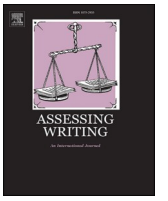
Assessing Writing 65 (2025) 100950

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw



The impact of self-revision, machine translation, and ChatGPT on L2 writing: Raters' assessments, linguistic complexity, and error correction

Minjoo Kim [a], Yuah V. Chon [b,*]

[a] Department or School: Graduate School of Education, University: Hanyang University, 222 Wangshimli-ro, Seongdong-Gu, Seoul 04763, South Korea
[b] Department or School: Department of English Education, University: Hanyang University, 222 Wangshimli-ro, Seongdong-Gu, Seoul 04763, South Korea

ARTICLE INFO

Keywords:
L2 writing
Machine translation
Inverse translation
ChatGPT
Proofreading

ABSTRACT

This study explores how learners in a South Korean high school English as a Foreign Language (EFL) context can effectively use neural machine translation (MT) and ChatGPT to enhance their L2 writing. While recent AI tools offer significant potential for supporting human writing feedback, a comparative analysis of how these tools impact writing outcomes—compared to when L2 writers independently proofread and revise their writing—has not been fully examined. To address this gap, a controlled experiment was conducted using three distinct proofreading interventions—self-proofreading (SP), MT-assisted proofreading (MAP), and ChatGPT-assisted proofreading (CAP). Learners were encouraged to first compose their texts in their L2 and then use either MT through inverse translation or ChatGPT through a structured proofreading process. The findings revealed that learners using MAP and CAP demonstrated substantial improvements in overall writing quality compared to those relying solely on SP. CAP users, in particular, produced longer texts, exhibited greater lexical diversity, and constructed more complex sentences, although this was accompanied by reduced verb cohesion. Both MAP and CAP significantly reduced grammatical errors, but did not affect prepositional errors. These findings provide practical recommendations for integrating MT and ChatGPT into L2 writing pedagogy.

1. Introduction

The role of artificial intelligence (AI) in education, particularly in language learning and writing assistance, is rapidly evolving (Steiss et al., 2024). Tools for machine translation (MT) and conversational agents such as ChatGPT are revolutionizing the teaching and assessment of second language (L2) writing. MT such as Google Translate excel at translating individual sentences, allowing students to refine specific linguistic elements (Mundt & Groves, 2016). The inverse translation feature (translating back and forth between languages) helps students detect errors by comparing translations with their native languages (Cancino & Panes, 2021; Chang et al., 2022). Meanwhile, ChatGPT supports deeper engagement by aiding in essay drafting, grammar explanations, and writing style refinement (Ali et al., 2023; Baskara, 2023; Kostka & Toncelli, 2023; Ray, 2023). Due to their different approaches to error correction and text refinement—MT focusing on direct translations and corrections based on linguistic rules, while ChatGPT provides more

* Corresponding author.
E-mail addresses: mjkim208@hanyang.ac.kr (M. Kim), vylee52@hanyang.ac.kr (Y.V. Chon).

https://doi.org/10.1016/j.asw.2025.100950
Received 17 April 2024; Received in revised form 17 March 2025; Accepted 28 April 2025
Available online 1 May 2025

context-aware and nuanced feedback—comparative analysis of these tools can foster critical awareness and help students balance their skills by using MT for sentence-level refinement and ChatGPT for broader feedback. Although MT is not inherently designed for automated writing evaluation (AWE), its comparative use provides indirect, model-based feedback, helping learners critically reflect on linguistic accuracy and complexity, thus functioning similarly to traditional AWE tools. For example, when students compare their self-written or self-translated texts with MT-produced texts, discrepancies become apparent, highlighting areas of grammatical error, incorrect word usage, and syntactic inaccuracy.

Despite high expectations, the impact of MT and ChatGPT in AWE is underexplored, particularly in supporting L2 writers. Key questions remain: Do human holistic assessments align with expectations of AI-assisted tool performance? Can these tools enhance language sophistication and effectively address errors while preserving intended meaning? These gaps motivated our study. Comparative research on neural MT and ChatGPT is still in its early stages, partly due to the perception of MT as outdated post-ChatGPT. While prior studies have focused on domains like news and social media (Hwang et al., 2023; Jiao et al., 2023; Sanz-Valdivieso & López-Arroyo, 2023; Son & Kim, 2023), their potential in L2 writing instruction remains largely untapped.

This study is also motivated by mixed findings in prior research on AWE, which have fueled skepticism about the accuracy and effectiveness of automatic written corrective feedback (AWCF) for L2 writers and teachers (Link et al., 2022). This skepticism is heightened when AWCF is compared to human feedback from peers or teachers (Fu et al., 2024). AWCF tools often focus on word- or sentence-level revisions, leaving content and meaning largely unaffected (Dikli & Bleyle, 2014; Li et al., 2015; Warschauer & Grimes, 2008). Additionally, Steiss et al. (2024) found human raters outperformed ChatGPT in most areas of assessment. Thus, validating the reliability and effectiveness of MT and ChatGPT as AWCF tools through comparative analysis remains essential.

Another persistent issue in empirical studies on AWE is the infrequent use of control groups (Barrot, 2021). Incorporating a control group is a crucial method for validating the effectiveness of AWCF tools. As Link et al. (2022) have emphasized, the absence of a control group complicates drawing robust conclusions about the effectiveness of AWE. Studies are needed regarding how learners perform with AWE in comparison to when the students receive no external feedback and have to improve self-written texts on their own.

This study addresses gaps and methodological flaws in prior AWE research by exploring the use of AI-powered proofreading tools in an English as a foreign language (EFL) context. Through a controlled experiment comparing self-proofreading (SP), MT-assisted proofreading (MAP), and ChatGPT-assisted proofreading (CAP), the study examines how these tools enhance autonomous learning and writing skills during the revision stage. SP fosters self-directed learning, a priority in Korean secondary schools, and serves as a control group without external resources (Barrot, 2021; Wang et al., 2013; Wilson & Czik, 2016). Including a control group addresses a key weakness in earlier studies, where its absence undermined validity (Warschauer & Grimes, 2008).The MAP group uses inverse translation to promote cross-linguistic awareness and message refinement, a technique underexplored in research (Cancino & Panes, 2021; Chang et al., 2022). However, its comparative effectiveness with ChatGPT remains unclear. CAP, offering dynamic and context-sensitive feedback, provides nuanced suggestions addressing linguistic form and meaning (Su et al., 2023). Yet, its effectiveness in tackling issues unresolved by earlier AWCF tools—such as superficial feedback, limited accuracy improvement (Chen & Cui, 2022), and challenges in higher-order writing features like coherence and argumentation (Li et al., 2015; Ranalli, 2018)—requires further examination. By comparing CAP with SP and MAP, this study investigates whether CAP's advanced capabilities improve holistic writing quality, linguistic complexity, and accuracy under these conditions.

1.1. Research questions

1. How do SP, MAP, and CAP compare in terms of improving the holistic scores of L2 writing?
2. How do these proofreading tools impact linguistic complexity measures in L2 writing?
3. How do these tools affect language accuracy in the texts produced by L2 learners?

The impact of SP, MAP, and CAP was researched through (1) human's holistic evaluations, (2) analyses of linguistic complexity, and (3) an examination of the types and frequencies of errors based on the following considerations: Previous AWE tools have shown inaccuracies, and the reliability of recent AWE tools—MT and ChatGPT—also needs to be validated. Therefore, the human-teachers' holistic evaluations were needed to assess the writing quality resulting from SP, MAP, and CAP. In addition to human raters' holistic assessments of writing products, incorporating measures of linguistic complexity was needed to capture aspects of language that go beyond basic correctness or clarity, including the use of contextually appropriate language. Finally, analyzing the extent to which errors are corrected with the assistance of MT and ChatGPT provides insights into their effectiveness in improving language accuracy. It also highlights their potential to introduce previously unrecognized errors. Since MT and ChatGPT are types of AWE, a review of all three was necessary, as detailed below.

2. Background

2.1. Theoretical foundation: role of automated writing evaluation

AWE refers to the use of computer-based systems to evaluate written texts, typically by providing scores, diagnostic feedback, or suggestions for improvement (Fu et al., 2024). Researchers have examined its effects from various angles, including its influence on writing practices (Li et al., 2015; Ranalli, 2018) and accuracy improvements across drafts (Bai & Hu, 2017; Barrot, 2021). Comparative studies highlight differences between AWE and human feedback (Dikli & Bleyle, 2014; Wang et al., 2013), while others explore instructor usage (Li et al., 2015) and student evaluations of AWE (Bai & Hu, 2017). Comprehensive reviews (Fu et al., 2024; Stevenson

& Phakiti, 2014) assess its overall impact on L2 writing. In peer feedback, studies show AWE facilitates review processes but underperforms in cohesive device use compared to peer feedback (Chen & Cui, 2022; Li & Li, 2018). However, few studies use control groups to assess AWE's efficacy (Barrot, 2021; Wang et al., 2013) or examine its impact on linguistic complexity. The following provides a more detailed review of studies based on recurring AWE themes.

Studies indicate that AWE tools significantly improve writing accuracy, particularly in addressing grammatical and mechanical errors. Bai and Hu (2017) found that Chinese university students using Pigai achieved 98 % accuracy in mechanics (e.g., spelling, capitalization, punctuation), with 100 % successful revisions. Similarly, Barrot (2021) showed that ESL students using Grammarly significantly reduced severe errors, though minor errors saw less improvement, reflecting a focus on grammar over mechanics. Liao (2016) demonstrated that integrating AWE with process-writing principles led to long-term reductions in grammatical errors, such as sentence fragments and subject-verb disagreements, among Taiwanese university students. Wang et al. (2013) further highlighted AWE's role in improving grammar, sentence structure, and word usage.

AWE corrective feedback influences students' writing practices, though its effectiveness depends on feedback specificity and focus. Li et al. (2015) found that using Criterion in an ESL process-writing class significantly reduced grammar and mechanics errors while encouraging frequent revisions. However, Criterion struggled with complex structures and nuanced errors, and generic feedback often caused confusion. Feedback primarily addressed lower-order concerns, offering limited support for coherence, content development, and argumentation. Similarly, Ranalli (2018) found that Criterion's specific feedback was more effective than generic feedback for most error types, except prepositions, where both were equally helpful. Students reported higher mental effort with generic feedback, highlighting the importance of clarity in facilitating effective revisions.

Comparative studies show that AWE feedback often falls short in addressing higher-order writing concerns compared to human-instructor or peer feedback. Dikli and Bleyle (2014) found that instructors identified significantly more errors than AWE tools, which frequently missed common L2 issues like verb tense and singular/plural agreement. Similarly, Link et al. (2022) observed that while AWE reduced teachers' workloads by addressing lower-level concerns, teacher-only feedback outperformed AWE in measures of complexity and fluency, as AWE tends to prioritize surface-level issues. Chen and Cui (2022) compared AWE with peer feedback in continuation writing tasks, finding that peer feedback significantly improved cohesion and coherence through greater use of cohesive devices and chains. In contrast, AWE feedback led to limited changes in these areas, with vague suggestions that failed to address higher-order concerns effectively. These findings suggest that while AWE supports accuracy improvements, it must be complemented by more nuanced feedback to promote comprehensive writing development.

In summary, AWE research highlights its strengths in reducing linguistic and surface-level errors, encouraging multiple drafts, and providing efficient feedback. As a supplementary tool, it allows instructors to focus on higher-order concerns like content and organization. However, findings on AWE's effectiveness compared to no feedback or teacher feedback are mixed (Fu et al., 2024; Stevenson & Phakiti, 2014). While effective for mechanical and linguistic aspects, AWE struggles with higher-order features like coherence and content development, and its accuracy varies by error type, excelling in mechanics but less so in grammar and collocations (Fu et al., 2024; Li et al., 2015). A persistent limitation in AWE research is the lack of control groups, making it difficult to draw conclusive results (Link et al., 2014). Even studies with control groups (e.g., Barrot, 2021; Wang et al., 2013) often fail to describe the teacher feedback provided.

2.2. *Machine translation in L2 writing: input quality, output complexity, and critical revision*

MT tools have significantly advanced as valuable resources for L2 writing, particularly following the transition from statistical machine translation (SMT) to neural machine translation (NMT) in 2016. NMT's ability to translate entire sentences rather than break them into phrases has increased accuracy and comprehensibility, drawing increasing academic attention (Chung & Ahn, 2021; Lee, 2022). This shift has prompted growing research on MT's enhanced role in foreign-language education.

The effectiveness of MT largely depends on input quality, specifically the accuracy and complexity of L1 texts entered for translation into L2 (Jiang et al., 2024). Chon et al. (2021) noted that poor input quality, such as ambiguous structures or incorrect grammar, including language-specific features such as pro-drop languages, results in mistranslations and awkward phrasing. Lee (2022) found that higher-proficiency learners who provided clearer input benefited more from MT than those with less precise input. Accurate and complex inputs lead to higher lexical and syntactic complexity in translations (Jiang et al., 2024), echoing Chung and Ahn's (2021) emphasis on training learners to craft clear input. Thus, the effective use of MT requires learners to provide clear and accurate input while recognizing cross-linguistic differences to mitigate MT's limitations. However, input quality alone does not guarantee successful communication, as the translation process also heavily depends on the quality of MT output.

The MT output refers to the text produced by these systems before human edits. Output quality is critical for effective L2 writing. When the output fails to convey the intended message accurately, post-editing at the lexical, sentential, and discourse levels is necessary (Jia et al., 2019; Konttinen et al., 2020; Shin & Chon, 2023). This underscores the importance of careful revision and critical engagement, especially because learners may struggle with complex linguistic features such as colloquial language and cultural nuances.

Research has indicated that MT outputs often demonstrate greater lexical and syntactic complexity than student writing, featuring more advanced and varied vocabulary (Chang et al., 2022; Tsai, 2019). For example, Chon et al. (2021) found that machine-translated writing enables learners to produce texts that are more lexically diverse and syntactically complex than those created through direct writing, or self-translated writing, suggesting that MT has the potential to enhance L2 writing sophistication. However, this raises an important question: is MT genuinely improving learners' L2 writing skills, or is it merely replacing their linguistic efforts with machine-generated proficiency? While it is unreasonable to expect human learners to perform at the same level of lexical and syntactic

sophistication as MT, this distinction is crucial in understanding the role of MT in language learning.

Studies such as those by Alrajhi (2023) and Cancino and Panes (2021) also show that MT, such as Google Translate, generally improves literacy levels and content quality across different genres but requires critical post-editing to ensure clarity. Furthermore, Chang et al. (2022) observed that while MT can enhance specific writing aspects, its overall impact on writing quality may be limited by the task type and complexity.

In summary, while MT significantly enhances lexical and syntactic complexity in L2 writing and provides advanced vocabulary and grammatical structures, it requires careful post-editing and critical engagement to ensure that the output aligns with the learner's intended meanings and academic conventions. The benefits of MT must be weighed against challenges, such as clarity and appropriateness, in the final text.

2.2.1. Reducing errors and improving accuracy through machine translation

A key advantage of MT in L2 writing is its ability to reduce errors and improve accuracy. Tsai (2019) demonstrated this by having EFL learners write in Chinese, translate their text into English using MT, and compare their English drafts with MT versions. The analysis showed that the machine-translated texts contained more words and fewer spelling and grammatical errors. Beiler and Dewilde (2020) noted that learners often use translation tools to verify their translations and cross-reference with other tools or peers to improve accuracy, particularly when dealing with complex language equivalents.

Chon et al. (2021) found that machine-translated writing had fewer errors than direct writing and self-translated writing, especially in articles and prepositions, indicating MT's effectiveness in correcting common grammatical mistakes. Lee (2022) similarly reported that MT significantly reduced grammatical errors across all proficiency levels, although the extent of error reduction varied, with lower-proficiency learners benefiting less from their limited ability to evaluate and revise MT outputs critically. Jiang et al. (2024) emphasize that while MT enhances grammatical accuracy, it is not foolproof and can introduce errors learners may miss if they do not critically engage in the revision process. These findings align with those of Chung and Ahn (2021) and Tsai (2019), highlighting the dual role of MT in reducing some errors while potentially introducing others that require careful attention.

Cancino and Panes (2021) found that writing accuracy improved significantly for groups using MT, with fewer grammatical errors than in the control group. There was no significant difference in accuracy between those who received instructions on using MT and those who did not, suggesting that MT can improve grammatical accuracy in L2 writing regardless of additional instruction. Chang et al. (2022) further investigated this by training an experimental group in recursive editing using MT, whereas the control group did not receive such training. The results showed that the MT-aided group performed better in error correction tasks, indicating that recursive editing with MT enhanced error correction abilities. However, the control group maintained higher fluency, likely because of less focus on meticulous editing, which may have slowed the writing flow.

The growing body of research on MT in L2 writing has shown that it is a valuable AI tool for reducing errors and improving accuracy. Nevertheless, with the rise of more advanced tools based on large language models and generative AI, such as ChatGPT, the role of MT in AWE must be reconsidered and redefined.

2.3. ChatGPT and L2 writing feedback

In November 2022, OpenAI launched ChatGPT, an AI that surpasses traditional MT by generating language through sophisticated natural language processing. This conversational AI chatbot produces immediate, contextually relevant, and human-like responses by leveraging extensive pre-acquired knowledge. Unlike existing AI-based tools, ChatGPT offers unique functions, including text generation for language learning (Mindner et al., 2023), translation (Cao & Zhong, 2023), correcting foreign language expressions (Mizumoto et al., 2024), and engaging in discussions (Kostka & Toncelli, 2023). These capabilities have led to significant changes in education.

Among its diverse functions, ChatGPT is particularly useful in L2 writing for tasks such as idea generation (Su et al., 2023), text correction (Harunasari, 2023), and providing personalized real-time feedback (Ali et al., 2023; Baskara, 2023; Ray, 2023). Researchers have begun empirically evaluating the effectiveness of ChatGPT in assisting L2 learners while critically examining its limitations and areas where it may fall short.

2.3.1. Benefits and challenges of ChatGPT as a writing tool

Several studies have explored the benefits and challenges of ChatGPT in L2 writing instruction, highlighting its potential to enhance language learning. Barrot (2023) assessed ChatGPT's role as an effective tutor and source of language input while raising concerns about its impact on writing pedagogy and academic integrity. One key benefit is the ability to grade written work automatically and provide detailed feedback on content relevance and clarity. Similarly, Ali et al. (2023) noted that ChatGPT enriches EFL students' experiences by offering immediate personalized feedback and boosting their confidence and motivation. Kohnke et al. (2023) emphasize ChatGPT's ability to provide rich linguistic input, facilitate authentic conversational practice, and offer immediate formative assessments, making it a valuable tool in language education.

Recent studies have highlighted the diverse benefits of ChatGPT for EFL learners at different stages of writing. Harunasari (2023) examined the integration of ChatGPT into an EFL writing class in Jakarta and demonstrated its effectiveness in idea generation, planning, drafting, and revision. For example, ChatGPT helped students overcome writer's block, develop coherent storylines, and improve their writing quality through grammatical corrections. Su et al. (2023) suggested using ChatGPT in argumentative writing to address dialogical, structural, and linguistic challenges at various stages, including outline preparation, content revision, and post-writing reflection.

Studies have shown that ChatGPT is highly effective in correcting linguistic errors, often surpassing human instructors and other proofreading tools. Algaraady and Mahyoob (2023) compared ChatGPT's performance with that of human instructors in detecting and analyzing writing errors among EFL learners and found that ChatGPT effectively identified most surface-level errors and provided suggestions for improving sentence structure, word choice, and clarity. Schmidt-Fajlik (2023) evaluated ChatGPT against Grammarly and ProWritingAid, noting ChatGPT's detailed feedback and user-friendly interface, particularly its ability to explain to lower-level Japanese learners. Athanassopoulos et al. (2023) explored ChatGPT's potential to enhance foreign language writing skills among refugees and migrants learning German, finding significant improvements in vocabulary and grammar.

2.3.2. Evaluating AI-generated feedback: comparing ChatGPT with human teachers and student self-proofreading

Research comparing AI-generated feedback with human teacher ratings and student self-proofreading has yielded mixed results. Cao and Zhong (2023) found that teacher feedback led to the highest translation quality, followed by self-feedback, while ChatGPT-based feedback scored the lowest. Escalante et al. (2023) observed no significant difference in learning outcomes between the AI and human tutor groups, although ChatGPT provided clear and specific feedback. Guo and Wang (2024) compared ChatGPT and teacher-generated feedback on EFL students' essays and found that ChatGPT was more directive, whereas teachers offered more informative feedback. Steiss et al. (2024) found that, while human raters outperformed ChatGPT in most areas, ChatGPT's feedback was surprisingly close in quality, suggesting its potential usefulness, especially in the early stages of writing.

Despite ChatGPT's significant advantages in L2 writing instruction, researchers highlight its challenges, including its limitations in providing comprehensive and nuanced feedback and its potential negative impacts on creativity and critical thinking (Barrot, 2023). Moreover, teachers and students require training in digital competencies and prompt engineering to maximize the benefits of ChatGPT (Harunasari, 2023). Kohnke et al. (2023) also highlight that ChatGPT's reliance on English-language training data can introduce cultural biases, as translations may fail to capture cultural nuances. These biases may go unnoticed, particularly if students assume ChatGPT is culturally neutral. These challenges must be considered carefully when using ChatGPT for L2 writing. The limitations and challenges of using ChatGPT for L2 writing are summarized in Table 1.

3. Methods

3.1. Participants and context

This study was conducted at a public high school in northwestern Seoul with 79 Korean-speaking students (35 females, 44 males), aged 16, learning English as a foreign language. One teacher led the study. Students, who began learning English at age 10 through communicative language teaching, experienced a shift to grammar translation in secondary education due to the College Scholastic Ability Test (CSAT). At the time, students had a B1 English proficiency level (CEFR) based on school assessments but struggled with writing despite active class participation and trust in their teachers.

Two years of remote middle school learning during the pandemic familiarized students with online platforms like Google Classroom and Google Forms. While most had experience with translation tools, few had used ChatGPT. This context was selected to reflect typical Korean EFL learners, supporting educational goals of fostering self-directed learning and autonomy (Joo et al., 2022).

Table 1
Limitations and challenges of using ChatGPT for L2 writing.

	Needs human oversight/ validation	Cannot replace nuanced human feedback	Struggles with deeper structural issues and pragmatics	Inconsistent feedback	Generates repetitive/ formulaic responses	Limits creativity/ critical thinking	Ethical concerns/ learning loss	Cultural biases	Requires training in digital skills
Barrot (2023)		O		O	O	O	O		
Ali et al. (2023)				O					
Kohnke et al. (2023)				O			O	O	O
Harunasari (2023)							O		
Su et al. (2023)	O		O						
Mizumoto et al. (2024)	O								
Cao and Zhong (2023)		O							
Escalante et al. (2023)	O								
Steiss et al. (2024)	O								

3.2. Research design

Learners in the pre-designated classes were randomly divided into three groups to ensure that all groups represented each class. A counterbalanced design was employed, allowing all the learners to use the three proofreading methods (SP, MAP, and CAP) in different sequences (Table 2). This design mitigates order effects, enabling a systematic comparison of the impact of proofreading tools on writing performance. In alignment with one of the educational goals of Korean secondary schools, which focuses on fostering self-directed and autonomous learning (Joo et al., 2022), SP group served as a baseline for comparing the quality of MT and ChatGPT outputs.

Each writing task followed a standardized procedure with clear instructions, specific prompts, and adequate time to ensure consistency across the groups. The interventions were uniformly applied, making the differences attributable to proofreading tools rather than task administration variations. The counterbalanced design ensured that each group received interventions in a different order while controlling for potential order effects. The basic characteristic of a repeated measures design (also known as a within-group design) is that multiple measurements come from each participant. In a counterbalanced design, because each individual does all tasks, and the order is different for each group, the issue of the possible lack of comparability due to ordering effects can be minimized (Mackey & Gass, 2015).

The study was conducted in line with learners' semester schedules, although writing topics followed textbook content progression and could not be counterbalanced. However, all topics were based on the Korean National Curriculum, which ensured a similar cognitive load. This allowed for a fair comparison of the proofreading methods while controlling for topic sequence effects.

Writing tasks and topics were derived from themes in the learners' textbook, "*High School English*" (Kim et al., 2018), covering roles and responsibilities, healthy living, and creating a better world. The prompts are presented in Table 2. To accommodate learners' English proficiency, writing tasks were conducted as guided writing with sentence frames/starters to scaffold the writing process, following practices to develop foundational writing skills (Echevarria, 2016).

3.3. Procedure

The main writing session (50 minutes) focused on idea generation, outlining, drafting, and revision (see Fig. 2). This study specifically targeted the revision stage, during which formative feedback can improve learners' texts most effectively. During this stage, we explored how learners could use AI-generated proofreading tools (Google Translate and ChatGPT) as formative feedback compared to their SP skills. Learners were advised that feedback from MT and ChatGPT might be inaccurate and were encouraged to make independent decisions about its use. Examples demonstrated how these tools often fail to account for idiomatic, cultural, or grammatical nuances, such as translating "시간이 빨리 간다" as "Time is running fast" instead of the natural English equivalent, "Time flies."

Writing tasks using the three proofreading tools were spaced approximately a month apart and aligned with the semester plan. Before the main writing sessions, learners were introduced to writing topics and proofreading tools (Google Translate and ChatGPT), forming background knowledge through textbook reading. Each main writing session lasted 50 min for task completion, with learners using their mobile phones to access the Google Translate and ChatGPT. When using these proofreading tools, learners were instructed to review the AI tools' output to determine its appropriateness and decide whether to accept, reject, or modify it before incorporating it into their final writing.

The instructions for each proofreading method were carefully staged. In the SP method, learners independently reviewed their drafts, identified errors, revised their work, and submitted the revised and final versions via Google Forms. Revising using the MAP involves a two-stage process. First, learners used a one-way translation, entered their English drafts into Google Translate, and checked the Korean translation for accuracy. Subsequently, they revised their English drafts accordingly. In the second stage, learners used inverse translation. Students enter the revised English draft into Google Translate again. They then reviewed the new Korean translation results for any remaining errors or inaccuracies. Using the "swap languages" function, learners translated the Korean translation result back into English. They then compared their revised English draft with the inverse-translated results, which were also recorded in Google Form, to identify any discrepancies or areas that still need improvement. CAP also uses a two-stage process: one-way translation and proofreading. The learners first translated their English drafts into Korean using ChatGPT with the prompt "Translate this paragraph to Korean." They then checked the translation for accuracy and then revised it accordingly. In the second stage, they used ChatGPT to proofread their revised English drafts with the prompt "Proofread this paragraph," accepting or rejecting corrections based on accuracy and appropriateness. The best prompts for eliciting ChatGPT feedback were determined through prompting and analysis cycles (Ingley & Pack, 2023; Steiss et al., 2024). The final and intermediate versions of the drafts were submitted to Google Forms.

Table 2
Research design for writing tasks with proofreading methods.

Writing Task	Writing topics	Group A	Group B	Group C
Writing Task 1	Q: What is more important to a sports player–talent or hard work?	SP	MAP	CAP
Writing Task 2	Q: What is the best way to control negative emotions?	CAP	SP	MAP
Writing Task 3	Q: Do you think volunteer work should be a requirement for graduation? Why?	MAP	CAP	SP

Note: SP = Self-proofreading, MAP = Machine translation- assisted proofreading, CAP = ChatGPT-assisted proofreading

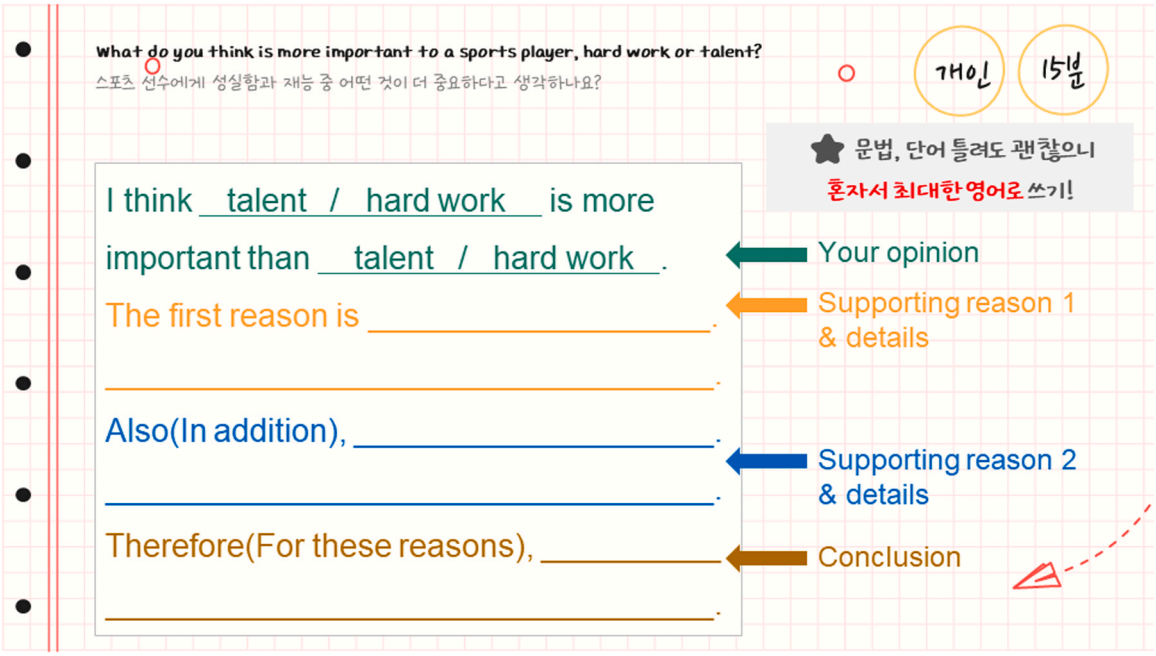


Fig. 1. Template used for Guided Writing.

3.4. Data analysis

3.4.1. Assessment of writing products

All statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS), version 26. Regarding Research Question 1, which focused on obtaining holistic writing scores from the students' writing products, the texts produced under SP, MAP, and CAP were compared. Two teachers holistically assessed the final writing products. Teacher 1 was the researcher, and Teacher 2, with a master's degree in English Language Teaching, had 15 years of teaching experience in South Korea. They used an adapted TOEFL iBT® Writing for an Academic Discussion Rubric, scoring up to 5 points, with intermediate scores (e.g., 4.5, 3.5) added for finer distinctions. All 237 essays (3 × 79) were double-scored independently for quality by the two educators. Prior to scoring, both teachers performed assessment tasks on benchmark essays to calibrate their ratings. The percentage of exact agreement was 86.5 %, while adjacent agreement was 100 %. The inter-rater reliability was high, with SP:.944, MAP:.883, and CAP:.898. The two raters' scores were calculated for mean scores which were submitted for statistical analysis using repeated-measures (RM) one-way ANOVA to assess the effects of the proofreading mode on the writing products.

3.4.2. Analysis of linguistic complexity

To address Research Question 2, the texts produced under SP, MAP, and CAP were analyzed for linguistic complexity using Coh-Metrix 3.0 (Graesser et al., 2011) for comparison. Linguistic features such as syntactic and lexical complexity are reliable indicators of L2 writing proficiency (Crossley & McNamara, 2009; Norris & Ortega, 2009). Syntactic complexity, measured through coordination, subordination, and nominalization, along with metrics such as mean clause length and dependent clause ratio, reflects the use of sophisticated structures (Lu, 2011). Lexical complexity is evaluated through lexical density, sophistication, and diversity, all of which indicate L2 developmental level (McNamara et al., 2014). Lexical diversity, assessed by measure of textual lexical diversity (MTLD) rather than the type-token ratio, accounts for text length (Jarvis, 2013).

Cohesion, including verb and referential cohesion, ensures a coherent text. Cohesion refers to the degree of connectedness and logical flow within a text, ensuring that the presented ideas are unified and coherent (McNamara et al., 2014). Referential cohesion involves repeating or overlapping words, phrases, or semantic references across different text parts, such as clauses, sentences, and paragraphs. This overlap helps create explicit connections between different segments of the text, making it easier for readers to follow and understand the content (McNamara et al., 1996). Verb cohesion, conversely, refers to the repetition or similarity of verbs within a text. McNamara et al. (2012) found that verb cohesion tends to be higher in texts aimed at younger readers and narrative genres, where the focus is often on events rather than objects.

Finally, cohesion (verb cohesion, referential cohesion), lexical diversity (type-token ratio, MTLD), and syntactic complexity (left embeddedness, sentence syntax similarity) were chosen as key measures to determine which proofreading method most effectively supported L2 learners. These measures effectively reflect students' ability to produce clear texts, diverse vocabulary, and sophisticated sentence structures. Misspellings and punctuation errors were corrected before submitting the text for computational analysis to ensure that the software functioned correctly. To examine the main effects of proofreading mode on linguistic complexity indicators,

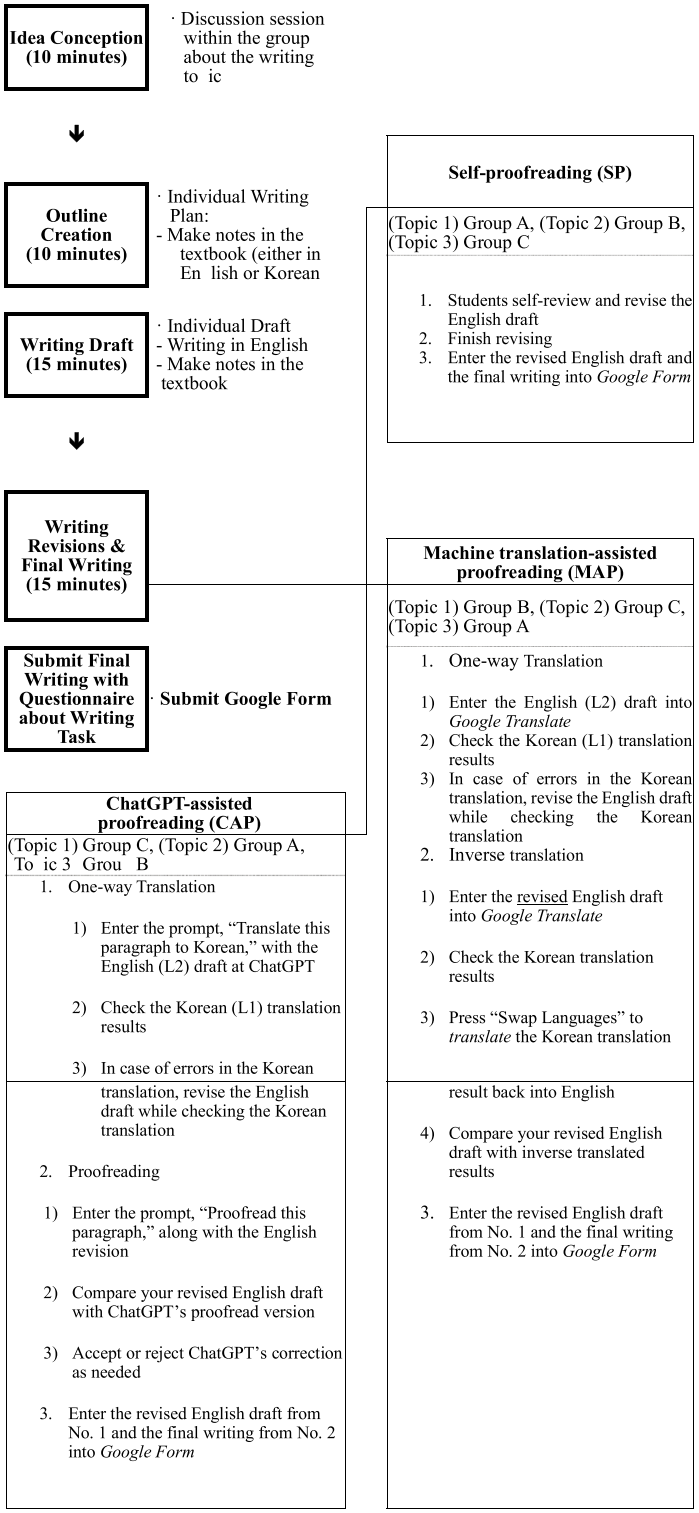


Fig. 2. Procedure of Writing Tasks.

RM one-way ANOVAs were conducted, and Bonferroni corrections were applied at.05/12 ($p < .004$) to control for the increased risk of Type 1 errors. Similarly, Bonferroni corrections were applied at.05/3 ($p < .017$) to control for inflated Type 1 error rates in all post-hoc multiple comparisons.

3.4.3. Analysis of writing errors

For Research Question 3, an error analysis was conducted on the texts produced under SP, MAP, and CAP using a taxonomy based on Ferris (2011) and Chon et al. (2021). The taxonomy included grammatical errors (e.g., articles and verb tense), lexical errors (e.g., word choice), and errors in sentence structure, mechanics, and mistranslation (see Appendix A for details). All errors were manually tallied and validated through consultation with another teacher, achieving 97 % agreement. Errors were normalized by calculating the occurrences per sentence. RM one-way ANOVAs were used to assess the effect of proofreading mode, with Bonferroni corrections applied to control for inflated Type 1 error rates.

4. Results

4.1. Rater's assessment of writing products

Each teacher independently assessed 237 essays written by learners using the SP, MAP, and CAP. The differences in their scores are presented in Table 3. Learners who used MAP and CAP significantly improved the completeness of their writing compared to SP. Texts produced with SP often contained incomplete sentences or paragraphs, inaccurately expressed arguments or irrelevant sentences. Conversely, learners using MAP achieved a mean score of 3.29, reflecting fewer incomplete sentences and clearer expressions. The texts written with CAP had the highest mean score of 3.85, indicating the highest level of completeness.

4.2. Analysis of linguistic complexity

4.2.1. Descriptive indices

Descriptive indices focused on sentence count, word count, average sentence length, and all indicators of writing fluency. The mean number of sentences across the three texts was approximately four, with no significant differences based on the feedback tool, likely because the guided writing activities limited the students to four sentences.

Depending on the feedback tool learners chose to use, there was a significant difference in word count but not in sentence length. Texts revised by learners with CAP had a mean word count of 53.92, higher than those revised with SP (45.49) or MAP (45.35). The nonsignificant difference in sentence length ($p = .385$) across tools reflects the controlled nature of the writing task and how learners used AI tools to modify text. Notably, the standard deviation for SP (8.64) decreased with MAP and CAP, suggesting that these tools helped learners improve their writing fluency, particularly those struggling to construct English sentences, even with sentence frames.

4.2.2. Cohesion

4.2.2.1. Text easability. As shown in Table 4, text ease was assessed through verb cohesion, and a significant effect of the proofreading tool was observed ($p < .0001$). Post-hoc tests revealed that texts produced by learners using SP (70.27 %) and MAP (67.69 %) had significantly higher verb cohesion than those produced using CAP (41.66 %). This suggests that users, through their interactions with ChatGPT, may see a reduction in verb frequency due to changes in sentence structure. For example, one student wrote, "Also you text with your friend make you **feel enjoy**," which ChatGPT revised to "Additionally, texting with your friend can bring you **enjoyment**."

4.2.2.2. Referential cohesion. Referential cohesion, which indicates text unity, was measured using noun repetition and content word overlap. Table 4 shows a significant difference in noun overlap between texts corrected using CAP (0.58) and MAP (0.42). The higher noun repetition in CAP texts suggests that users, with the help of ChatGPT, may compensate for lower overall cohesion by increasing noun repetition and improving text flow.

The proportion of content word overlap was higher in SP (0.25) and CAP (0.26) texts than in MAP (0.21) texts, although the main effect was not statistically significant. Descriptively, SP showed more content word repetition, whereas CAP texts included more content words to enhance cohesion. For instance, a learner wrote the incomplete sentence, "by having it [= by volunteering] we will grow," which, with the help of ChatGPT, was revised to "Volunteering can lead to personal growth and development, making it an essential experience," resulting in a more cohesive and contextually appropriate sentence, as presented in Fig. 3.

Table 3
Scoring of writing products by three proofreading tools.

(N = 79)	*M*	*SD*	*F* (2, 156)	*Post-Hoc*		
Self-proofreading	2.37	0.786	119.586***	1 < 2***	1 < 3***	3 > 2***
Machine translation-assisted proofreading	3.29	0.772				
ChatGPT-assisted proofreading	3.85	0.611				

Note: ***$p < .001$ (1 = self-proofreading, 2 = machine translation-assisted proofreading, 3 = ChatGPT-assisted proofreading).

Table 4
Linguistic complexity for three modes of feedback.

(N = 79)	Self-proofreading		Machine translation-assisted proofreading		ChatGPT-assisted proofreading		*F* (2, 156)	*Post-Hoc*		
	M	*SD*	*M*	*SD*	*M*	*SD*				
DESCRIPTIVE INDICES										
Sentence count (No. of sentences)	3.90	1.20	3.89	1.11	4.15	1.05	2.157	N/A		
Word count (No. of words)	45.49	12.88	45.35	13.99	53.92	15.33	14.364***	1 = 2	1 < 3***	2 < 3***
Sentence length (No. of words)	13.32	8.64	12.28	3.97	13.25	3.25	.960	N/A		
COHESION										
Text Easability										
Verb cohesion (percentile)	70.27	34.16	67.69	35.24	41.66	33.77	14.302***	1 = 2	1 > 3***	2 > 3***
Referential Cohesion										
Noun overlap (mean)	0.48	0.33	0.42	0.33	0.58	0.31	6.057*	1 = 2	1 = 3	2 < 3*
Content word overlap (proportional mean)	0.25	0.14	0.21	0.12	0.26	0.12	4.634	N/A		
Connective										
All connectives (incidence)	116.88	30.89	109.87	31.27	101.11	35.06	4.864	N/A		
LEXICAL DIVERSITY										
Type-token ratio (content word lemmas)	0.72	0.11	0.77	0.11	0.74	0.11	4.960	N/A		
Type-token ratio (all words)	0.70	0.09	0.72	0.10	0.71	0.09	2.327	N/A		
Measure of Textual Lexical Diversity [MTLD] (all words)	48.95	15.42	53.38	21.28	59.85	23.65	7.207*	1 = 2	1 < 3***	2 = 3
SYNTACTIC COMPLEXITY										
Left embeddedness, words before the main verb (mean)	2.87	1.46	2.98	1.25	3.57	1.60	5.683†	1 = 2	1 < 3*	2 = 3
Sentence syntax similarity, all combinations, across paragraphs (mean)	0.16	0.07	0.18	0.09	0.17	0.06	4.886	N/A		

Notes: (1 = self-proofreading, 2 = machine translation-assisted proofreading, 3 = ChatGPT-assisted proofreading)
Main Effects:
† p = .004 (Borderline significance level)
* p < .004 (Bonferroni correction)
Post-Hoc Tests:
* p < .017 (Bonferroni correction, post-hoc test)
General Significance:
** p < .001
*** p < .0001

(Learner no. 58)

<Sentences in the process of revision>
I think volunteer work should be one of the **requierments*** for graduation.
First of all we can other work to studying in school
also volunteer can be ensure because by **haveing*** it we will grow
(*Note*: the omission of periods is original)

<Final text>
I think volunteer work should be one of the requirements for graduation. First of all, it allows students to engage in other activities besides studying in school. Additionally, volunteering can lead to personal growth and development, making it an essential experience.

* = Error

Fig. 3. Texts proofread with ChatGPT.

4.2.2.3. Connectives incidence. Connective incidence refers to the frequency of all types of connectives (causal, logical, and contrastive). Descriptive statistics showed that learners using SP had a higher rate of connective occurrence in their texts (116.88) compared to those using CAP (101.11). Although not statistically significant, this difference may stem from the guided writing template, which included explicit connectives that learners used without modification. In contrast, users employing CAP produced more sophisticated sentences without relying heavily on explicit connectives. However, these results should be interpreted with caution because the proofreading method's main effect was insignificant.

4.2.3. Lexical diversity

Lexical diversity was measured using the type-token ratio (TTR) and MTLD. While the TTR results were not significant, the MTLD showed a significant effect ($p < .0042$), with learners producing more diverse vocabulary in CAP texts (59.85) than in SP texts (48.93). Unlike the TTR, the MTLD is a more valid measure because it accounts for text length. The higher word count in the CAP texts suggests that learners with ChatGPT assistance generated longer texts, naturally enhancing lexical diversity.

4.2.4. Syntactic complexity

Syntactic complexity was assessed using the number of words before the main verb (left-hand embeddedness) and syntactic structure similarity within a text. The proofreading method had a borderline significant effect on left-hand embeddedness ($p = .0042$), with CAP texts averaging 3.57 words before the main verb, higher than SP and MAP before the main verb. This indicates that ChatGPT may help learners add modifiers, thereby increasing sentence complexity. For example, ChatGPT revised "vounlteer* to no vounltly*" to "volunteering that is not done willingly is not genuine volunteering," demonstrating enhanced complexity. However, similarities in syntactic structure showed no significant differences, likely due to the guided nature of the writing task and genre constraints.

4.3. Analysis of errors

4.3.1. Grammatical errors

Significant differences in error frequency were found in articles, verb forms, and subject-verb agreement across the feedback tools. Although no difference was noted between the texts of learners using MAP and CAP, both groups significantly outperformed those using SP (see Appendix B for learner-applied corrections). As shown in Table 5, learners using the MAP and CAP improved their article usage accuracy. They also effectively corrected verb forms, such as changing "go outside and meet my friends" to "going outside and playing with my friends" and "talent cannot have everyone" to "talent cannot be had by everyone." CAP was particularly effective in helping learners correct subject-verb agreement errors, as seen in the change from "effort have" to "effort has."

Table 5
Differences in frequency of writing errors per sentence.

(N = 79)	Self-proofreading		Machine translation-assisted proofreading		ChatGPT-assisted proofreading		*F (df1, df2)*	*Post-hoc*		
	M	*SD*	*M*	*SD*	*M*	*SD*				
Article	0.18	0.37	0.01	0.05	0.01	0.05	15.50*** (1.03, 82.73)	1 > 2***	1 > 3***	2 = 3
Noun ending	0.06	0.24	0.03	0.09	0.00	0.04	2.41 (1.23, 95.66)	N/A		
Verb tense	0.02	0.12	0.01	0.11	0.00	0.00	0.82 (1.58, 123.25)	N/A		
Verb form	0.13	0.32	0.01	0.05	0.01	0.06	11.30** (1.09, 84.90)	1 > 2**	1 > 3**	2 = 3
Subject-verb agreement	0.05	0.13	0.00	0.03	0.00	0.02	10.51** (1.10, 86.01)	1 > 2*	1 > 3*	2 = 3
Prepositions	0.05	0.13	0.00	0.04	0.02	0.07	5.63 (1.29, 100.52)	N/A		
Pronouns	0.09	0.27	0.03	0.11	0.02	0.08	2.88 (1.33, 102.23)	N/A		
Word choice	0.08	0.26	0.10	0.18	0.07	0.17	0.43 (1.78, 137.36)	N/A		
Word form	0.06	0.16	0.01	0.05	0.00	0.00	8.88* (1.12, 87.39)	1 = 2	1 > 3**	2 = 3
Sentence structure	0.54	0.75	0.09	0.18	0.03	0.10	31.74*** (1.10, 85.48)	1 > 2***	1 > 3***	2 = 3
Mechanics	0.68	1.22	0.04	0.17	0.02	0.15	21.64*** (1.05, 82.22)	1 > 2***	1 > 3***	2 = 3
Mistranslations	0.33	0.39	0.11	0.22	0.01	0.054	31.07*** (1.46, 114.19)	1 > 2***	1 > 3***	2 > 3**

Notes: (1 = self-proofreading, 2 = machine translation-assisted proofreading, 3 = ChatGPT-assisted proofreading)
Main Effects:
† p = .004 (Borderline significance level)
* p < .004 (Bonferroni correction)
Post-Hoc Tests:
* p < .017 (Bonferroni correction, post-hoc test)
General Significance:
** p < .001
*** p < .0001

4.3.2. Lexical errors

Lexical errors are divided into word-choice and word-form errors. There were no significant differences in word-choice errors between the three feedback modes. With SP, learners often used familiar words to reduce errors. However, MT and ChatGPT did not always suggest the most appropriate words (see the Discussion). Word-form errors, such as incorrect parts of speech, occurred less frequently with CAP (0.00) than with SP (0.06). For example, ChatGPT helped learners correct "volunteer" to "volunteering" and "social" to "society" in the subject position. Although using MAP did not significantly differ from SP in reducing errors, it helped learners improve phrases like "kindness heart" to "warm and candid person." The MAP approach enables writers to check translations effectively and identify lexical and sentential errors.

4.3.3. Sentence structure errors

There were significant differences in sentence structure errors across the three feedback modes. The SP texts had a higher mean of 0.54 errors than MAP (0.09) and CAP (0.03) texts. Common errors include run-on sentences, misuse of commas, verb phrase errors, and incorrect word order, often due to a lack of basic sentence structure knowledge. Conversely, learners using the MAP or CAP effectively corrected these errors. For example, learner no. 68 used MAP to correct run-on sentences by adding "but" as a conjunction, while learner no. 25 connected main and subordinate clauses into a complete sentence after ChatGPT feedback. Although the difference between CAP and MAP did not meet the strict Bonferroni correction significance level, CAP generally resulted in fewer sentence structure errors, likely because of its ability to consider the context when generating sentences. Although the difference between CAP and MAP did not meet the strict Bonferroni correction significance level, learners using CAP generally made fewer sentence structure errors, likely due to the tool's capacity to consider context when generating sentences.

4.3.4. Mechanical errors

More errors were found in the SP than in the MAP or CAP. Spelling, punctuation, and capitalization errors appeared frequently in the SP texts. Capitalization errors were expected among these because the learners typed on their mobile phones, as in "Also, Music is relaxing my mind and body." Punctuation and spelling errors occurred, as seen in "Therefore oi* think best way is sleep." They did not consider these mistakes as sufficiently significant for correction. Learners were also unaware of punctuation errors, including periods. Conversely, the frequency of mechanical errors decreased significantly with MAP and CAP. With MAP, learner no. 35 who misspelled the English word "negative" as negatie* was able to correct spelling. Learner no. 52 wrote the conjunction "Also" at the beginning of a sentence without adhering to capitalization and punctuation rules, and the MT effectively corrected it. For learner 11, the capitalization and punctuation errors in the middle of the sentence were appropriately corrected using ChatGPT.

4.3.5. Mistranslation errors

As learners wrote their intended messages in L1, we assessed whether their L2 writing accurately conveyed them. The proofreading method significantly affected mistranslations ($p < .0001$). The SP texts had a higher mean of translation errors (0.33) than the MAP (0.11) and CAP (0.01). Learners who used MAP or CAP made fewer translation errors than those who used SP ($p < .0001$), with CAP being the most effective at accurately conveying the intended message. However, post-hoc tests revealed that learners using MAP made relatively more errors than those using CAP.

5. Discussion

This study uniquely contributes to the field by offering a comparative analysis of ChatGPT and neural MT as tools for AWE, an area rarely explored in L2 writing research. By incorporating a controlled experimental design with SP as a baseline, this research addresses gaps in prior studies by evaluating the tools' impact on linguistic complexity, accuracy, and holistic writing quality. These findings offer valuable insights into integrating AI tools into L2 writing pedagogy while emphasizing their strengths and limitations.

Both MAP and CAP significantly enhanced the overall quality of L2 writing compared to SP. In particular, the CAP resulted in the highest scores for completeness and coherence by helping to reduce incomplete sentences and improve the accuracy of argument expression. While MAP can help enhance writing quality, learners may find CAP more effective for producing polished, well-structured texts. The results align with findings that the use of AWE has shown positive effects on revision rates and editing efficiency, particularly when compared to conditions of no feedback (Fu et al., 2024). Additionally, ChatGPT has been found to enhance linguistic form and meaning (Athanassopoulos et al., 2023; Su et al., 2023).

CAP allowed learners to increase word count and lexical diversity, encouraging more elaborate and varied language usage. Learners also improved sentence structures with CAP (e.g., achieving higher left-hand embeddedness). However, they experienced lower verb cohesion, likely due to rephrasing verbs as nouns. This aligns with the observation that AWE feedback primarily supports the mechanical and linguistic aspects of writing but often falls short in enhancing coherence and content development, which are critical for higher-order writing features (Link et al., 2022; Stevenson & Phakiti, 2014).

Specifically, lower verb cohesion in ChatGPT-proofread texts may highlight several limitations of the AI tool, such as prioritizing sentence-level corrections over global cohesion, misinterpreting context, and introducing tense mismatches. For example, a student in the pre-CAP condition wrote: "I think eating food is the way to deal with a negative emotion First of all if i eat food l'm happy in addition while eat I don't think anything therefore when I take the stress,I prefer eating food than sleeping." CAP corrected this to: "I think eating food is a way to deal with negative emotions. Firstly, when I eat food, it makes me happy. Additionally, while eating, I can temporarily avoid thinking about anything else. Therefore, when I feel stressed, I often find solace in eating food rather than sleeping." While verbs like "is," "eat," "makes," "can temporarily avoid," and "feel stressed" maintain temporal consistency, the contrast between

"eating food" and "sleeping" lacks a logical link, reducing cohesion. Such cases, common with ChatGPT, show its reliance on assumed writer intent and its struggle with stylistic verb cohesion, highlighting the use of ChatGPT needing human oversight and validation to accurately express the writer's intent (Mizumoto et al., 2024; Steiss et al., 2024).

Regarding language accuracy, learners using MAP and CAP significantly reduced grammatical errors, particularly in articles, verb forms, and subject-verb agreement. CAP was slightly more effective than MAP at helping learners correct sentence structure errors, possibly due to its context-sensitive feedback, which aligns with the findings of Harunasari (2023). Learners also significantly reduced mechanical errors, such as spelling and punctuation, with both MAP and CAP, though CAP provided a slight advantage. Although the results cannot be directly compared to studies contrasting MT and ChatGPT due to the lack of such data, they align with the findings of Fu et al. (2024), which demonstrate that AWE systems are effective in improving grammar, mechanics, and sentence-level accuracy. Additionally, they are consistent with Algaraady and Mahyoob (2023), who highlighted ChatGPT's ability to identify surface-level errors.

Preposition errors remained consistent across all tools, highlighting learners' struggles with context-dependent features and the limitations of ChatGPT and earlier AWE tools in addressing nuanced issues (Barrot, 2023; Link et al., 2022; Stevenson & Phakiti, 2014). This highlights how AWE feedback effectively supports lower-order linguistic mechanics, such as grammar and syntax, but often falls short in addressing contextually complex issues like preposition use. Prepositions play a critical role in conveying relationships between ideas, objects, and actions—key elements of coherence and content development, which are essential for higher-order writing features. Although the absolute number of prepositional errors was relatively small (Pre-MAP: 6, Post-MAP: 4; Pre-CAP: 8, Post-CAP: 9), patterns reveal differences in how the two tools address them. Google Translate feedback often simplifies sentences to resolve prepositional errors directly. For example, "Volunteer will be just wasting time for them" (pre-MAP) was revised to "Volunteering will just waste their time," removing the ambiguous preposition "for." This restructuring results in a clearer expression. In comparison, ChatGPT may refine grammar and sentence structure but often overlook preposition errors, focusing instead on broader revisions. For instance, "For these reasons, I listen to music to deal with a negative emotion" (pre-CAP) remained largely unchanged in the post-CAP revision, retaining redundant use of "to." While grammatically correct, a more concise revision like "I listen to music as a way to cope with negative emotions" would enhance fluency. This highlights ChatGPT's strengths in improving lexical and structural aspects but its limitations in addressing finer stylistic redundancies (Steiss et al., 2024).

Neither MAP nor CAP proved effective in addressing errors related to noun endings, verb tense, pronouns, or word choice—areas often requiring human judgment and intervention. This supports previous observations that ChatGPT must be complemented by human oversight for nuanced understanding (Algaraady & Mahyoob, 2023; Escalante et al., 2023; Mizumoto et al., 2024). While ChatGPT handles surface-level errors well, it struggles with word choice and fully capturing the writer's intent. For instance, the uncorrected phrase "brain talent" and MT-revised awkward phrasing like "making a game" highlight the limitations of relying solely on AI. L2 writers may face challenges deciding whether to accept AI suggestions, potentially compounding errors. CAP-generated texts also did not consistently achieve the highest scores on holistic scales, underscoring AI's limitations in supporting higher-order writing without human guidance.

The nonsignificant impact of feedback on pronoun errors likely stemmed from the influence of L1. As a pro-drop language, Korean often omits subjects (Kwon & Sturt, 2013), making it challenging for both MT and ChatGPT to handle pronouns in translation. Additionally, ChatGPT's use of "they" as a gender-neutral singular pronoun, discussed by Alafnan and Mohdzuki (2023), may have contributed to errors in contexts where traditional pronoun use was expected. These findings highlight the importance of carefully considering how L1 phrases are constructed before inputting them into AI tools for translation (Jiang et al., 2024). By doing so, learners can better anticipate potential translation issues and adjust their input to align with the expected output more accurately.

Although this study did not explicitly assess learners' autonomy during the writing process involving MT and ChatGPT, their interaction with these tools demonstrated elements of independence and engagement. Students composed texts in English before using the tools, reflecting ownership of their learning. Their interaction with ChatGPT, in particular, highlighted a collaborative process. For example, learners used prompts like "Translate this" to evaluate whether their ideas were effectively expressed and interpreted by the AI.

Informal surveys by one of the researchers, also the learners' teacher, revealed positive outcomes. Learners noted that ChatGPT made their sentences smoother and more natural, resembling native-speaker writing. One student said, "It didn't sound too rigid but more like real English." Another shared, "I learned many words for 'happy' and realized that 'to' should not follow 'suggest.'" Learners also appreciated how ChatGPT expanded their vocabulary and corrected grammatical errors. However, close supervision is necessary to prevent overreliance on ChatGPT and ensure learners understand linguistic nuances (Cao & Zhong, 2023). Furthermore, the tool's multiple revision opportunities may inadvertently prioritize accuracy over other writing features, such as content development and coherence (Link et al., 2022).

What may go unnoticed is that instruction on linguistic complexity and error correction in L2 pedagogy is essential for effectively integrating MT and ChatGPT. These skills are particularly crucial for using ChatGPT, as evaluating AWE output requires advanced language proficiency. Critical instruction helps learners analyze cross-linguistic differences with MT, refine sentence-level accuracy with ChatGPT, and move beyond surface-level corrections, enabling them to enhance their overall proficiency and construct complex, contextually appropriate expressions.

Another often-overlooked issue in the use of ChatGPT is learning loss and cultural bias evident in ChatGPT's outputs (Jenks, 2024; Kohnke et al., 2023). L2 learners of English may unknowingly rely on culturally skewed information, as ChatGPT, heavily trained on English-language data, tends to reflect Western norms, idioms, and values, such as individualism over collectivism. Future research should focus on informing users about these biases, promoting critical evaluation, and fostering AI literacy and cross-cultural awareness to help learners recognize and mitigate such biases effectively. Jenks (2024) notes that "The biases that exist in AI-LLM

[artificial intelligence-large language models] are a real concern, as they are based on data gathered from existing societal ideologies and discourses that in part include prejudices, bigotry, and false representations" (p. 5).

6. Limitations and future directions

The study's product-oriented methodology offered valuable insights into the immediate outcomes of using MT and ChatGPT for proofreading but presented several limitations. First, it did not explore how learners perceived, analyzed, and adopted feedback or how they identified and revised issues in their drafts. The strategies used to engage with MT or prompts for ChatGPT were not learner-generated, limiting insights into independent corrective feedback practices. Second, the study focused on immediate changes without examining the delayed or sustained effects of using the tools. It remains unclear whether learners transferred the revisions suggested by MT and ChatGPT to improve future independent writing—an unresolved question common in AWE research.

These limitations highlight directions for future research. First, studies should investigate how learners perceive and apply feedback from MT and ChatGPT using methods like think-aloud protocols or screen tracking to uncover cognitive processes. Such research could reveal how learners identify issues and engage with feedback over time. Second, longitudinal studies are needed to evaluate the long-term impacts of MT and ChatGPT on L2 writing. Future research could explore whether repeated exposure to these tools improves subsequent writing tasks, second language acquisition, or metacognitive awareness, and whether learners develop independent proofreading strategies. Finally, since the error categories used for analysis focused heavily on surface-level grammar issues, future studies could benefit from an exploration of higher-level writing skills such as argumentation, coherence, or organization.

7. Conclusions

The rise of AI tools like ChatGPT, alongside earlier AWE tools (e.g., Grammarly, Criterion), has sparked both excitement and caution. While offering valuable support, these tools cannot replace human feedback due to inherent limitations. To enhance L2 writing development, learners should combine MAP and CAP methods with traditional proofreading techniques.

MAP with inverse translation functions can be used to sensitize L2 learners to cross-linguistic differences, while CAP is particularly effective for improving lexical diversity, syntactic complexity, and accuracy. However, human feedback remains essential for addressing areas where AI tools may fall short, such as prepositions, verb cohesion, and nuanced word choice. For instance, prepositions are highly context-dependent and difficult to translate directly. Additionally, ChatGPT may replace errors by generating entirely new sentences instead of correcting them, requiring learners to critically evaluate its suggestions.

Teachers can effectively integrate AI tools into language instruction by equipping learners with strategies to improve linguistic complexity and accuracy. By doing so, teachers can dedicate more attention to higher-order writing concerns, such as coherence, organization, and content development, which enrich the learning experience for L2 students. Additionally, this approach helps address the limitations of AI tools, including cultural biases and ethical concerns, ensuring their use remains balanced, responsible, and effective.

CRediT authorship contribution statement

Chon Yuah: Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Kim Minjoo:** Writing – original draft, Visualization, Project administration, Methodology, Investigation, Data curation.

Declaration of Competing Interest

The authors declare that there is no conflict of interest.

Appendix A. Major error categories and examples

Error	Subcategories	Error Description and Examples
Structural Errors	Article	· **Omission of Articles** E.g., Even if you are talent person with out some hard wort* your talent won't shine. (Self-proofreading, #7) (Target: Even if you are a talented person, your talent won't shine without some hard work.)
	Noun ending	· **Inaccurate Singular/Plural Forms of Nouns** E.g., In result, I think that we should do many volunteer works. (MT-proofreading, #36) (Target: Consequently, I think we should do a lot of volunteer work.)
	Verb tense	· **Incorrect or Inconsistent Verb Tenses** E.g., First of all when I *met* my friends, my emotion becomes relaxing (Self-proofreading, #19) (Target: First of all, when I met my friends, I became relaxed.)
	Verb form	· **Incorrect active/passive voice, confusion between infinitives and gerunds, etc.** E.g., First of all, the meaning of word "Volunteer" will be fades away if force and requirement contact. (Self-proofreading, #39) (Target: First of all, the meaning of the word "volunteer" will be faded away if students are forced and required to do volunteer work.)

(*continued on next page*)

(*continued*)

Error	Subcategories	Error Description and Examples
	Subject-verb agreement	· **Verbs not agreeing with the subject in singular/plural form** E.g., Also, volunteering should stem from a kind-hearted mindset, but neither of the students have a kind mindset. (ChatGPT proofreading, #18) (Target: Also, volunteering should stem from a kind-hearted mindset, but no student has a kind-hearted mindset.)
	Prepositions	· **Incorrect use or omission of prepositions** E.g., My friends are good listeners. And they're so funny. So when I talk my problem story, they make things fun and take away my worries. (MT-proofreading, #31) (Target: My friends are good listeners. And they're so funny. So, when I talk about my problem, they make me laugh and take away my worries.)
	Pronouns	· **Incorrect use of personal pronouns and relative pronouns** E.g., First of all, when someone eats spicy food, they forget about stressful situations. (ChatGPT proofreading, #10) (Target: First of all, when someone eats spicy food, he or she forgets about stressful situations.)
Lexical Errors	Word choice	· **Inaccurate word choice, unclear messages, awkward expressions, etc.** E.g., Therefore, being alone is the best way to vent bad feelings. (MT-proofreading, #38) (Target: Therefore, being alone is the best way to release bad feelings.)
	Word form	· **Inaccurate parts of speech or incorrect word forms** E.g., Also, there is no poiont* in doing voulunteer* work compulsory. (Self-proofreading, #28) (Target: Also, there is no point in doing volunteer work compulsorily.)
Sentence structure		· **Run-on sentences (including comma splices), fragments, verb phrase errors, word or phrase order errors, etc**. E.g., I think hanging out with friends is the best way to deal with negative emotions. First of all, because you can play with your friends without thinking. (MT-proofreading, #34) (Target: First of all, I think it is a good way because you can play with your friends without thinking.)
Mechanics		· **Spelling, punctuation, and capitalization errors** E.g., i think "Thinking again the reason why i am angry". (Self-proofreading, #53) (Target: I think it is "thinking again the reason why I am angry.")
Mistranslations		· **Incomprehensible and inaccurate translations, words, expressions** E.g., Also, sports talent often helps to overcome the gap of hard work rapidly. (ChatGPT proofreading, #49) (Target: Also, talent for sports often helps to overcome the difference in the amount of effort.)

Note: MT = Machine Translation

Appendix B. Errors and corrections with machine translation- and ChatGPT-proofreading

	Machine translation-proofreading	ChatGPT-proofreading
Article	**(#10)** (Before) Frist* of all, volunteer work is basic mind that you shoud* have. (After) Above all, volunteer work is **a** basic mindset.	**(#39)** (Before) Someone who studied with hard work couldn't get over the gap between talented student. (After) Someone who studied with hard work couldn't bridge the gap between **a** talented student.
Verb form	**(#30)** (Before) I think go outside and meet my friends is the best way to deal with a negative emotion. (After) Go**ing** outside and play**ing** with friends is a way to relieve stress.	**(#27)** (Before) The first reason is that talent cannot have everyone. (After) The first reason is that talent cannot **be had** by everyone.
Subject-verb agreement	**(#67)** (Before) First of all, We'll form a bond because volunteer work help our relation development (After) First of all, we form a bond because volunteer work develop**s** our interpersonal relationship.	**(#45)** (Before) Also efford* have a limit. (After) Additionally, effort **has** its limits.
Prepositions	**(#38)** (Before) Also, It helps to see that situation **in** objectively. (After) It also helps you to look at the situation objectively. **(#32)** (Before) Also if there is a character who is same me I can get encouragement from him or her. (After) Also, if there is a character **like** me, I can get strength from him.	**(#52)** (Before) because physical ablity* is important many sports. (After) because physical ablity* is important **in** many sports.
Word Form	**(#27)** (Before) First of all, exercising removes useless think. (After) First of all, exercise removes useless thoughts. **(#73)** (Before) Also They have kindness heart. (After) They can also be warm and candid.	**(#17)** (Before) First of all, If volunteer not involve in graduation requirements, We have little chance to approach about voluntary volunteer. (After) First of all, if volunteering is not involved in graduation requirements, we have little chance to learn about voluntary service. **(#22)** (Before) If many people doesn't do volunteer, social would destroy (After) If many people don't participate in volunteer work, society can easily deteriorate.

(*continued on next page*)

(*continued*)

	Machine translation-proofreading	ChatGPT-proofreading
Sentence structure	**(#68)** (Before) Also, anyone can do this, i think when they finished volunteer they feel proud themself*. (After) Also, anyone can do it, but I think I feel proud when I finish volunteering. **(#68)** (Before) Also, anyone can do this, i think when they finished volunteer they feel proud themself*. (After) Also, anyone can do it, but I think I feel proud when I finish volunteering.	**(#25)** (Before) I think talent is more important than hard work. Because the players who are respected to other of playground is good at playing. (After) I think talent is more important than hard work because players who are good at playing on the playground are respected by others. **(#25)** (Before) I think talent is more important than hard work. Because the players who are respected to other of playground is good at playing. (After) I think talent is more important than hard work because players who are good at playing on the playground are respected by others.
Mechanics	**(#35)** (Before) I think playing drum is the best way to deal with a negatie* emotion. (After) I think playing drums is the best way to deal with negative emotions. **(#52)** (Before) also doing hard working is limited. (After) Also, working hard is limited.	**(#11)** (Before) First, During sleeping, brain don't think anything. Also after wake up, body and brain get fresh (After) First, during sleep, the brain doesn't think about anything. Additionally, after waking up, the body and brain feel refreshed.

Data availability

Data will be made available on request.

References

Alafnan, M. A., & Mohdzuki, S. F. (2023). Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial Intelligence and Technology, 3*(3), 85–94. https://doi.org/10.37965/jait.2023.0267

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal*, (9), 3–17. Retrieved from ⟨https://ssrn.com/abstract=4534530⟩.

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. Q. (2023). Impact of ChatGPT on learning motivation: Teachers and students' voices. *Journal of English Studies in Arabia Felix, 2*(1), 41–49. https://doi.org/10.56540/jesaf.v2i1.51

Alrajhi, A. S. (2023). Genre effect on Google Translate-assisted L2 writing output quality. *ReCALL, 35*(3), 305–320. https://doi.org/10.1017/S0958344022000143

Athanassopoulos, S., Manoli, P., Gouvi, M., Lavidas, K., & Komis, V. (2023). The use of ChatGPT as a learning tool to improve foreign language writing in a multilingual and multicultural classroom. *Advances in Mobile Learning Educational Research, 3*(2), 818–824. https://doi.org/10.25082/AMLER.2023.02.009

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology, 37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, Article 100745. https://doi.org/10.1016/j.asw.2023.100745

Barrot, J. S. (2021). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning, 36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning, 5*(1), 44–55. https://doi.org/10.31763/ijele.v5i1.858

Beiler, I. R., & Dewilde, J. (2020). Translation as translingual writing practice in English as an additional language. *Modern Language Journal, 104*(3), 533–549. https://doi.org/10.1111/modl.12660

Cancino, M., & Panes, J. (2021). The impact of Google Translate on L2 writing quality measures: Evidence from Chilean EFL high school learners. *System, 98*, 1–11. https://doi.org/10.1016/j.system.2021.102464

Cao, S., & Zhong, L. (2023). *Exploring the effectiveness of ChatGPT-Based Feedback compared with Teacher Feedback and self-Feedback: Evidence from Chinese to English translation arXiv Preprint*. https://doi.org/10.48550/arXiv.2309.01645. arXiv:2309.01645. Retrieved from.

Chang, P., Chen, P. J., & Lai, L. L. (2022). Recursive editing with Google Translate: The impact on writing and error correction. *Computer Assisted Language Learning, 37*(7), 2116–2141. https://doi.org/10.1080/09588221.2022.2147192

Chen, M., & Cui, Y. (2022). The effects of AWE and peer feedback on cohesion and coherence in continuation writing. *Journal of Second Language Writing, 57*, Article 100915. https://doi.org/10.1016/j.jslw.2022.100915

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. (2021). 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. https://doi.org/10.1080/09588221.2020.1871029

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. https://doi.org/10.1016/j.jslw.2009.02.002

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006

Echevarria, J. (2016). *Are Language frames good for English learners? Reflections on Teaching English Language learners*. Retrieved from ⟨http://www.janaechevarria.com/?p=191⟩.

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), 57. https://doi.org/10.1016/j.asw.2014.03.006

Ferris, D. (2011). *Treatment of error in second language student writing*. University of Michigan Press.

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2024). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning, 37*(1–2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223–234. https://doi.org/10.3102/0013189X114132

Guo, K., & Wang, D. (2024). To resist it or to embrace it? Examining ChatGPT's potential to support teacher feedback in EFL writing. *Education and Information Technologies, 29*(7), 8435–8463. https://doi.org/10.1007/s10639-023-12146-0
Harunasari, S. Y. (2023). Examining the effectiveness of AI-integrated approach in EFL writing: A case of ChatGPT. *International Journal of Progressive Sciences and Technologies, 39*(2), 357–368. https://doi.org/10.52155/ijpsat.v39.2.5516
Hwang, K.-H., Heywood, D., & Carrier, J. (2023). The implementation of ChatGPT-assisted writing instruction in ESL/EFL classrooms. *New Korean Journal of English Language and Literature, 65*(3), 83–106. https://doi.org/10.25151/nkje.2023.65.3.004
Ingley, S. J., & Pack, A. (2023). Leveraging AI tools to develop the writer rather than the writing. *Trends in Ecology and Evolution, 38*(9), 785–787. https://doi.org/10.1016/j.tree.2023.05.007
Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*(1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x
Jenks, C. J. (2024). Communicating the cultural other: Trust and bias in generative AI and large language models. *Applied Linguistics Review, 16*(2), 787–795. https://doi.org/10.1515/applirev-2024-0196
Jia, Y., Carl, M., & Wang, X. (2019). How does the post-editing of neural machine translation compare with from-scratch translation? A product and process study. *Journal of Specialised Translation, 31*, 60–86. ⟨https://jostrans.soap2.ch/issue31/art_jia.php⟩.
Jiang, L., Yu, R., & Zhao, Y. (2024). Theoretical perspectives and factors influencing machine translation use in L2 writing: A scoping review. *Journal of Second Language Writing, 64*, Article 101099. https://doi.org/10.1016/j.jslw.2024.101099
Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Islem- ChatGPT a good translator? Yes, with GPT-4 as the Engine arXiv*. Retrieved from ⟨https://arxiv.org/abs/2301.08745⟩.
Joo, H., Kim, M., Kim, S., Bae, J., Kang, H., Kim, K., & Choi, W. (2022). *A study on the development of the 2022 revised national English curriculum*. Korea Institute for Curriculum and Evaluation.
Kim, S., Yoon, J., Koo, E., Jeon, H., Seo, J., Lee, H., & Shin, Y. (2018). *High school English*. NE Neungyule, Co., Ltd.
Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELC Journal, 54*(2), 537–550. https://doi.org/10.1177/00336882231162868
Konttinen, K., Salmi, L., & Koponen, M. (2020). Revision and post-editing competences in translator education. In M. Koponen, B. Mossop, I.S. Robert, & G. Scocchera (Eds.), Translation revision and post-editing: Industry practices and cognitive processes (pp. 187–202). New York: Routledge. https://doi.org/10.4324/9781003096962-15.
Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *Teaching English as a Second or Foreign Language–TESL-Ejemplo, 27*(3), n3. https://doi.org/10.55593/ej.27107int
Kwon, N., & Sturt, P. (2013). Null pronominal (pro) resolution in Korean, a discourse-oriented language. *Language and Cognitive Processes, 28*(3), 377–387. https://doi.org/10.1080/01690965.2011.645314
Lee, S.-M. (2022). Different effects of machine translation on L2 revisions across students' L2 writing proficiency levels. *Language Learning and Technology, 26*(1), 1–21. ⟨https://hdl.handle.net/10125/73490⟩.
Li, J., & Li, M. (2018). Turnitin and peer review in ESL academic writing classrooms. *Language Learning and Technology, 22*(1), 27–41. https://dx.doi.org/10125/44576.
Li, J., Link, S., & Hegelheimer, V. (2015). Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing, 27*, 1–18. https://doi.org/10.1016/j.jslw.2014.10.004
Liao, H. C. (2016). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal, 70*(3), 308–319. https://doi.org/10.1093/elt/ccv058
Link, S., Dursun, A., Karakaya, K., & Hegelheimer, V. (2014). Towards best ESL practices for implementing automated writing evaluation. *CALICO Journal, 31*(3), 323–344. https://doi.org/10.11139/cj.31.3.323-344
Link, S., Mehrzad, M., & Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. *Computer Assisted Language Learning, 35*(4), 605–634. https://doi.org/10.1080/09588221.2020.1743323
Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859
Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Routledge.
McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In In. J. Sabatini, E. Albro, & T. O.' Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Rowman & Littlefield Publishing Group Education.
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664
McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1
Mindner, L., Schlippe, T., & Schaaff, K. (2023). Classification of human- and AI-generated texts: Investigating features for ChatGPT. In In. T. Schlippe, E. C. K. Cheng, & T. Wang (Eds.), *Artificial intelligence in education technologies: New development and innovative practices* (pp. 152–170). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7947-9_12.
Mizumoto, A., Shintani, N., Sasaki, M., & Teng, M. F. (2024). Testing the viability of ChatGPT as a companion in L2 writing accuracy assessment. *Research Methods in Applied Linguistics, 3*(2), Article 100116. https://doi.org/10.1016/j.rmal.2024.100116
Mundt, K., & Groves, M. (2016). A double-edged sword: The merits and the policy implications of Google Translate in higher education. *European Journal of Higher Education, 6*(4), 387–401. https://doi.org/10.1080/21568235.2016.1172248
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578. https://doi.org/10.1017/S0958344020000191
Ranalli, J. (2018). Automated written corrective feedback: How well can students make use of it? *Computer Assisted Language Learning, 31*(7), 653–674. https://doi.org/10.1080/09588221.2018.1428994
Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems, 3*, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003
Sanz-Valdivieso, L., & López-Arroyo, B. (2023). Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? *In International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023)*, 97–107. https://doi.org/10.26615/issn.2683-0078.2023_008
Schmidt-Fajlik, R. (2023). ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal, 14*(1), 105–119. https://doi.org/10.54855/acoj.231417
Shin, D., & Chon, Y. V. (2023). Second language learners' post-editing strategies for machine translation errors. *27*(1), 1–25. https://hdl.handle.net/10125/73523.
Son, J., & Kim, B. (2023). Translation performance from the user's perspective of large language models and neural machine translation systems. *Information, 14*(10), 574. https://doi.org/10.3390/info14100574
Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction, 91*(101894), *1–15*. https://doi.org/10.1016/j.learninstruc.2024.101894
Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007
Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, Article 100752. https://doi.org/10.1016/j.asw.2023.100752
Tsai, S. C. (2019). Using google translate in EFL drafts: A preliminary investigation. *Computer Assisted Language Learning, 32*(5–6), 510–526. https://doi.org/10.1080/09588221.2018.1527361

Wang, Y. J., Shang, H. F., & Briody, P. (2013). Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning, 26*(3), 234–257. https://doi.org/10.1080/09588221.2012.655300

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal, 3*(1), 22–36. https://doi.org/10.1080/15544800701771580

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers and Education, 100*, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Minjoo Kim received her master's degree in English Education from Hanyang University. She currently teaches at Eunpyeong High School in Seoul, South Korea as a certified English teacher and has worked in public secondary schools for over eight years. Her interest lies in technology-assisted second language writing and mobile-assisted language learning.

Yuah V. Chon received her PhD in ELT from University of Essex. She is a professor in the Department of English Education at Hanyang University, South Korea. She has published peer-refereed journal articles on a wide range of topics, including second language writing with machine translators.