Effects of machine translation on L2 writing proficiency: The complexity, accuracy, lexical diversity, and fluency

*Sangmin-Michelle Lee, Kyung Hee University
Nayeon Kang, Sanggol Middle School

Abstract

With recent improvements in machine translation (MT) accuracy, MT has gained unprecedented popularity in second language (L2) learning. Despite the significant number of studies on MT use, the effects of using MT on students’ retention of learning or secondary school students’ use of MT in L2 writing has rarely been researched. The current study investigates the effectiveness of using machine translation on Korean middle school students’ L2 writing over an extended period of time. This study evaluated the complexity, accuracy, lexical diversity, and fluency of four versions of the students’ writing (pretest, MT-assisted version, posttest, and 2-week or 4-week retention tests) and measured errors in punctuation, spelling, vocabulary, and grammar in each version. The study employed a quantitative research method including descriptive statistics, repeated measures ANOVA, and paired t-tests. The results showed that fluency, accuracy, and complexity significantly increased in the MT-assisted version in every aspect of writing but decreased in the subsequent versions without MT (post- and retention tests). Decreases occurred more frequently with grammatical items than with lexical items. Despite the decreases, all of the items measured in the study scored higher in the retention tests than in the pretest, which indicates that the use of MT had a positive effect on L2 writing.

Keywords: Machine Translation, L2 Writing, Accuracy, Lexical Diversity

Language(s) Learned in This Study: English


Introduction

Second language (referred to hereafter as L2) writing is a daunting task for both learners and teachers. For L2 learners, it is a double burden because learners struggle with both the content and the language (Briggs, 2018; Tsai, 2022). This challenge is even greater for beginners, who do not have sufficient language skills to express their ideas in the L2 (Garcia & Peña, 2011; Tsai, 2022). L2 writing is also challenging from the teacher’s perspective. Teacher feedback is indispensable for improving L2 writing; however, it requires not only a great amount time and effort to facilitate L2 writing for students with different proficiency levels, but it is also never easy (Ferris & Hedgcock, 2014; Lee, 2020). Moreover, due to classroom constraints, such as large classroom size, limited class time, or shortage of teachers, it is often difficult for teachers to provide individualized feedback and facilitate L2 writing (Arifin, 2017). As a result, L2 writing practice is often limited, and students do not have sufficient opportunities to develop L2 writing skills (Tsai, 2022). In this L2 writing context, machine translation (MT) is a valuable means to aid L2 writing by helping learners reduce lexicogrammatical errors (Chon et al., 2021; Fredholm, 2015, 2019; Yang et al., 2023), focus more on content (Briggs, 2018; Chandra & Yuyun, 2018), and improve confidence in L2 writing (Fredholm, 2021).

* Corresponding author: Sangmin-Michelle Lee, sangminlee@khu.ac.kr
Prior to Google’s introduction of neural network MT in 2016, studies had reported a number of problems with MT output quality, such as literal translation, mistranslation of idiomatic expressions and colloquial language, and incomprehensibility of cultural nuances. More recent studies have confirmed that most of the previously reported MT errors have been resolved (Ducar & Schocket, 2018; Lee, 2022a; Tsai, 2022), while some errors, such as literal translation and mistranslation of less common words, have been significantly reduced (Ducar & Schocket, 2018). However, problems with MT still remain, such as different outcome accuracy among language pairs and translation irregularities (Lee, 2022a; Shadiev et al., 2019). Despite its imperfections, MT has gained increasing popularity due to this improvement and has now become standard in FL classrooms (Tsai, 2022). Accordingly, the number of studies in this field has been significantly growing over the last couple of years, and diverse pedagogical benefits have been reported. Regarding language, MT has been known to help enhance lexicogrammatical accuracy, increase vocabulary appropriateness, and enhance meaning delivery (Chang et al., 2022; Fredholm, 2019; Lee, 2022a; Niño, 2020; Tsai, 2019, 2022). From the metacognitive aspect, MT mediates the revision process and enables students to view writing as a process (Lee, 2020; Lee & Briggs, 2021), which is assumed to increase L2 learners’ metalinguistic awareness (Jolley & Maimone, 2015; Valijärvi & Tarsoly, 2019). MT also fosters independent self-directed learning (Valijärvi & Tarsoly, 2019) and cultivates critical thinking skills about language sources (Lee, 2022b; Valijärvi & Tarsoly, 2019). From the affective aspect, MT can lower language apprehension, particularly for lower-level students, and increase motivation and confidence toward L2 writing (Chon et al., 2021; Fredholm, 2021; Tsai, 2022).

Overall, previous studies have found that using MT is effective for L2 writing and revision (Chon et al., 2021; Lee, 2020, 2022a; Tsai, 2019; Yang et al., 2023); however, most of them limited their analysis to the effect of MT for the writing product, except for Fredholm (2019). In addition, although MT helped to improve the quality of the L2 writing product in these studies, it was not clear whether students gained L2 writing proficiency. For example, Fredholm (2019) showed that vocabulary gains with MT disappeared when students did not use MT. In other words, studies found that the quality of L2 writing improved when aided by MT, but they did not examine if the effect would last or transfer to learning. Hence, the current study examined the longer-term effect of MT on L2 writing by comparing students’ texts at different points of time up to four weeks after using MT.

**Literature Review**

Over the last few years, studies have found a high percentage of MT use in L2 classrooms (Hwang et al., 2022; Jolley & Maimone, 2022; Tourmen & Hoffmann, 2022), and this recent unprecedented popularity of MT was ascribed mainly to the improvement in MT accuracy. Unlike previous MT studies published prior to 2018, more recent studies have confirmed that MT accuracy has significantly improved by resolving most previous errors, such as grammatical inaccuracies, register, and inability to understand double-meaning words as well as the most notorious errors: literal translations (Chon et al., 2021; Kol et al., 2018; Stapleton & Kin, 2019). MT can better identify contextual and cultural nuances (Ducar & Shocket, 2018) and fill in missing subjects or objects in a sentence when translating from a pro-drop source language, such as Korean into English (Lee, 2021). As a consequence, a number of researchers have concluded that MT produces a better quality of output than high-intermediate English as a foreign language (EFL) learners in English writing (Stapleton & Kin, 2019; Tsai, 2019, 2022; Yang et al., 2023). The enhanced accuracy of MT, in turn, increased the reliability of its output, resulting in wide acceptance and approval from language teachers (Hwang et al., 2022; Tsai, 2022).

With this greater acceptance of MT in the language classroom, an increasing number of studies have been published regarding how MT use can facilitate student L2 writing in diverse aspects, such as reducing lexicogrammatical errors (e.g., Chung & Ahn, 2022; Lee, 2020, 2022a; Tsai, 2019; Yang et al., 2023) or enhancing fluency and complexity (e.g., Fredholm, 2015, 2019; Yang et al., 2023). Over the last decade, instructors have used MT for pedagogical purposes in different ways in L2 classrooms depending on its level of accuracy (or inaccuracy) at that time. Niño (2009) proposed four instructional models of MT use:
a bad model, a good model, vocational use, and computer-assisted language learning (CALL). In the earlier years of using MT, a bad model had been incorporated into the L2 classroom. For instance, in Niño’s study (2008) students learned language by detecting lexicogrammatical errors shown in MT outputs and correcting them. In more recent years, due to its increased reliability, MT has been used more frequently with CALL strategies, which assist L2 writing and the revision process (Niño, 2020). According to the literature, L2 writing is a specific domain where MT has diverse educational benefits (Briggs, 2018). Most frequently, prior studies have revealed that students with MT have demonstrated significantly better performance in terms of lexicogrammatical accuracy (Fredholm, 2015; Lee, 2022; Tsai, 2019, 2022), fluency (Fredholm, 2015; Lee, 2022), and complexity (Chang et al., 2022; Chon et al., 2021; Fredholm, 2015; Tsai, 2022; Yang et al., 2023). Fredholm (2015, 2019) investigated lexical diversity and complexity in students’ writing and reported that the MT group produced texts with higher lexical diversity. Considering that lexicogrammatical errors often negatively affect the overall quality of writing, improved accuracy can directly lead to quality improvement in L2 writing (Chang et al., 2022; Mín, 2006). Studies have also shown that MT facilitates L2 writing by reducing errors and suggesting better expressions for the intended meaning (Niño, 2020; Tsai, 2019). MT also provides a means of expression for beginner students who are otherwise unable to express their thoughts (Chang et al., 2022; Garcia & Peña, 2011). As beginner students struggle with both language and content during L2 writing, more frequently with language, using MT can help decrease their cognitive overload by supporting language and allowing students to focus more on meaning (Chandra & Yuyun, 2018; Chang et al., 2022; Tsai, 2022).

From a revision perspective, MT can provide written corrective feedback (WCF) to individual students by suggesting or replacing a lexical choice with one more appropriate for the context and by reformulating sentences (Lee, 2022a). However, previous studies have presented mixed results on which method of corrective feedback is the most effective (Bitchener & Knoch, 2009; Chandler, 2003; Lee, 2019). For instance, the literature on L2 writing distinguishes between written and oral, direct and indirect (i.e., whether feedback directly corrects the error), and comprehensive and focused feedback (i.e., whether feedback is on a limited number or a wide range of errors). Despite the controversy, more studies have favored direct corrective feedback because it offers explicit information about the error, avoids confusion and misunderstanding from students, and thus better facilitates L2 writing, especially grammar acquisition (Bitchener & Knoch, 2009; Zhang & Cheng, 2021). Concerning the effectiveness of each type of corrective feedback, previous studies were more inconclusive. Some have advocated for focused corrective feedback because it is friendlier and less demanding compared to comprehensive corrective feedback, which demands too much cognitive effort and attention from students (Bitchener & Storch, 2016; Shintani & Aubrey, 2016). On the other hand, others have argued that comprehensive corrective feedback is more effective (Zhang & Cheng, 2021), because providing focused corrective feedback can lead to correcting focused errors but does not necessarily lead to gains in the overall quality of L2 writing. Despite the disagreement on the most effective method for corrective feedback and the dispute on its ineffectiveness on L2 writing development (e.g., Truscott, 2007), L2 writing scholars generally concur that corrective feedback plays a positive role in L2 revision, particularly in increasing writing accuracy, and it is therefore indispensable for L2 writing development (Ferris, 2011; Zhang & Cheng, 2021). However, it requires a significant amount of time and effort from the teacher, and due to limited time and large classroom sizes, teachers’ corrective feedback is often unavailable in the L2 writing classroom (Tsai, 2019; Yang et al., 2023).

Under this situation, although it cannot replace teacher feedback, MT can serve as an alternative pedagogical tool to assist student L2 revision, and prior studies have found a positive impact of using MT on revision (Chang et al., 2022; O’Neill, 2016; Stapleton & Kin, 2019; Tsai, 2019, 2022; Yang et al., 2023). However, revision aided by MT is not always easy for L2 students due to the nature of the MT output. According to Lee (2022a), MT output lies between direct and indirect feedback. She states that MT reformulates sentences rather than providing one-on-one direct feedback on each error; thus, revising with MT requires students to notice their errors and correct them, which is challenging, especially for beginner students. Another study by Lee (2022b) showed that learners frequently translated larger chunks of words,
such as sentences (39.5%) or paragraphs (13.7%) using MT. This means that MT revises multiple sentences, a paragraph, or paragraphs, depending on how individual learner’s use MT, rather than focusing on target errors (Koltovskiaia, 2022). Hence, MT output can be overwhelming and frustrating to L2 students. Compared to lexical errors, syntactic errors are more difficult to find in sentences reformulated by MT (Niño, 2008), and beginner students have a hard time correcting grammatical errors during revision with MT because they usually do not have enough linguistic knowledge or confidence to choose between their own output and MT output (Fredholm, 2021; Lee, 2022a). Sachs and Polio (2007) also argued that reformulation is more difficult and less effective than direct feedback in L2 writing.

Despite the increasing number of MT studies, there are still research gaps. Most importantly, the long-term effects of using MT on L2 writing and learning have seldom been explored, and they remain largely unknown. Fredholm (2019) found that MT helped enhance lexical diversity and density, but the effect was valid only when MT was available to students; thus, he speculated that MT did not have a long-term effect on L2 writing. There are also inconsistencies in the findings of previous studies. For instance, while Chung and Ahn (2021) and Lee (2020) reported a significant increase in accuracy in L2 learners’ MT-assisted writing, but unclear benefits in complexity, Yang et al. (2023) reported increases in both accuracy and complexity. On the other hand, Fredholm (2019) reported increases in fluency and complexity, but not in accuracy. In addition, there is a lack of research on secondary school student writing. Most previous studies mentioned earlier examined college students’ use of MT, except for Fredholm’s studies (2015, 2019, 2021), which involved secondary school students. As more secondary school students begin to adopt MT in language learning (Hwang et al., 2022), investigating whether the use of MT is beneficial to them is imperative. Therefore, this study investigates the effect of MT use on the writing quality of EFL middle school students and their retention of learning in L2 writing. The following questions are addressed in the current study.

1. To what extent does MT use help improve the quality of L2 writing produced by EFL middle school students in terms of the complexity, accuracy, lexical diversity, and fluency?
2. To what extent do the students retain what they learn from their MT-assisted writing after two and four weeks of treatment (MT-assisted writing)?

Method

Participants

This study reports the findings of a quantitative research approach. The primary data consist of multiple versions of the writings of 59 Korean middle school students in the second grade during the first semester of a two-semester school year (M = 24, F = 35, age = 14). Their native language was Korean, and they had studied English for five years in public schools. Their English proficiency was at the beginner level (around an A2 in CEFR or below) based on the pretest. As most of them were not able to write a paragraph in English on their own, the students had taken fill-in-the-blank tests to assess their writings in the previous years when they did not use MT. However, since MT created an opportunity for the students to express their ideas about a topic, the English teacher decided to introduce MT and assigned an essay-type writing performance test.

Task Description

The students wrote four English compositions on one topic: pretest, MT-assisted version, posttest, and retention tests (two or four weeks after the MT-assisted writing), and one version in Korean on the same topic. Each version required about 120 words. For the pre-test, the students wrote for 20 minutes in English on the topic, "The person I respect" in handwriting without any additional resources (pretest). The teacher instructed the students to write the essay in English as much as possible, but if they did not know how to express sentences or words in English, they were allowed to write those sentences or words in Korean. This version was used as the pretest to measure the quality of the students’ initial L2 writing. Next, the students wrote an essay about the same topic using MT (treatment, MT-assisted version). For this, they first wrote
their essays in Korean, translated them into English using Google Translate to the full extent, compared their first version with the MT output, revised their first version based on the MT output (e.g., correcting lexicogrammatical errors, adopting newly learned English expressions and vocabulary), and submitted the revised version. Although they were allowed to use other resources, they used only MT. The teacher checked the revised versions, corrected the lexicogrammatical errors, and returned the essay to the students. The students were given one week to study their revised version as a model for their English essay. After a week, they wrote the same essay in class without MT (posttest) and wrote another essay without MT two or four weeks after the posttest (retention tests). For the retention test, approximately half of the students took it two weeks after the posttest (N = 31), and the remaining students took it four weeks after the posttest (N = 28) to further compare the delayed effect of using MT. We divided the students into two groups for the retention test in order to avoid a learning effect from the previous retention test. In addition, because they might have forgotten what they had written (content) in the MT-assisted version, the students were given the Korean text of their MT-assisted version. This method enabled the researchers to more accurately measure the effectiveness of using MT on the language rather than the content. The students were given 20 minutes each for the pre-, post-, and retention tests, and 45 minutes to write the MT-assisted revised versions. All the compositions were done under the teacher’s supervision during class time.

**Ratings**

All the versions of the English writing were scored from 0 to 5 using a holistic scoring rubric (see Appendix A for the rubric) by two trained raters, a middle school English teacher with an MA degree in English education and a researcher who works as a professor of English education. A holistic rubric was developed based on the TOEFL Writing Rubric (ETS, 2021). The present research focused more on language use in the students’ writing and did not include organization or content in the rubric, as MT use did not directly affect these areas. For example, as shown in the TOEFL writing rubric, if the essay failed or was severely limited in expressing relevant ideas due to inaccurate vocabulary and grammar, it would receive a low score (1 or 2). The average interrater reliability for scoring between the raters was 0.92.

**Data analysis**

In the current study, the four versions of the students’ English writing were analyzed in terms of the syntactic complexity, accuracy, lexical diversity, and fluency based on previous studies of MT and L2 writing. For syntactic complexity, this study measured the number of words per sentence (Lee, 2020) and the weighted distance between sentences (SYNMEDwrd) (Chon et al., 2021; Yang et al., 2023). For accuracy, error analysis in each version was conducted based on the unit of analysis: punctuation, spelling, vocabulary, and grammar. Lee (2020) argued that error analysis based on the t-unit can deflate the number of errors because L2 students tend to make more than one error per t-unit; therefore, she counted all the errors that occurred in the writing. This method was used in the current study because, especially in the writing of beginner students, more than two errors often occur in one t-unit. As the word count varied among the versions, the number of errors was nominalized by dividing by the total word count of each text to compare the versions. For lexical diversity, the number of different words per essay and the measure of textual lexical diversity (LDMTLD) were calculated (Yang et al., 2023). Last, the word count was calculated for fluency (Chon et al., 2021; Yang et al., 2023). Although previous studies with college students used more measures for complexity and lexical diversity, considering that the beginner-level EFL participants in this study produced mostly short and simple sentences with low-level vocabulary, more complicated measures were not included. The students were allowed to use Korean when they did not know how to write in English in the pretest, but Korean words were not included in the analysis.

Coh-metrix 3.0 was used to analyze syntactic complexity, lexical diversity, and fluency based on prior studies (Chon et al., 2021, Graesser et al., 2004; McNamara et al., 2014; Yang et al., 2023). For accuracy analysis, the two raters individually analyzed the total errors and frequency in each unit and discussed disagreements to reach consensus (interrater reliability = 0.91). After completing the initial analysis, a comparative analysis using inferential statistics was conducted to compare the versions (i.e., between the MT-assisted version and the posttest, between the posttest and the retention test, between the pretest and
the retention test, and between the 2-week and 4-week retention tests). IBM SPSS 26 was utilized to calculate the data at various levels. Descriptive statistics were employed to investigate the overall pattern of the data. Then, to investigate in which units (punctuation, spelling, vocabulary, or grammar) changes mostly occurred, the versions were compared, and repeated measures ANOVA was conducted to examine whether the differences among the four English versions were statistically significant. Paired t tests were also employed to determine whether there were significant differences between the versions. Mauchly’s test of sphericity was conducted in repeated measures ANOVA, and the homogeneity of the variances in a between-subject ANOVA was confirmed (> .850).

Results

The results clearly showed that there were differences among the versions in all the units measured. The results indicated that the students produced significantly better quality writing in the MT-assisted version, posttest, and retention tests compared to the pretest in terms of accuracy and fluency, as shown in Table 1 (< .001). The number of errors decreased significantly and the word count greatly increased in the MT-assisted and the subsequent versions. Concerning the score, the MT-assisted versions scored the highest, followed by the posttest, retention test, and pretest. The results for the syntactic complexity and lexical diversity showed similar patterns. It should be noted that approximately 45% of the students in the current study could not write in English in the pretest and wrote the essay mostly in Korean. Although the pretest also required 120 words, even those who could write in English produced very short texts (M = 43.00 words) with short and simple sentences (M = 3.57 words per sentence), such as “I will introduce her. Her name is Ko. She is kind. So I love her.” On the other hand, many students mixed Korean and English, shown as in the following example:

GTY is bighit music 5 team boygroups. 2019 year 3 월 데뷔하여 still 활동중이다. His 키는 177cm이며, his birthsyday is 2 월 5 일이다. (GTY is a member of a big hit boy group with 5 members. He made a debut in March, 2019. His height is 177cm, and his birthday is February 5th.)

Table 1

Descriptive Statistics and Results of the Repeated Measures ANOVA: CALF

<table>
<thead>
<tr>
<th>Item</th>
<th>Pretest</th>
<th>MT-assisted</th>
<th>Posttest</th>
<th>Retention</th>
<th>F</th>
<th>Sig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Complexity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DESSL</td>
<td>4.94</td>
<td>2.52</td>
<td>11.66</td>
<td>3.07</td>
<td>8.83</td>
<td>3.22</td>
</tr>
<tr>
<td>SYNMEDwrd</td>
<td>0.57</td>
<td>0.40</td>
<td>0.86</td>
<td>0.11</td>
<td>0.65</td>
<td>0.42</td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total errors</td>
<td>8.41</td>
<td>4.60</td>
<td>1.95</td>
<td>1.84</td>
<td>3.73</td>
<td>4.13</td>
</tr>
<tr>
<td>Lexical Diversity</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Different words</td>
<td>18.26</td>
<td>22.17</td>
<td>73.43</td>
<td>11.91</td>
<td>59.86</td>
<td>24.50</td>
</tr>
<tr>
<td>LDMTLD</td>
<td>32.85</td>
<td>22.89</td>
<td>57.81</td>
<td>18.95</td>
<td>53.41</td>
<td>20.78</td>
</tr>
<tr>
<td>Fluency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td>43.00</td>
<td>32.60</td>
<td>123.59</td>
<td>19.35</td>
<td>103.20</td>
<td>40.44</td>
</tr>
<tr>
<td>Score</td>
<td>1.61</td>
<td>.97</td>
<td>4.83</td>
<td>.70</td>
<td>3.66</td>
<td>1.66</td>
</tr>
</tbody>
</table>

Note. DESSL = the number of words per sentence, SYNMEDwrd = weighted distance between sentences, LDMTLD = measure of textual lexical diversity
Next, the current study analyzed the errors in punctuation, spelling, vocabulary, and grammar. The results showed that the students generated the most accurate text in the MT-assisted version. Regarding the unit of analysis, overall, grammar errors appeared most frequently (M = 2.17), more than twice as frequently as vocabulary errors (M = 0.98), followed by spelling (M = 0.89) and punctuation (M = 0.83) on average of the four versions per student. The number of total errors was highest in the pretest (M = 8.41), dramatically decreased in the MT-assisted version (M = 1.95) and the posttest (M = 3.73), and increased again in the retention test (M = 5.45), as shown in Figure 1. In particular, the number of grammatical errors decreased the most significantly from the pretest to the rest of the versions. A repeated measures ANOVA was conducted to examine whether there were significant differences in the unit among the versions (Table 2). The results revealed that the differences were statistically significant in all units except for vocabulary. Errors in punctuation decreased in the MT-assisted version but increased again in the later versions.

**Figure 1**

*Results of the Error Analysis*

![Graph showing error analysis results]

**Table 2**

*Descriptive Statistics and Repeated Measures ANOVA: Accuracy*

<table>
<thead>
<tr>
<th>Unit</th>
<th>Pretest</th>
<th>MT-supported</th>
<th>Posttest</th>
<th>Retention</th>
<th>$F$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Punctuation</td>
<td>1.06</td>
<td>1.73</td>
<td>.11</td>
<td>.48</td>
<td>1.18</td>
<td>2.14</td>
</tr>
<tr>
<td>Spelling</td>
<td>1.35</td>
<td>.52</td>
<td>.23</td>
<td>.65</td>
<td>.64</td>
<td>1.48</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>1.77</td>
<td>1.37</td>
<td>.56</td>
<td>1.16</td>
<td>.64</td>
<td>.84</td>
</tr>
<tr>
<td>Grammar</td>
<td>4.22</td>
<td>2.60</td>
<td>1.04</td>
<td>1.20</td>
<td>1.26</td>
<td>1.47</td>
</tr>
</tbody>
</table>

*Note.* The numbers were normalized (the number of errors/word count) for each comparison.
In addition to the repeated measures ANOVA, paired t tests were conducted to closely compare between the versions and examine the effect of using MT on the subsequent versions after the treatment (Table 3). The paired t test between the MT-assisted version and the posttest found significant differences in the word count, score, and punctuation, but not in the spelling, vocabulary, grammar. In the paired t test between the MT-assisted version and the retention test, significant differences were found in all aspects under investigation except for vocabulary. The results indicated that only vocabulary accuracy was maintained in both the post- and retention tests, while accuracy in the other units and fluency declined. However, the paired t test between the pretest and the retention tests showed that the gains were still significant except for punctuation errors; that is, the increases in fluency (word count), overall quality of writing (score), and the decreases in spelling, vocabulary, and grammar errors after the treatment (MT-assisted writing) were still maintained in the retention tests despite decreases of certain degrees.

Table 3

Results of Paired T Tests Between the Versions in Accuracy

<table>
<thead>
<tr>
<th></th>
<th>MT-assisted - Post</th>
<th>MT-assisted - Retention</th>
<th>Pretest – Retention</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t</td>
<td>p</td>
<td>t</td>
</tr>
<tr>
<td>WC</td>
<td>4.545</td>
<td>&lt;.001</td>
<td>8.931</td>
</tr>
<tr>
<td>Score</td>
<td>5.817</td>
<td>&lt;.001</td>
<td>11.873</td>
</tr>
<tr>
<td>Punctuation</td>
<td>-3.834</td>
<td>&lt;.001</td>
<td>-3.200</td>
</tr>
<tr>
<td>Spelling</td>
<td>-1.669</td>
<td>.100</td>
<td>-2.662</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.139</td>
<td>.890</td>
<td>-1.512</td>
</tr>
<tr>
<td>Grammar</td>
<td>-.733</td>
<td>.467</td>
<td>-3.240</td>
</tr>
</tbody>
</table>

The present study also examined whether there was a difference between the two-week and four-week retention tests by using the paired t test (Table 4). A statistically meaningful significance appeared only for grammar. Although there were differences found in other areas between the two retention tests, the differences were statistically insignificant. The results indicated that grammatical errors increased significantly between the 2-week and 4-week retention tests, but other areas remained quite steady.

Table 4

Descriptive Statistics and T Test Results of 2- and 4-week Retention Tests

<table>
<thead>
<tr>
<th>Item</th>
<th>2-week</th>
<th></th>
<th>4-week</th>
<th></th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word count</td>
<td>82.52</td>
<td>41.63</td>
<td>65.00</td>
<td>47.97</td>
<td>1.483</td>
<td>0.143</td>
</tr>
<tr>
<td>Score</td>
<td>2.81</td>
<td>1.47</td>
<td>2.25</td>
<td>1.50</td>
<td>1.453</td>
<td>0.152</td>
</tr>
<tr>
<td>Punctuation</td>
<td>1.19</td>
<td>3.16</td>
<td>1.22</td>
<td>2.62</td>
<td>0.045</td>
<td>0.964</td>
</tr>
<tr>
<td>Spelling</td>
<td>.81</td>
<td>1.42</td>
<td>1.41</td>
<td>1.82</td>
<td>1.406</td>
<td>0.165</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.94</td>
<td>1.52</td>
<td>1.00</td>
<td>1.21</td>
<td>0.172</td>
<td>0.864</td>
</tr>
<tr>
<td>Grammar</td>
<td>1.47</td>
<td>2.03</td>
<td>3.00</td>
<td>3.28</td>
<td>2.19</td>
<td>0.033</td>
</tr>
<tr>
<td>Error total</td>
<td>4.41</td>
<td>5.95</td>
<td>6.63</td>
<td>6.57</td>
<td>1.364</td>
<td>0.178</td>
</tr>
</tbody>
</table>
Finally, to show in more detail how using MT helped improve student writing, a representative sample set of one student’s writing is presented in this section. The students often used Korean for words and expressions that were beyond their ability in the pretest, but they were able to write those expressions in English with MT. Similarly, most of them were able to write sentences similar to the MT-assisted version in the posttest; however, many of the newly learned components decayed, and some of the students returned to using Korean in the retention tests. The sample in Appendix B shows a common pattern appearing in the students’ writings collected in this study. The student in this sample barely wrote the essay in English in the pretest, in which she mixed Korean and English to deliver her message. As shown in Appendix B, she used Korean not only for words (e.g., 인성 [good personality], 센스 [sense], 유머감각 [sense of humor]) that she did not know, but also for the grammatical components. For example, she used “이라면 [if]” in the last sentence, and this postposition signifies a conditional clause in Korean. Thus, she wanted to write, “if a person is Korean, the person would know him”; however, she did not know how to make a conditional clause in English and used Korean instead. Her writing was short, comprised of very short and simple sentences, which contained a number of punctuation, spelling, and lexicogrammatical errors. In contrast, her writing with MT significantly improved in fluency, accuracy, and complexity. The word count of the pretest was 39, but it greatly increased to 116 in the MT-assisted version. Although she made several lexicogrammatical errors in the MT-assisted version, the errors were local errors (e.g., he haven’t, this years), and her writing conveyed the meaning much better. Moreover, there were no incomprehensible sentences found in the MT-assisted version. She also used correct English words, “good personality,” “sense of humor,” and “wit” equivalent to the Korean words, 인성, 센스, and 유머감각, respectively, which were written in Korean in the pretest. She was also able to correct the grammar as well as the spelling, as shown in the following sentence: “The person I respect the most is YJS” from “My repect person is YJS.”

In the posttest, the word count (112) was similar to that of the MT-assisted version, and the student was able to write most of what she learned from the MT-assisted version. For instance, she still remembered the corrected sentence, “The person I respect the most is YJS,” and the newly learned vocabulary, “personality” and “sense of humor.” Moreover, the sample shows that the student could still use some sentence structures (e.g., The biggest reason I respect him is his good personality; The thing I want to learn the most from him is a sense of humor so that people around me feel happy), expressions (e.g., cope with, neighbors in need), and vocabulary (e.g., support, situation, mistakes), which were beyond her grade level of English in Korea. However, the retention test revealed that what she learned from the MT-assisted version significantly declined, and fluency, accuracy, and complexity also decreased. The word count decreased to 77, and her writing contained many spelling and lexicogrammatical errors. In particular, she could not correctly write the sentences with more complicated structures, such as embedded sentences, in the retention test (e.g., The thing I want to learn the most from him is a sense of humor so that people around me feel happy). As she could not remember how to make a when-clause (“he helped his juniors when they made mistakes”), she used Korean instead. In terms of punctuation, she missed periods and made mistakes in capitalization. The sample also showed that she did not know how to form a paragraph. She seemed to copy the paragraph format from the MT output, but she was unable to maintain it in the post and retention tests when MT was no longer available.

**Discussion**

The present study showed that the use of MT improved diverse aspects of writing, including fluency (measured by the word count), accuracy (measured by the number of total errors per essay and the number of errors in punctuation, spelling, vocabulary, and grammar), syntactic complexity (measured by the number of words per sentence and weighted distance between sentences), lexical diversity (measured by the number of different words used and the measure of textual lexical diversity), and the delivery of meaning. The number of errors was also significantly reduced in all units of punctuation, spelling, vocabulary, and grammar. The repeated measures ANOVA confirmed that the differences in scores
between the versions were statistically significant. Most prominently, MT allowed the students to write what they wanted to tell. As Peña and García (2011), Chang et al. (2022), and Tsai (2022) have claimed, the beginner students in the current study definitely benefited from using MT. While they could rarely express their ideas in English in the pretest and produced only a few short and simple sentences with errors, they were capable of writing their ideas in English with the help of MT. In previous years, the students took fill-in-the-blank tests for L2 writing because they were often unable to write an English essay on their own. The current study showed that 45% of the students could not write the essay in English in the pretest, but they could do so with the help of MT. These students would otherwise not be able to express their ideas in English, as the teacher could not translate all the essays into English for them.

The current study showed that the MT-assisted version was significantly improved in terms of fluency, accuracy, and complexity, and as a result, the students’ writing with MT was of higher quality. However, considering their low-level English proficiency, it is assumed that the students directly borrowed the MT output without critically thinking about it. This finding is consistent with L2 writing studies that have claimed that learners’ attitudes and strategies towards corrective feedback and revision vary depending on learner variables, such as language proficiency, motivation, and engagement (Chen et al., 2015; Tang & Liu, 2018; Zhang & Cheng, 2021). Particularly with MT, which does not offer one-on-one feedback on errors as the teacher does, low-level and less confident students often struggle to figure out what and how to correct their errors by comparing them with the MT output (Lee, 2022b). Without further explanations, low-level students are not able to understand the MT output and learn from it. Therefore, in the current study, although the students performed better in the posttest than in the pretest, it is fair to say that the students might have directly adopted the MT outputs, memorized them, and written essays in the posttest. Then, the question was whether they learned new language items from MT or remembered what they had learned from MT for a longer term. The results of the repeated measures ANOVA and t tests indicated that fluency, accuracy, and complexity decreased in the later versions, but the retention tests still scored twice as high as the pre-test in terms of fluency, accuracy, the number of different words and the number of words per sentence. Although we cannot rule out the possibility that the students’ English improved over the period for other reasons, it is difficult to assume that the students’ writing improved to such an extent over the two or four weeks (i.e., the students had two 45-minute English classes per week) without taking into account the impact of MT use.

In particular, the study showed that vocabulary remained quite steady in the later versions. Although there were slight decreases in vocabulary, they were not statistically significant, and the number of different words was still significantly higher in the retention test (M = 47.20) than in the pretest (M = 18.26). Prior studies have argued that lexical errors are easier to correct with MT than syntactic errors because vocabulary usually stands out more than sentence structures when comparing students’ own writing with MT outputs (Lee, 2022a; Niño, 2008). While it is difficult to understand a new grammatical element without further explanations, understanding or acquiring new vocabulary is easier with MT because MT presents the words in the L2 equivalent to the words in the L1. Moreover, while understanding a new grammatical feature requires understanding a string of words, learning vocabulary requires understanding just one segment in the sentence. For these reasons, the students in the current study seemed to remember new vocabulary better in the retention tests.

On the other hand, the current study revealed that there was a significant decrease in grammar. Both the repeated measures ANOVA and the paired t tests between the two retention tests showed that there were significant differences in the grammatical errors among the versions. Smith (1991, 1993) proposed three reasons why learners fail to grasp the target grammatical features presented in the input. First, L2 learners are generally not sensitive to the target language input. Second, certain grammatical elements in the language input are not noticeable to learners. Third, the learner’s L1 can impede his or her ability to detect certain linguistic features in the language input. Therefore, according to Smith (1991, 1993), due to learners’ low perceptual salience and/or the noticeability of grammatical elements, even with the language input (i.e., MT) present to them, students are not often able to learn the grammatical elements. As such,
the students in the current study may not have paid attention to the new grammatical elements shown by MT while copying and memorizing them so that grammar slipped away quickly. In particular, as shown in the sample writing set, grammar items that were beyond the students’ current level, such as embedded sentences, waned more frequently. Even though they were able to reproduce the sentences with new grammatical elements in the posttest, it seemed that they did not actually understand and internalize the grammar; thus, they could not write those sentences in the retention tests.

Concerning the second research question, to what extent the students retained what they learned from their MT-assisted writing, the present study concluded that the students remembered approximately half of what they learned from MT and that using MT had a positive impact on L2 learning over an extended period of time. Despite declines in accuracy, fluency, and complexity, the word count and the score on the retention tests were still significantly higher than those on the pre-test. The number of different words used and the number of words per sentence remained much higher in the retention test. More importantly, without MT, the students practiced mainly controlled writing (e.g., fill-in-the-blanks) in previous years, but with the help of MT, they could extend L2 writing to express their ideas. This, in turn, is expected to have a positive impact on the students’ confidence in L2 writing, as Fredholm (2021) argued that a more important insight of MT use was the increase in students’ self-esteem and confidence as L2 writers.

The results of the current study present a mixed picture when compared with previous MT studies. On the one hand, this study was consistent with prior studies that confirmed increases in complexity, accuracy, and fluency in MT-assisted student writing (e.g., Chang et al., 2022; Yang et al., 2023). On the other hand, while Chang et al. (2022) showed no increase in the word count with MT, the word count in the MT-assisted version in the present study increased significantly. This discrepancy may be due to the difference in the level of language proficiency of the subjects in the two studies. While the subjects in Chang et al.’s study (2022) were university students, the subjects in the present study were beginner-level middle school students, many of whom were unable to write sentences in English. Therefore, the students in the present study could write much longer responses only when they were aided by MT. The current study was consistent with Fredholm’s studies (2015, 2019, 2021) in that his studies and the current study confirm increases in complexity, fluency, and lexical diversity with the use of MT. However, accuracy did not improve with MT in his study (2015), and the positive effect of MT on lexical diversity disappeared when MT was not available (2019), which differed from the current study. In our work, accuracy increased significantly in MT-assisted writing and the posttest compared to the pretest, and although the impact of MT on vocabulary diminished in the later version, this study showed some lasting effects of MT use on vocabulary. Several explanations are possible for the discrepancy in accuracy between this study and Fredholm (2015), such as students’ different strategies for using MT, different accuracy levels of MT outputs between the years of the two studies (2015 and 2023), or different language pairs (Spanish–Swedish vs. English–Korean). Finally, regarding the effect of MT use on lexical diversity, students wrote essays on different topics in the pre- and posttests in Fredholm (2015), whereas the students in this study wrote about the same topic, which may have led to higher vocabulary retention.

Based on the results, the current study has the following pedagogical implications. As Tsai (2022) claimed, using MT can serve as an effective translingual CALL tool and bring pedagogical benefits to the L2 classroom by providing initial assistance to L2 writing and revision. As shown in the present study, MT enables students to use enriched vocabulary and more complex syntactic structures with fewer lexicogrammatical errors. However, since merely using technology does not guarantee learning (Chen et al., 2015), teacher intervention is necessary. While MT can assist student L2 writing and reduce the teacher’s workload in giving corrective feedback to a certain extent, MT cannot totally replace teacher feedback. As shown in this study, it is difficult to self-learn grammatical elements from MT outputs, particularly for beginner learners; thus, a follow-up Q&A session in which students and the teacher talk about student writing, MT outputs, and new grammatical elements would certainly be beneficial to develop accuracy and advance students’ L2 writing. Because MT does not provide corrective feedback targeting specific errors, students may be left confused, and the teacher’s explanations can clear the confusion and consolidate L2 learning by providing additional feedback complementary to MT.
Chang et al. (2022) reported that teachers’ corrective feedback can cause students’ anxiety, but MT is impersonal and cannot provide positive comments to encourage students. A teacher's ability to motivate students to learn L2 cannot be replaced by MT. Moreover, the comprehensive corrective feedback of MT can be too overwhelming and discouraging to students (Koltovskaya, 2022). It is suggested that teachers utilize MT as an overview of frequent errors in a student’s writing and a resource for more focused feedback. Furthermore, writing is more than just reaching accuracy or complexity. It requires logical thinking, content, and organization of ideas that only a human teacher, not MT, can provide. Studies have also shown that variables, such as student proficiency in L1 (source language) and L2 (target language), confidence in L2 (Groves & Mundt, 2015) and motivation and engagement toward L2 learning, affect learning with MT (Tang & Liu, 2018; White & Heidrich, 2013; Zhang & Cheng, 2021). Thus, the teacher needs to consider these variables and incorporate them during a feedback session.

In a similar vein, student training on using MT can also help them effectively utilize MT for their purposes. O’Neill (2019) showed that an MT group with training outperformed both an MT group without training and a non-MT group in L2 writing. This study implies the need for training when using MT for language learning. Training can include how to correctly write L1, critically adopt the MT output, and crosscheck the MT output with other reference resources, rather than viewing MT as a complete language learning source. Teaching strategies to use MT, such as back translation (from the target language to the source language), preediting (e.g., strategic use of the source language), and postediting strategies, can help students effectively use MT (Niño, 2004; Shin & Chon, 2023). Most importantly, the teacher should help students increase their autonomy and become successful and confident L2 writers (Fredholm, 2021). One-time use of MT during writing practice will not have a significant effect on writing development or writing proficiency; therefore, teachers need to develop a long-term MT-assisted L2 writing curriculum to facilitate student writing development.

The present study has the following limitations. First, the students’ writings were examined four weeks after using MT, but this period may not have been long enough to measure the longer-term effect of using MT. At the same time, this study examined the delayed effect of the MT use, and other teaching and learning effects over this period were not considered. The students in the current study were required to write the same text in the posttest and perform a retention test in order to more directly assess the effect of using MT. However, the study did not investigate whether the learning transferred to a subsequent text on a different topic. Therefore, further studies should investigate the effect of more prolonged use of MT on L2 writing and language learning (i.e., over longer periods of time, the effects of repeated use of MT) in order to examine whether language learning from MT is internalized and transferred to a new text.

Finally, we allowed the students to use Korean in the pretest to support their writing, but this skewed the results for the syntactic complexity and lexical diversity because the results did not include the writing of the students with the lowest English proficiency.

**Conclusion**

The current study showed that MT helped students deliver their meaning, reduce grammatical errors, find appropriate vocabulary, and use expressions and sentence structures beyond their current levels. As a result, compared with the pretest, the students generated much higher quality writing in the MT-assisted writing and the posttest in every aspect of writing. The significance of the study was to examine the effect of using MT on middle school students’ L2 writing proficiency in terms of the complexity, accuracy, lexical diversity, and fluency, and how much they remembered what they learned from MT. The present study found that using MT was effective to some extent for a longer period of time, but depending on linguistic items, the effectiveness varied.

Technologies such as AI and machine learning are significantly affecting L2 learning and teaching. For instance, Groves and Mundt (2015) and Briggs (2018) reported the reluctance of language teachers and students to use MT for language learning purposes. Briggs (2018) and Tsai (2019) warned that merely banning or ignoring MT in the language classroom would not be beneficial for learners. However, these
issues have rarely appeared in most recent studies. More language classrooms are now embracing this technology (Chang et al., 2022; Hwang et al., 2022), and MT has become omnipresent in the language classroom (Tsai, 2022). This technology has reshaped the perceptions and attitudes of language teachers and learners. Recently, ChatGPT has introduced a critical discussion to education. In the near future, more technologies may unexpectedly enter the language classroom. Amid the rapid development of technology, language teachers should carefully gauge and explore the potential benefits and drawbacks of using technology to maximize student learning.

References


Appendix A. EFL Writing Evaluation Rubric

0 – The student could not write the essay in English at all;
1 – The student could barely write the essay in English;
2 – The essay contained many major errors, which seriously undermined the meaning and overall quality of writing, including many incomprehensible sentences;
3 – The essay contained many serious major errors, but delivered the meaning;
4 – The essay contained a few major errors with several minor errors;
5 – The essay contained a few minor errors, which did not damage the meaning of the sentence
Appendix B. Student Writing Sample

Pretest

My respect person is YJS. He is Korea mc, comidan, singer. He is 50 years. Yoo is Korea person 이라면 know. He is very famous. I respect him. Because his 인성 is very good, and his 센스, 유미감각 is cool.

MT-supported version

The person I respect the most is YJS. He is Korean and his job is TV entertainer he turns 50 this years. The biggest reason I respect him is his good personality. He haven't made a mistakes in 30 years, and before becoming a celebrity, he helped and supported neighbors in need. Also, he has a good sense of humor and wits to cope with situation. In variety shows, he helped his juniors when they made mistakes, and there were many times when he moved on with a sense of humor. The thing I want to learn the most from him is a sense of humor so that people around me feel happy. Although she made many errors in the recall test, she still recalled some words or spellings that she learned, such as respect, support, and sense of humor as well as a few sentences.

Posttest

The person I respect the most is YJS.
He is Korean and his job is TV entertainer. He turns 50 this year.
The biggest reason I respect him is his good personality
He haven’t made mistakes in 30 years. Before becoming a celebrity, he helped and supported neighbors in need. Also, he has a good sense of humor and wits to cope with situation in variety shows. He helped his juniors when they made mistakes and there were many times when he moved on with a sense of humor
The things I want to learn the most from him is sense of humor So that people around me feel happy

Retention test

My person respect most is YJS. He is Korean and His job is TV entertainer. He is 50 this year I respect most him because he has very good personartiy. He never made mistakes in 30 years before he helped and support neibors he famous. Also he has good sense of humor. So in Vriate show, he helped his mistake 후배가 실수할때 I think earn to most form him is everyone happy sense of humor.

Note. All the versions are original, including the language errors and paragraph formatting. All versions were originally handwritten.
About the Authors

Sangmin-Michelle Lee is Professor of Global Communication at Kyung Hee University in Korea. She is currently working on developing an AI system to assist teachers. She has published extensively on language learning in immersive virtual reality, machine translation, L2 writing, game-based learning, and digital creativity. Her work has appeared in *LLT, BJET, CALL, ReCALL, Foreign Language Annals*, and *IEEE Transactions on Learning Technologies*, among others. Sangmin-Michelle Lee is the corresponding author.

E-mail: sangminlee@khu.ac.kr

ORCiD: https://orcid.org/0000-0002-7686-3537

Nayeon Kang has been teaching English for about 20 years at secondary schools in South Korea. She holds a BA in English Education from Ewha Womans University and a MA in TESOL from the University of Edinburgh, UK. Her research interests include the application of Artificial Intelligence (AI) in English education and the utilization of various educational technology tools for facilitating English language learning.

E-mail: nayeonkang24@gmail.com