THE ACCURACY OF CHATGPT IN TRANSLATING LINGUISTICS TEXT IN SCIENTIFIC JOURNALS

Rusmita Aeni[1], Baharuddin[2], Lalu Jaswadi Putera[3]
Boniesta Zulandha Melani[4]
English Education Program Language And Arts Education Department Faculty Of Teacher Training And Education University Of Mataram
[1]rusmitaa.aeni@gmail.com, [2]bahar@unram.ac.id, [3]elputra@unram.ac.id
boniestamelani@unram.ac.id

ABSTRACT

AI-generated Machine Translation, such as Neural Machine Translation, has transformed the traditional role of translators. There is a viral language model called ChatGPT. ChatGPT is a conversational variation of Natural Language Processing (NLP) Generative Pretrained Transformer (GPT) models. This research is aimed to analyze the accuracy of ChatGPT in translating linguistics text in scientific journals. The study employs qualitative approach and includes into quality assessment. The data source of this research is a National journal named Ranah focused on language and linguistics, Volume 12 number 1-13 published in 2023. The data then analyzed using few theories. Snover's theory in finding error rate (HTER) and translation quality index to measure translation accuracy by Schiaffino and Zearo. Through the analyzation result, it could be inferred that In the analysis of 2034 words using Snover's theory revealed specific categories of translation errors in ChatGPT with a total of 6% error, there were 15 insertion errors (0.73%), 22 deletion errors (1.32%), 72 substitution errors (3.54%), and 17 shifting errors (0.83%). With a total error of 6,4%, this brings ChatGPT's accuracy rate in translating linguistic scientific texts to 93,6%. Through the analyzation result, it could be inferred that ChatGPT successfully translates scientific text within excellent category.

Keywords: ChatGPT, HTER, Accuracy, Translation.

ABSTRAK

Terjemahan Mesin yang dihasilkan oleh AI, seperti Neural Machine Translation, telah mengubah peran tradisional penerjemah. Ada sebuah model bahasa viral yang disebut ChatGPT. ChatGPT adalah variasi percakapan dari model Natural Language Processing (NLP) Generative Pretrained Transformer (GPT). Penelitian ini bertujuan untuk menganalisis keakuratan ChatGPT dalam menerjemahkan teks linguistik pada jurnal ilmiah. Penelitian ini menggunakan pendekatan kualitatif dan termasuk ke dalam penilaian kualitas. Sumber data dari penelitian ini adalah jurnal nasional bernama Ranah yang berfokus pada bahasa dan linguistik, Volume 12 nomor 1-13 yang diterbitkan pada tahun 2023. Data tersebut kemudian dianalisis dengan menggunakan beberapa teori. Teori Snover dalam menemukan tingkat kesalahan (HTER) dan indeks kualitas terjemahan untuk mengukur keakuratan terjemahan oleh Schiaffino dan Zearo. Melalui hasil analisis, dapat disimpulkan bahwa dalam analisis 2034 kata dengan menggunakan teori Snover, ditemukan kategori kesalahan penerjemahan yang spesifik pada ChatGPT dengan total kesalahan 6%, yaitu 15 kesalahan penyisipan (0,73%), 22 kesalahan penghilangan

(1,32%), 72 kesalahan penggantian (3,54%), dan 17 kesalahan pergeseran (0,83%). Dengan total kesalahan sebesar 6,4%, hal ini membuat tingkat akurasi ChatGPT dalam menerjemahkan teks ilmiah secara linguistik mencapai 93,6%. Dari hasil analisis tersebut, dapat disimpulkan bahwa ChatGPT berhasil menerjemahkan teks ilmiah dengan kategori sangat baik.

Kata kunci: ChatGPT, HTER, Akurasi, Penerjemahan.

A. Introduction

Translation is important. Translation can assist people from various cultures and languages communicate effectively. It has the potential to connect between different languages and cultures. Thus, it is easier for individuals to communicate with one another. A human translation tends to make a fundamental mistake. (Windari & Al Hafizh, 2021) found that the translation students committed errors in translating text. These were lexical, contextual meaning mistakes, grammatical, textual respectively. The study found that the errors came from the inability of the students to comprehend the text. Furthermore, the students also showed the inability to perform correct grammar.

People prefer using Google Translate to human translation service. Fitria (2021) argued that Google Translate is more preferable because its ease of use, quick results, and inexpensive cost. Riadi, Gisella and Angelina (2020) found that Google Translate is unable to translate the texts' context. Google Translate also failed to translate scientific writing. Suryani and Fitria (2022) conducted a study to discover the ability of Google Translate to translate scientific paper. The errors made were comprehensive starting from lexico-semantic to the active-passive form. The results of translation were far from the desired result. There is a free translation service called BING by Microsoft. However, BING performance in translating text is as good as Google Translate (Jufriadi, Asokawati, and Thayyib, 2022). Amidst the unreliability of machine translation, recently, there is a viral language model called ChatGPT. ChatGPT is a conversational variation of Natural Language Processing (NLP) Generative Pretrained Transformer (GPT) models. These language models respond to conversational prompts in a relevant and coherent manner, making them extremely valuable for a variety of NLP applications. The ability of free

ChatGPT to translate still needs more empirical studies. Unfortunately, there is no available empirical studies conducted in Indonesia to date. This calls an immediate response to conduct and publish a scientific paper in this field to fill the research gap existed. Based on the previous study and the explanation above, there is a need to conduct a research about the error analysis on translation product of ChatGPT.

Evaluating the accuracy of ChatGPT requires a diverse set of translation samples and evaluate its output using established translation evaluation metrics. There is a need of conducting a study to assess the ability of ChatGPT in translating Indonesia to English. The study will employ free ChatGPT since it is widely used and free. The study also chooses a scientific linguistic paper since linguistic paper has a precise and technical language and translating it needs consistency in terminology. Furthermore, this is a specific field that many academia that read scientific linguistic paper. They need to know whether ChatGPT provide a reliable machine translation.

In order to assess the machine translation accuracy in translating scientific text, there are several methodologies as well as assessment available. There is an assessment namely HTER or Human Translation Edit Rate. This assessment relies heavily on the degree of editing once the translation machine finished the translation process (Snover, Dorr, Schwartz, Micciulla, and Makhoul, 2006). It measures how many errors in the translation product done by the machine. The HTER correlates higher than any other human assessment for the machine translation.

In simple words, HTER identifies as the minimum edits done by the human to match the source language. There are four possible edits should be done by the human. They are insertion, deletion, substitution, and shift.

There are several studies related to the accuracy of machine translation as well as how to assess them. Sutrisno (2020) conducted a study to assess accuracy of machine translation. The result showed that the machine excessed to the accuracy level of 60, in which was higher than any other Asian languages.

Siu (2023) found that while ChatGPT effectively translates from

English to Chinese, it could benefit from improvements in natural phrasing and term clarification. Khosafah (2023) observed that ChatGPT's Arabic to English translations are generally acceptable for simple texts but fall short in specialized domains and cultural nuances, often lacking the finesse of human translations. Işım (2023) reported a 68% success rate in ChatGPT's Turkish to English translations, with a notable number of grammatical and lexical errors in the sample analyzed. These studies collectively suggest that while ChatGPT shows promise in translation, it still requires enhancements to match the accuracy and cultural sensitivity of human translators.

B. Research Methods

The study employs qualitative approach. Qualitative approach seeks to accomplish a deeper knowledge of a phenomenon (Ary, Jacobs, Irvine, Walker, 2019). The key is to direct the objective of the study into the full picture and grasp a general understanding of a phenomenon. Finally, qualitative study is more flexible instead of rigid in terms of conducting the research. Despite being qualitative research, this study includes quantification when presenting data. There are tables, percentages and calculations in the process of analyzing the data.

Therefore, this study includes into quality assessment. Quality assessment in translation may investigate the process, context, and/or product and could be conducted either descriptive explanatory or evaluative (Saldanha and O'Brien, 2014). The data source of this research is a National journal named Ranah focused on language and linguistics, Volume 12 number 1 published in 2023. It was published by the National Agency for Language Development and Cultivation. To analyse the error rate of the text, the researcher employs HTER calculation. The researcher gives score for every error (insertion, deletion, substitution, and shift) as 1

To create a normalized score, the number of edits is divided by the average length of the total translations Snover (2006). HTER Perfomed using below formula.

$$HTER = \frac{INS + DEL + SUB + SHIFT}{Total\ Words}$$

In this example, translated text makes 5 errors in the average of 50 words reference text. It means that

$$HTER = \frac{5}{50}$$

$$HTER = 0.1$$

To measure translation accuracy of ChatGPT, The result of HTER Calculation and analyzation will then multiplied by 100.

Since HTER does not provide a standard for translation quality, this study uses evaluation criteria based on the Total Quality Index (TQI) by Schiaffino and Zearo (2005) to measure translation quality.

HTER × 100

= 0,1 × 100

Total Words	Total Error	HTER	TQI
2034	126	0,064	93,6

= 10

TQI = 100-10

= 90

The next step is classifying the percentage to the scoring criteria for Translation quality index.

Tabel 1 Translation quality index

TQI Score	Category
0	Negative
1-50	Poor
50-59	Low
60-69	Improvable
70-79	Average
80-89	Good
90- 100	Excellent

Therefore the result of translation accuracy for the above example is excellent.

C. Result dan Discussion

1. Accuracy rate & Error Percentage

In this section, the focus is on revealing the percentage distribution for insertion, deletion, substitution and shifting errors, as outlined by Snover's framework.

Tabel 2 HTER & TQI score

In a total of 2034 words, there were 126 total error found resulting in 0,064 HTER Score and 93,6 TQI. Referring to the TQI categories, the translation

results of ChatGPT in translating scientific linguistic texts are categorized as excellent.

Tabel 2 Error found in each category

NO	Error Category	Total
1	Insertion	15
2	Deletion	26
3	Substitution	72
4	Shift	17
	Total Error	**126**

From 2034 words, there were 126 total error found from each category errors. This included 15 insertion errors, which accounted for 0.73% of the total words; 22 deletion errors, which accounted for 1.32%; 72 substitution errors, which accounted for 3.54%; and 17 shifting errors, which accounted for 0.83%. Total error percentage found in all the samples is 6,4%

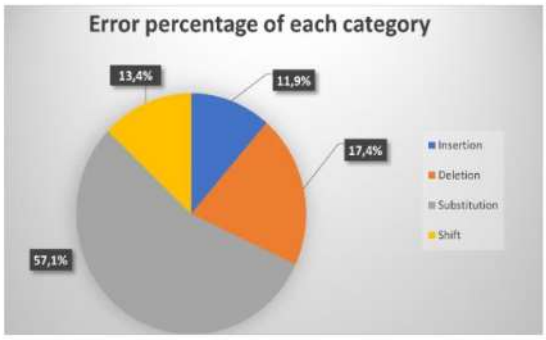

Figure 1 Error percentage of each category

In the presentation error analysis, a total of 126 errors were identified. Substitution errors dominated with an occurrence percentage of approximately 57,1%, indicating that the majority of errors are from the replacement of specific elements, word or phrase in the data. Following this, deletion errors accounted for around 17,4%, signifying a lower frequency compared to substitution errors. Insertion errors comprised approximately 11,9%, suggesting that the addition of supplementary elements in the presentation also contributed significantly to errors. Lastly, shift errors had a percentage of around 13,4%, indicating that misalignment or shifting of elements in the presentation was also a notable factor. This summary provides a detailed overview of the types of errors that prevail in the presentation, along with the relative proportions of each error type in the overall dataset.

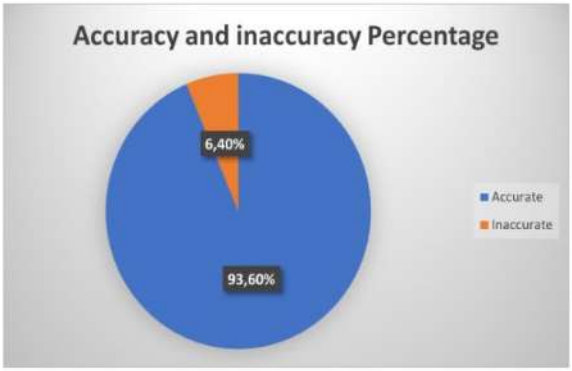

Figure 1 Accuracy and inaccuracy percentage of all the samples

Looking at the data in the table, the accuracy of ChatGPT's translation is the same as Schiaffino and Zearo's Translation Quality Index. This is because accuracy implies the proportion of words that do not contain errors in the translation.

Accuracy is a measure of correctness in this analysis, where higher accuracy indicates a higher proportion of unedited words. With a total error of 6,4%, this brings ChatGPT's accuracy rate in translating linguistic scientific texts to 93,6%.

2. Error Categories Found in the Analysis

a) Insertion

The editing process, which involves adding missing words to sentences that are incomplete due to the absence of important word units or phrases.

ST : *Dari perspektif psikolinguistik* ***remaja****…*

HT : From the perspective of **adolescent** psycholinguistics…

GPT : From a (INS) psycholinguistic perspective…

In translation result from ChatGPT the word "adolescent" is deleted. Theres a need to perform an edit which is insertion and added "adolescents" word.

The use of "adolescent psycholinguistics" accentuates that the discussion pertains specifically to the linguistic aspects intertwined with the psychological experiences of teenagers. If the word "adolescent" is deleted as by ChatGPT from the sentence, it would significantly alter the meaning and context of the statement.

b) Deletion

Deletion is the editing process due to the presence of insertional words or words that are unnecessary and not equivalent to the source text.

SL : *Kreatif dalam menemukan berbagai cara mencaci atau menghina*.

HT : Creative in finding different ways to berate or blaspheme.

GPT : ***This creativity is observed*** in finding various ways to insult or demean.

This sentence should be a direct continuation of the previous sentence. it conveys the creative aspect directly

without the need for the Insertional "observation" phrase. This maintains the focus on the creativity involved in linguistic expression without the extra phrase.

c) Substitution

Substitution is an editing process carried out by replacing incorrect word units in the machine translation with correct or acceptable one.

ST : *Penelitian ini menggunakan metode campuran dan model* ***campuran tidak berimbang****…*

HT : This research uses a mix method and ***concurrent embedded model…***

GPT : This research uses a mixed-method approach and an ***imbalanced mixed model…***

In the context of translation, ChatGPT failed in translating "model campuran tidak berimbang" with "imbalanced mixed model". Because the term that should have been used is "concurrent embedded model". So there is a need to perform substitution in the sentence.

This aligns with Sugiyono (2011) concurrent model combination method, which comprises three designs: Concurrent Triangulation Design (*campuran kuantitatif dan kualitatif secara seimbang*), Concurrent Embedded Design (*campuran tidak berimbang*), dan Concurrent Tranformative Design (*gabungan antara model* triangulation *dan*

embedded). An error in the translation of this research could potentially lead the reader to misinterpret its findings about research and scientific study.

d) Shift

Shift involves a change in form or structure between the source text and the target text.

ST : …*penggunaan maksim prinsip kesantunan pada tuturan langsung para mahasiswa.*

HT : …the use of ***politeness principle maxims*** in students' direct speech.

GPT : …the use of ***maxims in the principle of politeness*** in the direct speech of students.

ChatGPT replaces the term "Politeness principle maxims" with "maxims in the principle of politeness".

In the context of Geoffrey Leech's work, "politeness principle maxims" would be a more accurate and aligned term. It directly connects with his

theory of politeness principles and the specific maxims associated with them.

D. Discussion

There are several studies whose results contradict this study. Recent studies have delved into the translation capabilities of ChatGPT across various language pairs, revealing both its strengths and limitations. Khosafah (2023) focused on Arabic-English translation, finding that while ChatGPT can convey the core meaning, it sometimes falls short of the depth and nuance found in human translations. Işım (2023) explored Turkish-English translation, a challenging pair due to Turkish's agglutinative structure. The study reported a 68% accuracy rate, with ChatGPT translating 34 out of 50 paragraphs without errors.

In contrast with this study, involving Bahasa Indonesia to English translation, ChatGPT achieved a remarkable 93.6% accuracy rate, indicating a high proficiency in handling languages with less morphological complexity and suggesting effective training on Indonesian datasets. These findings highlight the varying performance of ChatGPT in translation tasks, with its effectiveness influenced by the linguistic characteristics of the source and target languages.

E. Conclusion

This thesis has critically examined the accuracy of ChatGPT in translating scientific texts from Bahasa Indonesia to English. The research findings reveal that ChatGPT achieves an impressive accuracy rate of 93,6%, indicating a high level of proficiency in this specific translation task. This performance surpasses the translation accuracy of other language pairs and highlights the potential of AI in facilitating scientific communication across linguistic barriers.

ChatGPT can offer multiple translation suggestions. Users should evaluate these and combine the best elements and prompt to enhance the translation quality.

Researchers and translators should be aware of the intricacies involved in machine translation, such as the challenges with conditional tenses and colloquial phrases. Researchers should continue to analyze translation accuracy and identify areas for improvement.

REFERENCES

Jufriadi, Amalia Asokawati, & Thayyib, M. (2022). The Error Analysis of Google Translate and Bing Translator in Translating Indonesian Folklore. FOSTER: Journal of English Language Teaching, 3(2), 69-79. https://doi.org/10.24256/foster-jelt.v3i2.89 Accessed on May28, 2023

Khoshafah, F. (2023). ChatGPT for Arabic-English Translation: Evaluating the Accuracy. Research Square. DOI: 10.21203/rs.3.rs-2814154/v1.

Işım, Ç., & Balcıoğlu, Y. S. (2023). CHATGPT: PERFORMANCE OF TRANSLATE. In 3rd International ACHARAKA Congress on Humanities and Social Sciences Proceedings Book (pp. 47-51).

Saldanha, G., and O'Brien, S. (2014). Research Methodologies in Translation Studies. Routledge. New York.

Schiaffino, R. and F. Zeano. (2002). The Measurement of Quality in Translation Using Statistical Methods. 43rd ATA Conference 2002. J.D. Edwards and Lionbridge Technologies, Inc. Atlanta: USA. Accessed on: 20 May 2023

Sugiyono. (2011). Metode Penelitian Kuantitatif, Kualitatif, dan Kombinasi (Mixed Methods). Bandung: Alfabeta.

Siu, Sai Cheong. (2023). ChatGPT and GPT-4 for Professional Translators: Exploring the Potential of Large Language Models in Translation. 10.2139/ssrn.4448091.

Snover, M., Dorr, B., Schwartz, R., Makhoul, J., Micciulla, L., Weischedel, R. (2006). A Study of Translation Error Rate With Targeted Human Annotation. LAMP-TR-126.

Suryani, N. Y., and Fitria, T. N. (2022). Error Analysis of Abstract Translation in Scientific Writing by Using Google Translate. JETAL: JOURNAL OF ENGLISH TEACHING & APPLIED LINGUISTICS. Vol. 3, No. 2. Pp. 33-40. https://doi.org/10.36655/jetal.v3i2.669

Windari, M., & Al Hafizh, M. (2021). Error Analysis on Students' Translation News Item Text from English to Indonesia.