# The Translator's Canvas: Using LLMs to Enhance Poetry Translation

**Natalia Resende** resenden@tcd.ie
Trinity Centre for Literary and Cultural Translation, School of Languages, Literatures and Cultural Studies, Trinity College Dublin, D02 CH22, Ireland
**James Hadley** hadleyj@tcd.ie
Trinity Centre for Literary and Cultural Translation, School of Languages, Literatures and Cultural Studies, Trinity College Dublin, D02 CH22, Ireland

## Abstract

We explore the potential of LLMs to enhance the translation process of rhymed and non-rhymed poetry. We examine LLMs' performance (ChatGPT-3.5, ChatGPT-4, Google Gemini) in terms of lexical variety, lexical density, and sentence length compared to human translations (HT). We also examine the models' abilities to translate sonnets while preserving the rhyme scheme of the source text. Our findings suggest that LLMs can serve as valuable tools for literary translators, assisting with the creative process and suggesting solutions to problems that may not otherwise have been considered. However, if the paradigm is flipped, such that instead of the systems being as tools by human translators, humans are used to post-edit the outputs to a standard comparable to the published translations, the amount of work required to complete the post-editing stage may outweigh any benefits associated with using machine translation in the first place.

## 1 Introduction

The translation of poetry has long been a contentious issue in the field of literary translation (Jones, 1986). The debate stems from the challenges inherent to translating poetry, which, depending on the specific poetic form in evidence, may require a delicate balancing act of content, style, tone, various types of phonetic devices, such as rhyme. Differences in language and poetic tradition may necessitate compromises and creative solutions with many competing constraints making the translation of poetry a highly complex activity. Literary translation has historically been regarded as the "last bastion of human translation" (Toral and Way, 2014), and poetry translation could be thought of as the most extreme example of this phenomenon. However, recent advances in the widespread availability of Large Language Models (LLMs) have shifted the conversation to ask in what ways human translators might make use of electronic tools in the negotiating of literary translation's stylistic and technical complexities.

Much of this work to date has focused entirely on prose, and while advancements have been substantial in this respect, much less attention has fallen onto poetry in general, and formal poetry in particular. Thus, formal poetry, simultaneously combining as it does many of the stylistic features that are known to complicate machine translation, remains an extreme challenge. Nonetheless, the emergence of web-based LLMs offers new opportunities. These models, such as ChatGPT and Google Gemini, enable the customisation of translated outputs through prompt engineering (Amatriain, 2024), whereby users can specify in detail aspects of a text to focus on, change or omit. This capacity sets LLMs apart from the web-based Neural Machine Translation (NMT) systems which have been the mainstay of the machine translation systems (MT) widely available to literary translators for the past decade or so. Generally speaking, such NMTs are limited to one or a small number of similar outputs for any given input, with little or no functionality to tailor the translation process around, for example, register, addressee, style, or tone. When it comes to complex operations such as rhyming or counting syllables, specialised systems would be required.

The focus in this discussion falls predominantly on web-based systems which are either free to use or financially accessible, because the funding associated with any given literary translation project is so limited (des Associations de Traducteurs Littéraires , CEATL). Thus, it is unrealistic to imagine literary translators being in a position to invest in expensive bespoke tools which may require training for the facilitation of their work, and therefore, the most realistic use cases when it comes to literary translation centre on systems which are easily and cheaply accessible.

For these reasons, the widespread emergence of LLMs such as ChatGPT and Google Gemini, with their free entry points, represents a significant opportunity for analysing whether such systems might be useful tools for literary translators. Here there is an important distinction to be made between Literary Machine Translation (LMT) and Computer-Assisted Literary Translation (CALT). On the one hand, LMT conceptualises the machine at the centre of the process of producing translated outputs, generally with one or more human beings supporting its work through pre-editing or post-editing. In this view, quality assessment reaches for the ultimate goal of producing outputs of the same standard as human translators (Koponen, 2016). On the other hand, CALT conceptualises the human translator as the primary agent, who makes use of the machine as a tool. In this view, the human translators may use the machine to translate only individual or isolated parts of the text, may use iteration to produce multiple versions of the same passage, and may wholly disregard the outputs of the machine if a better solution is found elsewhere.

The focus of each perspective is reversed. In LMT, the goal is to maximise the quality of the output to minimise editing work by the human. In CALT, the goal is to support human translators in their own idiosyncratic workflows, identifying and trailing possible solutions to translation challenges, and further stimulating human translators' creativity. Thus, whereas in an LMT workflow, producing multiple outputs of the same text, may be perceived as wasteful, because this would imply that each output would also need to be post-edited, in the context of CALT, producing multiple outputs of the same text or text fragment could perceivable be useful for a human translator who may use the machine's outputs more as inspiration than as something approaching a product to be refined.

While there is evidence supporting LLMs in the translation of prose works, especially novels (Karpinska and Iyyer, 2023), their impact on poetry translation remains under-explored. Thus, it is unclear how machine translations produced with the help of LLMs compare stylistically to human translations. Asking about these comparisons is fundamental to assessing whether and how LLMs might be made useful by practising translators of literature, and especially poetry.

To address this question, our initial step involves extracting and examining linguistic features at both the syntactic and lexical level from poems, as well as from translations of them produced by humans and by LLMs in Portuguese and Spanish.

## 2 Related work

The methodological approach used here is one which analyses and compares the stylistic features of translated text using Natural Language Processing (NLP) techniques. In each case, the candidate translations by each of the LLMs is compared side by side with previously published human translations of the same text. This approach is grounded in a body of literature that has developed since the 1980s to explore the distinctive stylistic characteristics of translated texts primarily on statistical terms. This literature is theoretically rooted in Toury's translation norms (Toury, 1980) which posits that translation is a culturally-bound phenomenon which functions different in different human contexts, and Baker's translation universals (Baker, 1996), which identify aspects of texts which anecdotal experience can allow us to identify translated from non-translated work. The approach responds to these two somewhat subjective theoretical constructs with corpus linguistics and NLP methods, which allow for the results to be statistics-based, and repeatable (Ilise et al., 2010; Ilisei and Inkpen, 2011; Pastor et al., 2008). The research conducted with these methods, has consistently shown that translated text does indeed tend to exhibit simpler syntax and less varied vocabulary than non-translated text (Laviosa, 2002;

Baroni and Bernardini, 2006; Pastor et al., 2008; Volansky et al., 2013a). This phenomenon, often referred to as *translationese* in the literature, is frequently associated with lower quality text, characterized by foreign-sounding and awkward wording and structure (Volansky et al., 2013b; Kunilovskaya and Lapshinova-Koltunski, 2019).

With the advent of MT systems, attention has shifted towards the stylistic features of machine-translated and post-edited texts (Daems et al., 2017; Toral, 2019; Castilho et al., 2019; Castilho and Resende, 2022). This research has gone on to show significant differences in style and content richness between human-translated text and machine-translated text, especially that produced by NMT systems (Castilho et al., 2019; Castilho and Resende, 2022).

Recent research has begun evaluating the translation capabilities of large language models (LLMs) in both literary and technical texts, often in comparison with NMT systems (Peng et al., 2023; Hendy et al., 2023; Karpinska and Iyyer, 2023). Preliminary studies (Cruys, 2023; Roos, 2023) have explored LLMs' ability to preserve the rhyme schemes in poetry translation, focusing on qualitative analysis of a single poem. However, there is a lack of research on how the range of stylistic features which come together to embody poetic texts are managed by LLMs, and how or whether LLMs might be made useful to practising human translators of poetry. This study represents an initial attempt to address this gap in the literature. Given the exploratory nature of the present study, the focus is on a limited corpus to provide a preliminary assessment of the the place these tools could have in a poetry translation workflow, setting the stage for more extensive research in the future.

## 2.1 Methodology

This analysis conducts a statistics-based stylistic comparison of features extracted from source poems, existing translations of the same poems produced by humans, and newly produced translations by three LLMs. The extracted features are both lexical and syntactic in nature, and designed to assess aspects of the formal qualities of the poems which a reader may not necessarily be consciously aware of when reading the texts, but which have an overall effect on the texts' literary qualities (Pynte et al., 2008). The specific questions addressed are:

1. How do the syntactic and lexical stylistic patterns of LLM-translated poetry compare to those of human-translated poetry?
2. How do these stylistic patterns vary between LLMs? Are there identifiable trends and/or deviations unique to each language model?
3. How do these qualities compare between formally constrained poems and free verse poems?

### 2.1.1 Corpus

In order to address these questions, this study draws on a small corpus of four published poems. Two of the poems are written in Spanish and two are written in Portuguese. Two of the poems are sonnets and two are free verse poems. A digital version of each poem was either collected from an online resource or was created by digitising a printed version. The choice to include two source languages allows for the comparison of similar features from different sources. The choice to include sonnets and free verse poems allows for the analysis of formal features both under the heavy formal constraint of a complex rhyme scheme, as found in the sonnet form, and under less constrained circumstances in the case of free verse. Before conducting the experiments, at least one published human translation into English of each poem was identified which was also collected in the same way as the source texts. To adhere to copyright laws and ethical standards, only texts not protected by copyright at the time of writing were included in the corpus. Table 1 shows the poems included in the corpus, along with each poem's short name, used in these experiments:

### 2.1.2 Examining the stylistic features

The poems in the corpus were translated using three large language models (LLMs) accessible online: ChatGPT-3.5, ChatGPT-4, and Google Gemini. A zero-shot approach was employed, instructing the models to translate the source poems into English without any prior training or fine-tuning. This method aimed to evaluate how the models perform in a realistic setting, assuming that most practising literary translators would not rely on advanced prompt engineering techniques. The following prompt was used with all the models under analysis, and with each of the poems in question:

| **Poem** | **Author** | **Year of Composition** | **Type** | **Language** | **Short Name** |
|---|---|---|---|---|---|
| *José* | Carlos Drummond Andrade | 1942 | Free verse | Portuguese | José |
| *Soneto da Fidelidade* | Vinícius de Morais | 1939 | Sonnet | Portuguese | Fidelidade |
| *Amor constante más allá de la muerte* | Francisco de Quevedo | 1648 | Sonnet | Spanish | Amor |
| *Corazón Coraza* | Mario Benedetti | 1939 | Free verse | Spanish | Corazón |

Table 1: Selected poems included in the corpus

**Prompt 1:**

*Translate this poem into English*

Next, the stylistic features of the translated output were compared with the versions in the human translations. This approach did not assume the human translations to be the correct, the only possible, or the only viable renditions of the poems in question into the target language. Nor was it assumed that all the features of the source texts were uniformly included in their human-translated versions. Rather, noting the features which were and were not included in the human-translated versions gives a basis of comparison between the versions translated by the various systems with what can be considered the current state of the art, in the form of the human translations. For this analysis the following features were extracted from the translated texts:

- Lexical richness
- Lexical density
- Sentence length in words
- Vocabulary overlap
- Rhyming patterns

The stylistic features were extracted from the texts using custom Python scripts. To assess lexical variety, which reflects the diversity of vocabulary in a text, the type/token ratio was calculated by dividing the total number of unique words (types) by the total number of words in each text (tokens).

$$\text{TTR} = \frac{N_t}{N_w} \tag{1}$$

where:

* $N_t$ represents the number of unique words (types) in the text.

* $N_w$ represents the total number of words (tokens) in the text.

Lexical density is a measure of the informational content within a text. It reflects the proportion of content words, relative to the total number of words. Content words are typically defined as nouns, verbs, adjectives, and adverbs, which carry the core message of a sentence.

Mathematically, the lexical density can be expressed as:

$$\text{LD} = \frac{N_c}{N_w} \tag{2}$$

where:

* $N_c$ represents the number of content words in the text.

* $N_w$ represents the total number of words (tokens) in the text.

Sentence length was calculated by counting the number of words between each set of sentence markers.

For the vocabulary overlap analysis, we identified the words present in the human-translated versions that were absent in the LLM-produced translations. This metric was chosen because, on our view, it provides a clearer, more intuitive understanding of the differences between texts in percentage terms. To complement this approach, we also calculated BLEU scores using the NLTK package (Bird et al., 2009), which provided valuable additional insights into the comparative performance of the translations. Finally, to assess the LLMs' ability to reproduce the rhyme schemes of the two sonnets in the corpus, the outputs in each case were categorised using the standard letter-based notation associated with line-end rhyming patterns. In this notation system, the final phonemes of each line of poetry are assigned an alphabetic value which marks the other lines in the same poem with a rhyming phoneme. An example can be seen in the opening stanza of Wordsworth's 1802 Lyrical poem, *Daffodils*:

I wandered lonely as a cloud<br>
That floats on high o'er vales and hills,<br>
When all at once I saw a crowd,<br>
A host, of golden daffodils;<br>
Beside the lake, beneath the trees,<br>
Fluttering and dancing in the breeze.

The lines can be annotated as: ABABCC, because the first and third lines rhyme phonetically, as do the second and fourth lines, and the fifth and sixth lines. However, it is important to note that as in this example, especially in languages with idiosyncratic orthographic conventions like English, rhyming phonetic values do not always correspond to similar spellings.

The first attempt at translating the sonnets with the straightforward prompt resulted in the rhyme scheme of each poem being ignored by the system. Therefore, the prompt was subsequently refined to target this aspect of the texts' stylistics more specifically. The subsequent prompt used for these tasks was:

**Prompt 2:**

*Can you improve the translation so that it maintains the same rhyme scheme as the source text?*

A few-shot prompting approach, complete with demonstration examples was also employed to facilitate comparison of the outputs achieved from each prompt technique. This strategy involved explaining the rhyme scheme of each the poems by providing examples of word that rhyme within the poem and also specifying the organisation of rhyming words in the poem flow. The advanced prompts designed for the translation of the sonnets can be found in Appendix A.

## 2.2 Results

The vocabulary overlap experiment asked: *How many words in the HT are not present in the version translated by the LLMs?* To address this question, the number of words present in the HT and not present in the outputs produced by the LLMs was calculated, shown in Table 2.

These results demonstrate that the proportion of vocabulary in the human translation that diverges from the model is higher in the sonnets than in the free verse poems. This finding is predictable, because of the additional formal constraints imposed by the sonnet form, compared to the free verse form. Key formal constraints in this respect include the need to rhyme and to fit lines into specific lengths. These constraints raise the complexity of the translation task, implying a higher synonym and paraphrase usage than in translation problems where these formal constraints are not present. Thus, it is possible to speculate that the greater variation in word choice observable in the models outputs in the case of the sonnets is linked to this additional layer of complexity.

Table 3 presents the BLEU scores, which are consistent with the findings from the vocabulary overlap experiment. The scores indicate that free verse poems consistently achieve higher BLEU scores than sonnets, suggesting closer lexical alignment observed between machine-translated free verse poems and their human-translated counterparts in contrast with greater vocabulary variation in machine-translated sonnets compared to their human translations. Notably, the GPT-4 model produces translations of free verse poems that are closest to the human versions. Conversely, for sonnets, the GPT-3.5 and Gemini models achieve higher BLEU scores, indicating the least variability in lexical choice relative to their human translations, a result also supported by the vocabulary overlap findings.

In terms of lexical variety, despite differences in word choices, the translations produced by the models is consistently narrower than the human translations. This pattern holds true regardless of the source language, the poetic form, the models provider (OpenAi or Google), or the version number of the LLM, as can be seen in Table 4.

The same pattern is also observable in terms of lexical density. Again, in this respect the human trans-

|  | **chatGPT-3.5 vs HT** | **chatGPT-4 vs HT** | **Gemini vs HT** |
|---|---|---|---|
| *José* (Free verse) | 38 (17%) | **31 (14%)** | 38 (17%) |
| *Fidelidade* (Sonnet) | 31 (26%) | 32 (27%) | **27 (23%)** |
| *Amor* (Sonnet) | 48 (41%) | 48 (41%) | 49 (42%) |
| *Corazón* (Free verse) | 14 (8%) | 11 (6%) | 11 (6%) |

Table 2: Vocabulary Overlap

| Poems | GPT-3.5 | GPT-4 | Gemini |
|---|---|---|---|
| José (free verse) | 0.2637 | **0.3461** | 0.33 |
| Fidelidade (sonnet) | 0.21 | 0.2768 | **0.2886** |
| Amor (sonnet) | **0.0198** | 0.0096 | 0.0197 |
| Corazón (free verse) | 0.3597 | **0.4092** | 0.3064 |

Table 3: Bleu scores

| **Poem** | HT | chatGPT-3.5 | chatGPT-4 | Google Gemini |
|---|---|---|---|---|
| *José* (Free verse) | 0.5 | 0.45 | 0.47 | 0.45 |
| *Fidelidade* (Sonnet) | 0.68 | 0.66 | 0.62 | 0.6 |
| *Amor* (Sonnet) | 0.75 | 0.7 | 0.64 | 0.68 |
| *Corazón* (Free verse) | 0.4 | 0.42 | 0.4 | 0.4 |

Table 4: Lexical Variety

| **Poem** | HT | chatGPT-3.5 | chatGPT-4 | Google Gemini |
|---|---|---|---|---|
| *José* (Free verse) | 0.51 | 0.43 | 0.47 | 0.47 |
| *Fidelidade* (Sonnet) | 0.43 | 0.4 | 0.4 | 0.4 |
| *Amor* (Sonnet) | 0.4 | 0.53 | 0.52 | 0.54 |
| *Corazón* (Free verse) | 0.54 | 0.38 | 0.4 | 0.45 |

Table 5: Lexical Density

lations consistently score more highly than the versions produced by the LLMs.

In terms of sentence length, the HT tended to produce longer sentences than the LLMs translations as shown in table 6, the noteworthy exception being the free verse poem, *José*, where the HT translations are shorter than those produced by the LLMs. One explanation for the LLMs' propensity to produce longer texts is found in their tendency to include optional pronouns as standard, whereas the human translator generally did not. This tendency constitutes a form of explicitation (Baker, 1993, 1996), or reducing ambiguities in translations, which inevitably contributes to an increase in sentence lengths.

### 2.2.1 Rhyme scheme reproduction

Table 7 shows the results obtained for the two sonnets and the few-shot approach used to design the prompts targeting the poems' rhyme schemes. In all but one case (Prompt 2. *Fidelidade*), the ChatGPT models appear to be more successful than Gemini in reproducing the rhyme scheme consistently. However, it is worth noting that the discrepancy, not only between the models' outputs within individual prompts but also across prompts, varies substantially and not always in intuitive ways. For instance, even though the words used in the translations differ, from the perspective of conveying the rhyme scheme, prompt 2, which simply asked the model to replicate the rhyme scheme of the source text; and prompt 4, which went into detail on the nature of that rhyme scheme, appear to have been precisely as successful for *Amor*, having successfully reproduced the rhyme in every line. However, prompt 2 is less successful in the case of *Fidelidade* for GPT-4 (64% of the source rhyme scheme), and much less successful for GPT-3.5 (29% of the source rhyme scheme). It is worth noting that the rhyme schemes of the two poems differ slightly. This implies that the models' ability to reproduce rhyme may be heavily variable, and possibly dependent on the contents of the poem, as well as the extent to which the rhyme schemes in question are represented in the training data. It is also worth considering the source language of the poems under analysis, and the impact this language may have on the results. In this case, *Amor* is written in Spanish, while *Fidelidade* is written in Portuguese. It is worth noting that both the GPT models were highly successful at reproducing the rhyme in the case of the Spanish text, even with a straightforward, zero-shot prompt. On the other hand, the models' success in reproducing rhyme in the Portuguese text was more varied. For the Portuguese sonnet, the GPT models did seem to benefit from the more complex prompts, improving their success rate by 14% (from 29% to 43%) and 15% (from 64% to 79%), respectively. Gemini appears to be much less successful across the board, and its success scores are so low that it is difficult to draw any meaningful conclusions based on this small dataset.

## 2.3 Discussion and conclusions

It is reasonable to ask whether the rapid emergence of LLMs which are either freely available or available for little cost on the internet for the first time might herald a shift in work practices when it comes to literary translators working with highly form-rich texts, such as poems. The systems clearly have far greater flexibility in terms of output style than the NMT systems which literary translation practitioners, especially those with limited technical expertise, are more likely to encounter. As what might be considered unusual use cases from the perspective of the majority of the work such widely available NMT systems do, addressing textual features such as rhyme or syllable count is seldom part of the systems' functionality. Thus, NMT systems usefulness as tools for human translators working with poetry is limited. In the case of LLMs, however, because prompts can be designed, tailored and used in an iterative fashion, their capacity to be useful in the translation of poetry is comparatively higher.

The experiments conducted here show that when using LLMs to translate both free verse poems and sonnets, the resulting texts differ significantly in terms of lexical variety, lexical density, and average sentence length from their human-translated counterparts. This matches findings from previous studies comparing translated and non-translated texts, as well as human-translated and machine-translated texts. Results show that human-translated texts tend to contain more varied vocabulary than LLM-translated texts and that human translations also tend to contain a higher information load as reflected by higher number of content words, compared to poems translated by LLMs. In addition, the mean sentence length of the human-translated poems is higher than

| **Poem** | HT | chatGPT-3.5 | chatGPT-4 | Google Gemini |
|---|---|---|---|---|
| *José* (Free verse) | 3.7 | 4 | 3.8 | 3.6 |
| *Fidelidade* (Sonnet) | 8.6 | 8.4 | 8.4 | 8.6 |
| *Amor* (Sonnet) | 8.4 | 7.0 | 8.4 | 7.7 |
| *Corazón* (Free verse) | 6.5 | 5.8 | 6 | 6 |

Table 6: Mean Sentence Length

| **Prompt 2. *Amor*:** | | | |
|---|---|---|---|
| | **Rhyme scheme** | **Overlap** | **Total count** |
| Source | ABBA—ABBA—CDC—DCD | | |
| Gemini | ABAB—CDCD—EEE—BFB | 1100—0000—000—000 | 2 (14%) |
| ChatGPT-3.5 | ABBA—ABBA—CD—DCD | 1111—1111—111—111 | 14 (100%) |
| ChatGPT-4 | ABBA—ABBA—CDC —DCD | 1111—1111—111—111 | 14 (100%) |
| **Prompt 2. *Fidelidade*:** | | | |
| Source | ABBA—ABBA—CDE—DEC | | |
| Gemini | ABAB—ABCA—DED—FEE | 1100—1100—000—010 | 5 (36%) |
| ChatGPT-3.5 | ABAB—CCCC—DDD—EEE | 1100—0000—100—010 | 4 (29%) |
| ChatGPT-4 | ABBA—ABBA—CCD—EFE | 1111—1111—100—000 | 9 (64%) |
| **Prompt 3. *Fidelidade*:** | | | |
| Source | ABBA—ABBA—CDE—DEC | | |
| Gemini | ABAC—DEFG—HIH—JKL | 1100—0000—000—000 | 2 (14%) |
| ChatGPT-3.5 | ABBA—CDDC—EFE—GEH | 1111—0000—001—010 | 6 (43%) |
| ChatGPT-4 | ABBA—ABBA—CDE—FGH | 1111—1111—111—/000 | 11 (79%) |
| **Prompt 4. *Amor*:** | | | |
| Source | ABBA—ABBA—CDC—DCD | | |
| Gemini | ABCA—DEFE—GHI—EJA | 1100—0000—000—000 | 2 (14%) |
| ChatGT-3.5 | ABBA—ABBA—CDC—DCD | 1111—1111—111—111 | 14 (100%) |
| ChatGPT-4 | ABBA—ABBA—CDC—DCD | 1111—1111—111—111 | 14 (100%) |

Table 7: Generated Rhyme Schemes

the mean sentence length of the poems translated by LLMs, suggesting syntactical differences between human and LLM renditions. (Baker, 1996; Baroni and Bernardini, 2006; Ilise et al., 2010; Ilisei and Inkpen, 2011; Volansky et al., 2013b; Toral, 2019)

When we analyse these results in the context of translationese literature, which posits that higher lexical diversity and density correlate with higher translation quality (Toury, 1980; Gellerstam, 1986; Baker, 1996; Volansky et al., 2013b), they suggest that human translators are more efficient at capturing stylistic nuances and content-based details in translations, whereas the outputs of LLMs tend to exhibit features of simplification as reflected by less varied vocabulary and lower lexical density. This might be because LLMs tend to translate with fewer departures from source text structures. Despite these differences, it can be inferred that the use of LLMs by poetry translators does not necessarily imply a substantial loss of output quality when it comes to lexical variety, lexical density and average sentence length. On the other hand, a reduction in quality could be expected if LLMs are used as part of a post-editing workflow.

In addition, when it comes to rhyme, the experiments here have demonstrated that LLMs can go some way to reflecting complex patterns of word choice based on word-final phonemes in their outputs, in special, GPT models showed a better performance in this specific task as compared to Gemini model. However, it should be noted that the success rate of such rhymed outputs is variable and not entirely predictable. It is likely closely tied to the amount of appropriate training data for the source and target languages that is available to the model (Hoffmann et al., 2024). In this case, ”appropriate” does not simply mean material which includes rhyme in general. Instead, as a general principle, the more similar the training data is to the form of the desired output, the more likely the system is to be successful (Sahu et al., 2022). Thus, if the target text is a sonnet, with a complex ABBA-ABBA-CDE-DEC rhyme scheme and the training data includes a wealth of examples of sonnets of the same format, it can be hypothesised that the output, in terms of rhyming, would likely be better than if the training data included a larger amount of poetry, but comparatively less with this specific rhyme scheme. Our experiments also show that prompt engineering can improve the models' ability to reproduce rhyme schemes, thus revealing the potential benefits for translators in receiving training in prompt engineering. This skill could be valuable not only for this specific translation task but also for addressing other translation challenges.

It is important to note that LLMs offer a new approach to choosing words, which differs from unassisted human translation, especially for poetic text types with strict forms, such as sonnets. The vocabulary overlap experiment here shows that LLMs often make different word choices from their human counterparts, which could be used by human translators for inspiration. LLMs' functionality also offers opportunities for facilitating the production of rhymed translation candidates, again, not with the goal of replacing the human translators, but for increasing the speed at which possible rhymed translations candidates can be suggested to the otherwise unaided human translator.

Indeed, follow-on research could investigate the creative potential associated with LLMs in terms of outputting a range of translation candidates for any given input and how or whether this affects the human translator's work. Thus, it may be that there is potential in LLMs for human translators working as part of a CALT (Computer-Assisted Literary Translation) workflow, in encouraging and developing creative outputs. In particular, the systems have the capacity to assist and speed the resolution of complex challenges, such as searching for rhyming pairs of specific lengths that encapsulate specific meanings. By contrast, the usefulness of the systems as the primary actors in poetry translation projects, coupled with human post-editing, is likely heavily limited, as seen in the results of the experiments here on lexical variety, lexical density, average sentence length, and especially, rhyme. Based on the findings gained from this small number of experiments, it appears likely that the post-editing work required to bring the systems' outputs to the standard observable in the published human translations would be so substantial and pervasive as to negate the benefits associated with using the models in the first place.

## References

Amatriain, X. (2024). Prompt design and engineering: Introduction and advanced methods.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In Francis, G. and Tognini-Bonelli, E., editors, *Text and Technology: In Honour of John Sinclair*, pages 233–252. John Benjamins Publishing Company, Netherlands.

Baker, M. (1996). chapter corpus-based translation studies: The challenges that lie ahead. In *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*, page 175–186, Amsterdam: John Benjamins Publishing Company.

Baroni, M. and Bernardini, S. (2006). A new approach to the study of translationese: machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Castilho, S. and Resende, N. (2022). Post-editese in literary translations. *Information*, 13(2):66.

Castilho, S., Resende, N., and Mitkov, R. (2019). What influences the features of post-editese? a preliminary study. In *Proceedings of the Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria.

Cruys, T. (2023). Up and about, or betwixt and between? In *Computer-Assisted Literary Translation*, pages 158–172. Routledge, New York.

Daems, J., De Clercq, O., and Macken, L. (2017). Translationese and post-editese: How comparable is comparable quality? *Linguistica Antverpiensia New Series - Themes in Translation Studies*, 16:89–103.

des Associations de Traducteurs Littéraires (CEATL), C. E. (2022). Survey on working conditions 2020. Accessed: 2024-06-10.

Gellerstam, M. (1986). Translationese in swedish novels translated from english. In *In Wollin, L. and Lindquist, H. Translation Studies in Scandinavia*, volume 4, pages 88–95, CWK Gleerup, Lund.

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How good are gpt models at machine translation? a comprehensive evaluation.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2024). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Ilise, I., Inkpen, D., Pastor, G. C., and Mitkov, R. (2010). Identification of translationese: a machine learning approach. In *In Gelbukh, A. F. (ed.), Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing, volume 6008 of Lecture Notes in Computer Science.*, pages 503–511.

Ilisei, I. and Inkpen, D. (2011). Translationese traits in romanian newspapers: a machine learning approach. *International Journal of Computational Linguistics and Applications*, 2(1–2).

Jones, J. (1986). *My First Book this Year*. John Doe.

Karpinska, M. and Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Koponen, M. (2016). Is machine translation post-editing worth the effort?: A survey of research into post-editing and effort.

Kunilovskaya, M. and Lapshinova-Koltunski, E. (2019). Translationese features as indicators of quality in English-Russian human translation. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)*, pages 47–56, Varna, Bulgaria. Incoma Ltd., Shoumen, Bulgaria.

Laviosa, S. (2002). Corpus-based translation studies: Theory, findings, applications. In *Approaches to translation studies*, Amsterdam & New York: Rodopi.

Pastor, G. C., Mitkov, R., and Pekar, V. (2008). V.: Translation universals: Do they exist? a corpus-based nlp study of convergence and simplification. In *In: Proceedings of the AMTA*.

Peng, K., Ding, L., Zhong, Q., Shen, L., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. (2023). Towards making the most of chatgpt for machine translation.

Pynte, J., New, B., and Kennedy, A. (2008). A multiple regression analysis of syntactic and semantic influences in reading normal text. *Journal of Eye Movement Research*, 2(1).

Roos, A. (2023). The experiment. In *Computer-Assisted Literary Translation*, pages 237–257. Routledge, New York.

Sahu, G., Rodriguez, P., Laradji, I. H., Atighehchian, P., Vazquez, D., and Bahdanau, D. (2022). Data augmentation for intent classification with off-the-shelf large language models.

Toral, A. (2019). Post-editese: an exacerbated translationese. In *Proceedings of Machine TRanslation Summit*, Dublin, Ireland.

Toral, A. and Way, A. (2014). Is machine translation ready for literature. In *Proceedings of Translating and the Computer 36*, London, UK. AsLing.

Toury, G. (1980). *In Search of a Theory of Translation*. The Porter Institute for Poetics and Semiotics. Tel Aviv University, Tel Aviv, Israel.

Volansky, V., Ordan, N., and Wintner, S. (2013a). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Volansky, V., Ordan, N., and Wintner, S. (2013b). On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

# A Advanced Prompts

**Prompt 3:**
*The rhyme scheme of the poem Soneto da Fidelidade is ABBA ABBA CDE DEC. Translate this poem into English reproducing the rhyme scheme of the source poem.*
*In this rhyme scheme:*
*example 1) A represents words ending in -ento: atento/pensamento*
*example 2) B represents words ending in -anto: encanto/tanto*
*example 3) C presents words ending -ure: procure/dure*
*example 4) D represents words ending in -ive: tive/vive*
*example 5) E represents words that ends in -ama: chama/ama*

*Soneto da Fidelidade*

*De tudo, ao meu amor serei atento (A)*
*Antes, e com tal zelo, e sempre, e tanto (B)*
*Que mesmo em face do maior encanto (B)*
*Dele se encante mais meu pensamento. (A)*
*Quero vivê-lo em cada vão momento (A)*
*E em louvor hei de espalhar meu canto (B)*
*E rir meu riso e derramar meu pranto (B)*
*Ao seu pesar ou seu contentamento. (A)*
*E assim, quando mais tarde me procure (C)*
*Quem sabe a morte, angústia de quem vive (D)*
*Quem sabe a solidão, fim de quem ama (E)*
*Eu possa me dizer do amor que tive (D)*
*Que não seja imortal, posto que é chama (E)*
*Mas que seja infinito enquanto dure. (C)*

**Prompt 4:**
*The rhyme scheme of the poem*
*Amor constante más allá de la muerte is*
*ABBA ABBA CDC DCD. Translate this poem into*
*English reproducing the rhyme scheme*
*of the source poem.*
*In this rhyme scheme:*
*example 1) A represents words ending in -era: postrera/lisonjera*
*example 2) B represents words ending in -ia: dia/mia*
*example 3) C presents words ending -ido: sido/ardido*
*example 4) D represents words ending in -ado:*

*dado/enamorado*

*Amor constante más allá de la muerte*

*Cerrar podrá mis ojos la postrera (A)*
*Sombra que me llevare el blanco día, (B)*
*Y podrá desatar esta alma mía, (B)*
*Hora a su afán ansioso lisonjera; (A)*
*Mas no de esotra parte en la ribera (A)*
*Dejará la memoria, en donde ardía: (B*
*Nadar sabra mi llama el agua fría, (B*
*Y perder el respeto a ley severa. (A)*
*Alma, a quien todo un dios prisión ha sido, (C)*
*Venas, que humor a tanto fuego han dado,(D)*
*Médulas, que han gloriosamente ardido, (C)*
*Su cuerpo dejará, no su cuidado; (D)*
*Serán ceniza, mas tendrá sentido; (C)*
*Polvo serán, mas polvo enamorado. (D)*