Prologue: Journal on Language and Literature
Vol. 10 No. 2 (2024)
https://prologue.sastra.uniba-bpn.ac.id/index.php/jurnal_prologue
ISSN: 2460-464X (Print)
ISSN: 2723-3952 (Online)

DeepL VS. ChatGPT: MACHINE TRANSLATION EVALUATION

Yulia Milka Nugraheni[1], Adi Sutrisno[2]
Gadjah Mada University [1], Gadjah Mada University [2]
yuliamilkanugraheni1999@mail.ugm.ac.id

Abstract

This research aims to evaluate DeepL and ChatGPT performance in translating acaddemic text, through human and machine evaluation. Furthermore, this research is expected to give readers an overview of the translation produced by DeepL and ChatGPT. DeepL and ChatGPT are two machine translation which are using the latet technology in machine translation called as Natural Language Processing (Jiao et al., 2023). The evaluation was conducted by using Koponen (2010) Error Analysis and Papineni (2002) automated machine translation evaluation called Bilingual Language Evaluation Understudy (BLEU). The evaluation was conducted by applying qualitative and quantitative method. Both methods used in order to draw a stronger conclusion. The result of the research concluded that DeepL evaluation showed a better performance than ChatGPT. On Error Analysis Evaluation, there are 25 errors found in DeepL translation and 26 errors in ChatGPT translation. On BLEU score evaluation, the final score of DeepL translation is 0.9446657236 and BLEU score of ChatGPT is 0.9211813372.

***Keywords**: Machine Translation Evaluation, Error Analysis, BLEU, DeepL, ChatGPT*

Article Information

Received:
Sept 02, 2024

Revised:
Sept 04, 2024

Accepted:
Sept 10, 2024

INTRODUCTION

Machine translation evaluation is a very important field of research nowadays. Machine translation (MT) can be a very useful tool for user in translating a text in a vast period. MT itself has been evolving during years. Six years ago, machine translation technology was still using a statistical framework which is called as statistical-based machine translation, which employed corpus of translation examples called "parallel or bilingual corpora" (Habash et al. 2011, p. 133). Today, machine translation has transformed employing a neural translation system which employs the context of sentence, grammar, and sentence structure to create a natural translation (Slatyer and Forget, 2019, p. 445). The rapid development of

MT nowadays has been a great advantage to the users of MT all over the world. Researchers have examined the implication of MT system development to the translation output and concluded that MT has been improving its translation output. Setiajid and Tirtayasa (2019), conducted diachronic research on the performance of *Google Translate* and concluded that its performance has improved significantly. However, the research pointed out that there are still some errors found in the translation output of *Google Translate*. It is shown that eventhough the present MT has given a significant improvement compared to its predecessor, MT continues to show translation error in its output (Lee and Briggs, 2020).

Translation error performed by MT triggers more research regarding the evaluation of MT system nowadays. Research on MT evaluation was conducted by several researchers. Rresearch on MT performance evaluation based applying a combination between manual and automated evaluation by Trigueros (2021). Analysis of performance between two MTs by Yulianto and Supriatnaningsih (2021), Automated MT evaluation development by Comelles and Atserias (2018). Longitudinal study of MT performance conducted by Lotz & Rensburg (2016), Setiajid and Tirtayasa (2019).

This research will examine the performance of two MTs applying error analysis as a manual assessment and applying BLEU metrics as an automated assessment. The two MTs that are analyzed for its performance are Chat GPT and DeepL. DeepL is a machine translation engine launched in 2017. DeepL employs deep machine learning and neural machine tranlsation system. Deepl has proven its performance by creating comparative experiment on translating a text with several other MTs. The experiment was assessed by professional translator and the result showed that DeepL translation was chosen as the best MT in performing the translation (Slatyer and Forget, 2019). The next translation engine, Chat GPT, was chosen because of its increasing popularity among current internet users. Chat GPT is an artificial intelligence that is a conversation engine designed by OpenAI under the instruction of InstructGPT (Ouyang et al., 2022). In use, ChatGPT integrates various natural language processing, including translation (Jiao et al., 2023). Nowadays, ChatGPT has become very popular because of its various features.

This research focuses on comparing the performance between DeepL and ChatGPT. DeepL and ChatGPT will be analyzed in this research since those MTs are relatively new to the machine translation development yet has been popularly used. The researchers seek to provide readers with an overview of the performance of the two MTs to see which MT perform a better translation. The researchers emphasize to evaluate both MTs based on the error analysis theory developed by Koponen (2010) and BLEU score. Machine translation is a massive inovation to translation, it eases the translation process. Moreover, the result produced bt MT nowadays turned out to be more human translation like and is proven reliable by researchers. The result of the research will carry out the machine translation with lower error rate, which can be concluded as a more reliable MT.

LITERATURE REVIEW

Machine Translation (MT) is an automated translation system using computation algorithm. The concept of MT had been carried out since 1950s, and had been made to realization in 1990s. DeepL is a machine translation launched in 2017. Machine Translation nowadays, has been developed to use the latest system called as Neural Machine Translation (NMT). NMT system is built to produce a human-like translation, which employs artificial neural networks. Chat GPT and DeepL are MTs which systems adopt the NMT system. DeepL is a new machine translation which adopts deep learning and neural machine translation system to translate automatically from one language to another (Slatyer and Forget, 2019, p. 447). Established in 2017, DeepL has become one of the best MT these days. Quoting from DeepL's official website (https://www.deepl.com/id/translator) there are some comments coming from companies around the world which compliment the performance of DeepL compared to any other MT. On the other hand, ChatGPT is an AI developed by OpenAI, it is a chatbot which is built to communicate with human using command prompt. The prompt given to the AI can also be a translate command prompt. ChatGPT can handle translation since it is an AI employing Neural Language Processor (NLP). ChatGPT is beneficial for translation purpose, because it capability to produce natural language, which is an esential ideology in translation (Kalla & Smith, 2023).

Machine translation evaluation is a concept of examining an MT in terms of its capability in translating a certain text. The capability here means how accurate can an MT translate a text. Machine translation evaluation is a very important aspect. Machine translation is currently growing rapidly both in terms of translation results and users who are increasingly in need of accurate and fast translation. The development of MT performance can be seen from the aspects of grammar, word selection, to the style of language in the translated text. However, the accuracy of machine translation has not been able to replace human translation which is capable to bring up the context and cultural background in the translation results. The results of machine translation still require evaluation to review the percentage of errors that are still found in the results of machine translation. According to research by Lee & Briggs (2020), eventhough MT has been evolving better than the earlier system, error can still be found in the translation product of MT.

Evaluating an MT can be done manually by applying error analysis to the evaluation process or automatically by using an automated translation evaluation metrics. Manual evaluation can be done by human by analyzing translated text to find error translation, and finally categorizing the error based on a particular error analysis theory. In this research, Koponen's error analysis will be applied to assess the translated text. Koponen (2010) in her journal entitled "Assessing Machine Translation Quality with Error Analysis" introduces the errors categories and the way the tested texts are selected. She divides the errors categories into two main classes, which are individual concept error along with the relation between concepts error (2010, pp. 4-5).

Individual Concept Error

Individual concept error is the first category of which is based on the accuracy of the relation between the individual concept in the source and target language. Individual concepts are concepts in target and source language which are represented through content words. Individual concept error is divided into 6 sub-categories: omitted concept, added concept, untranslated concept, mistranslated concept, substituted concept, and explicitated concept.

Relation between Concepts Error

The second category is the relation between concepts. Relation between concepts is represented through function words, inflection, and word order. This category is divided into 8 categories: omitted participant, omitted relation, added participant, added relation, mistaken participant, mistaken relation, substituted participant, substituted relation.

Evaluation on machine translation has been conducted by several researchers. The evaluation itself has been developing into an automated machine translation evaluation. A well-known automated machine translation evaluation was established by Papineni in 2002, and the automated evaluation itself is called as Bilingual Language Evaluation Understudy (BLEU). The development of BLEU was initiated due to the expensive expense on human evaluation (Hovy, 1999). BLEU is an automated machine translation evaluation tool which assists translator in examining translation result of an MT. According to Papineni (2002), BLEU combines a numerical "translation closeness" metrics and a corpus of good quality human reference translations. Eventhough, we call BLEU as an automated machine translation evaluation tool, BLEU still requires the knowledge of human translator. Papineni (2002) developed BLEU metric employing human translation reference to be considered as the correct translation to be compared to the machine translation to validate the evaluation result. The automation of the evaluation was established by simply counts the number of candidate translation words (unigrams) which occur in any reference translation and then divides by the total number of words in the candidate translation (Papineni, 2002).

RESEARCH METHODOLOGY

This research analyzes machine translation on academic text types. This research is expected to help users of machine translation in determining which translation engine can help the translation process. The researchers attempted to assess the performance of machine translation and provide an evaluation of machine translation performance, which in turn can serve as an indicator of which translation machine is better. This research is mixed research between qualitative and quantitative research. In addition, this research is longitudinal research using library and explicatory methods. Qualitative research is research that aims to describe the quality of something in an enlightening way (Williams & Chesterman, 2002, p. 64). To get sharper results, researchers also use quantitative research. Quantitative research according to William and Chesterman (2002) is to conclude something related to the generality, tendency, frequency, and distribution of a certain phenomenon.

The object of the study was an academic text which was translated from English to Indonesian using DeepL and ChatGPT. The academic text used were abstracts from some journal articles. The researchers collected the data from Humaniora, the journal of the faculty of cultural science of Gadjah Mada University. The researchers took three abstracts from each of the three volume of the journal which are Volume 33 No.1, Volume 34 No. 1, Volume 34 No. 2, Volume 35 No. 1. The 3 selected abstracts from each Volume were the abstracts which have the most viewers. Moreover, the researchers translated each abstract from English to Indonesian using DeepL and ChatGPT. The translated texts are then divided into sentences in the form of table to be coded.

FINDINGS AND DISCUSSIONS

DeepL Error Analysis Result

In the error analysis of machine translation DeepL, the researchers found a total of 24 errors out of a total of 105 sentences taken from 12 abstract texts. Thirty errors were derived from several machine translation error categories proposed by Koponen (2010). Based on the error categories, there are 6 concept omissions, 4 mis-translated concepts, 3 untranslated concepts, 8 concept substitutions, 1 element substitution, 1 element error, and 1 relationship error. The errors found are explained on the discussion part.

ChatGPT Error Analysis Result

In the analysis of ChatGPT translation errors, a total of 24 errors were found. The errors found are respectively divided into 2 concept omissions, 5 concept additions, 4 mistranslated concepts, 3 untranslated concepts, 5 concept substitutions, 1 element addition, 4 relationship errors, and 1 element substitution. Further explanation of the examples of errors found will be given in the explanation in the subsections below.

BLEU Score Calculation

BLEU score analysis is performed using a calculation algorithm developed by Natural Language Toolkit with the Python programming language. BLEU score calculation is done using Google Collab which is a cloud-based platform for writing, running, and sharing Python code through a web browser. The calculation of BLEU score was conducted on 105 translated data on 2 translation engines with analysis per sentence as the object unit of score calculation. The resulting BLEU score indicates the level of translation accuracy performed using machine translation. The resulting BLEU score range starts from 0 to 1. The score results that are close to 1 indicate that the translation results carried out by the machine have a higher level of accuracy, close to natural human translation. Of the 105 sentences analyzed, 19 sentences were found with scores below 1.

The BLEU score resulting error both on DeepL and ChatGPT is presented in the following tables:

Table 1. *DeepL* BLEU Score (<1)

No	Data Code	BLEU Score (NLTK)
1	2/ST/A1/2	0.5882432467
2	4/ST/A1/4	0.8968833071
3	17/ST/A2/8	0.5418220426
4	23/ST/A3/6	0.8801117368
5	26/ST/A3/9	0.7016035864
6	29/ST/A3/12	0.5384952356
7	31/ST/A4/2	0.539994081
8	36/ST/A4/7	0.7426141118
9	46/ST/A5/8	0.8176129039
10	57/ST/A6/9	0.5367088831
11	59/ST/A7/2	0.774403141
12	62/ST/A7/5	0.545246912
13	77/ST/A9/2	0.7071067812
14	85/ST/A9/10	0.6484115071
15	87/ST/A10/2	0.6434588842
16	91/ST/A10/6	0.5156626918
17	97/ST/A11/4	0.7419446627
18	102/ST/A11/9	0.9426151477
19	104/ST/A11/11	0.8316278416

Table 2. *ChatGPT* BLEU Score (<1)

No	Data Code	BLEU Score (NLTK)
1	1/ST/A1/1	0.8423626744
2	4/ST/A1/4	0.905278178
3	14/ST/A2/5	0.8232490472
4	15/ST/A2/6	0.8272321735
5	19/ST/A3/2	0.8403034716
6	21/ST/A3/4	0.8452785147
7	24/ST/A3/7	0.9051034982
8	25/ST/A3/8	0.6757784747
9	26/ST/A3/9	0.9353990131
10	29/ST/A3/12	0.8061898627
11	35/ST/A4/6	0.7624658586
12	38/ST/A5/1	0.8214535097

13	66/ST/A7/9	0.9193227152
14	70/ST/A8/3	0.8056920633
15	71/ST/A8/4	0.8096427216
16	84/ST/A9/9	0.9336510696
17	86/ST/A10/1	0.8049616863
18	87/ST/A10/2	0.9304899483
19	101/ST/A11/8	0.7611606003
20	106/ST/A11/13	0.7748137205

From the total results of the analysis of the calculation of the BLEU score DeepL as much as 105 data found as much as 19 data scored less than 1.0 and 86 data with a BLEU score of 1.0. Based on the analysis results, the total BLEU score generated is 0.9446657236. The score is obtained from the average of the total score of all data. The same analysis was applied to the translation of ChatGPT, 20 sentences with BLEU scores below 1.0 were found, while 85 other data produced a score of 1.0. From the calculation of the average total BLEU score of all data, the total BLEU score is 0.9211813372.

DISCUSSION

DeepL Errors Discussion

a. Omitted Concept Error

In the translation of DeepL, there are 6 concepts omitted from the source text. The omitted concepts found in the translation of DeepL can be in the form of words or phrases. The types of words and phrases found to be mistranslated are adjectives and adverbs. Examples of concept omission error categories in the target text in the translation of DeepL are elaborated in the next explanation.

Table 3. Omitted Concept *DeepL*

Datum No.	SL	TL *DeepL*	Error Category	Suggested Translation
7/ST/A1/7	The results are discussed with respect to the language maintenance …	Hasil penelitian ini membahas pemertahanan bahasa …	Omitted Concept – Adjective Phrase	Hasil penelitian didiskusikan sehubungan dengan pemertahanan bahasa …

The use of adjectives is very important in terms of writing, this is because adjectives provide descriptions that help readers understand certain concepts of a subject or subject. In datum number 7/ST/A1/7 above, the concept of the source language adjective phrase 'with respect' is not found in the target language. This can affect the intention of the source text writer who aims to provide information in more detail and also

emphasize the main topic in the sentence. For comparison, the suggested translation is the addition of the phrase 'terkait dengan' to match the overall concept contained in the source text.

b. Mistranslated Concept

The error analysis conducted on the translation of DeepL found 4 errors in the category of mistranslated concepts. The four mis-translated concepts were found. The mistranslation of a concept found in the translation of DeepL is in the form of mistranslated concepts of verbs, adjectives, and nouns. The explanation of the mistranslation will be explained in some examples in the next paragraphs.

Table 4. Mistranslated Concept *DeepL*

***Datum* No.**	**SL**	**TL *DeepL***	**Error Category**	**Suggested Translation**
31/ST/A4/2	… focusing on Javanese indentured laborers in Suriname, instead.	… berfokus pada para pekerja migran Jawa di Suriname.	Mistranslated Concept - Adjective	… berfokus pada para pekerja Jawa terikat kontrak di Suriname.

In datum number 31/ST/A4/2, there is an adjective translation error. Adjective is a class of words used to describe a certain subject or object. Of course, the right adjective is needed in a sentence so that the writer's message can be conveyed to the reader of the text. In terms of translation, the use of commensurate words, especially adjectives, is important to convey the writer's message in the source language into the target language. In the datum above, the use of the word 'migran' to describe the object of the worker is a different concept from the concept in the source language. In the source language text, the word 'indentured' should be translated as 'terikat kontrak' in the target language.

c. Untranslated Concept

The next category of error is untranslated concepts. In this category, errors are caused by the appearance of words in the source language text in the target language text. In the translation of DeepL, 3 untranslated concepts were found and among them were adjective and noun concepts. Here is an example of the untranslated concept error.

Table 5. Untranslated Concept *DeepL*

Datum No.	**SL**	**TL *DeepL***	**Error Category**	**Suggested Translation**
23/ST/A3/6	However, we argue that broadcasting neo-exotic narratives also created online news media discursive subjects that …	Namun, kami berpendapat bahwa penyiaran narasi neo-eksotisme juga menciptakan subjek diskursif media berita online yang …	Untranslated Concept - Adjective	Namun, kami berpendapat bahwa penyiaran narasi neo-eksotisme juga menciptakan subjek diskursif media berita daring yang …

In datum number 23/ST/A3/6, there is the word 'online' in the source language text and target language text. It is included in the untranslated concept category of

translation error. The word 'online' in the target language, namely English, already has an equivalent word in Indonesian, namely the word 'daring'. Therefore, in the translation suggestion given, the word 'online' is replaced with the word 'daring'.

d. Substituted Concept

The concept substitution error category is an error caused by the appearance of a lexical concept that is not equivalent in the target language but can be said to be a concept that can replace the concept in the source language with the appropriate context (Koponen, 2010). In the translation of DeepL, there are 8 errors in the concept substitution category, each in the form of noun, adjective, and preposition errors. Further explanation will be presented in the next paragraph.

Table 6. Substituted Concept *DeepL*

Datum No.	SL	TL *DeepL*	Error Category	Suggested Translation
57/ST/A6/9	As such, it is common for unregistered marriage (nikah siri) to be a “way out”.	Oleh karena itu, tidak jarang pernikahan yang tidak dicatatkan (nikah siri) menjadi "jalan keluar".	Substitusi Konsep – Adjektiva	Oleh karena itu, merupakan hal biasa ketika pernikahan yang tidak dicatatkan menjadi jalan keluar.

In datum number 2/ST/A1/2, there is a noun concept substitution category error. The concept substitution that occurs in the example is the substitution of the noun 'the case' in the source language text with the verb 'happened' in the target language. However, the context in the source language sentence can still be seen in the target language. In order to fulfill the concept of translation equivalence, the noun phrase 'the case' is translated into 'kasus' in the translation suggestion given.

e. Mistaken Participant

Mistaken participant errors are a category of translation errors that fall into the second category of errors promoted by Koponen (2010), namely errors of relations between concepts. This error considers the equivalence in terms of relations between concepts in a text in the source language and target language. Participant errors are errors caused by differences in the relationship between objects and their modifiers in the source and target texts. The difference is usually in the form of different types of elements used in the source and target texts. One participant error was found in the translation datum of DeepL which will be explained further in the next paragraph.

Table 7. Mistaken Participant *DeepL*

Datum No.	SL	TL *DeepL*	Error Category	Suggested Translation
104/ST/A11/11	… as well as caring women and girls victims at various stages of case handling system.	... serta merawat perempuan dan anak perempuan korban dalam berbagai tahapan	Mistaken Participant	… serta merawat korban wanita dan anak perempuan dalam berbagai tahapan sistem penanganan kasus.

		sistem penanganan kasus.		

The result of the datum analysis above is an example of an error in the relationship between concepts in a translation. In datum number 104/ST/A11/11, the target language text describes a concept with the phrase 'women and girls victims'. In the source text, the English word 'victims' is the head of the phrase and 'women and girls' meaning women and girls as the modifier, but the relationship is changed in the target language. In the translation of DeepL women and girls become the head and victims become the modifier of the phrase. This error can lead to misunderstanding of the message for the reader. Therefore, the translation suggestion adjusts the head and the modifier in the target language to those in the source language.

ChatGPT Errors Discussion

a. Omitted Concept

There are 2 concept omissions found in the ChatGPT translation. Below is one example of a concept omission error found in the ChatGPT translation:

Table 8. Omitted Concept *ChatGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation
29/ST/A3/12	Both movies can consequently be interpreted as cultural texts that …	Kedua film tersebut dapat ditafsirkan sebagai teks budaya yang …	Omitted Concept – Adverb	Kedua film dapat secara pasti ditafsirkan sebagai …

In the datum above, the error lies in the omission of the adverbial concept 'consequently' in the target language. This error is also found in the translation of DeepL. The omission of the adverbial concept can unravel the clarity of the concept in the sentence which affects the reader's understanding of the text. The translation suggestion is to add the phrase 'secara pasti' in the target text.

b. Added Concept

The addition of concepts can reduce the equivalence of a concept in a translation. In the data above, the addition of the prepositional concept 'against' is not appropriate because the equivalent term 'language attitudes' in Indonesian is 'language attitudes'. Therefore, in the suggested translation, the word 'against' is omitted to achieve the equivalence of the meaning of the source text.

Table 9. Added Concept *ChatpGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation

1/ST/A1/1	Language attitudes play an important role in …	Sikap terhadap bahasa memainkan peranan penting dalam …	Added Concept - Preposition	Sikap bahasa memainkan peranan penting dalam …

In datum number 1/ST/A1/1, the addition of the word 'towards' is a translation error because the standard translation of the phrase 'language attitude' in Indonesian is 'language attitude'. Therefore, in the suggested translation the word 'against' is omitted.

c. Mistranslated Concept

Table 10. Mistranslated Concept *ChatGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation
15/ST/A2/6	… development of their hometowns by using podcast as a way to establish the sustainable tourism.	…pengembangan kota halaman mereka dengan menggunakan podcast sebagai cara untuk menegakkan pariwisata yang berkelanjutan.	Salah Diterjemahkan – Nomina	Pengembangan kampung halaman mereka dengan menggunakan siniar sebagai cara untuk menegakkan pariwisata yang berkelanjutan.

The mistranslated concept error category is a fatal error because it can result in 2 fatal possibilities, namely a change in meaning and a void of meaning. In the data above, the direct equivalent of the word 'hometown' in Indonesian is ‘kampung halaman’. The error arises in the use of the phrase 'hometown' which does not convey the context in the source language and also has no meaning. This can cause confusion in the mind of the reader. Therefore, the translation suggestion is to replace the phrase 'kota halaman' with the phrase 'kampung halaman'.

d. Untranslated Concept

Table 11. Untranslated Concept *ChatGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation
19/ST/A3/2	… were positioned as the basis of various tourism events …	… diposisikan sebagai dasar berbagai event pariwisata …	Tidak Diterjemahkan – Nomina	… diposisikan sebagai dasar berbagai kegiatan pariwisata …

An untranslated category error appears in data number 19/ST/A3/2. The occurrence of the English word 'event' in the target language is ChatGPT's failure to translate the concept. The word 'event' can be translated into the word 'acara' to provide an equivalent and appropriate translation in the target language.

e. Substituted Concept

Table 12. Substituted Concept *ChatGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation
19/ST/A3/2	…representation theory with the discursive constructionist approach.	… teori representasi bersama pendekatan konstruksionis diskursif.	Substitusi Konsep – Preposisi	… teori representasi dengan pendekatan konstruksionis diskursif.

Substituted concept error is found in the ChatGPT translation in data number 19/ST/A3/2. The word 'with' in the source text is translated into 'bersama' in the target text. However, the word 'with' in the target text does not have the same meaning as the word 'with' in the source text. The word may not cause the reader's misperception of the source text. Therefore, the use of the word 'bersama' is replaced with the word 'dengan' in the source text to provide an equivalent translation.

f. Mistaken Relation

Table 13. Mistaken Relation *ChatGPT*

Datum No.	SL	TL *ChatGPT*	Error Category	Suggested Translation
25/ST/A3/8	… female characters who turn into ghosts in order to express their anger towards their male oppressors.	… karakter wanita sentral yang berubah menjadi hantu untuk mengekspresikan kemarahan mereka terhadap penindas pria mereka.	Kekeliruan Hubungan	… untuk mengekspresikan kemarahan mereka terhadap pria yang menindas mereka.

In the translation of data text number 25/ST/A3/8, there is a relationship error. The conclusion is drawn from the head and modifier relationship error in the phrase 'their male oppressors'. In the phrase 'their male oppressors' the possessive pronoun 'their' meaning 'mereka' in Indonesian refers to the female subject in the text. However, in the target text the word 'mereka' does not refer to the female subject in the sentence but 'men' who are the object in the sentence. Therefore, in the suggested translation the phrase 'penindas pria mereka' is replaced with the clause ‘pria yang menindas mereka' to achieve translation equivalence.

DeepL vs. *ChatGPT*

a. Error Analysis Comparison

In evaluating the two translation engines, an error comparison of the two translation engines, DeepL and ChatGPT, was conducted. In the previous chapter, the error analysis was conducted based on the theory of machine translation error analysis by Koponen (2010). In the translation of DeepL, a total of 24 errors were found, consisting of 6 errors in the concept omission category, 4 errors in the concept mistranslation

category, 3 errors in the concept not translated category, 8 errors in the concept substitution category, 1 error in the element substitution category, 1 error in the element error category, and 1 error in the relationship error category. In the ChatGPT translation, a total of 25 errors were found, each of which was categorized into each type of error. There are 2 concept omission category errors, 5 concept addition category errors, 4 concept mis-translation category errors, 3 concept untranslation category errors, 5 concept substitution category errors, 1 element addition category error, 4 relationship error category errors, and 1 element substitution category error. To see in detail, each error in both translation machines is presented in the tables as follows:

Table 14. *DeepL* vs. *ChatGPT* Error Analysis

No.	***DeepL***		***ChatGPT***	
	Data Code	**Error Category**	**Data Code**	**Error Category**
1.	2/ST/A1/2	Substituted Concept	1/ST/A1/1	Added Concept
2.	4/ST/A1/4	Mistaken Relation	4/ST/A1/4	Mistaken Relation
3.	7/ST/A1/7	Omitted Concept	14/ST/A2/5	Added Concept
4.	7/ST/A1/7	Substituted Participant	15/ST/A2/6	Mistranslated Concept
5.	17/ST/A2/8	Omitted Concept	15/ST/A2/6	Added Participant
6.	23/ST/A3/6	Untranslated Concept	19/ST/A3/2	Untranslated Concept
7.	26/ST/A3/9	Untranslated Concept	19/ST/A3/2	Substituted Concept
8.	29/ST/A3/12	Omitted Concept	21/ST/A3/4	Untranslated Concept
9.	31/ST/A4/2	Omitted Concept	24/ST/A3/7	Added Concept
10.	31/ST/A4/2	Mistranslated Concept	25/ST/A3/8	Substituted Concept
11.	36/ST/A4/7	Omitted Concept	25/ST/A3/8	Mistaken Relation
12.	46/ST/A5/8	Omitted Concept	26/ST/A3/9	Mistaken Relation
13.	57/ST/A6/9	Substituted Concept	29/ST/A3/12	Omitted Concept
14.	59/ST/A7/2	Mistranslated Concept	35/ST/A4/6	Added Concept
15.	59/ST/A7/2	Substituted Concept	38/ST/A5/1	Substituted Participant
16.	62/ST/A7/5	Substituted Concept	38/ST/A5/1	Mistaken Relation
17.	77/ST/A9/2	Substituted Concept	66/ST/A7/9	Mistranslated Concept
18.	85/ST/A9/10	Substituted Concept 1	70/ST/A8/3	Substituted Concept
19.	85/ST/A9/10	Substituted Concept 2	71/ST/A8/4	Mistranslated Concept 1
20.	87/ST/A10/2	Substituted Concept	71/ST/A8/4	Mistranslated Concept 2
21.	91/ST/A10/6	Mistranslated Concept 1	84/ST/A9/9	Untranslated Concept
22.	91/ST/A10/6	Mistranslated Concept 2	86/ST/A10/1	Omitted Concept
23.	97/ST/A11/4	Mistranslated Concept	86/ST/A10/1	Added Concept

24.	102/ST/A11/9	Substituted Concept	87/ST/A10/2	Substituted Concept
25.	104/ST/A11/11	Mistaken Participant	101/ST/A11/8	Mistranslated Concept
26.			106/ST/A11/13	Substituted Concept

From the error analysis data, it is known that the ChatGPT translation results found more translation errors. From these results, it can be concluded that the DeepL machine translation received a better evaluation score than the ChatGPT machine translation. The number of errors is the basis of the evaluation indication in the discussion section of this research.

b. BLEU Score Comparison

The results of the total score of the calculation of machine translation results *DeepL* and *ChatGPT* are obtained from calculating the average value of all data analyzed. The score obtained by the whole data can be seen in the appendix. From the results of the calculation of the BLEU score of *DeepL* and *ChatGPT*, it is found that the *ChatGPT* machine translation has a higher score than the *DeepL* machine translation. The score results obtained are not based on the numbers of data with scores below 1.0, but rather the overall errors found in a sentence which can be more than one. Furthermore, to further examine the performance of the two translation engines, a comparison of the results of the error analysis contained in the translated text will be provided.

From the two evaluations conducted, automatically using BLEU and directly by humans, showed that the *DeepL* machine translation scored higher. The machine translation *DeepL* can be said to be a machine translation that is able to perform the translation process from English to Indonesian more naturally. Natural because of the error analysis done by humans and evaluation by machines that also use human translation results as a comparison.

CONCLUSION

Machine Translation evaluation needs to be conducted more due to the quality of the translation it produces. Eventhough the vast development technology has made translation machine developed better, the translation machine cannot produce an accurate and human like translation. Based on the research findings, there are some mistakes found in the translation product of machine translation. However, this study also aims to seek for a transaltion machine producing a human-like translation. This study provides information regarding two machine translations: of *DeepL* and *ChatGPT*. According to its evaluation both by human and machine. The result of the analysis showed that there is significant difference both in the human and machine evaluation. Error analysis showed that *DeepL* produces lower error than *ChatGPT*. Furthermore, *DeepL* also showed higher score in BLEU evaluation conducted. Both evaluation result elaborates the better performance of *DeepL* compared to *ChatGPT*. Through this study, it is expected that the reader, especially machine translation machine user, to use a reliable machine translation. In which, a machine translation performing a better

translation with low error rates. This research can be a reference for further research regarding machine translation evaluation.

REFERENCES

Aflah, L. N. Komparasi Hasil Terjemahan Google Translate dan Bing Translator dalam Menerjemahkan Hedging Words. PRASASTI: Journal of Linguistics, 5(1), 68-75.

Amilia, I. K., & Yuwono, D. E. (2020). A Study of The Translation of Google Translate. Lingua: Jurnal Ilmiah, 16(2), 1-21.

Asokawati, A., & 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.

Conde, T. (2012). Quality and quantity in translation evaluation: A starting point. Across languages and cultures, 13(1), 67-80.

Craciunescu, O., Gerding-Salas, C., & Stringer-O'Keeffe, S. (2004, July). Machine Translation and Computer-Assisted Translation: a New Way of Translating? *Translation Journal, 8 (3)*. Retrieved from https://translationjournal.net/journal/29computers.html

Dewi, H. D., & Hidayat, R. S. (2020). The Effectiveness between Two Translation Assessment Models for English to Indonesian Translation of Undergraduate Students. Journal of Language and Literature, 20(2), 270.

Graham, Y., Haddow, B., & Koehn, P. (2019). Translationese in machine translation evaluation. arXiv preprint arXiv:1906.09833.

Hampshire, S., & Salvia, C. P. (2010). Translation and the Internet: Evaluating the Quality of Free Online Machine Translators. *Quaderns*, 197-209.

Hutchins, J., & Sommers, H. (1992). *An Introduction to Machine Translation.* London: Academic Press Limited.

Hutchins, W. J. (1995). Machine Translation: A Brief History. In R. E. Asher, & E. F. Koerner, *Concise History of the Language Sciences from the Sumerians to the Cognitivists* (pp. 431-445). Pergamon.

Koponen, M. (2010). Assessing Machine Translation Quality with Error Analysis. *Electronic proceedings of the KäTu symposium on translation and interpreting studies 4*, 1-10.

Lommel, A., Burchardt, A., Popović, M., Harris, K., Avramidis, E., & Uszkoreit, H. (2014). Using a new analytic measure for the annotation and analysis of MT errors on real data. In Proceedings of the 17th Annual conference of the European Association for Machine Translation (pp. 165-172).

Lotz, S., & Van Rensburg, A. (2016). Omission and other sins: Tracking the quality of online machine translation output over four years. Stellenbosch Papers in Linguistics, 46, 77-97.

Malenova, E. D. (2015). Translating subtitles–translating cultures.

Mehta, S., Azarnoush, B., Chen, B., Saluja, A., Misra, V., Bihani, B., & Kumar, R. (2020, April). Simplify-then-translate: Automatic preprocessing for black-box translation. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8488-8495).

Munday, J. (2016). *Introducing Translation Studies.* New York City: Routledge.

Roberts, R. P. (2002). Translation. In R. P. Roberts, *Translation and Interpretation. The Oxford Handbook of Applied Linguistics.* New York: Oxford University Press.

Saksono, M. B., Anindita, A., & Setiajid, H. H. (2022). THE PERFORMANCE OF GOOGLE TRANSLATE IN TRANSLATING THREE CATHOLIC FUNDAMENTAL PRAYERS. Prosiding Konferensi Linguistik Tahunan Atma Jaya (KOLITA), 20(20), 212-216.

Setiajid, H. H., & Tirtayasa, C. T. (2019). Google Translate's Quality in Translating an English Literary Text into Indonesian Performed in 2017 and 2019: A Diachronic Study. *Konferensi Linguistik Tahunan Atma Jaya 18*.

Sutopo, A. (2019, August). The Assesment and Reserach on translation studies. In Fifth Prasasti International Seminar on Linguistics (PRASASTI 2019) (pp. 10-16). Atlantis Press.

Wilks, Y. (2009). *Machine Translation: Its Scope and Limits.* Sheffield: Springer.

Williams, J., & Chesterman, A. (2002). *The Map; A Beginner Guide to Doing Research in Translation Studies.* Manchester: St. Jerome Publishing.