Universiteit Leiden

Master Creative Intelligence and Technology

[Differing Translation Methods of Japanese Onomatopoeia by Human and Machine Translators and Their Reception by Laypersons]

Name: [Payman Sipass]
Student ID: [S2929619]
Date: [09/09/2026]

Specialisation: [Creative Intelligence and Technology]

1st supervisor: [Max van Duijn]
2nd supervisor: [Alina Karakanta]

Master's Thesis in Creative Intelligence and Technology

Leiden Institute of Advanced Computer Science
Leiden University
Einsteinweg 55
2333 CC Leiden
The Netherlands

Contents

1. INTRODUCTION 3

2. PREVIOUS WORK 5

2.1 Translation of onomatopoeia / Ideophones 5
2.2 Machine translation 6
2.3 In-Context Learning 7
2.4 Evaluation of Texts by Laypersons 8

3. PRELIMINARY RESEARCH 9

3.1 NLP Analysis 9

4. METHODOLOGY 10

4.1 Text analysis 10
4.1.1 Choice of text 10
4.1.2 Preparing original Japanese text and extracting Japanese onomatopoeia 10
4.1.3 Choice of machine translation systems 11
4.1.4 Translating the text 11
4.1.5 Extracting onomatopoeia from human translation 12
4.1.6 Evaluating MT systems 12
4.1.7 Extracting onomatopoeia from MT systems 14
4.1.8 Labelling all Japanese onomatopoeia 14
4.1.9 Labelling all Human and MT onomatopoeia 15
4.1.10 Statistical Analysis using JASP 15
4.2 Survey Structure 15

5. RESULTS 17

5.1 Text Analysis Results 17
5.1.1 Japanese Onomatopoeia 17
5.1.2 Human and DeepL Translations 17
5.1.3 Translation Methods of HT and MT 18
5.1.4 Translating as an Adverb 19
5.1.5 Asymmetries in Translation 19
5.2 Survey Analysis 20
5.2.1 Higher preference for HT 20
5.2.2 Low perceived difficulty in ranking 22

6. DISCUSSION 23

6.1 Extracted expressions 23
6.2 MT evaluation 24
6.3 MT tendencies 24
6.4 HT tendencies 25
6.5 Perception of laypersons 25
6.6 Perceived beauty 26

7. CONCLUSION 27

8. SOURCES 27

1. Introduction

Onomatopoeia are generally associated with comic books in the West. While research shows that some Western languages also utilize them in multimodal contexts (Dingemanse, 2019), it is clear that onomatopoeia play a much broader role in the Japanese language. Called "*giongo",* they appear not only in comics, but are also regularly used in spoken conversations, literary texts, and even regional government meetings (Takamura et al., 2015). Their main function is to emulate specific sounds. An example of a *giongo* used in a sentence would be:

> *ame ga zaazaa to furu*
> *(it is raining hard)*

Here, the verb "*furu*" is preceded by the onomatopoeic expression "*zaazaa"*, which changes the meaning from simply "it is raining" to "it is raining hard". Of course, onomatopoeia is also used in a similar way to how it is usually used in Western comics (e.g., "*ban"* being used to emulate the sound of an explosion just like how "bang" is used in English).

Alongside these *giongo*, the Japanese language also contains other categories of phonomimes. Examples of these are "*gitaigo*", which emulate the texture or appearance of something, and "*gijougo*", which describe emotional feelings. *Gitaigo* in particular are frequently highlighted in research on Japanese onomatopoeia as mimetic expressions (Inose, 2008; Flyxe, 2002).

In linguistics, words like these which describe sounds or conditions are often referred to as ideophones and are used frequently in other languages, too. Examples include other East Asian languages such as Korean, but also in African languages such as Siwu, an indigenous language hailing from Ghana (Dingemanse, 2013). Research on ideophones is extensive and is often focused on the link between them and cultural context, linguistic structure, and iconicity (Dingemanse et al., 2016; Iida & Akita, 2024). The term iconicity is often related to this field of research, as it relates to how the form of an utterance matches its meaning (e.g., when pronounced, the word 'buzz' sounds like a bee buzzing). If a specific sound is associated with a particular meaning in one language but not in another, translation of a word containing the sound might contain the meaning element, but not the form element.

Translating ideophones is often cited as being particularly difficult, partly due to these cultural links and added layer of form and meaning (Sasamoto, 2021). This is especially the case when the target language makes less use of ideophones or completely lacks them altogether (Flyxe, 2002; Inose, 2008). However, the term ideophones is inexplicably seldom used in research surrounding the Japanese language. Instead, some researchers choose to use the word "onomatopoeia" or the Japanese word "*onomatope*" to refer to *giongo*, *gitaigo*, and other categories of these expressions. This paper will use the words "onomatopoeia" and "onomatopoeic expression" as umbrella terms to describe *giongo*, *gitaigo*, and all other categories.

At the same time, advancements in machine translation (MT) technology have quickened rapidly ever since Google Translate's transition from statistical machine translation (SMT) to neural machine translation (NMT) models in 2016 and more recently the increase in large language model-based (LLM) MT systems. MT technology has become increasingly better at translating texts from various languages, with some research even arguing that human parity (being of the same quality as human translation) has been achieved (Hassan et al., 2018),

although others do provide evidence to support the contrary (Toral et al., 2018; Läubli et al., 2018). Regardless, this improvement of MT technology has led to an increase in financial investment as the billion-dollar industry is projected to become a multibillion-dollar industry by 2030 (Grand View Research, 2026). This has in turn stoked fears among professional and aspiring translators about the future of their careers (Shin & Mah, 2024). Indeed, massive companies such as Netflix have introduced the usage of translation engines in the translation process of subtitles, for example (The Localization Institute, 2025).

Because MT systems lack the ability to process the aforementioned form element of onomatopoeia through human sensory experience and instead rely on internalizing this information through text-based datasets, it is particularly interesting to see how MT systems tackle the same challenges faced by human translators when translating onomatopoeia. However, research pertaining to this subject, especially when the source and target language are Japanese and Dutch, remains limited. This paper intends to fill this gap by answering the following research question:

> How do translations of Japanese onomatopoeia into Dutch differ between human translators and MT systems?

To further specify the exact goals of this research, the following sub-questions will aim to further our understanding of the translation methods utilized by human translators and MT systems when tackling onomatopoeia, and the extent to which laypersons differentiate between human and machine translations of these expressions.

I. Are MT systems more likely than human translators to maintain the word-class of an onomatopoeia from the source text when translating it?
II. Do laypersons perceive a difference in creativity, simplicity, and/or fluency in the human and machine translations of onomatopoeia from Japanese to Dutch?
III. Do laypersons consistently show a preference towards human or machine translations of onomatopoeia in terms of perceived beauty?

The perception of laypersons over professional translators is examined due to them ultimately being the final recipient and evaluator of most translated content such as literary work. This makes them an important actor in the economics of translation, as positive, neutral, or negative feedback on the translated content they consume can have an impact on the decisions of the organizations producing these translations.

Sub-question I will be answered by comparing the human and MT translations of a Japanese novel into Dutch to see what translation methods were primarily used and how the word-class of each onomatopoeic expression was altered. Sub-questions II and III will be answered by having laypersons rank human and MT translations on their creativity, simplicity, fluency, and beauty by means of a survey.

For each of the sub-questions, the following hypotheses and null-hypotheses have been formulated:

I. H_0: There is no difference between human translators and MT systems regarding the tendency to maintain the word-class of onomatopoeic expressions during translation.

H_1: Due to MT systems translating text more literally than human translators, a stronger tendency to maintain the word-class of onomatopoeic expressions is present.

II. H_0: Laypersons are unable to perceive a difference in creativity, simplicity, and/or fluency in the human and machine translations of onomatopoeia.
H_1: Laypersons can identify a difference in creativity, simplicity, and/or fluency in the different translations.

III. H_0: Laypersons do not show a strong preference in beauty towards either of the translations.
H_1: Laypersons show a significant preference in beauty towards one of the translations.

The next section of this paper attempts to paint a clear picture of the current state of all research fields related to this study. Section three describes a preliminary research approach taken to investigate more general differences between human and machine translations of Japanese text. Section four focuses on the chosen methodology and aims to explain the motivation behind all choices taken. Section five showcases the results of the analysis and section six discusses this data. Finally, section seven summarizes the main findings of this paper, mentions the limitations and provides suggestions for further research.

2. Previous Work

2.1 Translation of onomatopoeia / Ideophones

Inose (2008) and Minashima (2004) show that translators of Japanese text use a plethora of translation methods during the translation process. Inose tackled translations into Spanish and English while Minashima only analysed English translations.

Some of the most frequently used methods were translating onomatopoeia as verbs, adverbs, adjectives, nouns, or a combination of the four. Their results indicate that a standard translation method for most onomatopoeia is not available. In fact, in a significant number of instances, omitting the translation of the onomatopoeic expression is the preferred method. Consequently, the task of translating Japanese onomatopoeia is often cited as a particularly difficult hurdle for translators to overcome (Inose, 2008; Flyxe, 2002).

Sasamoto (2021) challenged the notion that a high omission rate is linked to the difficulty of translating onomatopoeia by proposing instead that the showing-meaning continuum might play a bigger role. The showing-meaning continuum refers to how communication carries a direct, non-coded element (showing) and an indirect, coded element (meaning). Sasamoto gives the example of someone telling you that they are doing the dishes by saying the words “I am in the middle of doing the dishes” being indirect and coded communication while them holding up soapy, gloved hands is direct and non-coded communication. Onomatopoeia, by their nature of emulating the sound or feeling they are describing, also contain this non-coded showing element. A full translation, then, requires both the non-coded element and the coded meaning it is assigned in a dictionary. In her research, Sasamoto provides evidence that onomatopoeia which have a stronger meaning element are translated more consistently than those with stronger showing elements, where the translations differ more

often. This strengthens the argument that the showing-meaning continuum plays a role in the difficulty of translating these expressions.

Flyxe's (2002) study emphasizing the lexicality of onomatopoeia can be linked to the showing-meaning continuum. Borrowing Flyxe's definition, lexicality refers to "the degree to which a word is part of the established vocabulary" (p. 56). Mimeticity, its counterpart which is mentioned in the following sentence, is defined as "the degree to which a word is a direct imitation of a sound, state, or condition" (ibid). Considering these definitions, it follows that an onomatopoeia which is further on the meaning side of the showing-meaning continuum would be more lexical. The same, then, can also be said about onomatopoeia with a stronger showing element being more mimetic. Indeed, Sasamoto (2021) frequently mentions Flyxe's writings on lexicality and writes that a word which is more mimetic than lexical would contain a stronger showing element, leading to a less consistent translation.

Another reason to explain the difficulty of translating onomatopoeia is the stylistic differences between the source and target language, as the usage of onomatopoeia can feel out of place or childish in certain languages (Flyxe, 2002). Finally, the meaning of an onomatopoeic expression can sometimes contain multiple possible interpretations. Even among Japanese speakers, a reaching of consensus is not always guaranteed (Fukushima et al., 2014). This further complicates the task of translating these expressions.

2.2 Machine translation

Machine translation has been evolving rapidly with new technologies such as large language models (LLMs) being incorporated as the next step since neural machine translation (NMT) systems were introduced. Research by Hassan et al. (2018) on Chinese to English translation claims that MT systems have already achieved human parity, meaning that the generated translations are on the same level of quality as translations produced by human translators. However, others have pointed out flaws with the study, mentioning that the data set was partly translated from English into Chinese before being translated back into English again (Toral et al., 2018; Läubli et al., 2018). Because translation tends to lead to the simplification of the content (Laviosa-Braithwaite, 1998), the data set was inconsistent in its complexity, arguably distorting the results of the study by unfairly providing the MT with easier texts to translate. Furthermore, the participants who were tasked with judging the MT output were evaluating on a sentence-level, meaning that participants saw the translated sentences out of order without being able to keep the context of the surrounding text in mind. Both Toral et al. (2018) and Läubli et al. (2018) argue that this makes it harder for participants to spot mistakes such as using different terminology for the same word, as this is only perceivable when evaluating on a document-level. When this was taken into account, evaluators ranked the HT to be of higher quality than the MT (Läubli et al., 2018).

Zhang et al. (2024), in research more relating to the topic at hand, produced results that indicated that machine translation methods were not yet capable of accurately translating onomatopoeia from Japanese to Chinese. Yang et al. (2025), although not focusing on translation per se, but rather on the evaluation of translation of idioms, reiterate these results by finding that LLM-based MT systems such as GPT-4 and Alma-13B frequently fail at translating idioms across multiple domains by being incorrect, incomplete, and too literal. Raunak et al. (2023) mention, however, that this is still an improvement over NMT systems, as LLM-based MT systems produce less literal translations across the board and particularly when translating idiomatic expressions.

Luo et al. (2024) find that MT has a stronger tendency than HT to be less morphosyntactically diverse, often being more convergent with the source text. This convergence refers to how the structure of the source text is frequently maintained in the translation (e.g., the word-class of the words in the source sentence remain unchanged), leading to more 'direct' translations than when humans translate a text.

The above findings provide insights into the translation tendencies of MT systems, with the foremost discovery being that MT systems tend to translate more literally and maintain the structure of source language text more frequently than human translators. This is relevant to the present study as this behaviour may also show itself in the specific condition of onomatopoeia translation from Japanese to Dutch.

2.3 In-Context Learning

In-Context Learning (ICL) refers to the concept of providing an artificial intelligence (in this case an LLM-based MT system) with additional context or instructions within a prompt as an attempt to increase output quality without the need to update any parameters or spend more time and resources on further training (IBM, 2026). Because the additional context is provided through the prompt, ICL is closely related to the field of prompt engineering. Khoboko et al. (2025) found that the usage of ICL techniques boosted the translation quality of low-resource languages by an LLM-based MT system, even producing competitive results with MT systems which were pre-trained on the language-pair in question. Their work is built on Bawden and Yvon's (2023) research where the prompt format "⟨*X*⟩ = ⟨*Y*⟩", where *X* is the text in the source language and *Y* is the translation into the target language, improved translation quality of low-resource languages. This same prompt format is also shown to achieve the highest translation quality scores during automatic evaluation over other prompt formats such as "[SRC]: ⟨*X*⟩ \n [TGT]: ⟨*Y*⟩", where SRC stands for the source language and TGT stands for the target language (e.g., [English] and [Japanese]) (Zhu et al., 2024). This further goes to show the importance of prompts when translating using an LLM-based MT system.

As Brown et al. (2020) describe, there are multiple types of ICL. Zero-shot ICL refers to not providing any examples and merely describing the translation task to the LLM. One-shot ICL concerns offering the LLM a single example in the prompt. Few-shot ICL entails supplying multiple examples. In each of these scenarios, an example refers to a source language text and its target language translation. The model is then provided with a source language sentence text to translate into the target language. Brown et al. (2020) indicates an increase in translation quality when few-shot ICL is applied compared to zero-shot ICL. Furthermore, they discover that as the size of the LLM model increases, the usage of ICL becomes more efficient. This means that providing an LLM model with a higher parameter count with ICL returns a bigger quality increase compared to a smaller LLM model.

Bertsch et al. (2025) and Baek et al. (2025) both provide evidence that providing more context than in few-shot ICL increases translation quality even further. This is called many-shot ICL or long context ICL. This method is competitive with traditional finetuning of the LLM model and is more efficient. However, given unlimited training data, finetuning is still capable of outperforming ICL (Bertsch et al., 2025). Their research builds on the findings of Khoboko et al. (2025), as they show that translation quality improves not only when introducing ICL techniques, but also when the ICL techniques utilize more context.

2.4 Evaluation of Texts by Laypersons

In research conducted by Ferstler et al. (2026), participants were asked to compare human and machine translations (LLM-based) of novels in various languages across two reading conditions. Their results indicate that, while the machine translation was sufficient, there was an overall preference for the human translation, especially regarding its clarity, and immersiveness. Furthermore, while the machine translation was deemed to be more immediately readable, only the human translation received praise for its literary merit and fluency. These praises were amplified when the translated text was not English, but rather another language. This could be explained by the fact that LLMs are much more proficient at English output than other languages due to there being more English content in its data set (Liu et al., 2025).

Another interesting result is that participants were not able to reliably identify which translation was made by a human or MT system. This is consistent with research done by Porter & Machery (2024), where participants were unable to distinguish between human and AI-generated poems. In fact, their research actually reported that the AI-generated poems were preferred by the participants across a variety of metrics such as emotional quality and creativity, contrasting with the results of Ferstler et al. (2026). The proposed reasoning is that the AI-generated poems were less complex and that the participants (non-experts) were better able to understand these poems compared to the human-generated poems. Furthermore, the participants thought that the poems they were rating higher were the human-written poems, indicating an expectation that the human-written poems would be better than the AI-generated poems.

The finding that AI-generated content, whether they be translations (Ferstler et al., 2026) or poems (Porter & Machery, 2024), is more immediately readable and less complex than human output is reflected in Niu & Jiang's (2024) research. They provide evidence that, while simplification of content is perceivable across human and machine translations, machine translations are more simplified than human translations. Simplification was measured by the lexical diversity of the translations. They go on to report that the genre of the content also impacted the simplicity, with machine translations of academic abstracts having a lower lexical diversity than machine translations of novels, for example. However, this last point is also true for the human translations, calling into question whether the genre of the text played a bigger role in the lexical diversity rather than the translator.

Finally, Arenas & Toral (2022) reported that, when presented with a human translation, a post-edited human translation, and a machine translation by an MT system trained on literary data, participants unanimously ranked the human translation as most creative, with the post-edited human translation being second and the machine translation being ranked last. This is consistent with Ferstler et al. (2026) but goes against the findings of Porter & Machery (2024). It is worth noting that the MT system used in this research was an NMT, not an LLM, which was used by both of the previously mentioned studies.

Overall, previous research provides mixed findings, with some supporting the claim that laypersons can distinguish between human- and machine-generated content across various metrics, while others assert the opposite. No research, however, focuses specifically on the Japanese to Dutch translations of onomatopoeia in literary novels, which is what this study addresses.

3. Preliminary Research

3.1 NLP Analysis

Before the topic of this research was focused on the translations of Japanese onomatopoeia in particular, preliminary research was first conducted to analyse whether potential differences between human and machine translations of Japanese novels in general could be found using NLP techniques. The reasoning for including this preliminary research in this paper is two-fold. First, using NLP techniques is a common way to analyse text as data. By including the researcher's experience of using these techniques and showcasing the results, other researchers can replicate, avoid, or improve upon the steps taken to produce further research on the topic. Secondly, it provides a more comprehensive picture of the research at hand and how it evolved over time, eventually shifting away from using NLP to analyse translations of Japanese text to Dutch and English in general to manual analysis of specifically the translations of Japanese onomatopoeia to Dutch.

For the preliminary research, four different translations of the prologue chapter of the Japanese novel *Kafka on the Shore* were used. These included the official Dutch and English translations by Jacques Westerhoven and Philip Gabriel respectively, and Dutch and English machine translations made with the online version DeepL. The translations were compared based on: (1) their average sentence length, (2) their lexical diversity, and (3) their readability using the Flesch Reading Ease score. Lexical diversity was defined as the ratio of unique words in each text to the total number of words present. For the Dutch translations, a modified formula called Flesch-Douma was utilized to calculate the readability. For these calculations, a lower score indicates a higher difficulty to read the text.

Sentiment analysis was also performed on the translations utilizing both a lexicon-based approach and an LLM-based approach. The lexicon-based approach used a dictionary with pre-determined sentimentality scores for each word it contained, assigning every word in the analysed text with the appropriate score. The LLM-based approach was inspired by research done by Suzuki et al. (2025) where LLM-based sentiment analysis was performed on overviews of Japanese businesses to ascertain whether a higher sentiment score would result in better stock price performance of a company. All analysis was performed with Python code scripts in a Jupyter Notebook environment.

The preliminary research did not yield any significant results. The only point worth mentioning was a higher lexicality and lower readability score for the DeepL translations compared to the human translations. These results, however, were statistically insignificant. Furthermore, other research provides evidence for the contrary, suggesting that translations done by MT systems tend to be more simplified than human translations (Niu & Jiang, 2024).

These lacklustre results might partly be explained by the small dataset. While four versions of the source text (Human Dutch, Human English, DeepL Dutch, DeepL English) were used, the source text only accounted for roughly 1400 words in each case. In addition, due to this preliminary research occurring at an early point of the research process before the final direction was even decided, the methods chosen were hasty and therefore might have been insufficient to produce any meaningful results. Research dedicated to this topic without changing its course along the way would have a better chance at uncovering any meaningful

differences between human and machine translations of Japanese text, much like Niu & Jiang (2024) have done for Chinese text.

4. Methodology

4.1 Text analysis

4.1.1 Choice of text

The chosen text for this research is Haruki Murakami's *Kafka on the Shore* and will be referred to with its English name for the sake of readability. Specifically, the original Japanese version released in 2002 (in Japanese, *umibe no kafuka*) and the Dutch translation by Jacques Westerhoven made in 2006 (in Dutch, *Kafka op het strand*) were the main texts used. *Kafka on the Shore* was chosen for multiple reasons. Firstly, the book is easily available in epub format in the relevant languages for this research. Secondly, Inose's research (2008), which this research is heavily inspired by, also uses Murakami novels for the analysis of onomatopoeic expressions. Finally, the researcher has read *Kafka on the Shore* multiple times in English and partially in Japanese, giving the author a better understanding of the source text.

It is important to note that the original Japanese version of *Kafka on the Shore* is split into two volumes, while overseas releases including the English, and more importantly, the Dutch translations are often combined into one novel. This research only used the first part of the Japanese release, corresponding to chapters one through 23 of the Dutch translation.

4.1.2 Preparing original Japanese text and extracting Japanese onomatopoeia

The Japanese text was prepared by first cleaning the epub file to isolate only the text of the chapters. This was done using a custom Python script.[1] After cleaning, JMdict (Electronic Dictionary Research and Development Group, 1999) and Nanoskript's Japanese Onomatopoeia Dataset (Nanoskript, 2023) were used to match onomatopoeic expressions in the source text. These matched expressions were automatically added to an Excel sheet alongside the sentence each appeared in and the sentences directly before and after it. This was done to provide the MT systems with more context for the translation of the expressions but also to help with extracting the expressions from the official human translation. While the JMdict dataset has all of its words tagged so that only onomatopoeic expressions were considered during extraction, many other words and characters were also incorrectly extracted. To combat this, an exclusion list was utilized which contained words or combinations of characters that either JMdict or the Nanoskript dataset would mistakenly target as an onomatopoeic expression. Nevertheless, not all extracted expressions were actually onomatopoeia. In these unclear cases, a dedicated Japanese dictionary on these expressions was consulted (Hida & Asada, 2018)

[1] All used code is publicly available on the researcher's portfolio website: https://psipass.github.io/

4.1.3 Choice of machine translation systems

Many different NMT and LLM MT systems were considered for this research. In the end, one NMT system and two LLM systems were chosen. These are (1) the legacy DeepL translation engine (NMT), (2) DeepL's Next-Gen LLM-based translation model, and (3) Google's 4B TranslateGemma LLM-based translation model. The TranslateGemma model was used twice, once with a standard translation prompt and once more with a customized prompt aimed at improving the translation of onomatopoeic expressions. While the usage of the Next-Gen DeepL model was not planned, including it required minimal effort, so it was added to the research. This also provided an opportunity to highlight the differences between an NMT system and LLM model in a more direct way, as both MT systems were developed by the same company.

4.1.4 Translating the text

The DeepL desktop application was used to generate the DeepL translations. Using a free trial for DeepL Pro, the entire excel file containing the individual onomatopoeia, the sentence containing it, and the sentences directly before and after it were translated. Both the legacy model and LLM model translations were performed using this method.

Ollama was used to locally run the 4B TranslateGemma model. First, a text file with each Japanese sentence containing an onomatopoeic expression alongside the sentences directly before and after it was prepared. A custom Python script was then used to feed every line of this text file to the TranslateGemma model opened in Ollama with a standard and custom prompt. Each line would then be translated and pasted to a new text file. The exact prompts used are as follows:

Standard prompt:

> "You are a strict translation machine. You must translate the given text accurately. Output ONLY the final translated text. Do not output conversational filler, explanations, formatting, or multiple options."

Custom prompt:

> "Context: You are a translation expert specializing in translating Japanese novels to Dutch. You must translate the given text accurately. Output ONLY the final translated text. Do not output conversational filler, explanations, formatting, or multiple options. Make sure that the amount[2] of sentences in the translation is the same as in the source text. The text which you will translate contains Japanese onomatopoeic and mimetic expressions[3]. Consider the translation of these expressions carefully. DO NOT leave the expressions untranslated. Make sure to reflect the meaning of the expression in the translation accurately. Here are some example Japanese sentences containing onomatopoeic and mimetic expressions and their translations
>
> Japanese: 'でもひとつだけはっきりしていることがある。'; Dutch: 'Eén ding staat echter vast.'

[2]The prompt displayed here and used during translation contained a grammatical error. The word 'amount' should have been 'number'.

[3] When the prompt was developed, the terminology used during research was "onomatopoeic and mimetic expressions" instead of "onomatopoeia" or "onomatopoeic expression", as it is now.

Japanese: 'でも彼女はじっと待っている。'; Dutch: 'Ze wacht echter geduldig af.'
Japanese: 'その優雅な部屋を僕はすっかりひとりじめにすることができる。'; Dutch: 'Ik heb het elegante vertrek helemaal voor mezelf.'
Japanese: '町を通り過ぎると、道はすっかり暗くなってしまう。'; Dutch: 'Als we eenmaal het stadje uit zijn, wordt het pikdonker langs de weg.'"
Japanese: 'すぐ隣には黒い犬が彫像のように、身動きひとつせずじっと座っていた。'; Dutch: 'Vlak naast hem, roerloos als een standbeeld, zat de zwarte hond.'

The custom prompt was constructed with in-context learning (ICL) in mind. ICL refers to the practice of providing sample problems and answers to improve task quality. Bertsch et al. (2025) found that ICL can increase translation quality of LLMs, especially in regard to low-resource languages (Akallouch & Fardousse, 2025). In the case of this research, ICL was provided to the LLM by including five sentence pairs from the original Japanese text and the human translation.

4.1.5 Extracting onomatopoeia from human translation

Extracting the onomatopoeic expressions from the human translation was done manually, as there was no straightforward way to automate the process. However, because the Japanese expressions were already extracted alongside the sentence in which they were contained in order of appearance, searching for the translated expressions was not difficult. Sometimes, one of the translations generated by an MT system was used to accelerate the process, as these translations would often contain similar translations of words that could be used to identify a particular sentence. All extracted sentences were transferred to an Excel sheet.

4.1.6 Evaluating MT systems

To limit the evaluation to translations of the onomatopoeia in particular as much as possible, only the sentences containing the onomatopoeia were included in the evaluation process. However, because previous research indicates that MT quality rises when provided with more context (Karpinska & Iyyer, 2023), the evaluated sentences were derived from a translation which included the sentence directly before and after as well. This means that each translation system was provided with three sentences to translate in total, with only the middle sentence containing the onomatopoeia and therefore being used for evaluation. However, during the automatic translation process, each MT system sometimes merged or split the three sentences in different ways, making it difficult to isolate the sentence which contained the onomatopoeia. The same issue was discovered in the human translation, where sentences which were split in the Japanese source text were merged during translation or vice versa. To combat this, the decision was made to first isolate all sentences from the human translation which were not merged or split. This would serve as the standard to compare with the four machine translations. Any translated sentences from the MT systems which did not contain exactly three sentences were excluded. Next, the middle sentence of each of these sentence groups was extracted to be used for evaluation. While shrinking the data set, automatic evaluation of translations necessitates that each translation is synchronized in the sense that each line should contain the translation of the same source text sentence. In the end, 308 sentences were available for evaluation. These sentences were copied to individual .txt files where each line would house one sentence. The Japanese sentences were also isolated and formatted as an .txt file.

All four translations were evaluated using the online automatic evaluation tool MATEO (Vanroy et al., 2023). The chosen evaluation metrics were BLEU, ChrF, COMET, and TER. The human translation was used as the reference file. This meant that the scores for each metric would be based on how close each machine translation was to the human translation. The Japanese text was used as the source file due to COMET necessitating one. Figure 1 and table 1 show the results of the evaluation.

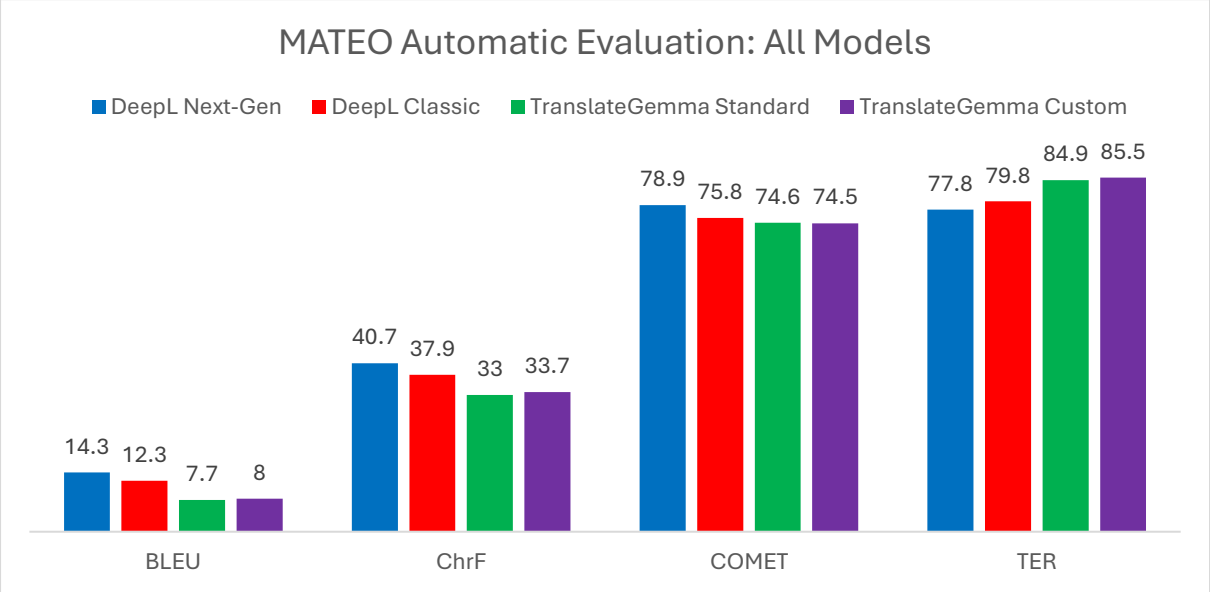


Figure 1: Automatic evaluation of all MT systems.

System	*BLEU (μ ± 95% CI)*	*chrF2 (μ ± 95% CI)*	*comet (μ ± 95% CI)*	*TER (μ ± 95% CI)*
DeepL Next-Gen (baseline)	14.3 (14.3 ± 1.8)	40.7 (40.7 ± 1.4)	78.9 (78.9 ± 1.1)	77.8 (77.9 ± 2.4)
DeepL Classic	12.3 (12.4 ± 1.6) (p = 0.0080)	37.9 (38.0 ± 1.4) (p = 0.0010)	75.8 (75.8 ± 1.3) (p = 0.0010)	79.8 (79.7 ± 2.4) (p = 0.0290)
TranslateGemma Standard	7.7 (7.7 ± 1.1) (p = 0.0010)	33.0 (33.0 ± 1.2) (p = 0.0010)	74.6 (74.6 ± 1.2) (p = 0.0010)	84.9 (85.0 ± 1.9) (p = 0.0010)
TranslateGemma Custom	8.0 (8.0 ± 1.3) (p = 0.0010)	33.7 (33.7 ± 1.3) (p = 0.0010)	74.5 (74.5 ± 1.2) (p = 0.0010)	85.5 (85.5 ± 2.1) (p = 0.0010)

Table 1: Metric scores of each MT system.

Due to MATEO necessitating a translation to function as a baseline, all of the p-values found in table 1 convey how statistically significant the difference is between each model and the baseline model (DeepL Next-Gen). This showed a clear superiority in quality of the DeepL Next-Gen translation model across all metrics tested. It outperformed not only its NMT predecessor, but also both of TranslateGemma's translation attempts. It is worth noting that for the TER metric in particular, lower scores indicate a better result.

Another evaluation was conducted to ascertain whether there were any significant differences between the TranslateGemma translations, as those were not perceivable in the first evaluation. Figure 2 and table 2 showcase these results. The only statistically significant difference found was between the ChrF scores of the standard and custom prompt translations, where the custom prompt translation scored slightly better than the standard prompt translation. Every other difference in scoring was not significant.

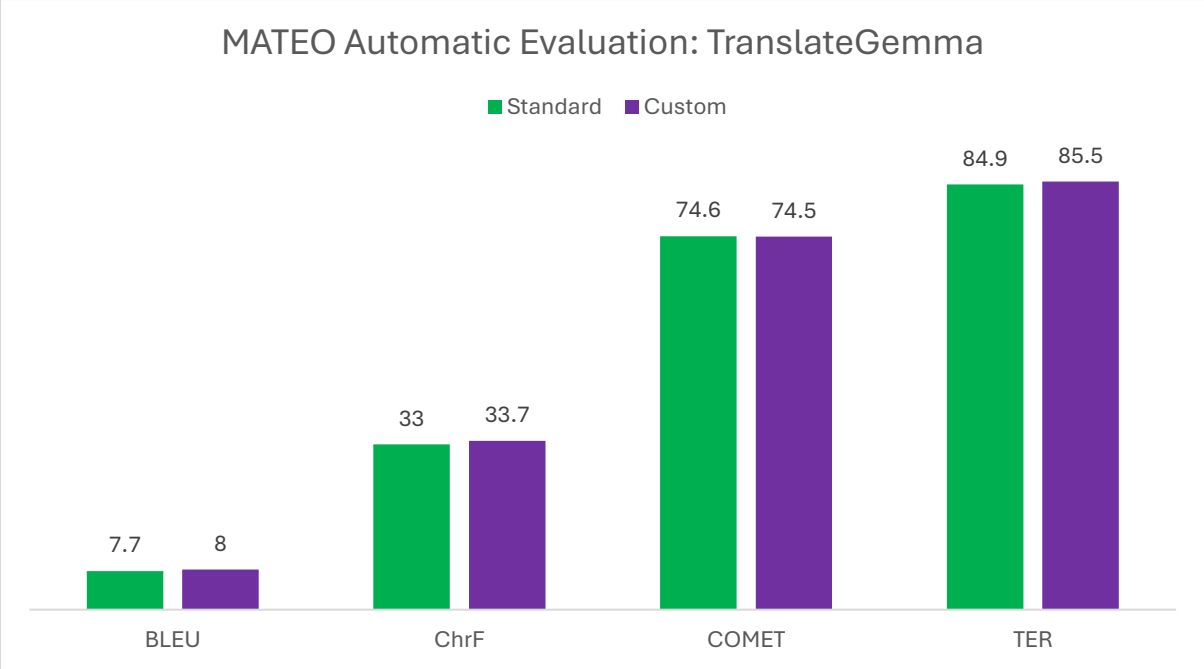


Figure 2: Automatic evaluation of the standard and custom prompt TranslateGemma translations.

System	*BLEU (μ ± 95% CI)*	*chrF2 (μ ± 95% CI)*	*comet (μ ± 95% CI)*	*TER (μ ± 95% CI)*
TranslateGemma Standard (baseline)	7.7 (7.7 ± 1.1)	33.0 (33.0 ± 1.2)	74.6 (74.6 ± 1.2)	84.9 (85.0 ± 1.9)
TranslateGemma Custom	8.0 (8.0 ± 1.3) (p = 0.1568)	33.7 (33.7 ± 1.3) (p = 0.0400)	74.5 (74.5 ± 1.2) (p = 0.2488)	85.5 (85.5 ± 2.1) (p = 0.1279)

Table 2: Metric scores of both systems. Only significant difference was found in the ChrF scores.

4.1.7 Extracting onomatopoeia from MT systems

After confirming that the DeepL Next-Gen translation performed the best out of all MT translations, it was chosen to be the representative machine translation for the remainder of this research. Extracting the onomatopoeia was done by isolating the sentence in which the expression appeared. These sentences would then be viewed in tandem with the sentences extracted from the human translation during the labelling process.

4.1.8 Labelling all Japanese onomatopoeia

Labelling the Japanese expressions was attempted with multiple methods. The part-of-speech tagger aimed specifically at labelling Japanese words called MeCab was considered due to its previous use in research pertaining to Japanese text, but its performance in accurately tagging the word-class of the expressions was unsatisfactory, as obvious adverbs or adjectives would be tagged incorrectly. Using the LLM Gemini was also attempted, but this resulted in similar unsatisfactory results in addition to being the slowest method. In the end, a custom Python script provided with the extracted expressions and a list of hard-coded rules iterating over each sentence containing the expression provided the best output. These rules were based on the moras directly behind the extracted expressions. An expression followed by the mora と, for example, strongly indicates that the expression functions as an

adverb. After running the code, a manual check was performed by the author to ensure accurate tagging.

4.1.9 Labelling all Human and MT onomatopoeia

All sentences from the human and DeepL translations containing onomatopoeia were viewed and labelled using an Excel spreadsheet. A modified version of Inose's (2008) labelling framework was used with two main differences. The translation method "translation using two adjectives, or combination of adverbs, adjectives or verbs" was excluded as differentiating it from "Translation using explicative paraphrases" was deemed too difficult. Furthermore, the translation method "No translation (omission or complete change of the phrase)" was split into "Omission" and "Change of phrase". If a sentence lacked the onomatopoeic expression entirely without a replacing word or phrase, it was counted as an "omission". If a replacing word or phrase was found with little to no shared meaning with the original expression, it was counted as a change of phrase. Considering these two modifications, the final list of translation methods used during labelling is as follows:

(1) Translation using adverbs
(2) Translation using adjectives
(3) Translation using verbs
(4) Translation using nouns
(5) Translation using explicative paraphrases
(6) Translation using idioms
(7) Translation using onomatopoeic expressions
(8) Omission
(9) Change of phrase

4.1.10 Statistical Analysis using JASP

After labelling each expression in the context of the sentence it appears in for both the human and DeepL translations, statistical analysis was performed on the data using the free program JASP. JASP was also used during the analysis of answers to the survey.

4.2 Survey Structure

For the second part of this research a survey was conducted using Qualtrics. After giving consent, receiving preliminary information about the project, and passing a screening question, the participants would rank the human and DeepL translations of the onomatopoeia based on the following four metrics: (1) creativity, (2) simplicity, (3) fluency, and (4) beauty. These metrics were briefly explained to the participants, but emphasis was made that no answer would be incorrect. The metrics were explained as follows:

> **Creativity**: How original and/or uniquely written you think the phrase is.
> **Simplicity**: How simple you think the phrase is written and how easy it is to read and understand.
> **Fluency**: How natural you think the Dutch phrase sounds.
> **Beauty**: How 'beautiful' you think the phrase is. The definition of 'beautiful' is for you to decide.

The metric 'beauty' in particular was the least defined, as participants were instructed to rank the translations on their own interpretation of beauty. This was done to increase

ecological validity, with the obvious drawback that the results would not be completely comparable. While preliminary information was provided to the participants, they were not told that they were evaluating the translations of onomatopoeia and that they were comparing human translations with automatically generated translations.

During the ranking section of the survey, the translated onomatopoeia were displayed next to each other in bold, black text while the surrounding sentence was greyed out and separated by a blank line (figure 3). Before the ranking portion, participants were given an explanation about the difference between the bolded black text and the greyed-out text, alongside the specific instruction to only consider the bolded black text when ranking. The participants were not told which translation was made by a human or MT system and the positions of the translations were randomized to avoid bias.

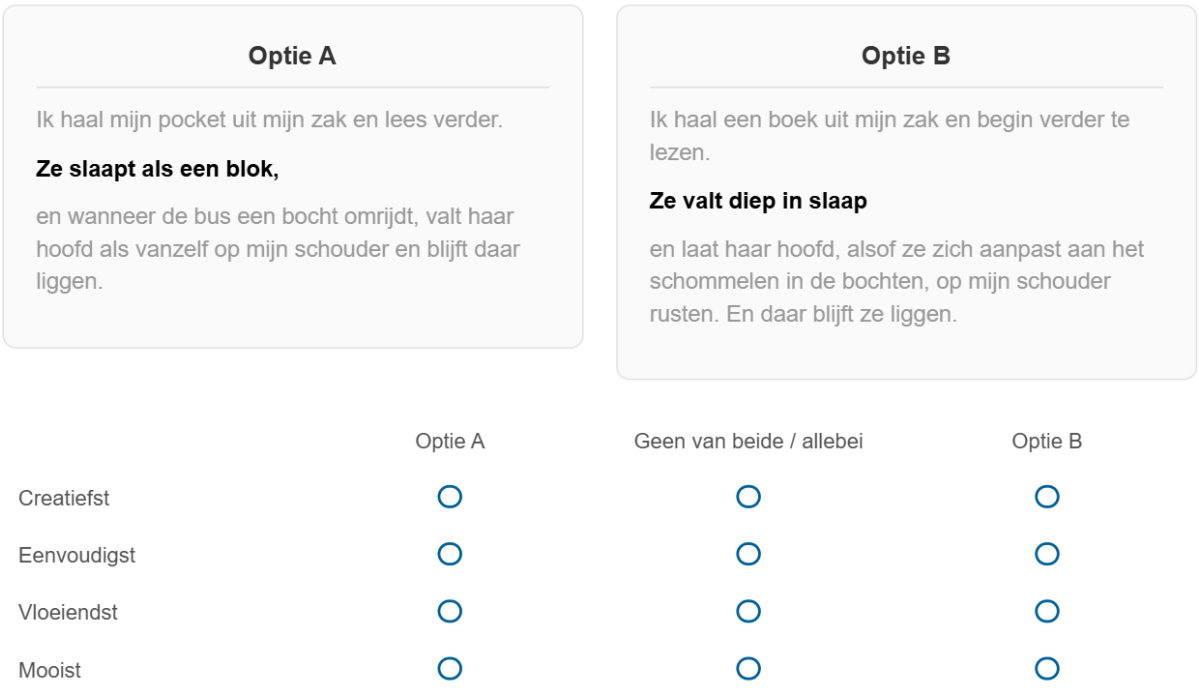


Figure 3: An example of one of the ranking questions. Human translation on the left, DeepL translation on the right. The positions of the human and DeepL translations were not the same for each question.

While the surrounding text could influence the ranking of the expressions, giving the participants a certain amount of context was argued to be more vital. Furthermore, a question at the end of the survey was included to ask participants how difficult they thought it was to rank the bolded black text purely on its own, and not the surrounding text.

Selecting the ten translation pairs was not done randomly, as this would introduce the possibility of there being not enough difference between the human and machine translations for any interesting results to be yielded. Instead, the author prioritized translation pairs where different translation methods were used while still being relatively equal in length.

After completing the survey, the participants were given a full explanation of the goal of the survey, including the information written above which was previously withheld from them.

The survey was first distributed to 18 people before being modified and distributed again after two weeks. Some of the 18 people were contacted again to redo the modified survey. Participants were mostly recruited through direct contact with the researcher and snowball sampling. The survey was also published on the survey sharing website SurveyCircle to reach more participants. Users of this website could complete the survey and receive a code to promote their own surveys.

5. Results

5.1 Text Analysis Results

5.1.1 Japanese Onomatopoeia

466 onomatopoeic expressions were extracted from the original Japanese text. Each of these expressions was tagged according to its function in the sentence it appeared in. Of the 466 expressions, 408 function as adverbs (87.7%), 50 as verbs (10.8%) and six as adjectives (1.3%). Of the remaining two expressions, one is used standalone as an explicative by a character to replicate the sound of a loud thump ("*zudon!*" in the original text) and was therefore tagged as an onomatopoeic expression. The final expression was not clearly identified as a specific function and therefore left out, resulting in a dataset of 465 expressions.

Japanese POS-Tags	*Frequency*	*Percent*
Adjective	6	1.3
Adverb	408	87.7
Onomatopoeic Expression	1	0.2
Verb	50	10.8
Total	*465*	*100.0*

Table 3: Breakdown of all extracted onomatopoeia from the Japanese text.

5.1.2 Human and DeepL Translations

The two most common translation methods found in the human translation (referred to as HT) and the DeepL translation (referred to as MT) are (1) translating as an adverb (HT: 158, MT: 248) and (2) translating as an adjective (HT: 75, MT: 95). In the HT, omitting the expression entirely (69) was the next most frequently used translation method, while for the MT it was translating as a verb (42).

Human POS-Tags	*Frequency*	*Percent*
Adjective	75	16.1
Adverb	158	34.0
Onomatopoeic Expression	3	0.6
Verb	51	11.0
Idiom	20	4.3
Noun	11	2.4
Omission	69	14.8
Change of Phrase	23	4.9
Explicative Paraphrase	55	11.8
Total	*465*	*100.0*

Table 4: Breakdown of all extracted onomatopoeia from the human translation.

DeepL POS-Tags	*Frequency*	*Percent*
Adjective	95	20.4
Adverb	248	53.3
Onomatopoeic Expression	3	0.6
Verb	42	9.0
Idiom	3	0.6
Noun	7	1.5
Omission	37	8.0
Change of Phrase	3	0.6
Explicative Paraphrase	27	5.8
Total	*465*	*100.0*

Table 5: Breakdown of all extracted onomatopoeia from the DeepL translation.

Overall, the HT contained a comparatively more even spread of translation methods compared to the MT. While the expressions translated as adverbs and adjectives in the HT account for only 50.1% of all translated expressions, this number is 73.7% for the MT. The third and fourth-most common translation methods take up 26.6% in the HT and 17.0% in the MT, further highlighting the uniformity of the MT.

The biggest disparity of usage rate between the two translations is found in the translation method of changing the phrase, where the HT text had 23 occurrences while the MT only had three. The translation method which was employed by both translators at the most similar frequency is translating the expression as an onomatopoeic expression, which happened three times in both translations.

5.1.3 Translation Methods of HT and MT

Table 6 shows a full comparison of how many times a specific translation method was preferred by both translations for each type of function available in the original Japanese text.

Japanese Word-class	*Translated Word-class*	*HT Count*	*MT Count*
Adjective	Adjective	2	4
	Explicative Paraphrase	3	0
	Idiom	0	1
	Noun	1	1
Adverb	Adjective	61	65
	Adverb	152	243
	Change of Phrase	18	3
	Explicative Paraphrase	46	25
	Idiom	16	1
	Noun	7	5
	Omission	65	36
	Onomatopoeic Expression	1	2
	Verb	42	28
Onomatopoeic Expression	Onomatopoeic Expression	1	1
Verb	Adjective	12	26
	Adverb	6	5
	Change of Phrase	5	0
	Explicative Paraphrase	6	2

	Idiom	4	1
	Noun	3	1
	Omission	4	1
	Onomatopoeic Expression	1	0
	Verb	9	14
Total		465	465

Table 6: A breakdown of each onomatopoeia and how it was translated in the HT and MT.

Of the 408 expressions labelled as an adverb in the original Japanese text, the HT translated 152 as adverbs, maintaining the word-class. This equals 96.2% of all adverbs in the HT or 32.7% of all expressions found in the HT. In the MT, the adverb word-class was maintained 243 times, being 98.0% of all translated adverbs in the MT or 52.3% of all expressions in the MT. This high consistency of source language word-classes between the original Japanese text and the HT and MT is only visible in the translations of the expressions marked as adverbs. The next biggest group of expressions in the Japanese text are verbs (50 in total), but the HT and MT translate these expressions as verbs only nine and 14 times respectively. Instead, translating the expression as an adjective was the most common translation method used in both translations (HT: 12, MT: 26). However, the HT once again shows a more even spread in the usage of different translation methods than the MT. While translating as an adjective accounted for 24% of all translated verb expressions in the HT, it was 52% for the MT. Unfortunately, too few expressions tagged as adjectives or onomatopoeic expressions in the Japanese text were available for any significant results to be seen.

5.1.4 Translating as an Adverb

In 128 instances, both the HT and MT chose to translate the same expression as an adverb. This accounts for 81.0% of all adverbs in the HT text and 51.6% of all adverbs in the MT text. Due to the overall higher adverb count of the MT, there are far more instances where the MT chose to label an expression as an adverb while the HT opted for another label (120) compared to the other way around (30). A Chi-squared test indicated that the imbalance in translating as an adverb between the HT and MT was statistically significant, χ^2 (1, N = 465) = 52.81, p < .001.

	Adverb in MT	*Not an adverb in MT*
Adverb in HT	128	30
Not an adverb in HT	120	187

Table 7: Each time the HT and MT agreed and disagreed on translating an onomatopoeic expression as an adverb.

5.1.5 Asymmetries in Translation

Table 8 exhibits the largest 'gaps' between the translation methods used by each translation. A substantial 'gap' signifies that both translations frequently chose a different, but particular, translation method when presented with an expression. For example, in 37 instances, the HT chose to omit the translation of an onomatopoeia while the MT preferred translating it as an adverb. The reverse scenario happened only 12 times, resulting in a gap-distance of 25. While this naturally indicates a higher tendency of the HT to omit onomatopoeia, it also highlights a pattern in the type of disagreements between the two translators. Incidentally, the top five translation method pairs all consist of the MT choosing to translate the expressions as adverbs while the HT defers to another method. For statistical analysis, a McNemar-Bowker test was chosen as the data concerned paired samples. Out of the total 36 possible pairs of

translation methods, ten were excluded due to there being no occurrences (e.g., not a single time was an onomatopoeia translated as an idiom versus being translated as a noun in either translation). The analysis was therefore performed on 26 pairs which resulted in an overall p-value of $p < .001$, indicating that the asymmetry across the entire dataset of 465 translated onomatopoeia was unlikely to have randomly occurred. To gain further insight into the statistical significance of each gap in the chosen translation methods, exact binomial tests were employed on the same 26 translation method pairs. To prevent any false positives from impacting the results, the Benjamini-Hochberg procedure was applied to calculate the False Discovery Rate (FDR). This provided adjusted p-values of $p < 0.01$ for the gaps of the top five pairs, indicating that these asymmetries in chosen translation methods are statistically significant. The 21 other gaps are not statistically significant as they all scored adjusted p-values above $p < 0.05$.

A	*B*	*Human A / DeepL B*	*Human B / DeepL A*	*Net Gap*	*Adjusted p-value*
Omission	Adverb	37	12	25	0.00244
Explicative Paraphrase	Adverb	24	4	20	0.00179
Change of Phrase	Adverb	16	1	15	0.00179
Verb	Adverb	16	1	15	0.00179
Idiom	Adverb	13	0	13	0.00179
Explicative Paraphrase	Adjective	11	4	7	0.46429
Idiom	Adjective	4	0	4	0.46429

Table 8: The biggest 'gaps' in chosen translation methods of HT and MT.

5.2 Survey Analysis

5.2.1 Higher preference for HT

In total, 98 participants were gathered. Of these 98, 37 had to be excluded from analysis due to either not finishing the survey or having filled in an earlier version of the survey. As mentioned earlier, a pilot-version of the survey was distributed to 18 participants. After modifying the survey, it was distributed once more roughly two weeks later. A portion of the original 18 participants were approached to fill in the final survey. All other participants had not filled in the survey before. This left a sample size of n=61. 29 participants were women, 25 were men, two were non-binary, four chose not to share this information, and one participant did not provide any answer. The largest age group was between 25 and 34 years old at 30 participants, with the next biggest group being 18- to 24-year-olds at 26 participants.

As each participant ranked ten translation pairs on four metrics, each metric has a total of 610 data points. The combined ranking results are displayed in figure 4 and table 9.

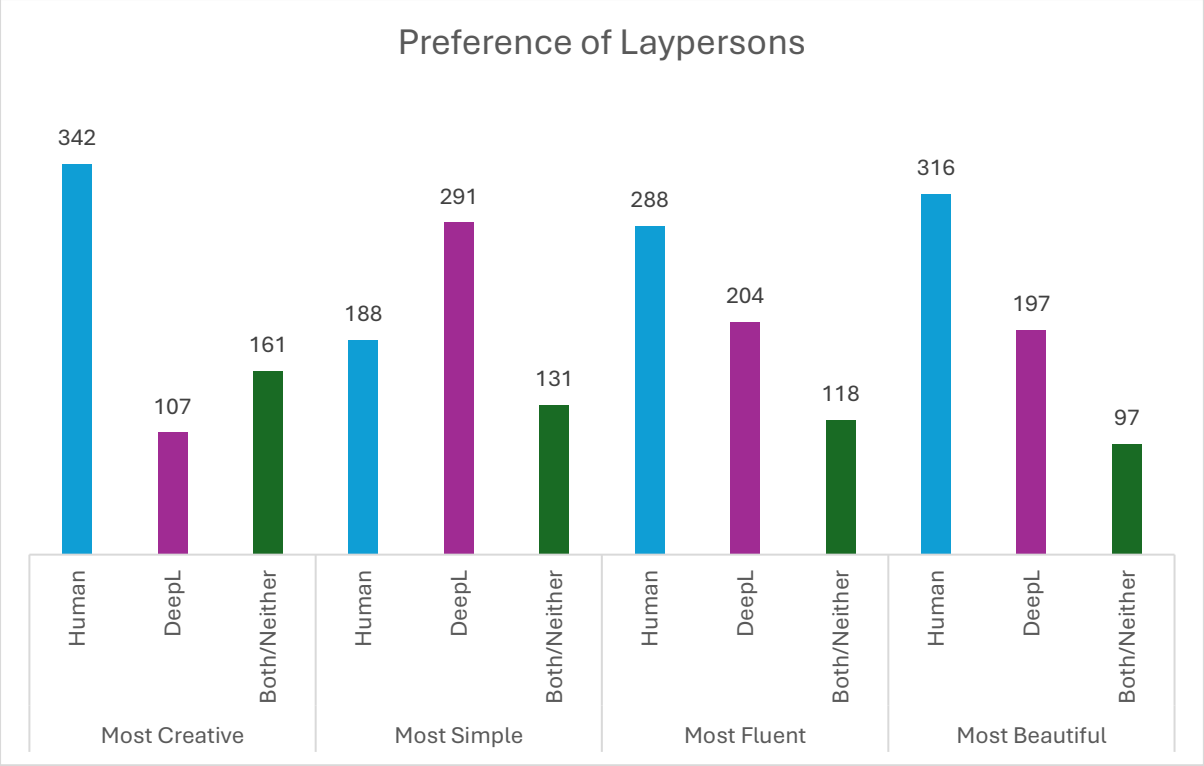


Figure 4: Preference of laypersons for each metric.

Metric		
Most Creative	Human	342
	DeepL	107
	Both/Neither	161
Most Simple	Human	188
	DeepL	291
	Both/Neither	131
Most Fluent	Human	288
	DeepL	204
	Both/Neither	118
Most Beautiful	Human	316
	DeepL	197
	Both/Neither	97

Table 9: The ranking results for each metric.

In each category exempting simplicity, the HT was consistently ranked higher than the MT. Consequently, the only metric where the MT ranked higher was simplicity.

For the creativity metric, excluding the instances where the human and DeepL translations were ranked equal (161), the human translation was ranked higher a total of 342 times, while the DeepL was only ranked higher 107 times, indicating a strong preference for the human translation. A Chi-square test confirmed that this was a statistically significant preference, χ^2 (1, N = 449) =123.00, p < .001. Similar tests for the fluency and beauty metrics also produced significant preferences for the human translations, χ^2 (1, N = 492) = 14.34, p < 0.001 & χ^2 (1, N = 513) = 27.60, p < 0.001. The only metric which showed a significant preference towards the DeepL translation was the simplicity metric, χ^2 (1, N = 479) = 22.15, p < 0.001.

Ranking both texts as equal was the least popular choice, except for the creativity metric, where it was the second most chosen ranking.

Bayesian statistical analysis was also employed. The probability of all possible answers (preference towards HT, preference towards MT, equal) being chosen a roughly equal number of times, which would indicate strong evidence for null-hypothesis II, was calculated for each metric individually (table 10). For each metric, an extremely low Bayes factor was yielded, providing strong evidence against the null-hypothesis.

Metric	*Bayes factor (BF_{01})*
Creativity	$2.557 * 10^{-29}$
Simplicity	$3.534 * 10^{-12}$
Fluency	$3.168 * 10^{-14}$
Beauty	$7.155 * 10^{-25}$

Table 10: The Bayes factor for each metric.

5.2.2 Low perceived difficulty in ranking

While the above results could indicate that participants can indeed spot differences in the translations of the onomatopoeia, there is a possibility that the surrounding, greyed-out context sentences might have influenced the participants' ranking. To address this, an extra question was asked concerning how difficult the participants themselves deemed ranking the translations based solely on the bolded segments of text. Participants were asked to rate this difficulty on a scale of one to five, with one being "very difficult" and five being "very easy". The results are displayed in figure 5. With 61 answers, the mean score was 3.89, indicating that most participants did not find the task difficult to complete. The median was 4 ("easy") and the SD was 1.14. Only 11 participants (18.0%) reported ranking the translations of the onomatopoeia "difficult" without also factoring in the surrounding context sentences. Ten participants (16.4%) reported a "neutral" level of difficulty and the remaining 40 (65.6%) found the ranking assignment either "easy" (15; 24.6%) or "very easy" (25; 41.0%). There was not a single participant who deemed the task "very difficult". A Chi-squared test verified that this perceived ease is statistically significant, χ^2 (4, N = 61) = 26.79, $p < 0.001$. Furthermore, to test if these results were significantly different from a random result, where each option was chosen an equal number of times, a Bayesian one-sample t-test was conducted. The yielded Bayes factor was ($BF_{01} = 7.189*10^{-6}$), providing evidence against this hypothesis.

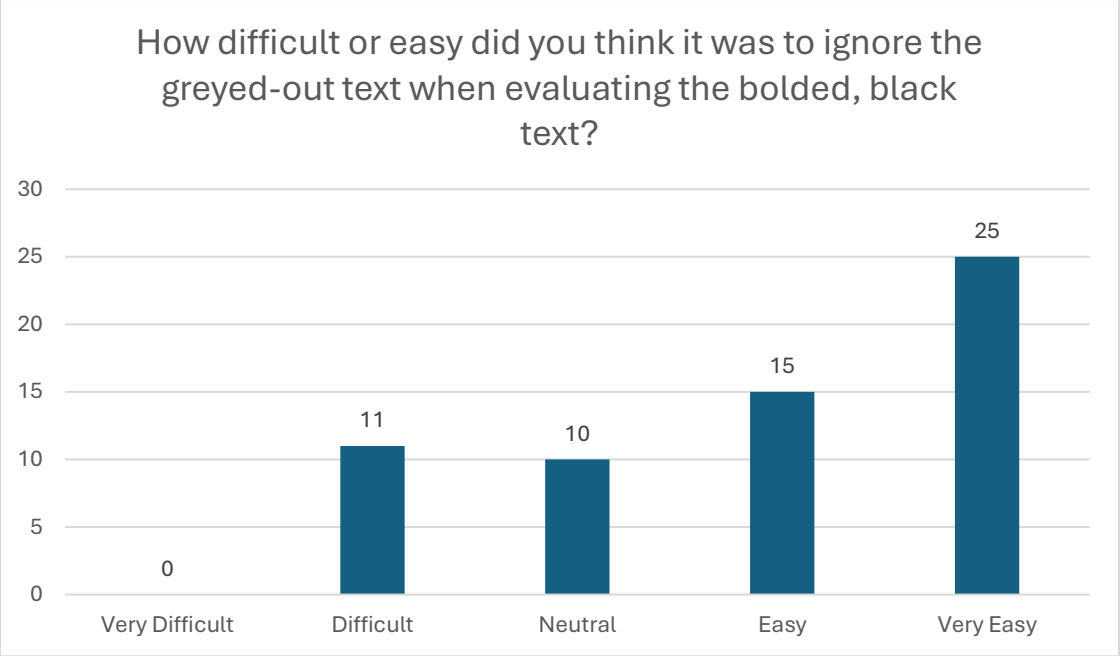


Figure 5: Most participants indicated little difficulty in ranking the bolded text on its own.

6. Discussion

6.1 Extracted expressions

There are several differences between the frequencies of certain translation methods in this study compared to previous work. Kubo (1995), for instance, reported a substantially higher 78% omission rate in the translation of Japanese to English onomatopoeia, while the number in this study is 14.8% in the case of the human translation and a mere 8.0% for the DeepL translation. Kubo's research seems to be an exception to the rule, however, as both Inose (2008) and Minashima (2004) report an omission rate of 19.3% and 16.3% respectively in Japanese to English translations. Finally, Flyxe (2002) documented a low 5.9% omission rate, although this was a Japanese to Swedish translation. Another peculiar difference was the frequency of translating onomatopoeia as adverbs. Inose (2008) extracted 295 onomatopoeic expressions and only 21% of those were translated as adverbs in Spanish, with the percentage being a mere 16.3% when translated into English. This research reported an adverb ratio of 34.0% for the HT and 53.3% for the MT, both being significantly more than the translations in Inose's research (the difference between the HT and MT is discussed in section 5.2). Because Inose also used novels written by Haruki Murakami, the disparity was unexpected. A source for this contrast may be found in the targeted language pairs, as this research covers Japanese to Dutch translation, not Spanish or English. Additional research focused on the Dutch translation of Japanese novels in particular could help further our understanding. Furthermore, although a similar classification system was used when tagging the translated expressions, having two different people classify the expressions introduces the possibility of inconsistent tagging.

6.2 MT evaluation

Before discussing the results of the analysis concerning the different translations, the decision to choose the DeepL Next-Gen translation as the representative translation for all MT systems should first be reflected upon. While all MT systems scored relatively low across all evaluation metrics, it is important to realize that a low score does not necessarily mean that the quality of the translation is poor. The BLEU metric, for example, rates translations higher the more the text is similar to the reference text (in this case, the human translation). This means that a difference in wording impacts the scoring negatively even if the meaning is maintained. Consequently, while the DeepL Next-Gen translation was chosen due to it scoring the highest across all metrics, a human evaluator might have preferred one of the other machine translations. This could have greatly impacted not just the frequency of certain translation methods, but also the ranking done by the survey participants. Further research would benefit from either employing human evaluation of the machine translations before deciding which should be the representative translation, or conducting the same analysis for all machine translations rather than just one.

6.3 MT tendencies

The results show a clear tendency to translate the Japanese adverbial expressions as adverbs in both the human and DeepL translations, therefore maintaining the word-class of the onomatopoeic expression. However, this tendency was significantly higher in the DeepL translation. This is consistent with research done by Luo et al. (2024) who discovered less morphosyntactic diversity (the altering of word-classes included) in machine translations by NMT systems compared to those done by humans, suggesting that the translations are more literal. LLM-based systems, like the one used in this paper, have shown to exhibit translation tendencies more like those of human translators, being less literal than their NMT counterparts, particularly when translating idiomatic expressions (Raunak et al. 2023). However, when comparing the output of LLM-based translation systems to that of human translators, Yan et al. (2026) still reported that the LLM translations were often more literal when compared to the human translators, who exhibited other types of translation quirks which are discussed in section 6.3.

The results of this paper provide evidence that DeepL is indeed more prone to maintain the word-class of onomatopoeia when translating. However, it is prudent to keep in mind that this was just in the case of adverbs. The most frequent translation method for onomatopoeic expressions used as verbs, for example, is translating as an adjective. Again, this applied to both the human and DeepL translation. Unfortunately, there were not enough expressions used as adjectives or other functions present in the original Japanese text to ascertain whether this tendency to maintain the word-class is also present concerning expressions that are not used as an adverb in Japanese. Additionally, the differing rates at which onomatopoeia were translated as adverbs in this research and that of Inose's (2008) could also be partly explained by the different language pairs used. Therefore, the only safe conclusion this paper can make is that MT systems are indeed more prone to maintain the word-class of onomatopoeic expressions in Japanese to Dutch translation if the expression was used as an adverb, which supports rejecting the null-hypothesis for sub-question I. Further research should strive to analyse datasets with a more varied mix of word-classes in the Japanese source text.

6.4 HT tendencies

While the MT showed a strong tendency to maintain the word-class of adverbial onomatopoeia in the source text, the HT was filled with a bigger variety of translation methods. The methods which clashed most often with those found in the MT are translating as (1) omission (no translation), (2) explicative paraphrase, (3) change of phrase, (4) verb, and (5) idiom. Possible explanations behind the different translation methodologies of the HT and MT are easier to speculate for some than others. Translating as an idiom, for example, seems like a translation method more often utilized by a human rather than a machine due to them being expressions that have meanings which are not always the sum of the meanings of the words they consist of (e.g., something being 'a piece of cake' does not necessarily mean it is an oven-baked treat). Verbs, on the other hand, are not known to contain double meanings in a similar fashion. Furthermore, idioms are also prone to contain cultural or historical links to the country where the language is mostly spoken. Take, for instance, the English idiom 'going postal', which means becoming extremely angry and/or violent. This idiom is derived from a series of fatal incidents in the United States in the eighties and nineties where postal delivery workers were involved in mass shootings (Polyezhayev et al., 2024). Research by Yang et al. (2025) indicates that MT systems have a harder time translating Chinese idiomatic expressions into English, although they do not mention this combination of double-meanings and cultural links as a possible reason. While the topic at hand concerns translating onomatopoeia into idioms, one possible interpretation is that these reasons might play a role in MT systems struggling to employ idioms as translations for non-idiomatic expressions.

The HT's comparatively more common reliance on omitting a translation or changing the phrase might also be explained by analyzing the translation tendencies of MT systems. Yan et al. (2026) compared the translation tendencies and common errors of two MT systems with those of human translators of three different skill-levels. Their results indicated that the human translators had a higher tendency to omit words in their translation compared to the MT systems, which concurs with the relatively high number of omissions and changes of phrase in this research. The same paper also identifies over-interpreting as another characteristic of the human translators, which might have a connection to the higher rate of explicative paraphrases in this research.

6.5 Perception of laypersons

The survey results show a clearly higher score for the human translation in creativity and fluency. As mentioned before, while definitions were given for both creativity and fluency, because participants were also instructed that no answer was necessarily incorrect, there likely exists a difference between the definitions of the qualities participants had in mind when ranking the translations. Nevertheless, the difference in ranking between the HT and MT was statistically significant in size and consistency. The same can be said for simplicity of the translations, as the MT was ranked simpler significantly more often than the HT.

Another interesting result is that the option to rank both translations as equal was the least chosen option in every case except for creativity, where it was the second most chosen option. This further indicates that, specifically regarding simplicity and fluency, a clear difference can be indicated by laypeople between the human and machine translations of Japanese onomatopoeia to Dutch. The question regarding the difficulty of ranking the translations purely on the words relating to the translation of the onomatopoeia (the bolded

text) provided extra support for this result, as most participants indicated little to no difficulty in completing this task. Consequently, the null-hypothesis for sub-question II can be rejected as a statistically significant and, most importantly, consistent difference in ranking was measured.

Arenas & Toral (2022) showed that participants rated human translations more creative than both machine translations and post-edited machine translations (although the post-edited machine translations did score higher than the non-post-edited translations). While not focused on the translation of onomatopoeia and not even tackling the same language pair, this does provide similar results to the findings of this research, namely that machine-translated texts are perceived as less creative than human translations. On the topic of fluency, Ferstler et al. (2026) showed that in a task where participants had to compare the reading experience of a human and machine-translated text, while both translations were sufficiently readable, the human translation was deemed more fluent while the machine translation was sometimes criticized for necessitating a rereading of sentences. Finally, concerning simplicity, Niu & Jiang's (2024) research corroborates a higher perceived simplicity rating for machine translations as they discover that, using lexicality as a proxy, Chinese to English machine translations were consistently simpler than human translations across three separate genres of text (contemporary novels, government documents, academic abstracts).

6.6 Perceived beauty

Survey results showed that participants ranked the human translation more beautiful than the machine translations significantly more often. This indicates a preference for the human translation when it comes to the subjective metric of beauty, disproving the null-hypothesis for sub-question III. What this beauty is, exactly, was purposefully left open to the interpretation of the participant. Ferstler et al. (2026) conducted research which also asked participants to rate a human and machine translation of literary texts. Their results showed that, while not focusing on onomatopoeic expressions, human translations received more literary praise and were often labelled as being more engaging and immersive. While terms such as literary, engaging, and immersive are not equal to beauty, it does show that this preference for human translations over machine translations outside the basic requirements of readability and such is not a discovery limited to this study. On the other hand, research done by Porter & Machery (2024) about how good people are at distinguishing between AI-generated and human-written poems showed that participants were not able to identify the author of the poem and, more importantly, ranked AI-generated poems higher than human-written poems in specific qualities including beauty. It is important to remember, however, that the poems that were the focal point of this study were not translated texts. One could therefore ponder that the ability to equal or even surpass humans in producing beautiful or literary texts is lost when translation is also a required step. Further research could be done on this topic and how this impacts the translation of onomatopoeia in particular.

It is important to mention that the results discussed in the previous two subsections were greatly influenced by the ten specific translation pairs chosen by the author for the survey. Other selection methods prioritizing not a difference in translation methods but another aspect of the translation such as a difference in complexity, length, etc., might have resulted in completely different rankings from the participants. The same can be said having the sentences be selected randomly, as translation pairs where the machine translation, for instance, makes an obvious mistake (e.g., different pronoun usage within the same sentence)

could negatively affect the ranking scores even though it would have no bearing on the quality of the translation of the onomatopoeic expression.

7. Conclusion

The purpose of this research was to examine the differences between the Japanese to Dutch translations of onomatopoeia by human and machine translators. By comparing translations of Haruki Murakami's *Kafka on the Shore*, a significant disparity in the frequency of translation methods was discovered. Firstly, while the two most chosen translation methods by both the human and machine translator were translating as an adverb or adjective, the human translator showed a much wider variety of translation methods than the machine translator, such as omitting the expression or replacing it with an idiom. Furthermore, the machine translator showed a significantly stronger tendency to maintain the word-class of an onomatopoeia if it was originally used as an adverb. These two points indicate that the machine translator is more inclined to rely on more literal translations compared to the human translator, who demonstrated a freer approach to the translation of the onomatopoeic expressions.

Survey analysis concluded that laypeople ranked the machine-translated onomatopoeia as less creative, less fluent, less beautiful, but simpler than the human translated onomatopoeia significantly more often, even without knowing who translated the text. This goes against some previous research which provides evidence for laypeople not being able to distinguish between human and machine translations.

This research was constrained by a limited research period, and further attempts would benefit massively from an even larger dataset across not just Japanese novels, but also other domains of text such as non-fiction works. This might allow trends in translation methods to be found in the translations of word-classes other than just adverbs. Furthermore, proper and thorough implementation of NLP techniques has the potential to provide a new standpoint to analyse this data from.

8. Sources

Akallouch, O., & Fardousse, K. (2025). In-Context learning for low-resource machine translation: A study on Tarifit with large language models. *Algorithms*, *18*(8), 489. https://doi.org/10.3390/a18080489

Baek, J., Lee, S. J., Gupta, P., Oh, G., Dalmia, S., & Kolhar, P. (2024). *Revisiting In-Context Learning with Long Context Language Models*. arXiv. https://doi.org/10.48550/arxiv.2412.16926

Bawden, R., & Yvon, F. (2023, March 3). *Investigating the translation performance of a large multilingual language model: the case of BLOOM*. arXiv. https://doi.org/10.48550/arxiv.2303.01911

Bertsch, A., Ivgi, M., Xiao, E., Alon, U., Berant, J., Gormley, M. R., & Neubig, G. (2025). In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational*

Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 12119–12149). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.605

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://doi.org/10.48550/arXiv.2005.14165

Dingemanse, M. (2013). Ideophones and gesture in everyday speech. *Gesture*, *13*(2), 143–165. https://doi.org/10.1075/gest.13.2.02din

Dingemanse, M., Schuerman, W., Reinisch, E., Tufvesson, S., & Mitterer, H. (2016). What sound symbolism can and cannot do: Testing the iconicity of ideophones from five languages. *Language (Baltimore)*, *92*(2), e117–e133. https://doi.org/10.1353/lan.2016.0034

Dingemanse, M (2019). 'Ideophone' as a comparative concept. In Akita, K., & Pardeshi, P. (Eds.), *Ideophones, Mimetics and Expressives* (Vol. 16, 13–33). John Benjamins Publishing Company. https://doi.org/10.1075/ill.16.02din

Electronic Dictionary Research and Development Group. (1999). *JMdict* [Online dictionary]. https://www.edrdg.org/wiki/index.php/JMdict-EDICT_Dictionary_Project

Ferstler, Y., Podoxin, A., Brassington, T., Grundkiewicz, R., Taboada, M., & Karpinska, M. (2026). *AI translation of literary texts is "fine", but readers still prefer human translations*. arXiv. https://doi.org/10.48550/arXiv.2606.26040

Flyxe, M. (2002). Translation of Japanese onomatopoeia into Swedish with focus on lexicalization. *Africa & Asia*, *2*, 54–73.

Fukushima, H., Araki, K., & Uchida, Y. (2014). Disambiguation of Japanese Onomatopoeias Using Nouns and Verbs. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue* (Vol. 8655, pp. 141–149). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-10816-2_18

Grand View Research. (2026). Machine translation market size, share & trends analysis report by technology (RBMT, SMT, NMT), by deployment (cloud, on-premise), by enterprise size, by application, by region, and segment forecasts, 2026–2033 (Report ID: GVR-1-68038-782-9). https://www.grandviewresearch.com/industry-analysis/machine-translation-market

Guerberof-Arenas, A., & Toral Ruiz, A. (2022). Creativity in translation: Machine translation as a constraint for literary texts. *Translation Spaces*, *11*(2), 184–212. https://doi.org/10.1075/ts.21025.gue

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T.-Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., … Zhou, M. (2018). *Achieving Human Parity on Automatic Chinese to English News Translation*. arXiv. https://doi.org/10.48550/arxiv.1803.05567

Hida, Y., & Asada, H. (2018). *Gendai giongo gitaigo yōhō jiten [Dictionary of contemporary onomatopoeia and mimetic word usage]*. Tōkyōdō Shuppan.

IBM. (2026). What is in-context learning (ICL)? https://www.ibm.com/think/topics/in-context-learning

Iida, H., & Akita, K. (2024). Iconicity emerges from language experience: Evidence from Japanese ideophones and their English equivalents. *Cognitive Science*, *48*(12), e70031-n/a. https://doi.org/10.1111/cogs.70031

Inose, H. (2008). Translating Japanese onomatopoeia and mimetic words. In A. Pym & A. Perekrestenko (Eds.), *Translation research projects 1* (pp. 97–116). Intercultural Studies Group. http://isg.urv.es/library/papers/InoseOnomatopoeia.pdf

Karpinska, M., & Iyyer, M. (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* (pp. 419–451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.wmt-1.41

Khoboko, P. W., Marivate, V., & Sefara, J. (2025). Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models. *Machine Learning with Applications*, *20*, Article 100649. https://doi.org/10.1016/j.mlwa.2025.100649

Kubo, A. (1995). Miyazawa Kenji no onomatope no sekai [The world of onomatopoeia in Miyazawa Kenji's work]. *Kobe Kaisei Review, 34*, 17–30.

Läubli, S., Sennrich, R., & Volk, M. (2018). *Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation*. arXiv. https://doi.org/10.48550/arxiv.1808.07048

Laviosa-Braithwaite, S. (1998). Universals of translation. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 288–291). Routledge.

Liu, C., Zhang, W., Zhao, Y., Luu, A. T., & Bing, L. (2025). Is translation all you need? A study on solving multilingual tasks with large language models. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 9594–9614). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-long.485

Luo, J., Cherry, C., & Foster, G. (2024). To diverge or not to diverge: A morphosyntactic perspective on machine translation vs human translation. *Transactions of the Association for Computational Linguistics, 12*, 355–371. https://doi.org/10.1162/tacl_a_00645

Minashima, H. (2004). Nichi-ei-go no onomatope [Onomatopoeias in Japanese and English*]. Fukui Daigaku Kyouiku Chiikikagakubu Kiyo I [Fukui University Faculty of Education and Regional Studies Journal I], 60*, 95–115. https://u-fukui.repo.nii.ac.jp/records/26543

Murakami, H. (2002). *Umibe no Kafuka [Kafka on the shore]* (Vol. 1). Shinchosha.

Murakami, H. (2006). *Kafka op het strand* (J. Westerhoven, Trans.). Uitgeverij Atlas. (Original work published 2002)

Nanoskript. (2023). *Japanese onomatopoeia search (jp-onomatopoeia)* [Data set]. GitHub. https://github.com/nanoskript/jp-onomatopoeia

Niu, J., & Jiang, Y. (2024). Does simplification hold true for machine translations? A corpus-based analysis of lexical diversity in text varieties across genres. *Humanities and Social Sciences Communications, 11*, Article 480. https://doi.org/10.1057/s41599-024-02986-7

Polyezhayev, Y., Maksymova, A., Tytar, O., Kulichenko, A., & Rukolyanska, N. (2024). Ethnolinguistics as a tool for studying the cultural heritage of the world's peoples. *Forum for Linguistic Studies, 6*(5), 287–302. https://doi.org/10.30564/fls.v6i5.6830

Porter, B., & Machery, E. (2024). AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports, 14*, Article 26133. https://doi.org/10.1038/s41598-024-76900-1

Raunak, V., Menezes, A., Post, M., & Hassan, H. (2023). Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 1041–1050). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-short.90

Sasamoto, R. (2021). Onomatopoeia, translation and relevance. *Pragmatics & Cognition*, *28*(2), 347–375. https://doi.org/10.1075/pc.21011.sas

Shin, J.-S., & Mah, S. (2024). Exploring user perspectives: Student evaluation of the evolution in neural machine translation from English to Korean and its implications. *FORUM. Revue Internationale d'interprétation et de Traduction / International Journal of Interpretation and Translation*, *22*(1), 77–95. https://doi.org/10.1075/forum.00036.shi

Suzuki, M., Ishikawa, Y., Teraguchi, M., & Sakaji, H. (2025) Sentiment works in small-cap stocks: Japanese stock's sentiment with language models. *International Journal of Information Management Data Insights, 5*(1), 1–13. https://doi.org/10.1016/j.jjimei.2024.100318

Telidis, D. (2025, July 22). Case study: Netflix's AI-powered multilingual content localization. *Localization Institute*. https://www.localizationinstitute.com/case-study-netflixs-ai-powered-multilingual-content-localization/

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). *Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation*. arXiv. https://doi.org/10.48550/arxiv.1808.10432

Vanroy, B., Tezcan, A., & Macken, L. (2023). MATEO: MAchine Translation Evaluation Online. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation* (pp. 499–500). European Association for Machine Translation (EAMT).

Yan, J., Yan, P., Chen, Y., Li, J., Zhu, X., & Zhang, Y. (2026). Benchmarking LLMs against human translators: A comprehensive evaluation across languages, domains, and expertise levels. *IEEE Transactions on Big Data, 12*(3), 801–813. https://doi.org/10.1109/TBDATA.2025.3644594

Yang, C., Dou, Y., Heineman, D., Wu, X., & Xu, W. (2025). *Evaluating LLMs on Chinese idiom translation*. arXiv. https://doi.org/10.48550/arXiv.2508.10421

Zhang, W., Chino, A., Yan, S., & Teraoka, T. (2024). Construction of Japanese-Chinese onomatopoeia corpus based on events and behaviors. In *Human Factors in Design, Engineering, and Computing* (Vol. 159, pp. 158–167). AHFE International. https://doi.org/10.54941/ahfe1005578

Zhu, S., Cui, M., & Xiong, D. (2024). Towards robust in-context learning for machine translation with large language models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16619–16629). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1444/