Machine Translation in Language Education: A Systematic Review of Open Access Articles

Louise Ohashi

Introduction

This article reports on a systematic review of research articles related to the use of machine translation（MT）in language education. The author's interest in MT stems from her experiences with it as a language learner, teacher, teacher trainer, and researcher. In all four roles, it is imperative to understand the affordances and limitations of MT, so it is anticipated that this systematic literature review could be of value to a broad cross-section of the foreign language education community. Due to the constant changes that developments in technology bring, it is vital to regularly engage with newly published research, but this can prove challenging. In recent years, the author has conducted MT training for teachers in her local context within Japan, as well as abroad. To give attendees in her sessions some background knowledge on empirical studies that focus on MT in language education, she has shared the findings of Lee's（2021）systematic review of articles on the use of MT in foreign language education, which draws on 87 studies published between 2000 and

2019. This is a very valuable study that draws together key findings from a wide range of contexts, but the fast pace of technological advancement has made it necessary for a more recent review to be conducted. In this article, the author reports on studies that were published in or after 2020, in order to post-date Lee's work. During preparation of this article, Klimova et al. (2023) contributed a 13-article review of language education studies published between 2018 and 2021 that focus on neural MT. Two articles in their study overlapped with those that had been targeted in the present study, so they were removed in order to make a worthwhile new contribution.

The main goal of this study is to provide a concise overview of key findings related to the use of MT in foreign language education, primarily to assist language teachers. Reading empirical studies can be an arduous task and language teachers often face barriers such as time constraints and paywalls that reduce their opportunities to do so. Therefore, it is desirable to have a concise summary of the key findings and implications of multiple articles. Ensuring that the articles reviewed are available through open access is also useful as any readers who want further details can freely read them in full. A secondary goal of the study is to summarise key research areas and methods in recent studies in order to identify trends and stimulate research that could fill current gaps. With these goals in mind, the following two research questions were posed:

> RQ1. What recent research trends can be seen in studies on the use of MT in language education? In particular, what are the key research contexts, focus areas and research methods?

RQ2. What are the key findings of recent research into the use of MT in language education?

The research questions are investigated through a review of 14 articles that were selected based on criteria that are explained below.

Methods

It is important to have clear parameters for article selection and transparent reporting methods when conducting a systematic review. This study draws on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) model (http://www.prisma-statement.org/), which enhances the transparency of reports when followed. This model, which was also adopted by Klimova et al. (2023) in their systematic review of MT research, offers guidance on many aspects of the review process. The key tool in this model is the PRISMA 2020 Item Checklist, which addresses the title, abstract, introduction, methods, results, discussion, and “other information”. Using this checklist can help researchers conduct and report rigorous systematic reviews and meta-analyses. The 11 items in the methods section of the PRISMA 2020 Item Checklist are addressed one by one below.

1. Eligibility criteria: This section outlines what was included and excluded. It is summarised in Table 1.

Table 1.　Inclusion and exclusion criteria

Inclusion Criteria
Accessible as a full-text article on June 12, 2023, in the ERIC database
Search term: “machine translation” AND（“language education” OR “language learning”）
Published from 2020 onwards
Reported primary research related to MT and language education or language learning
Exclusion Criteria
Does not fully meet the inclusion criteria
Was included in Klimova et al.’s（2023）systematic review of MT research

2. Information sources: The search was conducted through the ERIC (Education Resources Information Center) database (https://eric.ed.gov/) on June 12, 2023. ERIC, which describes itself as “a comprehensive, easy-to-use, searchable, Internet-based bibliographic and full-text database of education research and information” used by “five main user groups: academics, researchers, educators, policymakers, and the general public”, was chosen because it is a well-known, user-friendly database that does not require paid subscription.

3. Search strategy: An initial search with the term *“machine translation” AND*（*“language education” OR “language learning”*） yielded 102 results. This was reduced to 43 articles when “full text available in ERIC” was selected and further reduced to 23 articles when the date of publication was restricted to 2020 to 2023 to post-date Lee’s（2021）systematic review of MT-related studies. The 23 articles were downloaded and subjected to initial round screening by the author, who read the abstracts and looked for reports of results to check if the articles met the content requirements（i.e., presented primary research related to MT and language education or language learning). Articles that were outside of these boundaries were excluded, lowering the sample by six articles to 17. At this point, the author read all remaining articles from start to finish, except one that was excluded when closer reading showed it did not meet the inclusion criteria. This reduced the sample to 16 and as none of the included articles were published in 2023, the date parameter became 2020–2022. During preparation of this article, Klimova et al.（2023） published a systematic review that partially post-dates the articles included in Lee's（2021）review. Two articles in Klimova et al.'s study overlapped with those in the present study, so they were removed in order to make a worthwhile new contribution. This reduced the total number of articles in this systematic review to 14.

4. Selection process: The author independently screened each article against the inclusion and exclusion criteria to determine which ones would be part of the dataset.

5. Data collection process: The author read all articles that met the inclusion criteria in full. Initially notes were made in margins and after reading several articles the author began to tabulate key points. When new studies brought forth issues such as lack of data, the author made decisions on uniform ways to report it（for example, using the term "not stated"). After all data was tabulated, information was cross-referenced with the articles repeatedly to ensure accuracy and to fill gaps in data reporting.

6. Data items: The author made notes on each article's research areas (derived from the stated research questions or aims), context, research methods, MT tools, key findings, and additional points of interest unique to individual articles.

7. Study risk of bias assessment: This review was conducted by an individual researcher. The potential for bias in assessment may have been lowered by collaborating on the project or having a subsample assessed by another researcher. The latter was not done due to time constraints, but as all articles are available through open access, readers can refer to them directly to make their own assessments.

8. Effect measures: This study is not a meta-analysis, so this measure is not applicable. Rather than combining statistics of the selected studies, results are synthesised into written form in tables.

9. Synthesis methods: All 14 studies that form part of this systematic literature review were eligible for synthesis into tabular form. When compiling tables, the author aimed to be as concise as possible without dropping pertinent information. Within each section, consistency was sought by identifying key elements. For example, when reporting on the context, the author included information such as the research site, number of participants and languages involved.

10. Reporting bias assessment: Risk of bias is possible in interpretations of what is meant by the parameters “language education” and “language learning”. Some researchers may draw distinctions between studies that focus on developing language skills, those that teach content in the students’ L2, and those that report on MT use in formal education but do not specify the course type. However, all were included in this review as they were seen to meet the inclusion criteria in its broadest definition. The author increases transparency by reporting this here and providing information about the research context of each study.

11. Certainty assessment: Confidence（certainty）in the reliability of the outcomes drawn from the dataset was gained by carefully cross-referencing information presented in the data tables, summary of results, and discussion section with the source articles. This was done multiple times during data analysis and report writing.

Results

This section draws together the key points of the selected studies to address the research questions one by one.

> RQ1. What recent research trends can be seen in studies on the use of MT in language education? In particular, what are the key research contexts, focus areas and research methods?

The 14 studies covered a variety of research areas and used a diverse range of research methods. Table 2 provides a succinct summary of the context, research areas, research methods, and MT tools included in the studies. Key points are highlighted in the following sections, with articles referred to by the article numbers provided in the Table 2（“A” plus the article number; for example, “A1”).

Research Sites: Most of the studies were conducted in Asia（three in South Korea, two in Japan, one each in Taiwan and Indonesia). Single studies were also conducted in Turkey, France, Switzerland, the UK, Yemen, Iran, and New Zealand.

Student/Teacher Participants: Ten of the 14 studies collected data from students only, three reported on data from teachers and students, and one focused on teachers. The number of participants varied widely, with the smallest study involving only four students（A4）and the largest involving 1,926 students and 666 teachers（A7).

Target Languages: Data was mainly collected from students who were studying English (n=8), with three studies focusing on German and two on Chinese. One study addressed multiple languages (A2). Another study mentioned that students and teachers could participate in German, English or French, but it did not explicitly list the languages participants studied or taught（A7). The sole study that focused exclusively on teachers (A5) was conducted predominantly with English teachers, but also included a small number of French, Spanish, Italian, Indonesian and German teachers.

Participants' Foreign Language Proficiency: Of the 13 studies that collected data from students, four did not indicate their proficiency level in the target language. While the other nine student-based studies referred to their level, a lack of consistency in terminology and rationale for proficiency levels makes inter-study comparisons very challenging. The CEFR（Common European Framework of Reference）was most widely mentioned, but only appeared in five studies and the rationale for the chosen CEFR ratings was not always clear. A1 mentioned that first-year students usually enter university with CEFR A1-B2 level and later listed this as the level of students in that study. A4 listed equivalency scores（CEFR B2-C1） that were based on grades in the research site's language program and A14 described students as being at CEFR B1+ level, but did not explain how this was determined. A12 reported students' TOEIC (Test of English for International Communication) scores (average of 570) and noted this was equivalent to CEFR B1. A6 described proficiency based on students' course levels: beginner, elementary, pre-intermediate, intermediate and two repeater levels, with repeaters described as CEFR A2+ and B1 level. CEFR levels for the other four groups were not provided. A11 divided students into low, mid and high proficiency based on TOSEL（Test of Skills in the English Language） scores. In A10, students were enrolled in an intermediate course and a TOEFL（Test of English as a Foreign Language） range was given to indicate their level, but the author's investigation into this test showed that the score range provided in A11（310–677） is actually the lowest and highest score possible in the paper-based TOEFL test, so this figure does not actually represent an intermediate level. A13 used the term "high intermediate to advanced" based on a diagnostic writing test administered by the institution. Similarly, the terms beginner, intermediate, and advanced used in A2 were based on courses students were enrolled in. It is understandable that researchers use the proficiency tests that are most familiar in their context and refer to course levels from the research sites, but this prevents meaningful comparisons from being made between studies. The average teacher or researcher would not know how CEFR, TOEIC, TOEFL, TOSEL and institution-specific levels like beginner and advanced compare to each other, so including information on equivalency, even if only as an indicative range, would be beneficial, as will be elaborated upon in the discussion section.

MT Tools: Google Translate was the most common MT tool in the selected studies, featuring in 9 out of 14 of them. DeepL, Microsoft Translator, and Systran were included in two studies each, with Baidu, Youdao, Yandex Translate, Reverso Traduction, Al-Wafi, and Takarir appearing in one each. Four articles did not report on specific MT tools.

Research Focus: The majority of the studies focused on students' ability to assess MT's accuracy and respond appropriately. For example, they explored students' ability to notice and correct errors, their strategies for judging MT output, and their ability to reflect on differences in MT input/output（A1, A2, A4, A8, A9, A10, A11, A12, and A14). Opinions about MT from students（A2, A3, A8, and A14), teachers（A5）and both（A6 and A7）were explored in half of the articles and six of them investigated what MT was used for and/or how often it was used（students: A2, A3, and A8; teachers: A5; both: A6 and A7). Some articles contrasted different groups, making comparisons between teachers and students（A2, A6, and A7), level of proficiency（A2 and A11), and target language（A2). In terms of skills MT was used for, there was a tendency to focus on writing (A2, A3, A4, A10, A12, and A13), but reading was central to two studies（A1 and A3). Some articles focused on quite discrete areas. For example, A9 focused on neologisms and A10 focused on determiners, paraphrasing, and collocations. MT training needs were investigated in four articles (A5, A7, A8, and A10), with others drawing conclusions on training needs after conducting their studies.

Research Methods: A range of research methods were used in the selected studies. Surveys featured most prominently, with five reporting solely on survey data (A2, A3, A5, A6, and A7), two reporting on pre- and post-treatment surveys (A10 and A11), one that combined surveys with error correction tests（A8), and one that paired surveys on the researchers' evaluation of an MT tool with students' evaluation of it（A14). Tests were given in several studies, with two evaluating students' ability to notice and/or correct MT output errors (A8 and A11) and one evaluating their ability to produce accurate output with MT (A10). One study compared students' translations to MT translation (A9). Students' screen recordings and focus group discussions were used as data in one study（A4）and reflection videos were used in another（A12). Other data sources included teacher observation of students' discussions（A1）and comparisons of texts written with and without MT（A13). Additional details related to the context, research focus, data collection and MT tools are listed in Table 2 below and elaborated upon in the discussion section.

Table 2. Selected studies' context, research areas, research methods and MT tools

Authors, Year and Context	Research Areas	Research Methods and MT Tools
1. Bavendiek（2022） First-year German language learners at a university in the UK (site assumed from author's affiliation). Student numbers not stated (data through teacher observation). **Proficiency level:** CEFR A1-B2	1. Investigate students' ability to notice ungrammatical or incorrect output from Google Translate 2. Evaluate students' ability to meaningfully reflect on differences between the source text and Google Translate output 3. Assess whether Google Translate output is sufficient for engagement with literary extracts	**Data:** Teacher observation of students' in-class discussions during an activity in which they read the lyrics of a German song with MT-produced parallel text in English. The activity aimed to raise awareness of the transcultural nature of translation, with discussions focusing on the accuracy of form and content. **MT Tool:** Google Translate
2. Alm & Watanabe（2022） 150 students and 12 teachers (Chinese, French, German, Japanese, Spanish) at a university in New Zealand. **Proficiency level:** Beginner, intermediate, advanced	1. Investigate contexts in which L2 learners use MT, if this differs by language or level, and if their practices match teacher expectations 2. Investigate how L2 learners use MT for L2 writing and if this differs by language or level 3. Investigate MT post-editing practices and whether they differ by language or level 4. Assess how helpful MT is seen to be for L2 writing by teachers and students of different languages and levels	**Data:** A survey for students on MT use for L2 learning generally and L2 writing in particular, plus perceptions of MT's helpfulness. A survey for teachers on their personal experience with MT, perceptions of student use, and views on MT's usefulness. Likert scales, multiple-choice questions, and open-ended questions were included. **MT Tools:** Google Translate, DeepL
3. Powell et al.（2022） 100 graduate students in English STEM courses at a university in South Korea. **Proficiency level:** Not stated	1. Investigate the extent to which students use MT in and out of school 2. Investigate the extent to which students use MT to improve English writing and reading 3. Investigate students' views on MT's usefulness and acceptability 4. Investigate students' input strategies when using MT to aid with L2 writing and reading	**Data:** A survey for students with multiple-choice questions and Likert scales. Areas covered: Frequency of use for various tasks, input methods when using MT for reading and writing English texts, opinions on appropriate usage and usefulness. **MT Tools**: Not stated

4. Chang (2022) 4 students in an English-Chinese translation course at a university in Taiwan. **Proficiency level:** CEFR B2-C1	1. Investigate the strategies L2 learners use to make judgements about MT output for written translations	**Data:** Pre-task and post-task screen recordings of real-time translations in both directions (observation sheet used for analysis). The task was a course that included MT editing training. Focus group discussions elicited students' reflections/views. **MT Tools:** Google Translate, Microsoft Translator, Baidu, Youdao
5. Ohashi (2022) 153 foreign language teachers (6 languages) at universities throughout Japan. **Proficiency level:** Not relevant (teachers only)	1. Investigate language teachers' use of MT in their personal lives and their L2 courses 2. Investigate teachers' views on the use of MT in L2 education 3. Investigate teachers' knowledge of how to aid students to use MT for L2 development and their willingness to learn more	**Data:** A survey for language teachers that collected information about MT use, views and practices through multiple-choice questions and Likert scale items. **MT Tools:** Not stated
6. Ata & Debreli (2021) 462 EFL students and 34 teachers at a Turkish university. **Proficiency level**: CEFR A2+, B1, beginner, elementary, pre-intermediate, intermediate	1. Investigate the frequency and purposes of English learners' MT use and their views on its effectiveness and ethical use for learning English 2. Investigate the frequency and purposes of English instructors' MT use and their views on its effectiveness and ethical use for learning English 3. Investigate students' and instructors' beliefs on each other's views of MT use	**Data:** Different surveys for teachers and students on MT usage, perceptions of quality and ethical use, plus their perceptions of each other's views about MT. Multiple-choice questions, Likert scales, and open-ended questions were included. **MT Tools:** Google Translate, Yandex Translate, Microsoft Translator
7. Delorme Benites et al. (2021) 1,926 students and 666 teachers at four Swiss universities. Target languages not stated. **Proficiency level:** Not stated	1. Investigate teachers' and students' awareness of MT, user experiences, beliefs and attitudes, and training needs	**Data:** Similar surveys for students and teachers that covered the areas under investigation (partly with mirror questions) were completed in German, English and French. The surveys included 248 closed-ended and open-ended questions, but most were not reported in the article. **MT Tools:** Not stated
8. Loock et al. (2022) Pilot study: 159 English students at a French university who also studied another language. Follow-up study: Students from the same program (survey, n = 164; two error correction tasks, n = 196 and 158). **Proficiency level:** Not stated	1. Assess how students use MT in order to create an appropriate training program Note: This article reported on multiple studies so it is expected that more specific research questions and aims have been reported elsewhere	**Data:** A survey for students that addressed frequency of MT use, MT tools and methods used, satisfaction with MT, and beliefs about their ability to assess output quality (question types not stated). Two error correction tasks were also conducted (students corrected English-French MT output generated by DeepL). **MT Tools:** Google Translate, DeepL, Systran, Reverso Traduction
9. Awadh & Khan (2020) 55 students in a translation course (English-Arabic) at a university in Yemen. **Proficiency level:** Not stated	1. Investigate challenges faced by students in translating neologisms from English into Arabic 2. Investigate differences between human and MT translations of neologisms	**Data:** A 24-item test with English sentences that were translated to Arabic by a) participants and b) MT. Each item contained a neologism. Two translators evaluated the human/MT translations. **MT Tools:** Google Translate, Systran, Al-Wafi

10. Mirzaeian（2021） 20 students in an intermediate English writing course at a university in Iran (four-day workshop on MT). **Proficiency level:** TOEFL scores ranged from 310-677	1. Assess the effect of teaching students MT editing techniques for determiners, paraphrasing and collocations 2. Compare students' use of online bilingual dictionaries and MT for determiners, paraphrasing and collocations 3. Investigate how students use MT editing techniques in draft writing 4. Investigate students' attitudes and perceptions towards MT editing techniques	**Data:** 1. Pre-treatment survey (background information). 2. Treatment (overview of how to use Google Translate, training on editing input and output) 3. Pre-tests and post-tests to check the effect of training. 4. Post-treatment survey (results suggest Likert scales were used, but it was not explicitly stated). **MT Tool:** Google Translate
11. Yoon Chon（2022） 97 EFL middle school students in South Korea. **Proficiency level:** Low, mid and high proficiency (based on Test of Skills in the English Language)	1. Assess how well EFL students can correct MT errors by error type and English level 2. Investigate how the use of error correction strategies differs by English level 3. Investigate the types of error correction strategies that lead to successful correction of MT errors at different English levels	**Data:** Students did a survey on MT usage, received MT output correction training, then did an MT error correction test (based on errors in researcher-generated MT translations of text from the students' textbooks). Data on students' correction strategies was collected via the error correction test. **MT Tool:** Google Translate
12. Kennedy（2021） 40 students in an English writing course at a Japanese university. **Proficiency level:** Average TOEIC score of 570 (equivalent to CEFR B1)	1. Explore how students in an English for Academic Purposes (EAP) course completed written tasks Note: MT became relevant as it was found to be used by over half	**Data:** Three-minute reflective videos in which students described their strategies for improving their English writing (with MT and other tools) and the impact of these strategies on their learning. **MT Tools:** Not stated
13. Chon & Shin（2020） 65 university students studying English in South Korea. **Proficiency level:** high intermediate to advanced	1. Investigate differences in lexis, syntax and cohesion in three types of writing: direct writing (written in the L2 only), translated writing (written in the L1 and translated without help into the L2), and MT writing (written in the L1 and translated into the L2 with MT)	**Data:** Coh-Metrix software was used to analyse texts written in three ways: direct writing, translated writing, and MT writing (defined in the column to the left). The software provided linguistic analysis of lexis, syntax and cohesion. **MT Tool:** Google Translate
14. Kharis et al.（2021） 12 German majors at an Indonesian university and the researchers' themselves. **Proficiency level:** CEFR B1+	1. Evaluate the MT tool Takarir in terms of appearance/functions (from the researchers' perspective) 2. Assess the performance of Takarir in terms of accuracy, clarity and naturality when used in courses for simultaneous translation of Bahasa Indonesian to German and vice versa	**Data:** The researchers completed a user survey to rate the appearance and features of Takarir. Students completed a multiple-choice survey about its accuracy, clarity and naturality after using it for real-time translation in two seminars. Open questions were mentioned but not reported. **MT Tool:** Takarir

RQ2. What are the key findings of recent research into the use of MT in language education?

The level of detail given about the findings of the selected studies varied widely. This is partly due to differences in article length, which ranged from 5 to 25 pages. Furthermore, some articles reported on multiple research questions, while others were more focused in their scope. Due to this, the present study does not aim to provide a comprehensive overview of all findings. Rather, Table 3 draws together salient points from the articles selected for review. As not all points of potential interest are reported here, readers who wish to know more are encouraged to refer to the original articles for further information. Implications of the findings are addressed in the discussion section.

Table 3.　Selected studies' key findings

Article	Key Findings
1. Bavendiek (2022)	Students could find translation errors when L1 output made the errors obvious, and used their knowledge and other tools to correct them. Closer reading and teacher support was needed to spot literal translations that did not fit the context. Discussions about cultural background helped students to come up with more appropriate, nuanced English translations. MT was a valuable starting point.
2. Alm & Watanabe (2022)	MT often replaced dictionaries. Common usages (main users): Grammar corrections (beginners), vocabulary contextualisation with MT (advanced), proofreading (non-alphabet-based languages), multi-modal features (non-alphabet languages); reverse translation (widespread). Advanced learners detected/corrected errors better and viewed MT as more helpful for writing. Teachers' ideas about how students used MT did not strongly align with students' reported practices.
3. Powell et al. (2022)	Most students used MT at least sometimes for graded and ungraded work. MT was used more in reading and writing for sentence-level translations than word, paragraph or whole-text translations. The vast majority of students felt MT use should be permitted when preparing graded work in English as a medium of instruction (EMI) courses and found it useful for reducing the time needed for tasks. Over half felt it helped them to learn English.
4. Chang (2022)	Students' ability to judge MT output improved when they used multiple tools (which was a focal point of the study's MT training). Translations were checked more critically in post-tasks, with screen recordings showing increased use of multiple MT tools, online dictionaries, and search engines (e.g., image search, collocations) to cross-check MT output.
5. Ohashi (2022)	Teachers had high levels of personal MT use and evaluated it favourably as a learning tool, but course integration was limited. Most felt MT was used by a proportion of their students to cheat, but that use for L2 learning was more prevalent. Few felt it should be heavily restricted or banned. Most agreed students need MT guidelines, but few provided guidancce in all of their courses. Many lacked knowledge on how to support students to develop their L2 skills with MT and wanted to learn more.

6. Ata & Debreli（2021）	Almost all students used MT（mainly for single-word/phrase translation）, rated output quality low, and felt ethical use depended on how they used it, with longer translations seen as unethical. Most teachers used MT less often than students and over 25% never used it. Teachers' and students' views on ethical use of MT aligned and diverged for different uses. There were also gaps in perceptions of each other's beliefs and practices.
7. Delorme Benites et al.（2021）	Most teachers and students had used MT. It was used for academic work more than other purposes, but was not openly addressed in courses. Most teachers didn't conduct MT training and most students hadn't received any MT instruction. Teachers' and students' views on each other's awareness of MT's risks differed, with teachers more sceptical. Both groups saw MT as a relatively useful tool that won't reduce the need to learn languages. A need for MT literacy training for both groups was identified.
8. Loock et al.（2022）	**Pilot survey:** Most students used MT. Copying entire texts was rare（mainly sentence-level and word-level）. Some used MT for gist before translating or translated then checked MT. Almost all felt they could identify output errors. **Follow-up survey:** Students used MT for translation exercises, as a writing aid, for reading comprehension, and to check grammar. Most were sometimes or often satisfied with MT output. **Error correction tasks:** Most couldn't identify errors in MT output without help, but could correct many of them when underlined, so raising awareness of errors is vital.
9. Awadh & Khan（2020）	Neologisms（newly coined terms）posed great problems for both the human translators（students）and MT. Translation ratings from evaluators（accurate, acceptable, unacceptable and untranslated）showed that over half of students' translations were unacceptable/untranslated, with this figure rising substantially for MT. Specialized glossaries and dictionaries were recommended over MT for translation of neologisms.
10. Mirzaeian（2021）	After training students to edit MT input, post-test gains were found for determiners, paraphrasing and collocations, but only determiners reached statistical significance. Most participants highly evaluated their training and felt editing input improved their correction skills and writing. Paraphrasing was most challenging for low level learners. Training outcomes varied by proficiency, so training should target level-appropriate tasks.
11. Yoon Chon（2022）	Most students had used MT for schoolwork, mainly to check vocabulary. Tests on students' ability to correct output errors（based on translations of textbook content）showed over half were corrected, with greater success repairing word order and mistranslation errors（word and phrase level）than verb tenses. Higher level students had more success. Strategies: Students at all levels guessed from context and used literal translation. High level learners made more use of background knowledge.
12. Kennedy（2021）	Over half of the students used MT as a final step in drafting written work, translating from L2 to L1 to check accuracy or translating an L1 draft or L1 approximations of their already-written L2 draft. Students checked for differences and errors in drafts and tried to repair them using trusted sources（dictionaries, textbooks）. Some students reported an increase in their writing confidence and grades.
13. Chon & Shin（2020）	MT use facilitated fluency（longer sentences and texts）in students' writing. Higher rates of low frequency lexical items were found when translating unassisted and translating with MT than when writing directly in the L2. However, lexical diversity was slightly lower for MT-assisted translations. Texts translated with MT showed overall higher syntactic complexity and cohesion.
14. Kharis et al.（2021）	The researchers positively evaluated the appearance/features of Takarir, which supports 44 languages and dialects, but noted operating instructions were only given in Bahasa Indonesian and English. Students used Takirir for real-time translation in seminars. Half agreed the German-Indonesian translations were appropriate in terms of accuracy and clarity, but lacked naturality. For Indonesian-German translations, half felt they were clear, but fewer felt they were accurate or natural.

Discussion

This systematic review of studies on the use of MT in language education offers several points to note in terms of the research areas and methods. Firstly, there was a tendency for research to be conducted with students more than teachers, with only four studies covering teachers' perspectives. Given the important role teachers play in decisions related to language education, further investigation is warranted. For example, while some studies examined how students' practices changed after training, no studies investigated this with teachers. Such studies would make a valuable contribution, so research in this area is encouraged. Furthermore, there was a tendency towards survey data, and while this kind of research makes a valuable contribution, there is room for more work that examines MT from other angles and combines multiple data collection methods. Studies also tended to have single or short data collection periods. Some studies conducted pre- and post-intervention tests/surveys, but the follow-up data was collected soon after the intervention. Longitudinal studies that gather data from the same participants to ascertain long-term effects of intervention, including changes in practices and attitudes, are needed.

Another noteworthy issue was that students' language level was not always stated in studies. Moreover, it was difficult to compare students in different studies when it was included as levels were described using institution-specific course names/levels and a wide range of proficiency tests, generally without an indication of equivalency. Researchers are often restricted in the data they can gather about participants' proficiency level so there is no easy solution to this problem, but the situation could be improved if reference to a well-known framework such as the CEFR was made in articles. Although the CEFR is a European-based framework, it was also referred to in studies beyond Europe（Indonesia, Japan, Turkey and Taiwan), so it may have merit as a forerunner. Another advantage to using the CEFR is that there is a publicly available concise overview of the CEFR scales[1] that researchers could use to indicate the approximate CEFR level when reporting on courses that have generic terms like "intermediate". Readers who are unfamiliar with the CEFR could also refer to these scales. As some studies in this review highlighted the importance of MT being used for level-appropriate tasks, it is essential that steps be taken towards clarifying the language proficiency of participants so that teachers and researchers can make judgements about the relevance of research to their own contexts.

In terms of results, studies that assessed the usefulness of MT found that it can be a good starting point, but that students need guidance in manipulating the input, evaluating output, and correcting translation errors when they exist. There was evidence to suggest that students' ability to do this was partially connected to their proficiency level and that it could in some cases be improved with training. Furthermore, the target language（or L1-L2 pair）was shown to have an impact on the MT tools that students chose and the features within them that they used, so teachers should familiarise themselves with appropriate tools and functions. Teachers can play an important role by providing task-specific training that matches

[1] Available at https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

students' L1/L2 profile and their L2 proficiency level, but they need more skills and knowledge to do this. The reviewed studies suggest that teacher training is generally lacking, so institutions are urged to address this and teachers are encouraged to seek professional development opportunities.

Studies also highlighted the importance of teachers and students communicating with each other, as mismatches were found in their views on MT, their beliefs about how it was being used, and actual practices. MT is a tool that could potentially do the work for students, leading to loss of learning and academic dishonesty, but there was evidence to suggest that whole-text translation was not common, with studies that investigated this area finding that students' usage was more often limited to sentence-level and word-level translations. This is important to bear in mind, as there were reports of teachers concerns over students' unethical use of MT. Greater dialogue is needed between these two groups to facilitate mutual understanding, clarify expectations, and build trust.

Conclusion

This systematic review of research on the use of MT in language education has brought together findings from 14 open access articles. Its focus on articles published between 2020 to 2022 extends the work of previous systematic literature reviews conducted by Lee (2021) and Klimova et al. (2023), providing a valuable update in a domain that is constantly changing due to technological advances. It is imperative to revisit this topic in the near future, not only because tools dedicated to MT are continually evolving, but also due to the release of newly-developed AI tools such as ChatGPT that have advanced translation capabilities. A limitation of this study is that it only reviewed articles from one database, but this is countered in part by the level of detail that could be given about each reviewed article due to the small number included. Looking at the findings of these studies together offers valuable new insights that extend the original contributions made by each article. It is hoped that the findings from this systematic review will stimulate further research and be of benefit to language teachers and educational institutions.

Cite in English as: Ohashi, L. (2024). Machine Translation in Language Education: A Systematic Review of Open Access Articles *Kenkyu Nenpou: The Annual Collection of Essays and Studies, 70*, 105–125.

References

Alm, A. & Watanabe, Y. (2022). Online machine translation for L2 writing across languages and proficiency levels. *Australian Journal of Applied Linguistics, 5* (3), 135–157. https://doi.org/10.29140/ajal.v5n3.53si3

Ata, M. & Debreli, M. (2021). Machine translation in the language classroom: Turkish EFL learners' and instructors' perceptions and use. *IAFOR Journal of Education: Technology in Education, 9* (4), 103–122. https://files.eric.ed.gov/fulltext/EJ1318690.pdf

Awadh, A. N., & Khan A. S. (2020). Challenges of translating neologisms comparative study: Human and machine translation. *Journal of Language and Linguistic Studies, 16* (4), 1987–2002. https://doi.org/10.17263/jlls.851030

Bavendiek, U. (2022). Using machine translation as a parallel text to access literature for modern language learning. In C. Hampton & S. Salin (Eds), *Innovative language teaching and learning at university: Facilitating transition from and to higher education* (pp. 57–67). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.56.1373

Chang, L-C. (2022). Chinese language learners evaluating machine translation accuracy. *The JALT CALL Journal, 18* (1), 110–136. https://doi.org/10.29140/jaltcall.v18n1.59

Chon, Y. V. & Shin, D. (2022). Direct writing, translated writing, and machine-translated writing: A text level analysis with Coh-Metrix. *English Teaching,* (*75*) *1*, 25–48. https://doi.org/10.15858/engtea.75.1.202003.25

Delorme Benites, A., Cotelli Kureth, S., Lehr, C., & Steele, E. (2021). Machine translation literacy: A panorama of practices at Swiss universities and implications for language teaching. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation: Short papers from EUROCALL 2021* (pp. 80–87). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.54.1313

Kennedy. O. (2021). Independent learner strategies to improve second language academic writing in an online course. In N. Zoghlami, C. Brudermann, C. Sarré, M. Grosbois, L. Bradley, & S. Thouësny (Eds), *CALL and professionalisation: Short papers from EUROCALL 2021* (pp. 184–188). Research-publishing.net. https://doi.org/10.14705/rpnet.2021.54.1330

Kharis, M., Kisyani, Suhartono, & Yuniseffendri. (2021). Takarir: A new simultaneous translator voice to text to promote bi/multilinguality. *Journal of Language and Linguistic Studies, 17* (3), 1175–1183. https://www.jlls.org/index.php/jlls/article/view/2629/869

Klimova, B., Pikhart, M., Delorme Benites, A., Lehr, C., & Sanchez-Stockhammer, C. (2023). Neural machine translation in foreign language teaching and learning: A systematic review. *Education and Information Technologies, 28*, 663–682. https://doi.org/10.1007/s10639-022-11194-2

Lee, S.-M. (2021). The effectiveness of machine translation in foreign language education: A systematic review and meta-analysis. *Computer Assisted Language Learning*. Advance online publication. https://www.tandfonline.com/doi/abs/10.1080/09588221.2021.1901745

Loock, R., Lechauguette, S, & Holt, B. (2022). Dealing with the “elephant in the classroom”: Developing language students’ machine translation literacy. *Australian Journal of Applied Linguistics, 5* (3), 118–134. https://doi.org/10.29140/ajal.v5n3.53si2

Mirzaeian, V. (2021). The effect of editing techniques on machine translation-informed academic foreign language writing. *The EuroCALL Review, 29* (2), 33–43. https://doi.org/10.4995/eurocall.2021.13120

Ohashi, L. (2022). The use of machine translation in L2 education: Japanese university teachers' views and practices. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouësny, & M. J. Whelpton (Eds), *Intelligent CALL, granular systems, and learner data: Short papers from EUROCALL 2022* (pp. 308–314). Research-publishing.net. https://doi.org/10.14705/rpnet.2022.61.1476

Powell, N., Baldwin, J., & Manning, J. (2022). Graduate STEM student perspectives and implementation of machine translators in South Korea. *International Journal of Technology in Education and Science* (*IJTES*)*, 6* (2), 237–253. https://doi.org/10.46328/ijtes.322

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