English Teaching, Vol. 79, No. 4, Winter 2024, pp. 33-54
DOI: https://doi.org/10.15858/engtea.79.4.202412.33
http://journal.kate.or.kr

From No Aid to Neural Aids: Exploring the Influence of Machine Translation on L2 Reading Problems*

Gyeong Jin Park and Yuah V. Chon **

Park, Gyeong Jin, & Chon, Yuah. V. (2024). From no aid to neural aids: Exploring the influence of machine translation on L2 reading problems. *English Teaching, 79*(4), 33–54.

Neural machine translators (NMTs), such as *Google Translate*, may assist second language (L2) readers with general comprehension. However, previous empirical studies show mixed results regarding their effectiveness. In this study, 145 Korean English learners from a girls' high school were asked to solve three types of reading comprehension problems (*grammar judgment, inferring meaning from context, inferring main idea*) under three reading conditions (no aid, MT, glossary). Overall, when using MT, reading comprehension scores were higher than in either the no aid or glossary conditions individually. However, none of the reading aid conditions improved *grammar judgment*. Only mid-proficiency learners benefited from MT in both *inferring meaning from context* and *inferring main idea* tasks. The results suggest that the glossary may have interrupted the flow of the reading process. With the widespread availability of MT as an online reference tool, L2 teachers should consider incorporating MT as a legitimate reading aid for different proficiency levels and reading purposes.

Key words: machine translation, second language reading, reading problems, second language proficiency

*This work was supported by the research fund of Hanyang University (202200000003525).

**First Author: Gyeong Jin Park, Graduate Student, Department of English Education, Hanyang University

Corresponding Author: Yuah V. Chon, Professor, Department of English Education, Hanyang University; 222 Wangshimli-ro, Seongdong-Gu , Seoul 04763, Korea; Email: vylee52@hanyang.ac.kr

Received 30 September 2024; Reviewed 28 October 2024; Accepted 12 December 2024

1. INTRODUCTION

With the arrival of neural machine translators (NMTs), the quality of machine translators (MTs) has improved significantly (Shterionov et al., 2018; Van Brussel, Tezcan, & Macken, 2018). Nonetheless, users often receive *raw* machine translation (MT) output containing several translation errors, which leads to the questions: How useful are MTs (Macken & Ghyselen, 2018) and should they be used for carrying out MT-assisted reading tasks? Language teachers tend to be skeptical about the quality of MT output and tend to equate students' use of MT with academic dishonesty (Carré, Kenny, Rossi, Sánchez-Gijón, & Torres-Hostench, 2022; Ducar & Schocket, 2018). Despite this, there is ample evidence in the literature that students—by their own admission—are using it with increasing frequency to support language use and learning (Jolley & Maimone, 2022), including reading in a second or foreign language (Klimova, Pikhart, Benites, Lehr, & Sanchez-Stockhammer, 2023). However, it is not clear how helpful MT is for learners with different second language (L2) proficiency levels in tackling different types of reading comprehension problems.

The extent to which MT may be useful in L2 reading tasks has not been researched to the same degree as L2 writing (Maghsoudi & Mirzaeian, 2020) even though reading is a non-optional skill and the primary means of acquiring language and content knowledge. Hence, this study aims to examine whether MT can support L2 learners' academic reading, and to observe if L2 proficiency and types of reading problems may influence the learners' ability to benefit from MT. For this purpose, L2 high school learners were asked to solve reading comprehension problems counterbalanced for three types of reading aid conditions (no aid, MT, glossary) and three types of reading problems (*grammar judgment, inferring meaning from context, inferring main idea*). The study results are expected to elucidate whether and how raw MT output can be used to help L2 readers solve reading comprehension questions (RCQs).

2. BACKGROUND

2.1. Using Translation in Language Teaching

Recent years have seen a resurgence of translation in language teaching, indicating the need for a reevaluation of L1 use and translation in the L2 classroom (Jolley & Maimone, 2022). Cook (2010) viewed translation on a continuum, with the tightest word-for-word translation on one end and the loosest paraphrase and interpretations on the other, and took a broad approach to translation by using the term "translation in language teaching" (TILT). TILT serves as an umbrella term for all types of translation in the language classroom.

Accordingly, MT can be included as a type of TILT.

However, the case for and against translation remains a complicated issue given that the legitimacy of using it in the L2 classroom and its efficacy may vary with the social and linguistic relationships between a student's L1 and L2 (Carreres, 2014). Zojer (2009) argues that translation potentially causes interference errors owing to negative transfer from the mother tongue, and hinders the development of free-flowing self-expression as the student is required to translate specific texts. Nord (1996) also claimed that translation reduces the text to a string of disconnected and isolated sentences, thus ignoring intertextual devices in which the focus shifts from translational questions to the quest of finding mere semantic equivalences.

In contrast, a plethora of findings support the benefits of translation in language teaching. Translation and the use of L1 as the medium of instruction can promote reading comprehension through better perception of the written texts (Marzban & Azizi, 2011). Translation can be used to help learners notice idiosyncrasies in the two languages (Niño, 2009), realize that two cultures or languages may express similar items in different ways, notice different registers and the importance of appropriacy (Hatim & Munday, 2019), understand the importance of collocations, and see that acceptable L1 translation is associated with congruency (Sonbul, El-Dakhs, & Al-Otaibi, 2022). In contexts where independent and sustained reading is important, translations or parallel texts can help learners develop greater autonomy and reading motivation, thereby preparing them for future, fully independent reading (Weydt, 2009). Translation provides opportunities for focus on form (Hentschel, 2009), increases intake of L2, and enables teachers to perceive TILT as a positive pedagogical tool (Bruen & Kelly, 2017).

2.2. Machine Translation for Second Language Reading

Studies on MT in language teaching and learning have predominantly focused on L2 writing (Cancino & Panes, 2021; Chang, Chen, & Lai, 2022; Chon, Shin, & Kim, 2021; Chung & Ahn, 2022; Jolley & Maimone, 2022; Klimova et al., 2023; Kol, Schcolnik, & Spector-Cohen, 2018; Lee, 2023; Stapleton & Kin, 2019; Yoon & Chon, 2022), and there is a small under-researched body of literature on the use of MT for advancing L2 reading comprehension skills. There exists little empirical evidence regarding the usefulness of translations for understanding of L2 texts (Maghsoudi & Mirzaeian, 2020). Moreover, previous studies on MT-assisted L2 reading are limited to survey studies of language learners (Case, 2015) and quality evaluations of machine-translated output vs. human translated texts (Fuji et al., 2001; Jones et al., 2005; Pfafflin, 1965; Tomita, Shirai, Tsutsumi, & Matsumura, 1993).

Scarton and Specia (2016) used RCQs at different levels of complexity to assess the

quality of machine- and human-translated texts. When the texts were evaluated based on answers by fluent speakers of the target language, the researchers found that RCQs can provide valuable information about the quality of an entire machine translated document. Karnal and Pereira (2015) aimed to understand the cognitive strategies involved in reading L2 academic texts by assessing reading comprehension wherein the same academic text (abstracts) was read either using Google Translate (GT) or without any aid. Macken and Ghyselen (2018), inspired by Scarton and Specia (2016), compared the results of reading comprehension tests of human-translated and raw (unedited) machine-translated texts. Their results showed that 74% of the participants could identify whether a translation was produced by a human or a machine.

The paradigm shift in the MT landscape made it necessary to test the reading comprehension of NMT models by the end users of those translations. Castilho and Guerberof Arenas (2018) conducted a pilot study to measure the impact of the quality of two MT paradigms—NMT and statistical MT (SMT). Results showed that participants in the Spanish and simplified Chinese groups were able to complete more tasks successfully when using the NMT translations than when using SMT. Odo (2020) conducted a study to examine whether MT can support academic reading and writing development of L2 learners when the experimental group was provided with opportunities to read academic articles with MT support. After the treatment, the experimental group's L2 writing neither improved nor deteriorated. Kim and Cha (2020) examined whether MT can be effective in strengthening learners' performance in reading comprehension. Freshmen university students were assigned either to the word list group or the MT group. When both groups took multiple-choice reading comprehension tests before and after the experiment (with the same tests), within group analysis indicated significant differences in reading scores between the pre- and post-tests for the wordlist group, but not for the MT group. Between group analysis indicated that the wordlist group performed better than the MT group.

Maghsoudi and Mirzaeian (2020) found MT output to be as successful as human translation in terms of generating comprehensible texts; there was a negligible difference between the control group, who was given the human translation, and the experimental group, who was exposed to MT, when reading comprehension scores on identical items were compared. Chen (2020) used translation activities of pop songs to familiarize 16 English as a foreign language (EFL) Taiwanese university students with the potentials of MT tools. The results revealed that the majority of students (n = 11) input either paragraphs or whole texts into GT to obtain an initial rough translation. Klimova et al. (2023) indicated that NMT is an efficient tool for developing both productive (speaking and writing) and receptive (reading and listening) language skills, and provided a systematic review of neural MT; they found the NMT texts to be of comparable quality with non-NMT translated texts. They also reported that NMT tools are especially suitable for advanced L2 learners, whose higher

proficiency level enables them to critically reflect on the output of neural MT texts.

In summary, recent studies examining MT as a tool for L2 reading comprehension are scarce and yield ambiguous results. The effectiveness of MT for readers of varying proficiency levels across different types of comprehension problems remains largely unexplored. Additionally, some studies, such as those by Karnal and Pereira (2015), and Kim and Cha (2020) have methodological flaws including repetitive use of texts and comprehension questions, leading to potential practice effects that could skew results. Considering MT's current limitations, as noted by Chon et al. (2021), this study aims to assess whether MT aids in reading comprehension more effectively than no aid or alternative aids like glossaries.

2.3. Research Questions

Studies that investigated the efficacy of MT for L2 reading left unresolved questions regarding how learners of different proficiency levels use MT with different types of reading problems. The following questions guided our study:

1) How did reading aids (no aid, MT, glossary) affect the reading comprehension of L2 learners of different L2 proficiency levels (low, mid, high)?
2) How effective were the different reading aids for different reading problems (*grammar judgment, inferring meaning from context, inferring main idea*) for L2 readers of different L2 proficiency levels?

3. METHODOLOGY

3.1. Participants

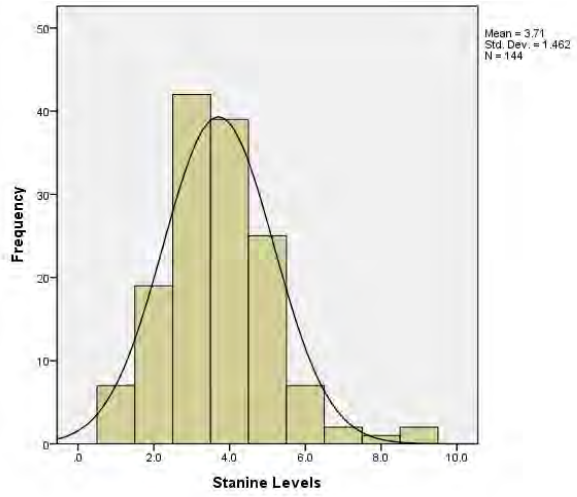
Participants of this study comprised 145 senior high school students attending a private girls' high school located in Daegu, South Korea. They were recruited by convenience sampling. However, the range of scores from the students' most recent mock Korean College Scholastic Ability Test (CSAT) (Suneung), conducted at the national level, indicated that the participants represented senior high school students in South Korea. As shown in Figure 1, the learners could be categorized as high- (stanine [1] 1-2: n = 26, 13.83%), mid- (stanine 3-4: n = 81, 56.25%), and low-proficiency learners (stanine 7-9: n = 37, 25.69%). In the

[1] Stanine (STAndard NINE) is a method of scaling test scores on a nine-point standard scale with a mean of five and a standard deviation of two.

context of the present study, stanine levels were used as the standard unit to assess students' performance on the Korean CSAT.

FIGURE 1

Stanine Levels of Participants



3.2. Instruments

RCQs in the form of multiple-choice questions were deemed valid measures for assessing learners' L2 reading ability and comparing their reading performance across three different reading aid conditions (no aid, MT, glossary) as displayed in Figures 2 to 4. MCQs provide a structured and standardized way to evaluate learners' comprehension and minimize subjective interpretation in scoring, ensuring a fair and consistent assessment of reading performance (Maghsoudi & Mirzaeian, 2020; Scarton & Specia, 2016). RCQs were selected from previous Korean CSATs (https://www.kice.re.kr/sub/info.do?m=0205&s=english) conducted by the South Korean Ministry of Education.

At the time of the study, 27 RCQs not featured on the national exam over the past seven years were selected. These were divided into three categories: 1) *grammar judgment* (9 items), where learners chose the most fitting language item; 2) *inferring meaning from context* (9 items), requiring identification of the most logical expression in text flow; and 3) *inferring main idea* (9 items), involving selecting an appropriate title for a text. Each passage was about 120 words (see Figures 2 to 4 for each item type). The selection of these item types was informed by previous L2 reading research emphasizing the importance of vocabulary (Schmitt, Jiang, & Grabe, 2011), grammar (Zhang, 2012), and discourse

knowledge (Nassaji, 2007) in reading comprehension.

FIGURE 2

Grammar Judgment with *No Aid*

25. No matter what we are shopping for, it is not primarily a brand we are choosing, but a culture, or rather the people associated with that culture. (A) Whatever / Whether you wear torn jeans or like to recite poetry, by doing so you make a statement of belonging to a group of people. Who we believe we are (B) is / are a result of the choices we make about who we want to be like, and we subsequently demonstrate this desired likeness to others in various and often subtle ways. Artificial as this process is, this is what becomes our 'identity,' an identity (C) grounded / grounding on all the superficial differences we distinguish between ourselves and others. This, after all, is what we are shopping for: self-identity, knowledge of who we are.

	(A)	(B)	(C)
①	Whatever	is	grounded
②	Whatever	are	grounding
③	Whether	is	grounded
④	Whether	are	grounding
⑤	Whether	are	grounded

FIGURE 3

Inferring Meaning from Context with *Machine Translation*

9. Psychologist Solomon Asch wanted to discover whether people's tendency to agree with their peers was stronger than their tendency toward independent thought and rational judgment. Asch assembled groups of twelve university students and announced that they were taking part in an experiment on visual perception. He showed them three line segments, and asked each one in turn which line was the longest. It was an easy task and the correct answer was obvious. However, Asch had secretly instructed all but the last person in each group, who was the real subject of the experiment, to say that the medium-length line was the longest. As it turned out, over 70 percent of the real subjects _____ and said that the medium-length line was the longest.

① caved in to group pressure
② figured out the correct answer
③ had problems with their vision
④ roped the other group members in
⑤ used rational judgment in their decision-making

9. 심리학자 솔로몬 애쉬(Solomon Asch)는 사람들이 동료와 동의하는 경향이 독립적인 사고와 합리적인 판단을 하려는 경향보다 강한지 여부를 확인하고 싶었습니다. Asch는 12명의 대학생으로 구성된 그룹을 구성하고 시각적 지각에 대한 실험에 참여한다고 발표했습니다. 그는 그들에게 세 개의 선분을 보여주고 어느 선이 가장 긴지를 차례로 물었다. 그것은 쉬운 작업이었고 정답은 분명했습니다. 그러나 애쉬는 실험의 진짜 대상인 각 그룹의 마지막 사람을 제외하고는 모두 중간 길이의 줄이 가장 길다고 비밀리에 지시했다. 그 결과 실제 피험자의 70% 이상이 _____하며 중간 길이의 줄이 가장 길다고 말했다.

At the time of the main study, it was not possible to provide computers to all learners (*N* = 145) simultaneously. To mimic the experience these learners would have when using online MT, and to allow for repeated access to machine translations, the test papers were designed to include GT translations (Figure 3). Equivalent conditions were created for the no-aid and glossary groups.

Before distributing the translations, they were quality checked via sentence adequacy evaluations (Moorkens, 2018). The 27 reading passages comprised 181 sentences, of which 140 produced acceptable translations (no errors) and the remaining 41 contained MT errors. However, as some of those 41 sentences contained more than one error, the total number of errors identified was 47 (missing word: 22; mistranslation: 23; ungrammaticality: 2). This meant that 22.65% of the 181 sentences had MT errors. When calculated by sentence as the unit of analysis, there were 0.25 errors per sentence. When the percentage of errors was calculated for each reading item based on the number of sentences for each item, 75.19% of the sentences produced acceptable translations.

Regarding the treatment of MT errors, these were left uncorrected to simulate real-life situations in which learners must grapple with imperfect translations. Correcting the errors would have created an artificial scenario for using machine translations. Therefore, it was crucial for MT users to extract the most useful information from the MT output, despite the presence of errors.

Providing different reading aid conditions allowed MT to be compared with a no-aid condition, establishing a baseline for learners' natural comprehension abilities without external support. Glossaries were included as they have traditionally been used in L2 reading to help learners understand vocabulary and contextual meaning. The learners could refer to the glossary at the right-hand side of the test items (Figure 4) as much as needed while solving the reading problems. The glossary offered definitions of words at the 3,000 to 7,000 word levels (3K ~ 7K) when analyzed with Compleat Lexical Tutor, Vocabprofilers (BNC-COCA 1-25k) (https://www.lextutor.ca/vp/comp/). This decision was based on the achievement standard of the Korean National Curriculum of English which aims to teach up to 3,000 word level by the end of high school year and the learners were in their final year of high school at the time of the study. The words in the glossary were arranged in order of their appearance in the reading passage. Whenever there were polysemies, the different definitions were provided. For all items, translations were not provided for the choices (distractors) of the multiple-choice items.

To remove any practice effect and to ensure all learners were solving the test items under identical conditions, three types of items (*grammar judgment, inferring meaning from context, inferring main idea*) and three reading aids were paired by counterbalancing. The order of the three item types was also alternated so that the order would not affect reading scores. This process produced three test batteries, each containing 27 RCQs (See Appendix

for configuration of item type and reading aid).

FIGURE 4

Inferring Main Idea with a *Glossary*

16. Some people work long hours even at very high levels of income. Have they got their priorities right? Most people would agree that, at a low level of income, an increase in income is likely to improve your quality of life, even if it means longer working hours. At this level, even if you have to work longer in your factory, higher income is likely to bring a higher overall quality of life by improving your health through better food, heating, hygiene and healthcare and by reducing the physical demands of household work through more household appliances. However, above a certain level of income, the relative value of material consumption in relation to leisure time is diminished, so earning a higher income at the cost of working longer hours may reduce the quality of your life.

① Does Working More Always Pay?
② Happier People Work Harder
③ Equal Pay for Equal Work
④ Consume Less, Save More
⑤ How Does Income Affect Health?

16.

priority	우선 사항, 우선권
factory	공장
overall	종합적인, 전체의
hygiene	위생
healthcare	의료, 건강 관리
household	가정
appliance	기기
relative	친척
consumption	소비
leisure	여가
diminish	줄어들다, 약해지다

3.3. Procedure

Once the test batteries were prepared for administration, they were pilot-tested with a group of learners similar to those in the present study. It was confirmed that these learners encountered no major difficulties in following the instructions or understanding the layout of the test batteries. In the main study, learners from intact classes were randomly assigned to one of the three test batteries.

3.4. Data Analysis

Statistical Package for Social Sciences (SPSS) 26.0 was used for quantitative analysis. The learners' responses to the RCQs were entered as "0" when incorrect and "1" when correct. Questionnaire responses were also entered into SPSS for analysis. Wherever relevant, scores were calculated for descriptive (Mean, SD) and inferential statistics.

For RQ1, a two-way mixed analysis of variance (ANOVA) was conducted to examine the combined effects of reading aid (independent variable; IV) and L2 proficiency (IV) on learners' reading score (dependent variable; DV). In RQ2, a follow-up analysis was

conducted for each type of reading problem by conducting a two-way mixed ANOVA (IV: type of reading aid, L2 proficiency; DV: reading score). The key assumptions for conducting two-way mixed ANOVA were carefully checked. The Shapiro-Wilk test was used to assess the normality of residuals for each group. All groups showed non-significant results ($p > .05$), indicating that the normality assumption was met. Mauchly's test of sphericity was conducted to assess the equality of variances of the differences between all combinations of related groups. The test was non-significant ($p > .05$), supporting the sphericity assumption. Levene's test was conducted to check for homogeneity of variances, and Box's M test was used to examine the homogeneity of covariances across groups. Both tests returned non-significant results ($p > .05$), confirming that these assumptions were met. Given that these assumptions were satisfied, the data were deemed appropriate for the two-way mixed ANOVA. When there were no interaction effects between the IVs, analyses were carried out to examine their main effects with appropriate inferential statistical tests.

4. RESULTS

4.1. Effects of Reading Aid and L2 Proficiency on Reading Comprehension

Regarding RQ1, the results of two-way mixed ANOVA (IV: type of reading aid, L2 proficiency; DV: reading score) showed that there was no significant interaction between reading aid and L2 proficiency (F (3.685, 250.582) = 1.033, p = .388) in their combined effect on reading scores.

Follow-up analysis was conducted for main effects of reading aid and L2 proficiency. Repeated measures one-way ANOVA indicated a significant main effect of reading aid (F (1.845, 254.576) = 19.420, p = .000, ηp^2 = .123) on reading scores. Post-hoc tests indicated that the learners performed better on the reading test with the MT (M = .575, SD = .227) than when not using any aid (M = .471, SD = .271) ($p < .001$). Learners also scored better with the MT than with the glossary (M = .455, SD = .232) ($p < .001$). However, there was no difference in reading scores between the no aid and glossary conditions (p = 1.00).

When one-way multivariate ANOVA (MANOVA) was conducted to analyze the main effect of L2 proficiency by type of reading aid, there was a significant main effect of L2 proficiency on reading scores (F (6, 268) = 26.215, $p < .001$; *Wilk's* Λ = 0.397, *partial* η^2 = .370). As shown in Table 1, L2 proficiency had a statistically significant effect on reading scores across all reading aid conditions when the alpha level was adjusted using Bonferroni correction (.05/3 = .017). Post-hoc tests indicated significant differences in reading scores between the proficiency groups; more proficient learners consistently performed better on the reading items ($p < .001$) regardless of the type of reading aid they

were using.

TABLE 1

Effects of L2 Proficiency by Reading Aid

(*N* = 139)		*M*	*SD*	*F* (2, 136)	*Partial* η^2	Post-Hoc
No Aid	High	.761	.168	45.053*	.399	Low < Mid*
	Mid	.478	.238			Low < High*
	Low	.241	.171			Mid < High*
	Total	.471	.271			
MT	High	.791	.158	38.978*	.364	Low < Mid*
	Mid	.591	.183			Low < High*
	Low	.378	.197			Mid < High*
	Total	.575	.227			
Glossary	High	.726	.167	51.570*	.431	Low < Mid*
	Mid	.450	.190			Low < High*
	Low	.263	.147			Mid < High*
	Total	.455	.232			

*p < .017 (Bonferroni correction); High: n = 26, Mid: n = 78, Low: n = 35

4.2. Effects of Reading Aid and L2 Proficiency on Reading Comprehension for Three Reading Problems

Regarding RQ2, a two-way mixed ANOVA was conducted for each type of problem (IV: Type of reading aid, L2 proficiency; DV: reading score). For each problem, there were no interaction effects between type of reading aid and L2 proficiency in their combined effects on reading scores (*grammar judgment*: F (4, 258) = .171, p = .953; *inferring meaning from context*: F (4, 254) = 2.411, p = .050; *inferring main idea*: F (4, 266) = .873, p = .480).

Based on repeated-measures one-way ANOVA and with the alpha level adjusted using Bonferroni correction (p < .017), the main effects of the reading aid on different types of reading problems showed varied results. The effect was not significant for *grammar judgment* (F (2, 262) = 3.170, p > .017). In contrast, a significant reading aid effect was observed for *inferring meaning from context* (F (2, 258) = 18.693, p < .017, $\eta p2$ = .127), and for *inferring main idea* (F (2, 272) = 5.034, p < .017, ηp^2 = .036) as shown in Table 2.

Post-hoc tests indicated that for both inferring meaning from context and inferring main idea, the MT was more effective than the ‘no aid’ condition (p < .017) in obtaining higher reading scores. However, there was no significant difference in reading scores between the glossary and ‘no aid’ conditions. For *grammar judgment*, the type of reading aid had no significant effect on reading scores.

TABLE 2

Effects of Reading Aid by Reading Problem

	M	*SD*	*F*(*df*)	*Partial* η^2	Post-Hoc
Grammar Judgment (*N* = 132)					
No aid	.487	.343	3.170 (2, 262)	.024	N/A
MT	.538	.315			
Glossary	.453	.322			
Inferring Meaning from Context (*n* = 130)					
No aid	.480	.345	18.693* (2, 258)	.127	1 < 2*
MT	.646	.301			1 = 3
Glossary	.444	.328			2 > 3*
Inferring Main Idea (*n* = 136)					
No aid	.451	.351	5.034* (2, 272)	.038	1 < 2*
MT	.555	.334			1 = 3
Glossary	.475	.325			2 = 3

*p < .017 (Bonferroni correction)

A one-way MANOVA was conducted to examine the main effects of L2 proficiency according to reading aid conditions within each type of reading problem (IV: L2 proficiency; DV: reading scores for reading aid). There were statistically significant differences in reading scores based on a learner's proficiency for all three problems (*grammar judgment*: F (6, 254) = 10.544, p < .001; *Wilk's* Λ = .641, *partial* η^2 = .199; *inferring meaning from context*: F (6, 250) = 12.866, p < .001; *Wilk's* Λ = .584, *partial* η^2 = .236; *inferring main idea*: F (6, 262) = 22.238, p < .001; *Wilk's* Λ = .439, *partial* $\eta 2$ = .337).

Post-hoc tests generally indicated that higher-proficiency learners scored higher than lower-proficiency learners regardless of reading aid for *grammar judgment* and *inferring main idea*. However, in the MT condition for *inferring meaning from context*, there were non-significant differences (p > .017) between high- (M = .769, SD = .206) and mid-proficiency learners (M = .653, SD = .328), and between mid- and low-proficiency learners (M = .522, SD = .258). This indicated that an MT, as a reading aid, can help narrow the proficiency gap between learners, in this case, for inferring meaning from context, even when unfamiliar words were present in the reading passages.

For a more detailed view of the role of reading aids in solving different types of problems, within group analysis was performed after splitting the file for the proficiency groups to conduct one-way repeated measures ANOVA for each problem (IV: Types of reading aid; DV: reading score). Regarding *grammar judgment*, results indicated no significant effect of reading aids for the high- (F (2, 50) = .664, p = .519), mid- (F (2,148) = 2.528, p = .083), and low-proficiency groups (F (2, 60) = .316, p = .730). That is, the reading aids did not affect the reading scores for *grammar judgment* in any of the proficiency groups. The results indicate that neither the MT nor the glossary was effective in helping learners make finer

grammar judgments in any of the proficiency groups.

In comparison, the effects of reading aids were significant on reading scores for *inferring meaning from context* and *inferring main idea* as indicated in Tables 3 and 4. For the former, detailed examination indicated significant differences between reading scores of the mid- (F (2,146) = 10.571, p < .001) and low-proficiency groups (F (1.537, 44.577) = 13.713, p < .001); both groups performed significantly better with the use of MT than in the no aid or glossary conditions (p < .017), but there were no significant differences in reading scores between the no aid and glossary conditions for both mid- (p = .576) and low-proficiency learners (p = 1.00). For inferring meaning from context, the MT seems to have helped learners consider the logical flow of the text and understand the context of the reading passage to infer the missing text whereas the glossary had not been helpful in doing so. Within the high-proficiency group, the effects of reading aids (F (2, 250) = .183, p = .833) were not apparent.

TABLE 3

Inferring Meaning from Context: Effects of Reading Aid on Reading Scores

(N = 130)		*M*	*SD*	*F*(*df*)	*Partial* p^2	Post-Hoc
High	No Aid	.744	.255	.183 (2, 250)	.007	N/A
	MT	.769	.206			
	Glossary	.731	.267			
Mid	No Aid	.495	.323	10.571*** (2, 146)	.126	1 < 2***
	MT	.653	.328			1 = 3
	Glossary	.428	.315			2 > 3***
Low	No Aid	.211	.270	13.713*** (1.537, 44.577)	.321	1 < 2***
	MT	.522	.258			1 = 3
	Glossary	.233	.217			2 > 3***

*p < .017 (Bonferroni correction); High: n = 26, Mid: n = 74, Low: n = 30

For inferring main idea, analysis revealed that only the mid-proficiency learners were impacted by the difference in reading aids (F (2, 152) = 5.928, p < .01). Post-hoc tests indicated a significant difference in reading scores between the MT and no aid conditions (p < .001), and a non-significant difference between the no aid and glossary conditions (p = 1.00). For the high- and low-proficiency group learners, reading aids had no effects (high: F (1.582, 39.550) = .428, p = .654; low: F (2, 64) = .408, p = .667). Overall, the results indicate that when the learners were asked to infer the main idea of a reading passage, the high- and low- proficiency learners did not benefit from the difference in reading aids.

TABLE 4

Inferring Main Idea: Effects of Reading Aid on Reading Comprehension

(*N* = 139)		*M*	*SD*	*F*(*df*)	*Partial* ηp^2	Post-Hoc
High	No Aid	.821	.270	.428 (1.582, 39.550)	.017	N/A
	MT	.833	.216			
	Glossary	.769	.279			
Mid	No Aid	.425	.309	5.928* (2, 152)	.067	1 < 2*
	MT	.580	.303			1 = 3
	Glossary	.470	.282			2 = 3
Low	No Aid	.222	.272	.408 (2, 64)	.013	N/A
	MT	.283	.278			
	Glossary	.253	.261			

*p < .017 (Bonferroni correction); High: n = 26, Mid: n = 78, Low: n = 35

5. DISCUSSION

This study demonstrates that MT can enhance L2 reading abilities, outperforming both no aid and glossary conditions. MT was particularly beneficial for lower proficiency learners (mid and low proficiency), aiding in comprehension at the sentence and discourse levels, though it was less effective for tasks requiring explicit language system knowledge, like grammar. These results are consistent with Chang et al. (2022), who found no significant improvement in accuracy or syntactic complexity with L1 translations. While Cook (2010) emphasizes the role of translations in improving metalinguistic knowledge, the current study suggests their primary utility lies in discerning subtle meaning differences. Resende and Way (2021) support this by showing implicit syntax learning through MT. However, Carré et al. (2022) warn that overreliance on MT might impede active grammar learning. Similarly, the study cautions against overreliance on MT, as it may impede active grammar learning. For mid- to low-proficiency learners, teachers could provide a balanced approach, supplementing MT with explicit grammar instruction to build their foundational language skills, and incorporating MT as a support tool, especially when comprehension questions require contextual understanding.

The study found that higher-proficiency learners consistently outperformed their lower-proficiency counterparts, regardless of the reading aid used. This advantage is attributed to their broader knowledge of L2 vocabulary (Qian & Schedl, 2004), reading strategies (Mokhtari & Sheorey, 2002), and reference search techniques (Chon, 2009). Higher proficiency enables learners to understand more, read faster, and effectively navigate polysemous words (Liou, 2000). Since high-proficiency learners demonstrated less reliance on MT and tend to employ a range of strategies for comprehension, teachers may focus on

encouraging them to further develop their reading strategies and minimize MT dependence. In contrast, learners with mid to low proficiency, constrained by limited L2 knowledge, focused more on understanding individual words rather than integrating the text with their existing knowledge (Graesser, Singer & Trabasso, 1994; Mesgarshahr & Alavi, 2019). The study underscores that the mere availability of MT does not guarantee improved comprehension. The efficacy of MT relies on learners' existing language skills, including their ability to recognize MT errors and to apply effective L1 or L2 reading strategies. For lower proficiency learners, incorporating strategy training sessions to help them navigate between L1 and L2 reading strategies may enhance their comprehension and enable them to utilize MT more effectively.

In particular, mid- and low-proficiency learners were able to utilize the MT to compensate for their lack of L2 knowledge for inferring meaning from context. That is, when reading problems were asked at a sentential level, they could benefit from using MT. In comparison, low-proficiency learners not being able to benefit from the MT for inferring main idea may be an issue related to their L1 comprehension skills. Even with access to MT output (L1 translations), their inability to ascertain the main idea of an L1 text can be attributed to their weak L1 literacy skills. Teachers could design exercises focusing on inferencing and main idea identification in L1, which may translate into stronger L2 reading performance, particularly for learners relying on MT. Another possible explanation for the low proficiency learners' inability to surpass their current proficiency level and minimize the difference in proficiency with the mid proficiency learners can be attributed to how the test instruments provided multiple choice options in L2. Unknown words or concepts may have interfered with their comprehension of the reading passage in all conditions, especially in the no aid condition. The results demonstrate that the content and the format of the reading aid should relate to reader's linguistic ability, rather than attempting to design a universal solution (Chen & Yen, 2013).

The study challenges the traditional view of glossaries as beneficial for vocabulary acquisition and reading comprehension in L2 learning (Leffa, 1992; Lee, Lee, & Lee, 2016; Nation, 2009). It found that glossaries, offering word-by-word definitions, did not significantly enhance reading comprehension compared to no aid. This ineffectiveness may stem from the disruption of reading flow caused by searching for words, and the potential mismatch of definitions with text context due to word polysemy. In contrast, MT provided complete translations, aiding in comprehension and maintaining reading focus, unlike the glossary which interrupted the reading process. The findings may encourage teachers to consider other vocabulary support strategies. For instance, providing glosses within the context of the text or using simplified definitions could reduce the interruption in reading flow. In addition, implementing vocabulary building exercises prior to reading could allow learners to focus on comprehension rather than word lookup during reading.

To summarize, the study suggests that MT, when used appropriately (Ducar & Schocket, 2018), can effectively assist in decoding L2 texts (Hall & Cook, 2012; Pintado Gutiérrez, 2021), reduce anxiety and cognitive load (Bruen & Kelly, 2017), and improve understanding of language structures (Hentschel, 2009). It also indicates that high-proficiency learners rely less on reading aids, likely employing various strategies to overcome linguistic challenges. Given that high-proficiency learners benefit less from reading aids and tend to outperform others, teachers could adopt a differentiated approach where MT is selectively used based on each learner's proficiency level.

6. CONCLUSION AND LIMITATIONS

The study reveals that the effectiveness of machine translations in L2 reading is influenced by learners' ability to interpret these translations, a skill dependent on their language proficiency. This research, one of the few empirical examinations of MT's role in L2 reading, suggests future areas for exploration, including the impact of various conditions (like time constraints), text genres, vocabulary levels, and reading objectives on the utility of MT for L2 learners. It also proposes investigating the most beneficial stages of L2 reading (pre-reading, during reading, post-reading) for MT use.

However, the study has limitations. It did not explore how learners perceive and handle mistranslations in MT, an aspect better suited to qualitative methods like think-aloud interviews. The study used short texts (120 words), raising questions about MT's effectiveness with longer, varied-genre texts for different reading purposes. Additionally, longitudinal studies are needed to assess if MT-assisted reading aids vocabulary acquisition, enhances metalinguistic grammar knowledge, and influences reading strategies, particularly in post-reading activities.

Another significant limitation of the study is that excessive dependence on MT may lead to an over-simplification of cognitive processing for L2 learners. While MT aids in achieving immediate text comprehension, it bypasses many cognitive steps essential for deeper language learning. Processes like perception, noticing, comparison, inference, and memory retrieval are often automated or skipped with MT, limiting the learner's active engagement with the language and potentially affecting their ability to internalize L2 structures or develop long-term language processing skills. MT is effective for comprehension, especially at sentence and discourse levels; however, since it may not actively contribute to grammar or vocabulary acquisition, L2 teachers will need to include activities that foster deeper cognitive processing for fluent and accurate use of L2.

Applicable levels: secondary

REFERENCES

Bruen, J., & Kelly, N. (2017). Using a shared L1 to reduce cognitive overload and anxiety levels in the L2 classroom. *The Language Learning Journal*, *45*(3), 368–381.

Cancino, M., & Panes, J. (2021). The impact of Google Translate on L2 writing quality measures: Evidence from Chilean EFL high school learners. *System*, *98*, 102464.

Carré, A., Kenny, D., Rossi, C., Sánchez-Gijón, P., & Torres-Hostench, O. (2022). Machine translation for language learners. In D. Kenny (Ed.), *Machine translation for everyone: Empowering users in the age of artificial intelligence* (pp. 187–207). Berlin, Germany: Language Science Press.

Carreres, Á. (2014). Translation as a means and as an end: Reassessing the divide. *The Interpreter and Translator Trainer*, *8*(1), 123–135.

Case, M. (2015). Machine translation and the disruption of foreign language learning activities. *eLearning Papers*, *45*, 4–16.

Castilho, S., & Guerberof Arenas, A. (2018). Reading comprehension of machine translation output: What makes for a better read? In J. A. Pérez-Ortiz, F. Sánchez-Martínez, M. Esplà-Gomis, M. Popović, C. Rico, A. Martins, J. Van den Bogaert, & M. L. Forcada (Eds.), *Proceedings of the 21st Annual Conference of the European Association for Machine Translation* (pp. 79–88). Alacant, Spain: Universitat d'Alacant.

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

Chen, W. (2020). Using Google Translate in an authentic translation task: The process, refinement efforts, and students' perceptions. *Current Trends in Translation Teaching and Learning E*, *7*, 213–238.

Chen, I., & Yen, J. (2013). Hypertext annotation: Effects of presentation formats and learner proficiency on reading comprehension and vocabulary learning in foreign languages. *Computers & Education*, *63*, 416–423.

Chon, Y. V. (2009). The electronic dictionary for writing: a solution or a problem? *International Journal of Lexicography*, *22*(1), 23–54.

Chon, Y. V., Shin, D., & Kim, G. E. (2021). Comparing L2 learners' writing against parallel machine-translated texts: Raters' assessment, linguistic complexity and errors. *System*, *96*, 102408.

Chung, E. S., & Ahn, S. (2022). The effect of using machine translation on linguistic features in L2 writing across proficiency levels and text genres. *Computer Assisted Language Learning*, *35*(9), 2239–2264.

Cook, G. (2010). *Translation in language teaching*. Oxford, England: Oxford University Press.

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.

Fuji, M., Hatanaka, N., Ito, E., Kamei, S., Kumai, H., Sukehiro, T., … & Isahara, H. (2001). Evaluation method for determining groups of users who find MT "useful." In B. Maegaard (Ed.), *Proceedings of Machine Translation Summit VIII* (pp. 103–108). Santiago de Compostela, Spain: European Association for Machine Translation.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Hall, G., & Cook, G. (2012). Own-language use in language teaching and learning. *Language Teaching*, *45*(3), 271–308.

Hatim, B., & Munday, J. (2019). *Translation: An advanced resource book for students*. Abingdon, England: Routledge.

Hentschel, E. (2009). Translation as an inevitable part of foreign language acquisition. In A. Witte, T. Harden, & A. Ramos de Oliveira Harden (Eds.), *Translation in second language learning and teaching* (pp. 15–30). Oxford, England: Peter Lang.

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.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., & Weinstein, C. (2005). Measuring human readability of machine generated text: Three case studies in speech recognition and machine translation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05)* (Vol. 5, pp. 1009–1012). Philadelphia, PA: IEEE Press.

Karnal, A. R., & Pereira, V. W. (2015). Reading strategies in a L2: A study on machine translation. *The Reading Matrix*, *15*(2), 69–79.

Kim, H.-S. & Cha, Y. (2020). Exploring the use of a machine translator on EFL learners' reading comprehension. *STEM Journal 21*(1), 119–143.

Klimova, B., Pikhart, M., Benites, A. D., 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.

Kol, S., Schcolnik, M., & Spector-Cohen, E. (2018). Google Translate in academic writing courses? *The EuroCALL Review*, *26*(2), 50–57.

Lee, H., Lee, H., & Lee, J. H. (2016). Evaluation of electronic and paper textual glosses on second language vocabulary learning and reading comprehension. *The Asia-Pacific Education Researcher*, *25*, 499–507.

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.

Leffa, V. J. (1992). Reading with an electronic glossary. *Computers & Education*, *19*(3), 285–290.

Liou, H. C. (2000). The electronic bilingual dictionary as a reading aid to EFL learners: Research findings and implications. *Computer Assisted Language Learning*, *13*(4–5), 467–476.

Macken, L., & Ghyselen, I. (2018). Measuring comprehension and perception of neural machine translated texts: A pilot study. *Translating and the Computer 40 (TC40): Proceedings* (pp. 120‑126). Geneva, Switzerland: Editions Tradulex.

Maghsoudi, M., & Mirzaeian, V. (2020). Machine versus human translation outputs: Which one results in better reading comprehension among EFL learners? *Japan Association for Language Teaching Computer Assisted Language Learning Journal (JALT CALL Journal)*, *16*(2), 69–84.

Marzban, A., & Azizi, A. (2011). The role of translation in promoting reading comprehension of Iranian high school students. *Procedia Social and Behavioral Sciences*, *18*, 526–532.

Mesgarshahr, A., & Alavi, S. M. (2019). Understanding L2 reading cognitive processes: The case of the L2 reader's goal. *The Reading Matrix: An International Online Journal*, *19*(1), 206–225.

Mokhtari, K., & Sheorey, R. (2002). Measuring ESL students' awareness of reading strategies. *Journal of Developmental Education*, *25*(3), 2–11.

Moorkens, J. (2018). What to expect from neural machine translation: A practical in-class translation evaluation exercise. *The Interpreter and Translator Trainer, 12*(4), 375–387.

Odo, D.M. (2020). Supporting pre-service English teachers' academic reading and writing with online machine translation. *STEM Journal, 21*(2), 123–143.

Nassaji, H. (2007). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language Learning*, *57*, 79–113.

Nation, I. S. P. (2009). New roles for FL vocabulary? In L. Wei & V. Cook (Eds.), *Contemporary applied linguistics* (Vol. 1, pp. 99–116). London: Continuum.

Niño, A. (2009). Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages. *ReCALL*, *21*(2), 241–258.

Nord, C. (1996). Text type and translation method: An objective approach to translation criticism. *The Translator*, *2*(1), 81–88.

Pfafflin, S. M. (1965). Evaluation of machine translations by reading comprehension tests and subjective judgments. *Mechanical Translation, 8*(2), 2–8.

Pintado Gutiérrez, L. (2021). Translation in language teaching, pedagogical translation, and code-switching: Restructuring the boundaries. *The Language Learning Journal, 49*(2), 219–239.

Qian, D. D., & Schedl, M. (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing*, *21*(1), 28–52.

Resende, N., & Way, A. (2021). Can Google Translate rewire your L2 English processing? *Digital*, *1*(1), 66–85.

Scarton, C., & Specia, L. (2016, May). A reading comprehension corpus for machine translation evaluation. In K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, … S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3652–3658). Portoro, Slovenia: European Language Resources Association (ELRA)

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43.

Shterionov, D., Superbo, R., Nagle, P., Casanellas, L., O'dowd, T., & Way, A. (2018). Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, *32*(3), 217–235.

Sonbul, S., El-Dakhs, D. A. S., & Al-Otaibi, H. (2022). Translation competence and collocation knowledge: Do congruency and word type have an effect on the accuracy of collocations in translation? *The Interpreter and Translator Trainer*, *16*(4), 409–427.

Stapleton, P., & Kin, B. L. K. (2019). Assessing the accuracy and teachers' impressions of Google Translate: A study of primary L2 writers in Hong Kong. *English for Specific Purposes*, *56*, 18–34.

Tomita, M., Shirai, M., Tsutsumi, J., & Matsumura, M. (1993). Evaluation of MT Systems by TOEFL. In M. Tomita, M. Shirai, J. Tsutsumi, & M. Matsumura (Eds.), *Proceedings of the Fifth Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages* (pp. 14–16). Kyoto, Japan

Van Brussel, L., Tezcan, A., & Macken, L. (2018). A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch. In N. Calzolari, K. Choukri, C. Cieri, T. declerck, S. Goggi, K. Hasida, … T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)* (pp. 3799–3804). Miyazaki, Japan: European Language Resources Association (ELRA).

Weydt, H. (2009). Reading books with translations: Greeting over the reading barrier. In A. Witte, T. Harden, & A.R., de Oliveira Harden (Eds.), *Translation in second language learning and teaching* (pp. 291–307). Oxford, England: Peter Lang.

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.

Zhang, D. (2012). Vocabulary and grammar knowledge in second language reading comprehension: A structural equation modeling study. *The Modern Language Journal*, *96*(4), 558–575.

Zojer, H. (2009). The methodological potential of translation in second language acquisition: Re-evaluating translation as a teaching tool. In A. Witte, T. Harden, & A.R., de Oliveira Harden (Eds.), *Translation in second language learning and teaching* (pp. 31–51). Oxford, England: Peter Lang.

APPENDIX

Configuration of Reading Aids, Item Types, and Item Numbers

Groups	Types of reading condition (Reading Test Battery No.)	Item Type	Item No.
Group A (n = 51)	No aid (Reading No. 1)	Grammar	1-3
		Vocabulary	4-6
		Main idea	7-9
	Machine Translation (Reading No. 2)	Vocabulary	10-12
		Main idea	13-15
		Grammar	16-18
	Glossary (Reading No. 3)	Main idea	19-21
		Grammar	22-24
		Vocabulary	25-27
Group B (n = 43)	Machine Translation (Reading No. 3)	Main idea	1-3
		Grammar	4-6
		Vocabulary	7-9
	Glossary (Reading No. 1)	Grammar	10-12
		Vocabulary	13-15
		Main idea	16-18
	No aid (Reading No. 2)	Vocabulary	19-21
		Main idea	22-24
		Grammar	25-27
Group C (n = 51)	Glossary (Reading No. 2)	Vocabulary	1-3
		Main idea	4-6
		Grammar	7-9
	No aid (Reading No. 3)	Main idea	10-12
		Grammar	13-15

		Vocabulary	16-18
	Machine Translation (Reading No. 1)	Grammar	19-21
		Vocabulary	22-24
		Main idea	25-27