issue 2/2024

tcworld Journal

Translation in Software. Software in Translation

L10N Journal: Translation in Software, Software in Translation. 2/2024

Publisher
STIMUL: the publishing and advisory center of the Faculty of Arts, Comenius University, Gondova 2, Bratislava, Slovakia, fif.stimul@uniba.sk, https://fphil.uniba.sk/stimul/

ISSN 2730-0757

Sponsors

*The publication of this issue was financially supported by the **Scientific Grant Agency VEGA**, under project No 1/0113/23 **Localization of Software Products Specifics in the Slovak – English Language Pair**.*

Scientific Committee

Contents

Kabát, Marián. 2024. Introduction. In: L10N Journal. 2(3), pp. 5-6.

Introduction

Marián Kabát

Comenius University in Bratislava

marian.kabat@uniba.sk

When we established the L10N Journal in 2021, the initial thought was to present Slovak thinking on localization and technologies to the scholarly community outside of Slovakia. The first two issues were symbolic; we wanted to show that localization research has its place in a country where the predominant research in translation studies was focused on literary translation. Looking back at the past three years gives me hope that what we started in 2021 continues today in 2024. This is evident, for example, in the program of both Winter Schools in Translation, where localization – and video game localization in particular – had its place.

The following issue, titled Slovak Research on Localization 3, is a continuation of the first two issues. It showcases research by young scholars who are interested in localization and technologies in translation, as well as their impact on English-to-Slovak translation.

The first article, by Alex Barák, compares three engines used for machine translation, one of which is a large language model. Research of this type was partly absent in Slovakia but is imperative in informing practicing translators about which tools are better suited for their day-to-day practice.

Radka Filkorová addresses the issue of consistency in spin-off literature related to video games. This reader-centric approach emphasizes the importance of cooperation between video game localizers and literary translators.

The next article, by Zuzana Hudáková, explores the impact of source text pre-editing for machine translation and machine translation post-editing, as well as the impact of education on these processes, their outcomes, and final translation quality.

Matúš Nemergut examines translation edit rate (TER) in English-to-Slovak machine translation post-editing. This research is first of its kind in this language pair; it began as a collaborative effort between the author and the industry, and the results could impact the way post-editors work today.

I consider the presented research to be highly significant, particularly for the English-to-Slovak translation community, as it addresses crucial practical challenges that have direct implications for the field. By exploring these important issues, the studies contribute to both theoretical discussions and real-world applications in translation practice. Furthermore, I sincerely hope that research of this nature will continue to develop in the coming years, fostering further discussions in the field.

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.

Filkorová, Radka. 2024. Consistency Analysis of Slovak Minecraft Video Game Terminology in Novels and Original Video Games. In: L10N Journal 2(3), pp. 29–41.

Consistency Analysis of Slovak Minecraft Video Game Terminology in Novels and Original Video Games

Radka Filkorová

Comenius University in Bratislava

filkorova5@uniba.sk

Abstract

This article examines the Slovak translations of fiction based on the video game *Minecraft*, with a focus on the adherence to the game's official terminology. The study aims to determine whether the terminology from the original game has been consistently applied in the translated fiction. The article begins by outlining key theoretical concepts related to video game terminology before presenting an analysis of the selected sample. The findings are discussed in terms of the accuracy of Slovak translations in relation to the official *Minecraft* terminology across six books. The Slovak translators under review are Slavomír Hrivnák, Lukáš Ondrejkovič, and Šimon Kotvas. The central research question explores the extent to which the translators preserved the game's terminology in their translations of the fiction.

Keywords: video games, terminology, localization, translation, Minecraft

1 Introduction

The increasing popularity of video games has significantly impacted contemporary culture. Once considered primarily a pastime for children, video gaming has evolved into a widespread hobby among adults. Some video games have even influenced literary works, inspiring the creation of novels and other written content. One notable example is *Minecraft*, a sandbox game originally designed for younger audiences, which has gained global popularity and spurred a wide range of merchandise. Initially, books related to *Minecraft* were limited to instructional guides, helping players optimize their in-game experience. However, since 2017, the game's developer, *Mojang Studios*, has commissioned various authors to create narratives set within the *Minecraft* universe. As a result, approximately 20 story-driven books have been published to date.

Given *Minecraft*'s fanbase in Slovakia, the local market has quickly responded, with *Fragment* publishing house releasing seven Slovak translations of these novels since 2017. Additionally, a Slovak localization of the game is available.[1] However, unofficial translations[2] have emerged due to the need for timely updates, leading to variations from the official version. This article aims to investigate whether Slovak translators of children's literature based on *Minecraft* have remained faithful to the game's terminology or adopted a more flexible approach, potentially diverging from the in-game lexicon.

2 Theoretical background

The topic of specialized video game vocabulary has been extensively addressed in international research, including the works of Mangiron and O'Hagan (2013), Bernal-Merino (2015), and Adams (2010), among others. Méndez González also published an article on video game terminology (2019). In Slovakia, Kabát (2022) contributed an article on video game terminology and its associated neologisms. Additionally, Koscelníková (2024) explored gaming vocabulary in a broader discussion on the translation of video games.

Méndez-González (2019) emphasizes the importance of understanding the unique characteristics of video games to produce high-quality products. The mechanisms and technologies that underpin games are crucial not only for developers but also for translators. To effectively localize video games, translators must be well-versed in the specialized terminology associated with gaming, including the frequent use of neologisms, which are a key component of this lexicon.

Kabát (2022) identifies three key areas of terminology found in video games. The first is platform-specific terminology, which refers to terms associated with hardware developers (e.g., Microsoft, Sony, Nintendo) and hardware systems (such as smartphones and tablets). The second area is video game-specific terminology, encompassing terms related to game software, such as character traits, weapon types, vehicle types, and other in-game elements. Lastly, there is industry-specific terminology, which pertains to marketing materials and releases in press. A competent video game translator must be proficient in all three areas to ensure accurate and effective localization.

Similarly, Méndez-González (2019) identifies three main categories of specialized video game terminology: terminology of the platform, game terminology, and industry terminology. In addition, Méndez González (2019) highlights the role of interpreters,

[1] Slovak localization should be available in options after downloading the game from the official Minecraft website: https://www.minecraft.net/en-us.

[2] For example: https://crowdin.com/project/minecraft.

because they are a crucial part of international events in the video game industry. While localization allows translators some time for fact-checking, interpreters must react instantly and be well-versed in the specialized language of video games, which is arguably one of the most terminologically complex fields in the entertainment industry. Following Hasani-Yasin's (2010) classification, Méndez-González (2019) found that the seven groups of neologisms work excellently for the video game industry, but some categories need to be added. According to Méndez-González (2019), the categories of neologisms (including those inspired by Hasani-Yasin) are: **scientific neologisms** (when fictional discoveries (and their names) become reality, they can be a source for real-life terms), **political neologisms** (these are meant to create some kind of political or rhetorical point), **pop culture neologisms** (they originate in forums or in-game chats and with time become a part of general knowledge), **imported neologisms** (their origin is in different language but they are commonly used by users and developers – mostly English terms), **trademark-related neologisms** (hardware-related terms that become a part of general language), **nonce words** (these are created to have special effect in a precise moment), **inverted neologisms** (they arise after playing with words and spelling words backwards to create a new term), **new species neologisms** (some of these terms can be misleading, because they can have another meanings besides the meaning connected with the particular videogame), **weapon and skill-related neologisms** (some can be based on real-life words; unfortunately they sometimes also have a different translation in every new game that is released in the franchise), and **item-related neologisms** (these terms are connected with items or power-ups that players can use for gaining some kind of advantage in the game).

Kabát (2022) presents a detailed overview of video game terminology, using examples from a corpus of 42,058 words drawn from the video game *Minecraft*. The largest category identified in the corpus consists of terms related to in-game realia. These are further divided into subtypes: **names of new species** (e.g., the mineral *redstone*, translated into Slovak as *rudit*), **weapons/abilities** (e.g., *diamond sword*, translated as *diamantový meč)*, **items** (e.g., *skin* as *vzhľad*, or *dispenser* as *výdajný blok* in Slovak), **scientific terms**, and **political terms** (with the only example being *Capitol*, in Slovak as *Kapitol*). The second major category comprises terms related to game mechanics. In *Minecraft*, one of the most fundamental mechanics is the *builder plate* or *stavebná platňa*, without which the game would lose its core feature – constructing buildings from various materials. A useful distinction between realia and game mechanics terminology is that realia typically appear in the narrative aspects of the game, while game mechanics terms primarily pertain to gameplay features.

Following Kabát's (2022) classification, the third group encompasses scientific and technical terms, which are sometimes left untranslated in the Slovak localization of the game because players adopt them before official translations become available. In the *Minecraft* corpus, a notable example is the term *Augmented Reality*, or *obohatená realita*.

However, when abbreviated as *AR*, the term is generally retained in its English form. The classification also includes legal terminology, primarily found in video game licensing agreements, as well as terms related to trademarks and copyrights, such as the names of operating systems like Android or iOS. Additionally, pop culture terms appear in the *Minecraft* corpus, with examples such as *Kraken* or *Steampunk*. Kabát's classification is further expanded by unique neologisms created for specific moments in the game, intended to captivate the player but not frequently repeated. One such example in the Slovak localization of *Minecraft* is the compound word *kravovrah* (*cow tipper*). It should be noted that the boundaries between these categories of terms are fluid, and some terms could be classified into more than one group.

3 Analysis

During this research, an analysis was conducted on the official terminology used in the video game *Minecraft* and compared with the corresponding terms found in six Slovak translations of *Minecraft*-themed books. To gain further insight, the English originals of these books were examined to identify the specific terms with which the translators worked. The focus was on game-specific terms and those that appeared repeatedly across multiple Slovak translations. The goal was to determine whether these terms were consistently translated within individual books and across the translations as a whole. Additionally, the study aimed to assess whether the translators adhered to official game terminology or exercised greater creative freedom in their translation choices.

The main source for verifying the accuracy of the Slovak translations was the official Slovak translation of *Minecraft* terminology. Access to the official *Minecraft* terminology was gained from a translator who was involved in its localization but did not want to be named. The Slovak versions of the game terms from the official terminology were compared with the terms found in the fiction. Sometimes, the term was translated correctly (or according to the official terminology) on one page, and several pages later, it was translated differently. A list of inconsistently or incorrectly translated terms can be found in Appendix A. Additionally, an examination of the current Slovak version of the game revealed that, in some cases, it does not fully adhere to the official translations, as would be expected.

4 Research evaluation

This research identified a significant lack of consistency in the translation of key terms, many of which did not align with the official *Minecraft* terminology. Additionally, numerous typographical errors and other mistakes were present in the books, likely due to factors such as translator oversight, inadequate proofreading, or time constraints

during the publication process. Attempts were made to establish contact with the publishing house, but all efforts were unsuccessful. As it was not possible to establish contact with the two translators responsible for most of the examined books, the underlying causes influencing the final quality of the translations can only be hypothesized. Additionally, no information could be obtained regarding changes the made during the proofreading process, preventing an analysis of this aspect. Despite these issues, the analysis revealed several creative translation solutions. The research sample consisted of 72 terms, varying in frequency across six English Minecraft-themed books and their Slovak translations.

In the book *The Island (Ostrov)*, 31 of the 72 terms from the sample appeared in the section where the protagonist learned the names of *Minecraft* realia from books he found. Otherwise, the protagonist invented names for objects, as he did not know where he was. Of these terms, 14 were translated accurately, while the remaining 17 were either inconsistent or did not correspond to the official game terminology. Additionally, multiple spelling errors were observed. The term *spawner* also presented a challenge, as it does not appear in the official Slovak *Minecraft* lexicon. Consequently, its translation could not be definitively assessed as correct or incorrect. However, it was rendered inconsistently across the text, with variations such as *zjavovač* (revealer) and *továreň na príšery* (monster factory).

Table 1. ***The Island***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The Island	31	14	12	4	1

In the book *The Crash (Havária)*, 52 relevant terms were identified, 20 of which were accurately translated. The remaining 23 terms were either inconsistently translated or did not align with the official *Minecraft* terminology. Additionally, 6 terms were not present in the official terminology, preventing an assessment of their accuracy. One example is the term *Overworld*, which was translated as *Povrch*. Several typographical and spelling errors were also noted in this text. These errors further highlight the need for careful proofreading and attention to detail in the translation process.

Table 2. ***The Crash***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The Crash	52	20	23	3	6

In the book *The End (Koniec)*, 34 terms were identified, 14 of which were translated correctly. Seventeen terms were either mistranslated or displayed inconsistent usage, while 3 terms were absent from the official *Minecraft* terminology, making it difficult to

assess the accuracy of their translations. Notably, the translator adopted the most flexible approach in this book compared to the others analyzed. This is evidenced by several loose translations, such as rendering *enderman* as *fantóm Konca*, and a tendency to modify sentence structure – frequently splitting or merging sentences differently than in the original, with occasional omissions of parts of sentences.

Table 3. ***The End***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The End	34	14	14	3	3

In the book *The Lost Journals (Stratené denníky)*, 52 terms were analyzed, 13 of which were accurately translated. A total of 33 terms were either mistranslated or displayed inconsistency, while the accuracy of 6 terms could not be assessed, as they were not part of the official *Minecraft* terminology. Noteworthy in this book are the creative translation solutions, particularly in rendering the names of sheep and certain characters. However, the translation also contained several typographical errors and misspellings, likely due to lack of attention and time.

Table 4. ***The Lost Journals***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The Lost Journals	52	13	27	6	6

In the book *The Shipwreck (Vrak lode)*, 44 terms were analyzed, with 22 translated correctly and 21 either mistranslated or inconsistently translated, including 9 inconsistencies. One term could not be verified, as it was absent from the official *Minecraft* terminology. The translation also contained typographical and spelling errors. Lukáš Ondrejkovič and Šimon Kotvas are the translators of this work, and Šimon Kotvas was the only one who could be contacted. He explained that his translation aimed to adhere closely to the official game terminology, while ensuring that the dialogue reflected the speech patterns of older school-aged children. In addition to maintaining accuracy in the terminology, the translator had to convey the distinct styles of the four text components: the main narrative, the online chat communication between the characters, the wizard's riddles, and the video game software-related text.

Table 5. ***The Shipwreck***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The Shipwreck	44	22	12	9	1

In the book *The Voyage (Výprava)*, 34 terms were analyzed, 14 of which were translated correctly, while 18 terms exhibited inconsistencies or inaccuracies, including 12 incorrect translations. Two terms could not be definitively assessed, as they were not included in the official *Minecraft* terminology.

Table 6. ***The Voyage***

Name of the book	Number of terms	Translated correctly	Translated incorrectly	Translated inconsistently	Failed to determine
The Voyage	34	14	12	6	2

The results indicate that the books *The Crash* and *The Lost Journals* contained the highest number of terms, with 52 terms each, representing 72% of the total sample analyzed. In contrast, *The Island* had the fewest terms, with 31 terms, accounting for 43% of the sample. The highest number of accurately translated terms, consistent with the official *Minecraft* terminology, was found in *The Shipwreck*, where 22 terms were correctly translated, corresponding to 50% of the 44 terms in the book. The lowest number of correctly translated terms, 13, was observed in *The Lost Journals*, representing 25% of the total terms in that book.

Table 7. ***Comparison 1***

	Name of the book	Number of terms	Percentage	Base number (100%)
Most of the terms	*The Crash; The Lost Journals*	52; 52	72%	72
Most terms translated correctly	*The Shipwreck*	22	50%	44
Most terms translated incorrectly	*The Lost Journals*	27	52%	52
Most terms translated inconsistently	*The Shipwreck*	9	20.4%	44
Most terms absent in official terminology	*The Crash; The Lost Journals*	6	11.5%	52

The Lost Journals also had the highest number of incorrectly translated terms (27 terms), accounting for 52% of the total. The fewest incorrect translations were found in *The Shipwreck*, with 12 terms, representing 27.3% of the total. However, *The Shipwreck* also had the highest number of inconsistently translated terms, with 9 terms, equating to 20.4% of the total. Conversely, *The Crash* exhibited the lowest number of inconsistently translated terms, with only 3 terms, representing 5.8%. Finally, the highest number of

terms for which accuracy could not be verified, due to their absence in the official terminology, was found in both *The Crash* and *The Lost Journals*, with 6 terms each, corresponding to 11.5% of the total.

Table 8. ***Comparison 2***

	Name of the book	**Number of terms**	**Percentage**	**Base number (100%)**
Least of the terms	*The Island*	31	43%	72
Least number of correctly translated terms	*The Lost Journals*	13	25%	52
Least number of incorrectly translated terms	*The Shipwreck*	12	27.3%	44
Least number of inconsistently translated terms	*The Crash*	3	5.8%	52
Least of the terms	*The Island*	31	43%	72

5 Conclusion

This article explored the consistency of video game terminology between fiction and the original game, with a focus on six English *Minecraft*-themed books and their corresponding Slovak translations by Slavomír Hrivnák, Lukáš Ondrejkovič, and Šimon Kotvas. The research aimed to determine the extent to which translators adhered to the official *Minecraft* terminology in their Slovak translations. Prior to the analysis, a review of the relevant theoretical framework was concluded, including video game terminology theory and classification.

In the practical section, 72 terms from the selected books and the official *Minecraft* terminology, as well as the current version of the game, were analyzed to evaluate the consistency between translations and the game's lexicon. While some creative translation choices were observed, occasional typographical errors and inconsistencies in terminology were also identified, suggesting that insufficient attention was paid to maintaining uniformity across the translations. More general terms were translated consistently and mostly correctly, whereas more specialized terminology was translated more freely. However, it was not possible to determine whether this was due to the translators lacking access to the official terminology of the game, or whether they chose to ignore the established terminology. Findings of this research indicate a need

for greater rigor in the translation and proofreading process, as the current approach lacks consistency in terminology adherence.

To address these issues, it is proposed that translators gain access to the official game terminology from the publishing house and collaborate with previous translators to ensure consistency across works. Furthermore, translators should develop their own glossaries and actively engage with proofreaders to explain translation choices, contributing to the entire translation process until publication. Raising awareness about video game localization and enhancing the quality of Slovak localizations could also contribute to improved consistency between fiction and original game terminology. While the establishment of a specialized publishing house dedicated to video game-inspired literature could potentially solve these challenges, it would require significant time and financial resources to compete in the current commercial landscape.

Filkorová, Radka. 2024. Consistency Analysis of Slovak Minecraft Video Game Terminology in Novels and Original Video Games. In: L10N Journal 2(3), pp. 29–41.

Gameography

Minecraft. 2011. USA: Mojang Studios.

Bibliography

Adams, Ernest. 2010. Fundamentals of Game Design. Berkeley: New Riders.

Baptiste, Tracey. 2018. Minecraft: The Crash. New York: Del Rey.

Baptiste, Tracey. 2018. Minecraft: Havária. Bratislava: Fragment. Translation: Lukáš Ondrejkovič.

Bernal Merino, Miguel Ángel. 2015. Translation and Localisation in Video Games: Making Entertainment Software Global. New York and London: Routledge.

Brooks, Max. 2017. Minecraft: The Island (The Zombie Survival Guide: Complete Protection from the Living Dead). New York: Del Rey.

Brooks, Max. 2017. Minecraft: Ostrov. Bratislava: Fragment. Translation: Lukáš Ondrejkovič.

Fry, Jason. 2020. Minecraft: The Voyage. New York: Del Rey.

Fry, Jason. 2020. Minecraft: Výprava. Bratislava: Fragment. Translation: Slavomír Hrivnák.

Hasani-Yasin, Ahmed. 2010. Neologism as a Linguistic Phenomenon in Mass Media Textbook with Reference to Translation. In: Journal of Historical and Cultural Studies. 2(6): pp. 243-264.

Kabát, Marián. 2022. Pár poznámok k terminológii a neologizmom v lokalizácii videohier. In: Nová filologická revue. 14(1): pp. 28–40. https://www.ff.umb.sk/app/cmsFile.php?disposition=a&ID=24306. Accessed on: 20 February 2024.

Koscelníková, Mária. 2024. Lokalizácia videohier na Slovensku. Nitra: Univerzita Konštantína Filozofa v Nitre.

Lafferty, Mur. 2019. Minecraft: The Lost Journals. New York: Del Rey.

Lafferty, Mur. 2019. Minecraft: Stratené denníky. Bratislava: Fragment. Translation: Slavomír Hrivnák.

Lee, C. B. 2020. Minecraft: The Shipwreck. New York: Del Rey.

Lee, C. B. 2022. Minecraft: Vrak lode. Bratislava: Fragment. Translation: Lukáš Ondrejkovič, Šimon Kotvas.

Méndez González, Ramón. 2019. Specialized Terminology in the Video Game Industry: Neologisms and their Translation. In: Vertimo studijos: 12: pp. 71-85. https://www.researchgate.net/publication/338839530_Specialized_Terminology_in_the_Video_Game_Industry_Neologisms_and_their_Translation. Accessed on: 15 January 2024.

O'Hagan, Minako, and Mangiron, Carme. 2013. Game Localization: Translating for the global digital entertainment industry. Amsterdam: Benjamins Translation Library.

Valente, Catherynne M. 2019. Minecraft: The End. New York: Del Rey.

Valente, Catherynne M. 2022. Minecraft: Koniec. Bratislava: Fragment. Translation: Slavomír Hrivnák.

Filkorová, Radka. 2024. Consistency Analysis of Slovak Minecraft Video Game Terminology in Novels and Original Video Games. In: L10N Journal 2(3), pp. 29–41.

Websites

Official website of Minecraft. https://www.minecraft.net/en-us. Accessed on: 10 December 2024.

Minecraft localization on Crowdin. https://crowdin.com/project/minecraft. Accessed on: 10 December 2024.

Appendix A

Table 9. ***Term translation comparison***

Term	Official Slovak translation	Other translations in the books	Slovak translation in the current version of the game
Beetroot	Cvikla	repa	repa
Blaze powder	prášok zo žiarivca	prach ohniváka	prach ohniváka
Boat	Čln	loď	čln
Bucket	Vedro	vedierko	vedro
Chest	Truhlica	truhla	truhlica
Crafting table	pracovný stôl	pracovný stolík; remeselný stôl; remeselnícky stôl	pracovný stôl
Creeper	Creeper	sliedič	creeper
Emerald	Smaragd	zafír	smaragd
Enchantment table	čarodejný stôl	čarovací stôl; zaklínačský stôl; stôl na očarovávanie; čarovný stôl	stôl očarovania
Ender chest	truhlica z Konca sveta	koncotruhla; endertruhlica; všadetruhla	Ender truhlica
Ender dragon	drak z Konca sveta	drak; drak záhuby; Ender drak; posledný drak; drak konca	Ender drak
Enderman	Enderman	fantóm konca; fantóm Endu; fantóm	enderman
Endermite	Endermit	koncormit	endermit

Furnace	Pec	vyhňa; ohnisko	pec
Ghast	Ghast	prízrak; mŕtvolák	ghast
Glowstone	žiarivý kameň	svietikameň	žiarivec
Lapis lazuli	Ultramarín	lapislazuli; lapis lazuli; lazurit	lazurit
Mob	Tvor	tlupa (príšer); banda; partia; (zvieracia) mafia; entita	tvor
Monster	Príšera	monštrum	príšera
Mooshroom	Mooshroom	hríbokravy; kravohríby	mooshroom
Nether	Nether	Podsvetie	Nether
Nether quartz	netheritový kryštál	podsvetný kremeň	Nether kremeň
Pressure plate	prítlačná doska	nášľapná doska; tlakový spínač	nášľapná doska
Redstone	Rudit	červenokameň; červený kameň	redstone
Redstone torch	ruditová fakľa	fakľa z červenokameňa; pochodeň z červenokameňa	redstonový prach
Sandstone	Pieskovec	vápenec	pieskovec
Silverfish	Rybenka	švehla; striebroryba; strieborník	švehla
Skeleton	Kostlivec	lukostrelec; kostra	kostlivec
Soul sand	pohyblivý piesok	piesok duší; prízračný piesok	piesok duší
Squid	Kalmár	sépia; kalamár; chobotnica	chobotnica
Stick	Palica	palička	palica
Survival mode	režim prežitia	mód prežitia; modus prežitia; mod prežitia	režim prežitia
Torch	Fakľa	pochodeň	fakľa

Village	Dedina	osada	dedina
Villager	Dedinčan	osadník	osadník
Witch	Striga	čarodejnica; ježibaba	bosorka
Wither effect	Chradnutie	efekt withera; chronické chradnutie	chradnutie
Wither skeleton	Wither kostlivec	witherkostlivec; uschnutý kostlivec	Wither kostlivec
Workbench	pracovný stôl	stôl; pracovný stolík	pracovný stôl
Zombie	Zombie	zombie; zombia; zombík; zombi	zombie

Hudáková, Zuzana. 2024. Comparing the Efficiency of Source Text Pre-editing vs. Machine Translation Post-editing. In: L10N Journal 2(3), pp. 42–59.

Comparing the Efficiency of Source Text Pre-editing vs. Machine Translation Post-editing

Zuzana Hudáková

Comenius University in Bratislava

zuzana.hudakova@uniba.sk

The publication of this article was financially supported by the Scientific Grant APVV-23-0539: Future-Proofing Translators and Interpreters: Establishing a Knowledge Platform for Language Professionals in Slovakia.

Abstract

As machine translation (MT) becomes increasingly embedded in professional workflows, researchers explore ways to improve quality and efficiency. Although neural MT systems like DeepL and Google Translate improve fluency, they still require human intervention. Two key strategies are pre-editing (PrE), which modifies the source text before MT to reduce errors, and post-editing (PoE), which refines MT output to meet quality standards.

This study compares PrE and PoE in MT workflows through a controlled experiment involving 20 translation students. One group used PoE alone, while the other combined PrE and PoE. Translation quality was assessed using the TAUS Dynamic Quality Framework, with time efficiency also analyzed.

Findings show PoE alone accelerates the process but increases error rates, particularly in accuracy and fluency. PrE enhances translation quality by reducing errors and cognitive load during PoE, though it requires more time upfront. The combination of PrE and PoE produced the highest-quality translations, suggesting that integrating PrE improves accuracy and consistency. These results highlight the importance of combining human expertise with MT to improve workflows, balancing speed and quality in professional translation.

Keywords: machine translation, pre-editing, post-editing, translation quality, efficiency in translation workflows

Hudáková, Zuzana. 2024. Comparing the Efficiency of Source Text Pre-editing vs. Machine Translation Post-editing. In: L10N Journal 2(3), pp. 42–59.

1 Introduction

As the translation industry grapples with growing demands for rapid and cost-effective solutions, MT has become an essential technology. Recent advancements in neural MT systems like DeepL and Google Translate have improved the fluency and efficiency of machine-generated translations. However, these systems are still not reliable enough to produce error-free results, making human intervention indispensable. In this context, two key strategies come into play: pre-editing (PrE) and post-editing (PoE).

PrE involves modifying the source text prior to MT to improve clarity, consistency, and compatibility with machine processing, thus reducing output errors. Conversely, PoE takes place after MT, where human translators refine the machine-generated text to meet quality standards. Each strategy has a distinct impact on the efficiency of the translation process and quality of the translation.

This study assesses the effectiveness of PrE and PoE in MT workflows, with a specific focus on determining which approach, or their combination, leads to higher-quality translations while optimizing speed. PrE has the potential to reduce the cognitive load in the PoE phase by producing clearer, more machine-compatible texts, but requires considerable investment of time in advance. On the other hand, PoE alone can speed up the translation process, but it often leads to more extensive revisions. Understanding these compromises is essential for balancing speed, accuracy, and fluency in professional translation.

This research holds particular relevance in light of the increasing reliance on MT tools within the translation industry, which aims to address escalating demands for efficiency without sacrificing quality. Translators often face the challenge of managing a large volume of content within tight deadlines while maintaining high-quality standards. Consequently, it is critical to identify the most effective applications of PrE and PoE to enhance translation workflows. Furthermore, as MT systems continue to advance, the role of human expertise in the oversight and enhancement of machine-generated translations remains a vital area for further investigation.

To examine these issues, this study uses a controlled experimental design involving 20 translation students divided into two groups: one group focuses exclusively on PoE and the other group utilizes both PrE and PoE. Participants' performance, in terms of translation accuracy and time efficiency, will be analyzed to assess the impact of each method.

2 Pre-editing and post-editing in translation

PrE is a strategic practice combining human expertise with machine efficiency to adapt source text for easier MT. The aim is to eliminate challenges for MT systems such as odd phrases, idioms, and typographical errors (Kokanova et al. 2022; Vieira 2019). PrE

involves editing the text based on certain guidelines, such as shorter sentences, simplified grammatical structures, and consistent terminology, which reduces the cognitive load of MT and results in clearer and more accurate translations (Arenas 2020). Optimizing the source text helps prevent errors and misinterpretations, as MT systems still struggle with semantic subtleties (Yang 2023).

PrE is a proactive approach to translation, as it optimizes the source text to improve MT output. While PrE does not rely on technological progress in principle, its role has become more relevant as MT systems benefit from clearer and more structured input. This is a step towards optimizing translation processes and creating source texts that are more suited to MT systems, thus reducing the need for extensive PoE. Studies show that PrE has significant impact on MT quality (Bounaas 2023; Mercader-Alarcón & Sánchez-Martínez 2016), with improvements like lower word error rates and fewer necessary corrections. However, PrE alone cannot prevent all errors without risking grammatical issues in the source text (Mercader-Alarcón & Sánchez-Martínez 2016).

On the other hand, PoE involves correcting MT output to meet certain language and style standards (Arenas 2020). Human translators refine machine-generated text, not only correcting errors but also ensuring that the content aligns with the audience's preferences and contextual needs. This collaborative process, known as the human-in-the-loop (HITL) approach, integrates human expertise into MT-driven translation workflows to improve quality and adaptability. In HITL frameworks, human feedback is essential to train and fine-tune MT models, so that they can adapt to specific linguistic nuances and cultural contexts. This symbiotic relationship leverages the efficiency of AI while retaining the accuracy and cultural relevance that only human translators can provide. By strategically allocating resources and scaling human involvement, organizations can ensure that MT systems improve over time, delivering translations of high-quality standards (Yang et al. 2023). The efficiency of PoE depends on factors such as the quality of the MT system, the complexity of the source text, and the expertise of the post-editor (Yang, 2023).

PoE is a dynamic and cognitively demanding process, particularly for texts with many stylistic elements, which sometimes requires more effort than MT-free translation (O'Brien 2022). Depending on the project objectives, organizations can choose between a light PoE that fixes only significant errors, and full PoE that ensures publication-ready quality (Daems et al. 2017). Full PoE involves addressing all linguistic issues, including cultural appropriateness (Vieira 2019).

PrE and PoE can be studied through the Skopos and functionalist theories, as they involve strategic interventions for translation optimization. However, MT limits the translator's role, reducing functional adaptation compared to traditional workflows. Their integration reflects a shift in translation practices to accommodate MT, however, the impact of PrE and PoE remains debated. Some argue that MT and PoE restrict

functional adaptation by limiting broader structural or stylistic changes, while others view them as necessary adjustments to industry demands. This debate aligns with Toury's (1995) concept of evolving norms, though further research is needed to assess their long-term impact.

Integrating PrE and PoE into translation workflows addresses both machine and human limitations, thus highlighting the collaborative nature of translation by combining human expertise with machine capabilities to produce translations that are accurate and functional.

3 Methodology

In this study, a comparative design was used to evaluate the efficiency and quality of two translation approaches: PrE followed by MT and PoE, compared to MT and PoE only. A review of previous studies highlights the increasing role of MT in translation workflows and its impact on human intervention. Calvo-Ferrer (2023) conducted a study on the ability of viewers to distinguish between machine-generated and human-translated subtitles, focusing on the implications of MT quality and audience perception. The study, which involved 119 translation students assessing ChatGPT-generated subtitles versus human translations, found that participants were generally unable to distinguish between the two, though lower-quality subtitles were more frequently attributed to non-human translation. These findings suggest that while MT has improved in fluency and readability, it still presents challenges in accuracy, particularly with complex linguistic features such as humor, cultural references, and idiomatic expressions. The study also indicates that translation expertise plays a role in identifying MT-generated content, as advanced students were more successful in distinguishing between human and machine outputs. These insights align with the current study's focus on the interplay between MT, human intervention through PrE and PoE, and the role of translation training in optimizing workflow efficiency and quality.

20 participants were recruited from university-level translation and interpreting programs in Slovakia, specifically from the Comenius University in Bratislava and the Matej Bel University in Banská Bystrica. Participants ranged from second to fifth year of study, with second-year students enrolled in bachelor's programs and the remaining participants pursuing their master's degrees. These participants were randomly assigned to two groups of ten to ensure representation from various stages of academic training. Participants were asked about their previous experience with PrE, PoE, and MT tools to assess how familiarity with these processes might impact translation quality and efficiency. The survey also collected information about whether they attended MT and PoE courses, which provided further insight into the impact of formal training on translation effectiveness.

The translation task involved translating a 370-word excerpt from a washing machine instruction manual, selected due to its technical complexity and linguistic challenges. The source text was in English, and participants translated it into Slovak. The text contained detailed instructions, safety warnings, and technical descriptions that are representative of the precision required in technical translations. The selection of a user manual was deliberate, as such documents require accuracy to ensure safety and clarity for end users. The text presented several challenges to both human translators and MT systems. It contained subordinate clauses, ambiguous wording, and misspelling errors, such as "The sensor automatically detects the quantity of a Detergent put by a user and the temperature and the quality of water to make the best washing algorithm for washing and rinsing." These problems are typical of the technical documentation and should reveal differences in translation quality between the two groups. Ambiguous terms such as "nails" and "downs" posed additional challenges and required careful handling in both PrE and PoE phases.

The translation results of both groups were assessed using the TAUS Dynamic Quality Framework (DQF), an established error typology that categorizes translation errors according to terminology, grammar, fluency, and style (TAUS 2017). This framework, which has been harmonized with the Multidimensional Quality Metrics (MQM) to form the DQF-MQM standard, enabled a comprehensive assessment of the quality of the translations produced by both workflows.

Two workflows were defined for the study: Group 1 (G1) was instructed to translate the source text using the NMT tool DeepL, followed by PoE to refine the translation output for accuracy, fluency, and coherence. Group 2 (G2) was asked to pre-edit the source text before MT. The PrE guidelines were designed to improve the clarity of the source text before it was processed by the MT system (Annex 1). Participants were instructed to shorten sentences, correct punctuation and spelling errors, and standardize terminology to improve translation consistency. After PrE, G2 used DeepL for translation and then post-edited the MT output.

DeepL was selected based on the findings of Petráš and Munková (2023), which highlighted its superior performance in English to Slovak translation compared to other tools such as ChatGPT. While ChatGPT is a large language model rather than a dedicated NMT tool, its translation capabilities are increasingly being integrated into professional workflows, making comparisons with specialized NMT tools relevant for evaluating practical translation quality. Additional support came from Agung et al. (2024), who demonstrated the effectiveness of DeepL in translating synthetic languages – such as Slovak and Indonesian – surpassing Google Translate in both accuracy and fluency.

To examine the relationship between translation quality and efficiency, the time taken by each participant for each phase (PrE, MT, and PoE) was recorded and correlated with

the error data. Analyzing the results with a focus on whether participants took a Machine Translation course provides a nuanced view of the impact of structured training on translation efficiency. While participant experience (year of study) was considered in assessing translation quality, its impact on task completion time was not explicitly measured. However, given that more experienced students might require less time, this remains an important variable for further investigation.

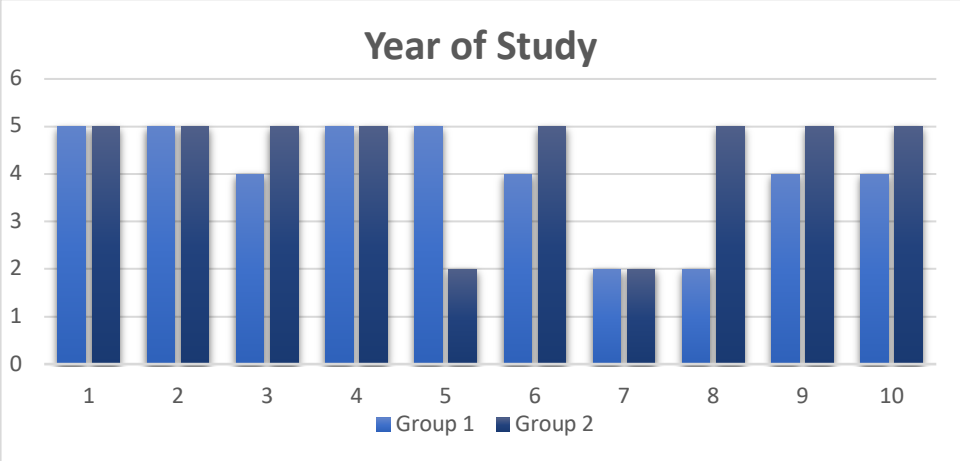
Descriptive statistics and correlation analysis were used to examine whether PrE had a measurable impact on reducing errors and improving translation speed. This approach also allowed for an examination into whether participants with prior education and practical experience in using MT systems performed better in either workflow.

This study acknowledges several limitations. The sample size of 20 participants limits the generalizability of the results, and the focus on a technical text may not reflect the broader challenges faced by other translation areas such as literary or legal translation. Additionally, differences in participants' experience with MT tools and their familiarity with PrE and PoE processes may have influenced the outcomes. Future research could expand the participant pool and diversify the text types to provide more informed conclusions about the applicability of PrE and PoE in different translation contexts.

4 Results

The analysis of G1 and G2 shows clear differences in efficiency and translation quality, which are influenced by both the academic training of the participants and the workflows used. G1, which focused solely on PoE, included a majority of participants in their 4th and 5th years of study, with 80% of participants being in these advanced stages. G2, with 80% of participants in their 5th year, had a slightly more experienced cohort overall. While this suggests that G2 may have had greater familiarity with PrE practices and translation technologies, no specific control was applied to ensure an equal distribution of experience across both groups. Therefore, although the workflow itself may have played a role in the observed differences, the impact of academic experience should be considered as a potential influencing factor (Fig. 1).

Figure 1. ***Participants: Year of Study***



In terms of time efficiency, G1 completed PoE tasks in an average of 28 minutes and 52 seconds (Fig. 2), However, this speed came at the expense of quality, as the average TAUS score was 13.9, reflecting higher error rate, particularly in categories such as accuracy and fluency. G2, which performed both PrE and PoE, took significantly longer – an average of 51 minutes and 19 (Fig. 3) – but their quality metrics were significantly better, with a TAUS score of 6.8, indicating higher quality with fewer errors.

Figure 2. ***G1: Time Spent PoE***

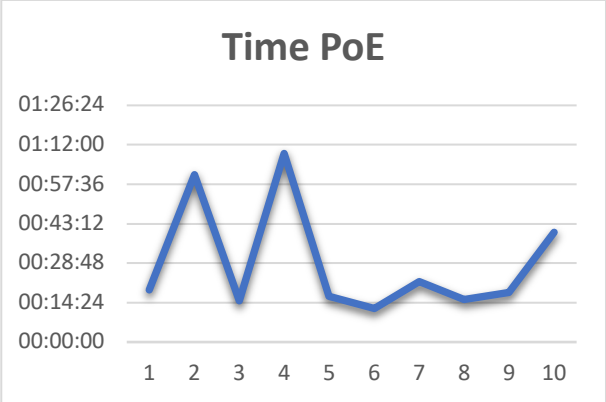
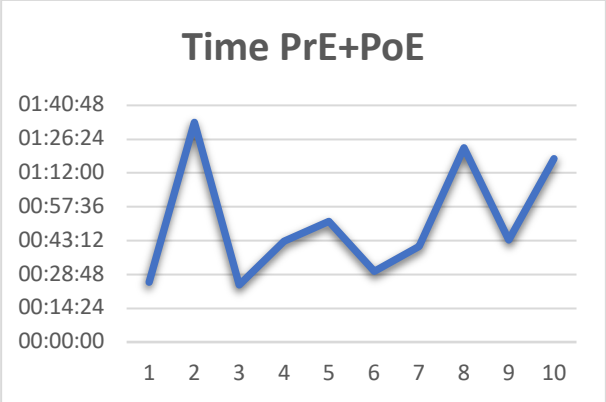
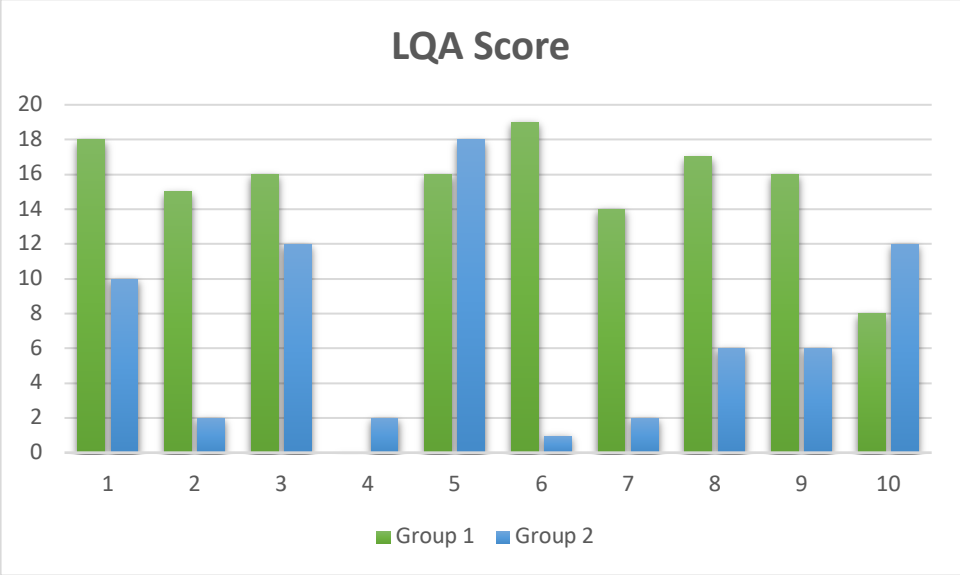


Figure 3. ***G2: Time Spent Total***



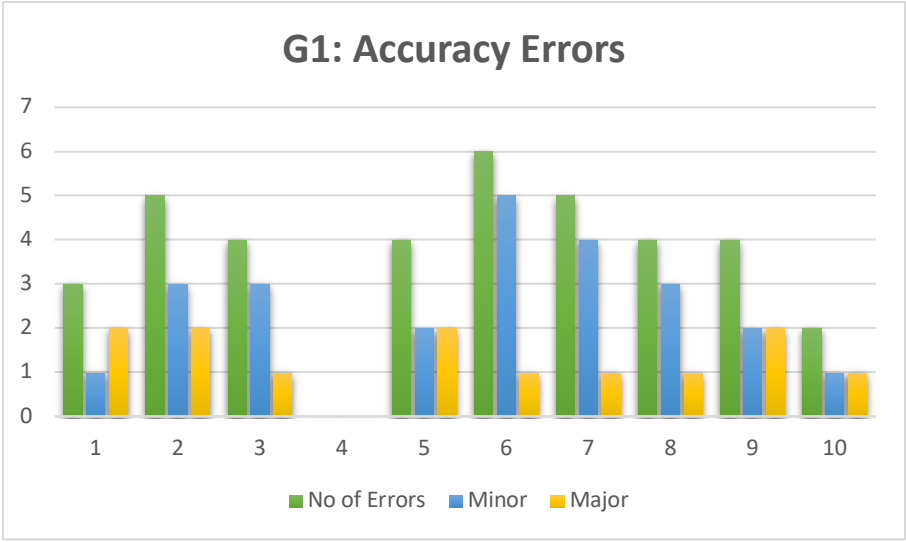
The results show that PrE can help reduce errors by addressing problems before MT, thereby improving the overall accuracy and coherence of the final output. While G1 completed their task faster, the higher frequency of major errors in accuracy, fluency, and terminology emphasizes the limitations of relying solely on PoE. In contrast, G2's PrE step appeared to reduce cognitive load during PoE, resulting in improved quality (Fig. 4).

Figure 4. ***LQA Scores***



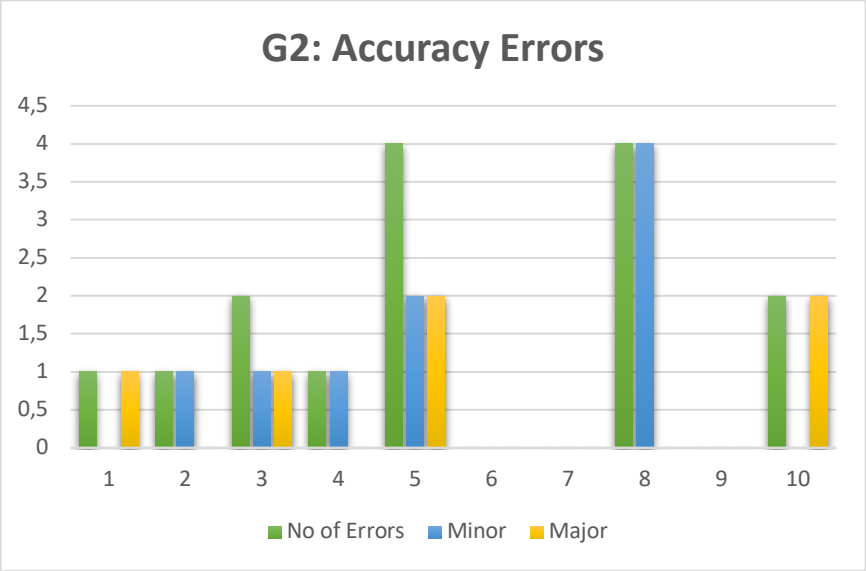
Accuracy was a crucial factor in the evaluation, with errors categorized by TAUS into subcategories such as Addition, Omission, Mistranslation, Over-translation, and Untranslated segments. G1's results showed a correlation between PoE experience and accuracy. Participants with extensive PoE training, such as Participant 4, recorded zero accuracy errors, suggesting that prior experience can help mitigate the risks associated with MT outputs. While this may be linked to specific PoE training, it could also reflect general translation experience, individual diligence and quality-consciousness, as more advanced translators tend to develop stronger revision and error-detection skills. However, participants without such training, like Participant 7, recorded five errors, emphasizing the challenges of dealing with machine-translated text without prior experience (Fig. 5).

Figure 5. ***G1: Accuracy Errors***



G2, which incorporated PrE, demonstrated superior accuracy performance overall. Participants with experience with both PrE and PoE, like Participant 6, achieved flawless results with zero accuracy errors. The structured approach to PrE enabled better control of translation quality and highlighted the importance of addressing potential issues in the source text before MT. Those with less experience in either PrE or PoE, such as Participants 5 and 8, recorded higher error rates, underscoring the role of comprehensive training in optimizing accuracy (Fig. 6).

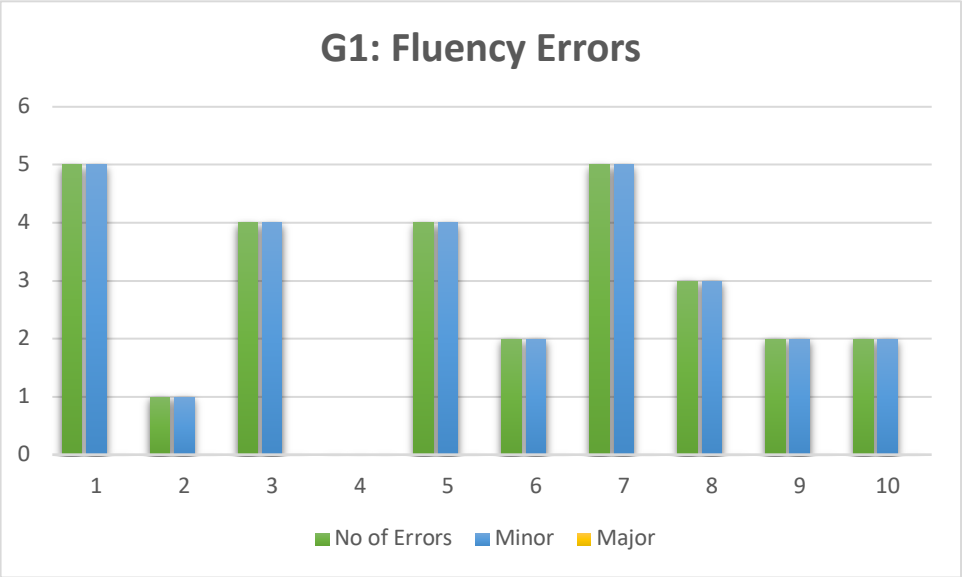
Figure 6. ***G2: Accuracy Errors***



A comparison of accuracy results between groups shows that PrE offers a clear advantage, particularly in reducing major accuracy-related errors.

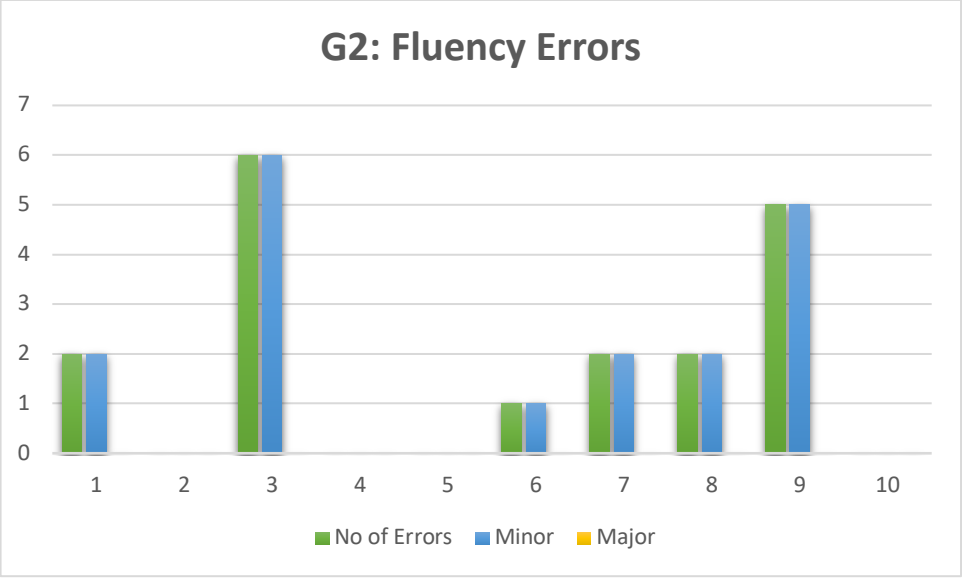
Another key dimension was fluency, which was assessed using subcategories such as Grammar, Punctuation, and Spelling. The performance of G1 in fluency varied greatly. Participants with PoE experience, such as Participant 4, produced translations with no fluency errors, though this may also reflect general translation experience rather than PoE training alone. However, those without such experience, like Participant 7, recorded a higher fluency error count, particularly in grammatical structure and punctuation (Fig. 7).

Figure 7. ***G1: Fluency Errors***



Including PrE in G2 resolved some fluency issues before they became problematic for MT, however, the data shows that fluency in translation is influenced by a combination of factors. Here, participants without previous PoE experience (5, 7, 8) recorded fewer fluency errors than their more experienced counterparts. (Fig. 8). The different results among participants with similar backgrounds suggest that individual skills, the specific nature of the translation tasks, and possibly the type of text may also play a crucial role.

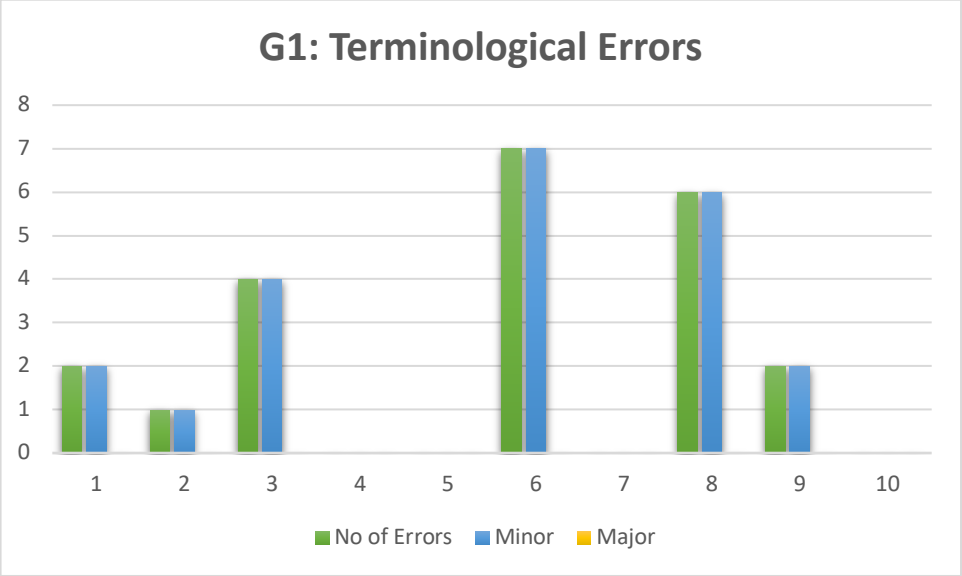
Figure 8. ***G2: Fluency Errors***



The use of specialized terminology was a central focus, particularly given the technical nature of the source text. G1's results showed that participants with PoE experience generally performed better in maintaining terminological consistency. For example,

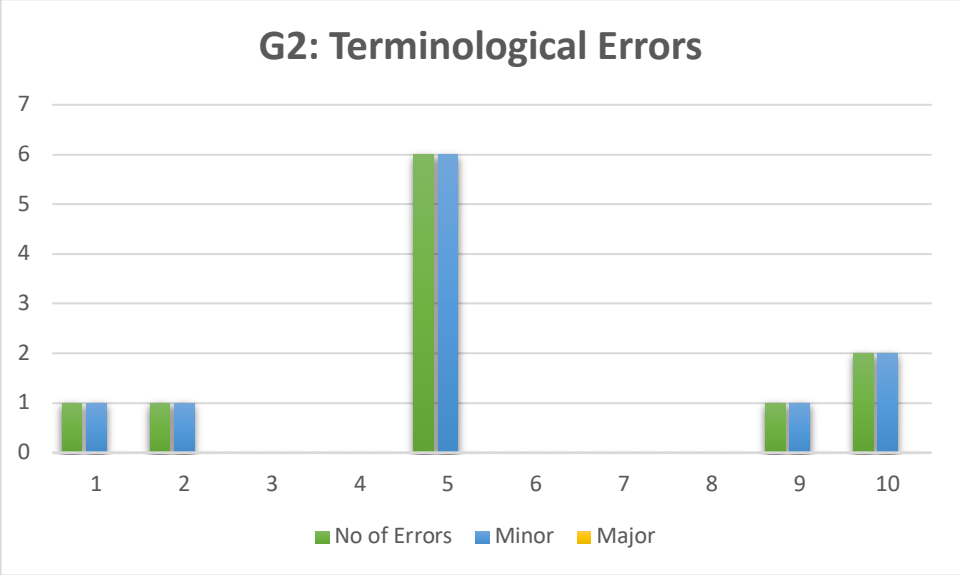
Participant 4, who had extensive PoE experience, recorded no terminological errors, while participants with less experience, such as Participant 6, experienced more issues adhering to the provided glossary (Fig. 9).

Figure 9. ***G1: Terminological Errors***



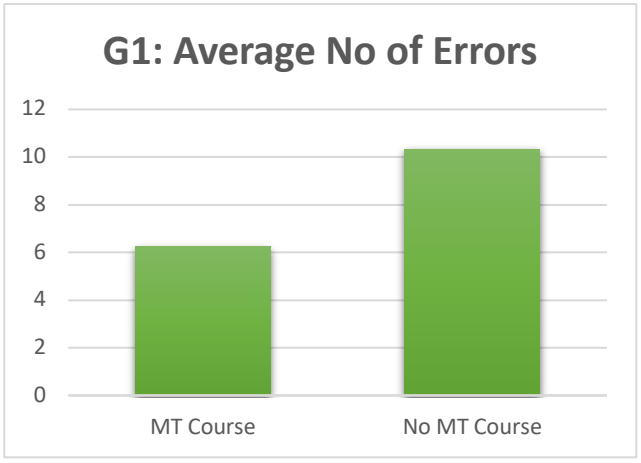
In G2, PrE allowed participants to standardize key terms before translation, resulting in fewer terminology errors during the PoE phase. Participants 3, 4, 6, 7 and 8 recorded an error-free performance, which could illustrate the advantages of addressing terminological consistency through PrE, taking into account the methodological limitations of this research. This step not only simplified the PoE process but also ensured that technical terms were handled correctly from the start (Fig. 10). The results indicate that PrE is especially effective in terminology management, particularly in technical translations where adherence to glossaries and terminological accuracy are crucial.

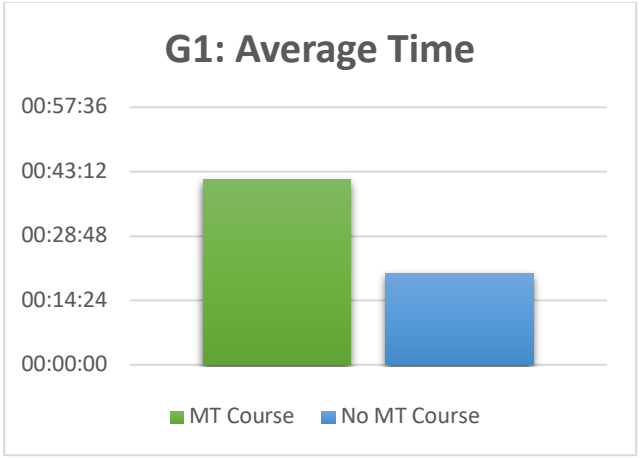
Figure 10. ***G2: Terminological Errors***



Analyzing the results with a focus on whether participants took a Machine Translation course provides a nuanced view of the impact of structured training on translation results. In G1, participants who completed the MT course generally showed mixed results in terms of errors (Fig. 11). It is noteworthy that Participant 4, who attended the course, made no errors, but spent over an hour on PoE, which may indicate thoroughness and application of the techniques learned.

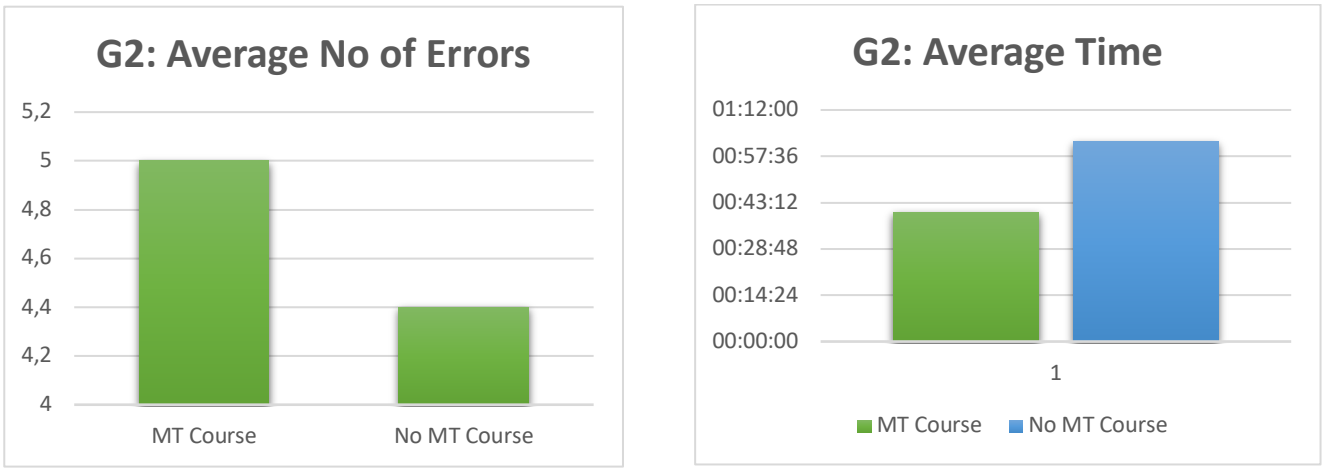
Figure 11. ***G1: Impact of MT course***





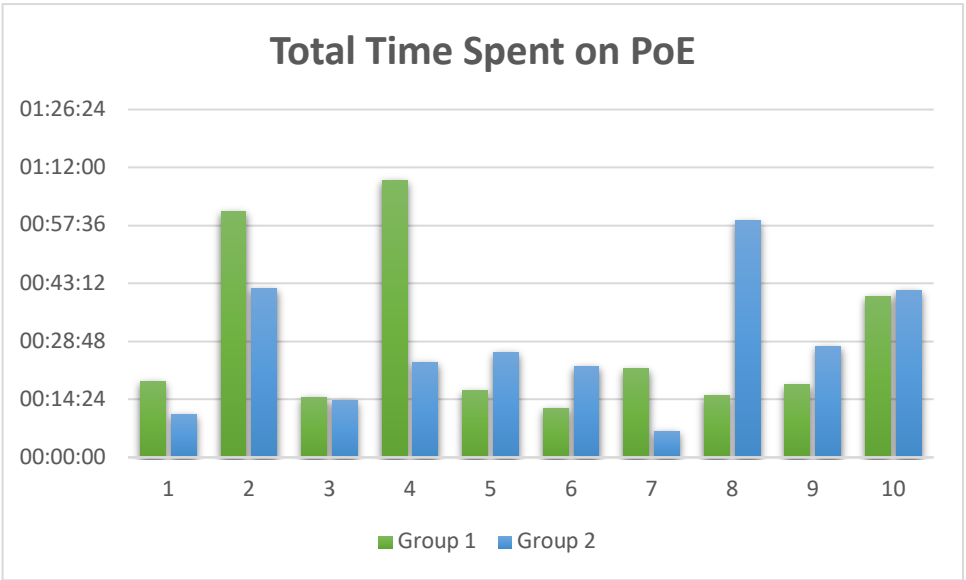
In contrast, G2 shows that those who had taken the MT course and engaged in PrE generally had better control over time management than their counterparts who had not taken the course. However, the untrained participants had a lower overall error rate (Fig. 12).

Figure 12. ***G2: Impact of MT course***



This indicates that although training speeds up the translation process compared to the untrained counterparts in this group, the error rate does not follow a straightforward pattern, suggesting that other factors might influence the quality of the output besides the MT training. Participants who took both MT and PoE courses generally demonstrated a more deliberate and thorough approach, reflected in longer completion times but fewer errors. In G1, trained participants spent an average of 41 minutes and 27 seconds on their task, compared to untrained participants, who completed the task faster but with more errors (Fig. 11). In G2, time efficiency was better balanced (Fig. 12), with trained participants completing tasks faster than their untrained counterparts, however, the error rate does not follow a clear pattern.

Figure 13. ***Total Time Spent on PoE***



5 Discussion

The comparative analysis of G1 and G2 offers valuable insights into the efficiency and quality dynamics of translation workflows when integrating PrE alongside PoE into MT processes. Considering that experience levels among participants varied, the findings of this study should be interpreted as hypotheses rather than definitive conclusions.

The results suggest that although PrE requires additional time investment, it may significantly improve translation accuracy and reduces the PoE workload, ultimately leading to potentially more efficient and higher-quality translation outcomes.

In G1, where participants relied solely on PoE machine-translated texts, significant differences in efficiency were observed. Participants without formal training in MT completed tasks faster, but at the expense of higher error rates. In contrast, participants with MT training – such as Participant 4, who recorded no errors – took longer but produced a significantly more accurate translation. These findings suggest that MT training may improve quality but could also lead to longer task completion times, as participants with more experience tend to invest additional effort into refining machine-generated outputs.

On the other hand, PoE alone can result in faster workflows, although the increased error rates suggest that lower-quality outputs may require more extensive corrections during the PoE phase itself, potentially offsetting the initial time savings. This underlines the trade-off between speed and accuracy, as higher error rates can lead to a more labor-intensive PoE process to achieve acceptable quality levels. These results align with findings by Sanchez-Torron and Koehn (2016), who observed that the quality of MT output directly impacts PoE efficiency, with lower-quality outputs requiring more effort and time during PoE.

G2, which used both PrE and PoE, showed a clear dependency between time and accuracy. Although the PrE phase increased the total time spent on translation tasks, the accuracy improvements were significant. For example, Participant 6 recorded only one error, illustrating that high levels of accuracy can be achieved when PrE is combined with MT training. This result calls attention to the importance of using PrE as a means to improve translation accuracy by addressing problems in the source text prior to MT. However, given the variation in participants' levels of experience, this trend should be explored further before drawing definitive conclusions. The study by Bounaas et al. (2023) supports this conclusion, as it found that PrE significantly improves the accuracy, appropriateness, and acceptability of translated texts.

Although PrE lengthens the initial phase of translation, it could provide strategic benefits for improving overall workflow efficiency, as it simplifies the source text, thereby reducing the cognitive load during PoE. This is particularly evident in the faster PoE times recorded by G2 (Fig. 13). By eliminating complex structures and ambiguities

in the source text during PrE, these participants were able to complete the PoE phase faster and demonstrated that PrE helps mitigate the typical challenges during PoE. This supports the assumption that PrE, when combined with PoE, can compensate for the additional time required upfront by optimizing the latter phase of the workflow.

The relationship between speed and quality, a well-documented phenomenon in translation workflows, is further confirmed by this study. Faster translation workflows, particularly those that skipped PrE, were often associated with higher error rates, as seen in G1. This trade-off between speed and quality reflects a common challenge in the translation industry, where time constraints can lead to a decline in translation accuracy and coherence. In contrast, G2 produced more accurate and consistent translations, highlighting the effectiveness of a more deliberate and structured approach. This result suggests that while fast translation is often a priority to meet tight deadlines, it can come at the expense of quality, especially for complex or technical content that requires precision.

In addition to improving accuracy and fluency, PrE also seems to contribute to improved terminological consistency. The ability to standardize terminology during PrE significantly reduced the likelihood of errors in PoE, further streamlining the translation process. This suggests that PrE is not just a time-consuming step but may serve as a valuable strategy for reducing PoE effort and improving overall translation quality. While the study suggests that PrE may enhance translation accuracy and streamline PoE efforts, further research is needed to confirm these trends across different translation tasks and professional contexts.

6 Conclusion

The main objective of this study was to determine whether PrE combined with PoE, or PoE alone results in a more effective translation process. The results show that PrE may improve translation quality by minimizing the need for extensive PoE, leading to potentially more accurate and consistent outputs. However, the study also confirms that while PoE alone is faster, it can result in lower translation accuracy and fluency.

The results show that PrE could help reduce errors by addressing problems before MT, thereby improving the overall accuracy and coherence of the final output. However, it is important to note that this study focuses on the efficiency of PrE and PoE within MT workflows, rather than comparing them to a fully human translation process. Since participant experience levels were not fully homogeneous, the results should be understood as indicative of possible trends rather than broadly generalizable conclusions.

A human-only workflow would introduce additional variables that are not directly comparable to MT-assisted workflows, making such a comparison beyond the scope of

this study. However, future research is needed to determine whether the lower error count observed with PrE results from the actual benefits of PrE or simply reflects the translator's deeper understanding of the source text before MT. A comparative study could assess whether reading and familiarizing oneself with the source text before translation – without explicitly performing PrE – yields similar improvements in translation quality.

Additionally, in professional settings, PrE is typically performed by source text authors, technical editors, or dedicated language professionals, rather than by the same individuals responsible for PoE. In this study, participants performed both PrE and PoE to ensure a controlled comparison of workflow efficiency and quality. This methodological choice may not fully reflect industry practices. Future research could explore how PrE affects translation quality when performed by different professionals within the workflow and whether its impact is distinct from the natural cognitive processing that occurs when translators engage with the source text before MT. Such studies could provide a clearer understanding of the specific contribution of PrE to overall translation accuracy and efficiency.

A closer look at the data shows that the combined use of PrE and PoE may deliver the most favorable results, especially for participants with experience in both areas. This combined approach resulted in the lowest error rates, which could indicate that expertise in both PrE and PoE improves the overall efficiency of the translation process. While this view is widely supported by language service providers and translation companies, some professional translators remain skeptical, arguing that MT can introduce errors that require extensive revision, potentially negating its efficiency benefits (Alvarez-Vidal et al. 2020; Cadwell et al. 2018). However, given the variability in participant experience, further research is needed to confirm whether this effect is consistent across different professional and educational contexts.

The study suggests that while neither PrE nor PoE alone consistently outperforms the other in every context, a combined approach that leverages the strengths of both methods may offer the best solution for achieving high-quality translations. The integration of PrE and PoE not only has the potential to improve translation fluency, accuracy, and consistency in terminology, but may also enable a smoother PoE phase by reducing the complexity of MT outputs. However, this approach requires a higher initial time investment, especially in the PrE stage. Looking forward, the study highlights the importance of translation training programs to equip future translators with MT and PoE skills. By fostering a deeper understanding of how MT tools can be effectively integrated with human editorial skills, translators can produce higher quality translations more efficiently. A strategic combination of PrE and PoE could therefore represent the most effective path forward for maximizing the benefits of MT in professional translation.

Hudáková, Zuzana. 2024. Comparing the Efficiency of Source Text Pre-editing vs. Machine Translation Post-editing. In: L10N Journal 2(3), pp. 42–59.

Bibliography

Agung, I. et al. 2024. Translating Performance of Google Translate and DeepL in Translating Indonesian Short Stories into English. Proceedings of the 2nd Linguistics, Literature, Culture, and Arts International Seminar (LITERATES). Universitas Mahasaraswati Denpasar. pp. 178-185. https://e-journal.unmas.ac.id/index.php/literates/article/view/8555/6443. Accessed on: 8 December 2024.

Alvarez-Vidal, S. et al. 2020. Post-editing for Professional Translators: Cheer or Fear? In: Tradumàtica Technologies de la Traducció, 18: pp. 49-69. https://doi.org/10.5565/rev/tradumatica.275

Arenas, A. G. 2020. Pre-editing and Post-editing. Bloomsbury Academic eBooks. London. https://doi.org/10.5040/9781350024960.0019.

Bounaas, C., Zemni, B., Shehri, F. A., and Zitouni, M. 2023. Effects of PrE operations on audiovisual translation using TRADOS: An experimental analysis of Saudi students' translations. Texto Livre. 16: pp. 1-5. https://doi.org/10.1590/1983-3652.2023.45539.

Cadwell, P. et al. 2018. Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. Perspectives. 26(3): pp. 301-321. https://doi.org/10.1080/0907676X.2017.1337210.

Calvo-Ferrer, J. R. 2023. Can You Tell the Difference? A Study of Human Vs Machine-translated Subtitles. Perspectives. 32(6): pp. 1115-1132. https://doi.org/10.1080/0907676x.2023.2268149.

Daems, J., Vandepitte, S., Hartsuiker, R. J., and Macken, L. 2017. Translation Methods and Experience: A Comparative Analysis of Human Translation and Post-editing with Students and Professional Translators. In: Meta. 62(2): pp. 243-484. https://doi.org/10.7202/1041023ar.

Kokanova, E. et al. 2022. Pre-editing English News Texts for Machine Translation into Russian. In: Language Studies and Modern Humanities. 4(1): pp. 25-30. https://doi.org/10.33910/2686-830x-2022-4-1-25-30.

O'Brien, S. 2022. How to Deal with Errors in Machine Translation: Post-editing. In: Kenny, D. (ed.), Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence. (Translation and Multilingual Natural Language Processing. 18). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6653406.

Petráš, P., and Munková, D. 2023. Strojový preklad založený na neurónových sieťach – sľubná cesta prekladu z analytických jazykov do flektívnej slovenčiny?. In: Slovenská reč. 88(1): pp. 74-89. ISSN 0037-6981.

Quah, C. K. 2006. Translation and technology. London: Palgrave Macmillan UK eBooks. https://doi.org/10.1057/9780230287105.

Sanchez-Torron, M., and Koehn, P. 2016. Machine Translation Quality and Post-Editor Productivity. In: Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track. Austin: The Association for Machine Translation in the Americas. pp. 16-26.

TAUS. 2017. Quality Evaluation using an Error Typology Approach. De Rijp: TAUS BV.

Vieira, L. N. 2019. Post-editing of machine translation. In: O'Hagan, M. (ed.), The Routledge Handbook of Translation and Technology. London and New York: Routledge. pp. 337-354. https://doi.org/10.4324/9781315311258.

Yang, X., et al. 2023. Human-in-the-loop Machine Translation with Large Language Model. arXiv preprint. https://doi.org/10.48550/arXiv.2310.08908.

Hudáková, Zuzana. 2024. Comparing the Efficiency of Source Text Pre-editing vs. Machine Translation Post-editing. In: L10N Journal 2(3), pp. 42–59.

Yang, Y., Liu, R., and Qian, X. et al. 2023. Performance and perception: machine translation post-editing in Chinese-English news translation by novice translators. In: Humanities and Social Sciences Communications. 10: 798. https://doi.org/10.1057/s41599-023-02285-7.

Annex 1

Suggested PrE rules:

- reduce sentence length
  - e.g., The software, which was developed to help with budgeting and has been used by many people since its launch last year, can also assist in tracking expenses effectively. -> The budgeting software, launched last year, also tracks expenses effectively.
- unify terminology if it is inconsistent
  - e.g., The handbook mentions guidelines on staff conduct, employee behavior, and worker regulations. -> The book mentions guidelines on employee conduct, behavior, and regulations.
- correct spelling and punctuation errors
  - e.g., The childrens' toys were scattered all over the living room floor. -> The children's toys were scattered all over the living room floor.
- simplify grammatical structures
  - e.g., There is a need for managers to be able to understand the data that is presented to them. -> Managers need to understand the presented data.
- remove ambiguities
  - e.g., He saw the man with a telescope. -> Using a telescope, he saw the man.

Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

Machine Translation Quality Based on TER Analysis from English into Slovak

Matúš Nemergut

Comenius University in Bratislava

nemergut.matus@gmail.com

Abstract

Translators are facing increased demand for their services and are increasingly required to utilize tools that can help them save time and increase their efficiency. Neural machine translation (NMT) has become the leading technology in the translation industry, and its utilization promises efficiency gains. However, this is not as straightforward as it may seem, and actual efficiency gains depend on a number of variables: the quality of the NMT output, the translator's skills, time, and the effort the translator expends on post-editing. This paper aims to analyze the number of edits required for the NMT output to meet quality requirements and determining the acceptability threshold of the neural machine translation output for post-editing based on this number. From a methodological perspective, the study uses TER, an automatic machine translation evaluation metric calculating the smallest edit distance required, to assess the number of edits needed. By analyzing samples from two experiments, it was found that with the TER score between 39% and 42.5%, i.e., when 39-42.5% of the machine translation output needs editing, post-editing ceases to be beneficial, and it is more efficient for the machine translation output with such score to be translated from scratch.

Keywords: post-editing, automatic machine translation evaluation metrics, TER, acceptability threshold for post-editing

1 Introduction

Modern technological advances are influencing and changing the set of demands users have for technology. It is expected to be innovative, fast, high quality, and to get even more innovative, faster, and of higher quality with every update. In translation technology, the past three decades have brought about some significant developments: statistical machine translation (SMT), computer-assisted translation (CAT) tools, and, most recently, neural machine translation (NMT), which, since its introduction in 2016, has been the most dominant among machine translation (MT) paradigms (Rothwell et al. 2023). Munková et al. (2019, in: Hudecová et al., 2021) note that the market has also

changed. Over the last decade, demand in the translation services sector has grown so significantly that it has far outstripped its supply. The increased demand for translation services places an invisible burden on translators to increase their efficiency and thus meet the demands of the market. The use of NMT can bring about efficiency increases, but the actual increase is dependent upon the quality of the output as well as the time and effort required to edit it. This work aims to investigate the effort required to edit an NMT output and to determine its acceptability threshold for post-editing based on TER analysis.

2 Machine Translation Post-Editing (MTPE)

Hudecová et al. (2021) note that the use of machine translation is becoming commonplace today, not only in the sphere of providing language services but also among non-professionals. The significant increase in demand in the translation industry over the last decade, which exceeded its supply (ibid.), underlines the fact that the question of machine translation is shifting from whether to use it to how best to use it, according to Koehn (2020, in: Hudecová et al. 2021). Hudecová et al. (2021) continue: "The suitability of utilizing machine translation is relative" (p. 193), and if we want it to be of high quality, accuracy, and fidelity, it needs to undergo further editing, i.e., *post-editing*.

Vieira (2019) explains that post-editing as an activity has been carried out since the beginning of the existence of MT technologies, but only in recent years it has evolved significantly as a service, practice, and research topic. In the early days, post-editing was seen as a step towards closing the gap between MT's purpose — a fully automatic translation of high quality — and its far insufficient output; for a long time, a model prevailed in which people supported the machine rather than the other way around. Vieira adds that there has been a significant improvement in MT integration in the field of professional translation in recent years, especially regarding CAT tools, changing the model of human-supported MT to computer/machine-supported human translation. However, he also sees the introduction of many technologies into CAT tools to have entailed some blurring of the boundaries between technology and support in the translation process, resulting not only in terminological inconstancy but also in ambiguity in post-editing procedures and tasks.

2.1 Error typology

The primary paradigm in the MTPE process is *static*—that is, MT is generated first and then it is statically edited in a separate step (Vieira, 2019). During editing, the post-editor works with three texts: (i) the source text, (ii) the MT output that is not

processed, and (iii) the MT output that is post-edited, i.e., the target text (Pavlíková 2022).

Situating MTPE within the models of the translation process, Absolon (2018) notes that MTPE is a combination of plain translation and revision (dual). This is due to the nature of errors, which require the post-editor to be flexible and able to shift from a process to process. After reading the MT output, the post-editor must assess its quality to know whether to edit the output or re-translate it. The ability to make quick decisions is therefore key — the work of a post-editor is always about deciding whether to continue with a given process or shift to another one.

The most common actions in MTPE are related to correcting punctuation, word order, number and gender correspondence, or incorrectly translated expressions (Pavlíková, 2022).

For the alternative paradigm in the MTPE process, Vieira (2019) uses the term *interactive*: there is interaction between the translator and the MT system during the production of the target text. He explains that such interaction may look like MT is completing or predicting the text of the human translation as it is being written or reciprocally reacting to and learning from the translator's edits on the fly and adds that while no tendency for post-editor's actions to speed up has been observed in static MTPE, there is such a tendency in interactive MTPE. At the same time, however, he notes that most productivity-focused research only marginally examines the quality of MTPE products and that the results are no longer clear-cut there, because while some research speaks of increased quality of interactive MTPE over static MTPE, other speaks otherwise.

2.2 MTPE vs. From-Scratch Translation

As such, MTPE research only addresses the issue of MTPE product quality in a complementary way in the context of comparing the productivity of from-scratch translation with MTPE (Vieira, 2019). Despite some negative perception of MT in the marketplace (Pavlíková 2022), Vieira (2019) writes about Screen's 2019 research (in: Vieira 2019), comparing a from-scratch translation with a post-edited translation, which concludes that the products of both procedures are largely equivalent; thus, along with other research, supporting the use of MT in professional practice. He also notes that with the advent of NMT, which is currently the cutting-edge MT technology, come the challenges of its post-editing: the higher fluency of the outputs of these systems makes identification of errors and their correction in mono- and bilingual texts more difficult. In a study comparing post-editing of phrase-based SMT and NMT, Yamada (2019, in: Vieira 2019) found that although the output of NMT contained fewer errors and the result of its post-editing was of higher quality, translation students achieved a lower error correction rate during post-editing but put more effort into it.

Vieira adds that research by Jia et al. (2019, in: Vieira 2019) concludes that, depending on the genre of the text, NMT post-editing may require less effort; with no difference in the quality of the target text between NMT post-editing and from-scratch translation, post-editing proved faster only in specialized texts.

Hudecová et al. note that "post-editing represents a specific skill" (2021, p. 195) and it may not be the case that a professional translator is also good at post-editing and vice versa (ibid.). Vieira (2019) recommends post-editing to be preferably done by professional translators, as those less experienced will hardly be aided by NMT in improving their performance and reminds that forming a proper understanding of the role of MT in professional translation is the task of translation study programs.

MTPE also raises the question of how to measure the efforts made by the post-editor. According to Koponen (2016, in: Hudecová et al. 2021), 3 indicators were defined for this purpose:

- *temporal indicator;*
- *technical indicator*—measured using automatic metrics;
- *cognitive indicator*—i.e., the efforts that the post-editor perceives to have made.

Although MTPE as an activity is in some cases demonstrably more productive, there is no known way by which it would be possible to determine "where post-editing is worthwhile and where it is not" (Hudecová et al. 2021, p. 195).

2.3 Translation Edit Rate (TER) and Translation Efforts

As mentioned above, three indicators have been defined to measure translation efforts, namely temporal, technical, and cognitive.

The aspect of time is prominent today, especially in the commercial sphere — do Carmo (2020), drawing on ISO 18587:2017, writes about time as one of the key determinants of MTPE. And since it is relatively easy to measure, it is at the center of much of the translation effort measurements. It should not, however, be left there alone. Since MTPE involves revision of the text, it is important to understand the effort required for the changes made. This is addressed by the technical indicator. When referred to, it can be a measurement of the actions performed with the mouse and keyboard on the one hand. These are recorded using various tools. At the same time, the technical indicator is the edit distance that needs to be performed on the hypothesis. That's where TER is utilized, and TER scores are calculated for these purposes. The lower the TER score, the less effort is required to post-edit a given MT output. For example, if the score is 0.4, i.e., 40%, it means that 40% of the MT output had to be changed to make it satisfactory. The third indicator is cognitive, and the most difficult to measure. It involves cognitive processes that cannot be directly seen or measured because they take place in the brain. These include reading, understanding and comparing the text, or following post-

editing guidelines. For instance, measuring the length of eye fixation or pupil dilation is one way of measuring the cognitive effort expended in MTPE. According to some, MTPE generates a greater cognitive load than translation itself; thus, it is important to take the cognitive dimension of MTPE into account when measuring translation efforts (O'Brien, 2022).

The methodological focus of this work is the evaluation of post-editing effort based on TER, therefore, the measurement of cognitive effort is not within the scope of this work.

3 Similar research

Given the complexity of examining translation efforts, one can encounter a variety of methods used in practice. The intention is to determine the threshold of usability and efficiency of MTPE in relation to from-scratch translation through TER analysis.

Guerrero (2020) works with the hypothesis that 50% edit distance is too high as an MT output acceptability indicator. Moreover, the results of her research show that at an edit distance between 30% and 40%, on a scale[3] of 1–4, MT outputs are already mostly rated as 2 by professional post-editors. Based on the comments of these raters, 2 is closer to the unacceptability of the MT output.

Also based on interviews with R. Tihlárik (2023), who at the time had been the director of a translation and localization services provider for 26 years, it can be stated that the current trend in the localization industry is for clients to consider the TER score of 50% as the limit; this means that the translator has to edit up to half of the entire MT output during MTPE, which in many cases can be more laborious than translating the text from scratch, but they get paid an MTPE rate that is partial to the rate for from-scratch translation.

Research conducted by e.g. Gueberof (2008, 2012), Guerrero (2003), or O'Brien (2006) on SMT has already shown a positive effect of MTPE on translator's productivity compared to translating from scratch or editing so-called "fuzzy matches"[4] and, in some cases, also on the quality of the final product (Temizöz 2013; O'Brien 2006).

In order to build on these conclusions today, it is necessary to consider the research findings of Koponen (2016), in which she emphasizes the existence of a relationship between the effort expended in MTPE and the characteristics of the source text or the MT error rate, the influence of the type of editing on the overall effort expended in MTPE, or the varying speed of post-editors themselves. She further talks about the trend to perceive the editing effort of longer sentences as higher, even if the number of

[3] A Likert scale was used, i.e., a scale used to measure behavior, opinions, and attitudes (scribbr.com, 2020).
[4] *Fuzzy match* represents a condition where a segment of the source text is partially identical to another already translated segment in the translation memory (thelanguageoflocalization.com, 2018).

edits is relatively low, or the effect of sentence length on the time required to post-edit it.

Stefaniak (2020) notes that although it might seem obvious that MT output with lower TER score will also require less time to post-edit and vice versa, this is not necessarily the case. This is a complex issue, and the research results are not sufficient to draw conclusions.

For example, Krings (2001, in: Temizöz 2013) has observed that the effort expended in post-editing a medium-quality MT output is higher than the effort expended in post-editing a low-quality MT output; he's attributed this to the fact that a medium-quality MT output contains a large number of elements that need to be extensively compared between the source text, the target text, and the MT output. He's compared post-editing of a low-quality MT output to the process of standard human translation, in which only the source and target texts are worked with, resulting in a lower cognitive load (Temizöz 2013).

It is important to note that the aforementioned research has been carried out on SMT systems; but as has already been mentioned, NMT systems produce smoother outputs, and thus, based on Yamada's observation (2019, in: Vieira 2019) that "although the output of NMT contained fewer errors and the result of its post-editing was of higher quality, translation students achieved a lower error correction rate during post-editing, but put more effort into it", it can be inferred that Krings' results (2001, in: Temizöz, 2013) can be applied to NMT as well.

Stefaniak's (2020) research on the English-Polish language pair has also shown no clear correlation between the TER score and the time required for post-editing a given MT output. These results are in line with the findings of Gaspari et al. (2014, in: Stefaniak 2020), who have noted that if there is a correlation between TER, but also BLEU or METEOR, and the time required for MTPE, it is only very weak (Stefaniak 2020).

However, in the same research, Krings (2001, in: Temizöz 2013) also found that MT outputs with higher quality assurance (QA) scores were post-edited faster (Temizöz 2013).

Additionally, O'Brien (2011), on the basis of her research, also on SMT, has preliminarily concluded that there is a correlation between TER and MTPE productivity and that TER can be a good indicator of MTPE speed for a set of segments.

Thus, Stefaniak's (2020) research has indeed revealed a correlation: the correlation between the TER score and productivity gains. In the research conducted, with a mean TER score of 0.39 (39%) and a median TER score of 0.375 (37.5%), the average MTPE speed vs. the average from-scratch translation speed was as follows: MTPE = 0.325 words/second (the median 0.295 words/second); from-scratch-translation = 0.215 words/second (the median 0.205 words/second). These results are in line with the

research of de Gibert Bonet (2018, in: Stefaniak 2020), who has set the TER score threshold up to which productivity gains occur at 0.33 (i.e., 33%) for the Spanish-English language pairs (Stefaniak 2020).

In their research on the English-Spanish language pairs, Parra Escartín and Arcedillo (2015) have reported an increase in MTPE productivity relative to from-scratch translation for the TER score ≤ 0.21 (i.e., 21%). However, for this research, it is important to note that only two translators were observed, the test set also contained 75%–100% fuzzy matches from a translation memory, and the MT tool they used to generate the machine translation was their home-grown tool that they had used and fine-tuned for three years (Parra Escartín and Arcedillo, 2015).

4 Methodology, research questions, and research methods

As mentioned previously, measuring the effort expended in MTPE consists of several aspects. As the methodological focus of this work, the technical aspect examined through TER analysis was chosen, and the goal was to determine the threshold of the acceptability of MT output for post-editing. In other words, the research tries to determine:

- what is the TER score at which it is more efficient to translate the source text from scratch than to post-edit its machine-translated form.

Zhechev (2014) notes that when comparing from-scratch translation and translation using MT, it is important to set a baseline. This is not easy if the translators do not translate and post-edit the same segment.

In the experiments, MA students majoring in philology with a specialization in translation and interpreting at Comenius University did not translate and post-edit the same segments because there would have been no way of solving the problem of them remembering what had already been translated/post-edited. Therefore, each of them translated one part of the text and post-edited the other. This eliminated the differences that might arise from the different levels of experience of the post-editors with MTPE and allowed for comparison of the productivity rates of each separately.

At the same time, according to Roberts (2007, in: Temizöz 2013), comparing from-scratch translation and MTPE is only justified if comparing outputs of the same person (Temizöz, 2013).

After defining the objectives and setting the measurement parameters, a decision was made to conduct two experiments: a pilot one—on a smaller sample and with students many of whom had no previous experience with MTPE—and a main one—on a slightly

larger sample and with students enrolled in Machine Translation Post-Editing course, in the middle of semester.

For the first experiment, an e-learning text from the environment of a transportation company was selected. The text was deemed adequate as it represented a common localization task with the need for MTPE, was heterogeneous with regard to its structure (running text, titles, numbering, paragraphs, choice selection, repetitions), and included different terms and abbreviations. It was slightly adapted so that it did not contain any elements characteristic of the industry that could reduce its comprehensibility. A glossary of terms to be adhered to was prepared, and the post-editors were instructed on which terms were designated as DNT *(Do Not Translate)*. The text contained a total of 516 words divided into 53 segments. The text with the glossary was then uploaded to the CAT tool Phrase, which the university had a license for and was accessible to all students. In Phrase, the text was translated by an MT system, in this case DeepL, and two files were created from the translated text: one preserved the MT output in the first half, this half was intended for MTPE, and the translation of the second half, which was intended for from-scratch translation, was deleted. In the second file, it was done the other way around to obtain data related to post-editing of the entire text. For each student, a separate project was prepared in Phrase, containing a source text with one half ready for MTPE and the other for from-scratch translation, and an attached glossary. In addition to translating from scratch and doing MTPE, the task of the post-editors was to record the time taken for each activity. After completing one part, they were not able to return to it again, thus the times they recorded were final. Nine post-editors took part in this pilot experiment.

For the second experiment, a support article for an iPhone stuck during transfer from a previous iPhone was chosen. The text had to be different, as some of the post-editors took part in both experiments. This time the text was rather homogeneous with regard to its structure but required greater attention to the syntax due to its instructive nature. The text was not edited, nor was it necessary to create a glossary for it. The text contained 566 words divided into 33 segments. The procedure was the same and a project was prepared for each post-editor. The experiment took place on university grounds during a class on Machine Translation Post-Editing (a compulsory elective course for MA students majoring in philology with a specialization in translation and interpreting, taught at the Department of British and American Studies at Comenius University). It was attended by 12 post-editors, all students of this course. The experiment was conducted in the middle of the semester; all post-editors had already had some experience with MTPE. In this case, they recorded the time themselves as well. They were not able to return to the completed task, the times were final.

The second step was to analyze the data using the TER metric. The program for TER analysis, along with the expertise required for its evaluation, was provided by exe a.s.

localization. Segments that were post-edited were analyzed and their TER scores were determined. Since the TER score of each segment is calculated based on the word count, it is important to factor this in when calculating the mean TER score of the entire post-edited section. Therefore, the weighted mean of each segment was calculated and exe's instructions for TER analysis evaluation was further followed. A table was created and the results of TER analysis for individual segments were entered into the table, along with the results for the entire post-edited part. Subsequently, both the individual times the post-editors took to perform MTPE and the times they recorded when translating from scratch were added into the table. These were compared and used to determine whether, in some cases, MTPE took more time than-from scratch translation, and if so, what was the TER score at which it happened. When evaluating the main experiment, an interesting phenomenon was observed: the results of one of the post-editors, who is known, to be actively working as a translator while still studying, differed substantially from all other post-editors. Thus, one more research question was formulated: Will the results of a professional post-editor be similar to the results of this experienced student post-editor? The professional post-editor was chosen on the basis of their familiarity with the translated subject. They proceeded in the same way: they post-edited and translated the text from the second experiment, i.e., the same text the experienced student post-editor post-edited and translated, and they recorded their times.

4 Machine Translation Quality Based on TER Analysis

Firstly, the outputs of the pilot experiment were analyzed, in which 9 post-editors participated. Five of them did MTPE first (they post-edited the first half of the text, which contained 230 words of MT output) and then they translated the second half of the text from scratch. The remaining four translated the first half of the text from scratch first and then they did MTPE (they post-edited the second half of the text, which contained 216 words of MT output). The procedure of having part of the post-editors post-edit one half of the text and the rest post-edit the other half was chosen in order to get as comprehensive a view of MTPE and from-scratch translation as possible.

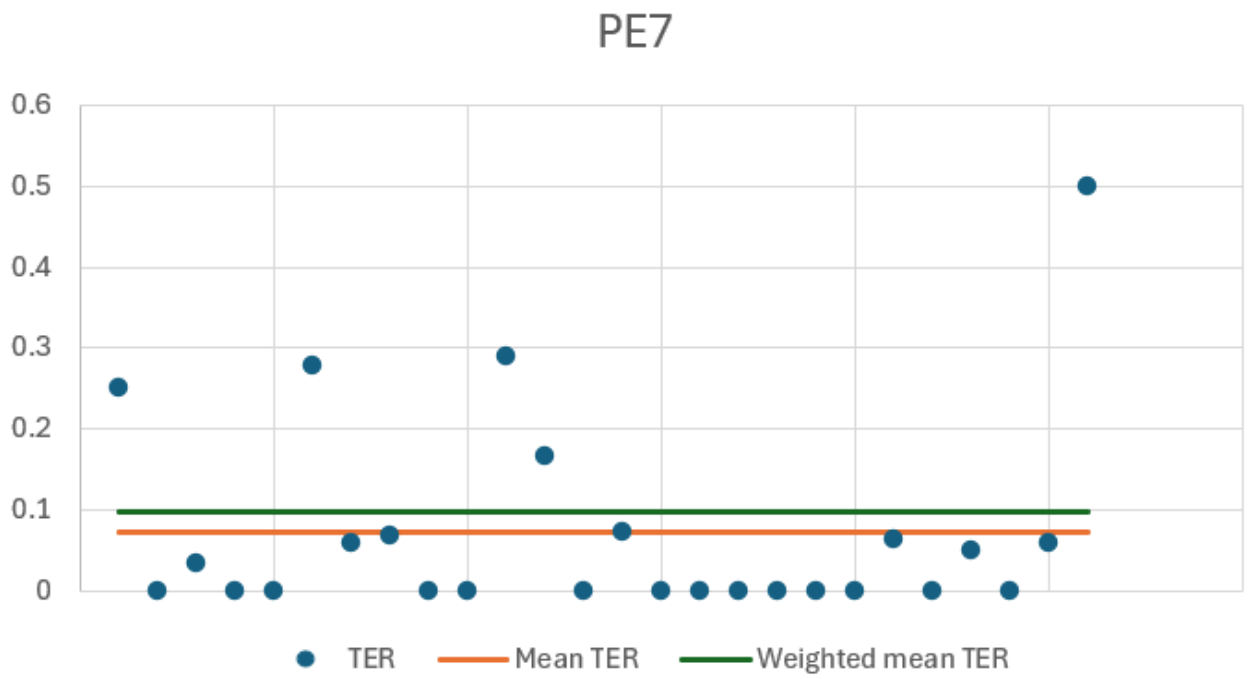
From-scratch translation was not faster than MTPE in any of the 9 cases. The variance of from-scratch translation (shown in Table 1) ranged from 20 minutes and 38 seconds to 1 hour, 1 minute, and 9 seconds; the variance of MTPE ranged from 10 minutes and 3 seconds to 31 minutes and 52 seconds.

Table 1. ***Results overview (pilot experiment): time***

Post-editor	MTPE	From-scratch
PE1	21 min 57 sec	44 min 53 sec
PE2	10 min 42 sec	61 min 9 sec
PE3	31 min 52 sec	43 min 44 sec
PE4	15 min	40 min
PE5	26 min	36 min
PE6	19 min 43 sec	43 min 27 sec
PE7	10 min 3 sec	27 min 34 sec
PE8	17 min 32 sec	54 min 26 sec
PE9	11 min	20 min 38 sec

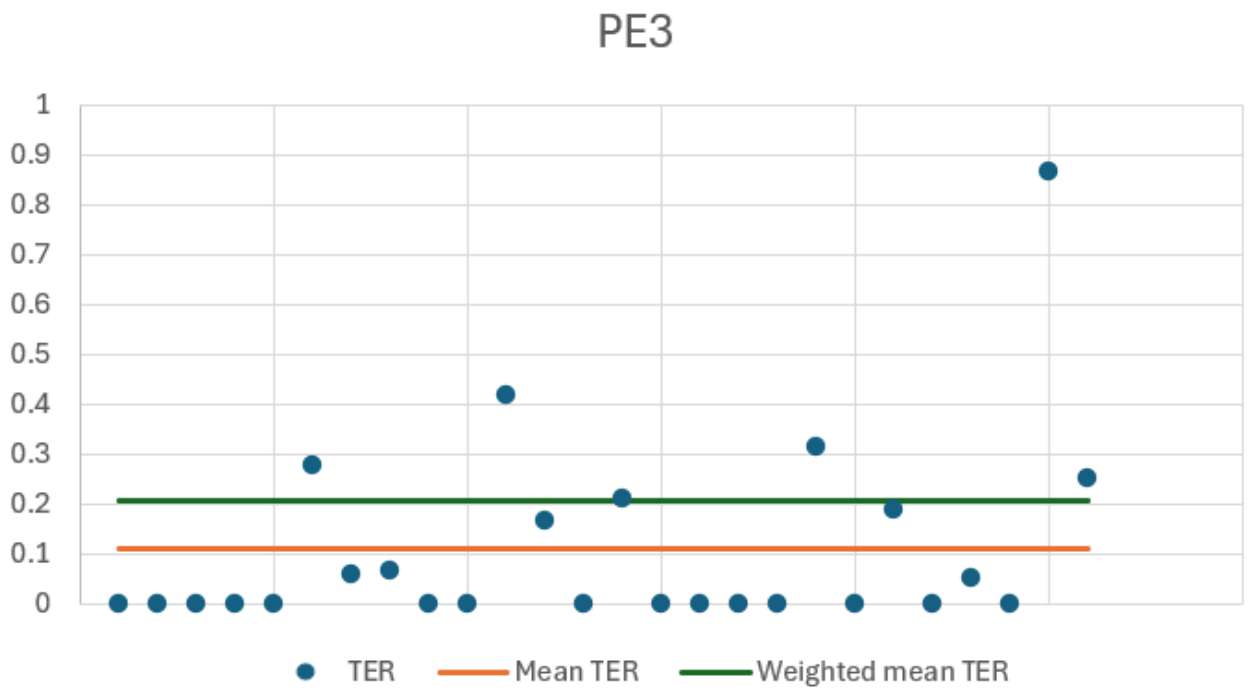
The fastest to post-edit was PE7 with a time of 10 minutes and 3 seconds. PE7 post-edited the MT output with a word count of 230 words and the mean TER score was 7% (see Figure 1). The weighted mean TER score, i.e., the score that takes into account the number of words in each segment and their contribution (weight) to the overall mean score, was 10% in this case. It took this post-editor 27 minutes and 34 seconds to translate from scratch, almost three times the time required for MTPE.

Figure 1. ***PE7's TER Analysis Chart***



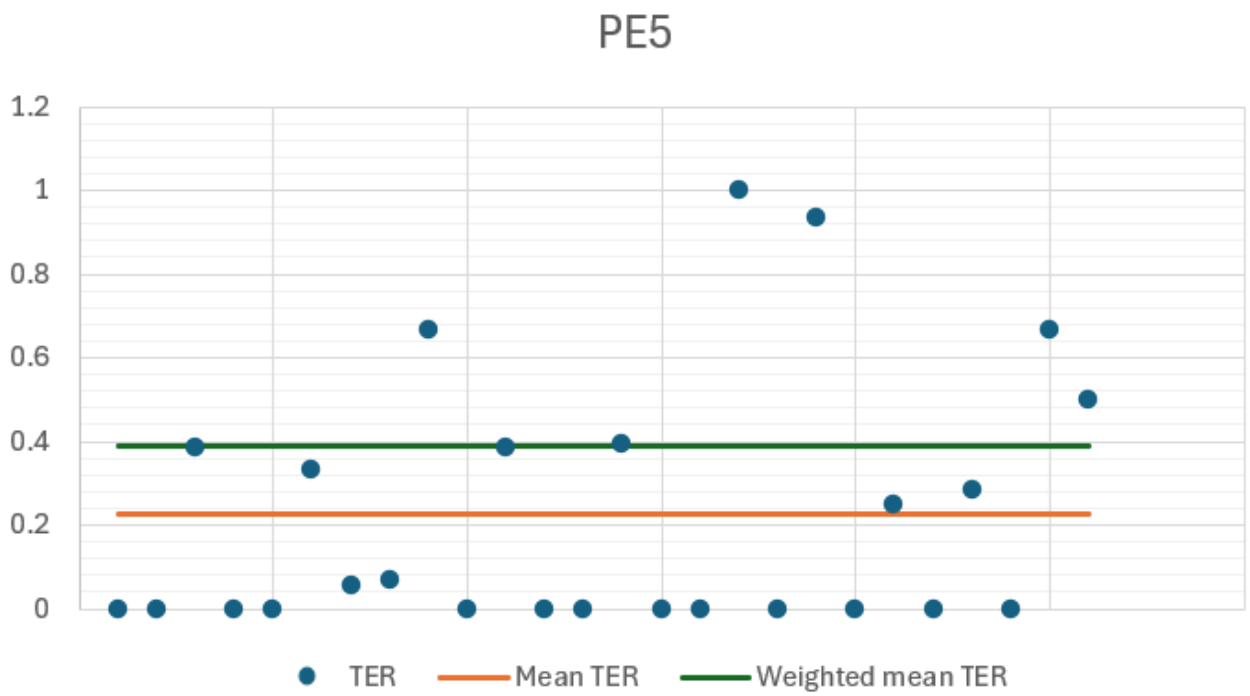
On the other hand, the slowest to post-edit was PE3 with a time of 31 minutes and 52 seconds. PE3 also posted the MT output with a word count of 230 words and the mean TER score was 11% (see Figure 2). The weighted mean TER score was 21%. From-scratch translation took this post-editor 43 minutes and 44 seconds, only about one third more compared with MTPE.

Figure 2. ***PE3's TER Analysis Chart***



The highest TER score of 23% (see Figure 3) with the weighted mean TER score of 39% were recorded by PE5, who took 26 minutes to do MTPE of 230 target words and 36 minutes to translate from scratch.

Figure 3. ***PE5's TER Analysis Chart***



The analysis showed that in none of the cases did the time required for MTPE occur to be longer than the time required for translation from scratch, i.e., in all cases it was demonstrated within the possible measurements (see Tables 2 and 3) that MTPE was (at least time-wise) more efficient than from-scratch translation. Thus, up to the highest recorded mean TER = 23% (weighted mean TER = 39%), there was no evidence that from-scratch translation was more efficient than MTPE.

Secondly, it can be observed that identification of and defining the threshold of the acceptability of MT output for post-editing was not possible in this experiment; one could only assume from the analysis that it must be at TER > 23% (weighted mean TER > 39%), as this is the highest TER score recorded where MTPE is still more efficient than from-scratch translation. Therefore, if at TER = 23% (weighted mean TER = 39%) MTPE is more efficient than from-scratch translation, the threshold of the acceptability of MT output for post-editing must be higher than this recorded TER score.

At the same time, the data does not indicate the existence of a trend that puts the time required for MTPE in a direct linear relationship with the TER score, or the time required for from-scratch translation.

Table 2. ***Results overview (pilot experiment): time, mean TER, weighted mean TER***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE1	21 min 57 sec	9%	12%	44 min 53 sec
PE2	10 min 42 sec	5%	8%	61 min 9 sec
PE3	31 min 52 sec	11%	21%	43 min 44 sec
PE4	15 min	17%	21%	40 min
PE5	26 min	23%	39%	36 min
PE6	19 min 43 sec	6%	7%	43 min 27 sec
PE7	10 min 3 sec	7%	10%	27 min 34 sec
PE8	17 min 32 sec	12%	20%	54 min 26 sec
PE9	11 min	7%	12%	20 min 38 sec

Table 3. *Ascending order of post-editors in each area (pilot experiment)*

PEMT	Mean TER	Weighted mean TER	From-scratch
PE7	PE2	PE6	PE9
PE2	PE6	PE2	PE7
PE9	PE7	PE7	PE5
PE4	PE9	PE1	PE4
PE8	PE1	PE9	PE6
PE6	PE3	PE8	PE3
PE1	PE8	PE3	PE1
PE5	PE4	PE4	PE8
PE3	PE5	PE5	PE2

Afterwards, the main experiment, in which 12 post-editors participated, was analyzed. While the pilot experiment involved students with relatively little or even no experience with MTPE, the main experiment involved students enrolled in Machine Translation Post-Editing course in the middle of the semester; in this way, we wanted to ensure that all post-editors had the same minimum background when it came to MTPE. The setup was the same: six post-editors did MTPE first (they post-edited the first half of the text, which contained 236 words of MT output) and then they translated the second half of the text from scratch. The remaining six post-editors translated the first half of the text from scratch first and then they did MTPE (they post-edited the second half of the text, which contained 272 words of MT output).

From the analysis, it is already evident at a glance that post-editors in this experiment were faster than post-editors in the pilot experiment; the variance of from-scratch translation (shown in Table 4) ranged from 18 minutes and 35 seconds to 35 minutes and 18 seconds, and the variance of MTPE ranged from 9 minutes to 21 minutes and 40 seconds.

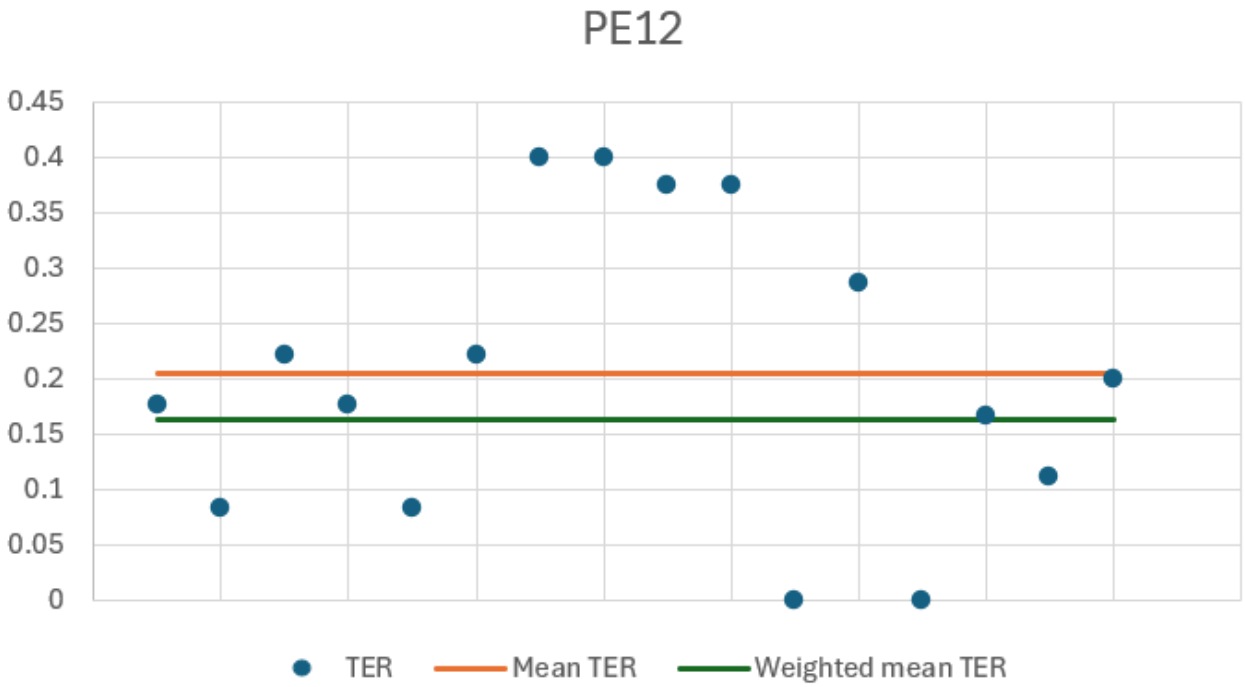
Table 4. ***Results overview (main experiment): time***

Post-editor	MTPE	From-scratch
PE1	21 min 40 sec	18 min 35 sec
PE2	15 min	25 min
PE3	16 min 34 sec	21 min 40 sec
PE4	11 min	27 min
PE5	21 min 40 sec	22 min
PE6	14 min	32 min 45 sec
PE7	13 min 51 sec	21 min 46 sec
PE8	14 min 48 sec	21 min 40 sec
PE9	17 min	27 min
PE10	14 min 30 sec	35 min 18 sec
PE11	17 min	33 min
PE12	9 min	25 min

The higher experience of the post-editors with MTPE could be partly responsible for this acceleration (as evidenced by the higher recorded values of the mean TER score compared to the pilot experiment), but there's no data to confirm this assumption. Moreover, if the post-edited text was the same in both experiments (some of the post-editors took part in both experiments, therefore the text had to be different), it would be possible to speak of a greater experiential complex of the post-editors. In this way, it can be assumed that the improvement in the temporal data was most likely due to the text itself, which was more general and comprehensible compared to the first text.

The fastest to post-edit was PE12, who took only 9 minutes to post-edit the 236-word MT output. The mean TER score (shown in Figure 4) was 20%, the weighted mean TER score was 16%. From-scratch translation took this post-editor 25 minutes.

Figure 4. ***PE12's TER Analysis Chart***



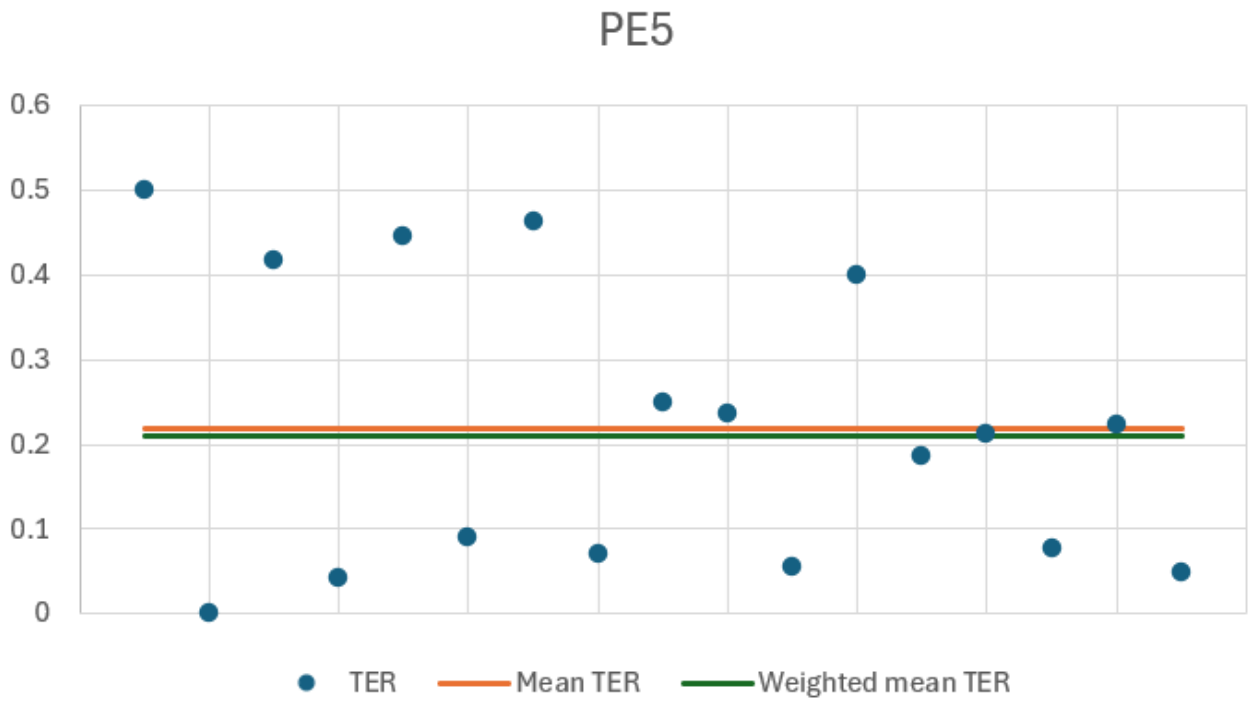
Similar results can be seen for almost all of the other post-editors (shown in Table 5); the time required for MTPE is lower than the time required for from-scratch translation, the differences between these times are heterogeneous and substantial, and there is no evidence of a tendency for the time required for MTPE to increase in direct proportion to the TER score.

Table 5. ***Results overview (main experiment): time, mean TER, weighted mean TER, sorted in ascending order by time required for MTPE***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE12	9 min	20%	16%	25 min
PE4	11 min	11%	14%	27 min
PE7	13 min 57 sec	8%	7%	21 min 46 sec
PE6	14 min	17%	26%	32 min 45 sec
PE10	14 min 30 sec	7%	12%	35 min 18 sec
PE8	14 min 48 sec	9%	10%	32 min 40 sec
PE2	15 min	9%	11%	25 min
PE3	16 min 34 sec	27%	27%	21 min 40 sec
PE9	17 min	15%	12%	27 min
PE11	17 min	17%	17%	33 min
PE5	21 min	22%	21%	22 min
PE1	21 min 40 sec	44%	41%	18 min 35 sec

For PE5, the difference between the times required for MTPE and from-scratch translation was only 1 minute (MTPE took 21 minutes and from-scratch translation took 22 minutes), indicating that MTPE was only slightly more efficient than from-scratch translation. The mean TER score (shown in Figure 5) was 22%, the weighted mean TER score was 21%.

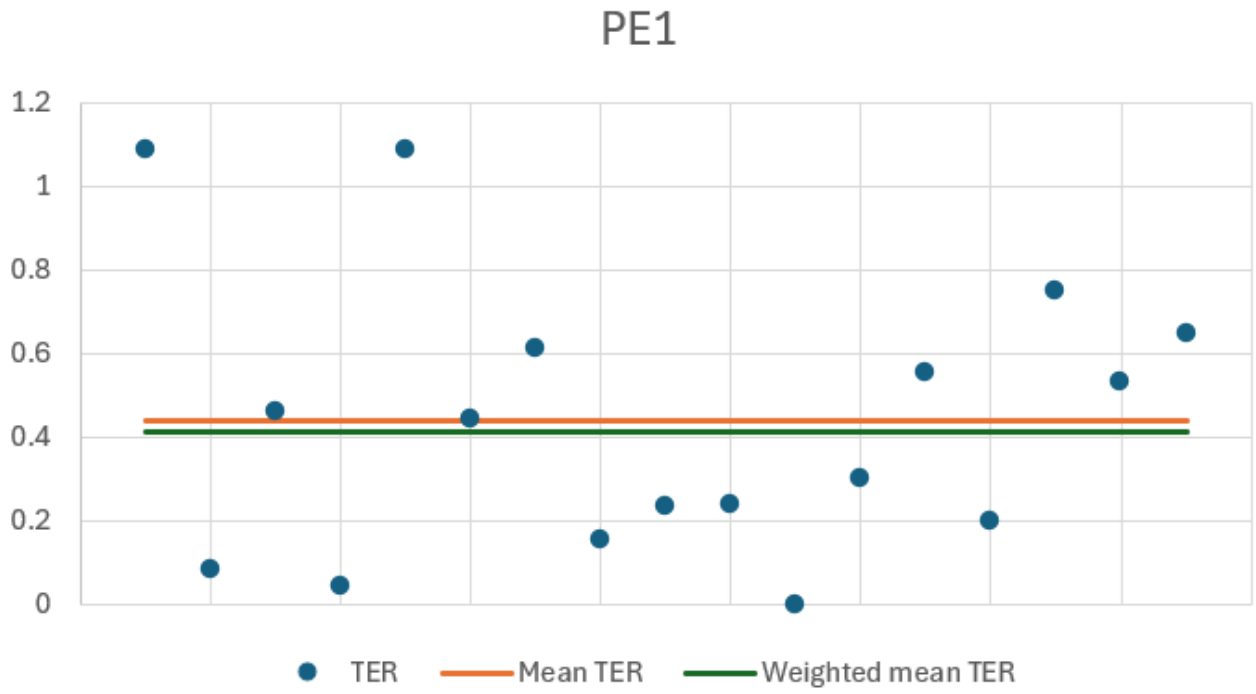
Figure 5. ***PE5's TER Analysis Chart***



With this post-editor, a hint of a tendency can already be seen that if the TER score were to rise to an even higher level, it could mean an increase in the time required for MTPE such that there would be a change in the efficiency of both processes in favor of from-scratch translation. However, it cannot be determined what TER score would be the tipping point based on this post-editor, because the only TER score recorded is the one at which MTPE is still more efficient than from-scratch translation. And since the recorded results do not show the existence of a correlation between the time required for MTPE and the TER score, there's no way of determining – based on the results of PE5 – even a hypothetical TER score where the change in efficiency would occur.

PE1's results are fundamentally different from all other post-editors. First of all, a higher MTPE time can be seen when compared to from-scratch translation, as well as a higher TER score rate. MTPE of the 272-word MT output took PE1 21 minutes and 40 seconds, which is 3 minutes longer than it took this post-editor to translate from scratch (18 minutes and 35 seconds). The recorded mean TER score was 44%, the weighted mean TER score was 41% (shown in Figure 6).

Figure 6. ***PE1's TER Analysis Chart***



What is significant about PE1's results is not only the highest recorded mean TER score, along with the weighted mean TER score (shown in Figure 7), and the higher MTPE time compared to from-scratch translation (shown in Figure 8), but also that both the MTPE time was the highest of all post-editors and the from-scratch translation time was the lowest of all post-editors. The MTPE time was 6 minutes above average, the from-scratch translation time was 8 minutes below average.

Figure 7. ***Chart comparing TER scores (main experiment)***

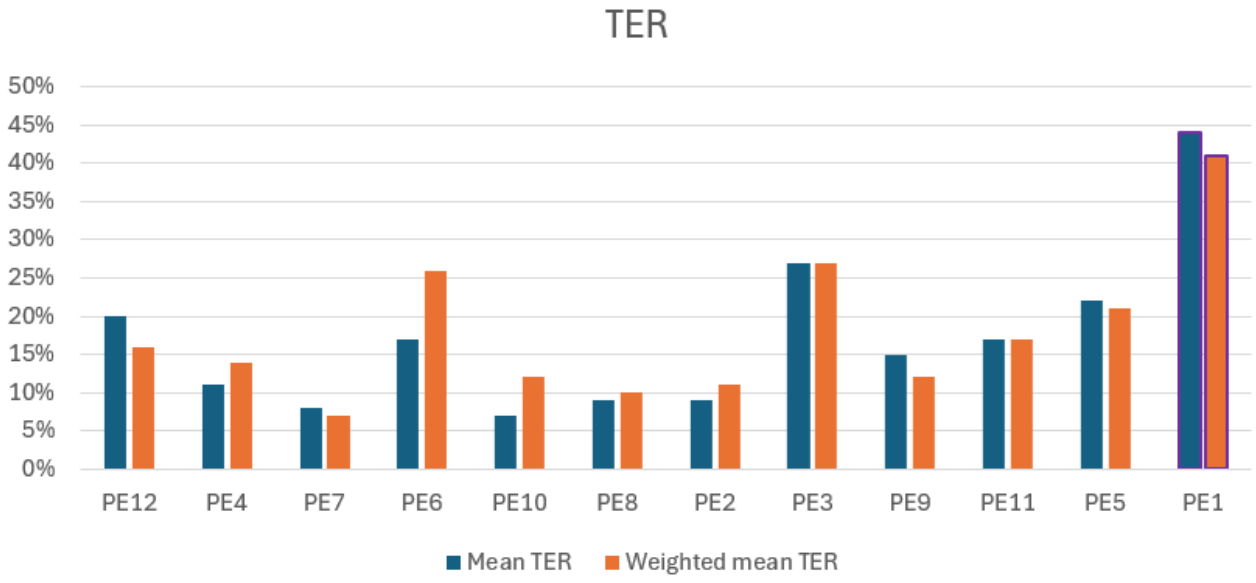
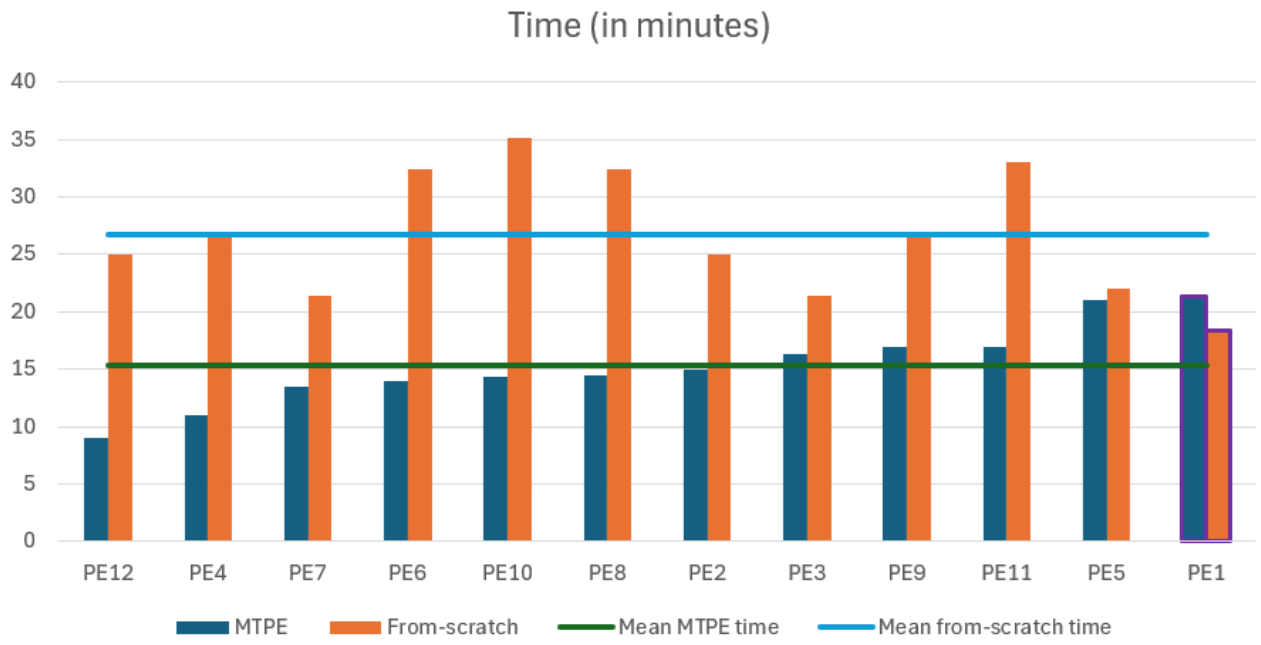


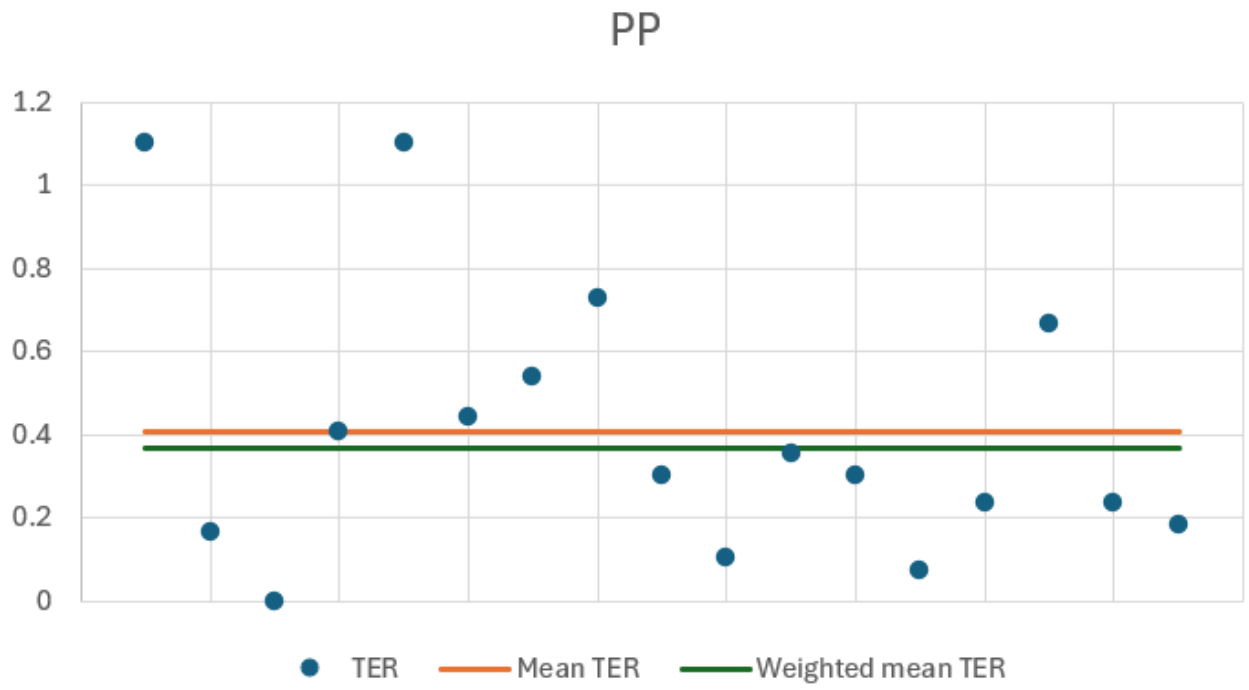
Figure 8. ***Chart comparing MTPE and from-scratch translation times (main experiment)***



On this basis, it is believed that the difference between PE1 and the other post-editors is due to their experiential complex, in which they surpass the other post-editors (PE1 have been active as a translator for a long time). It was decided to verify this assumption by assigning the same task to a professional translator/post-editor.

The professional translator/post-editor (PP) did the same task as PE1translating the first half of the text from scratch, then performing MTPE on the second half of the text, which contained 272 words of MT output. MTPE took them 22 minutes and 24 seconds, which is only 44 seconds more than it took PE1, and from-scratch translation took PP 17 minutes and 8 seconds, which is again just one minute and 27 seconds less than it took PE1. The mean TER score (shown in Figure 9) was 41% (PE1's TER score was 44%) and the weighted mean TER score was 37% (PE1's weighted mean TER score was 41%).

Figure 9. ***Professional translator and post-editor's TER Analysis Chart***



Based on the nearly identical temporal results and the comparable TER scores of PP and PE1 (shown in Table 6), it can be concluded that PE1's experiential complex exceeds that of the remaining post-editors and causes the difference between their results and the results of the remaining post-editors.

Table 6. ***PE1's and PP's results overview: time, mean TER, weighted mean TER***

Post-editor	MTPE	Mean TER	Weighted mean TER	From-scratch
PE1	21 min 40 sec	44%	41%	18 min 35 sec
PP	22 min 24 sec	41%	37%	17 min 8 sec

5 Discussion

Firstly, as mentioned above, from the analysis of the pilot experiment it can be concluded that MTPE was faster than from-scratch translation in all cases, i.e., it demonstrated higher temporal efficiency within the range of possible measurements compared to from-scratch translation in each case (see Table 1 in the previous section). The highest recorded mean TER scores in this experiment were TER = 23% (weighted mean TER = 39%) (see Figure 3 in the previous section). These results are in line with

the findings of Escartín and Arcedillo (2015), who have measured productivity gains for TER ≤ 0.21 (i.e., 21%). The remaining values are shown in Table 2 in the previous section.

In this experiment, it was not possible to identify and define a threshold of the acceptability of MT output for post-editing, as even at the highest recorded TER (23%), MTPE was more efficient than from-scratch translation. Based on this result, it might be assumed that the acceptability threshold must therefore be above the TER score of 23% (or 39% for the weighted mean TER).

At the same time, the measured results do not show a direct correlation between the time required for MTPE and the TER score, as has already been stated by Stefaniak (2020) in her research on English-Polish language pairs, nor between the time required for MTPE and the time required for from-scratch translation. Koponen's (2016) conclusions have shed light on the variation in the measured values: all of these are influenced by the characteristics of the source text, the type of editing, and the varying speed of the post-editors themselves.

The main experiment yielded comparable results: in almost all cases, MTPE was more efficient (faster) than from-scratch translation. The mean TER scores were higher despite the more general text, which can be interpreted to be due to the greater level of experience of the post-editors, especially their increased ability to identify errors in the MT output, which is also more and more difficult with the increasing quality of NMT systems, according to the findings of Yamada (2019, in: Vieira 2019). The participants were MA students at Comenius University majoring in philology with a specialization in translation and interpreting enrolled in Machine Translation Post-Editing course. Although it could not be guaranteed that all post-editors had the same experience, by selecting students enrolled in this course, it was ensured that all had comparable minimum skills. The argument about more experienced post-editors can also be supported by the temporal data, which is lower compared to the pilot experiment, i.e., post-editors in the main experiment post-edited and translated faster. Similarly with the results of the pilot experiment regarding a correlation between the time required for MTPE and the TER score, it can be concluded that such a correlation was not demonstrated here. Gaspari et al. (2014, in: Stefaniak 2020) have stated that such a correlation is weak, if it exists at all.

Results that already show a departure from the rest were found for two post-editors: PE5 and PE1. Although PE5's MTPE time was still lower than from-scratch translation time, the difference was only one minute (see Table 4 in the previous section). The recorded mean TER score was 22% (weighted mean TER = 21%) (see Table 5 or Figure 5 in the previous section). As mentioned earlier, it can be assumed that if the TER scores were to increase further, it could result in an increase of time required for MTPE in such a way that it could cause a change in the efficiency of MTPE vs. from-scratch translation in favor of from-scratch translation. However, since the results do not show a

correlation between the TER score and the time required for MTPE, there's no way of determining – even hypothetically – what TER score would be the threshold representing the tipping point in efficiency. The recorded TER score—at which MTPE was consistently more efficient—is consistent with the findings of Escartín and Arcedillo (2015), who, for English-Spanish language pairs, have set the threshold for the efficiency of MTPE versus from-scratch translation at TER ≤ 0.21 (i.e., 21%).

The most interesting were the results of PE1 (experienced student), whose MTPE time was the only one to be higher than the from-scratch translation time: MTPE = 21 minutes and 40 seconds; from-scratch translation = 18 minutes and 35 seconds (see Table 5 in the previous section). The times also represented the thresholds recorded within this sample: the MTPE time was the highest among all post-editors, the from-scratch translation time the lowest (see Table 5 or Figure 8 in the previous section). PE1 also had the highest recorded TER score, which significantly exceeded the scores recorded for the remaining post-editors (see Figure 7 in the previous section): the mean TER score of PE1 was 44% (weighted mean TER score was 41%) (see Figure 6 in the previous section).

These findings formed the basis for the belief that the results demonstrate PE1's higher experiential complex relative to the other post-editors, as PE1 is known, while still studying, to be actively working as a translator in a company that is among the best in the country in terms of its specialization.

To confirm or refute this assumption, an additional research question was therefore asked: Are the results of a professional post-editor similar to the results of PE1?

The results of the professional post-editor are almost identical to those of PE1 (see Table 6 in the previous section), and it can be concluded that PE1 possesses a significantly higher experiential complex compared to the other post-editors: the professional post-editor's TER score indicates that the MT output contained a significant number of flaws that needed to be identified and corrected, which PE1 handled significantly better than the other post-editors; PE1's from-scratch translation time was only one minute higher than the professional post-editor's, and yet it was still 8 minutes lower than average (see Figure 8 in the previous chapter).

The differences in the measured values between the professional post-editor and PE1 support both Stefaniak's (2020) and Gaspari et al.'s (2014, in: Stefaniak 2020) statements that the relationship between TER score and MTPE time is not directly proportional, or that no clear correlation between the two has been demonstrated.

Based on these claims and on the average of scores of PE1 and the professional post-editor, it can be concluded that at mean TER = 42.5% (weighted mean TER = 39%), MTPE is less efficient than from-scratch translation.

Stefaniak (2020), in her research on English-Polish language pairs, has noted a productivity increase in MTPE even at TER = 39%, therefore, it can be concluded that the threshold of the acceptability of MT output for post-editing is in the interval of TER = 39%–42.5%.

This is consistent with Guerrero's (2020) hypothesis that 50% edit distance is too high as an MT output acceptability indicator; and also with another conclusion in which she has argued that MT output with an edit distance between 30% and 40% is mostly rated as 2 on the acceptability scale (from 1 to 4) for post-editing, which corresponds to the threshold of the acceptability of MT output for post-editing in the interval of TER = 39%–42.5%

This finding is also slightly higher than that of de Gibert Bonet (2018, in: Stefaniak 2020), who, for Spanish-English language pairs, has determined the TER score threshold value up to which productivity gains in MTPE occur to be 33%.

6 Conclusion

The aim of this study was to determine the threshold of the acceptability of MT output for post-editing based on TER analysis. It compares the efficiency of MTPE against from-scratch translation and seeks to determine the threshold at which the efficiency scales tip towards from-scratch translation.

The work defines the concept of machine translation post-editing, presents approaches to it and the process itself, and includes an insight into the similar research carried out on other language pairs. Towards the end of the first part of this work, the research questions are introduced and the methodology of this work as well as the research methods are described.

The second, practical part presents the results of the research together with tables and graphs. The results are elaborated on in the discussion. The research showed that there was probably no clear correlation between the time required for post-editing a machine translation output and the TER score which would place these variables in a directly proportional relationship; furthermore, the comparison with a professional translator identified and confirmed a significantly higher experiential complex of one post-editor from the main group (PE1), whose results alone produced a deviation from the otherwise observed trend; finally, based on the results of these two post-editors, it was concluded that the threshold of the acceptability of MT output for post-editing lies in the interval of TER = 39%–42.5% These findings are in line with those of Stefaniak (2020) and Guerrero (2020).

From a practical point of view, these findings can serve as a guide for translation agencies or individual translators on how to approach, for example, ordering/accepting machine translation post-editing jobs, what expectations are realistic for

translators/agencies, as well as how to assess the value of their own effort, work, or time. From a didactic point of view, these findings can serve as an impetus for educational institutions that train translators to respond to the needs and trends of the market and to offer their students more courses that better prepare them for the realities of the translation profession.

In the future, this research could be enriched by the inclusion of language quality assurance (LQA) and the participation of a larger number of professional translators.

Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

Bibliography

Absolon, Jakub. 2018. Strojový preklad a posteditovanie. PhD dissertation. Nitra: DTS FA Constantine the Philosopher University in Nitra.

Bhandari, Pritha, and Nikolopoulou, Kassiani. 2020. What Is a Likert Scale? | Guide & Examples. https://www.scribbr.com/methodology/likert-scale/. Accessed on: 4 April 2024.

do Carmo, Félix. 2020. 'Time is money' and the value of translation. In: Translation Spaces. 9(1): pp. 35-57.

Escartín, Carla Para, and Arcedillo, Manuel. 2015. A Fuzzier Approach to Machine Translation Evaluation: A Pilot Study on Post-editing Productivity and Automated Metrics in Commercial Settings. In: Babych, Bogdan, Eberle, Kurt, Lambert, Patrik, Rapp, Reinhard, Banchs, Rafael E., and Costa-Jussà, Marta R. (eds.), Proceedings of the Fourth Workshop on Hybrid Approaches to Translation (HyTra). Beijing: Association for Computational Linguistics. pp. 40-45.

Guerrero, Lucía. 2020. In Search of an Acceptability/Unacceptability Threshold in Machine Translation Post-Editing Automated Metrics. In: Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track). Virtual: Association for Machine Translation in the Americas. pp. 32-47.

Hudecová, Elena, Stahl, Jaroslav, Benková, Lucia, and Munková, Daša. 2021. Porovnanie strojového, posteditovaného a ľudského prekladu technickej dokumentácie zo slovenčiny do nemčiny. In: Slovenská reč. 86(2): pp. 192-207.

Koehn, Philipp. 2020. Neural Machine Translation. Cambridge: Cambridge University Press.

Koponen, Maarit. 2016. Machine Translation Post-editing and Effort: Empirical Studies on the Post-Editing Process. PhD dissertation. Helsinki: University of Helsinki.

O'Brien, Sharon. 2006. Machine-Translatability and Post-Editing Effort: An Empirical Study using Translog and Choice Network Analysis. PhD dissertation. Dublin: School of Applied Language and Intercultural Studies, Dublin City University.

O'Brien, Sharon. 2011. Towards Predicting Post-Editing Productivity. In: Machine Translation. 25(3): pp. 197–215.

O'Brien, Sharon. 2022. How to Deal with Errors in Machine Translation: Postediting. In: Kenny, Dorothy (ed.), Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence. Berlin: Language Science Press. pp. 105–120.

Pavlíková, Diana. 2022. Porovnanie strojového a humánneho prekladu terminológie. In: L10N Journal. 1(1): pp. 64–101.

Rothwell, Andrew, Moorkens, Joss, Fernández-Parra, Maria, Drugan, Joanna, and Austermuehl, Frank. 2023. Translation Tools and Technologies. Abingdon, Oxon; New York: Routledge.

Stefaniak, Karolina. 2020. Evaluating the Usefulness of Neural Machine Translation for the Polish Translators in the European Commission. In: Martins, André, Moniz, Helena, Fumega, Sara, Martins, Bruno, Batista, Fernando, Coheur, Luisa, Parra, Carla, Trancoso, Isabel, Turchi, Marco, Bisazza, Arianna, Moorkens, Joss, Guerberof, Ana, Nurminen, Mary, Marg, Lena, and Forcada, Mikel L. (eds.) Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Lisboa: European Association for Machine Translation. pp. 263-269.

Temizöz, Özlem. 2013. Postediting Machine Translation Output and its Revision: Subject-Matter Expert Experts versus Professional Translators. PhD dissertation. Tarragona: Universitat Rovira i Virgili.

Nemergut, Matúš. 2024. Machine Translation Quality Based on TER Analysis from English into Slovak. In: L10N Journal 2(3), pp. 60–86.

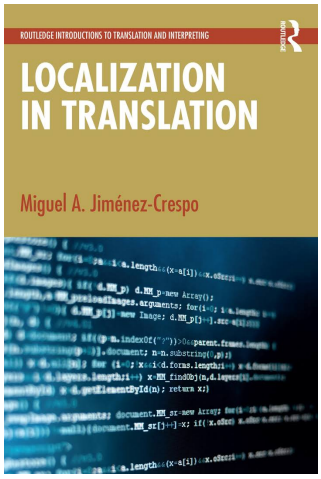
Vieira, Lucas Nunes. 2019. Post-Editing of Machine Translation. In: O'Hagan, Minako (ed.), The Routledge Handbook of Translation and Technology. London and New York: Routledge. pp. 319-335.

Walker, Julie. 2018. Term of the Week: Fuzzy Match. https://www.thelanguageoflocalization.com/2018/08/01/term-of-the-week-fuzzy-match/. Accessed on: 21 April 2024.

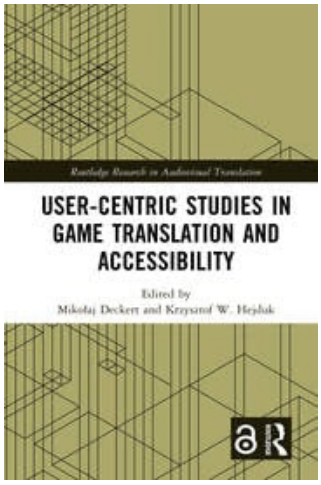
Zhechev, Ventsislav. 2014. Analysing the Post-Editing of Machine Translation at Autodesk. In: O'Brien, Sharon, Balling, Laura Winther, Carl, Michael, Simard, Michel, and Specia, Lucia (eds.), Post-editing of Machine Translation: Processes and Applications. Newcastle upon Tyne: Cambridge Scholars Publishing. pp. 2-23.

Final Variable

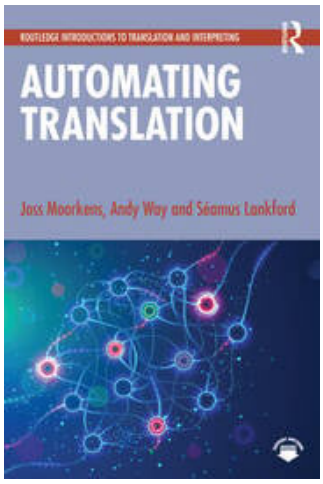
The Final Variable highlights important new publications in translation studies that focus on localization, technologies in translation, and related topics. This issue focuses on three new monographs that were published in 2024.



Localization in Translation by Miguel A. Jiménez-Crespo explores the process of localization, which goes beyond translation by adapting content to the cultural and linguistic norms of a target audience. The book examines theoretical and practical aspects of localization, including its role in software, websites, and multimedia content. Jiménez-Crespo discusses industry standards, translation technologies, and the influence of globalization. He also explores the challenges faced by translators and how localization reshapes traditional translation practices. The book is a valuable resource for students and professionals in translation and localization studies.



User-Centric Studies in Game Translation and Accessibility, edited by Mikołaj Deckert and Krzysztof W. Hejduk, explores how translation and accessibility impact video game experiences from a user-focused perspective. Divided into two main sections, the first part addresses theoretical challenges and research opportunities in game localization and accessibility, including issues with existing terminology, studies on minority languages, and the use of eye-tracking technology. The second part presents empirical studies examining topics such as streaming localized games, Arabic mobile game localization, gaming habits of visually impaired players, and the role of personality traits in localization testing.



Automating Translation by Joss Moorkens, Andy Way, and Séamus Lankford explores the role of machine translation (MT) and artificial intelligence in the translation industry. The book covers the origins and evolution of MT, the training data used in neural machine translation (NMT) and large language models (LLMs), and methods for evaluating their quality. It also discusses the integration of MT in audiovisual translation and localization, as well as ethical and sustainability concerns related to automation in translation. Additionally, the book provides practical insights on building and customizing MT models, making it a valuable resource for students, translators, and professionals in the field.