Barák, Alex. 2024. Comparing Machine Translation Effectivity of Selected Engines from English into Slovak on the Example of a Scientific Text. In: L10N Journal 2(3), pp. 7–28.

Comparing Machine Translation Effectivity of Selected Engines from English into Slovak on the Example of a Scientific Text

Alex Barák

Comenius University in Bratislava

barakalex20@gmail.com

Abstract

In the current age of rapid globalization and technological advancement, it is important to pay attention to machine translation engines. With the rise of artificial intelligence and machine learning, new and improved translation tools are emerging that promise more accurate and faster results. This study focuses on a comparison of the translations (from English to Slovak language) of three prominent tools: Google Translate, DeepL, and the new ChatGPT model. The free versions of these tools are used, except for ChatGPT where we also look at version 4.0, which, at time of writing, is the paid version. The study places emphasis on their capabilities and limitations in translating a specialized text. In the case of the ChatGPT model, the focus is also on how the glossary affects its translation quality. An analysis of not only the final translations but also of the underlying processes and technologies behind these tools is performed. The analysis and comparison of the translation quality of these tools are performed using the TAUS organization's template for evaluating the quality of machine translations. The key objective is to contribute to a better understanding of the advantages and disadvantages of these translation tools.

Keywords: machine translation, Google Translate, DeepL, ChatGPT, translation quality assessment

1 Introduction

In today's globalized society, machine translation plays a key role in overcoming language barriers and enabling effective international communication. With the growing importance of machine translation, various tools have emerged that promise accurate and efficient translations between different languages. These tools include the ChatGPT model, DeepL, and Google Translate, which are currently among the most popular machine translators. While ChatGPT itself is not considered a machine

translator but rather a generative artificial intelligence, throughout this paper, we will refer to the model as a translator or machine translator.

In parallel with this rise in popularity of machine translation, we are witnessing interesting developments in the field of translation. Translators are becoming post-editors – translation experts who can use machine translations efficiently while also correcting and improving the initial output from machine translators, thus speeding up, simplifying, and, in many cases, improving the translation process and the quality of the final translation. *"The fusion of technology and human proficiency in translation endeavors not only augments efficiency but also elevates the quality and cultural relevance of the final output*" (Wang 2024, p. 23).

The aim of this study is to systematically compare the translation quality of these three tools on a scientific text translated from English into Slovak. The study relies on the TAUS (Translation Automation User Society) DQF-MQM framework, which provides a standardized template for evaluating the quality of machine translations. Additionally, the study will examine differences between ChatGPT, DeepL, and Google Translate in terms of their ability to preserve semantics, grammatical correctness, correct terminology, and style in the target translation of a scientific text.

For the actual translation analysis, the study will look at translation error rates, translations correctness, error typology, and specific examples of errors made by the translators made. Finally, a comparison will be conducted on how ChatGPT with a glossary performed compared to ChatGPT without a glossary. Only the ChatGPT model will have a glossary available, since Google Translate does not support the use of glossaries, and DeepL does not have this option available for the Slovak language. Model 3.5 will be used for the primary analysis; however, there will also be an analysis of the translation capabilities of model 4.0. It is important to highlight that the study focuses on analyzing translations from English to Slovak language.

2 Theoretical background

Kenny (2022, p. 32) states that machine translation "involves the automatic production of a target-language text on the basis of a source-language text." It aims to produce a translation that retains the meaning of the original text in a way that is understandable to the reader in the target language.

There are multiple technological approaches to machine translation, such as the rule-based approach, data-driven machine translation, and the statistical approach. However, the approach currently used by many popular machine translators is machine translation, which is based on neural networks and deep learning. This approach allows for better context recognition and improves the overall quality of the translation.

2.1 Neural networks and deep learning

This state-of-the-art approach uses neural networks to model translation relationships between languages. Deep learning allows these models to automatically extract different levels of both semantics and context. Zhixing Tan et al. (2020), in their paper *Neural machine translation: A review of methods, resources, and tools,* describe that neural networks operate based on so-called neurons, layers, and learning.

As they state in the paper, the basic unit of a neural network is a neuron, which is modeled as a mathematical function. Each neuron has a weight and a threshold that determine its behavior. Neurons receive inputs, perform operations according to the weight of those inputs, and produce an output.

Neural networks are organized into layers, including an input layer, hidden layers, and an output layer. The input layer receives inputs, the hidden layers perform computations, and the output layer produces outputs. The hidden layers allow the network to extract different levels of abstraction from the data.

Neural networks are trained on data such as various language corpora, texts from the Internet, etc., and are aiming to minimize the difference between the network's predictions and the actual values. The quality of the texts from which the neural network learns has a great impact on the quality of the translation that the machine performs. Learning involves updating the weights and threshold values of the neurons to achieve the desired behavior of the network.

2.2 Machine translators

For this experiment, the following machine translators will be compared: Google Translate, DeepL, and ChatGPT.

2.3 Google Translate

Google Translate (GT) is an online machine translation tool developed by Google. It is one of the most popular and widely used tools for translating text and sentences between different languages. GT provides a fast and convenient way to translate texts and allows users to communicate and understand content written in other languages.

It is estimated that as early as 2018, around 500 million people used GT, and approximately 100 billion words per day were being translated (Fitriyani 2018). At the time of writing, GT supports 133 languages.

Caswell and Liang (2020) and Zhao (2019) explain that the GT architecture is based on so-called recurrent neural networks (RNNs) and transformers. Transformers are the most important component of the architecture, enabling models to efficiently process long sentences and capture context. RNNs allow the model to process sentences as

wholes, translating them without having to break them down into phrases or words. When translating text, GT breaks the input text into smaller parts. The input text is first preprocessed. This includes removing punctuation and normalizing the text. The text is divided into smaller units. These units are encoded into a vector, which allows the machine model to work efficiently with the text. It then creates context from the encoded text, allowing the model to understand the relationships between words, phrases, and sentences using a large language corpus that contains millions of parallel sentences in different languages. Transformer models can analyze the entire context of the text and, based on this analysis, decode the text in the target language. Machine translation models are trained to predict the next word in the target language based on the context in the source language. This process is repeated iteratively until the entire translation is generated.

GT also uses an automated machine learning system that allows it to continuously improve through user feedback.

It should also be noted here that all information entered into the compiler is processed on external servers. This means that this method of translation is unsuitable for translating sensitive information that must not be shared on external servers for legal reasons (Lukaszewicz 2020).

2.4 DeepL

DeepL entered the market in 2017. It was created by Linguee, which has been providing a database of parallel texts under this name since 2009 (Cambedda et al. 2021). At the time of writing, there is not much information about the principles on which DeepL works. The official website of DeepL (2021) states that DeepL operates on neural networks principles with a modified transformer architecture and has deep learning capabilities. DeepL also differs from other machine translators in its network topology, which allows it to provide better translations.

Regarding the training data, the official website of DeepL states that the translator has been trained on parallel texts of the Linguee corpus, which were generated from official protocols, laws and other documents of the European Parliament (EUR-Lex). On the webpage, it is also stated that the company has also developed special tools that crawl texts on the internet and assess their quality. The neural network was trained by repeatedly showing it different examples of translations. The network then compared these translations with its own translations, and if there were discrepancies, the network's weights were adjusted as necessary. Subsequently, the site notes only that other machine learning methods were also used.

Since 2017, DeepL has become an extremely popular translation tool for many people. As reported by Phrase (2023), at the time of writing, the translator had been used by

more than 1 billion users, supports 31 languages for translation, and includes more than 650 possible language combinations for translation. Users can choose between free and paid versions of the service, as well as between a web interface and a standalone translator. The free version is suitable for personal use, while the paid version offers more features for businesses.

2.5 ChatGPT

ChatGPT (Generative Pre-trained Transformer) is a type of generative AI developed by OpenAI. According to Ray (2023, 121), "Generative AI models rely on deep learning techniques and neural networks to analyze, understand, and generate content that closely resembles human-generated outputs." Deng and Lin (2022) further state that ChatGPT is a system capable of processing natural language and considering the context of the conversation when generating text to produce the most appropriate response. ChatGPT claims that it can reply in over 100 languages, including English, Spanish, German, French, Chinese, Japanese, Russian, Arabic, Portuguese, Dutch, and many more, and that its ability to work in different languages is based on the training data on which it was trained. An approximate number would be over 100 languages, but ChatGPT does not have an exact list of all supported languages (ChatGPT 2023). For the purposes of the study, following models will be used: GPT-3.5 and GPT-4, as they are the latest models, and the GPT-3.5 is currently the only one available for free.

2.6 ChatGPT-3.5

According to Yenduri et al. (2023), GPT-3.5 is a smaller, updated version of the GPT-3. GPT-3.5 was trained on mixed data containing text and code. From the vast amount of data collected from the internet, including thousands of Wikipedia entries, social media posts, and news stories, GPT-3.5 learned to recognize relationships between words, sentences, and different linguistic components. OpenAI has used it to create systems tailored for specific purposes. In addition to being able to translate text, it can also perform basic mathematical operations, write programming codes, and engage in human-like conversations

2.7 ChatGPT-4

Ray (2023) and Yenduri et al. (2023) also describe OpenAI's latest GPT model, ChatGPT-4. This model is a large multimodal language model. It was released on March 14, 2023, and is now available to the general public in a limited capacity through the subscription-based ChatGPT Plus. With this model, OpenAI has made significant progress in improving deep learning. The model can accept both image and text inputs

and generate text outputs. The GPT-4 model has demonstrated the ability to perform many tasks at a similar level to that of humans. For example, in a simulated test, it achieved results comparable to the top 10% of students who took the test. In comparison, GPT-3.5 achieved results comparable to the worst 10% of students. ChatGPT-4 is considered a significant improvement over the 3.5 model in every aspect.

2.8 Previous research

This section provides an overview of previous studies on machine translation (MT) and translation quality assessment. The purpose of this chapter is to contextualize the research by examining existing work in this field, which will help identify gaps and opportunities for the study.

Sanz-Valdivieso and López-Arroyo (2023) compared the effectiveness of ChatGPT and Google Translate in translating specialized texts, specifically on wine and olive oil tasting. Their experiment, which involved translations from Spanish to English, aimed to assess whether the models could accurately handle domain-specific terminology. Standard translation quality assessment (TQA) methods and automated metrics on 50 sentences were used. ChatGPT-3.5 outperformed Google Translate in terminology accuracy, with 12.57% fewer errors and 36% of the text translated without mistakes (compared to Google's 14%). However, both models often replaced terminology with more generic equivalents, and the authors concluded that neither tool is currently accurate enough to work without a domain expert.

Jiao et al. (2023) examined the translation quality of ChatGPT, DeepL, and Google Translate across multiple language pairs (Chinese, English, German, and Romanian). ChatGPT's performance was comparable to that of Google and DeepL for widely spoken languages, but it struggled with languages with fewer training data, such as Romanian, where its BLEU1 score was 46.4% lower than Google's for English-to-Romanian translations.

Petráš and Munková (2023) analyzed Google Translate (both statistical and neural models), DeepL, and ChatGPT, focusing on journalistic texts. They found that while neural models produced smoother translations, they still lacked semantic adequacy. ChatGPT also showed limitations, especially with morphologically rich languages. The authors noted improvements in machine translation, but human oversight is still needed for high-quality translations.

Widiatmika et al. (2023) explored the performance of DeepL, ChatGPT, and Google Translate in translating linguistic texts from English to Indonesian. Using a descriptive-qualitative approach, they found that ChatGPT was most effective in preserving meaning and context. The model was better at identifying examples, abbreviations, and technical terms.

Ogundare and Araya (2023) highlighted that GPT-4 performs similarly to commercial translators for high-resource languages but struggles with low-resource languages. They proposed a method involving intermediate translations into high-resource languages to improve quality for low-resource language pairs.

Wang et al. (2023) and Karpinska and Iyyer (2023) noted that ChatGPT matches the performance of other tools for document-level translation. Similarly, Bang et al. (2023) found that while ChatGPT competes with commercial tools for high-resource languages, it suffers from a significant drop in performance (up to 50%) for low-resource languages.

Yulianto (2021) compared Google Translate and DeepL for French-English translations, demonstrating DeepL's superiority in readability and translation quality. Newcomer (2024) also emphasized that DeepL provides more natural-sounding translations and handles idioms better, though Google Translate supports more languages and excels in specific combinations, such as Arabic, Korean, and Mandarin.

Key findings from the aforementioned research can be summarized as follows:

- Compared to other translators, ChatGPT performs better in translation of terminology and adhering to terminology.
- ChatGPT is better at preserving the meaning of the text and considering the context during translation.
- ChatGPT performs better in tasks such as distinguishing examples, clarifying examples, recognizing abbreviations, identifying synonyms, and differentiating sentence structures.
- Google Translate and DeepL handle translations of languages with limited training data more effectively.
- The quality of GPT 3.5 translations can be improved by translating languages with low amounts of training data first into a high-resource language, and then into the target low-resource language.
- ChatGPT, DeepL, and Google Translate have similar translation quality at the document level.
- Google Translate achieves better translations with larger language combinations.
- None of the translators are yet sophisticated enough to produce high-quality translations without the assistance of a human post-editor knowledgeable in the subject matter.

3 Methodology

The aim of this study is to compare two popular online machine translators (Google Translate and DeepL) and the new ChatGPT generative AI to find out which of them can produce a more successful (accurate) and higher-quality translation from English to Slovak and what are their strengths and weaknesses in translation of specialized texts. The study also aims to assess whether and to what extent the use of a glossary in the case of ChatGPT would improve the quality of its translation. The study will also compare the translations generated by GPT-3.5 both with and without a glossary, as well as those produced by GPT-4 with a glossary. This comparison aims to assess whether ChatGPT models can effectively utilize a glossary and to evaluate the extent to which the glossary improves translation quality relative to other translations.

During the analysis, the focus will be on answering the four research questions:

- Which translator was more successful based on error rate?
- Which translator was more successful based on the number of penalty points obtained?
- What types of errors did each translator make most often?
- How does the glossary improve the translation quality of ChatGPT, and to what extent are the GPT models able to use the terminology correctly and consistently?

To evaluate the translation quality of each translator (including ChatGPT), an excerpt was chosen from a blog post by an author with the username FALLENANGEL. This text was selected because it contains sophisticated use of language, including nuanced vocabulary, complex sentence structures, metaphors, and specialized terminology. This makes it a challenging test case for machine translators, which must handle both literal translation and contextual nuances. The blog discusses various literary aspects of the famous work The Divine Comedy by Dante Alighieri. The text size had to be chosen accordingly so that all the translators could process it in a single prompt, so that the text did not need to be inserted in parts but could be inserted as a whole. Google Translate has the smallest prompt size, stating a limit of 3,900 characters. However, after inserting the text, it was found that it can actually accept a maximum of 2,711 characters. Therefore, the excerpt used in this study consists of 2,510 characters, including spaces, or 414 words (blog post available at: https://stottilien.com/2015/02/09/9306/).

The texts were then translated into Slovak by all three translators (as mentioned, we will also refer to ChatGPT as translator). In the case of ChatGPT, a suitable prompt had to be created to trigger its translation capabilities. This prompt was provided in Slovak: "Prelož tento text do slovenčiny:" (in English: "Translate this text into Slovak.") The TAUS table was then used to evaluate the translation quality of each translator.

For evaluation of the translations, the TAUS quality assessment table was used (template available at: https://info.taus.net/dqf-mqf-error-typology-template-download).

After evaluating the translations, a terminology list of the terms present in the text was developed in Slovak. Both ChatGPT and DeepL have the ability to use a glossary in translation. However, at the time of writing, this feature in DeepL is not available for the Slovak language (the glossary only supports combinations of English, German, Spanish, French, Italian, Polish, Chinese, Danish, Russian, and Portuguese). Nevertheless, it is possible to create a prompt for ChatGPT that serves as a glossary during translation. After the initial attempts to examine how and whether different prompts affected ChatGPT's ability to work with a glossary, a final prompt was created: "Prelož tento text do slovenčiny" (in English: "Translate this text into Slovak:", (inserted original text in English), followed by: "Tu sú termíny z textu a preklady termínov ktoré použi pri preklade" (in English: "Here are the terms from the text and the translations of the terms to use in the translation:"), followed by the listed terms and their translations, and the prompt was ended as follows: "Tieto termíny môžeš v texte skloňovať a používať ich plurálové formy" (in English: "You can inflect these terms in the text and use their plural forms"). With this prompt, it was ensured that ChatGPT understood to use the terms from the glossary for translation. Pilot experiments confirmed that if the terms were not present in the text, ChatGPT would not try to artificially add them to the text. This process ruled out various defective prompts and resulted in the best prompt for this experiment – one that best helps the model understand what is expected of it. A new chat was created so that ChatGPT did not have access to (and was not influenced by) previous translations and translate the same text into English using a glossary. The translation was then re-analyzed using the TAUS quality assessment table.

4 Analysis and comparison

4.1 Translation error rate

First, the number and the severity of errors will be examined. During the research, only two severity levels were identified – major and minor.

Table 1. ***Translation error rate according to the severity levels.***

	Error rate				
Severity level	**Google Translate**	**DeepL**	**ChatGPT 3.5**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Major	8	5	9	4	0
Minor	29	23	36	31	22
Total	37	28	45	35	22

ChatGPT 3.5 made the greatest number of errors in its translation. It also made the highest number of minor and major errors. On the other hand, ChatGPT 4 made the fewest number of errors out of all the translators. It also made zero major errors, making it the only translator that has achieved this in this study. ChatGPT 3.5 with glossary is comparable to DeepL, however DeepL made fewer minor mistakes. However, it must be noted that mistakes in terminology were considered major mistakes, and since only the ChatGPT models had a glossary at their disposal, it is understandable why they made the fewest major errors. Even ChatGPT made an error in terminology, except for model 4. This will be further analyzed in Chapter 3.3.4 Terminology.

Comparing the 3.5 models with and without a glossary makes it evident that a glossary improves the quality of the translation. However, it must also be noted that each time ChatGPT translates the same text, the translation will differ slightly, and thus, the quality of the translations will vary. occurs because ChatGPT is not designed solely as a translator; rather, it is intended to imitate human responses and communication.

4.2 Translation correctness

Translation correctness was evaluated based on the number of penalty points assigned to each translator using the TAUS template.

Table 2. ***Translation correctness***

	Translation correctness				
	Google Translate	**DeepL**	**ChatGPT 3.5**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Number of penalty points	69	48	81	51	22

Since translation correctness is closely tied to the category of translation error rate, it is possible to observe similar results. The most accurate translation was produced by GPT-4. It is evident that DeepL and GPT-3.5 with a glossary do not differ significantly from each other in terms of translation correctness. However, even though GPT-3.5 with a glossary and Google Translate made a very similar number of errors, they differ much more in translation correctness. This is because Google Translate made more major errors, which have the greatest impact on the final translation correctness score. GPT-3.5 produced the least successful translation. Thus, a significant improvement in the translation quality of the GPT models is already apparent, as in only one generation, it has progressed from being one of the weakest translators to competing with the better ones. However, the TAUS template deemed all translations a failure. A translation is considered to have passed only if it contains fewer than 50 errors in 1,000 words, a threshold that all translations in this study (approximately 330 words) far exceeded.

4.3 Error typology

Here, each category and its subcategories in which the translators made errors are presented. The TAUS quality assessment template contains 8 basic error categories, but in this experiment, the translators made errors in only four of them: accuracy, fluency, style, and terminology. Only the GPT-3.5 and GPT-4 made errors in the terminology category, as they were the only translators that had access to a glossary.

Table 3. ***Error categories***

	Errors				
Error category	**Google Translate**	**DeepL**	**ChatGPT**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Accuracy	16	12	17	5	2
Fluency	11	9	20	24	11
Terminology	-	-	-	1	0
Style	10	7	8	5	9
Design	0	0	0	0	0
Locale convention	0	0	0	0	0
Verity	0	0	0	0	0
Other	0	0	0	0	0

It is evident that the GPT-4 was the most accurate of all the translators, meaning it made the fewest errors that impacted the meaning of the text. In terms of fluency, it is comparable to Google Translate, but there is also a significant improvement over the previous models. DeepL was the most fluent, meaning it made the fewest grammatical errors. In the style category (which includes errors where the translation sounded unnatural), GPT-3.5 with glossary performed the best. The GPT-4 model made 4 more errors, but again, this could be due to the inconsistent text generation of ChatGPT (meaning that if the same text was translated again, the results could vary to some extent). As previously mentioned, in the terminology category, only the two GPT models were capable of making errors, but only one of them actually did. GPT-3.5 with glossary was the only model that ignored a term from the glossary. A closer analysis of this particular error will be provided in Chapter 3.3.4 Terminology. The translators did not make any errors in remaining categories. It could be argued that this was due to the nature of the text, which did not allow for such types of errors.

4.3.1 Accuracy

The accuracy category covers errors in translation that alter the meaning or purpose of the text or otherwise misrepresent the source text.

Table 4. ***Accuracy errors***

	Errors				
Error subcategory	**Google Translate**	**DeepL**	**ChatGPT**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Addition	0	0	0	0	0
Omission	0	1	0	0	0
Mistranslation	8	6	11	5	3
Over-translation	0	0	0	0	0
Under-translation	1	0	0	0	0
Untranslated text	7	5	6	0	0
Improper exact TM match	0	0	0	0	0

In the omission category, only DeepL made an error by failing to translate the first part of a sentence. A great advantage of machine translators is their ability to translate everything, since the machine typically processes text sentence by sentence. However,

it appears that even this feature cannot be relied on 100% of the time. This particular sentence also posed a challenge for ChatGPT, as its translation sounded very unnatural and awkward. This issue will be further analyzed in Chapter 4.3.3 Style.

Mistranslations were often caused by word-for-word translation. For example, the term “Big assortment”, which in the text refers to the English translation of the title of Ptolemy’s book *Megale Syntaxis*, should be translated into Slovak as "Veľká kniha"(Big book). However, all translators were influenced by the English phrase and translated it as "Veľký výber" or "Veľký sortiment"(both meaning Big selection), except for the ChatGPT model with a glossary, as this expression was included in its glossary.

ChatGPT’s mistranslations were often caused by the fact that it translated certain words into Czech instead of Slovak. This occurs because these languages are very similar and mutually intelligible. Additionally, it is possible to find Czech words in Slovak texts on which the machine translators are learning. Every model made this error, but GPT-4 made it only once.

The text contained many expressions from a third language for which Slovak has its own equivalents. An example of such a word is “Canto” (in Slovak: spev). Only ChatGPT with a glossary correctly translated it as "spev", but again, it must be noted that this term was included in its glossary. DeepL retained the original word but slovakized it by changing the initial "c" to "k" and further inflecting it as a Slovak word. Other translators also inflected the original form but did not change the initial letter.

The under-translation subcategory refers to errors where the translation is less specific than the source text or where the full meaning is not correctly translated into the target language. Only Google Translate made an error in this subcategory. Google Translate was misled by the source text and retained the name “Mount” in its original form. The translator likely followed the naming convention of Mount Everest and similar cases, because this name is used in Slovak in this form. However, the issue is that even in Dante’s work itself, the mountain is referred to in Slovak as "hora Očistec" and not "Mount Očistec". Clearly, Google Translate correctly recognized that it needed to translate this name but failed to translate the full name correctly.

The untranslated subcategory pertains to text that remains untranslated in the target text. In this case, it must be noted that almost all the translation errors were caused by expressions written in a third language in the source text, such as “Purgatorio” or “Paradiso”. These names refer to the titles of the different parts of the *The Divine Comedy*. In the source text, these names were also left in the third language, even though Slovak has its own translations of these terms, which are used in the official translations of the *The Divine Comedy* by Jozef Felix and Viliam Turčány.

Additionally, ChatGPT-3.5 incorrectly left the English title of the book in the translation (“Comedy” instead of the Slovak “Komédia”). It was likely confused by the quotation marks and did not attempt to translate the expression within them.

4.3.2 Fluency

This subcategory primarily deals with errors such as grammatical mistakes, spelling errors, and similar issues.

Table 5. ***Fluency errors***

	Errors				
Error subcategory	**Google Translate**	**DeepL**	**ChatGPT**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Punctuation	3	6	3	9	0
Spelling	0	0	2	1	0
Grammar	8	3	15	14	11
Grammatical register	0	0	0	0	0
Inconsistency	0	0	0	0	0
Link/cross-reference	0	0	0	0	0
Character encoding	0	0	0	0	0

Punctuation errors involve missing or incorrectly used punctuation. Most of the errors made by the translators were missing quotation marks. In certain parts of the source text, closing quotation marks were likely omitted by mistake. As a result, translators like DeepL or Google Translate also omitted the closing quotation marks. Interestingly, the ChatGPT models correctly added these quotation marks in the translation. However, even though they inserted them, they used the English-style quotation marks (" ") instead of the Slovak variant („ “). On the other hand, Google Translate was the only translator that consistently and correctly replaced the English quotation marks with Slovak quotation marks. However, it was unable to independently add quotation marks where they were missing in the source text.

The spelling subcategory addresses incorrect spelling, inflection of words, typographical mistakes, and similar issues. Only the GPT-3.5 models made errors in this category, struggling with the inflection of the word “Ptolemaic” in Slovak language.

The fluency category was dominated by errors in the grammar subcategory. This subcategory includes mistakes such as incorrectly case usage, sentence syntax, and overall incorrect sentence construction. The ChatGPT models made the most errors in this subcategory, with the GPT-4 having the fewest errors (11) and GPT-3.5 without a

glossary having the most errors (15). The DeepL translator made only 3 errors in this subcategory.

4.3.3 Style

This category highlights the stylistic issues in the text. It consists of five subcategories, but errors were found in only one – the awkward subcategory.

Table 6. ***Style errors***

	Errors				
Error subcategory	**Google Translate**	**DeepL**	**ChatGPT 3.5**	**ChatGPT 3.5 with glossary**	**ChatGPT 4 with glossary**
Awkward	10	7	8	5	9
Company style	0	0	0	0	0
Inconsistent style	0	0	0	0	0
Third-party style	0	0	0	0	0
Unidiomatic	0	0	0	0	0

The awkward subcategory addresses parts of the text that sound strange or unnatural in the target language. Most of these errors were caused by the use of words that did not fit the context in terms of meaning. Additionally, many errors resulted from machine translators attempting to translate a complicated compound sentence without breaking it down in the target language, resulting in convoluted sentence structures and, at times, nonsensical sentences. Google Translate had the most errors in this category; however, the other translators did not perform significantly better, except for the GPT-3.5 model with a glossary. Notably, this model made only 5 errors. Interestingly, GPT-4 produced more errors despite being a more advanced version than its predecessor. Once again, this highlights the inconsistent nature of the outputs of the ChatGPT models.

4.3.4 Terminology

This category highlights the stylistic issues in the text. It contains five subcategories, but errors were found in only one – the awkward subcategory.

Table 7. ***Style errors***

	Errors

Error subcategory	Google Translate	DeepL	ChatGPT	ChatGPT 3.5 with glossary	ChatGPT 4 with glossary
Inconsistent with term base	0	0	0	1	0
Inconsistent use of terminology	0	0	0	0	0

As previously mentioned, only ChatGPT could make mistakes in this category, as it was the only translator with access to a glossary. The model was provided with a glossary that contained 13 terms in total. Although it had to work with a relatively short text and a small number of terms, GPT-3.5 failed to remain consistent with the glossary in one instance. It had issues with the term "The Prayer and Purification passage" (which should be translated into Slovak as "Priechod modlitby a očistenia"). The term "Priechod" (meaning "passage") was incorrectly translated as "cesta" (meaning "road"). The rest of the terms from the glossary were translated correctly. It is unclear why the model ignored this particular term in the translation. However, the GPT-4 model was able to translate every term correctly and consistently.

5 Discussion

After analyzing the results of the experiment, answers to the research questions posed in Chapter 3 are presented below.

Which translator was more successful based on error rate?

Based on the research findings, ChatGPT-4 produced the fewest errors (22), followed by DeepL (28). Google Translate and ChatGPT-3.5 with a glossary had a similar number of errors (37 and 35, respectively). The highest number of errors was recorded for ChatGPT-3.5. Therefore, in terms of error rate, ChatGPT-4 was determined to be the most successful translator.

Which translator was more successful based on the number of penalty points obtained?

Since the number of penalty points is relatively closely correlated with the category of translation error rate, it is possible to observe some similarities. However, this criterion provides an insight into the severity of errors made by the translators. For example, while DeepL produced significantly fewer errors than ChatGPT-3.5 with a glossary (DeepL: 28, ChatGPT-3.5 with a glossary: 35) (see Table 1), the difference in penalty points is less pronounced. DeepL accumulated 48 penalty points and the ChatGPT model with a glossary received 51. This result indicates that the ChatGPT model made more minor errors, while DeepL made more major errors, as major errors have the

greatest impact on the final number of penalty points. Regarding Google Translate, a total of 69 penalty points was recorded. The least successful translation was produced by ChatGPT-3.5 without a glossary, with 81 penalty points. These findings further demonstrate that, although ChatGPT model produced the least successful translations, its performance improved significantly when provided with a glossary. Even when some terms were not translated correctly, the glossary contributed to a substantial improvement, allowing it to compete with the best-performing translators. Notably, ChatGPT-4 achieved the highest level of success in this regard, with only 22 penalty points.

What types of errors did the translators make the most often?

The most common errors made by the translators occurred in the categories of accuracy, fluency, and style. Other categories, such as design or locale convention, could not be tested due to the nature of the translated text. This topic presents an opportunity for future research).

In the category of translation accuracy (which focuses on the correct transfer of meaning from the source to the target text), ChatGPT-4 made by far the fewest errors (see Table 4), and none of these errors were classified as major. Surprisingly, DeepL ranked third, with 12 errors, meaning that even ChatGPT-3.5 with a glossary produced a more accurate translation. This result may be attributed to the glossary used by both ChatGPT models, as ChatGPT without a glossary made 17 errors in accuracy. Google Translate made only one less error than ChatGPT-3.5. A closer examination of accuracy errors reveals that the greatest number of errors in the mistranslation and the untranslated text subcategories. However, GPT-3.5 with a glossary and GPT-4 made 0 errors in these subcategories. Additionally, DeepL was the only translator that made an error in the omission subcategory, while Google Translate was the only one with an error in the under-translation subcategory.

In the fluency category, which addresses formal aspects of the language (such as grammar, syntax, etc.), ChatGPT-3.5 with a glossary made the most errors (24), while DeepL made the fewest (9) (see Table 5). ChatGPT-4 followed with 11 errors, while ChatGPT-3.5 model made 20 errors and Google Translate also made 11 errors. These findings showcase the strengths and weaknesses of the translators. While DeepL was initially expected to perform best in terminology, ChatGPT-3.5 with a glossary and GPT-4 outperformed it in this aspect. However, it should be noted that without the option of using a glossary, ChatGPT models would likely not have achieved this level of accuracy, and DeepL might have been the best-performing translator in this case as well. It is also worth noting that DeepL has been trained on parallel texts from the Linguee corpus, which includes official protocols, legal documents, and other documents from the European Parliament (EUR-Lex). Thus, it can be assumed that if

the experiment had been conducted on legal texts, DeepL would likely have demonstrated superior performance in terminology accuracy.

In the style category (which addresses stylistic problems in the text), ChatGPT-3.5 with a glossary performed better, making only 5 stylistic errors, while ChatGPT model without glossary made 8 (see Table 6). Thus, the model performed comparably to the DeepL translator, which had 7 errors in this category. However, the latest GPT-4 made 9 stylistic errors, almost as many as Google Translate (10 errors), once again demonstrating the variable output of ChatGPT.

Next, the study aimed to determine how the glossary improves the translation quality of ChatGPT and to what extent GPT models are able to use terminology correctly and consistently. The glossary contained 13 terms. ChatGPT-3.5 with a glossary correctly used 12 terms, achieving a 92.3% success rate in translating the terms correctly. In contrast, GPT-4 had no issues with the glossary and successfully translated all 13 terms.

To what extent the glossary improved the quality of the translation was already partially addressed. As previously established, translation accuracy is the category most affected by the glossary. ChatGPT with a glossary made significantly fewer accuracy errors than the model without a glossary (see Table 3). However, in the fluency category, a slight deterioration was observed in the model with glossary (24 errors) compared to the model without glossary (20 errors). As mentioned earlier, this result can likely be attributed to the model's inability to generate consistent translations of the same text. A similar trend was observed in the style category (ChatGPT-3.5: 8 errors, ChatGPT-3.5 with a glossary: 5 errors, ChatGPT-4: 9). GPT-4 was tested only with the glossary, but it produced by far the fewest errors in all categories except for the category of style.

Regarding the overall number of errors, GPT-3.5 with a glossary made 35 errors, while the model without a glossary made 10 more (45 errors). GPT-4 made only 22 errors (see Table 3). However, GPT-3.5 without a glossary produced significantly more major errors (9) compared to the model with a glossary (4), whereas GPT-4 made no major errors. Due to this, a significant difference in the number of penalty points assigned to each model was observed. ChatGPT-3.5 accumulated 81 penalty points, while the glossary model received considerably fewer (51 points). Since GPT-4 only made minor errors, it received just 22 penalty points (see Table 2). Thus, it can be concluded that the glossary had a significant impact on the translation quality of the model, particularly in terms of translation accuracy and in the number of penalty points. In other areas, the difference was not significant enough to confidently attribute it to the glossary alone rather than other factors, such as inconsistent translation outputs. Additionally, there is a notable improvement of overall translation capabilities between the GPT-3.5 and GPT-4 models.

Throughout the research, the focus has been on identifying which translator produced the most successful translation with the lowest error rate. However, it must be noted that even the best-performing translator has not yet reached a level where it can reliably translate texts without the intervention of a human post-editor.

6 Conclusion

The goal of this research was to compare and evaluate selected translators based on their ability to translate the selected specialized text.

This study analyzed the performance of Google Translate, DeepL, and the ChatGPT model across multiple aspects of translation quality, using the TAUS quality assessment template.

First, the study examined the number of errors in the translations. The analysis showed that the fewest number of errors was made by ChatGPT-4.0. In contrast, ChatGPT-3.5 without a glossary produced the greatest number of errors. However, the glossary improved its translation quality, making its error count comparable to Google Translate. DeepL was the second-most successful translator in this regard.

Next, the study assessed the number of penalty points obtained based on the severity of errors. Although DeepL made significantly fewer errors than ChatGPT-3.5 with a glossary, in terms of penalty points the difference was minimal. This finding demonstrates that the quality of the translation is not only determined solely by the number of errors but also by their severity. ChatGPT-4 again received the fewest penalty points.

Another important aspect of the research was the analysis of the types of errors that the translators made. The study found that translators most frequently made errors in the categories of translation accuracy, translation fluency, and style. Additionally, errors in terminology were observed, including the incorrect translation of glossary terms and inconsistent translation of the same term throughout the text.

In terms of accuracy, ChatGPT-4 produced the best translation. Among the 3.5 models, the version with a glossary made significantly fewer errors than the version without. DeepL made more than twice as many accuracy errors as ChatGPT-3.5 with a glossary. ChatGPT-3.5 and Google Translate made almost the same number of errors in this category. The most fluent translation was produced by DeepL, followed by ChatGPT-4. The ChatGPT-3.5 models performed similarly, indicating that a glossary does not impact the fluency of the translation. Google Translate made the same number of fluency errors as ChatGPT-4.

In terms of style, the best translation was produced by ChatGPT-3.5 with a glossary, followed by DeepL. GPT-3.5 and 4 had a similar number of stylistic errors, while Google Translate made the most stylistic errors.

Finally, the study examined how the glossary affects the quality and success of ChatGPT's translation. The results indicate that the glossary significantly improves translation quality in the category of translation accuracy but has limited impact on other areas, such as fluency and style.

Thus, the two best-performing translators in this experiment were DeepL and ChatGPT-4. The advantage of DeepL lies in its ability to generate consistent translation quality, a characteristic that cannot be attributed to the other translators studied. ChatGPT-4 demonstrated good potential, outperforming even DeepL in translation accuracy. However, its writing style and fluency still require improvement. Additionally, because ChatGPT generates different translations of the same text, its translation consistency cannot be fully relied upon. It can also be concluded that, at the time of writing, none of the translators are capable of generating sufficiently high-quality translations without human post-editing. Each translator has its own advantages and disadvantages, and all can serve as valuable tools when used appropriately by human translators.

This study provides insight into the performance and limitations of various machine translators. The findings present opportunities for further research and underscore the importance of considering multiple factors when evaluating and selecting machine translators.

Barák, Alex. 2024. Comparing Machine Translation Effectivity of Selected Engines from English into Slovak on the Example of a Scientific Text. In: L10N Journal 2(3), pp. 7–28.

Bibliography

Bang, Y., et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. The Hong Kong University of Science and Technology. In: arXiv. https://arxiv.org/abs/2302.04023. Accessed on: 2 December 2023.

Caswell, Isaac, and Liang, Bowen. 2020. Recent advances in Google Translate. https://blog.research.google/2020/06/recent-advances-in-google-translate.html. Accessed on: 30 November 2023.

DeepL, 2021. How does DeepL work? https://www.deepl.com/en/blog/how-does-deepl-work. Accessed on: 23 November 2023.

Deng, J., and Lin, Y. 2022. The benefits and challenges of ChatGPT: An overview. In: Frontiers in Computing and Intelligent Systems. 2(2): pp. 81-83. ISSN: 2832-6024. https://scholar.google.com/scholar?hl=sk&as_sdt=0%2C5&q=A+Multitask%2C+Multilingual%2C+Multimodal+Evaluation+of+ChatGPT+on+Reasoning%2C+Hallucination%2C+and+Interactivity&btnG=. Accessed on: 15 November 2023.

Fallenangel. (9 February, 2015). Dante's Divine Comedy – symbolism and archetypes. [Blog]. StOttilien. https://stottilien.com/2015/02/09/9306/.

Fitriyani, Dian Zelina. 2019. Translation Process and Product of Google Translate in Translating Health Articles from English into Indonesian. In: UNNES International Conference on English Language Teaching, Literature, and Translation (ELTLT 2018). Atlantis Press. pp. 361-365.

Jiao, Wenxiang, et al. 2023. Is ChatGPT a good translator? A preliminary study. Tencent AI Lab. In: arXiv preprint. https://arxiv.org/abs/2301.08745. Accessed on: 26 December 2023.

Karpinska, Marzena; Lyyer, Mohit. 2023. Large Language Models Effectively Leverage Document-level Context for Literary Translation, but Critical Errors Persist. In: Proceedings of the Eighth Conference on Machine Translation. Singapore: Association for Computational Linguistics. pp. 419-451.

Kenny, Dorothy. 2022. Machine Translation for Everyone: Empowering Esers in the Age of Artificial Intelligence. [online]. Berlin: Language Science Press. ISBN: 978-3-96110-348-5. https://langsci-press.org/catalog/book/342. Accessed on: 23 December 2023.

Lukaszewicz, S. 2020. Google Translate: The Unwitting Confidentiality Flaw. Imperialcrs. https://www.imperialcrs.com/blog/business-insights/google-translate-the-unwitting-confidentiality-flaw. Accessed on: 1 January 2025.

Newcomer, C.; 2024. TranslatePress. DeepL Translator Review: Is It Better Than Google Translate? [online]. https://translatepress.com/deepl-translator-review. Accessed on: 30 January 2023.

Ogundare, Oluwatosin and Araya, Gustavo Quiros. 2023. Comparative Analysis of ChatGPT and the evolution of language models. In: arXiv preprint. arXiv:2304.02468. https://arxiv.org/abs/2304.02468. Accessed on: 25 December 2023.

Petráš, P., Munková, D. (2023): Machine Translation Based on Neural Networks – a Promising Way to Translate from Analytic Languages into Flective Slovak? In: Slovenská reč, 88/1, p. 74-89.

Phrase, 2023. Exploring DeepL for Machine Translation: How It Works, and How Accurate It Is. https://phrase.com/blog/posts/deepl/#how-does-deepl-work. Accessed on: 27 October 2023.

Barák, Alex. 2024. Comparing Machine Translation Effectivity of Selected Engines from English into Slovak on the Example of a Scientific Text. In: L10N Journal 2(3), pp. 7–28.

Phrase, 2023. Machine Translation Explained: Types, Use Cases, and Best Practices. https://phrase.com/blog/posts/machine-translation/#how-does-machine-translation-work. Accessed on: 27 October 2023.

Ray, Partha Pratim. 2023: ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. In: Internet of Things and Cyber-Physical Systems, Vol. 3: pp. 121-154. ISSN 2667-3452.

Sanz-Valdivieso, Lucía, and López-Arroyo, Belén. 2023. Google Translate vs. ChatGPT: Can non-language professionals trust them for specialized translation? In: International Conference Human-informed Translation and Interpreting Technology (HiT-IT 2023). pp. 97-107. ISSN 2683-0078.

Tan, Zhixing, et al.: Neural machine translation: A review of methods, resources, and tools. 2020. In: AI Open, Vol.1: 5-21. ISSN: 2666-6510.

TAUS: About us [online]. https://www.taus.net/company/about-us. Accessed on: 30 January 2023.

TAUS: Machine Translation Post-editing Guidelines. https://info.taus.net/dqf-mqf-error-typology-template-download. Accessed on: 10 November 2023.

TAUS: Start tracking errors with DQF-MQM. [online]. https://info.taus.net/dqf-mqf-error-typology-template-download. Accessed on: 10 November 2023.

Wang, Longyue et al.: Document-Level Machine Translation with Large Language Models. 2023. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, p. 16646-16661, Singapore. Association for Computational Linguistics.

Wang, Y. 2024. The Impact of Technology on Human Translators and Translation Quality: A Study on Machine Translation and Computer-Assisted Translation Tools. In: English Linguistics Research, 13, 19. https://doi.org/10.5430/elr.v13n1p19.

Widiatmika, Putu Wahy et al. 2023. Examining the result of machine translation for linguistic textbook from English to Indonesian. In: Proceeding the second english national seminar: exploring emerging technologies in english education. 2023. LPPM Press STKIP PGRI PACITAN. p. 54-65. ISSN 2986-6456.

Yenduri, Gokul, et al. 2023: Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. In: arXiv preprint. arXiv:2305.10435, https://arxiv.org/abs/2305.10435. Accessed on: 24 December 2023.

Yulianto, Ahmad and Supriatnaningsih, Rina. 2021. Google Translate vs. DeepL: A Quantitative Evaluation of Close-language Pair Translation (French to English). AJELP: Asian Journal of English Language and Pedagogy. Vol. 9 No .2, p. 109-127. ISSN 2289-8697. https://ojs.upsi.edu.my/index.php/AJELP/article/viewith6087. Accessed on: 28 December 2023.

Zhao, Tianyi, 2019. CCTP-607: “Big Ideas”: AI to the Cloud. The AI Powers Behind Google Translate. https://blogs.commons.georgetown.edu/cctp-607-spring2019/2019/05/01/the-ai-powers-behind-google-translate/. Accessed on: 26 December 2023.