Journal of Historical, Cultural and Social Sciences, 2025, 1(2), 10-22
https://doi.org/10.71204/7mdsqs66

Post-editing of AI-assisted Translation in Translator Education: A Genre-based Classroom Study

Qiushi Gu [1,*], Shiyan Wang [2], Xinyao Ren [1], Xinyu Ji [1]

[1] School of Japanese Studies, Beijing International Studies University, Beijing 100024, China

[2] Japanese Department, University of International Relations, Beijing, 100091 China

* **Correspondence**:

Qiushi Gu

guqiushi@bisu.edu.cn

Received: 9 December 2025/ Accepted: 21 December 2025/ Published online: 21 December 2025

Abstract

The emergence of generative AI in translation classrooms has shifted instructional priorities away from text production alone toward issues of quality and judgement. This study reports a within-class longitudinal investigation of post-editing of AI-assisted translation in Japanese–Chinese translation training. Over a single semester, twenty-eight undergraduate Japanese majors completed six translation assignments, each involving an AI-generated draft, manual revision, and a short-written justification of their revision choices. Two learning outcomes were examined: overall translation quality (score_total, 0–100) and the quality of students' rationales (rationale_quality, 0–2), the latter capturing the extent to which revision decisions were supported by explicit, text-based evidence. Results indicated gradual and non-linear improvement in translation quality, alongside more noticeable changes in students' justification practices over time. Genre also played a mediating role: news tasks tended to elicit cue-based rationales related to modality and attribution, whereas literary tasks prompted broader but less readily verifiable stylistic reasoning. These findings suggest that the pedagogical value of post-editing of AI-assisted translation lies not only in improving translation products, but also in fostering evaluative judgement through routine, scaffolded post-editing and justification activities.

Keywords: AI-assisted Translation; Post-editing; Evaluative Judgement; Translation Pedagogy

1. Introduction

The widespread use of generative artificial intelligence and machine translation has reshaped everyday practices in translation teaching. In many translation classrooms, the key issue is no longer whether students should be allowed to use AI-based tools, but how such tools can be incorporated without weakening the development of professional judgement. Previous studies have shown that machine translation systems substantially reduce the effort required to produce an initial draft, enabling learners to generate linguistically fluent target texts with relative ease (Bowker, 2021; Kenny, 2022). This convenience, however, has also prompted concerns among educators, particularly regarding students' ability to evaluate translation quality once the initial draft is no longer produced by the students themselves.

In professional contexts, translation competence is defined less by speed or surface fluency than by the capacity to evaluate competing solutions and to take responsibility for translational choices. As AI systems increasingly assume the role of text generation, evaluative skills become more central rather than less. Although post-editing has long been recognised as a pedagogically meaningful way to integrate machine translation into translator training (O'Brien, 2002), classroom-based research has paid limited attention to how students' decision-making processes develop over time when AI tools are used on a routine basis. Moreover, the potential role of genre in shaping students' engagement with AI-generated drafts has received little systematic attention. The present study therefore reports a longitudinal classroom investigation of Japanese–Chinese translation tasks, examining changes in both translation quality and students' ability to articulate evidence-based rationales for revision, with particular attention to differences between news and literary texts.

2. Literature Review

Research on machine translation in translator education has undergone a notable reorientation over the past two decades. Rather than focusing primarily on whether machine translation should be restricted in the classroom, recent work has increasingly addressed how it can be integrated in pedagogically meaningful ways. Central to this discussion is the concept of post-editing. O'Brien (2002) described post-editing as a process that goes beyond surface correction and involves active judgement and decision-making on the part of the translator. Later studies in translator education have drawn on this view, arguing that machine translation supports learning only when students are encouraged to question and revise AI-generated output rather than rely on it uncritically (Bowker, 2021; Kenny, 2022; O'Brien & Ehrensberger-Dow, 2020).

This line of research foregrounds judgement as a core learning outcome, linking post-editing practice with broader discussions of evaluative judgement and feedback literacy in higher education. Evaluative judgement refers to learners' capacity to make informed assessments of quality and to explain the basis of those assessments in complex tasks (Tai et al., 2018). Related work on feedback literacy further suggests that feedback supports learning only when students are able to internalise criteria and apply them in subsequent decisions (Carless & Boud, 2018). In translation education, where multiple acceptable solutions often coexist, these perspectives underscore the importance of making students' reasoning processes visible, rather than evaluating final products alone.

Genre has received comparatively little attention in studies of AI-assisted translation pedagogy. Existing research indicates that source-text characteristics influence machine translation output quality and may shape subsequent revision behaviour (Lee, 2022; Loock & Léchauguette, 2021).

At the same time, work on critical AI literacy in language education suggests that students are more likely to engage critically with AI tools when their use is embedded in task-specific risks and constraints (Tacelosky et al., 2025; Krüger, 2023). In translation classrooms, contrasting genres such as news and literary texts provide a natural context for this form of situated engagement, as they foreground different dimensions of responsibility, stance, and stylistic coherence. The present study draws on these insights by examining genre as a mediating factor in students' engagement with AI-assisted post-editing.

3. Methodology

3.1. Research Design and Research Questions

To capture changes that unfold through repeated classroom practice, this study followed a single cohort of students across multiple AI-assisted translation tasks over the course of one academic semester. The design did not compare different groups or experimental conditions, but focused on within-class development. This decision presupposes that skills such as post-editing judgement and the ability to justify revision decisions are not developed in reaction to a single task, but are formed through repeated interactions with translation standards and feedback (Tai et al., 2018; Carless & Boud, 2018).

In translation education, the role of machine translation has shifted noticeably in recent years. Rather than being treated solely as a shortcut to be avoided, MT is increasingly incorporated into instruction through post-editing activities. Importantly, post-editing is not understood as routine correction, but as a process that requires evaluative judgement, decision-making, and an awareness of communicative responsibility (O'Brien, 2002; Bowker, 2021). Research on machine translation literacy similarly suggests that learning outcomes depend less on whether students use AI tools and more on how they are guided to question and revise AI-generated output in context (O'Brien & Ehrensberger-Dow, 2020; Kenny, 2022). These considerations make a longitudinal classroom design particularly suitable for examining students' engagement with AI-assisted translation.

Against this background, the study addressed the following research questions:

(1) Does students' overall translation quality change over the course of AI-assisted translation instruction?

(2) Does students' ability to provide evidence-based rationales for post-editing decisions change over time?

(3) Do news and literary genres shape these developmental patterns in different ways?

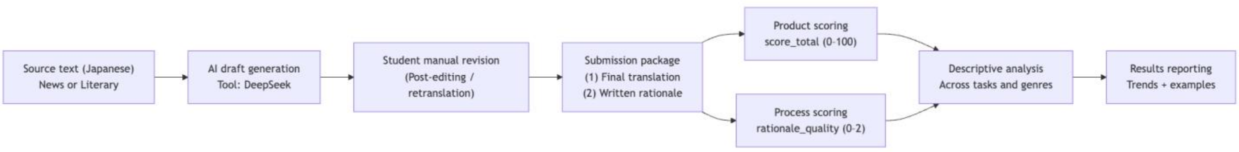

Figure 1. Workflow of the AI-assisted Translation and Data Collection Process

3.2. Participants and Instructional Context

The participants were 28 undergraduate students majoring in Japanese at a public university in China. All were enrolled in a compulsory Japanese–Chinese translation course taught over a 16-week semester. The course met twice a week, with each session lasting 90 minutes. Before taking the course, students had completed general Japanese language training and introductory translation exercises, but none had received systematic instruction in AI-assisted translation or post-editing.

AI use was explicitly built into the course design. For all assignments, students relied on the same AI translation system (DeepSeek) to generate initial drafts. This was a deliberate instructional decision. Allowing multiple tools would have introduced unnecessary variation and made it difficult to interpret students' revision behaviour. Similar concerns have been raised in previous work on machine translation literacy, which emphasises the importance of controlling tool-related variables in classroom-based studies (Kenny, 2022).

Throughout the semester, students were repeatedly reminded that AI-generated drafts were not to be treated as final translations. Classroom discussion addressed typical limitations of MT output, such as overgeneralisation, inappropriate certainty, and stylistic flattening, issues that have been noted in earlier studies of MT use in language and translation classrooms (Lee, 2023). The instructional goal was not to discourage AI use, but to help students develop a more reflective and responsible stance toward AI-assisted translation.

Table 1. Overview of Indicators Included in Land-Use Scale, Development Intensity

Component	**Specification**
Research design	Within-class longitudinal
Setting	Undergraduate Japanese–Chinese translation course (public university, China)
Participants	28 undergraduate Japanese majors
Course duration	16 weeks
Contact hours	2 × 90 minutes per week
Language pair	Japanese to Chinese
AI tool	DeepSeek (used consistently across all tasks)
Assignments	6 translation tasks (T1–T6)
Genres	3 news texts; 3 literary texts
Unit of analysis	One submission per student per task (max. 168 submissions)
Outcome measures	score_total (0–100); rationale_quality (0–2)
Analysis approach	Descriptive trends and qualitative illustration

3.3. Tasks and Instructional Procedure

Over the semester, students completed six translation assignments, consisting of three news texts and three literary excerpts. The use of two contrasting genres was intentional. Previous research has shown that machine translation output and post-editing strategies are sensitive to text type, and that genre characteristics influence the kinds of problems learners attend to during revision (Lee, 2022; Lee, 2023). In addition, work on critical AI literacy suggests that students engage more critically with AI tools when tasks highlight concrete communicative risks rather than abstract technical skills (Ducar & Schocket, 2018).

The assignments were scheduled from Week 3 onward, with one task assigned every two weeks. The basic procedure of each task is similar. Students first generated an AI draft using DeepSeek. After generating the initial draft, students were asked to revise the translation on their own. In doing so, they decided which parts required modification and which could be retained. In addition to the revised version, students provided a short written explanation commenting on some of their revision decisions as well as why the revisions were made.

The task procedure was identical, but the focus of classroom discussion changed depending on the text type. With news translation tasks, evidentiality, modality, and responsibility attribution were considered, and tasks in this direction examined whether the AI-generated expressions demonstrated excessive confidence or ambiguous information sources. In the translation of literary texts, discussion focused on narrative voice, tone, and consistency. Rather than correcting isolated errors, students were encouraged to consider how local revisions affected the coherence of the text as a whole. This use of post-editing to foreground different aspects of translation quality is consistent with earlier classroom-based research (Niño, 2008; Yang, 2023).

Table 2. Task Schedule and Genre Coverage

Task	Week	Genre	Source text length	Instructional focus	Deliverables
T1	Week 3	News	~700 JP characters	evidentiality; stance; attribution	final translation + rationale
T2	Week 5	Literary	~1,000 JP characters	narrative voice; tone	final translation + rationale
T3	Week 7	News	~750 JP characters	modality; hedging	final translation + rationale
T4	Week 9	Literary	~1,100 JP characters	rhythm; stylistic coherence	final translation + rationale
T5	Week 11	News	~800 JP characters	factual caution; responsibility	final translation + rationale
T6	Week 13	Literary	~1,200 JP characters	perspective; voice stability	final translation + rationale

3.4. Measures and Rating Procedures

To capture both learning outcomes and learning processes, two measures were used in the analysis: overall translation quality (score_total) and students' rationale quality (rationale_quality).

3.4.1. Translation quality (score_total)

Translation quality was measured using the total score awarded for each assignment according to the course assessment rubric (score_total). The rubric included dimensions commonly used in translation pedagogy, such as accuracy, coherence, appropriateness of expression, and genre conformity. Scores ranged from 0 to 100 and were assigned by the course instructor as part of regular coursework assessment. Using course-based scores as outcome measures is consistent with classroom-oriented research on MT post-editing, where ecological validity is often prioritised over experimental control (Niño, 2008; Yang, 2023).

3.4.2. Rationale quality (0-2)

Beyond product scores, the analysis also considered how students explained their revision decisions. This aspect was captured through a process-oriented measure referred to as rationale quality, which assessed whether students could articulate clear and text-based reasons for the changes they made. The use of this measure was informed by Tai et al. (2018)'s work on evaluative judgement, which views quality assessment as a form of disciplinary competence, and by research on feedback literacy that stresses the importance of internalising criteria through use rather than through instruction alone (Carless & Boud, 2018).

Rationale quality was rated on a three-point scale:

(1) 0 (Impression-based): Statements based on intuition or preference, without identifiable evidence.
(2) 1 (Partially grounded): References to textual or contextual factors that remain vague or weakly connected to the revision.
(3) 2 (Evidence-based): Clear justification linked to identifiable source-text cues, pragmatic considerations, or genre conventions.

Table 3. Measures and Rating Criteria

Measure	Construct	Scale	High-score definition	Example (anonymised)
score_total	Overall translation quality (product)	0–100	Strong performance across rubric dimensions	"92/100"
rationale_quality	Evidence-based justification (process)	0–2	Explicit, verifiable reasoning	"Because 'とみられる' signals hedging, I used '据称/可能' to avoid overstating certainty."

***Note.** *score_total* refers to the overall translation score assigned according to the course assessment rubric (range: 0–100). *rationale_quality* was coded on a three-point scale (0–2), reflecting the extent to which revision decisions were supported by explicit and verifiable justification.

Table 4. Coding Examples for Rationale_quality

Level	Typical wording	Coding rationale	Example (anonymised)
0	“It sounds more natural.”	No explicit evidence	“I revised the sentence because it reads better.”
1	“The source seems cautious, so I adjusted the tone.”	Partial awareness, unclear linkage	“The Japanese is uncertain, so I softened the Chinese.”
2	“The source uses ‘とみられる’, so I retained hedging to preserve stance.”	Explicit cue and justified revision	“Cue: とみられる, Revision: 据称/可能, Reason: avoid categorical claims in news.”

***Note.** Examples are anonymised excerpts selected to illustrate typical patterns observed in student rationales. Coding focuses on the presence and clarity of evidence-based justification rather than linguistic accuracy.

3.5. Data Analysis

Data analysis focused on developmental trends across tasks and genres. For each assignment, means and standard deviations of score_total and rationale_quality were calculated. The data were analyzed from two perspectives: first, results were traced across six tasks (T1–T6) over time to examine change; second, patterns between news and literary tasks were compared to determine whether text type was associated with different developmental tendencies.

To supplement these quantitative results, a sample of anonymized rationale extracts was used to illustrate general over-time changes in students’ justification practices. It is typical in classroom-based research in translation pedagogy to combine descriptive statistics with qualitative description in order to examine learners’ reasoning, rather than to claim causal effects under strictly controlled conditions (Bowker, 2021; Yang, 2023).

4. Results

This section reports patterns observed across the six AI-assisted translation tasks. Rather than treating the results as evidence of linear improvement, the focus is placed on how students’ performance and justification practices changed, stabilised, or fluctuated over time. The results are organised around overall translation quality, rationale quality, and genre-related differences.

4.1. Overall Patterns in Translation Scores across Tasks

Students’ overall translation performance was examined using the course-based total score (score_total, 0–100). Mean scores and standard deviations were calculated for each task to provide an overview of changes across the semester.

Table 5 indicates that average scores increased across the six tasks, although the pattern was not uniform. However, the change was gradual and uneven. In the early tasks, particularly T1 and T2, score differences between students were relatively large. Some students achieved relatively high scores from the outset, while others remained clustered around the mid-range.

In later tasks, mean scores were higher on average, but this increase did not occur uniformly from one task to the next. In several instances, mean scores plateaued or changed only marginally, especially when the genre of the task shifted. Each practice material belongs to a different genre (in this case, news or literature), and the topics, difficult words, and styles, etc. in different materials also vary across materials.

Standard deviations provide additional insight. While variation across students remained visible throughout the semester, dispersion tended to narrow slightly in later tasks. This indicates that, for some students, performance became more stable over time. At the same time, a small number of students continued to display irregular performance patterns, suggesting that the AI-assisted workflow did not affect all learners in the same way.

Table 5 Mean Score_total and Standard Deviation across Six Translation Tasks

Task	Mean score_total (M)	SD
T1	71	9.5
T2	72	9.0
T3	75	8.5
T4	78	8.0
T5	79	7.5
T6	80	7.5

***Note**. Mean values represent average *score_total* across students for each task. SD refers to standard deviation, indicating variation in performance within the class.

4.2. Changes in Students' Justification Practices

To examine how students explained their revision decisions, written rationales were analysed using the three-level rationale_quality scale (0–2). The distribution of rationale levels was calculated for each task.

In the initial tasks, a large proportion of rationales were coded at Level 0. These explanations typically relied on general impressions, such as claims that a revision made the translation "sound better", without reference to specific textual features. Level 1 rationales, which showed partial awareness of textual or contextual factors, were present but less common.

When Table 6 is examined across tasks, changes in students' rationales become more apparent in the later stages of the experiment. The percentage of Level 0 rationales decreased, whereas Level 1 and Level 2 rationales rose. In the later tasks, Level 2 rationales—those providing explicit and verifiable justification—ceased to be exceptional. Learners increasingly resorted to recognizable clues in the source text or to genre-related conventions in their textual justifications for revision (Koponen, 2016). This pattern did not apply to all students. Some students, in a number of assignments, continued to rely on impression-based explanations even after being reminded several times to justify their choices.

Another observation is that the quality of the rationale did not necessarily change in a parallel manner to translation scores. Cases were identified in which, even when overall translation quality remained relatively high, students produced more coherent explanations of why particular decisions were made. This suggests that justification practices may develop before improvements in final products become visible.

Table 6. Distribution of Rationale Quality Levels across Tasks (%)

Task	Level 0 (%)	Level 1 (%)	Level 2 (%)
T1	45	40	15
T2	38	42	20
T3	30	45	25
T4	22	46	32
T5	18	44	38
T6	12	42	46

***Note**. Percentages may not sum to 100 due to rounding. Rationale levels correspond to the coding scheme described in Table 4.

4.3. Genre-related Differences in Performance and Reasoning

To explore the role of genre, results were examined separately for news and literary translation tasks. Mean score_total and mean rationale_quality were calculated for each genre.

Table 7. Comparison of Translation Performance and Rationale Quality by Genre

Genre	Mean score_total	Mean rationale_quality
News	78.5	1.45
Literary	76.5	1.30

In the case of news translation tasks, mean translation scores showed comparatively stable variation across assignments. Students' rationales in such exercises more often alluded to concrete linguistic cues, such as modality or attribution patterns. In some instances, students explicitly associated revision with distrust of overstatement or responsibility, suggesting an awareness of communicative risk.

Literary tasks were more varied. On the one hand, some students produced high-quality translations and elaborate rationales; on the other hand, some students were unable to sustain a high level of performance. Rationales in literary tasks tended to focus on tone, narrative voice, or stylistic flow, but were at times less precise in identifying what prompted a revision. By comparison, news tasks appeared to attract more cue-based justifications, whereas literary tasks encouraged more general yet less readily correctable forms of explanation.

5. Conclusion

This study aims to examine how students' engagement with AI-assisted translation tasks developed over time, focusing not only on changes in translation quality but also on how students justified their revision decisions. The results do not point to dramatic improvement in final products. Instead, they reveal a more uneven and gradual pattern of change, particularly in the ways students reasoned about translation choices. In this section, the discussion addresses two closely related issues: how learning becomes visible beyond product scores, and how genre shapes students' evaluative engagement with AI-generated drafts.

5.1. Looking Beyond Products: What Changes Before Quality Improves

A notable feature of the results is the contrast between relatively minor changes in translation scores and more noticeable improvement in students' written rationales. Although there were some gradual gains in overall translation quality throughout the semester, these were not uniform and did not always occur in task order. By comparison, changes in how students explained their revision preferences were more apparent and tended to emerge earlier.

Students' performance did not show immediate improvement, which suggests the presence of a learning curve in AI-assisted post-editing. Rather than demonstrating linear gains in translation quality, students appeared to develop increasing awareness of what constitutes a justified decision. Analysis of the submissions indicated that some students provided more explicit arguments for their revisions despite inconsistencies in translation quality. A similar pattern is described by Tai et al. (2018), who argue that evaluative judgement develops through disciplinary engagement rather than through short-term performance gains.

This observation raises the need to reconsider how learning outcomes are defined and evaluated in translation classrooms. Product-based scores, although necessary, tend to mask early forms of development that are cognitive or reflective in nature. Comparable concerns have been raised in the feedback literacy literature, which emphasizes that students may internalize quality standards before applying them consistently in performance (Carless & Boud, 2018). From this perspective, improvement in the quality of rationales represents a valuable, albeit incomplete, indicator of learning prior to visible gains in translation output.

Nevertheless, the findings caution against assuming that progress is automatic. Development in justification remained uneven, indicating that the use of AI tools requires instructional designs that consistently foreground reasoning, evidence, and responsibility.

5.2. Genre, Constraint, and the Nature of Evaluative Engagement

The findings show that students interacted with AI-generated drafts in relation to genre. Distinctions between news and literary translation tasks were evident not only in translation scores, but more clearly in the nature of students' rationales.

During news translation exercises, students more often based their justifications on recognizable linguistic cues, including markers of modality or attribution. These tasks encouraged forms of assessment that were relatively tangible and measurable. News discourse is more likely to be characterized by explicit signs of stance and responsibility, offering clearer points of reference for evaluative judgement. Previous studies have noted that machine translation output is less reliable in informational genres, making deviations and risks easier for students to detect during post-editing (Lee, 2022).

By contrast, literary translation assignments exhibited a different pattern. Students frequently referred to stylistic concerns such as tone, narrative voice, or overall coherence. These aspects are central to literary translation but are not easily tied to specific textual triggers. As a result, students' rationales were not always as explicit as those observed in news-related tasks. This inconsistency highlights limitations of AI-generated drafts in handling interpretive complexity and places a greater interpretive burden on students.

Because news and literary texts differ in their characteristics, the present study does not suggest that one genre is more effective than the other. Rather, the findings indicate that translation pedagogy may benefit from the use of multiple genres. Alternating between news and literary tasks appears to expose students to different forms of evaluative challenge: one grounded in constraint and risk management, the other in interpretive judgement and stylistic sensitivity. From this perspective, genre variation functions not simply as content diversification, but as a means of shaping how students learn to engage critically with AI tools.

5.3. Implications and Limitations

Previously, translation pedagogy primarily aimed at enhancing the quality of translations. The results of this paper indicate that the pedagogical benefit of AI-aided translation does not lie in short-term quality improvement but in the development of evaluative judgement. Where learners are required to justify the necessity of a revision, AI-generated drafts serve as prompts for reflection rather than directing them toward a satisfactory conclusion. This view supports previous arguments on the cognitive load of post-editing and its nature as an activity that requires assessment and decision-making rather than mechanical correction (O'Brien, 2002; Bowker, 2021).

The current research has a number of drawbacks. It is based on one class and does not aim to make causal claims. There is also a risk of limited extrapolation beyond similar instructional contexts due to the use of course-based assessment measures. Future studies may investigate the effects of different types of scaffolding on justification practices or integrate rationale analysis with interview data to gain deeper insight into students' decision-making processes.

Irrespective of these constraints, the study offers a grounded description of students' interaction with AI-generated drafts over time. The results do not allow AI to be considered either a threat or a solution; its pedagogical influence is determined by how it is integrated into tasks that foreground judgement, evidence, and responsibility.

Funding

This work was supported by a university-level general project on educational and teaching reform "Intelligent Translation Technologies and Communication" (Grant No.1111129137) of Beijing International Studies University.

Author Contributions:

Qiushi Gu contributed to the conceptualization, methodology of the study, and supervised the overall project, and coordinated the research process of the study. Shiyan Wang provided guidance on theoretical framing and critical revisions of the manuscript. Xinyao Ren and Xinyu Ji performed data analysis and conducted the first draft of the manuscript. All authors have read and agreed to the published version of the manuscript.

Informed Consent Statement:

Not applicable.

Data Availability Statement:

Not applicable.

Conflict of Interest:

The authors declare no conflict of interest.

References

Bowker, L. (2023). Machine translation. In De-mystifying translation (1st ed.). Routledge.

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. Assessment & Evaluation in Higher Education, 43(8), 1315–1325.

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

Kenny, D. (Ed.). (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence (Translation and Multilingual Natural Language Processing 18). Language Science Press.

Koponen, M. (2016). Is machine translation post-editing worth the effort? A survey of research into post-editing and effort. The Journal of Specialised Translation, 25, 131–148.

Krüger, R. (2023). Some reflections on the interface between professional machine translation literacy and data literacy. Journal of Data Mining and Digital Humanities. https://doi.org/10.46298/jdmdh.9045

Lee, S. M. (2022). An investigation of machine translation output quality and the influencing factors of source texts. ReCALL. 34(1), 81-94.

Lee, S. M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. Computer Assisted Language Learning, 36(1–2), 103–125.

Loock, R., & Léchauguette, S. (2021). Machine translation literacy and undergraduate students in applied languages: report on an exploratory study. Revista Tradumàtica. Tecnologies de la Traducció, 19, 204–225.

Niño, A. (2008). Evaluating the use of machine translation post-editing in the foreign language class. Computer Assisted Language Learning, 21(1), 29–49.

O'Brien, S. (2002). Teaching post-editing: A proposal for course content. In Proceedings of the 6th EAMT Workshop: Teaching Machine Translation (pp. 99–106). European Association for Machine Translation.

O'Brien, S., & Ehrensberger-Dow, M. (2020). MT literacy – A cognitive view. Translation, Cognition & Behavior, 3(2), 145–164.

Tacelosky, K., Kasun, G. S., Shapiro, B. R., Liao, Y.-C., & Harris, K.-L. (2025). Exploring critical AI literacy in language education: A case study. Foreign Language Annals. Advance online publication. https://doi.org/10.1111/flan.70029

Tai, J., Ajjawi, R., Boud, D., Dawson, P., & Panadero, E. (2018). Developing evaluative judgement: Enabling students to make decisions about the quality of work. Higher Education, 76, 467–481.

Yang, Y. (2023). Performance and perception: Machine translation post-editing in Chinese-English news translation by novice translators. Humanities and Social Sciences Communications, 10, 285.