to English by comparing it with the English-translated texts of intermediate English as a foreign language students. The study also examined the factors within the source texts that affect the quality of MT outputs. Five trained evaluators examined multiple aspects of MT output samples ($N = 104$) and students' English texts ($N = 104$), including mechanics, vocabulary, grammar, and context. The results showed that both texts were equally comprehensible, but MT outperformed the students in most aspects under investigation. The study further found that only two factors in the source texts – punctuation and sentence complexity – influenced MT output quality, whereas lexical and grammatical accuracy, lexical diversity, and contextual understanding did not affect it. Based on the results, the study presents classroom implications for using MT for educational purposes.

Keywords: machine translation; quality; source text; influencing factors

1. Introduction

Recent years have seen an increase in the number of students using machine translation (MT) for diverse purposes in educational settings as well as in their daily life (Alhaisoni & Alhaysony, 2017; Briggs, 2018). They use MT to communicate with others with different language backgrounds, enhance their vocabulary, facilitate speaking, improve reading and writing, and complete their assignments. MT is fast, cost-effective, and convenient; thus, it attracts many foreign language (FL) learners (Alhaisoni & Alhaysony, 2017). Moreover, with the introduction of the Google Neural Machine Translation (NMT) in late 2016, MT quality and accuracy have significantly improved (Sun, 2017; Tsai, 2019). Advances in artificial intelligence and machine learning have been making the quality of MT more reliable, more accurate, and more human-like (Godwin-Jones, 2019). Although language teachers have been reluctant in allowing their students to use MT in their language classrooms because of issues surrounding ethics, overdependence, and reliability, MT is undeniably becoming more popular among FL students (Briggs, 2018).

Previous studies have conducted preliminary research and presented tentative conclusions about the effectiveness of using MT in that, despite its inaccuracies, it remains useful for FL education (Briggs, 2018; Garcia, 2016; O'Neill, 2016; Sun, 2017). However, the quality of MT

Cite this article: Lee, S-M. (2021). An investigation of machine translation output quality and the influencing factors of source texts. *ReCALL* FirstView, 1–14. https://doi.org/10.1017/S0958344021000124

outputs has been rarely investigated in FL education; the majority of studies examined MT accuracy based on users' perceptions, and only a few empirical studies have been conducted that directly assess the accuracy of MT (Lee, 2020). Moreover, these studies presented mixed results regarding MT's lexical and grammatical accuracy (Fredholm, 2015; Lee, 2020). Accordingly, teaching implications, such as whether MT would be more beneficial to beginner or advanced FL students, are also inconsistent (Briggs, 2018; Garcia & Pena, 2011; Kazemzadeh & Kashani, 2014; O'Neill, 2013; Tsai, 2019). Although MT has been widely used in FL education, we still have insufficient empirical evidence to firmly support whether language teachers adopt or reject it. Although numerous studies on MT accuracy were published in translation studies, translation studies frequently use paid MT software for professional translators; have different research foci (i.e. architecture of MT, tagging, or segmentation); or employ automatic evaluation methods to measure MT accuracy, such as bilingual evaluation understudy (BLEU) or metric for evaluation of translation with explicit ordering (METEOR) (Maruf & Haffari, 2018; Yang, Chen, Wang & Xu, 2018). Therefore, they are difficult to apply to FL education. Without understanding the quality of MT from the perspective of FL education, it would be impractical to determine whether to use it or how to incorporate it into the FL classroom to maximize student FL learning. Therefore, the current study aims to investigate the quality of MT outputs and the factors within the source texts (L1) that shape it from the perspective of FL education.

2. Literature review

Overall, the majority of studies on MT in FL education conducted over the last decade have reported positive results. Previous studies reported that students who used MT performed second language (L2) writing better and more effectively, with fewer errors in lexical and grammatical aspects, than those who did not use it (Garcia, 2016; Giannetti, 2016; Tsai, 2019). More specifically, Kol, Schcolnik and Spector-Cohen (2018) and O'Neill (2014) showed that students who used MT outperformed those who did not use it in terms of grammar, spelling, content, and overall comprehensibility in L2 writing. Their studies found that MT helped students use more words, longer sentences, and linguistic structures beyond their current levels. Fredholm (2015) also discovered that the MT group produced fewer errors in spelling and grammar (but more errors in syntax) than the non-MT group, and found in a later study (Fredholm, 2019) that the MT group produced texts with higher lexical diversity and density. Similarly, students significantly reduced lexical and grammatical errors in their revisions by using MT (Lee, 2020; Tsai, 2019), and, as a consequence, they found MT an effective pedagogical supplement to FL learning (Bahri & Mahadi, 2016; Murtisari, Widiningrum, Branata & Susanto, 2019; Stapleton & Kin, 2019). It also has been known that MT facilitated students' metacognitive knowledge toward L2 (Clifford, Mershel & Munné, 2013; Thue Vold, 2018) and provided a more comfortable L2 learning environment (Bahri & Mahadi, 2016; Niño, 2008).

The effectiveness of MT on FL learning and learner satisfaction can, in fact, depend on the MT output quality. A number of MT studies indirectly evaluated MT output quality based on users' perceptions of using MT. Earlier studies have revealed mixed results. On the one hand, students found MT to be helpful in improving the accuracy and fluency of L2 writing (Bahri & Mahadi, 2016; Niño, 2009). On the other hand, they showed concerns about the inaccuracy of MT, such as literal translation and lexical, syntactic, and discourse inaccuracies; thus, they perceived MT to be only somewhat useful or even not useful at all (Clifford *et al.*, 2013). Since the advances in NMT increased in 2016, it has been reported that Google Translate's accuracy has been greatly enhanced and has achieved good fluency (Sun, 2017). Accordingly, more recent studies have reported that MT has either outperformed English as a foreign language (EFL) students at least in certain aspects or has already surpassed the production level of many EFL students (Briggs, 2018; Tsai, 2019). Specifically, Briggs's (2018) study indicated that students trusted MT more than their

own English abilities. However, this does not mean that all users are satisfied with the quality of MT outputs; students still exhibit mixed impressions of or uncertainty regarding the accuracy of MT (Im, 2017; Lee, 2019). A discrepancy was also found between instructors' and students' perceptions of MT accuracy. Although Clifford *et al.* (2013) discovered that instructors were more doubtful about the accuracy and usefulness of MT than students, Jolley and Maimone (2015) revealed that 51% of the instructors considered MT to be accurate, and only 15.6% of students trusted its accuracy. A more recent study indicated that an increasing number of instructors have a positive attitude toward using MT in their language classrooms (Marinac & Barić, 2018). In addition, whereas several researchers suggested that MT would be more helpful to beginner or intermediate FL students (Briggs, 2018; Garcia, 2016; Garcia & Pena, 2011; Kazemzadeh & Kashani, 2014), others claimed that it is more appropriate for advanced students (Niño, 2009; O'Neill, 2013; Stapleton & Kin, 2019; Tsai, 2019).

Although there are still few studies, several empirical studies have investigated the accuracy of MT based on error analysis. Earlier studies showed that MT produced a number of errors, particularly in word choice, syntax, and discourse; the most common errors were found to be in word choice, followed by sentence structure (Groves & Mundt, 2015; Niño, 2008). White and Heidrich (2013), in their investigation of errors produced by MT (from English to German), found that sentence structure errors appeared most frequently, and verbal errors were the least frequent. They also discovered that phrasing, declension, word choice, and word order were dominant error subcategories in MT outputs. More recent studies, however, indicated that MT accuracy has significantly improved. In Stapleton and Kin's (2019) study, 12 teachers evaluated MT outputs and students' outputs (from Chinese to English) in terms of comprehensibility, vocabulary, and grammar, with the MT outputs achieving higher scores. Additionally, some of the lexical problems of the earlier iterations of MT have since been solved. According to Ducar and Schocket (2018), MT can now translate less commonly used words, idioms, misspelled words, and colloquial language quite well owing to the large databases used. O'Brien, Simard and Goulet (2018) even recommended that L2 learners write in L1 first, use MT, and edit the output to reduce cognitive load and maximize efficiency.

Prior studies have also found several factors that influence MT output quality. First, different language pairs can lead to different output quality and error types. According to Shadiev, Sun and Huang (2019), O'Brien *et al.* (2018), and Ruiz and Federico (2014), translation difficulty increases between distant language pairs. The results also vary depending on the tool (Groves & Mundt, 2015). For instance, BabelFish has been known to work better with Asian and minority languages, whereas Google Translate functioned better with Western languages (Aiken & Balan, 2011). However, since Google's introduction of the NMT method, Google Translate has been known to work best with almost all languages (Stapleton & Kin, 2019; Sun, 2017). The quality of the source text also determines the quality of MT outputs, including length, text difficulty, lexical and structural complexity, and punctuation (Clifford *et al.*, 2013; Jolley & Maimone, 2015; Shadiev *et al.*, 2019).

As an increasing number of L2 learners are using MT, it has become imperative to assess the quality of MT. Depending on the quality of MT, the effectiveness of L2 learning and strategies for using MT may vary. Despite its significance, this issue has been under-researched, and furthermore, the majority of previous studies assessed MT quality mainly based on user perception and were rarely engaged in evaluating actual MT outcomes (Lee, 2020). In addition, as the accuracy of MT is improving every year, more updated studies are needed.

Therefore, the present study evaluated the quality of MT outputs by comparing them with student-generated texts in various areas such as comprehensibility, vocabulary, grammar, and punctuation. This study also examined which factors in the source text influenced the MT output quality. Overall, this paper addresses the following questions:

1. Which of the two versions – students' own English translations or MT outputs – is more credible? In which areas does one outperform the other?
2. Which factors in the source texts (Korean) influence MT output (English) quality?

3. Method

3.1 Participants and procedures

The current study employed a quantitative approach (the evaluators' scores of the texts) as its primary method and a qualitative approach (their written reflections) as its secondary method. The text samples included three versions of essays of 104 college students in Korea. The students wrote essays on the topic "technology and language," first in Korean (first language; L1 texts), and then they translated them into English on their own (L2 texts). They used MS Word for the task and can therefore utilize the spell checker. Finally, they used MT to translate them into English again (MT texts) and revised the L2 text by comparing it with MT outputs. Google Translate was selected in this study because it was the students' most commonly used MT. The students' L1 was Korean, and their English writing proficiency was 4.1 on average (out of 6, intermediate) according to the TOEFL essay scoring scale. The task aimed to provide the students with opportunities to find mechanical, lexical, and grammatical errors in the L2 texts; improve the quality of the texts; and learn the target language from the errors. In other words, as Lee (2020) mentioned, the students were expected to notice the differences between their texts and MT texts, which would raise awareness about the features of the target language and provide the students with opportunities to ponder the differences, to correct their errors, and to improve the quality of the revised texts. The students' revised texts were excluded from the study because it did not serve the research purpose.

As translation continues to be one of the most commonly used methods of learning L2 writing, particularly in Asian countries, studies have confirmed that translation helps L2 learners with L2 writing in diverse ways (Cook, 2012; Kim, 2011; Kim & Yoon, 2014). However, writing and translation should be viewed separately because they require different mental processes in terms of language, and generation and organization of ideas and contents (Choi & Lee, 2006; Kim & Yoon, 2014). Although direct L2 writing and translation have distinct benefits, translation was selected in the current study for two reasons. First, from a pedagogical perspective, it provided opportunities to raise the students' linguistic awareness by comparing their own translation with MT outputs (Lee, 2020; Tsai, 2019). It could also give the students insight into effective methods of using MT during L2 writing. Second, from a research perspective, the outcomes of translation allowed the researchers to trace the process, evaluate the quality of MT, and locate the origin of errors in MT. As a result, the researchers were able to investigate MT output quality and the influencing factors of source texts.

3.2 Data analysis

To evaluate the texts, two sets of rubrics were developed based on literature: one for the L1 texts and the other for L2 texts translated by the students and via MT. Previous studies pointed out that MT's most common flaws are lexical and grammatical inaccuracy, ambiguity, lack of cultural understanding, contextual errors, and literal translations (Briggs, 2018; Ducar & Schocket, 2018; Stapleton & Kin, 2019). Based on these studies, the following 10 aspects were selected to evaluate the MT texts and the students' L2 texts: spelling, punctuation, grammar, vocabulary appropriateness, vocabulary level and diversity, comprehensibility, unambiguity, idiomatic expressions, authenticity, and contextual understanding (see Appendix A in the supplementary material). The unit of analysis remained flexible while dependent upon the aspect of evaluation. That is, word was a unit of analysis for vocabulary, sentence for grammar, and paragraph or an entire text for authenticity and comprehensibility. All these aspects were evaluated on a 5-point scale except for spelling and punctuation, which were assessed using a 2-point scale because errors in these aspects were simpler and appeared only in a narrow range of error frequencies.

With regard to source texts, studies have found that misspellings, incorrect grammar, long and complex sentences, vocabulary misuse, and ambiguity in L1 writing undermine the

comprehensibility and quality of MT outputs (Clifford *et al.*, 2013; Jolley & Maimone, 2015; Kim, 2019; Shadiev *et al.*, 2019). MT was also reported to produce more flawed outputs when translating texts with polysemous lexical items and pragmatic information and texts requiring a high level of cultural understanding (White & Heidrich, 2013). Park (2017) and Kim (2019), in particular, examined the MT English outputs translated from Korean language and pointed out that lengthy and complex sentences and culture-specific expressions led to inadequate translations of MT. Based on previous studies, the current study evaluated the L1 versions according to nine aspects: spelling, punctuation, grammar accuracy, sentence complexity, vocabulary appropriateness, vocabulary level and diversity, comprehensibility, unambiguity, and level of Korean contextual information included in the L1 text (see Appendix B in the supplementary material). Similar to L2 text evaluation, all aspects were scored on a 5-point scale, except for spelling and punctuation (2-point scale) in the L1 text evaluation.

Five trained EFL writing instructors (three Korean and two native English speakers) evaluated the students' L2 texts and MT outputs. Prior to evaluation, the instructors were briefed on the evaluation process, with detailed explanations about the rubric. They were not told that one set of the English versions was translated by Google until they finished their evaluation. Each rater evaluated the first five sample texts, checked their understanding of the rubric, and drew a consensus on the acceptable levels for vocabulary appropriateness and authenticity. The inter-rater reliability was calculated using intra-class correlation coefficient (ICC, Cronbach's alpha): 0.89 for the students' texts and 0.92 for the MT texts. Upon completing their assessments, the evaluators wrote one-page reflections on their general impressions of the texts and the idiosyncratic features of the MT texts and the students' L2 texts. The students' L1 texts were evaluated by two of the Korean writing instructors. They were asked to examine grammar more strictly than usual because, as a pro-drop language, pronouns are often missing in Korean language, which does not necessarily result in incorrect grammar but negatively influences the MT outputs (Kim, 2019). ICC for the inter-rater reliability turned out to be 0.80. No reflection was conducted in the L1 evaluation. In addition to the raters' evaluation, lexical density, sentence length, and readability level (Gunning Fog Index, GFI) were calculated using a text analyzer for a more thorough comparison between the MT texts and the students' L2 texts.

Multiple quantitative methods were employed for data analysis. First, descriptive statistics and *t*-test were used to find general data patterns and examine statistical significance between the versions. Next, regression analysis was conducted to discover which factors in the L1 texts influenced MT output accuracy. In addition, the evaluators' reflections were coded based on emerging themes.

4. Results

4.1 Quantitative analysis of the texts

The study found that the MT outputs and the students' texts were equally comprehensible but that MT obtained a higher total score (out of 44; $M = 28.38$) than the students ($M = 25.82$). The MT texts were scored higher than the students' L2 texts in all areas under investigation except for contextual understanding (Table 1). The *t*-test confirmed that MT was superior in spelling ($t = -4.209$), vocabulary appropriateness ($t = -3.946$), vocabulary level and diversity ($t = -3.404$), grammatical accuracy ($t = -3.427$), idiomatic expressions ($t = -3.109$), and authenticity ($t = -2.800$). Although MT obtained higher scores in comprehensibility and unambiguity, the differences were not statistically meaningful.

The text analysis results showed that the average word count of the MT outputs was slightly higher than that of the students' L2 texts. Sentence length was greater in the students' L2 texts, but lexical density and GFI were higher in the MT texts, which implied that the MT texts had higher text readability than the students' L2 texts. The *t*-test results showed that differences in sentence

Table 1. Results of the independent *t*-test

		Students' L2 texts		MT texts					
		M	*SD*	*M*	*SD*	*t*	*p*	*M* difference	*SE* difference
Mechanics	Spelling	1.81	.396	1.98	.138	−4.209	.000	−.173	.041
	Punctuation	1.68	.468	1.77	.423	−1.399	.163	−.087	.062
Lexico-grammatical accuracy	Vocabulary appropriateness	2.72	1.000	3.25	.932	−3.946	.000	−.529	.134
	Vocabulary level and diversity	2.96	.823	3.36	.847	−3.404	.001	−.394	.116
	Grammatical accuracy	2.55	1.190	3.08	1.031	−3.427	.001	−.529	.154
General understanding and expressions	Idiomatic expression	2.62	1.135	3.08	1.002	−3.109	.002	−.462	.148
	Comprehensibility	3.23	1.125	3.36	1.123	−.802	.424	−.125	.156
	Authenticity	2.42	1.049	2.84	1.080	−2.800	.006	−.413	.148
	Contextual understanding	3.26	1.043	3.11	1.004	1.084	.280	.154	.142
	Unambiguity	2.87	1.212	3.04	1.123	−1.009	.314	−.163	.162
Text statistics	Word count	196.87	98.031	216.63	109.17	−1.373	.171	−19.760	14.388
	Sentence length	14.705	74.515	13.365	5.5129	2.007	.046	1.3298	.6625
	Lexical density	56.729	7.577	56.377	12.376	−.209	.835	−.2757	1.320
	Gunning Fog Index	11.892	2.101	13.805	2.766	−5.614	.000	−1.9126	.340

Table 2. Evaluation results of the L1 texts

	Aspect	*M*	*SD*
Mechanics	Spelling	1.92	.268
	Punctuation	1.75	.435
Lexico-grammatical accuracy	Vocabulary appropriateness	3.94	.846
	Vocabulary level and diversity	3.52	.914
	Grammatical accuracy	3.87	1.005
General understanding and expressions	Comprehensibility	4.41	.705
	Unambiguity	3.81	1.141
	Korean contextual knowledge	3.39	1.464
Text statistics	Sentence complexity	3.02	.881
Total		20.22	3.162

Table 3. Regression analysis results

		Unstandardized coefficients		Standardized coefficients		
		B	*SE*	Beta	*t*	*p*
Mechanics	Spelling	.541	.297	.141	1.820	.072
	Punctuation	.693	.209	.292	3.308	.001
Lexico-grammatical accuracy	Vocabulary appropriateness	−.044	.133	−.036	−.333	.740
	Vocabulary level and diversity	.175	.114	.155	1.538	.127
	Grammatical accuracy	.066	.132	.064	.500	.618
General understanding and expressions	Comprehensibility	−.069	.155	−.048	−.447	.656
	Unambiguity	.217	.119	.240	1.822	.072
	Level of contextual information	.096	.052	.136	1.827	.071
Text statistics	Sentence complexity	.303	.087	.259	3.465	.001

Note. Dependent variable: grammatical accuracy of the MT outputs.

length ($t = 2.007$) and GFI ($t = -5.614$) were statistically meaningful, whereas that in lexical density was not.

Regarding the L1 texts, the results showed that, overall, the students had a good command of mechanics, language, and comprehensibility in L1. The students seldom committed errors in spelling ($M = 1.92$) but produced more errors in punctuation ($M = 1.75$). The results showed that most of the L1 texts did not have issues with comprehensibility ($M = 4.41$). The results of the L1 text analysis are summarized in Table 2. The scores of students' L1 texts ranged between 13 and 25 (out of 39; $M = 20.22$). However, it should be noted here that a higher total score does not necessarily indicate a higher quality of L1 writing because two items – level of Korean contextual information and syntactic complexity – are unrelated to text quality.

Next, the associations between the students' L1 texts and MT accuracy were examined. As the correlation analysis indicated that most of the areas evaluated in the L1 texts were correlated with those in the MT texts, multiple regression analysis was conducted to examine which factors of the L1 texts influenced MT text quality (Table 3). The results showed that only punctuation and

sentence complexity affected MT output quality. Grammatical accuracy; vocabulary appropriateness, level, and diversity; cultural understanding; and language ambiguity in the L1 texts did not influence it.

4.2 Evaluators' reflections

Four themes emerged in the evaluators' reflections about the students' L2 and MT texts. First, all evaluators, with the exception of one, did not realize during evaluation that one set was translated via MT. Second, they regarded the MT texts as better than the students' in terms of overall quality and lexico-grammatical aspects; hence, two of them thought that the MT texts were revisions of the students' texts. Overall, they agreed that vocabulary was more accurate and diverse in the MT outputs, and MT demonstrated higher accuracy of grammar. On the other hand, they also found that MT sometimes produced completely inaccurate or incomprehensible sentences (i.e. "Second, to use for frequent among teens these days can be seen") or awkward vocabulary choices, such as "abbreviated abbreviation," which even beginner college students might not use. The evaluators further mentioned that only one or two such extreme errors were enough to undermine the authenticity of the texts and their overall perceptions of them.

Third, the MT texts showed irregularities in grammar and vocabulary. That is, although lexical and grammatical accuracy was consistent in the students' texts, it was inconsistent in the MT texts. The evaluators reported that the vocabulary level used in a single MT text was not consistent as well. For instance, they sometimes found the most advanced vocabulary among low-level vocabulary in the MT texts, in which they assumed that the students used a dictionary. Additionally, the evaluators noted the disparity between grammar level and vocabulary level in the MT texts; that is, mismatches were found between vocabulary and grammar levels, such as more advanced words in poorly structured sentences.

The last theme pertained to the evaluators' suspicions regarding MT usage. The irregularities found in the MT texts led an evaluator to suspect that "some students seemed to get help from MT, at least in some parts" because those parts sounded too good or too bad. Two other evaluators also mentioned that they "thought it was strange that some parts in the MT outputs, which [they] originally considered revisions, contained atypical and unusual errors and were worse than the original versions (students' L2 texts)," and they came to understand how this happened when they were told that one of the sets was produced via MT.

5. Discussion

5.1 Credibility of MT outputs

Regarding the first research question, the results revealed that MT generated more qualified English texts than the students in most aspects under investigation. Previous studies showed that MT achieved lower accuracy compared to intermediate human translators and is predisposed to produce more lexical and grammatical errors (Abraham, 2009; Fredholm, 2015; White & Heidrich, 2013). In contrast, the current study, as also claimed in the most recent studies (Briggs, 2018; Ducar & Schocket, 2018; Stapleton & Kin, 2019; Tsai, 2019), indicated a significant improvement in the quality of the MT outputs in terms of lexicon and grammar and, as a result, outperformed the intermediate EFL students in almost all investigated aspects. Although it was not surprising that MT produced fewer errors in mechanical aspects such as spelling and punctuation, it was impressive that it was assessed higher in the lexical areas. Resembling a recent comparative MT study between MT and EFL students in Hong Kong (Stapleton & Kin, 2019), this study showed that the MT outputs contained more diverse and advanced words than the students' texts. Stapleton and Kin (2019) reported that, compared to students, MT produced "not only a correct translation, but also one with more advanced language and nuanced meaning"

(p. 27). However, they also pointed out that MT sometimes produced literal translations because it does not consider a word's context. These errors were found in the MT texts in this study as well; despite this, MT was evaluated higher than the students in both lexical aspects, appropriateness, and level and diversity. The MT outputs' use of more advanced and diverse words seemed to result in a higher level of readability (GFI). Also, no significant differences in lexical density were found between the two versions, which implies that the lexical density of the source texts determined that of both translated versions in the first place.

Another noticeable finding in this study was the grammatical accuracy of MT. Grammatical inaccuracy has been a major concern of language teachers when using MT for educational purposes (Barr, 2013; Fredholm, 2015; Lee, 2020). However, in this study, MT earned a significantly higher score in grammar compared with the students' L2 outputs. Considering that this study's participants had an intermediate level of English proficiency, MT seems to have already surpassed the intermediate level. Although the scores in grammar indicated MT's overall superiority over the students, the evaluators' reflections revealed that MT also produced serious grammatical errors. In addition, while the language proficiency was consistent within a student's English text, the MT outputs sometimes contained inconsistencies. Particularly, as expressed by O'Neill (2014, 2016), mismatches between the vocabulary and grammar levels, such as more advanced words in ill-formed sentences were found, which implied that the grammatical accuracy of MT is lower than its lexical accuracy. Ducar and Schocket (2018) argued that such irregularities in the students' L2 texts can be a telltale sign of the students' use of MT. In the current paper, they somewhat affected the evaluators' perceptions of authenticity. Nonetheless, MT scored higher in terms of authenticity, proving that MT's overall language proficiency was higher than that of the students despite the errors. The MT outputs also obtained a significantly higher score in idiomatic expressions. When the students did not know a certain idiomatic expression, they would literally translate it using a dictionary. In contrast, as Google can process a large amount of data, MT is assumed to have become more capable of finding more accurate expressions. Particularly in the present study, the source texts contained many Korean slang and abbreviations used in youth culture, and MT translated well in most of the cases.

5.2 L1 factors influencing MT quality

The second research question examined which factors in the source texts influenced MT quality. The study found that these factors were punctuation and sentence complexity. Although the students did not make many punctuation errors in the source texts, it turned out that even a small number of punctuation errors led to undesirable results in the MT outputs. For instance, missing periods greatly diminished MT output accuracy. Missing commas, which increased sentence complexity, had the same effect on the quality of the MT outputs, as sentence complexity was negatively correlated with MT output accuracy. This is consistent with prior studies that claimed that sentence length and structure determined MT outcomes (Jolley & Maimone, 2015; Kim, 2019; Ruiz & Federico, 2014; Shadiev *et al*., 2019). The study revealed that the more complex the L1 sentence, the more errors produced by MT. It is assumed that long and complex sentences make it more difficult for MT to find and match equivalent L2 translations from its database.

On the other hand, this study found that the grammatical accuracy of the source texts did not affect the quality of the MT outputs. Two explanations are possible: one is that the students did not commit many serious grammatical errors in their source texts; thus, these grammatical errors hardly affected MT output quality. The other explanation is that MT detected some simple errors in the source texts and filled the gaps in its translations. For instance, sentences with missing pronouns, particularly personal pronouns, such as "I," which are common in Korean language (Kim, 2019), are not necessarily ungrammatical in Korean but would be in English, thus affecting the MT outputs. Such cases were often encountered in the source texts, and it was found that MT

correctly added missing subjects and objects in the majority of cases, as shown in the following example:

L1: 어린 학생들이 (young students) 줄임말을 (abbreviations) 사용하는 것은 (to use) 옳지 않다고 (is not right) 생각한다 (think).
MT output: I think it is not right for young students to use abbreviations.

In this example, the verb "think" has no subject in L1, but MT put "I " in its translation. MT also put "it" as the subject for "is not right." Unlike earlier studies (Niño, 2008, 2009; Steding, 2009) that pointed to MT's inability to understand context as one of its major weaknesses, this study showed that MT was capable of finding context-based referents in some cases even when denotative elements were missing. The result was contradictory to that of Kim (2019), wherein the pro-drop feature of the input language (Korean) led to incorrect translation in the MT output (English). The inconsistency between his study and the current study was attributed to the use of different units for translation. The students in Kim's study translated a single sentence owing to which Kim's results did not provide any context for missing pronouns. In the present study, the students translated entire texts; thus, MT was able to correctly draw out the missing pronouns in the sentences. Moreover, Correa (2014) and Goulet, Simard, Parra Escartín and O'Brien (2017) pointed out that translations between syntactically distant language pairs would more likely produce a large number of syntactical errors. In a recent study, Shadiev *et al.* (2019) confirmed that intelligibility and accuracy vary depending upon the similarities or differences between language pairs. Considering the striking structural differences between Korean and English, the results in this study were outstanding. In addition, this study discovered that the accuracy, vocabulary diversity, and level of Korean contextual information included in the source texts did not significantly affect MT output quality. This result, again, seems to be attributed to MT's use of enormous corpus data to easily match the words and expressions of the source texts with English words and expressions appropriate for the given context (Bowker & Ciro, 2019).

5.3 Pedagogical implications

The mixed results of previous studies on MT output quality and effectiveness in FL education may hold back language teachers from using it. Studies published prior to 2017 often reported the partial effectiveness of using MT in FL classrooms with a warning that it remains far from perfect despite its usefulness (Garcia & Pena, 2011; Niño, 2008, 2009). However, more recent studies presented improved MT outcomes, with some even arguing that MT is superior to their EFL students participating in their studies (Briggs, 2018; Lee, 2020; Stapleton & Kin, 2019; Sun, 2017; Tsai, 2019). The results of the present study supported these most recent studies. Lee (2020) compared the L2 revisions of MT with those of students' peers whose FL command is not perfect but still helpful. She found that MT helped the EFL students fix lexical and grammatical errors in their revisions, as did their peers. In the same vein, the present study indicated that MT can be a useful tool for EFL students, as it produced better-quality outputs, particularly in vocabulary and grammar.

There are a few issues that language instructors should consider when using MT to mitigate its accuracy problems and maximize its effectiveness. First, they should select a tool that is compatible to the language pair, as each MT tool works differently with certain language pairs (Goulet *et al.*, 2017; Shadiev *et al.*, 2019). Second, as MT is becoming more accurate, teachers and researchers need to reframe its instructional model. Niño (2009) proposed four models of MT use: a bad model, a good model, vocational use, and a computer-assisted language learning (CALL) tool. MT had been initially used as a bad model, where students detected errors and fixed them through post-editing (Kliffer, 2008; Niño, 2008, 2009). However, recent studies have shown that MT is now more often used as a CALL tool, in which MT serves as a facilitator to help

students solve their own language problems (Lee, 2020; O'Brien *et al*., 2018; Stapleton & Kin, 2019; Tsai, 2019). Considering the increasing accuracy of MT, it is important for teachers to find the best instructional model for it in their classroom. In doing so, teachers also need different models appropriate for different language proficiency groups.

In addition, strategies on how to use MT affect MT outcomes and educational effectiveness. Prior studies suggested translating short segments, such as word or phrase units, for better-quality MT outcomes (Jolley & Maimone, 2015; Stapleton & Kin, 2019); thus, it is preferable to translate a small chunk each time for more accurate outcomes. Conversely, the current study showed that large portions of texts can provide a context for translation, through which MT can correctly fill in the missing pronouns in the source texts. This study also suggests that longer and complex sentences are more error-prone; using short and concise sentences in the source text or refining them through pre-editing will increase the accuracy of MT outcomes. For better results of MT outputs, Bowker and Ciro (2019) suggest making L1 writing simple and MT-friendly. More specifically, their guidelines include using short sentences, avoiding wordiness, using nouns instead of personal pronouns, choosing unambiguous words, and avoiding idiomatic expressions and cultural references. Last, O'Neill (2013) emphasized the importance of training on the effective strategies of using MT in the language classroom. He confirmed that students who underwent a training session prior to using MT outperformed those with no training. Therefore, teachers should provide guidelines for using MT and explicitly teach effective strategies to students prior to using it.

It is also important for teachers to consider various factors, options, and strategies to help their students effectively and appropriately use MT. Thus, teachers should develop an effective instructional model and strategies to best utilize it, as MT remains a capable tool despite its imperfections. For instance, modifying or simplifying the source language and selecting MT that is accurately targeted for a specific language pair can increase MT accuracy and the quality of MT outputs. Additionally, different language skills and proficiencies may require different strategies for using MT. More importantly, language instructors should be familiar with the tool first and be aware of its strengths, weaknesses, and potential dangers in the classroom and inform students of such issues beforehand. As MT is set to become more accurate in the future, instructors should also be prepared to deal with its corresponding challenges such as academic dishonesty and student demotivation toward FL learning.

6. Conclusion

As MT has become more popular in FL education, researchers have warned language teachers against merely denying or entirely banning it (Briggs, 2018), but their main concern regarding MT – its output quality – has not been examined yet in FL education. The purpose of the present study was to evaluate MT output quality and the factors in the source texts that influence it. It found that MT outperformed the students in most aspects. In particular, this study's significant finding was MT's superiority in lexical and grammatical aspects, indicating that it can serve as a pedagogical tool better than before in the language classroom. In Lee's (2020) terms, this means that EFL students can recruit a more capable peer to help their language learning. In EFL classrooms with inadequate teacher feedback (Amaral & Meurers, 2011; Briggs, 2018; Tsai, 2019), MT has a great potential to facilitate students' language learning; hence, teachers should not ignore or deprive the students of this tool.

This study has a few limitations as well as implications for future studies. As the present study included only one language pair, Korean and English, the results cannot be generalized to other language pairs. Also, while this study evaluated MT outputs according to nine aspects, it did not analyze the grammatical errors of MT outputs through error analysis. This necessitates an empirical study that adopts systematic error analysis to assess MT's grammatical accuracy and

idiosyncratic errors. Additionally, although this study investigated long passages translated via MT, future studies would benefit from examining translations of short segments, which may show different results on MT accuracy.

Although the current study confirmed the potential of MT as a pedagogical tool in the FL classroom, there are many issues that are yet to be investigated. For instance, although MT can facilitate the L2 writing and revising process, the longer-term effect of MT on L2 learning still remains uncertain. Students may simply copy or adopt MT outputs in their writing, which will not lead to L2 learning. Moreover, the effectiveness of using MT may vary due to learners' different language proficiencies or different cognitive levels. Learners' strategies of using MT will also influence outcomes of using MT, and in the same vein, teachers' guidance will also play an important role. Last, as MT is constantly evolving, more updated research on MT accuracy in various language pairs is essential. Future research on these issues will provide more insight into MT as a pedagogical tool for language learning.

Supplementary material. To view supplementary material referred to in this article, please visit https://doi.org/10.1017/S0958344021000124

Ethical statement. There are no conflicts of interest in this paper. All participation in the study was voluntary, and informed consent was provided prior to commencement of the assignment. The anonymity of the participants was maintained throughout the study.

References

Abraham, L. B. (2009) Web-based translation for promoting language awareness: Evidence from Spanish. In Abraham, L. B. & Williams, L. (eds.), *Electronic discourse in language learning and language teaching*. Amsterdam: John Benjamins, 65–83. https://doi.org/10.1075/lllt.25.06abrl

Aiken, M. & Balan, S. (2011) An analysis of Google Translate accuracy. *Translation Journal*, 16(2). http://translationjournal.net/journal//51pondering.htm

Alhaisoni, E. & Alhaysony, M. (2017) An investigation of Saudi EFL university students' attitudes towards the use of Google Translate. *International Journal of English Language Education*, 5(1): 72–82. https://doi.org/10.5296/ijele.v5i1.10696

Amaral, L. & Meurers, D. (2011) On using intelligent computer-assisted language learning in real-life foreign language teaching and learning. *ReCALL*, 23(1): 4–24. https://doi.org/10.1017/S0958344010000261

Bahri, H. & Mahadi, T. S. T. (2016) Google Translate as a supplementary tool for learning Malay: A case study at Universiti Sains Malaysia. *Advances in Language and Literary Studies*, 7(3): 161–167. https://doi.org/10.7575/aiac.alls.v.7n.3p.161

Barr, D. (2013) Embedding technology in translation teaching: Evaluative considerations for courseware integration. *Computer Assisted Language Learning*, 26(4): 295–310. https://doi.org/10.1080/09588221.2012.658406

Bowker, L. & Ciro, J. B. (2019) *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. *Bingly: Emerald Publishing*. https://doi.org/10.1108/9781787567214

Briggs, N. (2018) Neural machine translation tools in the language learning classroom: Students' use, perceptions, and analyses. *The JALT CALL Journal*, 14(1): 3–24. https://doi.org/10.29140/jaltcall.v14n1.221

Choi, Y. H. & Lee, J. (2006) L1 use in L2 writing process of Korean EFL students. *English Teaching*, 61(1): 205–225.

Clifford, J., Merschel, L. & Munné, J. (2013) Surveying the landscape: What is the role of machine translation in language learning? *The Acquisition of Second Languages and Innovative Pedagogies*, 10: 108–121. https://doi.org/10.7203/attic.10.2228

Cook, G. (2012) *Translation in language teaching: An argument for reassessment*. Oxford: Oxford University Press.

Correa, M. (2014) Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content and Language Integrated Learning*, 7(1): 1–20. https://doi.org/10.5294/laclil.2014.7.1.1

Ducar, C. & Schocket, D. H. (2018) Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google Translate. *Foreign Language Annals*, 51(4): 779–795. https://doi.org/10.1111/flan.12366

Fredholm, K. (2015) Online translation use in Spanish as a foreign language essay writing: Effects on fluency, complexity and accuracy. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, 18: 7–24.

Fredholm, K. (2019) Effects of Google Translate on lexical diversity: Vocabulary development among learners of Spanish as a foreign language. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, 13(26): 98–117. https://doi.org/10.26378/rnlael1326300

Garcia, I. (2016) Can machine translation help the language learner? *International conference ICT for language learning*. https://conference.pixel-online.net/conferences/ICT4LL2010/common/download/Proceedings_pdf/TRAD02-Garcia.pdf

Garcia, I. & Pena, M. I. (2011) Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning*, 24(5): 471–487. https://doi.org/10.1080/09588221.2011.582687

Giannetti, T. R. (2016) *Google Translate as a resource for writing: A study of error production in seventh grade Spanish*. St. John Fisher College, master's thesis.

Godwin-Jones, R. (2019) In a world of SMART technology, why learn another language? *Educational Technology & Society*, 22(2): 4–13.

Goulet, M.-J., Simard, M., Parra Escartín, C. & O'Brien, S. (2017) Using machine translation for academic writing in English as a second language: Results of an exploratory study on linguistic quality. *ASp: la revue du GERAS*, 72: 5–28. https://doi.org/10.4000/asp.5045

Groves, M. & Mundt, K. (2015) Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, 37: 112–121. https://doi.org/10.1016/j.esp.2014.09.001

Im, H.-J. (2017) The university students' perceptions or attitudes on the use of the English automatic translation in a general English class: Based on English writing lessons. *Korean Journal of General Education*, 11(6): 727–751.

Jolley, J. R. & Maimone, L. (2015) Free online machine translation: Use and perceptions by Spanish students and instructors. In Moeller, A. J. (ed.), *Learn language, explore cultures, transform lives: Selected papers from the 2015 Central States Conference on the Teaching of Foreign Languages*. Richmond: Robert M. Terry, 181–200.

Kazemzadeh, A. & Kashani, A. (2014) The effect of computer-assisted translation on L2 learners' mastery of writing. *International Journal of Research Studies in Language Learning*, 3: 29–44.

Kim, E.-Y. (2011) Using translation exercises in the communicative EFL writing classroom. *ELT Journal*, 65(2): 154–160. https://doi.org/10.1093/elt/ccq039

Kim, S. (2019) Playing with machine translation in language classroom: Affordances and constraints. *Multimedia-Assisted Language Learning*, 22(2): 9–28.

Kim, Y. & Yoon, H. (2014) The use of L1 as a writing strategy in L2 writing tasks. *GEMA Online Journal of Language Studies*, 14(3): 33–50. https://doi.org/10.17576/GEMA-2014-1403-03

Kliffer, M. D. (2008) Post-editing machine translation as an FSL exercise. *Porta Linguarum*, 9: 53–67. https://doi.org/10.30827/Digibug.31745

Kol, S., Schcolnik, M. & Spector-Cohen, E. (2018) Google Translate in academic writing courses? *The EUROCALL Review*, 26(2): 50–57. https://doi.org/10.4995/eurocall.2018.10140

Lee, S. (2019) Korean college students' perceptions toward the effectiveness of machine translation on L2 revision. *Multimedia-Assisted Language Learning*, 22(4): 206–225. https://doi.org/10.15702/mall.2019.22.4.206

Lee, S.-M. (2020) The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, 33(3): 157–175. https://doi.org/10.1080/09588221.2018.1553186

Marinac, M. & Barić, I. (2018) Teachers' attitudes toward and use of translation in the foreign language classroom at institutions of higher education in Croatia. *Theory and Practice in Language Studies*, 8(8): 906–915. https://doi.org/10.17507/tpls.0808.02

Maruf, S. & Haffari, G. (2018) Document context neural machine translation with memory networks. In Gurevych, I. & Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long papers)*. Stroudsburg: Association for Computational Linguistics, 1275–1284.

Murtisari, E. T., Widiningrum, R., Branata, J. & Susanto, R. D. (2019) Google Translate in language learning: Indonesian EFL students' attitudes. *The Journal of AsiaTEFL*, 16(3): 978–986. https://doi.org/10.18823/asiatefl.2019.16.3.14.978

Niño, A. (2008) Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, 21(1): 29–49. https://doi.org/10.1080/09588220701865482

Niño, A. (2009) Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, 21(2): 241–258. https://doi.org/10.1017/S0958344009000172

O'Brien, S., Simard, M., & Goulet, M.-J. (2018) Machine translation and self-post-editing for academic writing support: Quality explorations. In Moorkens, J., Castilho, S., Gaspari, F. & Doherty, S. (eds.), *Translation quality assessment: From principles to practice*. Cham: Springer, 237–262. https://doi.org/10.1007/978-3-319-91241-7_11

O'Neill, E. (2013) Online translator usage in foreign language writing. In Swanson, P. B. & Hoyt, K. (eds.), *Dimension 2013. World language learning: Setting the global standard*. Decatur: Southern Conference on Language Teaching, 74–88.

O'Neill, E. M. (2014) Real-life technology and the L2 French classroom: Online translation usage among intermediate French students. *Proceedings of the 2014 American Association of Teachers of French Convention*. New Orleans, LA, 19–22 July.

O'Neill, E. M. (2016) Measuring the impact of online translation on FL writing scores. *The IALLT Journal*, 46(2): 1–39. https://doi.org/10.17161/iallt.v46i2.8560

Park, J. (2017) Analysis of the types of translation errors by the Google Translator. *Studies on English Language & Literature*, 59(4): 279–299.

Ruiz, N. & Federico, M. (2014) Complexity of spoken versus written language for machine translation. In Cettolo, M., Federico, M., Specia, L. & Way, A. (eds.), *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*. Allschwil: European Association for Machine Translation, 173–180.

Shadiev, R., Sun, A. & Huang, Y.-M. (2019) A study of the facilitation of cross-cultural understanding and intercultural sensitivity using speech-enabled language translation technology. *British Journal of Educational Technology*, 50(3): 1415–1433. https://doi.org/10.1111/bjet.12648

Stapleton, P. & Kin, B. L. K. (2019) Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. *English for Specific Purposes*, 56: 18–34. https://doi.org/10.1016/j.esp.2019.07.001

Steding, S. (2009) Machine translation in the German classroom: Detection, reaction, prevention. *Die Unterrichtspraxis/Teaching German*, 42(2): 178–189. https://doi.org/10.1111/j.1756-1221.2009.00052.x

Sun, D. (2017) Application of post-editing in foreign language teaching: Problems and challenges. *Canadian Social Science*, 13(7): 1–5. https://doi.org/10.3968/9698

Thue Vold, E. (2018) Using machine-translated texts to generate L3 learners' metalinguistic talk. In Haukås, Å., Bjørke, C. & Dypedahl, M. (eds.), *Metacognition in language learning and teaching*. New York: Routledge, 67–97. https://doi.org/10.4324/9781351049146-5

Tsai, S.-C. (2019) Using Google Translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning*, 32(5–6): 510–526. https://doi.org/10.1080/09588221.2018.1527361

White, K. D. & Heidrich, E. (2013) Our policies, their text: German language students' strategies with and beliefs about web-based machine translation. *Die Unterrichtspraxis/Teaching German*, 46(2): 230–250. https://doi.org/10.1111/tger.10143

Yang, Z., Chen, W., Wang, F. & Xu, B. (2018) Unsupervised neural machine translation with weight sharing. In Gurevych, I. & Miyao, Y. (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long papers)*. Stroudsburg: Association for Computational Linguistics, 46–55.

About the author

Sangmin-Michelle Lee is a professor of Global Communication at Kyung Hee University in the Republic of Korea. She earned her PhD from the Pennsylvania State University in language education. She has published papers on language learning in a technology-enhanced learning environment, machine translation, L2 writing, game-based learning, and digital creativity.

Author ORCiD. Sangmin-Michelle Lee, https://orcid.org/0000-0002-7686-3537