ReCALL (2023), 35: 2, 211–224
doi:10.1017/S0958344023000022

RESEARCH ARTICLE

Assessing the effectiveness of machine translation in the Chinese EFL writing context: A replication of Lee (2020)

Yanxia Yang
Nanjing University, China; Nanjing Agricultural University, China (yanxiayang@njau.edu.cn)

Xiangqing Wei
Nanjing University, China (weixq@nju.edu.cn)

Ping Li
Nanjing Agricultural University, China (liping5110@njau.edu.cn)

Xuesong Zhai
Zhejiang University, China (xszhai@zju.edu.cn)

Abstract
With the dramatic improvement in quality, machine translation has emerged as a tool widely adopted by language learners. Its use, however, has been a divisive issue in language education. We conducted an approximate replication of Lee (2020) about the impact of machine translation on EFL writing. This study used a mixed-methods approach with automatic text analyzer Coh-Metrix and human ratings, supplemented with questionnaires, interviews, and screen recordings. The findings obtained support most of the original work, suggesting that machine translation can help language learners improve their EFL writing proficiency, specifically in strengthening lexical expressions. Students generally hold positive attitudes towards machine translation, despite some skeptical views regarding the values of machine translation. Most students express a strong wish to learn how to effectively use machine translation. Machine translation literacy instruction is therefore suggested for incorporation into the curriculum for language students.

Keywords: machine translation; EFL writing; Coh-Metrix; human ratings; replication

1. Introduction

The quality of machine translation (hereafter MT) has significantly improved with technological advancements in recent years. Improving quality and free accessibility have led to the presence of MT in foreign language learning activities (Briggs, 2018; Yang & Wang, 2019), particularly in EFL writing (e.g. Chung & Ahn, 2022; Lee, 2020). More recently, studies suggest that MT can help students write more fluently and accurately (cf. Tsai, 2019), as well as gain confidence and motivation (cf. Lee, 2023; Lee & Briggs, 2021). It seems as if MT has become a tool for EFL learning. However, the performance of MT may fluctuate depending on different MT systems, text types, and language pairs (Daems, Vandepitte, Hartsuiker & Macken, 2017). MT on the one hand is used as a convenient tool for different learning purposes by a number of students, while on the other hand it is seen as a disservice to language learning by some teachers and

Cite this article: Yang, Y., Wei, X., Li, P. & Zhai, X. (2023). Assessing the effectiveness of machine translation in the Chinese EFL writing context: A replication of Lee (2020). *ReCALL* 35(2): 211–224. https://doi.org/10.1017/S0958344023000022

educators (Klekovkina & Denié-Higney, 2022). The omnipresent use of MT and its effects have raised educational concerns (Ducar & Schocket, 2018).

Most prior studies have focused on the effect of MT on language writing and translation. One of the outstanding examples is Lee (2020), an exploratory study on the impact of MT on EFL students' writing. The results demonstrate that MT can help improve students' writing quality. However, due to the imperfections and fluctuations of MT quality, as well as the varieties of research samples, the effectiveness of MT use varies in different situations. Studies are still required to confirm the effectiveness of MT in EFL writing. Replication is considered to be an important approach to test the robustness of the original research. To this end, the purpose of the present study is to construct an approximate replication (Porte & Richards, 2012) of Lee's work (2020) to further verify the impact of MT on EFL writing while allowing for minor changes to the methodology of the original study.

2. Literature review

2.1 Machine translation in language education

Online MT is a free language resource accessible to all users (Ducar & Schocket, 2018; Niño, 2020). The use of MT is largely associated with the quality of MT (Lee, 2023). In the early years, MT quality was far from satisfying. Its output was often used as a resource for error correction and revision practice (La Torre, 1999; Niño, 2008; Somers, 2004). More recently, with continuing quality improvement, MT has served as a potential reference for students' learning queries (Briggs, 2018; Stapleton & Kin, 2019). However, the current MT quality is not perfect. Frequent use of imperfect and defective language may possibly generate unwanted habits for language learners (Bowker, 2020; Yamada, 2019).

The controversial issue of using MT has drawn attention to MT's effectiveness. For instance, Garcia and Pena (2011) find that MT may help second language (L2) writing beginners to communicate more and better. It is also found that MT can help students reduce lexico-grammatical errors (Lee, 2020; Tsai, 2019) and increase lexical accuracy and complexity (Fredholm, 2019; Kol, Schcolnik & Spector-Cohen, 2018). Furthermore, MT can narrow the writing ability gap between skilled and less skilled learners (Chon, Shin & Kim, 2021). Regarding the perception of using MT, some students exhibit positive attitudes, whereas other students are skeptical about the values of MT (Dorst, Valdez & Bouman, 2022; Lee, 2023). Teachers are either reluctant to use or are ignorant of the MT application in foreign language education (Briggs, 2018). Tsai (2019) believes that MT use is not equivalent to language acquisition, given the fact that most MT users do not intentionally memorize the lexical and syntactical information for learning purposes. Yamada (2019) suggests a careful incorporation of MT into language learning. There is a pressing need for instruction for users to optimize the use of MT (Kenny, 2022). In this regard, MT literacy instruction is proposed to help students know "how MT works, how it can be useful in a particular context, and what the implications are of using MT for specific communicative needs" (O'Brien & Ehrensberger-Dow, 2020: 146). Altogether, proactive efforts are required to accumulate evidence that confirms the effectiveness of MT.

2.2 Research to be replicated

Replication is a vital approach to verify prior findings, advance understanding of methodological practice, and consolidate the generated knowledge (McManus, 2022). The study to be replicated here was published by Sangmin-Michelle Lee (2020) in *Computer Assisted Language Learning*. Lee has published a series of papers about the effect of MT on language learning. In this study, Lee investigated the impact of MT on Korean college students' EFL writing. The methodological procedures comprised three steps. First, students wrote essays in Korean and then manually

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

translated them into English. Second, they got the English version of their Korean essays with the help of Google Translate. Finally, they revised their manual translations by referring to the MT output. Lee's findings suggested that MT could facilitate writing strategies use and help to decrease lexico-grammatical errors. Students generally held positive attitudes towards using MT.

The decision to replicate Lee's (2020) work was based on the following reasons. First, the increasing popularity of MT among language learners has drawn attention from academia to investigate its benefits and limits as a pedagogical tool (e.g. Jolley & Maimone, 2022; Kelly & Hou, 2022). It is necessary to replicate research about the impact of MT on EFL writing to confirm the effectiveness of MT. Second, replication studies on EFL writing remain scarce (Porte & Richards, 2012). EFL writing can be influenced by a corrective feedback style in the CALL environment (e.g. Brudermann, Grosbois & Sarré, 2021). This is particularly true for studies involving MT in EFL writing, since the MT quality may vary in different language pairs. Third, having different populations in replications is helpful to increase the generalizability of the original findings (Handley, 2018). In Lee's study, Korean EFL learners were the research sample. Chinese EFL learners were considered in the present study, a new but roughly comparable sample to the original one. English is taught as the foreign language for both Korean and Chinese students. A replication with a comparable sample is expected to increase the study's explanatory power on MT's effectiveness in EFL writing.

3. Method

Following replication guidelines (Porte & McManus, 2019), the authors of the present study conducted an approximate replication in the Chinese context. Lee's (2020) research design and experimental procedures were rigidly followed to contextualize variables of MT and EFL writing. A comparable sample of Chinese EFL learners was invited in an EFL writing task. To accurately gauge the impact of MT on EFL writing, objective and subjective measures were taken to triangulate the quality assessment, including human rating and automatic text evaluation. The automatic text analyzer Coh-Metrix was employed for fine-grained analysis. Screen recordings were utilized to watch the writing revision process. Semi-structured interviews were conducted to investigate students' perceptions in using MT. Thus, it is hoped that this study makes a significant contribution to assess the feasibility and applicability of MT in EFL writing.

3.1 Research questions

The research questions are formulated as follows:

1. Are there any writing quality differences between Chinese EFL learners' first manual version and their revised version with the help of MT?
2. If so, how and to what extent does MT have facilitating effects on Chinese EFL learners' writing?
3. How do Chinese EFL learners perceive and evaluate the MT use in their EFL writing process?

3.2 Participants

Thirty-one first-year English majors were invited to participate in the present study. Their ages ranged from 18 to 20 ($M = 19.06$, $SD = 0.63$). All students were Chinese native speakers and took English as a foreign language. They enrolled in a 1.5-hour English writing course per week for 18 weeks. Most of them had passed College English Test Band 4 (CET-4), a nation-wide English proficiency test serving as a benchmark for English language teaching and learning in China (Zheng & Cheng, 2008). In that regard, they were considered at the lower-intermediate level

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

of English proficiency. With respect to MT use frequency, 87% of students reported that they frequently used MT in daily language learning activities. Participants' demographic information in the present study roughly corresponds to that in Lee's work, which is helpful in minimizing the impact of individual variables on the results.

Further, we carefully determined the sample size before initiating the study to validate the statistical power. The required sample size was calculated by using G*Power (G*Power 3.1.9.7; Faul, Erdfelder, Lang & Buchner, 2007). A test power of $1\text{-}\beta = 0.80$ and significance level at $\alpha = 0.05$ were considered to disregard the smaller differences or effect. The recommended minimum sample size was suggested as 28. In the event of any risks of invalid data or withdraw cases, the sample size of 31 participants was considered basically adequate and appropriate.

3.3 Task description

The experiment was carried out at three stages in the classroom setting. At the first stage, participants were asked to write a descriptive essay in Chinese of around 350 words on a given topic within 30 minutes. The word count and time limit were controlled with reference to CET-4 requirement (Jin & Yang, 2006). At the second stage, students were required to manually translate their initial Chinese texts into an English-language version within 30 minutes. The manual translations were hereafter regarded as T1. At the third stage, participants were first required to use Google Translate[1] to translate the initial Chinese essays into English, as in Lee's (2020) study. They were then instructed to revise their manual translations by referring to the MT version and generate the final version of T2 in 20 minutes. Digital resources were allowed. It is helpful to create a quasi-real situation for EFL writing. Additionally, it is hoped that participants can be motivated to produce quality texts by searching maximum background information at the drafting stage. Screen recording was used to capture the real-time process for in-depth data analysis, such as the detail revision process with the aid of MT. In addition, the recording data is also helpful to detect any possible cheating behaviors during the whole experimental process.

3.4 Data collection and analysis

This study adopted a mixed-methods approach to obtain qualitative and quantitative data for in-depth analysis. Coh-Metrix, an automatic text analyzer, was used for text feature analysis. It is a robust tool, which can help researchers "acquire a deeper understanding of language-discourse constraints" (Graesser & McNamara, 2011: 372). Coh-Metrix offers a broad range of features that are useful for discriminating between high- and low-quality writing. These features have been widely used in writing assessment studies (e.g. Crossley, Salsbury, McNamara & Jarvis, 2011; Latifi & Gierl, 2021). Based on Coh-Metrix Version 3.0 Indices,[2] five indicators of lexical, syntactical, and textual features were carefully selected, as presented in Table 1.

Prior research showed that successful writers tended to produce linguistically longer texts (McNamara, Crossley & Roscode, 2013). In this regard, three indicators of lexical features were considered, including DESWLlt, LDTTRc, and PCCNCz. DESWLlt refers to the average number of letters in words. Lexical diversity is a significant predictor of writing quality (Wiley *et al.*, 2017). LDTTRc was adopted here since it is an important indicator for lexical diversity analysis in the Coh-Metrix framework. Moreover, PCCNCz, an indicator for text easability component scores, was considered to assess word concreteness and meaningfulness. In terms of syntactical features, SYNMEDwrd is an indicator for syntactic complexity. It is believed to correlate with measures of referential and semantic cohesion (McNamara, Graesser, McCarthy & Cai, 2014). The textual feature CNCCaus is the casual connectives incidence. According to Crossley and McNamara

[1]https://translate.google.cn

[2]http://cohmetrix.memphis.edu/cohmetrixhome/documentation_indices.html

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

Table 1. Selected text features for the writing assessment

Features	Description	Indicators
Lexical	Word length	DESWLlt
	Word diversity	LDTTRc
	Word concreteness	PCCNCz
Syntactical	Syntactic complexity	SYNMEDwrd
Textual	Causal connective	CNCCaus

Table 2. Writing quality assessment rubrics

Components	Description	Score weight
Content	Knowledgeable, substantive, and relevant to the assigned topic	30%
Organization	Fluent expression, logically sequenced, and cohesion in writing	20%
Vocabulary	Effective word choice and usage, word form mastery, and appropriate register in writing	20%
Language use	Effective complex constructions, agreement, tense, word order, and prepositions in writing	25%
Mechanics	Mastery of conventions like spelling, punctuation, capitalization, and paragraphing	5%

(2009), causality is important for constructing relations between events and actions. Causal connectives could discriminate the high cohesion texts from the low cohesion texts (McNamara *et al*., 2014).

Human ratings were used to triangulate the results generated by Coh-Metrix using the ESL Composition Profile (Jacobs, Zingraf, Wormuth, Hartfiel & Hughey, 1981). This assessment framework is a widely used analytic scale in writing research. The grading rubrics consist of content, organization, vocabulary, language use, and mechanics. Descriptions of the rating rubrics are presented in Table 2. The rubrics have been extensively tested in EFL writing assessment (e.g. Lam & Pennington, 1995; Liu & Brantmeier, 2019). Two highly proficient raters were invited to separately grade the text quality of T1 and T2. They were instructed to attend primarily to text features of content, organization, vocabulary, lexical use, and mechanics. Disagreements in scoring were resolved by discussions or by consultations with a third rater. R language was used as the statistical tool for data analysis. Descriptive statistics, paired sample *t*-test, and correlation analysis were considered to explore the quality differences between the initial manual version (T1) and the final revised version (T2).

Finally, questionnaires and interviews were administered to participants after they had completed the whole writing test. They were conducted to collect participants' opinions on MT use and identify specific difficulties during the writing process. The questionnaire included demographic information of participants' age, computer use experience, MT use experience, language proficiency, self-assessment on Chinese and English writing ability, as well as the perception of using MT in the writing revision process. Semi-structured retrospective interviews were conducted with six students. The purpose of the interviews was to collect information about writing self-assessment and to elicit interviewees' points of view on MT use, as well as their expectations of using MT in writing activities. The interview questions were developed with reference to the interview items in Lee's (2020) work. Compared to the original, wordings of interview items in

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

Table 3. Comparisons between the present and the original study

Study	Participants	English proficiency	MT	Research tools
Lee's work	34 Korean students majored in English	Intermediate 70–95 TOEFL	Google	Mixed methods: quality assessment, interview, reflection paper
Present study	31 Chinese students majored in English	Lower-intermediate CET 4	Google	Mixed methods: quality assessment, questionnaire, interview, screen recording

the present study were slightly adjusted in view of the research purpose. Additional items (1–4 below) were added to investigate students' expectations of using MT:

1. What is your writing ability like?
2. What are the advantages and disadvantages of using MT during revision?
3. What's your perception of using MT in the writing process?
4. What's your expectation of using MT in the writing process?

A peer interviewer who had received adequate training carried out the interview. The interview was conducted separately, lasting about 15 minutes for each interviewee. Mandarin was used in an attempt to increase interviewees' comfort and induce their real thoughts as much as possible. The interviews were recorded, transcribed into texts, and finally categorized into separate files.

Apart from the questionnaires and interviews, screen recording was adopted to understand the individual trajectory during the writing process. Students' screens were captured by BB FlashBack, a compact and easy-to-use recording tool. The recording data is like a multimodal corpus, containing timeline and specific revising moves. The data were viewed repeatedly and then transcribed by the authors, before being carefully coded, based on Enríquez Raído's (2014) coding schema, which is a time frame of behaviors and query related to the search information. Selected segments analysis was guided by moment analysis, an approach aiming to capture the individual and their cognitive processes surrounding the critical moment of action (Li, 2011).

Successful replications rely on transparency in terms of methodology (Tschichold, 2019). Hence, we have developed a comparative framework to have a rounded view on the research design between Lee's (2020) work and the present study. The framework included the number of participants, participants' majors, English proficiency, MT systems, and research tools (see Table 3). Both studies were based on a small sample size of around 30 students. Students were EFL learners who took English as their majors. Their English proficiency was approximately comparable at the intermediate levels. Google MT systems were used in the two studies. Language pairs were set as Korean to English in Lee's work and Chinese to English in the present study. Both studies used a mixed-methods approach with writing quality assessment and interviews. The present study also conducted objective automatic ratings with Coh-Metrix and adopted screen recordings to gather multimodal process information, while Lee used a reflection paper. With different language pairs and different language students, the present study can not only serve as a replication study but also more importantly contribute to a better understanding of the MT's effect on EFL writing in the Chinese context.

4. Results

4.1 Quantitative data analysis

4.1.1 Writing quality assessment by Coh-Metrix

A descriptive analysis was conducted on the selected features and indicators between T1 and T2. T1 was the initial manually translated version of Chinese writing. T2 was the final revised version

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

Table 4. Descriptive analysis of selected text indicators

	Word length		Word diversity		Word concreteness		Syntactic complexity		Causal connectives	
Text	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
T1	4.54	0.24	0.78	0.07	−0.13	0.70	0.85	0.12	94.55	16.91
T2	4.60	0.25	0.79	0.05	−0.05	0.68	0.89	0.13	99.99	15.73

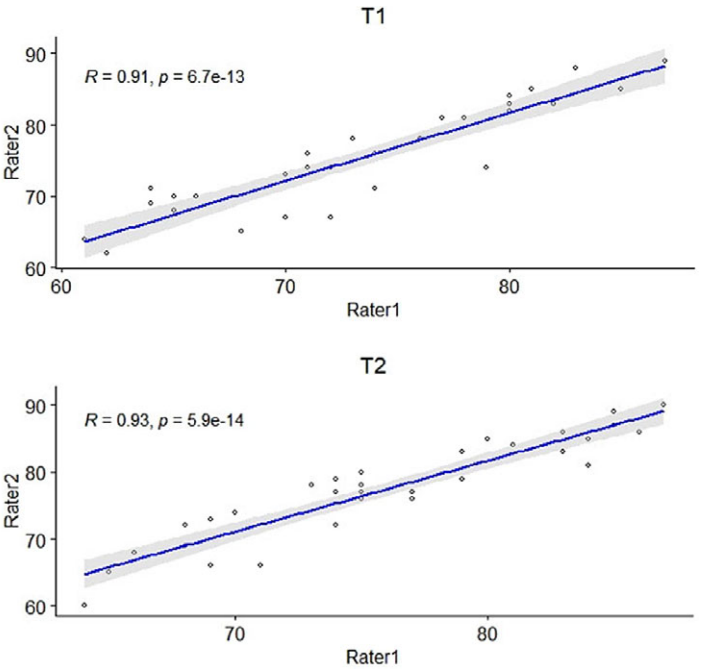

Figure 1. Correlation coefficients of raters between T1 and T2

of T1 with reference to the MT output. As is shown in Table 4, the mean scores of the listed indicators in T2 were generally higher than the scores in T1, especially for word length and causal connectives. The descriptive analysis has suggested that the writing quality of T2 was improved in comparison with T1.

To further examine whether there was a significant difference between the text features of T1 and T2, we conducted mean value comparisons for paired samples. Shapiro–Wilk test was first used to check for data normality. It was shown that the *p*-value for the test was less than 0.05, suggesting that the data were not normally distributed. Since the normality assumption was not satisfied, Wilcoxon signed-rank test, a non-parametric statistical method, was considered to detect the differences at 0.05 level of significance. Effect sizes for Wilcoxon signed-rank test were calculated by converting the z scores into effect size estimates. According to Coolican (2009), the effect size r of 0.10 was interpreted as a small effect, 0.30 as a medium effect, and 0.50 as a large effect. There were significant differences between T1 and T2 with respect to word length ($z = -2.55$, $p < 0.05$, $r = -0.32$), word concreteness ($z = -2.29$, $p < 0.05$, $r = -0.29$), syntactical complexity ($z = -2.66$, $p < 0.05$, $r = -0.34$), and causal connectives ($z = -2.61$, $p < 0.05$, $r = -0.33$). However, no significant difference was observed in word diversity ($z = -1.56$, $p = 0.12$, $r = 0.20$). The obtained values indicate that using MT has a positive and

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

Table 5. Scoring between T1 and T2

Components	T1		T2	
	M	*SD*	*M*	*SD*
Content	23.03	2.22	23.47	2.15
Organization	15.55	1.46	16.02	1.50
Vocabulary	14.81	1.49	15.47	1.60
Language use	17.26	2.43	17.73	2.42
Mechanics	3.84	0.62	4.03	0.68
Total score	74.48	7.01	76.71	7.05

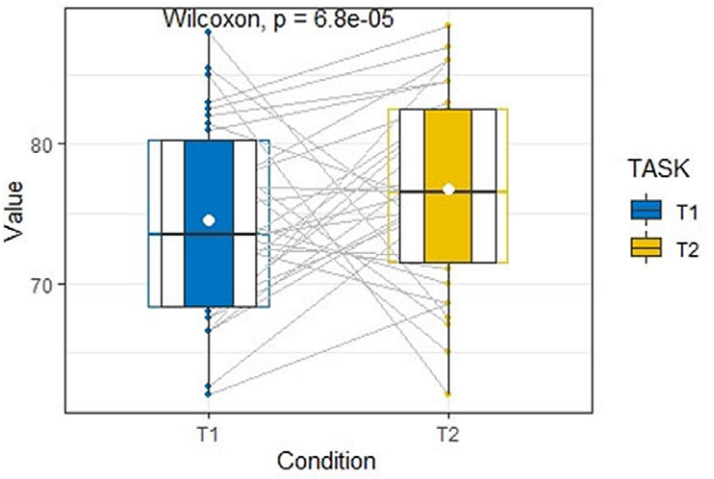

Figure 2. Comparative analysis between the scores of T1 and T2

medium effect on Chinese students' EFL writing in regard to word length, word concreteness, syntactical complexity, and causal connectives. The obtained results imply that MT basically can help students improve their writing, especially on linguistic use and syntactical complexity.

4.1.2 Writing quality assessment by human raters

Interrater reliability was considered to reduce rater bias in writing quality assessment. Pearson correlation coefficient was calculated to investigate the degree of interrater reliability. Correlation coefficients below 0.30 were considered as weak, between 0.30 and 0.60 as moderate, and above 0.60 as strong (Dancey & Reidy, 2004). It was found that the calculated interrater reliability coefficient was 0.91 for T1 and 0.93 for T2 ($p < 0.05$), suggesting a good agreement on the writing quality assessment (see Figure 1).

Table 5 shows the descriptive analysis performed on specific scoring metrics between T1 and T2. The scoring metrics consisted of content, organization, vocabulary, language use, and mechanics. The mean scores of the specific metrics in T2 were all higher than those in T1, especially for the vocabulary part.

Wilcoxon signed-rank tests were conducted to detect differences between the writing proficiency of T1 and T2 on human ratings (see Figure 2). The results suggest that there was a significant difference between T1 and T2 in terms of writing content ($z = -3.51$, $p < 0.05$, $r = -0.45$),

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

organization ($z = -3.13$, $p < 0.05$, $r = -0.40$), vocabulary ($z = -4.18$, $p < 0.05$, $r = -0.53$), language use ($z = -2.40$, $p < 0.05$, $r = -0.30$), and mechanics ($z = -2.36$, $p < 0.05$, $r = -0.30$). Statistically significant differences were found for the metrics of content, organization, vocabulary, language use, and mechanics. Akin to the finding generated by Coh-Metrix, lexical use was the one that showed striking improvement. The total score of T2 was higher than the score of T1, with a statistically significant difference ($z = -3.99$, $p < 0.05$, $r = -0.51$). The increased mean scores coupled with the significant quality differences suggest that writing proficiency has improved significantly with the help of MT.

4.2 Qualitative data analysis

4.2.1 Questionnaire data

A post-test questionnaire was immediately distributed to the participants after they had completed the whole tests in order to explore students' attitudes towards and opinions on the use of MT in EFL writing. Thirty-one valid responses were collected for analysis. The questionnaire consisted of items on a 5-point scale, including writing ability self-assessment, difficulties and challenges, as well as attitudes towards using MT in EFL writing. The collected data provided an interesting and mixed evidence for the use of MT.

For the self-assessment of Chinese writing ability, 22% of students reported that they were good at writing Chinese essays, while 74% believed their native language writing ability to be at the moderate level. With respect to specific difficulties in writing Chinese essays, students ranked content organization as the biggest challenge ($M = 3.55$, 35.48%), followed by word use ($M = 3.39$, 29.03%) and logical relations ($M = 3.06$, 9.68%). In terms of the difficulty in manual translation from Chinese into English, 58% of students considered limited vocabulary as the biggest challenge ($M = 3.94$), followed by lexical use ($M = 3.52$) and syntactical use ($M = 2.42$). On the difficulties during the writing revision process with reference to MT output, 45% of students reported that they could identify the errors produced by MT but they did not know how to revise or edit them ($M = 3.00$). Some students believed that MT might pose negative effects on cognitive processing ($M = 2.74$, 35%). Some students expressed their inability to identify the MT errors ($M = 2.10$, 13%). With regard to the favorable choice based on the MT suggestions, 68% of students put lexical suggestions in first place ($M = 3.42$, 68%), syntactical expressions in second place ($M = 2.58$, 19%), and grammatical suggestions in third place ($M = 2.06$, 13%).

Regarding the advantages of using MT, translation speed was put first ($M = 4.35$, 68%), followed by lexical help ($M = 4.1$, 26%) and syntactical help ($M = 2.71$, 6%). As for the drawbacks of using MT, laziness encouragement was placed top ($M = 4.26$, 39%), accompanied by cohesive problems ($M = 3.97$, 29%) and low feasibility ($M = 3.87$, 19%). Quality self-assessment of T1 and T2 based on a 10-point scale found that the mean score was 5.52 for T1 and 7.52 for T2, respectively. Nearly 84% of students thought the MT version of their writings was satisfying compared to their own manual translation. In this regard, it was no wonder that 90% of students believed that MT has helped them to improve their English writing.

4.2.2 Interview data

Semi-structured interviews were conducted to further triangulate the data from the writing quality assessment and the questionnaire. The interviewees were first asked to self-assess their Chinese writing ability and English writing ability. Interestingly, it was found that all students perceived their Chinese native writing ability to be at the medium level rather than the expected advanced one. This information matched the questionnaire response. For example, one student reported that "sometimes, I cannot stay on-track in my Chinese writing and cannot focus on a unique point of view. What I write is not always the same as what I think at the pre-writing stage. Off-topic

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

often happens in my writing process. These problems could also happen in my English writing. I often produce grammatical errors. In addition, the sentences that I make are simple and plain." Another one expressed that "although I can make up a good story in Chinese, as a matter of fact the wordings are not good enough. Limited vocabulary actually restricts my opinions and thoughts expression in English writing process." Although students felt more confident in native language writing, they were not actually satisfied with their native writing ability. Based on the reporting data, it is noteworthy that students' EFL writing ability is closely related to their native language writing ability.

In order to collect students' perceptions on the use of MT, the interviewees were asked to describe what their writing revision was like with reference to the MT output. They reported a number of advantages and disadvantages. Most of them believed that MT was a convenient and quick tool for generating translations. As one interviewee said, "MT can provide good suggestions for people who have poor translation ability or limited vocabulary." Negative comments were primarily associated with the quality of MT. For example, according to one interviewee, "the wordings offered by MT were accurate but not on a consistent basis." Another mentioned that "productions generated by MT often contain defects, such as blunt translation of ideas and style transfer problems." From the interviewees' comments, it could be seen that some students were aware of the strengths of MT. However, some of them were still skeptical about the values of MT for education purposes. This sort of hesitation and uncertainty were partially reflected in the recording videos where students frequently navigated between the production of MT and their manual ones.

In view of the improving quality of MT, most interviewed students showed a positive attitude and expressed a wish to use MT in their writing revision process. For example, one proposed that "MT can provide different translation choices for a given expression." Another reported that "MT outputs are sometimes even better than the ones translated by myself." However, worries and concerns on the use of MT were still observed: "It is difficult for me to break the addiction to MT use", a student noted. The interviewees were ignorant of the working principles of MT and did not know how to critically and effectively use the MT output, but used MT for its convenience and speed. In light of their poor literacy for MT, they expressed a strong wish to learn how to make use of MT effectively.

4.2.3 Screen-recording data

Screen recording generated about 40 hours of observational data. It took an average of 85 minutes for each student to complete the whole writing task. Drafting and translating consumed nearly 85% of the processing time, while revising took up 15% of the processing time.

A detailed look at the writing process showed that most students searched the online resources for writing materials at the drafting stage. For instance, they searched the life experience of the character involved in the writing theme. At the revising stage, recording data revealed that students were more apt to make micro-changes rather than macro-changes. For example, students did not simply adopt the raw output of MT at the first glance. They tended to accept specific lexical words or syntactical structures rather than directly select the massive and lengthy parts that MT offered. The "accepted-to-be" lexical words from MT were sometimes double-checked before the final decision was made. Students' uncertainty with reference to MT was revealed, reinforcing the need to deliver MT literacy instruction. It was also found that students preferred to choose syntactically complex sentences from the MT output. Taking one student's revision as an example, she adopted the MT production of "Learning English and communicating with others broadened his horizons and laid a solid foundation for his future career in the Internet industry" instead of her original one: "By learning English and communicate with others, he has largely broadened his horizon. This is this experience that lays a solid foundation for his establishment". In the interview, the student confirmed the superior MT quality compared to her own translation.

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

Finally, the screen-recording data generally echoed the findings by Coh-Metrix, supporting the view that MT is helpful for language learners in improving their EFL writing.

5. Discussion

The present study approximately replicated the work by Lee (2020), which showed that MT can help reduce lexico-grammatical errors, and MT use was perceived positively in EFL writing. Our findings therefore broadly support Lee's study. It was found that (1) the revised version (T2) was significantly improved compared to the initial manual version (T1); (2) lexical use improved more prominently in comparison with the syntactical changes; and (3) students generally believed that MT was helpful for their EFL writing and expressed a strong wish to have MT literacy training.

Automatic ratings and human ratings were both considered in this study to determine the impact of MT on EFL writing. Data generated by Coh-Metrix showed that MT could help students produce longer and more logical sentences. Word length, word concreteness, syntactical complexity, and discourse cohesion all significantly improved in the final version of T2, but not the word diversity. Specific revising moves in screen recordings showed that students preferred lexical choices to syntactical substitutions when they consulted the MT production in revising process. This finding is in agreement with the previous research (e.g. Lee, 2020; Tsai, 2019). Lack of lexical proficiency may result in global errors and lead to breaks in L2 communication (de la Fuente, 2002), especially in timed writing (Santos, 1988). Linguistic sophistication is a predictive indicator for L2 writing proficiency (Crossley & McNamara, 2012). Additionally, it was also found that syntactical complexity was significantly improved with reference to MT. This finding is contrary to Lee's (2020) work, but is consistent with the findings in Chon *et al.* (2021). One possible explanation may lie in the mixture of different language pairs and text types. MT can produce higher style levels in descriptive genres (Alrajhi, 2022). However, no statistically significant difference was found in lexical diversity. A plausible explanation may be that MT offers a range of commonly used words, which is perhaps another reason to explain students' uncertainty in selecting words and phrases from MT output (cf. Chung & Ahn, 2022).

Human ratings indicated that MT could improve the writing quality regarding content, organization, vocabulary, language use and mechanics, suggesting that MT serves not only as a meaningful language resource but also as a helpful tool in the writing revising process. This finding is in line with the work of Garcia and Pena (2011), who found that MT could help improve the L2 writing performance of writing beginners. Thus it is possible to use MT as a pedagogical tool for obtaining lexical resources during the writing process.

Qualitative data showed that students had mixed feelings about using MT in the writing process, although they acknowledged that MT use was beneficial to their writing. Most of them reported frequent use of MT in learning activities. However, they were not familiar with the working principles of MT, or the MT error patterns. Double-checking of the MT output has revealed students' uncertainty and hesitation, which has supported prior findings (Dorst *et al.*, 2022; Lee, 2023; O'Neill, 2019). MT literacy instruction appears urgent, and adequate training on the limits and strengths of MT should be provided for students to help them use MT effectively. Several studies have addressed the integration of MT literacy instruction in language education (Bowker & Ciro, 2019; O'Brien & Ehrensberger-Dow, 2020). O'Brien, Simard and Goulet (2018) suggested that post-editing MT errors could be taught in L2 writing. In sum, this study mainly supports the findings of the study replicated (Lee, 2020). It also partially echoed other prior studies where MT was used in the writing process (e.g. Bowker, 2020; Chung & Ahn, 2022; Garcia & Pena, 2011; Lee & Briggs, 2021).

https://doi.org/10.1017/S0958344023000022 Published online by Cambridge University Press

6. Conclusion

This study has found that MT could facilitate the English writing process for Chinese EFL learners. Most students perceived MT as a useful tool and wanted to be instructed to effectively use it. Furthermore, new findings were discovered from Coh-Metrix and screen recordings. Students preferred micro-changes to macro-changes and their syntactical complexity was improved after referring to the MT feedback. There is a real need to integrate MT literacy instruction into foreign language training. The obtained findings have led us to believe that Lee's work is a significant step forward in identifying the role of MT in educational settings. MT could be tentatively used as a pedagogical tool for EFL writing with adequate MT literacy training.

This replication study strives to underscore the importance of examining MT's impact in the educational field. Despite that, we still need to address its limitations. First, due to the limited sample size, we should be careful not to generalize the obtained findings to student samples with different language proficiency levels or cultural backgrounds. Second, the genre of the task materials was limited to descriptive writing. Different genres of writing tasks may impose different effects on the MT performance. Third, the language pair is another variable that can influence the MT quality. Therefore, to further strengthen the validity of the obtained findings, it is important to consider a larger sample of participants with different language proficiency levels and cultural backgrounds in future studies. Additionally, different genres of writing tasks in different language pairs should be considered in order to increase the generalization power of the findings. The research on the impact of MT on EFL writing has only just begun.

Acknowledgements. This study was supported by the Scientific and Technical Projects from the Center for Translation Studies at Guangdong Universities of Foreign Studies (CTS202107), Philosophy and Social Science Projects at Colleges and Universities in Jiangsu Province (2021SJA0068), and Special Research Project on Foreign Language Teaching Reform of High-Quality in Jiangsu Province (2022WYYB036).

The authors thank all the participants and raters in this study. Special gratitude is extended to Ma Wentao for her help in data collection. We are also grateful to Dr Cornelia Tschichold and the anonymous reviewers who provided constructive and thought-provoking comments on the revision of the paper.

Ethical statement and competing interests. Ethical permissions were obtained from institutions. All participants voluntarily participated in this study. Anonymity of the participants' responses was preserved. The authors declare no competing interests.

References

Alrajhi, A. S. (2022) Genre effect on Google Translate-assisted L2 writing output quality. *ReCALL*. Advance online publication. https://doi.org/10.1017/S0958344022000143

Bowker, L. (2020) Chinese speakers' use of machine translation as an aid for scholarly writing in English: A review of the literature and a report on a pilot workshop on machine translation literacy. *Asia Pacific Translation and Intercultural Studies*, 7(3): 288–298.

Bowker, L. & Ciro, J. B. (2019) *Machine translation and global research: Towards improved machine translation literacy in the scholarly community*. Bingley: 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.

Brudermann, C., Grosbois, M. & Sarré, C. (2021) Accuracy development in L2 writing: Exploring the potential of computer-assisted unfocused indirect corrective feedback in an online EFL course. *ReCALL*, 33(3): 248–264.

Chon, Y. V., Shin, D. & Kim, G. E. (2021) Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, 96: Article 102408. https://doi.org/10.1016/j.system.2020.102408

Chung, E. S. & Ahn, S. (2022) The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, 35(9): 2239–2264.

Coolican, H. (2009) *Research methods and statistics in psychology* (5th ed.). London: Routledge.

Crossley, S. A. & McNamara, D. S. (2009) Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2): 119–135.

Crossley, S. A. & McNamara, D. S. (2012) Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2): 115–135.
Crossley, S. A., Salsbury, T., McNamara, D. S. & Jarvis, S. (2011) Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, 28(4): 561–580.
Daems, J., Vandepitte, S., Hartsuiker, R. J. & Macken, L. (2017) Translation methods and experience: A comparative analysis of human translation and post-editing with students and professional translators. *Meta*, 62(2): 243–484.
Dancey, C. P. & Reidy, J. (2004) *Statistics without maths for psychology: Using SPSS for Windows* (3rd ed.). Harlow: Pearson Education Limited.
de la Fuente, M. J. (2002) Negotiation and oral acquisition of L2 vocabulary: The roles of input and output in the receptive and productive acquisition of words. *Studies in Second Language Acquisition*, 24(1): 81–112. https://doi.org/10.1017/S0272263102001043
Dorst, A. G., Valdez, S. & Bouman, H. (2022) Machine translation in the multilingual classroom: How, when and why do humanities students at a Dutch university use machine translation? *Translation and Translanguaging in Multilingual Contexts*, 8(1): 49–66. https://doi.org/10.1075/ttmc.00080.dor
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
Enríquez Raído, V. (2014) *Translation and web searching*. New York: Routledge. https://doi.org/10.4324/9780203798034
Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2): 175–191.
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.
Garcia, I. & Pena, M. I. (2011) Machine translation-assisted language learning: Writing for beginners. *Computer Assisted Language Learning*, 24(5): 471–487.
Graesser, A. C. & McNamara, D. S. (2011) Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2): 371–398.
Handley, Z. (2018) Replication research in computer-assisted language learning: Replication of Neri et al. (2008) and Satar &x04E6;zdener (2008). *Language Teaching*, 51(3): 417–429.
Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F. & Hughey, J. B. (1981) *Testing ESL composition: A practical approach*. Rowley: Newbury House.
Jin, Y. & Yang, H. (2006) The English proficiency of college and university students in China: As reflected in the CET. *Language, Culture and Curriculum*, 19(1): 21–36. https://doi.org/10.1080/07908310608668752
Jolley, J. R. & Maimone, L. (2022) Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal*, 14(1): 26–44.
Kelly, R. & Hou, H. (2022) Empowering learners of English as an additional language: Translanguaging with machine translation. *Language and Education*, 36(6): 544–559. https://doi.org/10.1080/09500782.2021.1958834
Kenny, D. (ed.) (2022) *Machine translation for everyone: Empowering users in the age of artificial intelligence*. Berlin: Language Science Press.
Klekovkina, V. & Denié-Higney, L. (2022) Machine translation: Friend or foe in the language classroom? *L2 Journal*, 14(1): 105–135. https://doi.org/10.5070/L214151723
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
Lam, F. S. & Pennington, M. C. (1995) The computer vs. the pen: A comparative study of word processing in a Hong Kong secondary classroom. *Computer Assisted Language Learning*, 8(1): 75–92.
Latifi, S. & Gierl, M. (2021) Automated scoring of junior and senior high essays using Coh-Metrix features: Implications for large-scale language testing. *Language Testing*, 38(1): 62–85. https://doi.org/10.1177/0265532220929918
La Torre, M. D. (1999) A web-based resource to improve translation skills. *ReCALL*, 11(3): 41–49.
Lee, S.-M. (2020) The impact of using machine translation on EFL students' writing. *Computer Assisted Language Learning*, 33(3): 157–175.
Lee, S.-M. (2023) The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 36(1–2): 103–125. https://doi.org/10.1080/09588221.2021.1901745
Lee, S.-M. & Briggs, N. (2021) Effects of using machine translation to mediate the revision process of Korean university students' academic writing. *ReCALL*, 33(1): 18–33. https://doi.org/10.1017/S0958344020000191
Li, W. (2011) Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics*, 43(5): 1222–1235. https://doi.org/10.1016/j.pragma.2010.07.035
Liu, H. & Brantmeier, C. (2019) "I know English": Self-assessment of foreign language reading and writing abilities among young Chinese learners of English. *System*, 80: 60–72.
McManus, K. (2022) Are replication studies infrequent because of negative attitudes? Insights from a survey of attitudes and practices in second language research. *Studies in Second Language Acquisition*, 44(5): 1410–1423. https://doi.org/10.1017/S0272263121000838

McNamara, D. S., Crossley, S. A. & Roscode, R. (2013) Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2): 499–515.
McNamara, D. S., Graesser, A. C., McCarthy, P. M. & Cai, Z. (2014) Automated evaluation of text and discourse with Coh-Metrix. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664
Niño, A. (2008) Evaluating the use of machine translation post-editing in the foreign language class. *Computer Assisted Language Learning*, 21(1): 29–49.
Niño, A. (2020) Exploring the use of online machine translation for independent language learning. *Research in Learning Technology*, 28: 1–32. https://doi.org/10.25304/rlt.v28.2402
O'Brien, S. & Ehrensberger-Dow, M. (2020) MT literacy—A cognitive view. *Translation, Cognition & Behavior*, 3(2): 145–164.
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. M. (2019) Online translator, dictionary, and search engine use among L2 students. *CALL-EJ*, 20(1): 154–177. 'http://callej.org/journal/20-1/O'Neill2019.pdf
Porte, G. & McManus, K. (2019) *Doing replication research in applied linguistics*. New York: Routledge. https://doi.org/10.4324/9781315621395
Porte, G. & Richards, K. (2012) Focus article: Replication in second language writing research. *Journal of Second Language Writing*, 21(3): 284–293. https://doi.org/10.1016/j.jslw.2012.05.002
Santos, T. (1988) Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly*, 22(1): 69–90. https://doi.org/10.2307/3587062
Somers, H. (2004) Does machine translation have a role in language learning? Paper presented at *UNTELE 2004: L'autonomie de l'enseignant et de l'apprenant face aux technologies de l'information et de la communication, Teacher and Learner Autonomy vis-a-vis Information Communication Technology*. Universiteé de Technologie de Compiegne, 19 March.
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 Purpose*, 56: 18–34.
Tsai, S.-C. (2019) Using Google Translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning*, 32(5–6): 510–526.
Tschichold, C. (2019) CALL replication studies: Getting to grips with complexity. In Meunier, F., Van de Vyver, J., Bradley, L. & Thouësny, S. (eds.), *CALL and complexity: Short papers from EUROCALL 2019*. Research-publishing.net, 362–366.
Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J. Hughes, S., Wallace, P., Griffin, T. D. & Britt, M. A. (2017) Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education*, 27(4): 758–790.
Yamada, M. (2019) The impact of Google Neural Machine Translation on post-editing by student translators. *The Journal of Specialised Translation*, 31: 87–106. https://jostrans.org/issue31/art_yamada.pdf
Yang, Y. & Wang, X. (2019) Modeling the intention to use machine translation for student translators: An extension of technology acceptance model. *Computers & Education*, 133: 116–126. https://doi.org/10.1016/j.compedu.2019.01.015
Zheng, Y. & Cheng, L. (2008) Test review: College English Test (CET) in China. *Language Testing*, 25(3): 408–417.

About the authors

Yanxia Yang is a postdoctoral fellow at the School of Foreign Studies, Nanjing University, and an associate professor at the School of Foreign Studies, Nanjing Agricultural University. Her research interests include computer-assisted language learning, machine translation post-editing and translator training.

Xiangqing Wei is a professor at the School of Foreign Studies, Nanjing University. Her research interests are foreign language writing, lexicology and translation.

Ping Li is a professor at the School of Foreign Studies, Nanjing Agricultural University. His research focus is translation studies.

Xuesong Zhai is a senior researcher at the College of Education, Zhejiang University. His research interests include but not limited to the area of interactive learning, construction of smart learning environment and emerging technology-enhanced learning.

Author ORCiD. Yanxia Yang, https://orcid.org/0000-0001-5543-0065
Author ORCiD. Xiangqing Wei, https://orcid.org/0000-0002-6340-1341
Author ORCiD. Ping Li, https://orcid.org/0000-0001-7455-9403
Author ORCiD. Xuesong Zhai, https://orcid.org/0000-0002-4179-7859