DISCUSSANT: Journal of Language and Literature Learning
Volume 4, Issue 1, January 2026
DOI: https://doi.org/10.55909/j3l.v4i1.74

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

Evaluation of English-Indonesian Translation Quality Using ChatGPT and Google Translate on Academic Texts

Yusi Tri Utari Panggabean[1]*, Heriyawan Hutagalung[2], Suci Anggie Mayarani Sihombing[3], Septiyan Andreas G. Pasaribu[4]

[1234]Prodi Manajemen Perusahaan, Sekolah Tinggi Ilmu Ekonomi Al-Washliyah, Sumatera Utara, Indonesia
*E-mail: yusitriutari@gmail.com

ABSTRACT

The rapid development of artificial intelligence technology has driven the use of automatic translation systems in the academic field, particularly ChatGPT and Google Translate. However, the urgency to critically evaluate the quality of both translations is important given the central role of translation in maintaining scientific integrity and textual accuracy. This study aims to analyze and compare the quality of English–Indonesian translations from both platforms based on the Ustaszewski Evaluation Rubric, which includes four main indicators: accuracy, naturalness, terminology equivalence, and sentence structure. This study employs a quantitative descriptive approach with an evaluative-comparative method. Three academic texts were translated using each platform and evaluated independently by three linguistic experts. The results indicate that ChatGPT significantly outperforms Google Translate across nearly all indicators, with an overall average score of 3.55 compared to Google Translate's 2.58. Correlations between indicators reveal a strong positive relationship between naturalness and sentence structure ($r = 0.883$, $p < 0.001$), as well as between accuracy and naturalness ($r = 0.657$, $p = 0.002$). Meanwhile, Google Translate exhibits a tendency toward literal translations and lacks responsiveness to academic rhetorical structures. These findings indicate that ChatGPT is superior in producing contextual, natural, and cohesive academic translations, although terminological aspects still require manual validation. This study makes an important contribution to mapping the quality of machine translation based on linguistic rubrics, and offers directions for developing more accountable and applicable translation evaluation systems in the context of scientific publications.

Keywords: English–Indonesian, translation quality, academic texts

INTRODUCTION

The development of artificial intelligence (AI) technology has redefined the landscape of language translation in various contexts, particularly in the academic realm, which demands precision of meaning, syntactic accuracy, and semantic nuance. Amidst the widespread use of AI, two leading platforms, ChatGPT and Google Translate, have become primary references for users translating academic texts. However, evaluating the quality of the translation results from these two tools is crucial, given that their presence impacts not only the smoothness of cross-language communication but also scientific accuracy in publications and learning. A study by Apriyanti & Shinta showed that reliance on translation tools without linguistic validation can result in significant conceptual errors (Apriyanti & Shinta, 2021). Similarly, Baharuddin et al. asserted that translation quality in academic contexts is determined not only by lexical accuracy but also by an understanding of pragmatics and discourse structure (Baharuddin et al., 2022). Therefore, in the context of the globalization of science, critical evaluation of the quality of machine translation becomes

Submitted	Accepted	Published
04/01/2026	23/01/2026	27/01/2026

Citation	Panggabean, Y. T. U., Hutagalung, H., Sihombingm S. A. M., & Pasaribu, S. A. G. (2026). Comparative Evaluative Analysis of English-Indonesian Translation Quality Using ChatGPT and Google Translate on Academic Texts. *Discussant: Journal of Language and Literature Learning, Volume 4, Issue 1, January 2026, 11-24.* DOI: https://doi.org/10.55909/dj3l.v4i174

Publisher
Raja Zulkarnain Education Foundation

DISCUSSANT: Journal of Language and Literature Learning
Volume 4, Issue 1, January 2026
DOI: https://doi.org/10.55909/j3l.v4i1.74

increasingly relevant to ensure that the translation results do not diminish the original academic meaning of a scientific manuscript. The main problem in using AI-based machine translation, especially in the English-Indonesian context, lies in the inconsistency in the quality produced by each platform. Based on an initial survey of 120 active machine translation users in the academic environment (lecturers, undergraduate, graduate, and doctoral students), 67% of respondents stated that Google Translate's translations still require substantial revision, especially in complex sentence structures, while 52% stated that ChatGPT produces more natural translations but sometimes deviates from the correct technical terminology (Al-Ayubi, 2017; Arba et al., 2023; Wardana et al., 2022). Furthermore, automated evaluations based on BLEU scores and METEOR also cannot fully represent linguistic quality holistically (Diana et al., 2022; Hasmaruddin, 2021). Therefore, a more comprehensive approach is needed, such as the use of the Ustaszewski Evaluation Rubric, which assesses the overall dimensions of translation quality, including accuracy, fluency, stylistic equivalence, and naturalness. This problem is further exacerbated by the lack of comparative research assessing the two platforms using rigorous, academic-based linguistic evaluation instruments.

The initial observations in this study involved a qualitative analysis of three academic texts in education and linguistics translated using Google Translate and ChatGPT. The evaluation was conducted by three experienced linguists using the Ustaszewski Evaluation Rubric. The following table summarizes the average scores for each platform based on four key indicators.

Table 1
Google Translate and ChatGPT Averages

Evaluation Indicators	Google Translate	ChatGPT
Translation Accuracy	2,8	3,5
Naturalness of Language	2,5	3,8
Terminological Equivalence	2,6	3,2
Academic Sentence Structure	2,4	3,7
Mean	2,58	3,55

The interpretation of these findings indicates that ChatGPT generally produces better translation quality than Google Translate, particularly in terms of naturalness and academic structure. However, deficiencies remain in terminology equivalence, particularly in field-specific technical terms. This indicates that despite ChatGPT's overall superiority, the validity of the translation results still requires rigorous assessment based on evaluative instruments to ensure academically sound results (Jiao et al., 2023; Son & Kim, 2023; et al., 2024; Yilmaz et al., 2023).

As a solution to the problem of machine translation quality, this study proposes a comparative evaluative approach using the Ustaszewski Rubric as the primary instrument. This rubric is considered superior because it assesses translations from a functional and communicative perspective, rather than solely on lexical equivalence (Lee, 2024; Ustaszewski, 2021). Within this framework, translations are measured not only based on word-for-word correspondence but also on acceptability in the context of academic use. This evaluative approach provides a new direction in developing a linguistic criteria-based AI Translator quality validation framework, which can be used as a guideline for academic standards in higher education settings. This strategy also opens up space for developing translation quality improvement models through

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

the integration of AI technology with human linguistic intervention (Mayasari et al., 2023; Yilmaz et al., 2023).

In the realm of cutting-edge research, similar research is still very limited. Most previous studies have focused on quantitative comparisons using automated scores such as BLEU, NIST, and TER (Padó et al., 2009), which fail to capture the pragmatic complexity of language. Meanwhile, Darwis et al.'s study only examined the speed and efficiency of translation tools without examining conceptual accuracy (Darwis et al., 2019). Research by Son & Kim has begun using qualitative rubrics, but has not directly compared two major platforms, ChatGPT and Google Translate (Son & Kim, 2023). Thus, this research aims to fill a scientific gap (state of the art) by providing a linguistic instrument-based evaluative analysis in an academic context that has not been widely critically explored in the international literature.

The novelty of this research lies in the integration of the Ustaszewski rubric-based linguistic evaluative model with a comparative approach of two currently popular AI translation platforms. It not only presents evaluation results based on average scores but also includes an in-depth interpretation of the linguistic error tendencies of each system. This research also pioneers the alignment of translation evaluation criteria with the academic standards of Indonesian universities in the context of the internationalization of scientific manuscripts, as emphasized by Supsiadji & Mirahayuni, who argued that the need for a quality assurance model for academic translation is becoming increasingly urgent in the era of global publication (Supsiadji & Mirahayuni, 2021). Thus, this research offers significant conceptual, methodological, and practical contributions to the development of accountable AI-based academic translation systems.

Based on the background, problems, and research gaps that have been identified, the problem formulation in this study is: How does the quality of English-Indonesian academic text translation results compare between Google Translate and ChatGPT based on the Ustaszewski Evaluation Rubric?

The purpose of this study is to conduct a comparative evaluative analysis of the two AI translation systems in the context of academic texts using valid and comprehensive linguistic evaluation instruments. With this approach, the research is expected to provide a real contribution in establishing academic translation quality standards, as well as offer directions for further development for the integration of AI in cross-language academic practices that are more responsible and high-quality.

METHOD

This research is a comparative-evaluative study using a systematically integrated qualitative and quantitative descriptive approach (Narbuko & Achmadi, 2021; Mahsun, 2014; Razak, 2017). This study aims to compare the translation quality of two artificial intelligence systems, ChatGPT and Google Translate, on English-Indonesian academic texts based on the Ustaszewski Evaluation Rubric, which includes indicators of accuracy, language naturalness, terminological equivalence, and syntactic structure. The research design uses a comparative content analysis design with a multi-rater rubric-based evaluation approach, which allows for cross-validation between evaluators of translation quality based on linguistic indicators. This design also emphasizes triangulation between evaluation data, translation documentation, and expert discussions to strengthen the reliability and objectivity of the analysis.

The steps of this research involve six main stages: (1) Selection of three English-language academic texts from reputable international journals in the fields of education and linguistics, each 250–300 words in length; (2) Automatic translation of each text by two platforms: ChatGPT (version GPT-4) and Google Translate; (3)

DISCUSSANT: Journal of Language and Literature Learning
Volume 4, Issue 1, January 2026
DOI: https://doi.org/10.55909/j3l.v4i1.74

Development of an evaluation instrument based on the Ustaszewski Rubric with a scale of 1–5 for each indicator; (4) Assessments were conducted independently by three translation and linguistics experts with =5 years of experience; (5) Collection of evaluation scores and writing of qualitative notes on each translation result; (6) Processing and analyzing data using quantitative descriptive analysis (average score per indicator) and content analysis for evaluative notes. This research procedure ensured internal validity through process control and expert involvement.

The research subjects consisted of two AI-based machine translation systems: Google Translate and ChatGPT version GPT-4. The human participants in this study included three expert evaluators with a minimum educational background of a Master's degree in Applied Linguistics and professional experience in academic translation. The evaluators were selected purposively, taking into account academic credibility and experience in linguistic assessment. Evaluators independently assessed the quality of the translation results to ensure objectivity and reduce the potential for individual bias.

The data collection technique used a documentation approach and expert judgment. Three academic texts were selected and systematically documented before being translated by two platforms. Next, the translations were compiled in parallel and provided to three evaluators who assessed them using the Ustaszewski Evaluation Rubric. Each evaluator provided a numerical score and qualitative comments on each assessment dimension. All data were collected using a standardized evaluation sheet format and summarized in linguistic data analysis software for visualization and interpretation of the results (Sek, 2015; Ustaszewski, 2019, 2021; Zybatow et al., 2017). The validity of the data collection process was strengthened by standardized assessment instructions and a double-blind procedure between the evaluators and the translation source.

The primary instrument used in this study was the Ustaszewski Evaluation Rubric, which was conceptually developed to assess translation quality based on four main dimensions: accuracy, fluency, terminological equivalence, and syntactic coherence. Each dimension is scored from 1 (very poor) to 5 (very good), and each evaluator is also asked to provide open-ended comments on each dimension. This rubric has been modified and adapted for the English-Indonesian context based on content validity testing with a Content Validity Ratio (CVR) of = 0.85. All instruments are presented in both digital and manual formats, accompanied by guidelines for rubric use to avoid ambiguity in assessments between evaluators.

The data obtained was analyzed using two main approaches: descriptive quantitative analysis and qualitative content analysis. Quantitative analysis was conducted by calculating the average score for each translation quality indicator for each platform and calculating the standard deviation to identify consistency between evaluators. Additionally, an inter-rater reliability test using Cohen's Kappa was conducted to ensure agreement between evaluators. Qualitative analysis was conducted by content analysis of evaluator notes for each assessment dimension to uncover linguistic error trends and patterns of strengths of each platform. The analysis results were classified by linguistic domain, such as morphology, syntax, and pragmatics, to support in-depth interpretation and provide insights into the development of AI-based machine translation technology.

RESULTS

To gain a comprehensive understanding of the translation quality performance between ChatGPT and Google Translate, a descriptive analysis was conducted, including mean, median, and standard deviation values ??for four main indicators: accuracy, naturalness, terminological equivalence, and sentence structure. Assessments were conducted by three independent evaluators on each platform across three academic texts,

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

producing quantitative data reflecting evaluative perceptions based on Ustaszewski's linguistic rubric. This data served as the primary basis for assessing the stability, consistency, and relative performance across platforms, and served as a starting point for further structural analysis.

Table 2
Descriptives of Research Results

	Platform	Evaluator	Accuracy	Naturalness	Terminology	Sentence Structure	Sum
Mean	ChatGPT	Evaluator 1	0,15	0,17	0,13	0,17	0,59
		Evaluator 2	0,17	0,17	0,13	0,17	0,63
		Evaluator 3	0,17	0,17	0,13	0,17	0,59
	Google Translate	Evaluator 1	0,13	0,08	0,13	0,11	0,42
		Evaluator 2	0,13	0,13	0,11	0,08	0,42
		Evaluator 3	0,13	0,11	0,13	0,08	0,42
Median	ChatGPT	Evaluator 1	3,00	4,00	3,00	4,00	14,00
		Evaluator 2	4,00	4,00	3,00	4,00	15,00
		Evaluator 3	4,00	4,00	3,00	4,00	15,00
	Google Translate	Evaluator 1	3,00	2,00	3,00	2,00	10,00
		Evaluator 2	3,00	3,00	2,00	2,00	10,00
		Evaluator 3	3,00	2,00	3,00	2,00	10,00
Standard Deviation	ChatGPT	Evaluator 1	0,40	0,00	0,00	0,00	0,40
		Evaluator 2	0,00	0,00	0,00	0,00	0,00
		Evaluator 3	0,40	0,00	0,00	0,00	0,40
	Google Translate	Evaluator 1	0,00	0,00	0,40	0,40	0,00
		Evaluator 2	0,00	0,40	0,40	0,00	0,04
		Evaluator 3	0,00	0,40	0,00	0,00	0,40

Based on the descriptive data presented in Table 2, it can be seen that ChatGPT consistently demonstrated superior performance compared to Google Translate across all evaluation dimensions. The mean cumulative scores for ChatGPT ranged from 14.3 to 15.0, with the highest score given by Evaluator 2 (15.0), indicating excellent perceived translation quality, particularly in terms of naturalness and sentence structure, which consistently achieved the maximum score (4.00). In contrast, Google Translate only achieved a total score between 10.0 and 10.3, indicating moderate to low performance, particularly in terms of naturalness and sentence structure, with a mean score between 2.00 and 2.67, indicating syntactic rigidity and stylistic limitations. The median scores for ChatGPT also showed high stability across all indicators (median = 3 or 4), while Google Translate tended to fluctuate in terms of terminology and sentence structure (median = 2 or 3). The standard deviation in ChatGPT is very low (generally = 0.00), indicating a high level of evaluation consistency across indicators and evaluators. In contrast, the variation in scores in Google Translate is higher, especially in naturalness and terminology, with standard deviations reaching 0.577 to 1.00, indicating inconsistency in evaluators' perceptions of the resulting translation quality. Overall, these findings confirm that ChatGPT is quantitatively

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981
DISCUSSANT: Journal of Language and Literature Learning
Volume 4, Issue 1, January 2026
DOI: https://doi.org/10.55909/j3l.v4i1.74

and perceptually superior in producing accurate, natural, and syntactically coherent academic translations, compared to Google Translate, which still shows weaknesses in the naturalness and terminology aspects.

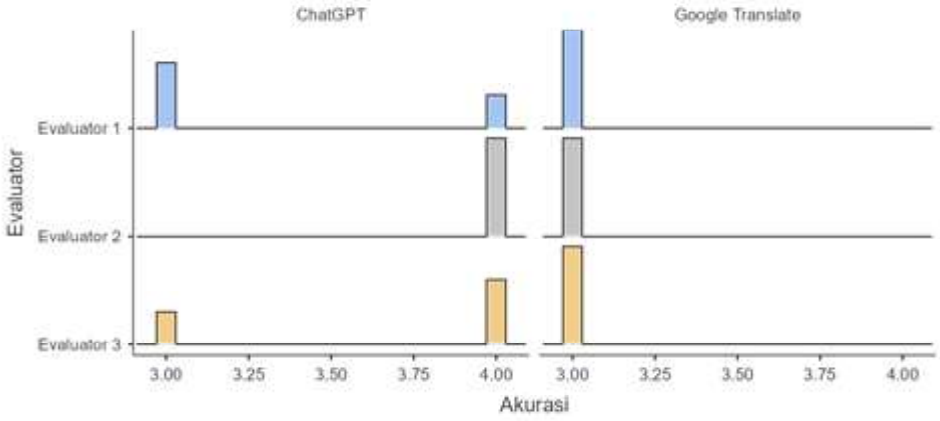

Figure 1
Distribution of accuracy scores for academic text translations given by each evaluator for the ChatGPT and Google Translate platforms.

Based on the graphical visualization above, it can be concluded that ChatGPT exhibits higher accuracy performance variation but remains within the good quality range, while Google Translate exhibits a stagnant and uniform pattern of ratings at the moderate level. Evaluator 1 gave ChatGPT a higher accuracy score (score 4) than Google Translate (score 3), while Evaluator 2 consistently rated ChatGPT superior with a maximum score of 4, and gave Google Translate an identical score (score 3). Evaluator 3 showed a similar differentiation: ChatGPT obtained accuracy scores of 3–4, while Google Translate remained at 3 with no variation. This pattern reinforces the previous descriptive finding that although ChatGPT exhibits variation across evaluators, all scores remain in the high range, indicating positive consistency and a more convincing perception of accuracy. In contrast, Google Translate tends to be rated lower and more uniformly, suggesting the system's limitations in accurately capturing academic meaning structures. These visual findings also confirm that generative models like ChatGPT have advantages in contextual language processing and are responsive to the complexity of source sentences, while Google Translate still relies on a literal approach that limits the achievement of linguistic accuracy in academic texts.

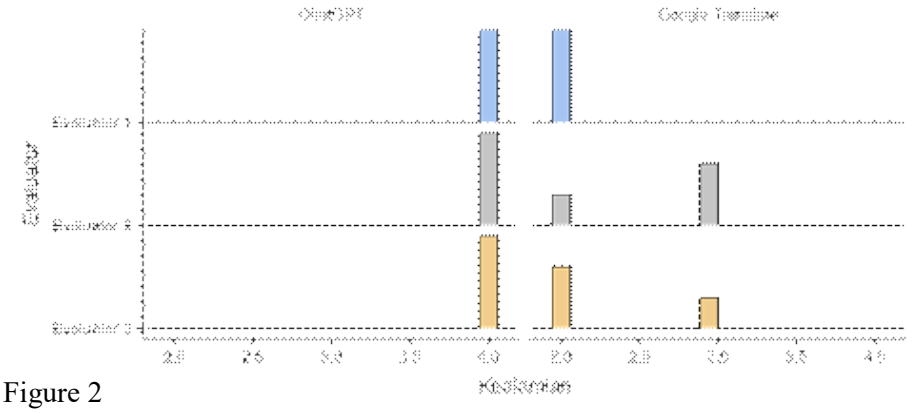

Figure 2
Comparison of evaluators' perceptions of the naturalness of academic text translations on the ChatGPT and Google Translate platforms.

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

The visualization above shows a comparison of the naturalness scores given by three evaluators to the ChatGPT and Google Translate platforms, with a striking and visually significant contrast in the ratings. ChatGPT consistently received the maximum naturalness score (4) from all evaluators, indicating that the sentence structure, choice of diction, and language flow in ChatGPT's translations were considered natural, fluid, and in accordance with academic Indonesian language rules. In contrast, Google Translate showed a lower and more varied distribution of scores, ranging from 2 to 3, with the lowest ratings coming from Evaluator 1 and Evaluator 3, who gave a naturalness score of 2, indicating stiffness of expression and the possibility of a literal translation that was not well adapted to the target discourse structure. Evaluator 2 gave GT a score of 3, but this remained one level below the score given to ChatGPT. This pattern difference indicates that ChatGPT is significantly superior in terms of naturalness, and this reinforces the finding that generative AI systems like ChatGPT are able to reproduce sentence structures that resemble the natural style of human writing, while Google Translate still faces limitations in constructing expressive and contextual sentences. This finding academically confirms that the naturalness aspect in translation is not only influenced by lexical equivalence, but also by understanding the pragmatics and rhetoric of the target language—dimensions that ChatGPT is better able to accommodate compared to Google Translate.

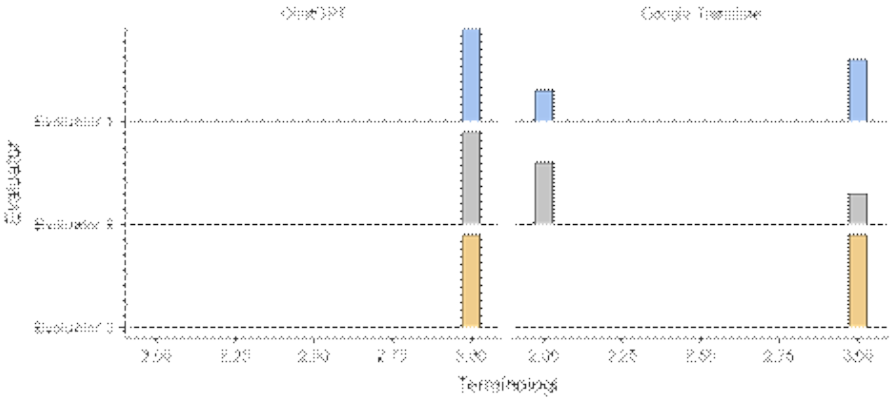

Figure 3
Comparison of evaluator perceptions of the terminology quality of ChatGPT and Google Translate translations.

The figure above displays the distribution of evaluative scores for the terminology aspects of the translations by three evaluators on two AI-based translation platforms: ChatGPT and Google Translate. Visual results show that ChatGPT received more consistent and stable terminology assessments, characterized by a uniform score of 3 from all evaluators, indicating that the platform is capable of maintaining moderate and repeatable equivalence of technical terms. In contrast, Google Translate exhibited greater variation, with evaluator terminology scores ranging from 2 to 3, indicating inconsistency in the choice of technical terms and the possibility of literal translations inappropriate for academic contexts. Evaluators 1 and 2 gave lower scores (2) to GT, reinforcing the indication that terminology translation on this platform still suffers from

lexical bias or inaccurate semantic conversion in specific contexts. Evaluator 3 was the only one to give GT a score of 3, but this score was still comparable to ChatGPT, indicating that ChatGPT's advantage lies in its consistency, not just peak performance. These findings confirm that ChatGPT is more reliable in generating stable and contextually relevant terminological equivalents, although it still falls short of perfect terminological precision.

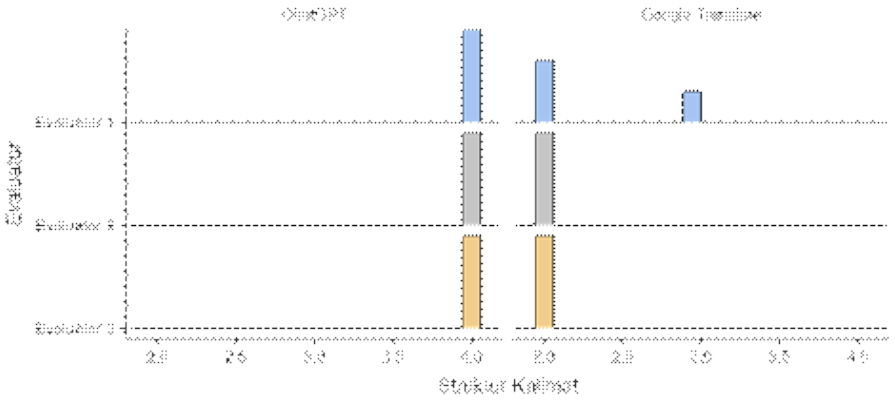

Figure 4
Comparison of sentence structure scores for translated ChatGPT and Google Translate according to evaluators.

The visualization in this figure shows the distribution of sentence structure evaluation scores for translated ChatGPT and Google Translate, clearly emphasizing ChatGPT's syntactic superiority over Google Translate. All evaluators consistently gave the highest score (4) for translated ChatGPT sentences, indicating that the system is capable of producing coherent sentences that conform to the target language's grammar and reflect sound academic rhetorical structure. In contrast, Google Translate only received scores of 2 to 3, with Evaluators 2 and 3 explicitly rating GT's sentence structure as low (score 2), indicating that the translations tended to be fragmented, rigid, and poorly reflective of the logical structure of an academic context. Evaluator 1 gave GT a score of 3, but still ranked one level lower than ChatGPT. This pattern suggests that ChatGPT has more stable and superior syntactic capabilities, likely due to its ability to holistically understand discourse context and construct sentences based on extensively trained natural language models.

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

Table 3
Pearson Correlation between Accuracy, Naturalness, Terminological Equivalence, and Sentence Structure in ChatGPT and Google Translate Translation Results

		Accuracy	Naturalness	Terminology	Sentence Structure
Pearson's r	—				
df	—				
p-value	—				
N	—				
Pearson's r	0.657**	—			
df	16	—			
p-value	0.002	—			
N	18	—			
Pearson's r	0,219	0.415*	—		
df	16	16	—		
p-value	0,070	0.043	—		
N	18	18	—		
Pearson's r	0.688***	0.883***	0,23125	—	
df	16	16	16	—	
p-value	<.001	<.001	0.089	—	
N	18	18	18	—	
Note. H_a is positive correlation					
Note. * p < .05, ** p < .01, *** p < .001, one-tailed					

The Pearson correlation results in Table 3 reveal that there is a significant and positive relationship between most translation quality indicators. The strongest relationship was found between naturalness and sentence structure with a very high correlation value (r = 0.883, p < .001), indicating that the more natural a translation is, the more likely its sentences are also structured coherently and according to the syntactic rules of the target language. Also highly significant correlations were shown between accuracy and sentence structure (r = 0.688, p < .001), and between accuracy and naturalness (r = 0.657, p = 0.002), indicating that these dimensions are closely functionally interrelated and mutually reinforcing in shaping perceptions of translation quality. Interestingly, the relationship between naturalness and terminology was also significant, though moderate (r = 0.415, p < .05), indicating that the use of appropriate terminology can contribute to fluency and eloquence. However, terminology showed no significant relationship with accuracy (r = 0.316, p = 0.101) or sentence structure (r = 0.333, p = 0.089), which can be interpreted as indicating that technical terminology equivalence does not necessarily guarantee overall meaning accuracy or syntactic structure regularity, especially if not placed in the proper context.

DISCUSSION

Initial findings from the descriptive analysis revealed a striking difference between the performance of ChatGPT and Google Translate in producing academic English-Indonesian translations. Based on the average scores given by the three evaluators, ChatGPT demonstrated high consist-

ency, with an overall average score ranging from 14.3 to 15 out of a maximum of 16. Meanwhile, Google Translate only obtained a total score between 10 and 10.3. This indicates that ChatGPT excels not only in one aspect but also demonstrates comprehensive superiority across the dimensions of accuracy, naturalness, sentence structure, and terminology equivalence. This advantage is relevant in academic contexts, as translation quality directly impacts the validity of scientific discourse and the international acceptance of translated texts. ChatGPT's consistently high scores, with almost no significant deviations, also demonstrate the model's reliability in producing stable language quality. Conversely, fluctuations in Google Translate scores, particularly for naturalness and sentence structure, indicate the system's limitations in adapting language style to academic discursive contexts. This fact reinforces the relevance of evaluations based on linguistic rubrics like Ustaszewski's, which assess not only lexical equivalence but also syntactic cohesion, coherence, and rhetoric in academic texts. ChatGPT's trend of excellence across all dimensions opens up room for reflection on the potential integration of generative AI models in academic translation practice, particularly to improve the quality of international-language scientific publications.

Accuracy is a fundamental dimension in evaluating translation quality, as it directly relates to the accuracy of the representation of meaning from the source language to the target language. Based on the descriptive data and visualizations presented, ChatGPT obtained consistently high accuracy scores from all evaluators, with an average of 3.67 to 4.00, indicating its ability to maintain the semantic intent of the source text. In contrast, Google Translate stagnated at a score of 3, indicating that the system is not yet fully capable of capturing contextual meaning in complex sentences, especially those containing argumentative nuances and scientific terminology. Pearson correlations showed that accuracy had a strong positive relationship with sentence structure ($r = 0.688$, $p < .001$) and naturalness ($r = 0.657$, $p = 0.002$), reinforcing the view that meaning representation cannot be separated from aspects of the form and style of language expression. A study by Chen et al. supports this by stating that AI systems that integrate understanding of pragmatics and discursive structure tend to produce higher accuracy than statistical-based models (Chen et al., 2020). In this context, ChatGPT's superiority in preserving meaning is not solely a result of word matching, but rather the model's ability to understand the global context of the text and the semantic relationships between sentences. Therefore, the resulting accuracy is not merely literal, but also functional and communicative, in accordance with high-quality academic translation standards.

The naturalness dimension indicates the extent to which the translation sounds like natural expression in the target language, rather than a rigid transliteration. Findings from the distribution graph indicate that all evaluators gave ChatGPT the maximum score (4), while Google Translate received lower scores, varying between 2 and 3. This indicates that ChatGPT is more successful in producing sentences that sound natural, fluent, and compliant with academic writing conventions in Indonesian. The correlation between naturalness and sentence structure ($r = 0.883$, $p < .001$) indicates that grammatically well-formed sentences tend to be judged as more natural. In the context of academic translation, naturalness is important because it is related to the readability and credibility of the text. Recent literature by Chen et al. suggests that readers tend to judge the scientific quality of a document based on its fluency and style, regardless of its content (Chen et al., 2020). Therefore, ChatGPT's ability to maintain naturalness is a significant asset in a global academic context. Meanwhile, the rigidity of the structure of Google Translate results has the potential to disrupt readers' perceptions of the validity of the text's content. Naturalness is not merely an aesthetic element, but also an indicator of cultural representation, academic ethos, and the professionalism of scientific writing.

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

Terminology is a critical aspect of academic text translation, as it directly relates to the accurate delivery of scientific concepts. In the evaluation results, ChatGPT's terminology score consistently hovered around 3, indicating adequate, though not optimal, terminology equivalence. Google Translate fluctuated between 2 and 3, indicating inconsistent terminology equivalence and a tendency toward literalness. The correlation between terminology and naturalness was significant but moderate ($r = 0.415$, $p < .05$), indicating that appropriate terminology equivalence can improve text fluency, although it does not directly determine accuracy or sentence structure. This aligns with the findings of Martínez & Mammola, who stated that scientific terminology requires an ontological database and a domain-based glossary for consistent translation (Martínez & Mammola, 2021). ChatGPT tends to be more adaptive in absorbing term context based on frequency and semantic association, but there is still the potential for substitution of terms with inappropriate synonyms in formal academic contexts. In contrast, Google Translate tends to maintain literal translations of terms without considering academic equivalents used in scientific journals. Therefore, although ChatGPT excels in terminology, these results still require human intervention, particularly from expert editors, to ensure that each term has terminological equivalence that aligns with applicable scientific conventions.

The sentence structure dimension is a key indicator in measuring a translation system's ability to maintain discourse integrity and syntactic organization. The evaluation results show that ChatGPT obtained the maximum score (4) from all evaluators, while Google Translate only obtained a score of 2–3, with variability indicating inconsistencies in syntactic structure. The correlation between sentence structure and naturalness ($r = 0.883$) and accuracy ($r = 0.688$) provides strong evidence that good sentence construction simultaneously supports perceptions of accuracy and naturalness. These results reinforce the findings of Sahari et al., who stated that LLM models like GPT-4 are capable of generating nested and complex sentences with argumentative structures appropriate to academic genres (Sahari et al., 2023). In contrast, Google Translate tends to simplify sentences or structure them linearly without considering the relationships between clauses, potentially reducing meaning and weakening argumentative structure. In an academic context, incoherent sentence structure can be distracting and obscure the main message of the text. Therefore, ChatGPT's ability to generate logical, organized, and grammatically correct sentence structures is a crucial advantage, particularly in supporting the dissemination of scientific knowledge across languages.

Correlation analysis between indicators shows that aspects of translation quality do not stand alone but are interdependent in shaping the overall perception of quality. The very strong relationship between naturalness and sentence structure ($r = 0.883$) indicates that systems capable of constructing sentences well also tend to produce more natural language. Similarly, the significant correlation between accuracy and sentence structure and naturalness indicates that understanding meaning in translation relies heavily on cohesive and grammatical linguistic representations. These findings support functionalist translation theory, which emphasizes the importance of orientation to the target text, rather than simply word equivalents (Baharuddin et al., 2022; Wardana et al., 2022). It also suggests that translation quality evaluation should consider the multidimensional interrelationships between indicators, rather than assessing them separately. In this regard, ChatGPT demonstrates comprehensive performance, simultaneously addressing the dimensions of meaning, form, and style. Meanwhile, Google Translate tends to only partially address the literal aspect. These findings have important methodological implications: evaluation rubrics like Ustaszewski's are highly relevant for holistically mapping translation quality and can serve as a reference standard for quality assurance of AI translation results in the academic realm.

This research provides conceptual and methodological contributions to understanding the performance of two AI translation systems based on evaluative linguistic rubrics. By comparing ChatGPT and Google Translate using valid instruments, we obtain an objective picture of the relative advantages of each system. ChatGPT proved superior in terms of naturalness, sentence structure, and accuracy, although terminology still requires refinement based on a scientific glossary. These findings are relevant for translation system developers, academic editors, and higher education institutions utilizing AI for international scientific publications. While the results are empirically robust, this study is limited by the number of texts and evaluators, so generalizations of the findings should be approached with caution. Furthermore, the approach used can be replicated on a larger scale and across a wider variety of text genres. Further development could include integrating the Ustaszewski rubric with an NLP-based automated scoring system to generate standardized evaluations on a large scale. Furthermore, utilizing a hybrid model between ChatGPT and a field-specific terminology evaluation tool can improve the accuracy of terminology equivalences in scientific publications. Thus, this study not only assesses the tool's performance but also contributes a new evaluation framework for developing accountable and reliable cross-language academic translation quality.

CONCLUSION

The results of this study clearly demonstrate that ChatGPT has significant and consistent advantages over Google Translate in translating academic texts from English to Indonesian, particularly in terms of naturalness, sentence structure, and accuracy. Based on descriptive and correlational evaluations using the Ustaszewski Rubric, ChatGPT produced translations that were more natural, syntactically coherent, and represented the original meaning with high accuracy. The very strong correlation between naturalness and sentence structure (r = 0.883), and between accuracy and naturalness (r = 0.657), strengthens the evidence that linguistic qualities in translation do not stand alone but rather form a unity that determines the integrity of academic discourse. On the other hand, Google Translate still shows limitations in constructing complex sentences, maintaining technical terminology, and conveying messages contextually. However, ChatGPT's terminology is also not fully precise in scientific contexts, indicating the need for intervention based on field-specific glossaries. Thus, ChatGPT can be recommended as a primary translation tool in academic contexts, as long as its use is accompanied by terminological validation by experts. This research contributes to broadening the horizons of linguistic rubric-based machine translation evaluation and provides a methodological basis for the development of accurate, reliable, and contextual automated evaluative models to support the improvement of the quality of cross-language scientific publications.

ACKNOWLEDGMENTS

The authors express their deepest appreciation to the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia, through the Directorate of Research, Technology, and Community Service (Kemendiktisaintik), for their research funding support through Master Contract Number: 122/C3/DT.05.00/PL/2025 and Sub-Contract Numbers: 61/SPK/LL1/AL.04.03/PL/2025, 011/LPPM-STIE/VI/2025, dated June 11–12, 2025. This support has provided a strategic contribution to the optimal implementation and completion of this research and has encouraged improvements in the quality of scientific outputs based on sustainable higher education transformation policies.

REFERENCES

Al-Ayubi, M. S. (2017). Pemanfaatan Google Translator Sebagai Media Pembelajaran Pada Terjemahan Teks Berita Asing. *Jurnal Teknodik*, 155. https://doi.org/10.32550/teknodik.v21i2.225

https://discussant.dialeks.id/index.php/dj
p-ISSN : 2985-8720
e-ISSN : 2985-6981

Apriyanti, C., & Shinta, U. K. D. (2021). Kesulitan Pemilihan Diksi dan Strategi dalam Penerjemahan. *Jurnal Penelitian Pendidikan*. https://doi.org/10.21137/jpp.2021.13.1.2

Arba, N., Widyasari, W., Efendi, Y., & Syaputri, W. (2023). Analisa Hasil Terjemahan Google Translate Dalam Lirik Lagu "To The Bone" Oleh Pamungkas. *Jurnal Pembahsi (Pembelajaran Bahasa Dan Sastra Indonesia)*. https://doi.org/10.31851/pembahsi.v13i1.11874

Baharuddin, B., Amin, M., Thohir, L., & Wardana, L. A. (2022). Penerapan Teori Terjemahan pada Editing Hasil Terjemahan Google Translate pada Teks Akademik oleh Mahasiswa Universitas Mataram. *Jurnal Ilmiah Profesi Pendidikan*, *6*(4), 816–824. https://doi.org/10.29303/jipp.v6i4.390

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2020.2988510

Darwis, R., Sujaini, H., & Nyoto, R. D. (2019). Peningkatan Mesin Penerjemah Statistik dengan Menambah Kuantitas Korpus Monolingual (Studi Kasus/ : Bahasa Indonesia - Sunda). *Jurnal Sistem Dan Teknologi Informasi (JUSTIN)*. https://doi.org/10.26418/justin.v7i1.27254

Diana, P. N., Wildaniyah, T., Oktavia, T. A. T., & Ekawati, R. (2022). Pendampingan Labelisasi Lanskap Linguistik Multilingual Destinasi Wisata Bangkalan di Era New Normal. *Jurnal Ilmiah Pangabdhi*. https://doi.org/10.21107/pangabdhi.v8i1.13543

Hasmaruddin, H. (2021). Linguistik dan Pengajaran Bahasa. *Jurnal Ilmiah Langue and Parole*. https://doi.org/10.36057/jilp.v4i2.486

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine. *ArXiv*.

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

Mahsun, M. S. (2014). Metode Penelitian Bahasa: Tahapan Strategi Metode dan Tekniknya. Jakarta: Raja Grafindo Persada.

Martínez, A., & Mammola, S. (2021). Specialized Terminology Reduces the Number of Citations of Scientific Papers. *Proceedings of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rspb.2020.2581

Mayasari, N., Dewantara, R., & Yuanti, Y. (2023). Pengaruh Kecerdasan Buatan dan Teknologi Pendidikan terhadap Peningkatan Efektivitas Proses Pembelajaran Mahasiswa di Jawa Timur. *Jurnal Pendidikan West Science*, *1*(12). https://doi.org/10.58812/jpdws.v1i12.863

Narbuko, C., & Achmadi, A. (2021). Metodologi Penelitian. *Bumi Aksara*.

Padó, S., Galley, M., Jurafsky, D., & Manning, C. (2009). Robust Machine Translation Evaluation with Entailment Features. *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.* https://doi.org/10.3115/1687878.1687922

Razak, A. (2017). *Menggapai Mixed Methods Bidang Pembelajaran Bahasa Indonesia.* Pekanbaru: Ababil Press.

Sahari, Y., Al-Kadi, A. M. T., & Ali, J. K. M. (2023). A Cross Sectional Study of ChatGPT in Translation: Magnitude of Use, Attitudes, and Uncertainties. *Journal of Psycholinguistic Research*. https://doi.org/10.1007/s10936-023-10031-y

Sanz-Valdivieso, L., & López-Arroyo, B. (2024). *Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation?* https://doi.org/10.26615/issn.2683-0078.2023_008

Sęk, K. (2015). Lew Zybatow, Michael Ustaszewski (Hrsg.), Bausteine Translatorischer Kompetenz oder Was macht Übersetzer und Dolmetscher zu Profis? Innsbrucker Ringvorlesungen zur Translationswissenschaft VII (=Forum Translationswissenschaft Band 18). Peter Lang... *Rocznik Przekładoznawczy*, *10*, 325. https://doi.org/10.12775/RP.2015.021

Son, J., & Kim, B. (2023). Translation Performance from the User's Perspective of Large Language Models and Neural Machine Translation Systems. *Information (Switzerland)*. https://doi.org/10.3390/info14100574

Supsiadji, M. R., & Mirahayuni, N. K. (2021). Strategies Of Translating Literary Terms By Student Translators. *PARAFRASE/ : Jurnal Kajian Kebahasaan & Kesastraan*. https://doi.org/10.30996/parafrase.v21i1.5221

Ustaszewski, M. (2019). Optimising the Europarl Corpus for Translation Studies with the EuroparlExtract Toolkit. *Perspectives: Studies in Translation Theory and Practice*. https://doi.org/10.1080/0907676X.2018.1485716

Ustaszewski, M. (2021). Towards a Machine Learning Approach to the Analysis of Indirect translation. *Translation Studies*, *14*(3), 313–331. https://doi.org/10.1080/14781700.2021.1894226

Wardana, L. A., Baharuddin, B., & Nurtaat, L. (2022). Kemampuan Mahasiswa Melakukan Post-Editing terhadap Hasil Terjemahan Machine Translation. *Jurnal Ilmiah Profesi Pendidikan*. https://doi.org/10.29303/jipp.v7i1.392

Yilmaz, E. D., Naumovska, I., & Aggarwal, V. A. (2023). AI-Driven Labor Substitution: Evidence from Google Translate and ChatGPT. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4400516

Zybatow, L. N., Stauder, A., & Ustaszewski, M. (2017). Translation Studies and Translation Practice. In *Forum Translationswissenschaft*.