Chapter 2: How good is MT at translating Latin religious texts?

Contents

- Introduction
- Previous Work
- Methods & Data
- Experiments
- Results & Discussion
- Conclusion
- Bibliography
- Appendix
- Deleted:

Stefano Staffa, Andrea Scheck

Introduction

In 1382, when the Latin Bible was first translated English by John Wycliffe, this process required enormous human effort, extensive knowledge of both Latin and the vernacular, and years of labor by many religious scholars. The result of this translation - a religious text which could be understood by the common population - had groundbreaking and far-reaching impacts on culture and religion. One can hardly imagine how history might have changed had the Bible never been translated into English—or translated less carefully.

Almost 650 years later, Machine Translation (MT) has reduced the effort required for translation processes from years to minutes. Even low-resource languages, like Latin, are increasingly translated with the assistance of these tools, with studies suggesting that some

MT systems can achieve a reasonably good translation quality [Volk *et al.*, 2024]. However, while MT systems perform efficiently on many genres, they still face challenges when dealing with more creative works [Cespedosa Vázquez and Mitkov, 2023], of which the Bible with its poems and psalms contains many.

Given that there are many Latin religious texts which remain untranslated to this day, this chapter raise the question: Does MT serve as a fitting tool for translating Latin religious text to English? To explore this, we examine the performance of four advanced MT systems (GPT-4o, Gemini, Google Translate, and Yandex) when handling Latin religious texts compared to more neutral, descriptive Latin passages. By comparing the results to each other and to the gold standard human translation, we aim to shed light on the efficacy and limitations of MT in translating Latin, identifying which tools are better suited for specific genres of texts. By understanding these distinctions, we hope to contribute to future advancements in MT for low-resource languages and support the translation of the vast untranslated Latin texts which could provide valuable insights into the historical and intellectual evolution of the Western world.

Previous Work

Work on MT and language modelling for Latin has progressed significantly since the advent of MT, leveraging both traditional and modern approaches. Martínez Garcia and García Tejedor Martínez Garcia and García Tejedor [2020] utilized the Bible as a parallel corpus to build a Transformer-based Neural Machine Translation (NMT) system for Latin-Spanish, addressing challenges associated with Latin's complex morphology and low-resource nature. Similarly, Christodouloupoulos and Steedman Christodoulopoulos and Steedman [2014] demonstrated the utility of the Bible as a parallel corpus in over 100 languages, including Latin, highlighting its structure and consistency as beneficial for multilingual NLP tasks. Liu et al. Liu *et al.* [2021] further confirmed the effectiveness of using biblical texts for improving MT in low-resource settings, underscoring their relevance for Latin translation tasks.

Beyond religious texts, Fischer et al. Fischer *et al.* [2022] explored translating 16th-century Latin letters into German, emphasizing the importance of tailored training data for domain-specific translation. Bistafa Bistafa [2024] examined the challenges of translating Latin scientific texts, particularly the works of mathematician Leonhard Euler, using artificial intelligence and revealing complexities in specialized vocabulary and syntax. These studies highlight the diversity of Latin translation tasks and the potential of MT systems in addressing them.

Recent advancements in Large Language Models (LLMs) have further expanded the possibilities for Latin MT. Volk et al. Volk *et al.* [2024] evaluated GPT-4o's performance on both translation and summarization of Latin texts, achieving superior results compared to

traditional MT systems and showcasing LLMs' capabilities in handling historical and low-resource languages. Riemenschneider and Frank Riemenschneider and Frank [2023] trained multilingual LLMs on Latin corpora, achieving state-of-the-art results for part-of-speech tagging and lemmatization, evaluated against the EvaLatin 2022 dataset. These works collectively provide a strong foundation for investigating LLM-based translation for Latin texts across religious and non-religious domains, as undertaken in this project.

Methods & Data

We constructed a corpus of Latin texts spanning both religious and non-religious genres to evaluate translation performance across diverse stylistic and thematic categories. The dataset contains 1'398 unique Latin words and 1'685 unique English words. Of these, religious texts contribute 566 unique Latin words and 937 unique English words, while non-religious texts comprise 832 unique Latin words and 748 unique English words. The dataset includes approximately 180 sentences drawn from texts written between the 1st century BCE and the 12th century CE.

Data

Religious texts were sourced from the Biblia Sacra iuxta Vulgatam Clementinam and comprise 60 passages distributed across 4 songs, 3 poetic passages, and 4 neutral passages. Each Latin text is paired with three English translations from the English Standard Version (ESV, *ESV* [2001]), the Douay-Rheims Bible (DRB, at Rheims and Douay [1899]), and the King James Version (KJV, [*KJV*, 1611]). The inclusion of three Bible versions captures theological, cultural, and stylistic differences, providing a nuanced basis for comparison. For instance, the DRB was translated directly from the Latin Vulgate and, as a result, adheres more closely to Latin phrasing. The KJV and the ESV drew from Hebrew, Greek and Latin and adopt a modernized style.

To allow for an optimal comparison of MT performance, we aimed to chose neutral, descriptive Latin passages as sources for the non-religious excerpts. As many well-established Latin works with English translations tend to be religious, philosophical, or fictional, identifying a truly neutral text was a challenge. We selected Cicero's De Legibus, a key legal text, and Geoffrey of Monmouth's Historia Regum Britanniae, which includes descriptive historical narratives.

This combination of literal and non-literal translations was expected to highlight interpretative variations for MT systems. Poetic passages, in particular, require systems to balance semantic accuracy with stylistic complexity, while neutral texts test straightforward syntactical translations.

Metrics

Translation quality was assessed using four widely recognized metrics: BLEU, ROUGE-L, METEOR, and chrF. Each metric captures a distinct dimension of translation quality, providing a comprehensive evaluation framework. All scores were calculated as percentages from 0 to 100. Thresholds were set to classify scores, with scores below 30 seen as faulty translations and scores exceeding 60 seen as high-quality translations.

Bilingual Evaluation Understudy (BLEU)

BLEU measures the overlap of n-grams (sequences of 1 to 4 words) between the MT and one or more reference translations [Papineni *et al.*, 2002]. It calculates precision for these n-grams and includes a brevity penalty to discourage overly short translations. The metric is particularly suited for evaluating literal translations where exact word matches are critical. However, BLEU is less effective for assessing translations with valid paraphrasing or synonym use, as it does not account for semantic similarity or contextual nuance. A BLEU score of 75 implies high fidelity, while scores below 30 suggest significant deviations from the reference.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE-l)

ROUGE-L evaluates the longest common subsequence (LCS) between the machine-generated translation and the reference text [Lin and Och, 2004]. This emphasizes structural similarity, focusing on recall—the proportion of the reference that appears in the generated text. Unlike BLEU, ROUGE-L is not limited to contiguous n-grams, making it useful for evaluating texts with more flexible word order, such as Latin prose and poetry. It excels in identifying translations that preserve the overall structure and flow of the reference, even if individual word choices differ.

Metric for Evaluation of Translation with Explicit ORdering (METEOR)

METEOR extends beyond BLEU and ROUGE-L by incorporating precision, recall, and additional linguistic features such as stemming (reducing words to their root forms) and synonym matching [Banerjee and Lavie, 2005]. It aligns words semantically rather than strictly lexically, enabling a better assessment of idiomatic expressions, paraphrasing, and non-literal translations. METEOR also assigns higher weights to exact matches while still rewarding partial matches and correct word order. This makes it particularly effective for evaluating poetic and figurative texts where semantic equivalence outweighs literal fidelity. Higher scores reflect greater similarity to the reference, factoring in synonyms and rephrased segments.

Character n-gram F-score (chrF)

chrF operates at the character level, comparing sequences of character n-grams between the machine translation and the reference text [Popović, 2015]. This fine-grained approach is especially advantageous for highly inflectional languages like Latin, where slight morphological differences (e.g., verb conjugations or noun declensions) can significantly alter meaning. By focusing on characters rather than words, chrF provides sensitivity to subtle grammatical nuances that may not be captured by word-based metrics. It also avoids penalizing legitimate variations in word segmentation or inflection. For Latin translations, chrF is particularly valuable for detecting morphological accuracy and alignment with the reference text.

Tools

Machine translations were generated using GPT-4o, Google Translate, Gemini, and Yandex Translate. These systems leverage either pre-trained language models or statistical algorithms to translate Latin texts into English. Each system has strengths suited to specific text types: GPT-4o excels at contextual and semantic nuances, while Google Translate often delivers consistent outputs for literal translations.

Automated scoring scripts, written in Python, were used to evaluate the MT against gold-standard references. By combining diverse data sources, detailed metrics, and advanced translation systems, this methodology provides a comprehensive framework for evaluating machine translation performance on Latin texts of varying complexity and stylistic nuance.

Experiments

The experimental setup began with selecting original Latin excerpts from the aforementioned sources and their corresponding gold-standard translations. The excerpts were translated from Latin to English individually using the web interfaces of Google Translate and Yandex. For GPT-4o and Gemini, translations were conducted in separate conversations, preceded by a standardized prompt to limit the influence of prior knowledge or external context on the outputs[1]. Each translation was then scored against the gold standard using the four metrics (BLEU, ROUGE, METEOR, and chrF), resulting in a matrix with four translations per excerpt and four scores per translation.

Table 1 allows for a look into the translation results: (*here will follow an interactive code block which the reader can run to see a random example of one Latin excerpt and its translation by all MT systems*)

```
import pandas as pd
import random
from IPython.display import display, Markdown

csv_path = "data/translations.csv"  # CSV with all translations
data = pd.read_csv(csv_path, delimiter=';')

random_row = data.sample(n=1).iloc[0]  # Select a random row

# Show in a table
table = "| Header | Content |\n"
table += "|--------|---------|\n"
for column, value in random_row.items():
    # Only show the first 400 characters
    truncated_value = str(value)[:400] + "..." if len(str(value)) > 400 else str(va
    table += f"| {column} | {truncated_value} |\n"
display(Markdown(table))
```

Low BLEU scores

Examining the scoring results, we observed low BLEU scores across texts, with an overall BLEU average of 24.86 %. Of the 49 translations, 35 received a BLEU score below the threshold of 30 %, indicating notable errors in lexical or syntactic accuracy. Since the scores of all metrics were averaged, these low BLEU scores negatively influenced the overall results of almost every translation. To address this, we also considered the median of the metrics alongside the average, mitigating the impact of outliers caused by low BLEU scores.

While seemingly low, these scores align with prior research: Volk et al. (2024) observed a BLEU score of 25.22 % for Google Translate and 34.50 % for GPT-4o. Our findings, with BLEU averages of 25.32 % for Google Translate and 56.69 % for GPT-4o, indicate no major errors during the experiments but consistency in the outputs, while also highlighting a notable improvement in GPT-4o's performance compared to earlier projects.

Table 2 allows for a look into the scores for a random translation: (*here will follow an interactive code block which the reader can run to see a random example of one Latin excerpt and its scores in all metrics*)

```
import pandas as pd
import random
from IPython.display import display, Markdown

csv_path_scores = "data/scores.csv"  # CSV with all scores
data_scores = pd.read_csv(csv_path_scores, delimiter=';')

random_row_scores = data_scores.sample(n=1).iloc[0]  # Select a random row

# Show in a table
table_scores = "| Source | BLEU average | ROUGE average | chrF average | METEOR ave
table_scores += "|--------|--------------|---------------|--------------|----------
for column, value in random_row_scores.items():
    # Only show the first 400 characters for the Source (although it shouldn't excee
    truncated_value = str(value)[:400] + "..." if len(str(value)) > 400 else str(va
    table_scores += f"| {truncated_value} " if column == 'Source' else f"| {truncate
table_scores += "|\n"
display(Markdown(table_scores))
```

Error proofing

Six translations received an overall average score below 30 %, indicating significant errors in the translation. These included Psalm 88:3-7 (DRB), Psalm 23:4-6 (DRB), Book 1 Chapter 13 of The History of the Kings of Britain, Job 3:11-13 (in both the ESV and KJV), and Book 1 Section 40 of De Legibus. Upon review, we identified issues in three of the corresponding gold standard translations, where they had been either incorrectly or incompletely processed. After addressing these discrepancies, the average score for the affected excerpts increased to slightly above 30 %, marking them as acceptable translations.

Retranslations

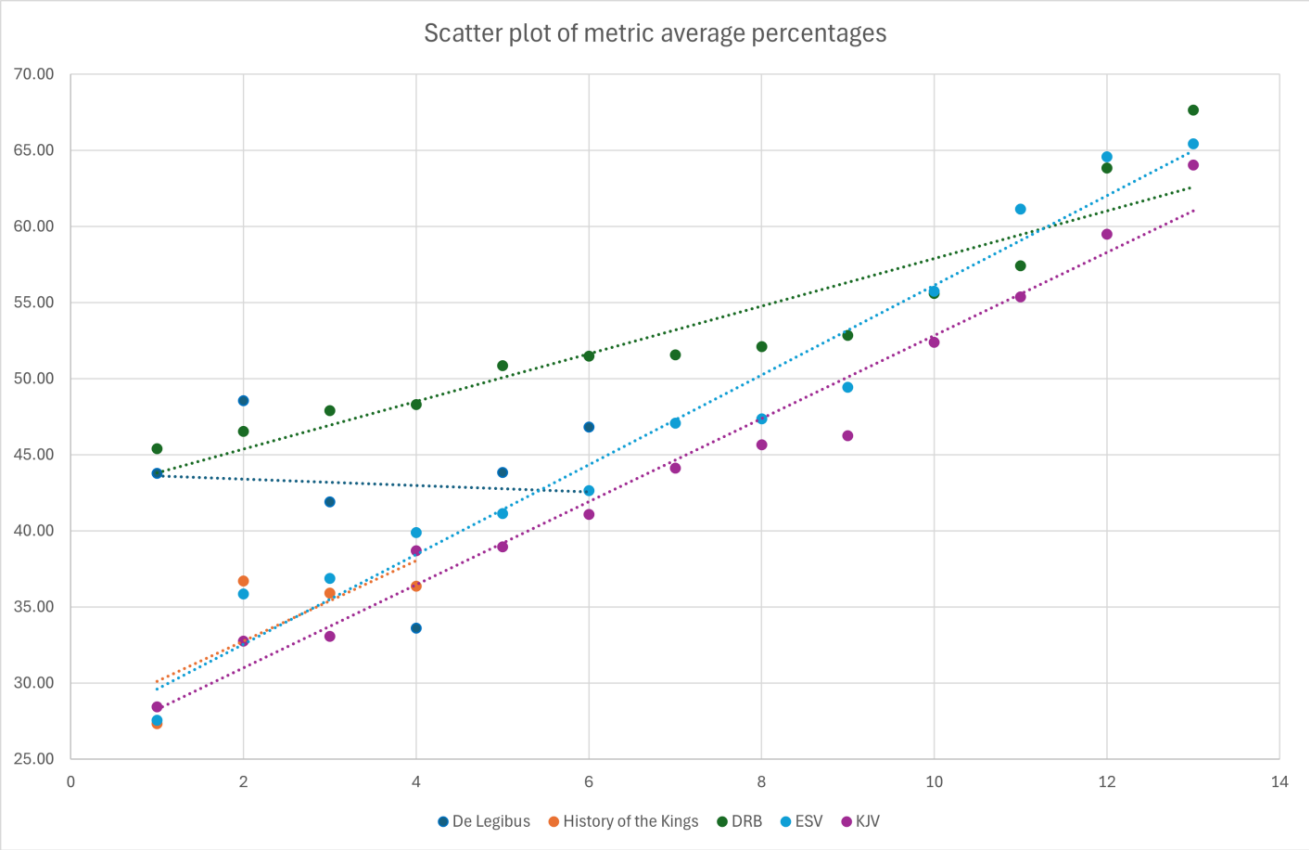
After these corrections, three texts remained with an average score below 30 %: Book 1 Chapter 13 of the History of the Kings of Britain and Job 3: 11-13 in both ESV and KJV. To address this, we adjusted the translation workflow by translating each sentence individually rather than entire paragraphs. This approach aimed to mitigate the challenges posed by long, syntactically complex structures, which are particularly difficult for MT systems to handle (He, 2023). While this method improved the score for The History of the Kings of Britain by 1.8 %, it did not result in significant changes for Job 3:11-13. Ultimately, the average scores for all three excerpts remained below 30 %. Possible reasons for these results are explored in detail in the following chapter.

Results & Discussion

Comparing scores within sources

Considering the average and median values of all chosen metrics, we compared the translation quality between the different genres of text. The highest and lowest scores per text source were examined further to be sure there were no processing errors.

Figure 1 shows a scatter plot of the average scores for the text excerpts of each source, arranged in ascending order of values:



De Legibus

All translations of De Legibus lie above the threshold of 30 %, with five of six excerpts scoring above 41 % when considering the average, and above 45 % when considering the median. The average and median scores for De Legibus generally fall within a narrow range, indicating consistent performance across different sections. The translation for Book 1 Section 60 received the highest scores (average of 46.82 %, median of 48.73 %), possibly due to its closer adherence to the original text in both structure and meaning. However, there is a notable outlier in Book 1 Section 40, which achieved an average of only 33.61 %.

Examining **Book 1 Section 40** in detail, it becomes apparent that the gold standard

translation diverges significantly from the Latin source, using fewer details and replacing long descriptive clauses with brief summaries. This abbreviation of content could cause fewer overlapping n-grams when compared to translations that more closely adhered to the original text. Additionally, the gold standard uses a more modernized tone, losing certain phrases such as the justification of crimes through “naturae iure” and softening the vivid imagery of torture. This semantic difference could reduce matches in metrics like METEOR and chrF.

History of the Kings of Britain

The translations for History of the Kings of Britain show a wider range of scores, with three of four translations achieving average scores above 30 %, but none exceeding 40 %. The average scores range from 27.33 % to 36.69 % with median scores falling between 30.30 % and 40.59 %. Notably, Book 1, Chapter 13 received the lowest average score and fell below the threshold of 30 %.

Examining **Book 1, Chapter 13** in detail, it is notable that the Latin source uses a vivid, descriptive language, explicitely depicting the violence of Corineus’s actions. For example, the original describes him severing arms, decapitating enemies, and engaging in physical combat. The gold standard translation simplifies this text by consolidating actions and reducing the intensity. Instead, the gold standard translation uses phrases like “terrible slaughter” and omits the specific actions, which could lead to lower BLEU scores.

As this text fell below the threshold of 30 %, it was one of the retranslations mentioned in the preceding chapter. This marginally improved the scores, with the average rising from 27.33 to 29.13, a percentage-wise improvement of approximately 6.57 %. Even the improved score from sentence-to-sentence-translation was not adequate enough to reach the 30 % threshold.

Religious texts

For the Bible translations using the KJV as the gold standard, 12 of 13 translations achieved an average score greater than 32.75 %, with the highest score reaching 64.05 %. Among these high-performing texts were neutral descriptions of law and genealogy, stories as well as songs and poems. The lowest score, recorded for Job 3:11-13, was 28.44 %, which deviated significantly from the other translations.

For the Bible translations using the ESV as the gold standard, 12 out of 13 texts received an average score above 35.85 %, with the highest average score being 65.43 %. Most of the translations scored against the ESV gold standard showed consistently high scores across the BLEU, ROUGE, chrF, and METEOR metrics. The average scores for these texts generally

remained well above 35 %, with scores consistently improving for passages like Psalm 23:4-6 (65.43 %) and Exodus 7:20-24 (61.25 %). Notably, three translations scored against the ESV gold standard achieve an average score above 60 %, making it the translation project with the largest number of "good" results. Within the translations, there was no clear trend of higher scores for neutral texts as opposed to songs or poems, however, the lowest score was recorded for the poem Job 3:11-13, at 27.54 %.

For the Bible translations using the DRB as the gold standard, all texts received a score average above 45.41 %, with the highest average reaching 67.65 %, and the corresponding median at 73.06 %. Remarkably, even Job 3:11-13, which presented challenges and low scores for the other Bible versions, achieved a significantly better score of 46.52 % in the DRB translation. The DRB translations consistently achieved higher scores across all the MT systems and less variation in performance across the passages compared to both the KJV and ESV. Notably, this was the only translation project in which no translation had an average score under 30 %, suggesting that the language style and structure of the DRB translation may be more amenable to machine translation systems, particularly in the semantic equivalence captured by the evaluation metrics (e.g., ROUGE, METEOR). The lowest score was recorded for Psalm 88:4-8, but at 45.41 %, it still performed significantly better than the same passage in the KJV and ESV translations.

Job 3:11-13 in detail

The Latin text of Job 3:11-13 contains highly poetic language that reflects emotional distress, characterized by the repetition of rhetorical "Why"-questions. This type of structure can pose challenges for any translators, as they may struggle to preserve the nuanced tone and rhetorical structure. This is amplified by the fact that the Latin word order is flexible and emphasizes certain elements for rhetorical effect, which is not always directly translatable into English. These challenges become evident when comparing the gold standard translation from KJV, ESV and DRB:

While both KJV and ESV are faithful to the original meaning and easily readable for modern eyes, they use a modernized language structure that simplifies the emotional weight. For example, the poetic expression like "dormiens silerem" becomes the less expressive "lain still/down and been quiet". The Latin differentiation between "in vulva" (approximately: "in the womb") and "ex utero" ("out of the womb") might have been a additional challenge for the translators, being translated to "at birth" or "from the womb". Additionally, the expression "give up the ghost" in the KJV translation is a stylistically unique choice.

In the DRB translation, the translation for the same excerpt received an acceptable average score, as the language and repetition from the Latin are preserved well in the DRB, allowing for more accurate semantic mappings. Notably, however, this translation does not read as a well-formed text to a modern reader, with some questions entirely lacking the subject "I".

While this might be close to the Latin original and also close to the MT results, it is hardly how we would form a sentence today.

Verse	Latin	ESV	KJV	DRB
Job 3: 11-13	Quare non in vulva mortuus sum? egressus ex utero non statim perii? Quare exceptus genibus? cur lactatus uberibus? Nunc enim dormiens silerem, et somno meo requiescerem	Why did I not die at birth, come out from the womb and expire? Why did the knees receive me? Or why the breasts, that I should nurse? For then I would have lain down and been quiet; I would have slept; then had I been at rest,	Why died I not from the womb? Why did I not give up the ghost when I came out of the belly? Why did the knees prevent me? Or why the breasts that I should suck? For now should I have lain still and been quiet, I should have slept: then had I been at rest,	Why did I not die in the womb, why did I not perish when I came out of the belly? Why received upon the knees? Why suckled at the breasts? For now I should have been asleep and still, and should have rest in my sleep.

The same challenges which led to deviation between the different gold standards also seem to have affected the MT systems. The repetition of questions seems to have been highly difficult for MT, with systems appearing to interpret the words literally. Google Translate, Gemini and Yandex abandon the “Why” structure in the second sentence, asking “if” the speaker died; in the following sentences, they also lose the reference to the subject (“I”). Yandex was unable to reproduce the emotional tone, producing with very practical sounding sentences. Additionally, vocabulary ambiguities also occur due to the very limited context, with “perii” (perished) translated as “ruined” and “exceptus” (received) as “caught”.

Latin	GPT-4o	Google Translate	Gemini	Yandex
Quare non in vulva mortuus sum?	Why did I not die in the womb?	Why did I not die in the womb?	Why not in the womb did I die?	Why am I not dead in the womb?
Egressus ex utero non statim perii?	Having left the uterus, why did I not perish immediately?	Did I not immediately perish when I came out of the womb?	Having gone forth from the womb not immediately did I perish?	Going out of the womb was not immediately ruined?
Quare exceptus genibus?	Why was I received upon the knees?	Why except the knees?	Why caught by the knees?	Why knees?
Cur lactatus uberibus?	Why was I nursed at the breasts?	Why did he breastfeed?	Why suckled by the breasts?	Why breastfeed?
Nunc enim dormiens silerem,	For now, sleeping, I would be silent,	For now, sleeping in silence,	Now indeed sleeping I would be silent,	Silent sleeping for now,
Et somno meo requiescerem.	And in my sleep, I would rest.	And I would rest in my sleep.	And in my sleep I would rest.	I need my sleep.

A retranslation experiment, in which the text was translated sentence by sentence, showed an improvement in overall scores when compared to ESV and KJV, particularly in BLEU. This suggests that translating smaller units may help maintain structure and preserve n-gram matches, even in poetic or emotionally dense texts.

Comparing scores across sources

Overall, there seems to be no clear correlation between the scores and the "neutrality" of the text, neither within the religious texts nor across sources. Non-religious texts, on average, did not perform better than religious texts: While De Legibus had average scores mostly between 33.61 % - 48.54 % and History of the Kings of Britain scores ranged between 27.33 % - 36.69 %, religious texts showed higher overall scores and consistency across versions. Even the lowest-performing religious texts (Job 3: 11-13) had scores comparable to or better than the poorest-performing non-religious texts.

When directly comparing neutral tone excerpts to poems and songs, there is no consistent difference in score averages. The best-performing descriptive text translations, such as Exodus 7: 20-24 from the DRB, did score higher than most poetry. Yet, other descriptive texts like History of the Kings of Britain had lower scores, often around 30 - 40 %, being outperformed by poetry like Ecclesiastes 3: 7-8 at 63.84 % - 64.56 %.

It is notable that translations scored against the DRB (e.g., Exodus 7: 20-24, Psalm 23, Ecclesiastes 3: 7-8) performed particularly well, often exceeding 50 % averages.

Source	Average of all translation scores for the source(%)	Median of all translation scores for the source (%)	General Performance Ranking
Douay-Rheims Bible (DRB)	**52.35**	**55.85**	Highest-performing source
English Standard Version (ESV)	47.96	48.88	Strong, consistent
King James Version (KJV)	45.06	46.93	Good, slightly behind ESV
De Legibus	43.42	47.16	Moderate, varying
History of the Kings	**33.08**	**37.11**	Lowest-performing source

A possible explanation for this is that the DRB, being the oldest Bible translation we included, was translated directly from Latin with strict adherence to the source text. Later translations like ESV and KJV included other sources (Hebrew and Greek) and adapted to the contemporary language of their time, which decreases literal similarity with the source. MT is therefore more likely to score better with the DRB as a target text, even if the resulting language may seem archaic.

While the Bible has a formal, well-established vocabulary and highly consistent structure, De Legibus and History of the Kings both feature complex sentence structures, with subordinate clauses, jargon and references. This is a known challenge for MT systems. Additionally, religious Latin texts may have been more represented in the training data, as they are among the best known and most translated Latin sources.

Comparing scores across tools

GPT-4o

GPT-4o emerges as the most versatile and reliable performer across all evaluated metrics. For biblical texts, it achieves the best results with a BLEU score of 28.74 % and a ROUGE score of 58.27 %, reflecting high accuracy in lexical and structural reproduction. The chrF score of 54.44 % and METEOR score of 57.28 % further underscore its ability to maintain semantic and stylistic alignment. For non-biblical texts, while its performance decreases slightly, with a BLEU score of 19.10 % and a ROUGE score of 47.94 %, GPT-4o remains robust and consistent, managing stylistic diversity better than other tools. Its particularly strong results for biblical texts suggest an aptitude for formal, repetitive, and structured content, though some challenges arise in handling more stylistically varied material.

Google Translate

Google Translate delivers solid translation results but consistently ranks just below GPT-4o across all metrics. Its performance with biblical texts, reflected in a BLEU score of 26.98 %, ROUGE score of 57.43 %, and chrF score of 53.35 %, highlights its fluency and adequacy in reproducing structured content. A METEOR score of 53.36 % indicates reasonably strong semantic alignment.

However, its capabilities diminish more noticeably with non-biblical texts, where it scores a BLEU of 18.37 % and a ROUGE of 46.08 %. The chrF score of 49.35 % and METEOR score of 47.02 % indicate greater difficulty in managing nuanced meanings and stylistic variability. This performance disparity suggests that Google Translate is particularly effective for formal, structured text but struggles with the complexities of informal or stylistically diverse material.

Gemini

Gemini delivers balanced results and strong competitiveness, particularly when compared to Google Translate. For biblical texts, it performs well, achieving a BLEU score of 27.22 %, a ROUGE score of 56.54 %, and a chrF score of 52.71 %, indicating high fluency and coherence. Its METEOR score of 54.71 % further reflects strong semantic alignment, making it a reliable option for structured content like Bible texts.

When translating non-biblical texts, Gemini faces slightly greater challenges but still performs admirably. Its BLEU score of 19.39 % and ROUGE score of 45.84 % indicate it handles stylistically diverse content on par with or slightly better than Google Translate in certain aspects. However, its chrF score of 48.37 % and METEOR score of 43.64 % highlight areas for improvement in capturing lexical richness and nuanced meaning.

Yandex

Yandex consistently underperforms relative to other tools, showing significant limitations in both precision and coherence. For biblical texts, it achieves a BLEU score of 18.68 % and a chrF score of 46.05 %, suggesting weaker lexical and structural fidelity. Its METEOR score of 47.52 % reflects moderate semantic alignment, though it struggles to preserve deeper meanings effectively.

The tool's challenges become more pronounced with non-biblical texts, where it scores only 10.65 % in BLEU and 45.19 % in chrF, with a METEOR score of 43.13 %. These results highlight its difficulty in adapting to diverse stylistic and syntactic demands, making it less suitable for high-quality translations across various content types.

Comparing scores across metrics

When comparing the metrics BLEU, ROUGE, chrF, and METEOR across the translations, distinct patterns emerge. For Bible texts, BLEU scores show the widest variability, ranging from 18.68 % (Yandex) to 28.74 % (GPT-4o). METEOR follows closely with a range from 47.52 % (Yandex) to 57.28 % (GPT-4o). ROUGE scores are slightly narrower, ranging from 51.92 % (Yandex) to 58.27 % (GPT-4o), while chrF ranges from 46.05 % (Yandex) to 54.44 (GPT-4o). This indicates that while BLEU and METEOR are more sensitive to performance differences, ROUGE and chrF provide more stable evaluations.

For non-Bible texts, BLEU again shows the largest variability, with scores spanning from 10.65 % (Yandex) to 19.39 % (Gemini). ROUGE follows, ranging from 40.74 % (Yandex) to 47.94 % (GPT-4o). chrF and METEOR, while showing less declines, still demonstrate meaningful gaps, with chrF ranging from 45.19 % (Yandex) to 51.31 % (GPT-4o) and METEOR from 43.13 % (Yandex) to 48.80 % (GPT-4o). Thus, while chrF and METEOR are slightly more conservative

than BLEU and ROUGE, they still reflect noticeable drops in performance, particularly for more challenging text types.

Across all tools, the metrics reveal a consistent trend: translations of texts with greater stylistic variability score lower. BLEU and ROUGE, which emphasize precision and recall for specific lexical and syntactic features, show sharper declines for such texts, highlighting the difficulty in preserving structural elements. Meanwhile, chrF, which balances character-level precision and recall and METEOR, which incorporates synonym matching and semantic alignment, show relatively smaller variations between text types. This suggests that even when lexical and stylistic accuracy falters, tools maintain a fair degree of meaning preservation.

Conclusion

This chapter highlights the varied performance of MT systems in translating Latin texts, revealing distinct strengths and weaknesses across different text types and evaluation metrics. GPT-4o consistently outperformed other systems, particularly with structured and repetitive content like religious texts, achieving the highest scores across BLEU, ROUGE-L, METEOR, and chrF. In contrast, systems like Yandex struggled significantly, especially with stylistically diverse and complex texts.

Religious texts demonstrated higher overall translation quality compared to neutral texts, likely due to the structured and consistent nature of their source material. Notably, the DRB provided the most favorable benchmark, suggesting that older, more literal translations align better with MT capabilities. Here, it is important to note that the best scoring translation does not automatically equate to the most readable translation for modern readers. Additional challenges persisted with poetic and highly emotional passages, as MT systems often failed to replicate nuanced tone and rhetorical complexity.

The findings underscore the importance of text type and evaluation metric selection in MT research. While MT has advanced significantly, translating low-resource languages like Latin still demands refinements, particularly for creative and interpretive content. Future work could explore improved contextual understanding in MT systems and expand training datasets to include more diverse examples, aiding the broader goals of digital humanities and classical studies.

Tools

- ChatGPT, Version 4o, OpenAI: https://chat.openai.com/
  - Translation of text passages
  - Help with creation of text structure
  - Help with interpretation of results
- Google Translate, Google: https://translate.google.com/
  - Translation of text passages.
- Gemini, Google: https://gemini.google.com/app
  - Translation of text passages
- Yandex, Yandex: https://translate.yandex.com/
  - Translation of text passages

Bibliography

{cite:p}``

[1] **missing editor in KJV1611**

[2] **missing editor in ESV2001**

[3] English College at Rheims and Douay. *DRB*. Douay Bible, 1899.

[4] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL: https://aclanthology.org/W05-0909.

[5] Sylvio R. Bistafa. Translating scientific latin texts with artificial intelligence: the works of euler and contemporaries. 2024. URL: https://doi.org/10.48550/arXiv.2307.07520, arXiv:2307.07520, doi:10.48550/arXiv.2307.07520.

[6] Ana Isabel Cespedosa Vázquez and Ruslan Mitkov. Machine translation of literary texts: genres, times and systems. In Raquel Lázaro Gutiérrez, Antonio Pareja, and Ruslan Mitkov, editors, *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, 48–53. Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL: https://aclanthology.org/2023.nlp4tia-1.7.

[7] Christos Christodoulopoulos and Mark Steedman. A massively parallel corpus: the bible in 100 languages. *Language Resources and Evaluation*, 49:1–21, 06 2014. URL:

https://doi.org/10.1007/s10579-014-9287-y, doi:10.1007/s10579-014-9287-y.

[8] Lukas Fischer, Patricia Scheurer, Raphael Schwitter, and Martin Volk. Machine translation of 16th century letters from latin to german. In *Proceedings of the Workshop on Computational Methods for Historical Texts*. 01 2022.

[9] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, 605–612. Barcelona, Spain, 2004. URL: https://doi.org/10.3115/1218955.1219032, doi:10.3115/1218955.1219032.

[10] Ling Liu, Zach Ryan, and Mans Hulden. The usefulness of bibles in low-resource machine translation. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, 1:, 01 2021. URL: https://doi.org/10.33011/computel.v1i.957, doi:10.33011/computel.v1i.957.

[11] Eva Martínez Garcia and Álvaro García Tejedor. Latin-Spanish neural machine translation: from the Bible to saint augustine. In Rachele Sprugnoli and Marco Passarotti, editors, *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, 94–99. Marseille, France, 2020. European Language Resources Association (ELRA). URL: https://aclanthology.org/2020.lt4hala-1.14.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. URL: https://doi.org/10.3115/1073083.1073135, doi:10.3115/1073083.1073135.

[13] Maja Popović. ChrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392–395. Lisbon, Portugal, 2015. Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/W15-3049, doi:10.18653/v1/W15-3049.

[14] Frederick Riemenschneider and Anette Frank. Exploring large language models for classical philology. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15181–15199. Toronto, Canada, jul 2023. Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/2023.acl-long.846, doi:10.18653/v1/2023.acl-long.846.

[15](1,2) Martin Volk, Dominic Philipp Fischer, Lukas Fischer, Patricia Scheurer, and Phillip Ströbel. Llm-based machine translation and summarization for latin. In *LT4HALA*.

2024. URL: https://doi.org/10.5167/uzh-259369.

Appendix

[1] Prompt used for GPT-4o and Gemini during the experiments: “Approach this sentence translation without drawing on any pre-existing knowledge or examples you’ve encountered. Use only the specific sentence structure and vocabulary present, rather than referencing broader linguistic context, cultural knowledge, or past translations of similar phrases.” (followed by Latin excerpt)