Journal of Modern Languages
Vol. 36(1) 2026
https://doi.org/10.22452/jml.vol36no1.8

An Exploratory Case Study on Translating Performance: ChatGPT vs. NMT vs. Human Translators

Kayo Tsuji*
tsujikayo@omu.ac.jp
Faculty of Liberal Arts, Sciences and Global Education
Osaka Metropolitan University
Osaka 599-8531, Japan
https://orcid.org/0009-0005-3956-9602

Benjamin Neil Smith
bsmith@kwansei.ac.jp
School of Economics, Kwansei Gakuin University
Hyogo 662-8501, Japan
https://orcid.org/0009-0003-6920-7704

Hideki Oshima
hioshima@edu.shiga-u.ac.jp
Faculty of Education, Shiga University
Shiga 520-0862, Japan
https://orcid.org/0009-0007-8604-0722

Received: 27 March 2025 Accepted: 20 November 2025

Abstract

This exploratory case study compares the translation output of a Japanese-language academic paper to English from four sources: the original author, Large Language Model (LLM, namely ChatGPT), Neural Machine Translation (NMT) and a professional, focusing on the differences in the output of ChatGPT and NMT. The recent development of these systems as a means of generating human-like translations has received increased

*Corresponding author

scholarly attention comparing their strengths and weaknesses. However, there is a lack of research with respect to their performance with the Japanese-English language pair in the context of academic research publications. Two researchers in the field of Applied Linguistics were provided four samples to quantitatively rate in the categories of Accuracy, Fluency, Terminology and Style, and gave justifications for each score. The results showed that the author's own translation rated highest, followed by NMT, then ChatGPT and, finally, the professional translation. Raters felt the author's translation captured the intent behind the original text whilst the latter samples took a direct approach that more strictly adhered to the structure and literal meaning of the original text, which hurt their style and terminology but gave them relatively high accuracy scores. Moreover, raters considered the ChatGPT and NMT samples ill-fitting for the intended context: an academic publication. Regarding ChatGPT and NMT, specifically, both raters considered ChatGPT to provide a stricter, word-by-word translation of the text than that generated by NMT. However, given the influence of prompt design in ChatGPT's output, it may be possible to provide a more successful translation with a revised input.

Keywords: Academic Text, ChatGPT, Japanese-English Translation, Human Translation, Large Language Models, Neural Machine Translation

1. Introduction

The advent of Neural Machine Translation (NMT) in the 2010s marked a significant evolution in translation technology, gradually superseding Statistical Machine Translation[1] (SMT) and Rule-based Machine Translation[2] (RMT). NMT systems, such as DeepL and Google Translate, emulate neural networks akin to human cognitive processes, leading to rapid advancements (Goto, 2017). Although NMT has greatly enhanced translation fluency compared to its predecessors (Nakazawa, 2017), issues persist. Problems include omissions or additions (Castilho et al., 2017), untranslated segments (Castilho et al., 2017; Goto & Tanaka, 2018; Nakazawa, 2017; Wu et al., 2016),

[1] SMT operates by segmenting the input sentence into smaller components, translating each segment, and aggregating these translations to form the complete output (Nakazawa, 2017).
[2] RMT employed grammatical rules and lexical dictionaries for translation, but its efficacy was constrained by insufficient computational resources and data (Son & Kim, 2023).

mistranslations (Castilho et al., 2017; Van Brussel et al., 2018), and duplicated sentences (Ehara, 2018; Nakazawa, 2017; Tsuji, 2024).

Concurrently, Large Language Models (LLMs) have undergone rapid development. LLMs, defined as models built with extensive datasets and deep learning, demonstrate human-like fluency and perform natural language tasks with high precision (Doi et al., 2024). Notable LLMs include BERT and ChatGPT, introduced in 2018 and 2022, respectively. BERT, developed by Google, excels in text comprehension (Lund & Wang, 2023) but is not publicly accessible. In contrast, ChatGPT version 3.5 is freely available (Sakaguchi, 2023), and by January 2023, it had amassed 100 million active users (Hulman et al., 2023). ChatGPT has gained attention for accuracy, performing complex tasks and generating high-quality texts (Ikeda, 2023; Lund & Wang, 2023).

ChatGPT's functionalities encompass grammar correction, summarization, memo writing, and translation (Kaya, 2024). Despite not being designed for translation, LLMs like ChatGPT are increasingly used for this (Araújo & Aguiar, 2023; Lee, 2024). Research on ChatGPT's translations includes assessments of accuracy (Ghosh & Caliskan, 2023; Hendy et al., 2023; Jiao et al., 2023) and integration with MT tools for academic papers (Yanase, 2023). Findings indicate that ChatGPT excels in generating contextually relevant translations (Hendy et al., 2023) and performs comparably to Google Translate for high-resource languages, though it is lagging for low-resource languages (Jiao et al., 2023). Yanase (2023) suggests combining machine translation with ChatGPT's stylistic enhancements for academic writing.

Nevertheless, critiques of ChatGPT's translation abilities highlight its limitations in user-friendliness and advanced knowledge, particularly for low-resource languages (Hendy et al., 2023; Peng et al., 2023; Yamada, 2023). The quality of translations can vary with the prompts used (He, 2024) and comparative studies with leading NMT systems are still inconclusive (Hendy et al., 2023; Widiatmika et al., 2023). This study investigates how ChatGPT's translation performance compares to DeepL, a representative NMT system.

2. Previous Studies

Previous research has extensively examined ChatGPT's diverse functions, particularly evaluating ChatGPT as a writing tool by comparing its English essays with those authored by native speakers (Fujiwara, 2023) and investigating its efficacy in drafting academic papers (Zhai, 2023). Studies have also emphasized prompt engineering in optimizing its effectiveness for various applications,

including translation (Gu, 2023; He, 2024; Yamada, 2023), summarization (Shi et al., 2023; Soni & Wade, 2023), document creation (Giray, 2023), and correction functions (Terashima et al., 2023). In the realm of translation, research has examined ChatGPT's role in producing English papers (Jiao et al., 2023; Yanase, 2023).

2.1 Advantages and disadvantages of using ChatGPT for translation

Research on ChatGPT's use as a translation tool has yielded various insights. Studies have assessed its translation functions (Işım & Balcıoğlu, 2023; Khoshafah, 2023), patent translation (Larroyed, 2023), comparisons with human translations (Cao & Liu, 2024), and its role in translation education (Fan et al., 2023).

Işım and Balcıoğlu (2023) investigated ChatGPT's translations from Turkish to English by analyzing 50 paragraphs of the educational literature. They identified 70 errors in 3,350 words, including 38 grammatical and 42 lexical errors, with 19 significantly altering the meaning. Although most errors were lexical and did not severely impact overall meaning, the study concluded that ChatGPT could be a reliable tool for educational translations, recommending its use alongside other software. Similarly, Khoshafah (2023) evaluated its accuracy in Arabic-English translation across history, literature, media, law, and science. While ChatGPT generally produced accurate translations, it struggled with technical texts, idioms, jokes, and colloquial expressions, lacking the nuanced understanding of a human translator. However, it showed potential for translating shorter documents fluently.

Another study on the use of ChatGPT for translations was by Larroyed (2023) who investigated patent translation, assessing 20 English-to-Portuguese texts using a scale covering mistranslation, accuracy, terminology, language, and style. The study found that ChatGPT performed well in language aspects, particularly syntax and sentence structure, comparable to an experienced human translator. Conversely, Cao and Liu (2024) conducted a comparative analysis of Chinese-English political discourse with results indicating that ChatGPT had difficulties with latent elements, such as agents and tense, but could assist human translators effectively. The study underscored the importance of prompt design on ChatGPT's translation output. Fan et al. (2023) explored ChatGPT's role in translation education, noting that while it could act as a "human-like teacher" (p. 51), it struggled with emotionally charged texts, like poetry. They emphasized that traditional translation education principles and human teachers would remain essential.

In summary, ChatGPT exhibits several advantages, including fluency in short documents (Khoshafah, 2023), high performance in language use (Larroyed, 2023), and its supportive potential (Cao & Liu, 2024). However, it also has notable drawbacks, such as issues with lexical items (Işım & Balcıoğlu, 2023), limitations in translating complex texts (Khoshafah, 2023), and difficulties with idioms, jokes, and contextual nuances (Cao & Liu, 2024; Khoshafah, 2023). As a general-purpose NLP model, ChatGPT is not specifically engineered for translation, highlighting the need for comparative studies with NMT systems to fully assess its quality as a translation tool (Larroyed, 2023).

2.2 ChatGPT vs. NMT

Recent studies have compared the validity and fluency of NMT systems and ChatGPT across various criteria. Widiatmika et al. (2023), for example, assessed ChatGPT's translation quality from English to Indonesian. Their findings indicated that while DeepL and Google Translate demonstrated high translation effectiveness, ChatGPT excelled in capturing the essence of the original text. The study analyzed translations of English-language linguistics textbooks produced by Google Translate, DeepL, and ChatGPT. Linguistic evaluations revealed that ChatGPT outperformed both Google Translate and DeepL in terms of adequacy. In a subsequent fluency assessment by five linguistics majors, Google Translate effectively ensured intelligibility, whereas DeepL exhibited issues with sentence structure and terminology, affecting the accuracy of academic texts. ChatGPT's translations were noted for their naturalness and overall quality, leading to the conclusion that ChatGPT provided superior translations in this context.

Another study found that while ChatGPT exhibited style issues, these could be mitigated by providing appropriate context, whereas NMT systems face more persistent accuracy challenges without recrafting the source text (Jiang & Zhang, 2024). In their study, Jiang and Zhang (2024) compared the translation capabilities of major NMT systems (Microsoft Translator, Google Translate, DeepL) with ChatGPT, using three tailored prompts for Chinese-to-English translation of 6,878 remarks from press conferences. Automated evaluations using BLEU [3] (Bilingual Evaluation Understudy) and Chrf[4] (Character n-gram F-score) metrics indicated that ChatGPT scored lower than the NMT systems, suggesting challenges in accurately reflecting reference

[3] BLEU scores MT output by n-gram overlap with references, penalizing overly short translations. (Jiang & Zhang, 2024).
[4] "ChrF computes the F-score of character n-gram overlap instead" (Jiang & Zhang, 2024, p. 4).

phrasing. However, evaluations with BERT and COMET[5], which assess semantic similarity and fluency, showed that ChatGPT had a strong semantic affinity with the references. Human evaluations further highlighted that contextual information significantly improved ChatGPT's translation quality, underscoring the importance of prompt design in optimizing outputs.

Karpinska and Iyyer (2023) explored whether translating entire paragraphs with ChatGPT yields higher-quality translations compared to the traditional sentence-by-sentence approach. Google Translate served as a baseline for this study, which focused on 18 language pairs and utilized literary works as source texts. Human assessors evaluated translations based on mistranslation, grammar, untranslated segments, inconsistency, register, and format. The study found that ChatGPT's paragraph-level translation approach better integrated contextual information compared to sentence-level translation, resulting in fewer mistranslations, grammatical errors, and stylistic inconsistencies than Google Translate.

In terms of language-related performance, Son and Kim (2023) found that GPT-4[6] (Generative Pre-trained Transformer 4) showed superior performance with certain language pairs, such as German-English, suggesting potential improvements in future GPT models. They compared the translation performance of LLM (ChatGPT) and NMT systems (Google Translate, Microsoft Translator) across various content types, including news articles and reports. Their study involved parallel corpora and evaluated 18 language pairs using BLEU, Chrf, and TER (translation edit rate) scores[7]. NMT systems outperformed ChatGPT across all metrics, indicating that ChatGPT is not yet on par with specialized translation systems. Although BLEU, Chrf, and TER metrics provide quantitative assessments of translation quality, they do not capture qualitative aspects such as fluency and cultural appropriateness. Thus, a comprehensive evaluation combining these metrics with human assessment is necessary.

Li (2024), investigating whether LLMs outperform NMT systems in Chinese-to-English translation for Chinese literary and non-literary texts, found that LLMs and NMT systems performed similarly when translating texts of the same genre in the same direction. However, LLMs showed no significant advantage, while DeepL performed better with non-literary texts.

[5] BERTScore assesses cosine similarity between translations and references, while COMET also relies on them but factors in the source sentence. (Jiang & Zhang, 2024).
[6] Multimodal large-scale language model developed by OpenAI.
[7] TER complements BLEU and Chrf by judging translation quality through the edits required to match the reference. (Son & Kim, 2023).

The study compared the quality of Chinese-to-English translations using LLMs (ChatGPT and Wenxin Yiyan, or ERNIE Bot) with NMT (DeepL) across various text genres.

In summary, the comparison between ChatGPT and NMT reveals several areas where ChatGPT outperforms NMT. These include higher ratings in adequacy (Widiatmika et al., 2023), better semantic accuracy (Jiang & Zhang, 2024), and fewer mistranslations, grammatical errors, and stylistic inconsistencies (Karpinska & Iyyer, 2023). Conversely, Jiang and Zhang (2024) indicate that ChatGPT faces difficulties with stylistic inaccuracies. Li (2024) stresses the future potential of ChatGPT and similar LLM systems as machine translators and their capacity to eventually outperform NMTs. However, at present, it struggles to understand domain-specific terminology and cultural context (Khoshafah, 2023).

3. Purpose of the Study

NMT has been demonstrated to produce translations comparable to human translators for certain language pairs (Nakazawa, 2017). However, its performance is stronger with similar language pairs than distant ones, such as Japanese-English (Sun et al., 2021). Its accuracy and fluency may also vary from pair to pair (Yang et al., 2023). Accordingly, the above-mentioned findings cannot be simply extrapolated to Japanese-English translation. Research on the comparison between ChatGPT and NMT in Japanese contexts is relatively unexplored. This study is an exploratory case study examining the translation tool of ChatGPT when translating from Japanese to English. It analyses two raters' perspectives on translations from four sources (published materials, DeepL, ChatGPT, and a professional translator) to further elucidate the benefits and limitations of ChatGPT as a translation tool, considering the following research question:

> What are the differences between translation tools (amongst author, DeepL, ChatGPT, and professional translator) when translating a published academic text from Japanese to English?

In exploring this question, the reliability of ChatGPT as a translator can be assessed alongside its differences and/or similarities with output from the other sources.

4. Methodology

The text analysed was an extract of a published Japanese-written, peer-reviewed work written by the first author of this study: "Developing module-focused scoring rubrics for argumentative

essays" (Tsuji, 2019). The translations were prepared by said author[8], DeepL, ChatGPT and a professional translator. Due to the high cost associated with employing the services of a professional translator, the scope of data collection was necessarily limited to a single text at this time.

The author is a native Japanese speaker with a degree from an English-medium university, fluent in both languages with a Japanese-English translation capability comparable to a professional translator. Their text was proofread by a native English speaker, who made minor edits to articles and punctuation. The text was entered into DeepL by paragraph unit following Yoshida (2020), who found that another NMT, Google Translate, performed better with Japanese-English when inputting text in this way rather than by document. ChatGPT was likewise utilised, in accordance with the recommendation of Karpinska and Iyyer (2023), and was given a prompt adapted from the most highly rated prompt in He (2024)[9] in Japanese. He (2024) assessed translation output based on four different prompts, using keywords from the domain of translation studies. Prompts identifying the role as "translator" performed the strongest and one was adapted for this study. The translation was generated with version 3.5 via OpenAI in March 2024 with the following prompt in Japanese:

> あなたは学術翻訳の専門家です。以下の日本語で書かれた研究論文の一部［原文］を英語に翻訳してください。この翻訳は研究者によって読まれることを想定しています。
>
> [English Translation: You are a professional academic translator. Please translate the extract of the following Japanese-written research paper [original text] into English. This translation will be read by a researcher.]

The professional translation service was provided by a private organisation. The translation involved two individuals who specialize in language education: One, a native English speaker, and the other, a native Japanese speaker. The translation was a collaborative effort to ensure that it was made in accordance with the original Japanese text. It was fully paid for, and task-specific directions were provided. Their first draft was analyzed for the study.

[8] The author of the paper analysed is a native Japanese speaker with a high level of English proficiency. The English translation was published as "Developing and evaluating a scoring rubric for argumentative essays: A module-based approach" (Tsuji, 2021).
[9] He (2024) tested "four prompts in this experiment, including one basic prompt functioning as a baseline for comparison, and three other prompts featuring three keywords in the scholarship of translation studies: translation brief, author, and translator (p. 3)." "Two quality evaluation metrics were adopted in this study (p. 4)" and human evaluation. The translation quality of the prompts with translators was the best.

Two raters were provided with four texts and rating criteria: accuracy, fluency, terminology, and style. These four were the primary criteria in most studies and, details of each criterion are described in Jiang and Zhang (2024). The raters were native-English and native Japanese professors of Applied Linguistics who volunteered to assist with the study. The raters were asked to blind-evaluate the translations based on said criteria and were provided definitions of these terms as follows:

> Accuracy: How closely the text reflects the source text. While evaluating, raters should note additional or missing information, increased or lacking specificity, and overall representation of the content (Bhattacharyya, 2015 cited in Widiatmika et al., 2023; Jiang & Zhang, 2024).
>
> Fluency: Whether a native speaker would accept the text, including word choice, placement and register (Bhattacharyya, 2015 cited in Widiatmika et al., 2023). Raters should also note issues relating to spelling, grammar and syntax (Jiang & Zhang, 2024).
>
> Terminology: How accurately the text used domain-specific words. Raters should note inconsistent usage within the text along with whether a particular term is not one that would be used by an expert in the field (Jiang & Zhang, 2024).
>
> Style: How appropriate and consistent the translation is for its intended context. Raters should note instances of grammatical accuracy but lacking idiomaticity (Jiang & Zhang, 2024).

Raters were asked to score each text in each category from 0-5, with half-points being accepted. They were also asked to provide a written explanation of their scores and overall impressions of each text. Their scores were collected and analysed quantitatively, while their written rationales were analysed by the authors to identify the salient differences in translation quality between the texts.

5. Results

5.1 Ratings of Translation Quality

This section will focus on a quantitative analysis of the raters' evaluation of each writing sample. Table 1 shows the individual ratings assigned to each sample by the two raters for the four scoring

categories: accuracy, fluency, terminology and style. Writing sample A was a translation prepared by the author of the original text, B was created by DeepL, C by ChatGPT, and D, a professional translator.

Table 1. Results of Ratings for Writing Samples

Writing	Accuracy		Fluency		Terminology		Style	
Sample	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
A	4.5	5.0	4.5	5.0	4.5	4.0	4.5	4.0
B	4.5	4.0	4.5	4.0	4.5	3.0	4.5	3.0
C	3.5	3.0	3.5	3.0	3.5	3.0	3.5	2.5
D	2.5	3.0	2.5	3.5	2.5	2.5	3.0	2.0
Correlation	0.82		0.66		0.76		0.88	

Table 2 shows the average rating of each sample within each category based on the above results. The correlation coefficient between the two sets of data was 0.82 for accuracy, 0.66 for fluency, 0.76 for terminology, and 0.88 for style, which indicated a moderate or high correlation between the two raters (see Table 2).

Table 2. Average Score of Writing Samples

Writing	Accuracy	Fluency	Terminology	Style
Sample	Average Score	Average Score	Average Score	Average Score
A	4.75	4.75	4.25	4.25
B	4.25	4.25	3.75	3.75
C	3.25	3.75	3.25	3
D	2.75	3	2.5	2.5
κ	0.00	- 0.07	0.20	- 0.07

Table 2 also displays Cohen's Kappa coefficients (κ) for each category, which tests the concordance of evaluation measurements by the two raters. This revealed a fairly weak negative concordance for the evaluation of fluency and style (κ = - 0.07, p = 0.51 and κ = - 0.07, p = 0.51, respectively) and a fairly weak positive concordance for the evaluation of terminology (κ = 0.20, p = 0.05). For the evaluation of accuracy, κ = 0.00, showing no statistically significant

concordance was observed. The results showed that a statistically significant concordance was found for the evaluation of terminology only.

Finally, in order to aid comparison, Figure 1 shows a visual representation of how the raters viewed each sample within the respective categories.

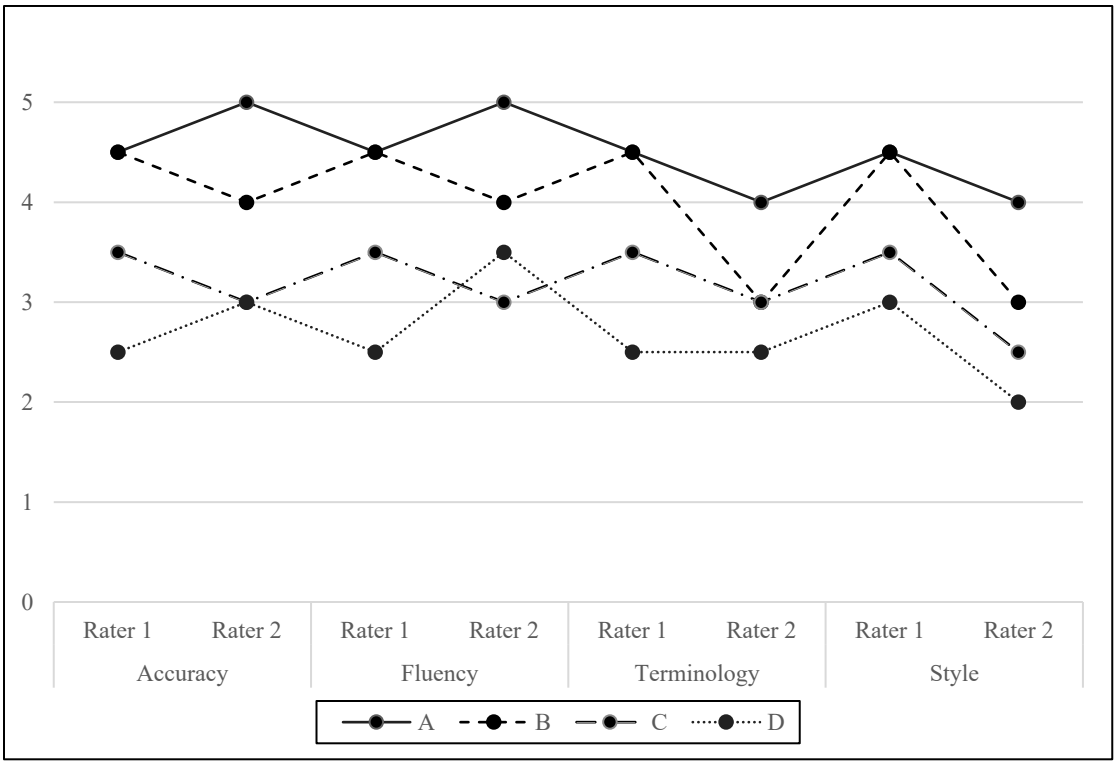

Figure 1. Ratings of Writing Samples

Overall, the raters agreed that sample A, prepared by the original author, was the best translation, with its lowest score being 4.0 and the only sample to receive a score of 5 (albeit from the second rater only). Rater 1 scored this sample 4.5 across all evaluation categories, pointing towards its overall high quality. Sample B, created by DeepL, followed closely behind – although rater 1 scored this equally with sample A, the second rater gave it lower scores for terminology and style. Next, the translation output by ChatGPT was third, with scores ranging from 3.0-3.5, suggesting that it is competent without excelling in any particular area. The lowest rated was sample D, prepared by a professional translator. This was the only sample to score below 3.0, however, raters 1 and 2 disagreed as to its performance within the style category: it received 3.0 from rater 1, its highest score, and 2.0 from rater 2, its lowest score.

Despite disagreements within the ratings, the overall trend and ranking was the same between the two participants. Considering this, and the inherent potential for each rater to ascribe different values to different scores, it is necessary to explore the rationale behind each rater's decisions in order to obtain a better understanding of the strengths and weaknesses of each sample.

Moreover, in the present study, a small number of writing samples were analysed and the concordance of evaluation measurements by the raters did not reach relatively statistically significant levels across the four scoring categories. Indeed, the low Cohen's kappa values indicated high subjectivity in the evaluations of the two raters. To mitigate such rating variation and achieve higher reliability, more raters and a larger writing corpus are required in future work. This may suggest that further research involving a greater number of samples and expert raters would be beneficial in providing more generalizable conclusions.

5.2 Qualitative Analysis of Evaluation Results

5.2.1 *Accuracy*

Accuracy was scored based on how closely the translated text compares with that of the source text, i.e., is it a reflection of it or are their additions, alterations, omissions or errors. The author's translation, sample A, scored highly: 4.5 from rater 1 and 5.0 from rater 2, the difference potentially being explained by a differing interpretation of accuracy between them. Both raters comment that this text seems to take an 'author's intent' approach to translation, rather than offering a direct one-to-one version of the source text. Rater 1 deducted half a point due to certain additions to the text (such as the use of ambiguous, unused terms like 'univocal'; the elongation and elaboration of some concepts; or the reordering of sentences within a paragraph). While these points prevent the sample from being a true and accurate reflection of the source text, rater 2 believed that they instead more accurately captured the author's intent in the target language. Whether one believed that the goal for translation is to produce a direct retelling of the original message or to translate the 'spirit' of the author's message in the target language seemed to be the distinction in the scores here.

In contrast with sample A, the output from DeepL, ChatGPT and the professional (B, C and D, respectively), all adopted a 'direct' approach. All but D receive an overall score above 3, indicating their competency. B was rated highly by both participants due to its consistency in linearly transferring the text from Japanese to English with a lack of errors. The inconsistency in

scores may again be explained by the rater's different interpretations of what an accurate translation ought to be. C suffered from a stricter application of the direct approach, leading to a text rater 1 described as more 'bookish' and lacking nuance, focusing on translation from word to word. Rater 2 noted specific inaccuracies in C, such as the incorrect translation of the name 'Mochizuki' as 'Motomura' or the incorrect use of 'guidance' rather than guidelines, which may account for the overall lower score when compared with rater 1. Finally, with regard to D, rater 1 notes that D suffers from applying the direct approach to such an extent that the words and sentences are sometimes structured in ways that they would be in Japanese rather than in English, for example:

> "*This essay … is substantially and formally similar to the argumentative essay in this paper.*"

Here, it appears that the second referenced 'essay' is contained within the paper rather than being the subject of the paper. While this may function within the source language, further clarification is needed in English, leading to a sentence that deviates from its original meaning. It is only 'accurate' in that it is a complete transplantation of the original text using the vocabulary of the target language. That said, both raters comment that parts of D are highly competently translated, leading to an inconsistent translation with a mixed overall output. This could explain why neither gave below 2.5, as the competent sections offset those which were translated poorly.

5.2.2 *Fluency*

Fluency was scored based on how acceptable the text would be as a product of native written language, and includes elements such as word choice, word order and register. Both raters agreed that the author's translation captured the appropriate register of the chosen context – that of an English-language academic publication. Not only does it convey semantic and communicative properties of the source text well, it also elaborates and makes certain unclear concepts from the original text clearer. While rater 1 assigned the DeepL text the same score for providing an, overall, highly fluent text, it was criticised on this point. A's 'author's intent' approach helps it achieve a certain flow from one idea to the next at both the paragraph, sentence and sub-sentence level. Rater

2 was somewhat more critical of B, commenting that the translation included some poorly structured sentences:

> "*…that learners of English in an English as a Foreign Language (EFL) environment may have…*"

The above underlined section, for example, would be improved by simplifying it (e.g., "…English as a Foreign language (EFL) learners…"). A further example is:

> "*…the perspectives and explanations of the perspectives based on the learners' tasks.*"

Likewise, the repetition in the underlined section hinders readability and it would be improved with simplification (e.g., "perspectives and their explanations"). These issues were deemed relatively minor, with rater 2 still giving DeepL a high score for fluency despite its more direct approach to translation.

ChatGPT, with its greater adherence to the original structure, scored relatively low for fluency compared with B. Rater 1 commented that it displays "insufficient organisation of content connections at the sentence and paragraph level", while rater 2 pointed out certain grammatical inconsistencies (e.g., incongruent plural verb conjugations: "rubrics … is…"). Although the text remained understandable, its stricter adherence to the source text meant that it was regarded as comparable yet of lower quality than the text produced by DeepL.

Both raters agreed, once again, that the professional translator's text was inconsistent; however, this negatively affected rater 1's evaluation more so than rater 2's. Rater 1 criticised its incoherent, fragmented expressions, with poor relationships between ideas:

> "*For the validity survey, IWR (ETS, 2004, as cited in Tsuji, 2021) was used for external evaluation as it is widely accepted in English writing education, was used for external evaluation*."

The repeated underlined section is, firstly, redundant, but also fragmented in its latter placement, as it does not grammatically connect with the prior section. Rater 2 remarked that parts

of the translation felt like a bullet-point version of the source text, rather than an attempt to write an academic paper, since sentences could be far too succinct and simple for their intended context. Still, parts of the text were competently translated, leading to the resulting scores.

Overall, in terms of fluency, the author's translation proved the best performing in terms of presenting a genuine piece of academic literature. The direct approach displayed in the other samples hindered their fluency, with DeepL outperforming ChatGPT due to a less strict adherence to the source text. The professional translation suffered due to some inconsistency and an overly succinct tone which was deemed inappropriate for the target register.

5.2.3 *Terminology*

Terminology was rated based on the consistent and appropriate use of terminology related to the target domain. In this case, the source text was a piece of academic literature within the field of applied linguistics. As above, A scored highly due to a consistent use of terminology and adaptation of the source text for the target domain. Its elaboration of certain concepts compared with the original also improves readability by not only using the terminology correctly but briefly explaining it for a potentially unaware reader. Rater 1 pointed out its use of "task" to both discuss an activity undertaken by the students (participants within the research), as well as the author's activity of testing the rubric as part of their research. This could potentially create confusion, albeit minimal, and could easily be avoided by using distinct terms. It also presents terms not directly used in the source text but which help establish the academic tone – e.g., "univocal", mentioned above, or "milieu".

B, C and D were deemed of equal quality with rater 2, while rater 1 graded them 4.5, 3.5 and 2.5 respectively. Rater 2 comments that they use domain-specific terminology in a largely competent way. However, their interchangeable use of the terms "evaluation" and "assessment" could create confusion within the context of the text: it is important to differentiate between the evaluation/assessment of the students' work and the evaluation/assessment of the rubric in its capacity to do this. Certain sources were also incorrectly translated: for example, while the correct reference, used in A, is "Council on International Educational Exchange [CIEE] Japan", both B and C opt for "TOEFL® Test Japan Office", while D uses "TOEFL® Test Japan Secretariat". The lack of consistency in the use of certain terms and the inaccurate translation of others hurt the score; however, otherwise their use of terminology (e.g., "rubric", "criteria") was acceptable.

Given the importance of accurately referencing and citing sources within academic publications, this is a potentially significant drawback to the use of DeepL, ChatGPT or a professional, and suggests that a degree of domain-specific knowledge will be needed in order to remedy these issues in post-editing. The lack of this specialised knowledge and context amongst these three translators may explain the inconsistencies and inaccuracies when compared with the author's work, as the author necessarily possesses this information.

5.2.4 *Style*

Style ratings were based on the consistency and appropriateness of the text, considering elements such as its idiomaticity and any potential awkwardness. The author's text was highly consistent and maintained an academic tone throughout without being 'bookish' (as rater 1 described the translation by DeepL). Although certain phrases could be improved (for example, "English-medium university" rather than "English speaking university"), this did not significantly impact its scores by both raters. Rater 1 scored DeepL equally with the author for style, while rater 2 scored it a single point lower, with the reasoning being that, despite coming together as a passable text with ideas that connect linearly, they do not do so as smoothly as they could and the awkward phrases mentioned in 5.2.2., above, also had a negative effect on the score assigned for style. ChatGPT, with its stricter direct approach, received lower scores from both raters due to its inappropriacy as a piece of academic literature.

> "*According to Yamauchi (2010, as cited in Tsuji, 2021), in today's era of digitalization, a common means of communication to individuals of different languages and cultures is 'writing.*"

In the underlined example, above, the positioning of 'writing' comes across awkwardly and, potentially, even sarcastically, due to the structure of the sentence. Were the word positioned earlier, this may be remedied. Rater 1 commented that its focus is on translating the text word-by-word, rather than conveying its semantic content, and this is perhaps illustrated by the sentence, above.

Similarly, the professional translation suffers due to its inconsistency and it likely would not be accepted as a piece of academic literature due to being overly succinct and direct:

> "*This study focuses on argumentative essays, which state opinions and contribute to 'cultivating the ability to actively utilize English skills and express ideas independently (Central Education Council, 2014, as cited in Tsuji, 2021),' This concept is the foundation of this article.*"

Not only are there punctuation errors in the above (e.g., including the citation within the quotation marks, forgetting a period following the end of the sentence), but the phrases come off as very direct and matter-of-fact. Compared to the same passage translated by the author, there is a clear difference in quality for its given context (despite the lack of closing quotation marks or page number):

> "*Accordingly, this article focuses on an effective way of improving the writing skills required for an argumentative essay; the target task can contribute to cultivating learners' "ability to actively use English skills and assertively express [their] ideas (Central Council for Education, 2014, as cited in Tsuji, 2021).*"

This could be explained by the varying proficiency of the translators and their knowledge of English language academic publication requirements and style. This also likely effects the output of DeepL and ChatGPT. DeepL currently has no way to instruct it to translate a text for this purpose. While ChatGPT allows detailed prompts to be written, further studies would need to be undertaken in order to consider prompt design and how this impacts particular elements of its translations. In any event, an author with intimate knowledge of English language academic literature would need to consider the appropriacy of a third-party translation for this purpose and potentially conduct extensive post-editing to improve it.

6. Discussion of NMT vs. ChatGPT

These results support the strength of DeepL when compared with ChatGPT. Prior studies found that DeepL provided translations of high fluency (Jiang & Zhang, 2024; Son & Kim, 2023; Widiatmika et al., 2023) and this study was no different. Despite its direct approach to translating the text, it output few and relatively minor errors. This contrasts with ChatGPT, which raters felt

took a stricter word-by-word approach, more closely adhering to the word order and sentence structure of the source text and resulting in a translation the raters found lacking compared to that provided by DeepL, particularly in the Fluency and Style categories. While Karpinska and Iyyer (2023) found that ChatGPT outperformed an NMT (Google Translate, in that study) at paragraph-level translations, it appears that this may not apply to document-level translation tasks. That said, even though the ChatGPT translation analysed in the present study may be lacking, raters believed that it was, overall, still a competent translation of the meaning and content of the source text (Işım & Balcıoğlu, 2023), and both commented on the similarities between the translations produced by DeepL and ChatGPT. Interestingly, Widiatmika et al. (2023) noted the strength of ChatGPT in providing natural texts, which was not the case in this study – raters noted the rigidity of its output to the source structure as a weakness. This difference could be ascribed to the differing contexts, both the source language and the target register (in this case, Japanese and an academic publication, respectively). This perhaps further highlights the influence of prompt design as a potentially powerful factor in its output, as noted by Jiang and Zhang (2024) – if the prompt had been altered to provide the additional context, the output translation of ChatGPT would have differed from that analysed as part of this study. It may be that in order to obtain the most appropriate translation possible; the target register must be specifically and adequately defined within the prompt. This could prove a strong advantage over NMT systems, as there is no way to 'instruct' the NMT on the target context or intended purpose of the output text. In this study ChatGPT was informed of the intended reader (a researcher), but not the intended purpose (a translation fit for peer review and publication). It might be considered that the current output is (largely) successful in providing its reader with the content of the source text, and it is possible that with refinement of the prompt, the output would improve so that it is more suitable to achieve its intended purpose. Future research into prompt design and its effects on the quality of a translation for a particular purpose would offer potentially valuable insights in this area. Moreover, although raters thought the translation would not be accepted for publication in its current form, it may provide a solid base with which a proficient author could prepare a more domain-appropriate text. This could be by editing it to more closely reflect the style and requirements of a publishable academic text, by revising the prompt in order to produce a more satisfactory output, or a combination thereof. The key point, however, is that a relatively high level of L2 proficiency may be required in order to successfully produce a translation acceptable for this purpose. Rather than independently completing the task

for those unable to do so, ChatGPT provides a way to potentially expedite it for those who already can.

7. Conclusion

The purpose of this case study was to compare and analyse the quality of translations of an academic text output by its author, NMT (DeepL), LLM (ChatGPT) and a professional, when translating from Japanese to English, with particular attention paid to the differences between the NMT and LLM texts. Raters scored the author's translation highest in all categories (accuracy, fluency, terminology and style), attributed to them possessing the requisite domain-specific knowledge and being completely informed of the target context. However, this is highly dependent on the individual and their L2 capabilities and may not be of help to less proficient learners who prefer to utilise tools such as DeepL, ChatGPT or a professional translation service. While these results may demonstrate the advantages of developing L2 proficiency over relying on these tools, the raters agreed that DeepL offered a somewhat superior translation to the latter two based on all four evaluation criteria. This is despite both systems taking a similarly direct approach to translating the source text, producing translations that largely mirrored its original sentence structure and word choice. In particular, raters felt ChatGPT suffered from issues relating to accuracy, terminology and style, but scored it comparatively well in fluency. The terminology and style issues displayed by both DeepL and ChatGPT could be ascribed to the lack of context provided to the systems, as they were unaware of the target context (an academic research publication) or the nuance of particular domain-specific terminology, using certain terms interchangeably where an expert would not. Although ChatGPT was informed that a researcher would read the text, it was not told that the text would be, for example, peer reviewed. As such, modifications to the prompt with added context could have resulted in a different output. With regard to accuracy, raters commented that the ChatGPT translation lacked the nuance of that generated by DeepL and also made some fairly egregious mistranslations for this context (such as completely mistranslating the name in a citation). Based on the participants' views, then, DeepL offers a slightly better tool for translating academic research texts than ChatGPT. This ignores a possible strength of ChatGPT, however, in that its output could potentially be refined and improved by providing further information on the intended context of the output text in the prompt. Finally, although the professional translation performed the lowest of all texts evaluated in this

study, this is highly dependent on the service used and cannot be extrapolated to the quality that a different professional would provide.

8. Limitations

This case study is limited by both its context and its scope. It offers insights to the translation quality of these four translators, but only with the Japanese-English language pair and only in the domain of an academic research publication. Furthermore, only two raters analysed translations of an extract from a single paper authored by the lead researcher. The results could, therefore, vary given a different language-pair and a different context, as well as a different text written by an author of differing L2 proficiency. Future research could expand the scope of this study in order to investigate its applicability to Japanese research papers or papers written in a different language, particularly by adding more participants. A further limitation is the lack of context provided to ChatGPT. Future studies could provide valuable insights into using ChatGPT for specific purposes by comparing different prompts or by investigating its ability to improve upon its use of style and terminology in translations after its initial output. Finally, the number of texts available for analysis was limited due to practical difficulties in obtaining multiple professional translations. Future research will seek to refine strategies for data collection in order to expand the dataset both quantitatively and qualitatively in order to offer a more comprehensive and methodologically robust analysis.

Acknowledgment

We would like to express our sincere gratitude to Kiyo Okamoto of Osaka Metropolitan University for valuable assistance as a research assistant.

References

Araújo, S., & Aguiar, M. (2023). Comparing ChatGPT's and human evaluation of scientific texts' translations from English to Portuguese using popular automated translators. In M. Aliannejadi, G. Faggioli, N. Ferro & M. Vlachos (Eds.), *Working notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), September 18−21, 2023, Thessaloniki, Greece* (pp. 2908−2917). https://ceur-ws.org/Vol-3497/paper-243.pdf

Cao, H., & Liu, S. (2024). The effectiveness of ChatGPT in translating chunky construction texts in Chinese political discourse. *Journal of Electrical Systems*, *20*(2), 1684−1698. https://doi.org/10.52783/jes.1616

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, *108*(1), 109–120. http://dx.doi.org/10.1515/pralin-2017-0013

Doi, H., Ishida, H., Nagasawa, D., Tuboi, Y., Kikuchi, R., Ichino, N., Akiyama, H., & Saitou, K. (2024). Performance of generative pretrained transformer on the national licensing examination for medical technologist in Japan. *Japanese Association of Medical Technologists*, *73*(2), 323–331. https://doi.org/10.14932/jamt.23-80

Ehara, T. (2018). Hybrid system of neural machine translation and statistical machine translation. *Japio Year Book 2018*, 300–303. https://japio.or.jp/00yearbook/files/2018book/18_4_08.pdf

Fan, P., Gong, H., & Gong, X. (2023). The application of ChatGPT in translation teaching: Changes, challenges, and responses. *International Journal of Education and Humanities*, *11*(2), 49–52. http://dx.doi.org/10.54097/ijeh.v11i2.13530

Fujiwara, T. (2023). A Comparison between ChatGPT-generated essays and essays written by native speakers of English: An analysis from the perspective of corpus linguistics. *Regional Studies*, *24*(1), 61–73. http://purl.org/coar/version/c_970fb48d4fbd8a85

Ghosh, S., & Caliskan, A. (2023). ChatGPT perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across Bengali and five other low-resource languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, *Canada*, 901–912. https://doi.org/10.1145/3600211.3604672

Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, *51*, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Goto, I. (2017, August). *Nyu-raru nettowa-ku ni yoru kikai honyaku gijyutu* [Neural machine translation technology utilizing neural networks]. *NHK Science & Technology Research Laboratories*, *8*(149). https://www.nhk.or.jp/strl/publica/giken_dayori/149/4.html

Goto, I., & Tanaka, H. (2018). Detecting untranslated content for neural machine translation. *Journal of Natural Language Processing*, *25*(5), 577–597. https://doi.org/10.5715/jnlp.25.577

Gu, W. (2023). *Linguistically informed ChatGPT prompts to enhance Japanese-Chinese machine translation: A case study on attributive clauses*. arXiv:2303.15587. https://doi.org/10.48550/arXiv.2303.15587

He, S. (2024). Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In C. Scarton, C. Prescott, C. Bayliss, C. Oakley, J. Wright, S. Wrigley, X. Song, E. Gow-Smith, R. Bawden, V. M. Sánchez-Cartagena, P. Cadwell, E. Lapshinova-Koltunski, V. Cabarrão, K. Chatzitheodorou, M. Nurminen, D. Kanojia & H. Moniz (Eds.), *Proceedings of the 25th Annual Conference of the European Association for Machine Translation*, *June 24-27, 2024, Sheffield, United Kingdom - Volume 1: Research and implementations & case studies* (pp. 316–326). https://aclanthology.org/2024.eamt-1.27/

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Yong, J. K., Afify M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation*. arXiv:2302.09210. https://doi.org/10.48550/arXiv.2302.09210

Hulman, A., Dollerup, O. L., Mortensen, J. F., Fenech, M. E., Norman, K., Støvring, H., & Hansen, T. K. (2023). ChatGPT-versus human-generated answers to frequently asked questions about diabetes: A Turing test-inspired survey among employees of a Danish diabetes center. *PLoS ONE*, *18*(8), 1–10. https://doi.org/10.1371/journal.pone.0290773

Ikeda, T. (2023). *ChatGPT saikyou no shigotojutsu* [ChatGPT's most powerful working techniques]. Forest Publishing.

Işım Ç., & Balcıoğlu, Y. (2023). ChatGPT: Performance of translate. In M. E. Jones, C. Arenas & E. R. O. Agayev (Eds.), *Proceedings of 3rd International Acharaca Congress on Humanities and Social Sciences*, March 11-13, 2023, Izmir, Turkiye (pp. 47–51). BZT Academy Publishing House. Acharaca Book of Proceedings

Jiang, Z., & Zhang, Z. (2024). *Can ChatGPT rival neural machine translation? A comparative study*. Article 2401.05176v1. https://arXiv.org/abs/2401.05176v1

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. Article 2301.08745v4. https://arXiv.org/abs/2301.08745v4

Karpinska, M., & Iyyer, M. (2023). *Large language models effectively leverage document-level context for literary translation, but critical errors persist*. arXiv:2304.03245. https://doi.org/10.48550/arXiv.2304.03245

Kaya, T. (2024). Generative AI in English education: ChatGPT for enhanced learning. *Bulletin of Gakushuin Women's College*, *26*, 59–77. http://hdl.handle.net/10959/0002002776

Khoshafah, F. (2023). ChatGPT for Arabic-English translation: Evaluating the accuracy. *Research Square*, 1–18. https://doi.org/10.21203/rs.3.rs-2814154/v2

Larroyed, A. (2023). Redefining patent translation: The influence of ChatGPT and the urgency to align patent language regimes in Europe with progress in translation technology. *GRUR International*, *72*(11), 1009–1017. https://doi.org/10.1093/grurint/ikad099

Lee, T. K. (2024). Artificial intelligence and posthumanist translation: ChatGPT versus the translator. *Applied Linguistics Review*. *15*(6), 2351–2372. https://doi.org/10.1515/applirev-2023-0122

Li, X. (2024). Comparison of translation quality between large language models and neural machine translation systems: A case study of Chinese-English language pair. *International Journal of Education and Humanities*, *4*(2), 121–128. http://i-jeh.com/index.php/ijeh/index.

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News*, *40*(3), 26–29. http://dx.doi.org/10.1108/LHTN-01-2023-0009

Nakazawa, T. (2017). New paradigm for machine translation: How the neural machine translation works. *Information & Documentation*, *60*(5), 299–306. http://doi.org/10.1241/johokanri.60.299

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., & Tao, D. (2023). *Towards making the most of ChatGPT for machine translation*. arXiv:2303.13780. https://doi.org/10.48550/arXiv.2303.13780

Sakaguchi, T. (2023). Evaluations of a large language model using university exam questions and discussions for the improvement of its output accuracy. *Journal of Economic Sciences*, *27*(1), 13–24. https://shudo-u.repo.nii.ac.jp/record/2000035/files/KK27102.pdf

Shi, Y., Ren, P., Wang, J., Han, B., ValizadehAslani, T., Agbavor, F., Zhang, Y., Hu, M., Zhao, L., & Liang, H. (2023). Leveraging GPT-4 for food effect summarization to enhance product-specific guidance development via iterative prompting. *Journal of Biomedical Informatics*, *148*, 1–9. https://doi.org/10.1016/j.jbi.2023.104533

Son, J., & Kim, Y. (2023). Translation performance from the user's perspective of large language models and neural machine translation systems. *Information 2023*, *14*(10), 1–18. https://doi.org/10.3390/info14100574

Soni, M., & Wade, V. (2023). *Comparing abstractive summaries generated by ChatGPT to real summaries through blinded reviewers and text classification algorithms*. arXiv:2303.17650. https://doi.org/10.48550/arXiv.2303.17650

Sun, H., Wang, R., Utiyama, M., Marie, B., Chen, K., Sumita, E., & Zhao, T. (2021). Unsupervised neural machine translation for similar and distant language pairs: An empirical study. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *20*(1), 1–17. http://dx.doi.org/10.1145/3418059

Terashima, H., Inada, E., Itai, Y., Sumii, S. (2023). An attempt to utilize ChatGPT feedback in composition. *Japanese Language Education Methods*, *30*(1), 68–69. https://doi.org/10.19022/jlem.30.1_68

Tsuji, K. (2019). Developing module-focused scoring rubrics for argumentative essays. *Journal of Japan Association for College and University Education*, *40*(2), 64–71. https://researchmap.jp/tsuji_kayo/published_papers/26073130/attachment_file.pdf

Tsuji, K. (2021). Developing and evaluating a scoring rubric for argumentative essays: A module-based approach. *Urban Scope*, *12*, 1–13. https://urbanscope.lit.osaka-cu.ac.jp/journal/pdf/vol012/01-tsuji.pdf

Tsuji, K. (2024). Identifying MT errors for higher-quality target language writing. *International Journal of Translation, Interpretation, and Applied Linguistics*, *6*(1), 1–17. http://dx.doi.org/10.4018/IJTIAL.335899

Van Brussel, L., Tezcan, A., & Macken, L. (2018). A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, May 7-12, 2018, Miyazaki, Japan* (pp. 3799–3804). European Language Resources Association.

Widiatmika, P. W., Segara, I. B. M. A., & Kusuma, N. M. Y. W. (2023). Examining the result of machine translation for linguistic textbook from English to Indonesian. *Proceedings of the*

Second English National Seminar, *Pacitan*, 54–65. http://repository.stkippacitan.ac.id/id/eprint/1332

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M, Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. arXiv:1609.08144v2, https://doi.org/10.48550/arXiv.1609.08144

Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. *Proceedings of Machine Translation Summit XIX, 2*, *China*, 195–204. https://doi.org/10.48550/arXiv.2308.01391

Yanase, Y. (2023). Challenges and solutions for Japanese speakers utilizing AI to produce academic papers in English. *Journal of Information Science and Technology Association*, *73* (6), 219–224. https://doi.org/10.18919/jkg.73.6_219

Yang, Y., Liu, R., Qian, X., & Ni, J. (2023). Performance and perception: Machine translation post-editing in Chinese-English news translation by novice translators. *Humanities and Social Sciences Communications*, *10*, 1–8. http://dx.doi.org/10.1057/s41599-023-02285-7

Yoshida. K. (2020). Class practice of physics experiments in English using Google Translate. *Japanese Journal of Applied Physics Education*, *44*(1), 25–28. http://id.ndl.go.jp/bib/030497036

Zhai, X. (2023). ChatGPT user experience: Implications for education. *Social Science Research Network*, 1–18. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4312418