Scaffolded AI Feedback for L2 Writing: Fostering Self-Correction in Japanese University Students

Atsushi Nakanishi

Proceedings of the 39th Pacific Asia Conference on
Language, Information and Computation (PACLIC 39)

Emmanuele Chersoni, Jong-Bok Kim (eds.)

2025

Scaffolded AI Feedback for L2 Writing: Fostering Self-Correction in Japanese University Students

Atsushi Nakanishi
Osaka Institute of Technology
atsushi.nakanishi@oit.ac.jp

Abstract

This study introduces AI-assisted Translation Learning and Search (ATLaS), an AI-powered system designed to enhance Japanese university students' English writing through translation-based learning. Grounded in the constructivist learning theory and Vygotsky's Zone of Proximal Development, the system uses scaffolded feedback using Ferris's hierarchical error taxonomy, promoting metalinguistic awareness through staged intervention rather than direct correction. A case study of 26 Japanese university students demonstrated significant improvements, with holistic writing scores increasing from 2.3 to 3.3 points (43% improvement). The system identified an average of 6.7 corrections per student, with 67% of the flagged errors being independently corrected by the learners. The error analysis of 175 instances revealed distinctive patterns: elevated lexical errors driven by word-choice difficulties, in contrast to reduced morphological errors in traditional contexts. The results suggest that AI-assisted feedback systems effectively supported L2 writing development when incorporating appropriate pedagogical scaffolding.

1 Introduction

1.1 Automated written corrective feedback

Traditional L2 writing instruction struggles to provide detailed, individualized, and timely feedback due to large class sizes and limited teacher availability. Although teacher feedback remains valuable, it is often not sufficiently scalable to guide learners through the recursive processes of drafting, reflection, and revision. This pedagogical gap highlights the potential for technology-enhanced learning environments that supplement conventional instruction. Sophisticated AI, particularly large language models (LLMs), enables immediate, data-driven, and personalized feedback (Mizumoto, 2025; Woo et al., 2024).

Although students often perceive teacher feedback to be authoritative and reliable, it has inherent limitations. Teachers, frequently constrained by time and large class sizes, tend to focus on local issues, particularly grammar, and sometimes neglect higher-order concerns such as content and organization. This has generated interest in automated writing evaluation (AWE) systems that provide automated written corrective feedback (AWCF) to reduce teacher workload and offer students immediate support (Feng et al., 2025).

However, AWCF's pedagogical effectiveness remains debated, with complex implementation challenges. Some empirical studies have indicated that the AWCF does not necessarily improve writing quality. For example, Fan (2023) found no significant difference in writing quality between lower-proficiency EFL students who received combined Grammarly and teacher feedback and those who received teacher feedback alone. This lack of improvement can be attributed to several factors, including learners' low proficiency level, which hinders their ability to understand and process feedback, and their general unfamiliarity with AWE tools.

Beyond feedback accuracy and comprehensibility, a more critical pedagogical concern is the risk of learners developing over-reliance on automated systems. For instance, Karatay and Karatay (2024) highlighted that students may develop trust levels that lead them to accept automated suggestions uncritically, without engaging in thoughtful analysis and deliberation essential for skill development. This passive "correction" behavior can inhibit the growth of learners' autonomy and their ability to self-edit.

These limitations indicate that neither teachers nor automated feedback can provide a complete solution. Instead, a growing body of research has identified the importance of an integrated approach that leverages the strengths of both approaches.

A quasi-experimental study by Cheng and Zhang (2024) demonstrated the potential synergy of AWE-teacher integrated feedback. In their model, the AWE system first addressed local-level language errors, allowing teachers to focus their feedback on global issues such as content and organization. This integrated approach not only led to significantly greater improvements in all aspects of writing performance compared to teacher feedback alone but also promoted a deeper behavioral and cognitive engagement from students. By creating a writing-feedback-revision cycle, such a model helps learners move beyond the belief that writing is a one-time task and, thus, recognize the importance of revision.

This study suggests that the most promising path forward lies in a thoughtful human–machine partnership. The goal is not to replace the expertise of human instructors, but augment their capabilities, thus creating a more efficient and effective feedback ecosystem. The AI-assisted Translation Learning and Search (ATLaS) system proposed in this study is based on this principle. It is designed to address the shortcomings of conventional AWCF by providing a scaffolded learning environment that not only offers corrective feedback but also supports the deeper learning processes necessary for long-term writing development.

1.2 Error analysis

The systematic analysis of learner errors constitutes a fundamental component of second language acquisition research, facilitated by the development of large-scale learner corpora and assessment datasets, such as the ICNALE Global Rating Archives (Ishikawa, 2023). These resources enable a comprehensive investigation of the linguistic characteristics of learners' language. Recent advances in AI have further enhanced this field through automated error analysis tools, such as the Auto Error Analyzer (Mizumoto, 2025), which automates accuracy metric calculations in learner texts. These developments underscore the necessity for a systematic and theoretically grounded classification framework to effectively categorize learner errors.

For the development of the ATLaS system, a robust error taxonomy is essential to provide structured and comprehensible feedback to learners. As such, this study adopted the comprehensive error classification framework proposed by Ferris (2011), which is widely recognized for its application in analyzing and treating errors in L2 students' writing.

Error Category	Subcategory
Morphological	Verbs [Tense, Form, Subject-verb agreement], Nouns [Articles/determiners, Noun endings (plural/possessive)]
Lexical	Word choice, Word form, Informal usage, Idiom error, Pronoun error
Syntactic	Sentence structure, Run-ons, Fragments
Mechanical	Punctuation, Spelling
Miscellaneous	Unclassified errors

Table 1: Error classification framework.

This framework organizes errors into five primary domains integrated into the ATLaS error analysis engine, as detailed in Table 1.

By operationalizing this established taxonomy, ATLaS was designed to provide learners with feedback that is both accurate and pedagogically organized, enabling them to understand the specific nature of their errors and facilitate targeted improvement strategies.

1.3 Research objectives

This study introduces the ATLaS system. This system was designed not only to correct errors but also to promote deeper metalinguistic awareness and encourage self-regulated learning. By leveraging a powerful AI model, ATLaS provides users with detailed feedback on their Japanese-to-English translations and classifies errors into a hierarchical system of grammatical, lexical, structural, and stylistic categories.

The development of the system serves two primary purposes: (1) the enhancement of self-correction abilities in translation learning through a gradual feedback provision system, and (2) the verification of error-type specific learning support effectiveness. This study also reports on the implementation of ATLaS with 26 Japanese university students in a classroom setting, examining its impact on translation accuracy and learner engagement in error-correction processes.

2 System development

2.1 ATLaS design

ATLaS was developed in accordance with the constructivist learning theory and Vygotsky's Zone of Proximal Development (ZPD), wherein AI-generated feedback functions as a mediating instrument to facilitate the transition between students' existing translation competencies and their poten-

tial performance. Instead of offering instantaneous corrections, the system employs a staged intervention approach that fosters reflective self-correction processes. This methodology aligns with existing research on indirect corrective feedback, which enhances reflective thinking and problem-solving abilities.

ATLaS consists of two fundamental operational modes, each tailored to address distinct facets of translation learning: the Translation Feedback Mode and the Error Search Mode. The former offers structured AI-mediated correction assistance for individual translation assignments, while the latter allows learners to investigate aggregated error patterns and examples organized in accordance with established linguistic taxonomies. These complementary modes function synergistically to facilitate immediate learning requirements and foster long-term metalinguistic development. The system leverages GPT-4 as its core language processing engine, selected for its advanced multilingual capabilities and JSON-structured output support. The model operates with a temperature setting of 0.7 to balance creativity and consistency in feedback generation, with a maximum token limit of 4,000 to ensure comprehensive explanations while maintaining processing efficiency.

2.2 Translation Feedback Mode

The Translation Feedback Mode facilitates structured correction, helping learners progressively identify and address errors. This mode functions via a systematic workflow that includes text input, AI-assisted analysis, structured feedback provision, and revisions initiated by the learner.

Figure 1: Interface of the Translation Feedback Mode.

The system analyzes Japanese source texts and English translations using GPT-4 to identify up to 10 significant issues. Carefully engineered prompts establish the AI as a bilingual instructor prioritizing educational scaffolding over direct correction. The prompt engineering incorporates two key components: (1) a system prompt that defines the AI's instructional role and output format requirements, and (2) a user prompt that provides the specific translation task context. For example, the system prompt instructs the AI to "identify up to 10 important issues" and constrains error classification to predefined taxonomic categories, ensuring consistency with Ferris (2011) framework. The error type constraint is implemented through explicit enumeration: "The 'error_type' field MUST be exactly one of the predefined error types. Do not create new error type names."

A typical user prompt structure follows this template: "Please evaluate the following translation based on the instructions. Japanese Text: ```[source text]``` English Translation: ```[student translation]```" This format provides clear task boundaries while maintaining consistency across all system interactions.

The system generates structured JSON responses that consist of three components: marked_text, feedback_message, and correction_table. The structure of the correction_table enables comprehensive feedback delivery through the following components (Table 2).

This structured format enables the system to present feedback in a pedagogically organized manner, with each correction including contextual information, explicit error categorization, and scaffolded guidance questions.

Component	Description
japanese	Complete original Japanese sentence providing source context
original	User's complete original English sentence for comparison
correction	Proposed correct English sentence demonstrating target form
explanation	Detailed explanation in Japanese clarifying linguistic principles
error_type	Predefined error classification based on Ferris taxonomy
prompting_question	Guiding question in Japanese with error marker references

Table 2: Correction_table structure.

The system generates numbered markers within the original translation text, thereby establishing

clear visual connections between the identified issues and their corresponding feedback elements, as shown in Figure 2. This marking system enables learners to focus their attention on specific problematic segments while maintaining overall text coherence and understanding the translation.

The hint-generation mechanism produces culturally appropriate prompting questions in Japanese, which assist learners in self-correction without explicitly providing answers. These questions were designed to elicit relevant linguistic knowledge, while fostering metacognitive reflection on translation decisions. For example, verb tense errors may prompt questions regarding the temporal relationships between events, whereas article errors may pertain to patterns of noun countability and definiteness.

The correction submission interface illustrated in Figure 2 provides learners with an editable version of their original translation, facilitating direct text modification while maintaining the associations with the original error markers. This approach enables focused attention on the identified issues while concurrently allowing for a comprehensive revision of the entire translation. The system uses both the initial and revised translation versions for comparative analysis and learning assessment.

Upon correction submission, ATLaS provides comprehensive feedback, including corrected versions of each identified error, detailed explanations of the underlying linguistic principles, and error-type classifications. The explanations are provided in Japanese to ensure accessibility and comprehension by integrating examples and contrasts that clarify the rationale for the suggested corrections.

The feedback presentation shown in Figure 3 employs a structured table format that delineates the original text, corrections, and explanations, enabling systematic comparison and analysis. Error-type classification enables learners to recognize patterns in their translation challenges and develop targeted improvement strategies for future learning activities.

2.3 Error Search Mode

The Error Search Mode organizes accumulated error patterns using Ferris (2011) framework, enabling systematic error exploration for skill development. Users interact with the system via a hierarchical interface and select the main error categories from a dropdown menu that updates subcategory options contingent on the available data. Upon the selection of specific error types, the system queries its database of correction instances and presents comprehensive concordance data.

Figure 2: Feedback interface.

The search results present the Japanese source text, original erroneous translations, corrected versions, and detailed explanations in a structured format (Figure 4). This systematic structure facilitates pattern recognition across multiple instances of similar errors while promoting both individual and collaborative learning through shared error analysis.

Each result entry provides comprehensive contextual information, enabling learners' comprehension of not only error identification but also the rationale for specific corrections within particular contexts. The concordance display presents multiple examples of similar error types, allowing learners to identify common patterns and the underlying

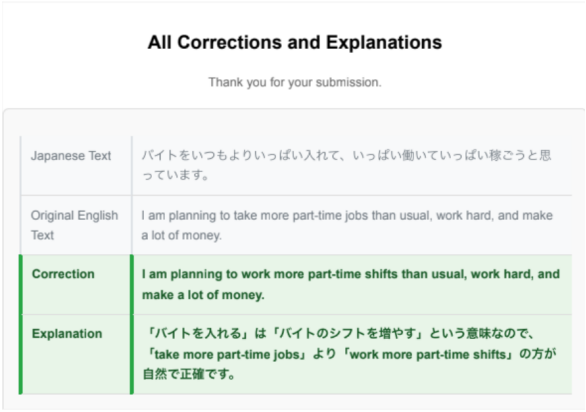

Figure 3: Final feedback interface.

Figure 4: Search results interface of "Tense."

linguistic principles.

This mode supports autonomous learning by enabling learners to search for personally encountered error types to reinforce their learning experience or explore unfamiliar categories to develop broader linguistic awareness. By providing access to the accumulated error patterns, the Error Search Mode reduces the dependence on immediate feedback and simultaneously promotes independent learning capabilities and metalinguistic awareness essential for long-term language development in line with the system's pedagogical goals.

3 Methodology

3.1 Research design

This study employed a single-group case study design to investigate the effectiveness of ATLaS in enhancing the English writing performance of Japanese university students. The design was deliberately selected to examine the learning processes and system interactions within an authentic classroom context, thereby facilitating the comprehensive documentation of improvement patterns and error-correction behaviors.

The study was conducted over two months (May–June 2025) with 26 Japanese university students enrolled in an "English Usage" course. A pre-post design was implemented to measure improvements in writing quality using holistic scoring rubrics, and a systematic error analysis was conducted on 175 correction instances categorized according to Ferris (2011) taxonomic framework.

Data collection focused on quantitative measures, including: (1) holistic writing scores using a five-point rubric, (2) correction frequency per student, and (3) error type distribution across morphological, lexical, syntactic, and mechanical categories. The qualitative analysis examined error patterns and improvement trajectories based on the systematic content analysis of the translation samples.

Ethical considerations were addressed through integrated informed consent procedures within the web application interface to ensure voluntary participation without academic coercion.

Several limitations should be noted. The single-group design limits causal inferences, as improvements may reflect the combined effects of AI feedback and concurrent instructor guidance rather than that of ATLaS alone. Additionally, the single writing topic constrains generalizability across different genres and discourse types.

3.2 Participants

The study employed a convenience sampling methodology by recruiting participants from an existing "English Usage" class at a private Japanese university. Initially, 28 third-year students who enrolled in the course were invited to participate. However, the final sample consisted of 26 students due to attrition: one student discontinued class attendance during the study period and another failed to submit the required assignments. An attrition rate of 7.1% was considered acceptable for educational research.

Participants' English proficiency levels were assessed using TOEIC Listening and Reading scores, which provide standardized measures of language ability. The proficiency distribution revealed considerable variation, with scores ranging from 165 to 635. The mean score was 377.5 points (SD = 125.84), whereas the median score was 367.5

points. This distribution indicates a predominantly intermediate-low to intermediate proficiency level, with some participants demonstrating more advanced abilities.

All participants were native Japanese speakers studying at a private university where instruction is conducted primarily in Japanese. The students were from the Faculty of Information Science, indicating that English was not their primary academic focus, but rather a supplementary skill requirement.

3.3 Implementation procedures

The implementation followed a four-session protocol advancing translation learning through AI feedback. Students worked on "Summer Vacation Plans" to maintain consistency while encouraging natural expression.

Session 1 entailed the development of original Japanese compositions (maximum 500 characters) alongside engaging in pre-editing activities to clarify ambiguous expressions and simplify complex structures. Session 2 focused on translating pre-edited texts into English using Microsoft Word, with dictionary access permitted; however, the use of AI translation tools prohibited the assessment of authentic linguistic competence.

Session 3 implemented a core intervention using the Translation Feedback Mode. Students submitted their initial translations and received AI-generated scaffolded feedback, which included error markers, contextual hints, and pedagogical explanations. Following the review of the feedback, the students revised their translations and resubmitted the corrected versions through the platform.

Session 4 used the Error Search Mode, enabling students to explore systematic error patterns in their work and in peer examples. This analytical phase developed students' metalinguistic awareness of common translation challenges while familiarizing them with a comprehensive error-classification framework.

Throughout the process, the students submitted four text versions: the original Japanese text, the pre-edited Japanese version, the initial English translation, and the ATLaS-revised English text. Instructor feedback complemented the system's local linguistic focus by addressing global issues, including content organization, coherence, and communicative effectiveness, thus creating an integrated feedback environment that addresses both surface-level accuracy and higher-order writing concerns.

3.4 Data collection and analysis

Data collection used multiple mechanisms to capture information about learning processes and system effectiveness. The primary data source consisted of automatically logged user interactions, including original Japanese texts, initial English translations, AI-generated correction feedback, and final revised versions. The quantitative metrics focused on measurable learning outcomes and system usage patterns. The key variables included (1) the number of corrections per student, (2) the distribution of error types according to Ferris (2011) taxonomy, and (3) holistic writing quality scores using a five-point rubric.

Writing quality was assessed using a standardized five-point holistic scoring rubric administered through OpenAI's GPT-4, evaluating both the initial and revised translations for grammatical accuracy, lexical appropriateness, syntactic complexity, and overall coherence. To determine if the change in writing quality was statistically significant, a paired-samples t-test was used to compare the holistic scores before and after the intervention.

A qualitative analysis was conducted through the systematic content analysis of the translation samples, focusing on error categorization and improvement patterns. A hierarchical error taxonomy based on Ferris (2011) framework categorized 175 correction instances into four primary domains. The error pattern analysis employed frequency distribution comparisons with established research and qualitative examinations of the characteristic difficulties faced by Japanese learners.

4 Results

4.1 Learning effectiveness

The statistical analysis demonstrated consistent improvement patterns across multiple writing quality dimensions. The mean number of corrections per student was 6.7 (SD = 2.4, range = 2–10), indicating that the system successfully identified meaningful improvement opportunities for learners across different proficiency levels. The distribution of correction frequencies showed that 46.2% of the participants received 4–6 corrections, which indicates an optimal cognitive load for learning effectiveness.

A quantitative assessment of learning effectiveness revealed substantial improvements across multiple writing quality dimensions. Comparative assessment of initial and revised translations using a five-point holistic scoring rubric administered

Error Type	Ferris (2011)	ATLaS %(n)
Morphological Errors		
Verbs		
Tense	10.9	5.1 (9)
Form	7.8	0.6 (1)
Subject-verb agreement	2.9	1.7 (3)
Nouns		
Articles/determiners	6.6	6.3 (11)
Noun endings	8.9	0.6 (1)
Lexical Errors		
Word choice	11.5	44.0 (77)
Word form	6.5	4.6 (8)
Informal usage	0.3	0.0 (0)
Idiom error	0.8	0.6 (1)
Pronoun error	2.9	0.6 (1)
Syntactic Errors		
Sentence structure	22.5	30.9 (54)
Run-ons	2.9	0.0 (0)
Fragments	1.8	1.1 (2)
Mechanical Errors		
Punctuation	6.8	0.6 (1)
Spelling	5.9	2.3 (4)
Miscellaneous	0.9	1.1 (2)

Table 3: Error distribution comparison.

through OpenAI's GPT-4 showed significant advancement from a mean initial score of 2.3 to a mean revised score of 3.3, representing a 43% improvement in the overall writing quality. A paired-samples t-test was conducted to verify the significance of this improvement. The results confirmed that the increase in scores was statistically significant, $t(25) = 6.43, p < .001$.

The analysis identified distinctive learning patterns based on initial proficiency levels. Advanced learners (initial scores of 4–5) demonstrated sophisticated refinements in stylistic choices and idiomatic expressions, whereas intermediate learners (scores of 2–3) showed substantial improvements in grammatical accuracy and sentence structure. Beginning learners (scores of 1–2) exhibited fundamental corrections in basic grammatical construction and vocabulary selection, although the improvement margins were modest.

4.2 Error pattern analysis

An analysis of 175 correction instances revealed distinct patterns that diverged from the established error distribution findings. The error distribution showed notable deviations from Ferris (2011) findings, reflecting the intersection of Japanese learners' characteristics, translation-based learning, and AI-mediated error detection (Table 3).

Most significantly, lexical errors dominated the ATLaS corrections (49.8%), substantially exceeding Ferris (2011) correction (22.0%). This was likely driven by word-choice difficulty (44.0% vs. 11.5%). For instance, the AI flagged subtle collocational errors typical for Japanese learners, such as correcting "I joined the exam" to "I took the exam." This suggests that AI systems are highly sensitive to semantic nuances that human instructors may overlook, based on leveraging vast linguistic databases to detect contextually inappropriate vocabulary.

Syntactic errors were the second-largest category (32.0%), approximating Ferris (2011) findings (27.2%). However, the AI captured different phenomena. For example, it corrected nuanced prepositional choices, such as changing "finish the work until tomorrow" to "by tomorrow," which affects grammatical precision more than immediate comprehensibility. This highlights AI's systematic identification of structural deviations, whereas instructors may prioritize communicative effectiveness.

Conversely, morphological errors showed a markedly lower frequency (14.3%) than in Ferris (2011) study (37.2%). This reduction may reflect the pattern-recognition capabilities of AI in identifying morphological consistency, in which source-text cues facilitate accurate grammatical choices. Mechanical errors also showed a substantial reduction, which was attributable to the digital writing environment and sophisticated AI checking of spelling.

These findings suggest that AI-mediated detection produces different error distributions compared to traditional human analysis, emphasizing the need for pedagogically informed AI training that balances systematic accuracy with communicative priorities in L2 writing instruction.

5 Conclusions

5.1 Summary

This study provides preliminary evidence that the ATLaS system may enhance Japanese university students' English writing performance through structured AI-mediated feedback. The scaffolded learning approach, grounded in the constructivist learning theory and Vygotsky's ZPD, appeared to promote metalinguistic awareness and increased learner autonomy among the 26 participants during the two-month intervention. Quantitatively, mean holistic scores increased from 2.3 to 3.3

(43% improvement), and a paired-samples t-test indicated this change was statistically significant, $t(25) = 6.43, p < .001$. However, because the study used a single-group pre–post design and instructor feedback was provided alongside ATLaS (see Methods 3.3), improvements cannot be unambiguously attributed to ATLaS alone. These results should therefore be interpreted as promising but preliminary, and future randomized controlled trials are required to isolate the specific effects of the system.

A systematic analysis of 175 correction instances revealed distinct error patterns. Lexical errors dominated the corrections (49.8%), substantially exceeding Ferris (2011) reported frequency of 22.0%, primarily driven by word-choice difficulties (44.0% vs. 11.5%). Syntactic errors constituted the second largest category (32.0%), whereas morphological errors showed a markedly lower frequency (14.3%) than in Ferris (2011) study (37.2%).

These findings suggest that AI-assisted feedback systems can contribute to L2 writing development when designed with appropriate pedagogical scaffolding. Nevertheless, given the single-group design and potential instructor–system interaction, further controlled research is needed to confirm causal mechanisms.

5.2 Further research directions

This study shows promising results for AI-assisted writing instruction, but several areas need further investigation. First, future research should include larger and more diverse participant groups to improve the generalizability of these findings. Randomized controlled trials would help isolate the specific effects of AI feedback and test the system with different writing tasks beyond personal narratives.

Second, longitudinal studies across multiple semesters are needed to determine whether the observed improvements persist over time. Such studies would reveal whether learners maintain self-correction skills after the intervention ends and how AI feedback affects long-term writing development.

Third, research should examine how cultural and institutional factors influence AI-assisted feedback effectiveness. Studies in different educational contexts would help us understand how various pedagogical approaches and technologies affect system adoption and success.

Finally, future versions of ATLaS should expand beyond Japanese-to-English translation to support multiple language pairs. This multilingual approach would make the system applicable to broader language learning contexts and enable comparative analyses across different first languages. Such expansion could lead to more inclusive instructional design and enhance the system's global relevance.

Acknowledgments

This research was supported by the Grant-in-Aid for Young Scientists (Grant Number: 23K12254).

References

Xiaolong Cheng and Lawrence Jun Zhang. 2024. Examining second language (L2) learners' engagement with AWE–teacher integrated feedback in a technology-empowered context. *The Asia-Pacific Education Researcher*, 33:1023–1035.

Ning Fan. 2023. Exploring the effects of automated written corrective feedback on EFL students' writing quality: A mixed-methods study. *SAGE Open*, 13(2).

Haiying Feng, Kexin Li, and Lawrence Jun Zhang. 2025. What does AI bring to second language writing? a systematic review (2014–2024). *Language Learning & Technology*, 29(1):1–27.

Dana R. Ferris. 2011. *Treatment of Error in Second Language Student Writing*. University of Michigan Press, Ann Arbor, MI.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge, Abingdon, UK.

Yasin Karatay and Leyla Karatay. 2024. Automated writing evaluation use in second language classrooms: A research synthesis. *System*, 123:103332.

Atsushi Mizumoto. 2025. Automated analysis of common errors in L2 learner production: Prototype web application development. *Studies in Second Language Acquisition*, pages 1–18.

David James Woo, Hengky Susanto, Chi Ho Yeung, Kai Guo, and April Ka Yeng Fung. 2024. Exploring AI-generated text in student writing: How does AI help? *Language Learning & Technology*, 28(2):183–209.