IDEAS
Journal of Language Teaching and Learning, Linguistics and Literature
ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)
Volume 14, Number 2, December 2026
pp. 9434 - 9456

Copyright © 2026 The Author
IDEAS is licensed under CC-BY-SA 4.0 License



Issued by English study program of UIN Palopo

The Measurable Effects of Machine Translation on L2 Writing Performance: A Systematic Literature Review

Mimo Lalito Erisyerico[1], Ni Made Ratminingsih[2], Made Hery Santosa[3], I Putu Indra Kusuma[4]
[1,2,3] Doctoral Program in English Language Education, Postgraduate Program, Universitas Pendidikan Ganesha, Singaraja

Article Info

Received: 2026-04-07
Revised: 2026-05-19
Accepted: 2026-11-07

Keywords:
machine translation; L2 writing; writing performance; learner proficiency; instructional guidance

DOI:
10.24256/ideas. v14i2. 12617

Corresponding Author:
Mimo Lalito Erisyerico
mimo.lalito@gmail.com
Doctoral Program in English Language Education, Postgraduate Program, Universitas Pendidikan Ganesha, Singaraja

Abstract

This systematic review investigates the measurable effects of machine translation (MT) on second language (L2) writing performance. It asks how MT is used in EFL/ESL writing instruction and revision, what effects have been reported, and how learner proficiency and instructional guidance shape these effects. Scopus was searched for peer-reviewed studies published in 2022–2026, and a procedure adapted from PRISMA was followed. Fourteen studies met the criteria. Studies were coded, appraised for quality, and counted by outcome. MT served as free access, a revision tool, feedback, a base for post-editing instruction, and a comparison with other tools. Nine of the 10 studies that measured accuracy reported better accuracy. Four of the six studies that measured fluency reported gains. Only four of nine studies reported gains in syntactic complexity, and each gain had limits. Lexical results were mixed to positive (five of nine studies reported clear gains), but only one study tested vocabulary directly. In the studies that tested this, gains made with MT fell when MT was removed, but stayed above pretest levels in the clearest case. Studies that added explicit instruction, such as post-editing training, reported gains mainly in accuracy. Guidance seems to shape the effect of MT, but no study clearly isolated guided from unguided MT use. Only three studies compared proficiency levels, and effects were not simple. Lower-proficiency learners gained most in syntactic complexity in one study, while higher-proficiency learners kept vocabulary gains longer in another.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

Study quality was high in two, moderate in five, and low in seven. The evidence shows short-term writing performance more than long-term language development. Therefore, MT should be taught with guidance and assessed with and without MT. Future studies need longitudinal designs with delayed tests without MT, guided versus unguided comparisons, and comparisons across proficiency levels and contexts.

1. Introduction

Machine translation (MT) has a long history. Early rule-based systems appeared in the late 1940s (Hutchins, 2010). For decades, MT stayed in government, military, and business settings. Few language classrooms used it (Jolley & Maimone, 2022). This changed with neural machine translation (NMT). Google moved to a neural system in 2016, and its output was more fluent and accurate than that of earlier systems (Wu et al., 2016). Mobile devices then made MT easy to reach for many learners. Today, learners use MT when they draft, revise, and edit texts in a second language (L2). MT may also shape learners' language. Resende (2023) found that L2 English learners who saw Google NMT output began to use some of its syntactic structures. This raises a practical question for teachers: does MT change how well learners write? A growing set of experimental and classroom studies tries to answer it.

This review needs a clear boundary. The term MT is used for tools that translate text from one language to another, such as Google Translate and DeepL. Generative artificial intelligence (GenAI), such as ChatGPT, is a different tool. It can write, explain, and revise text. In this review, GenAI is context, not focus. GenAI studies were not searched. When an included study also tested ChatGPT, only its MT condition was used. GenAI research is growing. Song and Song (2023) reported better academic writing among EFL learners who used GenAI, while Erisyerico et al. (2025) reported positive student perceptions of AI-generated feedback. A review of 69 studies from 2021 to 2026 linked GenAI tools to short-term gains in writing accuracy, with weaker evidence for long-term transfer (Klimova & Carvalho, 2026). Loock and Holt (2024) found that MT and GenAI together helped learners analyze English compounds. Alharbi (2023) found that university EFL students used AI-enabled MT often and with little supervision.

Teachers disagree about MT. Some treat unauthorized MT use as academic dishonesty that should be detected and prevented (Correa, 2011; McCarthy, 2004; Steding, 2009). Others see MT as a resource that should be taught and used (Correa, 2014; Ducar & Schocket, 2018; Niño, 2020; Vold, 2018). GenAI makes detection harder (Alharbi, 2023). Therefore, the key question is not whether to allow MT. It is how to guide its use so that learners use it responsibly. For this reason, instructional guidance is central to this review. MT also has limits. Its output often contains errors, especially in verb forms and idioms (Yoon & Chon, 2022).

Polysemy and syntactic ambiguity can cause mistranslation even in neural systems (Qin, 2025). MT can repeat gender bias (Farkas & Németh, 2022; Savoldi et al., 2021). Fan et al. (2025) also warned about "metacognitive laziness" with GenAI: essay scores rose, but knowledge transfer did not. These points suggest that gains depend on how learners use MT.

To evaluate these outcomes clearly, the main terms are defined as follows. Accuracy is how free a text is from errors, usually measured as the number or rate of errors. Fluency is the amount of text produced, usually word count or units per text. In untimed written tasks this is mostly a measure of text length, so it is treated as a limited proxy for fluency. Syntactic complexity is the length and structure of clauses and sentences. Lexical complexity is the range, diversity, and sophistication of vocabulary. Vocabulary development is the gain in words learners can use, including retention. Overall writing quality is a holistic score given by raters. The writing process is what writers do when they plan, draft, and revise. An effect is called measurable when a study reports a change in rater scores or in text-based indices for these outcomes. Three things that studies often mix are also separated. The first is performance with MT: the quality of a text written or revised with MT. The second is performance without MT: the quality of a later text written alone. Only the second can show that learning took place. The third is learners' perceptions of MT. Perceptions are not effects, and they are not counted as outcomes.

Earlier reviews give useful maps, but a gap remains. Alones (2025) reviewed 83 studies published from 2020 to 2024. The research covered only a few EFL contexts, such as Saudi Arabia, China, and Thailand. Jolley and Maimone (2022) reviewed three decades of work. They found that few studies link MT activities to second language acquisition (SLA) theory. Lee (2023) reviewed 87 studies published from 2000 to 2019 and pooled 12 of them in a meta-analysis. She reported a small overall effect of MT on foreign language learning ($g = 0.345$), with significant effects in all writing subcategories except lexical complexity. She also noted that, apart from two studies, no study had tested effects beyond the task at hand, and that almost none included teacher intervention or training. Her review ends in 2019. Jiang et al. (2024) took a different angle. Their scoping review of MT use in L2 writing grouped MT into three models: a linguistic processor, a mediational artifact, and a translanguaging process. They found that most studies focused on linguistic factors, and that the effect of MT depends on individual, contextual, and ideological factors. As a scoping review, it mapped models and factors. It was not built to count how many studies support each outcome or to appraise study quality. Much of this literature deals with perceptions or with detecting MT use.

To the authors' knowledge, few reviews bring together the experimental and quasi-experimental studies that measure writing outcomes after 2019. In the reviews cited here, four steps done together were not found: counting how many studies support each outcome, appraising study quality, separating MT effects from

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

instructional effects, and separating performance with MT from performance without MT. This review fills that gap. It synthesizes recent experimental, quasi-experimental, longitudinal, and classroom studies, including small case studies, on the measurable effects of MT on the writing of EFL and ESL learners. It asks three research questions:

1. What is the role of MT in EFL/ESL writing instruction and revision practices?
2. What measurable effects of MT on writing performance have been reported (accuracy, fluency, syntactic and lexical complexity, vocabulary, and overall quality), and how many studies support each effect?
3. How do learner proficiency and instructional guidance shape the effects of MT on writing performance, with and without MT?

The review adds three things. For research, it shows how strong the evidence is for each outcome and where design weaknesses limit causal claims. For writing pedagogy, it shows which MT uses have support and which need teacher guidance. For assessment, it separates what learners can do with technology from what they can do alone.

2. Method

A systematic review design was used. The procedure was adapted from the PRISMA 2020 guideline (Page et al., 2021). It had four steps: identification, screening, eligibility assessment, and inclusion.

Eligibility Criteria

Table 1 lists the criteria. A study was included if it met six conditions. It was a peer-reviewed article in English. It was published between 2022 and 2026. It came from a journal indexed in Scopus. It took place in an EFL or ESL setting. It used MT in teaching or learning activities. It reported measurable writing outcomes or observable writing or revision behaviors. Studies that only reported perceptions were excluded. Study design was not restricted, so the sample includes experiments, quasi-experiments, single-group studies, and case studies. If a study compared MT with GenAI tools such as ChatGPT, it was included and only the MT condition was used. “Scopus-indexed” means indexed in Scopus when the search was run. Publication year means the year of the journal issue or of first online publication.

Table 1. Eligibility Criteria

Criteria for inclusion	Criteria for exclusion
1. Peer-reviewed article in English. 2. Published 2022–2026 (issue or first online year).	1. Non-peer-reviewed articles in languages other than English. 2. Published outside 2022–2026. 3. Not indexed in Scopus.

3. Indexed in Scopus when the search was run. 4. Conducted in EFL/ESL contexts. 5. MT used in teaching or learning activities (only the MT condition if other tools were also tested). 6. Reports measurable writing-related outcomes or observable writing/revision behaviors.	4. Non-EFL/ESL contexts. 5. Perception-only studies. 6. Does not report measurable writing-related outcomes or observable writing/revision behaviors, or does not involve the pedagogical use of MT.

Search and Selection Procedure

One database, Scopus, was searched. It was chosen for three reasons. It covers many applied linguistics and education journals. It gives clear indexing information. It lets other researchers repeat the search. The search was limited to 2022–2026 because Lee (2023) covers studies up to 2019, Jolley and Maimone (2022) review three decades of work, and the aim was to capture the period of wide classroom use of NMT. This choice leaves out studies published in 2020 and 2021, so the review should be read as an update on Lee (2023) and not as a full history.

The search used the following keyword combinations: “machine translation” AND “L2 writing”; “machine translation” AND “writing performance”; “Google Translate” AND “EFL writing”; “DeepL” AND “writing”; “machine translation” AND “post-editing”; and “machine translation” AND “ESL writing”. The fields searched were title, abstract, and keywords. The search was run from 3 to 14 August 2026. It returned 132 records. After 14 duplicates were removed, 118 records were screened by title and abstract. Of these, 71 were excluded and 47 full texts were assessed. The first author independently screened all records at both stages, initially by title and abstract and subsequently by full-text assessment, applying the eligibility criteria presented in Table 1. No second reviewer conducted independent screening.

A total of 33 full-text articles were excluded for the following main reasons: 18 were perception-only studies (Criterion 5); 8 did not report measurable writing outcomes or observable writing or revision behaviors, or did not involve the pedagogical use of MT (Criterion 6); and 7 were not conducted in EFL or ESL contexts (Criterion 4). No articles were excluded because they were non-peer-reviewed articles or published in languages other than English (Criterion 1). Similarly, no articles were excluded because they fell outside the publication period of 2022–2026 (Criterion 2) or were not indexed in Scopus at the time of the search (Criterion 3). Each article was assigned to one primary exclusion category to avoid double counting, resulting in a total of 33 excluded articles. Fourteen studies were included (Figure 1). Using one database may miss studies indexed elsewhere and may favor published, positive results. This is discussed again in the limitations.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

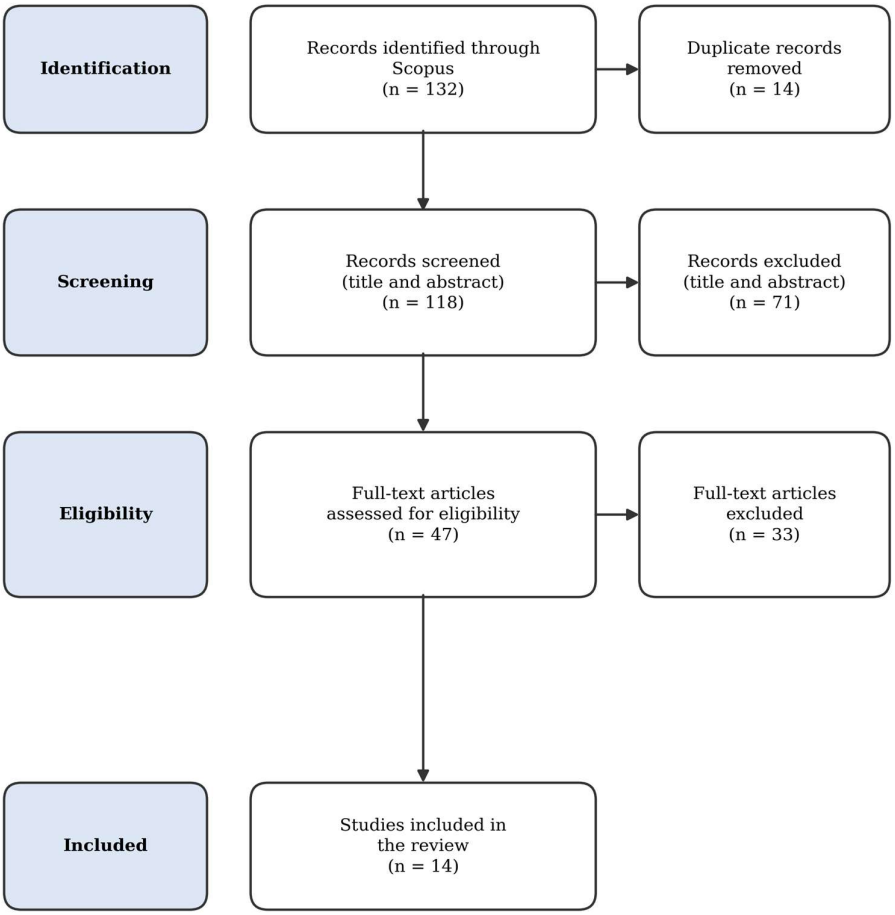


Figure 1. PRISMA Flow Diagram of the Study Identification and Selection Process

Data Extraction

Each study was coded for participants, education level, country, first language (L1), MT tool, writing task, length of intervention, comparison condition, outcome measures, main findings, and limits. Whether the results were tested statistically or reported only as counts or means was also recorded. The full text of all studies was read.

Quality Appraisal

Each study was appraised with a four-item checklist adapted for this review. The items were: (1) a comparison or control condition; (2) a total sample of 30 or more; (3) inferential statistics; and (4) a measure written without MT, such as a posttest or delayed test. All included studies reported outcomes from raters or text-based indices, so this was a condition for inclusion and not a scored item. A study was rated High if it met all four items, Moderate if it met three, and Low if it met

two or fewer. A rating was lowered by one level when the teacher corrected the MT output before learners used it, or when the reported analysis had clear problems. No study was excluded by quality. Instead, the ratings decided how strongly each conclusion was stated.

Synthesis

A narrative synthesis was used, following the SWiM guideline (Campbell et al., 2020), with thematic coding of the findings (Braun & Clarke, 2006). All studies were read, their findings were coded, the codes were grouped into themes, the themes were checked against the studies, and the themes were named. The themes were: MT and writing quality; MT-assisted revision, editing, and post-editing; proficiency and instructional guidance; and vocabulary development. They are reported under the three research questions. For each outcome, the number of studies that reported gains, no gains, or mixed results was also counted. No meta-analysis was run because designs and outcome measures differed too much. The counts are unweighted: each study counts once, whatever its sample size or design. For this reason, results from studies with a comparison condition are also noted. The first author coded all studies. The other authors checked all the codes. Disagreements were resolved by discussion until all authors agreed, and by checking the source articles again.

3. Result

Study Characteristics

The 14 studies were published between 2022 and 2026. Thirteen appeared in 2023 or later. Twelve came from East and Southeast Asia: Korea (6), China (3, one of them in Macao), Taiwan (1), Japan (1), and Thailand (1). One came from the Czech Republic and one from India. None took place in Indonesian higher education. All 14 studies were set in EFL contexts, and none in an ESL context. Nine studies involved university students and five involved secondary school students (Almashy et al., 2024; Kim & Chon, 2025; Lee & Kang, 2024; Lee & Oh, 2025; Namfah, 2024).

Where reported, sample sizes ranged from 6 to 213, and six studies had fewer than 40 participants. Interventions ranged from one session to a full semester. Five lasted 10 weeks or more (Almashy et al., 2024; Chang et al., 2024; Kang, 2026; Poláková & Klimová, 2023; Wang & Ke, 2022). Seven studies had a comparison condition (Almashy et al., 2024; Chang et al., 2024; Kim & Chon, 2025; Lee & Oh, 2025; Namfah, 2024; Sasaki et al., 2025; Wang & Ke, 2022). The other seven used a single group or a pre–post design. Table 2 summarizes each study.

Table 2. Overview of the Included Studies

No.	Study	Context and sample	Design and MT use	Outcomes and main findings

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

1	Poláková & Klimová (2023)	Czech university; 16 students, C1	Pre–post, 10 weeks. Weekly summaries: L1 text, English text without MT, then with DeepL. Pre- and posttest written without MT.	Errors fell from 149 to 60. No control group and no inferential statistics.
2	Kim & Chon (2025)	Korean high school; 79 students, age 16, B1	Controlled experiment: self-proofreading, MT-assisted proofreading, ChatGPT-assisted proofreading. 237 essays.	MT and ChatGPT groups scored higher overall than the self-proofreading group and made fewer grammar errors. Preposition errors did not change. Gains in lexical diversity and sentence complexity were reported for the ChatGPT group, not for the MT group.
3	Lo (2025)	Macao university; 63 law undergraduates, higher (30) and lower (33) proficiency	Quasi-experimental, 3 weeks. L1 text, own translation, then comparison with Google Translate and revision of 10 legal terms. Posttests without MT, immediate and after 2 weeks.	Vocabulary scores were high in both groups (9.57 and 9.27 of 10). Delayed retention: 9.43 (higher) vs. 5.42 (lower). No comparison without MT, so the scores cannot be credited to MT alone.
4	Namfah (2024)	Thai secondary school; 29 students, grade 9, low proficiency	Ten tasks, with MT allowed on some and not on others. Teacher notes and interviews.	Mean score 4.70 with MT and 1.16 without, but many no-MT papers were blank. Students used MT to plan and draft, and skipped reviewing.

5	Chang et al. (2024)	Taiwan university; 34 trained students and 33 controls	Quasi-experimental, 18-week semester. Recursive editing with Google Translate for the trained group.	Trained group did better on error correction, accuracy, and mean length of clause. Writing scores did not differ. Control group had higher fluency.
6	Kang (2026)	Korean university; 130 learners at three levels	Longitudinal, one general English course. Essays written without MT in the first and last class.	Accuracy and fluency improved at all levels. Lexical gains were larger for advanced learners, syntactic gains for lower-level learners.
7	Yang et al. (2023)	China; 31 first-year English majors, lower-intermediate	Replication study. One session: manual translation, then revision with Google Translate. No control group.	Higher ratings after MT revision. Word use and syntactic complexity improved; lexical diversity did not.
8	Lee & Kang (2024)	Korean middle school; 59 students, A2 or below	Single group: pretest, MT-assisted text, posttest, retention test at 2 or 4 weeks. Teacher corrected the MT-assisted text.	Best scores in the MT-assisted text. Later scores fell but stayed above pretest, except punctuation. Grammar gains fell more than vocabulary gains.
9	Sasaki et al. (2025)	Japan university; 23 students, B1	Crossover, 4 weeks. DeepL feedback vs. teacher feedback. Rewrite without feedback.	Accuracy and fluency improved after MT feedback; complexity improved after teacher feedback. No significant difference between conditions.
10	Kim & Chang (2024)	Korean EFL learners; 6 students	Ten tasks. Manual translation, then MT-aided revision.	Accuracy and fluency improved. Complexity gains were less clear. Individual paths

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

				varied.
11	Almashy et al. (2024)	India secondary school; 213 students, grade 12, age 16–17	Experiment, 12 weeks, four sessions a week. Classes used ChatGPT, Grammarly, or Google Translate (write in Hindi, translate, compare, revise). Control class had traditional teaching.	All classes, including the control, made fewer errors; the class-by-test interaction was not significant. Google Translate class improved most on articles, capitalization, and possessives (raw means). Errors counted with Grammarly.
12	Kim & Han (2023)	Korean university; 12 students	Case study, 4 days, four drafts. MT output, then post-editing workshops.	Students copied most corrections from MT in Draft 3. In Draft 4, 80% of paraphrases were their own. Counts only.
13	Wang & Ke (2022)	China; 82 English majors, two intact classes, lower-intermediate	Quasi-experimental, 16 weeks. MT class followed a write–translate–compare–revise procedure.	MT class scored higher, with better accuracy and lexical complexity. No group difference in fluency or syntactic complexity.
14	Lee & Oh (2025)	42 Korean middle school learners	Two groups. 90-minute Google Translate lesson vs. none. Picture-based narrative, 3 days.	Instruction group improved spelling and grammar accuracy. No significant effect on syntactic complexity.

The Role of MT in Writing Instruction

The studies used MT in five ways. First, one study gave learners MT access without training (Namfah, 2024). Second, six studies used MT-assisted revision. In most of them, learners wrote or translated a text, compared it with MT output, and revised it (Kang, 2026; Kim & Chang, 2024; Lee & Kang, 2024; Lo, 2025; Wang & Ke, 2022; Yang et al., 2023). Third, one study used MT output as feedback and

compared it with teacher feedback (Sasaki et al., 2025). Fourth, four studies added explicit instruction or post-editing training (Chang et al., 2024; Kim & Han, 2023; Lee & Oh, 2025; Poláková & Klimová, 2023). Fifth, two studies compared MT with other tools, such as ChatGPT or Grammarly (Almashy et al., 2024; Kim & Chon, 2025).

Three studies looked at the writing process. Revisions were mostly small. In Wang and Ke (2022), about half of all revisions (7,192 of 14,589) were at word level. Learners did not always review MT output. In Namfah (2024), students used MT mostly to plan and draft, and they skipped reviewing. In Kim and Han (2023), students copied most of the corrections from MT in the third draft (61 of 67 changes). After post-editing workshops, 80% of their paraphrases in the fourth draft were their own.

Study Quality

Table 3 shows the quality ratings. Two studies were rated High (Lee & Oh, 2025; Wang & Ke, 2022), five were rated Moderate, and seven were rated Low. Four weaknesses were common. Samples were small. Half of the studies had no comparison condition. Only seven studies had a measure written without MT. In several studies, MT was combined with a set procedure, training, or teacher input, so the MT effect could not be separated from the effect of instruction. In Almashy et al. (2024), the control class improved as much as the tool classes, errors were counted with Grammarly, and the degrees of freedom did not match the reported sample of 213.

Table 3. Quality Appraisal of the Included Studies

No.	Study	Comparison	Sample	Statistical test	Measure without MT	Rating
1	Poláková & Klimová	None	16	No (counts)	Yes	Low
2	Kim & Chon	Self-proofreading	79	Yes	No	Moderate
3	Lo	Proficiency groups only	63	Yes	Yes (immediate and 2 weeks)	Moderate
4	Namfah	Same students, MT vs. no MT	29	No (means)	No	Low
5	Chang et al.	Control group	67	Yes	No	Moderate

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

6	Kang	None reported	130	Yes	Yes	Moderate
7	Yang et al.	None	31	Yes	No	Low
8	Lee & Kang	None; teacher corrected text	59	Yes	Yes (≤ 4 weeks)	Low
9	Sasaki et al.	Teacher feedback (crossover)	23	Yes	Yes (same prompt)	Moderate
10	Kim & Chang	None	6	Yes	No	Low
11	Almashy et al.	Control class (also ChatGPT, Grammarly)	213	Yes (MANOVA)	No	Low
12	Kim & Han	None	12	No (counts)	No	Low
13	Wang & Ke	Control class	82	Yes	Yes	High
14	Lee & Oh	Comparison group	42	Yes	Yes	High

Measurable Effects of MT

Table 4 counts the studies for each outcome. Eleven studies reported statistical tests or significant differences. Three reported only counts or means in their results (Kim & Han, 2023; Namfah, 2024; Poláková & Klimová, 2023). Studies measured the same outcome in different ways. Accuracy was measured as error counts (Poláková & Klimová, 2023), errors per 100 words (Sasaki et al., 2025), or a rate formula (Wang & Ke, 2022). Fluency was word count in most studies, but sentence, clause, and T-unit counts in Chang et al. (2024). Those counts overlap with complexity measures, so the fluency result in that study needs care. For this reason, studies were counted and effects were not pooled. In short, accuracy and fluency improved more often than syntactic complexity.

Table 4. Number of Studies Supporting Each Outcome

Outcome	Studies that measured it	Studies with gains	Studies with no or mixed gains	Note
Accuracy or	10	9	1 (Almashy: no	8 tested; 1

errors			difference from control); 2 of the 9 had error types that did not improve	counts only
Fluency	6	4	2 (control equal or higher)	Chang et al.; Wang & Ke
Syntactic complexity	9	4 (each with limits)	5 (no or weak gain)	Gains limited to one index, one level, or the MT-assisted text
Lexical outcomes (vocabulary or lexical complexity)	9	5	4	Vocabulary was the main outcome in one study (Lo); Yang et al. counted as mixed
Overall quality (rater scores)	6	4	2	Chang et al. no difference; Namfah (2024) not counted because of blank papers
Measure without MT	7	6 kept in full or in part	1 limited to higher proficiency	Longest delay: 4 weeks

Accuracy. Ten studies measured accuracy or errors. Nine reported better accuracy or fewer errors (Chang et al., 2024; Kang, 2026; Kim & Chang, 2024; Kim & Chon, 2025; Lee & Kang, 2024; Lee & Oh, 2025; Poláková & Klimová, 2023; Sasaki et al., 2025; Wang & Ke, 2022). One of the nine reported counts only (Poláková & Klimová, 2023). Some error types did not improve. Prepositions did not change in Kim and Chon (2025), and punctuation did not improve in Lee and Kang (2024). In Almashy et al. (2024), all four classes, including the control class, made fewer errors, and the class-by-test interaction was not significant. The Google Translate class improved most on articles, capitalization, and possessives, but only in raw means. This study is therefore not counted as showing a gain for MT. Among the studies with a comparison condition, five reported fewer errors (Chang et al., 2024; Kim & Chon, 2025; Lee & Oh, 2025; Sasaki et al., 2025; Wang & Ke, 2022) and one did not (Almashy et al., 2024).

Fluency. Six studies measured fluency. Four reported gains (Kang, 2026; Kim & Chang, 2024; Lee & Kang, 2024; Sasaki et al., 2025). In two studies, MT did not

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

help. In Chang et al. (2024), the control group had higher counts of sentences, clauses, and T-units. In Wang and Ke (2022), the groups did not differ, and the control class gained more words.

Syntactic complexity. Nine studies measured it. Four reported gains, each with limits. Chang et al. (2024) found a gain in mean length of clause only. Yang et al. (2023) found a gain in one index. Lee and Kang (2024) found a gain in the MT-assisted text, which did not last. Kang (2026) found a gain for the lowest proficiency group. Five studies found no gain or a weak one (Kim & Chang, 2024; Kim & Chon, 2025; Lee & Oh, 2025; Sasaki et al., 2025; Wang & Ke, 2022). In Kim and Chon (2025), the sentence-complexity gain was reported for the ChatGPT group rather than the MT group.

Vocabulary and lexical complexity. Nine studies had a vocabulary or lexical outcome. Five reported clear gains (Kang, 2026; Lee & Kang, 2024; Lo, 2025; Poláková & Klimová, 2023; Wang & Ke, 2022). Yang et al. (2023) reported mixed results: word use improved but lexical diversity did not. Kim and Chang (2024) found less clear lexical gains, and word choice errors remained in Kim and Han (2023). In Kim and Chon (2025), greater lexical diversity was reported for the ChatGPT group, not for the MT group. Only Lo (2025) used vocabulary as the main outcome, and it tested ten legal terms. The other eight studies measured word use, lexical complexity, or word-choice errors, not vocabulary development.

Overall quality. Six studies gave rater scores. Four reported higher scores with MT (Kim & Chon, 2025; Lee & Kang, 2024; Wang & Ke, 2022; Yang et al., 2023). In Chang et al. (2024), writing scores did not differ between groups. Namfah (2024) is not counted as a gain. The mean score was 4.70 with MT and 1.16 without MT, but many papers without MT were blank. The gap reflects task completion more than language quality.

Gains With and Without MT

Seven studies had a measure written without MT (Kang, 2026; Lee & Kang, 2024; Lee & Oh, 2025; Lo, 2025; Poláková & Klimová, 2023; Sasaki et al., 2025; Wang & Ke, 2022). Gains were kept in full or in part in six of them. Lee and Kang (2024) give the clearest picture. The mean score was 1.61 at pretest and 4.83 in the MT-assisted text. It fell to 3.66 at the posttest and 2.51 at the retention test, two or four weeks later. All later scores stayed above the pretest, except for punctuation. Vocabulary errors held, but grammar errors rose. In Lo (2025), delayed retention after two weeks was 9.43 of 10 for the higher-proficiency group and 5.42 for the lower-proficiency group. In Sasaki et al. (2025), the rewrite without feedback was done on the same prompt, so it shows only a short-term effect. The longest delay found was four weeks. In Wang and Ke (2022) and Lee and Oh (2025), the MT or instruction group did better on accuracy in texts written without MT. Wang and Ke (2022) also reported better performance on lexical complexity, but not on syntactic complexity. In Lee and Kang (2024), the tests used the same topic, so transfer to

new topics was not tested.

Learner Proficiency

Seven studies involved one proficiency level. They were C1 (Poláková & Klimová, 2023), B1 (Kim & Chon, 2025; Sasaki et al., 2025), lower-intermediate (Wang & Ke, 2022; Yang et al., 2023), low (Namfah, 2024), and A2 or below (Lee & Kang, 2024). Four studies did not report levels clearly (Almashy et al., 2024; Chang et al., 2024; Kim & Chang, 2024; Lee & Oh, 2025). Only three compared levels. In Lo (2025), higher-proficiency learners kept vocabulary gains longer. In Kang (2026), accuracy and fluency improved at all three levels. Advanced learners gained more in lexical complexity, and lower-level learners gained more in syntactic complexity. Learners at the high-intermediate level gained least. In Kim and Han (2023), beginners found it hard to spot errors in MT output. The link between proficiency and benefit was therefore not simple.

Instructional Guidance

Four studies added explicit instruction or training. In Chang et al. (2024), the trained group did better on error correction, accuracy, and mean length of clause. However, writing scores did not differ, and the control group had higher fluency. In Lee and Oh (2025), a 90-minute Google Translate lesson improved spelling and grammar accuracy but not syntactic complexity. In Kim and Han (2023), learners became more independent after post-editing workshops. In Poláková and Klimová (2023), guided DeepL use went with fewer errors, but the study had no control group. In Sasaki et al. (2025), MT feedback and teacher feedback did not differ significantly. The two worked on different outcomes: accuracy and fluency improved after MT feedback, and complexity improved after teacher feedback. Chang et al. (2024) and Lee and Oh (2025) compared trained and untrained groups, but the reports do not make clear whether the untrained groups also used MT. As far as the reports show, no included study therefore isolated guided from unguided MT use.

4. Discussion

The Main Pattern

The clearest finding is about accuracy. Nine of the 10 studies that measured accuracy reported better accuracy or fewer errors. Fluency gains were less common (4 of 6 studies). Gains in syntactic complexity were rare and conditional (4 of 9 studies, each with limits). Lexical results were mixed to positive (5 of 9 studies), but only one study had vocabulary as a main outcome. This pattern fits the way MT works. MT output gives learners a model at word and phrase level. Word-level differences are easy to see when learners compare their text with the MT text. In Wang and Ke (2022), about half of the revisions were at word level. Changes in sentence structure need more knowledge. Lee and Kang (2024) suggest that grammar items are harder to notice and use without explanation. One study

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

with a control class found that all classes, including the control class, made fewer errors (Almashy et al., 2024). Some accuracy gains may therefore come from instruction and practice, and not from MT. This reading of the pattern is an interpretation. The reviewed studies did not test it directly. Fluency may rise because learners can write first in their L1 and then translate. Yet the control group had equal or higher fluency in two studies (Chang et al., 2024; Wang & Ke, 2022). So MT does not guarantee fluency gains. This pattern partly agrees with Lee (2023), whose meta-analysis of 12 studies found significant effects on accuracy, orthography, and writing quality, but not on lexical complexity. Her meta-analysis also found a significant effect for syntactic complexity (three studies), which the newer studies in the sample do not confirm.

MT Effects and Instructional Effects

The gains cannot be credited to MT alone. In most studies, MT came with a set procedure, training, or teacher input. Three conditions can be separated. The first is access to MT. Namfah (2024) shows that access helps learners finish tasks, but not necessarily write better. The second is a pedagogical procedure, such as write–translate–compare–revise (Lee & Kang, 2024; Wang & Ke, 2022; Yang et al., 2023). The third is explicit post-editing instruction (Chang et al., 2024; Kim & Han, 2023; Lee & Oh, 2025; Poláková & Klimová, 2023). Teacher input can also confound results. In Lee and Kang (2024), the teacher corrected the MT-assisted text before learners studied it. In Sasaki et al. (2025), teacher feedback and MT feedback gave comparable gains, though on different outcomes. Therefore, claims that guided MT works better than unguided MT should be treated with care. No included study clearly isolated this contrast. Earlier evidence is clearer. Lee (2023) described a study that compared a non-MT group, an MT group without training, and an MT group with prior training, and the trained MT group did best. No study in the sample repeated this three-group design. What the studies support is narrower: MT plus a structured procedure or training is linked to gains, mainly in accuracy.

Learner Agency

A tentative reading is that gains depend on what learners do with MT output. This rests on three small and mostly descriptive studies, so it is a hypothesis and not a finding. When learners copied it, the benefit was weak. Students in Namfah (2024) skipped reviewing. Students in Kim and Han (2023) copied most corrections in Draft 3. Gains were stronger when learners had to evaluate, change, or question the output. After post-editing workshops, learners in Kim and Han (2023) produced more of their own paraphrases. In Wang and Ke (2022), learners revised at many levels. This fits the view that MT supports learning when it prompts noticing and decisions, not when it replaces them.

Theoretical Perspectives

The reviewed studies rarely name a theory, but their procedures fit two well-known SLA accounts. The first is the Noticing Hypothesis (Schmidt, 1990), which holds that learners acquire the features of input that they consciously notice. In write–translate–compare–revise tasks (Lee & Kang, 2024; Wang & Ke, 2022; Yang et al., 2023), MT output sits next to the learner's own text, so differences at word and phrase level are easy to notice. This may help explain why accuracy gains were the most frequent. It also fits the weak benefit when learners copied output without reviewing it (Kim & Han, 2023; Namfah, 2024), because copying does not require noticing. The second account is the output hypothesis (Swain, 1995), in which producing language helps learners notice gaps in what they can say. In most MT-assisted revision studies, learners wrote a draft before they saw the MT text, so MT worked as a reformulation of their own output. Interaction accounts (Long, 1996) fit less well, because MT gives a reformulation but no negotiation of meaning. Teacher guidance may supply some of this missing interaction, which is consistent with the gains linked to training.

Jiang et al. (2024) add that MT can be treated as a linguistic processor, a mediational artifact, or a translanguaging process. Most of the 14 studies measured text outcomes, which suits the processor view, while the learner-agency pattern above fits the mediational view better. No included study tested these accounts directly, so this reading is an interpretation. Future studies could test it by measuring what learners notice, for example with think-aloud protocols or stimulated recall.

Proficiency

The studies do not show a simple link between proficiency and benefit. Higher-proficiency learners can judge MT output better. In Kim and Han (2023), beginners found it hard to spot errors. In Lo (2025), higher-proficiency learners kept vocabulary gains longer. Lower-proficiency learners may rely on MT more. Namfah (2024) found many blank papers without MT, and Lee and Kang (2024) studied beginners. For them, MT can allow tasks they could not do alone. That is a gain in performance, and it does not show a gain in ability. Kang (2026) found that the lowest level gained most in syntactic complexity and the high-intermediate level gained least. Proficiency effects depend on the outcome and on the context. Lee (2023) made a similar point for studies up to 2019: most included only one proficiency level, so claims about who benefits more could not be compared.

Performance and Learning

Many studies measured performance with MT. Fewer measured performance without MT. The two can differ. In Lee and Kang (2024), scores were highest with MT and lower at the posttest and retention test. Vocabulary held, and grammar did not. Some evidence points to learning: Kang (2026) used essays written without

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

MT at the end of a course, and Poláková and Klimová (2023) used a posttest without MT. Lee and Oh (2025) and Wang and Ke (2022) also report accuracy gains on measures written without MT. Still, the delays were short, and half of the designs had no control group. Lee (2023) reported the same for earlier work: apart from two studies, none tested longer-term effects, and in one of them the gain in lexical diversity vanished when MT was no longer used. Fan et al. (2025) warned about "metacognitive laziness" with GenAI, where scores rise without knowledge transfer. The evidence for MT is not strong enough to confirm or reject this concern.

A Simple Framework

Figure 2 summarizes this reading as a working framework, not a tested model. MT availability can raise performance with MT by itself, as Namfah (2024) shows for task completion. The studies that reported more than this combined MT with guided interaction and learner evaluation. Independent development is the last and least supported component. Evidence for the first four components comes from Chang et al. (2024), Kim and Han (2023), Lee and Kang (2024), Lee and Oh (2025), Namfah (2024), and Wang and Ke (2022). Evidence for the last component comes mainly from Kang (2026), Lee and Kang (2024), and Lo (2025), and it is partial.

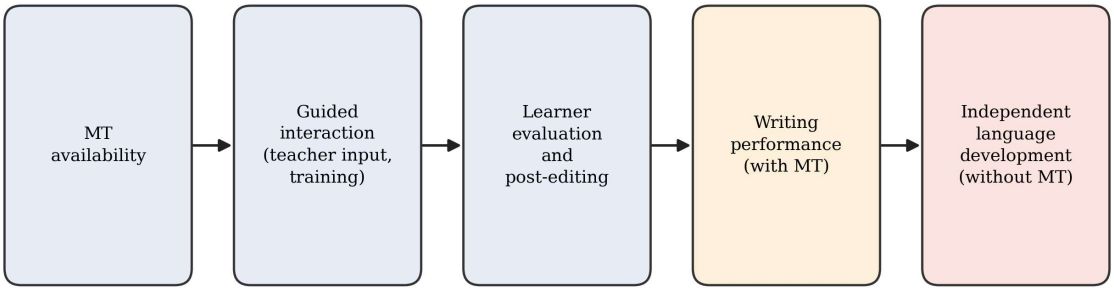


Figure 2. A Simple Framework for the Effects of MT on L2 Writing

Implications

For teaching, six practices are suggested. Evidence supports the first five. Teach learners to evaluate MT output before they accept it (Kim & Han, 2023). Make post-editing a required step (Chang et al., 2024; Lee & Oh, 2025). Ask learners to compare their own draft with the MT-assisted draft (Wang & Ke, 2022; Yang et al., 2023). Add teacher feedback, which gave comparable gains in Sasaki et al. (2025). Follow MT-assisted tasks with a task without MT (Lee & Kang, 2024). Ask learners to explain why they accept or change each MT suggestion. This is not tested in the reviewed studies, but it fits the evaluation step in Figure 2. For assessment, teachers should decide what they want to measure: writing with technology, or writing alone. Task instructions should say whether MT is allowed.

A task without MT is needed to check independent ability.

Limitations and Future Research

This review has limits. It covers only 14 studies. Twelve came from East and Southeast Asia, so the findings may not fit other contexts. One database was searched, which may miss studies and may favor published, positive results. The first author alone screened all records and full texts, so no second reviewer checked the screening decisions. Some eligible studies may have been missed or wrongly excluded. Interventions were mostly short. Outcome measures differed, so studies were counted and effect sizes were not estimated. The counts give every study equal weight. Because most designs lack a control condition or a measure without MT, MT is described as linked to gains, not as their proven cause. The quality ratings rest on a short checklist and on the authors' judgment. Quality was low in seven studies, so the counts should be read with care.

Future studies should do five things. They should use delayed tests without MT, including new topics. They should compare guided and unguided MT directly. They should study the writing process, for example with screen recordings. They should test whether learners understand grammar after MT use. They should include more contexts, such as Indonesian higher education, and more learner levels.

5. Conclusion

This review asked three questions. First, MT was used in five ways, namely as free access, as a revision tool, as feedback, as the base for post-editing instruction, and as a tool compared with others such as ChatGPT or Grammarly. Studies with explicit instruction or structured procedures gave the clearest results. Second, the evidence is strongest for accuracy (9 of the 10 studies that measured it). It is weaker for fluency (4 of 6) and syntactic complexity (4 of 9, with limits). Clear lexical gains appeared in 5 of 9 studies, but only one study tested vocabulary directly. Third, the effect of proficiency was not simple, and guided MT could not be compared with unguided MT because no study clearly isolated the two. In the studies that tested this, gains with MT fell when MT was removed, but they stayed above pretest levels in the clearest retention test. Most of the evidence is short-term and of low to moderate quality. Therefore, MT should be taught with guidance and evaluated with tasks both with and without MT. Better studies are needed to show whether MT leads to lasting gains in independent L2 writing.

6. References

Alharbi, W. (2023). The use and abuse of artificial intelligence-enabled machine translation in the EFL classroom: An exploratory study. *Journal of Education and e-Learning Research, 10*(4), 689–701. https://doi.org/10.20448/jeelr.v10i4.5091

Almashy, A., Ahmed, A. S. M. M., Jamshed, M., Ansari, M. S., Banu, S., & Warda, W. U.

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

(2024). Analyzing the impact of CALL tools on English learners' writing skills: A comparative study of errors correction. *World Journal of English Language, 14*(6), 657–667. https://doi.org/10.5430/wjel.v14n6p657

Alones, J. (2025). The role of machine translation in English language learning: A systematic review. *Journal of English Language and Pedagogy (JELPA), 3*(1), 46–67. https://doi.org/10.51826/jelpa.v3i1.1388

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ, 368*, l6890. https://doi.org/10.1136/bmj.l6890

Chang, P., Chen, P. J., & Lai, L. L. (2024). 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

Correa, M. (2011). Academic dishonesty in the second language classroom: Instructors' perspectives. *Modern Journal of Language Teaching Methods, 1*(1), 65–79.

Correa, M. (2014). Leaving the "peer" out of peer-editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content & Language Integrated Learning, 7*(1), 1–20. https://doi.org/10.5294/3568

Ducar, C., & Schocket, D. H. (2018). Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google Translate. *Foreign Language Annals, 51*(4), 779–795. https://doi.org/10.1111/flan.12366

Erisyerico, M. L., Perdana, I., & Karani, E. (2025). Students' perceptions of Grammarly application feedback on writing skill at SMAN 4 Palangka Raya. *PROJECT (Professional Journal of English Education), 8*(2), 292–305. https://journal.ikipsiliwangi.ac.id/project/article/view/26252

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology, 56*(2), 489–530. https://doi.org/10.1111/bjet.13544

Farkas, A., & Németh, R. (2022). How to measure gender bias in machine translation: Real-world oriented machine translators, multiple reference points. *Social Sciences & Humanities Open, 5*(1), 100239. https://doi.org/10.1016/j.ssaho.2021.100239

Hutchins, J. (2010). Machine translation: A concise history. *Journal of Translation Studies, 13*(1–2), 29–70.

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*, 101099. https://doi.org/10.1016/j.jslw.2024.101099

Jolley, J. R., & Maimone, L. (2022). Thirty years of machine translation in language teaching and learning: A review of the literature. *L2 Journal, 14*(1), 26–44. https://doi.org/10.5070/L214151760

Kang, D.-M. (2026). The effects of machine translation (MT) use on the writing quality and perceptions of L2 writers across proficiency levels: A longitudinal examination. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688261439276

Kim, M., & Chon, Y. V. (2025). The impact of self-revision, machine translation, and ChatGPT on L2 writing: Raters' assessments, linguistic complexity, and error correction. *Assessing Writing, 65*, 1–18. https://doi.org/10.1016/j.asw.2025.100950

Kim, M., & Han, S. (2023). A case study on integrating Google Translate into college EFL writing instruction. *Korean Journal of English Language and Linguistics, 23*, 1111–1135. https://doi.org/10.15738/kjell.23..202312.1111

Kim, T., & Chang, S. M. (2024). Machine translation and EFL writing development: A focus on complexity, accuracy, and fluency. *The Sociolinguistic Journal of Korea, 32*(4), 185–220. https://doi.org/10.14353/sjk.2024.32.4.07

Klimova, B., & Carvalho, A. N. P. (2026). A systematic review of innovative methodologies in foreign language pedagogy in the era of artificial intelligence. *Social Sciences & Humanities Open, 14*, 103170. https://doi.org/10.1016/j.ssaho.2026.103170

Lee, J., & Oh, S.-Y. (2025). Effects of Google Translate instruction on young learners' L2 writing revisions. *The Journal of AsiaTEFL, 22*(1), 41–64. https://doi.org/10.18823/asiatefl.2025.22.1.3.41

Lee, S.-M. (2023). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning, 36*(1–2), 103–125. https://doi.org/10.1080/09588221.2021.1901745

Lee, S.-M., & Kang, N. (2024). Effects of machine translation on L2 writing proficiency: The complexity, accuracy, lexical diversity, and fluency. *Language Learning & Technology, 28*(1), 1–19. https://doi.org/10.64152/10125/73585

Lo, S. (2025). Neural machine translation in EFL classrooms: Learners' vocabulary improvement, immediate vocabulary retention and delayed vocabulary retention. *Computer Assisted Language Learning, 38*(3), 592–611. https://doi.org/10.1080/09588221.2023.2207603

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). Academic Press.

Loock, R., & Holt, B. (2024). Augmented linguistic analysis skills: Machine translation and generative AI as pedagogical aids for analyzing complex

ISSN 2338-4778 (Print)
ISSN 2548-4192 (Online)

English compounds. *Technology in Language Teaching & Learning, 6*(3), 1–27. https://doi.org/10.29140/tltl.v6n3.1489

Masruddin, M., Amir, F., Langaji, A., & Rusdiansyah, R. (2023). Conceptualizing linguistic politeness in light of age. International Journal of Society, Culture & Language, 11(3), 41-55.

Masruddin, Hartina, S., Arifin, M. A., & Langaji, A. (2024). Flipped learning: facilitating student engagement through repeated instruction and direct feedback. Cogent Education, 11(1), 2412500.

McCarthy, B. (2004). Does online machine translation spell the end of take-home translation assignments? *Computer-Assisted Language Learning Electronic Journal, 6*(1), 26–39.

Namfah, N. (2024). "Should I submit a blank assignment?!!": The effect of machine translation on the writing process and performance among Thai EFL students with low English proficiency. *LEARN Journal: Language Education and Acquisition Research Network, 17*(2), 134–162. https://doi.org/10.70730/PRNM2789

Niño, A. (2020). Exploring the use of online machine translation for independent language learning. *Research in Learning Technology, 28*, 1–32. https://doi.org/10.25304/rlt.v28.2402

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery, 88*, 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Poláková, P., & Klimová, B. (2023). Using DeepL translator in learning English as an applied foreign language–An empirical pilot study. *Heliyon, 9*(8), e18595. https://doi.org/10.1016/j.heliyon.2023.e18595

Qin, Y. (2025). Semantic role labeling in neural machine translation addressing polysemy and ambiguity challenges. *Journal of Technology Informatics and Engineering, 4*(1), 41–55. https://doi.org/10.51903/jtie.v4i1.274

Resende, N. (2023). Investigating syntactic priming cumulative effects in MT-human interaction. *Open Research Europe, 1*, 93. https://doi.org/10.12688/openreseurope.13902.2

Sasaki, M., Mizumoto, A., & Matsuda, P. K. (2025). Machine translation as a form of feedback on L2 writing. *International Review of Applied Linguistics in Language Teaching, 63*(4), 2301–2326. https://doi.org/10.1515/iral-2023-0223

Savoldi, B., Gaido, M., Bentivogli, L., Negri, M., & Turchi, M. (2021). Gender bias in machine translation. *Transactions of the Association for Computational Linguistics, 9*, 845–874. https://doi.org/10.1162/tacl_a_00401

Schmidt, R. W. (1990). The role of consciousness in second language learning.

Applied Linguistics, 11(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology, 14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Steding, S. (2009). Machine translation in the German classroom: Detection, reaction, prevention. *Die Unterrichtspraxis/Teaching German, 42*(2), 178–189. https://doi.org/10.1111/j.1756-1221.2009.00052.x

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford University Press.

Vold, E. T. (2018). Using machine-translated texts to generate L3 learners' metalinguistic talk. In A. Haukås, C. Bjørke, & M. Dypedahl (Eds.), *Metacognition in language learning and teaching* (pp. 67–97). Routledge. https://doi.org/10.4324/9781351049146-5

Wang, J., & Ke, X. (2022). Integrating machine translation into EFL writing instruction: Process, product and perception. *Journal of Language Teaching and Research, 13*(1), 125–137. https://doi.org/10.17507/jltr.1301.15

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*. https://doi.org/10.48550/arXiv.1609.08144

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

Yoon, C. W., & Chon, Y. V. (2022). Machine translation errors and L2 learners' correction strategies by error type and English proficiency. *English Teaching, 77*(3), 153–175. https://doi.org/10.15858/engtea.77.3.202209.153