Breaking Boundaries: Neural Approaches to Interlinear Translation of Classic Texts

Anonymous ACL submission

Abstract

Machine Translation (MT) is a crucial field in Natural Language Processing, with recent advancements like the transformer architecture revolutionizing the task. While MT typically aims for accurate and natural translations, there are instances, such as educational translations, where maintaining the original syntactic structure and meaning is paramount. Interlinear translation, exemplified by its application to ancient texts like the Iliad and the Bible, emphasizes this fidelity to the source text's structure.

Despite the importance of interlinear translation, research in automating this process remains limited, particularly for ancient texts. Our work aims to address this gap by evaluating state-of-the-art neural machine translation models on the task of interlinear translation from Ancient Greek to Polish and English. We compare the performance of general-purpose multilingual models with dedicated language models and assess the impact of Part-of-Speech (POS) tags as well as data preprocessing strategies on model performance.

Our contributions include constructing a word-level-aligned parallel corpus of interlinear translations of the Greek New Testament. We fine-tune four base models in various conditions, totaling 144 models, the best of which we make publicly available. Last, but not least, we suggest three approaches for encoding morphological information via dedicated embedding layers, which outperform solutions that do not utilize tags by up to 20% (BLEU score) on an interlinear translation task into both of the target languages.

1 Introduction

Machine translation (MT) is a well-established subfield in Natural Language Processing, primarily focused on producing accurate and natural translations. In typical scenarios, MT systems have the flexibility to reorder words or go beyond literal meanings to account for syntactic differences between source and target languages.

However, there are exceptional cases where maintaining the original syntactic structure and meaning of the source language is of greater importance. One such scenario arises in educational translations, where preserving a close word-to-word correspondence facilitates better comprehension of the relationships between texts. Interlinear translation serves this purpose by aiming to retain the syntactic structure of the original text. Figure 1 presents an example of an interlinear translation. This type of translation holds particular significance in the study of ancient texts, such as the Iliad, Odyssey, works of ancient philosophers, and religious scriptures like the Bible.

Despite the significance of interlinear translation for scholars interested in ancient texts, there has been limited research on automating this process. This may be attributed to the pre-existing translations for many influential texts. However, we believe that this issue remains pertinent, particularly for individuals lacking expertise in ancient languages, cultures, and histories, such as those seen in the Bible.

In our research, we aim to achieve the following objectives:

- Evaluate the performance of state-of-the-art MT models in interlinear translation from Ancient Greek to Polish and English,
- Compare the effectiveness of general-purpose multilingual models with dedicated language models trained specifically on ancient languages and the given target language,
- Assess the impact of Part-of-Speech (POS) tags on model performance by comparing different tag sets and various approaches to their integration, namely encoding them within the input text or via a dedicated embedding layer,

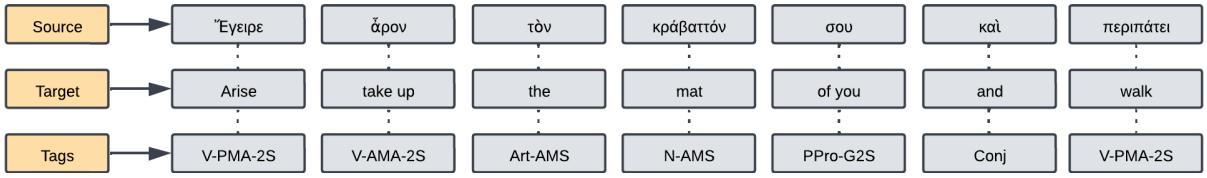


Figure 1: Example of interlinear translation. Snippet from John 5:8 taken from the Bible Hub corpus. The dataset comprises three sequences: words in the source language (Ancient Greek), their respective translations in English, and morphological tags for each source unit.

- Investigate the influence of *pre-processing strategies*, specifically focusing on the common techniques of lower-casing input text and removing diacritics.

Regarding the source corpus, we focus on the study of the full text of the Greek New Testament. We take here into account the fundamental importance of this text for the international society, the fact of Ancient Greek being the original language of the New Testament as well as the existence of numerous translations of it.

For the targets of our translations, we examine differences in performance of the models with respect to languages with a different syntactic character – positional English and inflectional Polish.

Our contributions The contribution of our research is threefold. Firstly, we construct a word-level-aligned parallel corpus of two interlinear translations of the Greek New Testament – to English and to Polish using scraped data from *Bible Hub*[1] and *Oblubienica*.[2][3] Secondly, we conduct fine-tuning experiments for an interlinear translation task using four base models – PhilTa, GreTa (Riemenschneider and Frank, 2023a) and mT5 (Xue et al., 2020) (in two sizes), in 36 setups each, totaling 144 fine-tuned models – the best of which we publish for others to examine and further experiment with[4]. Lastly, our experiments show that including diacritics and morphological tags in interlinear machine translation improves model's performance on the task. We suggest novel ways of encoding the tags within the model's input, improving the baseline score by 20%.

[1]https://biblehub.com/interlinear/

[2]https://biblia.oblubienica.eu/

[3]As we have not received a response regarding the possibility of publishing the datasets, we have chosen not to do so for now. However, we are happy to share them with other researchers upon request.

[4]We are eager to share the models with the reviewers, but have not provided the link to a repository, to simplify the double-blind review process, since the models are heavy.

2 Related Work

Recent years have witnessed a substantial increase in scholarly output focusing on the intersection of machine learning and the study of ancient languages, including Ancient Greek (Sommerschield et al., 2023). Nevertheless, there still remains a significant scope for novel contributions in this area.

In this section, we cover related work in the following fields of study:

- NLP for Ancient Greek,
- Machine Translation for the Bible,
- Interlinear Machine Translation,
- Usage of POS tags in low-resource settings.

The majority of recent research addressing Machine Learning for Ancient Greek focuses on solving problems using encoded forms of the Greek language. Scholars commonly apply encoder models of the BERT family (Devlin et al., 2019) to problems such as Part-of-speech (POS) tagging and lemmatization (Singh et al., 2021a), translation alignment (Yousef et al., 2022; Keersmaekers et al., 2023) and dependency parsing (Nehrdich and Hellwig, 2022) to name a few.

At the same time, encoder-decoder models applied within the machine translation domain enjoy a much smaller popularity. A recent survey on Machine Learning for Ancient Languages lists only a single work covering Ancient Greek in the section devoted to machine translation (Sommerschield et al., 2023). The study, however, does not directly deal with MT – specifically, it addresses the problem of corpus alignment between Ancient Greek and Latin and applies an encoder model to do so.

In our review of the literature, we found only one instance where the state-of-the-art encoder-decoder models were trained specifically for Ancient Greek. Riemenschneider and Frank (2023a) train two T5-family (Raffel et al., 2023) models, *GreTa* and

PhilTa – respectively a monolingual model trained on Ancient Greek and a trilingual one pre-trained on Ancient Greek, Latin and English.

Literature suggests two ways of applying machine learning to the domain of NLP for Ancient Greek – the more popular one takes general-purpose multilingual models such as *BERT* and further trains them via tasks such as Masked Language Modelling (MLM) on an Ancient Greek corpus (e.g. Singh et al. (2021b)). The other approach is to train a dedicated model from scratch, without using an existing pre-trained model as the base. An example of this approach is the previously mentioned work of Riemenschneider and Frank (2023a). While the latter approach allows for training an optimized tokenizer for working with Ancient Greek, it may result in the model's impaired performance in machine translation into the target language.

As reported by Gerner (2018), the number of languages into which the Bible has been translated is growing exponentially. Some, such as Hurskainen (2018) discuss machine translation's usefulness in the Bible translation, but the focal point of their study is translation of the Bible from one modern language to another – oftentimes a low-resource one. Indeed, such case studies were reported for Navajo, Basque and English (Ling et al., 2023) or Mizo and English (Devi et al., 2022).

Some studies use the Bible within NLP for Ancient languages. Martínez Garcia and García Tejedor (2020) use parallel Bible corpora to train a model for translating from Latin to Spanish, Riemenschneider and Frank (2023b) use the Bible data for Greek-English corpora alignment, Krahn et al. (2023) use parallel Ancient Greek-English Bible corpora to evaluate the translation bias of multi-language sentence embedding models, by measuring the distance of embedded Ancient Greek text of the Bible to its embedded English translations. It seems that most scholars use the Bible for its parallel-corpus features and little attention is paid to assessment of how the state-of-the-art MT models would perform on translating the Bible itself from its Ancient Greek manuscripts (in case of the New Testament or the Septuagint) to modern languages especially in settings other than free translation.

Interlinear glossing has been a subject of many extant works. While the glossing may happen on either word- or morpheme- level, the latter one is much more popular as a research area, possibly due to its application in language documentation and preservation. The former is more commonly used as a means of providing readers with a deeper understanding of a given text in the source language even if the reader does not necessarily know it (Carter, 2019).

Some works study the possibility of incorporating source language glosses in generation of free translations in the target language (Zhou et al., 2020), but a reverse scenario, where glosses are part of the algorithm's output enjoys much more attention (Moeller and Hulden, 2018; McMillan-Major, 2020; Zhao et al., 2020). Last year has seen a shared task on interlinear glossing introduced at SIGMORPHON as a part of which participants produced grammatical descriptions of input sentences on morpheme-level.[5]

Part of speech tagging has been present in the history of NLP since its beginning. In the recent years, thanks to better neural architectures and more resources, the need for manual feature engineering has diminished, especially in well-resourced languages. However, discarding morphological metadata might not be such an obvious choice in low-resource settings. Moeller et al. (2021) report that the presence of POS tags does not necessarily impact the performance of Transformer models on selected morphological tasks. At the same time Perera et al. (2022) report that injection of morphological features into their English-to-Sinhala Transformer resulted in a performance boost for one of two tested models. Hence, we aim to evaluate tag-injected model performance.

3 Methodology

In this section we discuss our corpora, including gathering, alignment and preprocessing of the data. Further, we cover models employed and our approaches for encoding the morphological metadata in their inputs. Finally, we describe how the models were fine-tuned.

3.1 Datasets

For our fine-tuning dataset, we prepared two corpora comprising interlinear translations of the Greek New Testament: one into Polish and one into English.

Data Acquisition Both datasets were scraped from distinct sources – Oblubienica and Bible Hub, respectively. The corpora utilize different textual variants of the Greek text. Specifically, the

[5] https://github.com/sigmorphon/2023GlossingST

BH: Ἐγένετο δὲ, ἐν τῷ τὸν Ἀπολλῶ εἶναι ἐν Κορίνθῳ...

OB: εγενετο δε εν τω απολλω ειναι εν κορινθω..

Figure 2: A passage (*Acts 1:19*) showing differences between the source texts in both corpora. The first line originates from Bible Hub (BH) while the the second from Oblubienica (OB). Differences include casing (BH varies casing, OB uses only lowercase), diacritics (used in BH, but not in OB), and an extra article (τον) in Bible Hub's version.

Greek text in the Oblubienica corpus follows Nestle Aland Novum Testamentum Graece 28 (NA28), while Bible Hub merges multiple textual variants, including NA27, Byzantine Majority Text, Scrivener's Textus Receptus 1896, Westcott and Hort, SBLGNT, and Nestle 1904, and marks each variant using special quotes. Although the primary disparity between the two corpora lies in the textual variant used, there are additional distinctions, which include varying casing, usage of diacritics, and punctuation, as depicted in Figure 2. The tag sets of the two corpora differ as well. While in majority of the sentences there is a cross-corpus agreement on a morphological form of a word, there are also places where the two corpora diverge – e.g. in *Rev 5:5* δαυιδ is tagged by Bible Hub as *N-GMS* (Noun – Genitive, Masculine, Singular), but Oblubienica tags the same word as *ni proper* (noun, indeclinable, proper), thus focusing on a different aspect of the word.

Corpus Alignment To evaluate model performance based on the tag set used, we aligned the two corpora at the word level, thus establishing a mapping and exchange between their respective tag sets. Initially, we removed marked textual variants other than NA27 from the Bible Hub version to align it with Oblubienica (NA28). Subsequently, we applied heuristics to match each word from one corpus with its counterpart in the other, prioritizing rules such as exact matching, within-verse matching, and selecting the closest match in cases of multiple candidates. These steps successfully matched over 99% of the words in the corpora. For unmatched words, we mapped morphological tags from one tag set to the other using the statistically most common counterpart. Remaining edge cases, like proper nouns, were mapped manually.

Final Dataset: After alignment, each word in both corpora carries two morphological tags: one original and one cloned from the corresponding word-level counterpart in the other corpus. It's worth noting that the tag sets vary not only in quality but also in quantity. Refer to Table 1 for specific volume details of each corpus.

Corpus	Oblubienica	Bible Hub
Verses	7,940	7,940
Words (GR)	137,390	137,317
Words (PL/EN)	133,581	185,722
Tag Set Size	1,073	684

Table 1: Corpus Statistics. The rows display number of rows, words in the source and target language and the count of unique morphological tags in the tag set.

3.2 Data Preprocessing

There are two schools of thought when it comes to preprocessing texts in Ancient Greek. The first one advocates for keeping all diacritics and training the tools to learn from them. This approach was used by Riemenschneider and Frank (2023a) while training PhilTa and GreTa. The other approach is much more popular and it normalizes the data by stripping the texts from diacritics e.g. Yamshchikov et al. (2022). Within our experiments, we test both of these paradigms. For the *diacritics* version of our dataset, we use the spelling from Bible Hub. We benefit from the fact that the datasets are aligned and replace Greek words in Oblubienica with their Bible Hub counterparts.

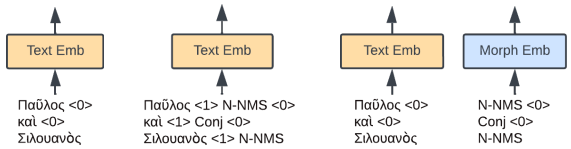
One common measure of a tokenizer's efficiency on a given corpus is the average number of tokens per word (Yamshchikov et al., 2022), which we calculate and report in Table 2. While there is a visible discrepancy in tokenization performance for Ancient Greek with diacritics – here, mT5 requires twice as many tokens to represent a word compared to PhilTa or GreTa – this gap disappears when tokenizing the normalized source. In all other cases (Polish, English, and tags), mT5 outperforms the others. Additionally, it is worth mentioning the much higher token numbers per morphological tag in the Oblubienica dataset compared to its Bible Hub counterpart. This is mainly due to the longer tags used in Oblubienica, e.g. when Bible Hub tags ἀρχῇ as *N-DFS*, Oblubienica tags it as *n_ Dat Sg f*.

3.3 Base Models

Our study employs four base models: GreTa, PhilTa (Riemenschneider and Frank, 2023a) and mT5 in two sizes – base and large (Xue et al., 2020),

Tokenizer Dataset	GreTa	PhilTa	mT5
GR – diacritics	**1.57**	1.58	3.23
GR – normalized	2.50	**2.36**	2.37
PL	4.08	4.20	**2.37**
EN	3.51	**1.92**	1.99
Tags (OB)	7.26	6.94	**5.45**
Tags (BH)	5.06	5.26	**3.82**

Table 2: Overview of tokenization metrics. The consecutive rows display the average number of tokens required by each tokenizer for: a Greek word with diacritics, a normalized Greek word, a Polish word, an English word, a tag from the Oblubienica (OB) tag set, and a tag from the Bible Hub (BH) tag set, respectively.



(a) t-o (b) t-w-t (c) emb-*

Figure 3: Comparison of three input sequence encoding methods. The first method (t-o, baseline) omits the morphological metadata. The second method (t-w-t) includes these tags as part of the input. Lastly, the third method (emb-*) utilizes a dedicated embedding layer to encode the tags as a separate sequence.

all belonging to the T5 model family (Chung et al., 2022). Both GreTa and PhilTa are T5-base-sized models, with GreTa trained on Ancient Greek corpora and PhilTa trained on Ancient Greek, Latin, and English. mT5 was trained on the mC4 corpus, which comprises 101 languages including English and Polish – the target languages for our translations. Ancient Greek was not reported to be part of the pre-training data for the model. We select mT5-base to match the size of the other models and mT5-large to explore whether increasing the number of parameters improves performance.

3.4 Model Inputs

In our experiments we assess whether inclusion of morphological tags leads to an improved performance on interlinear translation task. To do so, we implement five scenarios which can be grouped into three categories (as seen in Figure 3) based on how the tags are encoded. We discuss them in this section.

In *Text Only (t-o)* – the baseline scenario – no morphological information is passed to the model. Each Greek word is separated with a dedicated sentinel token, and this sequence of words and separators constitutes the model's input.

In *Text With Morphological Tags (t-w-t)* we encode POS tags as part of the model's text input. Greek words and tags are encoded with the help of two sentinel tokens: one to separate word-tag pairs and another to demarcate the end of the Greek word and the beginning of the tag within each pair.

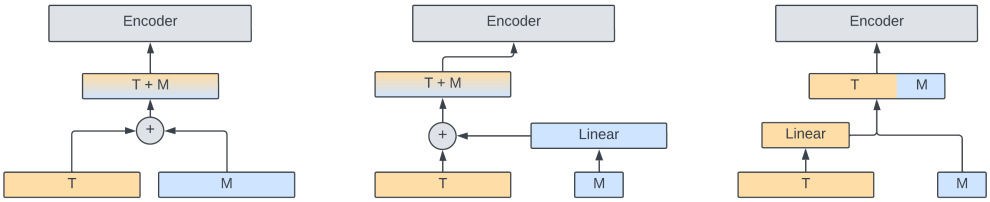
The third group (*emb-*) comprises the remaining three scenarios, which involve introducing a dedicated embedding layer trained during the model fine-tuning process. Initially, we tokenize the text and one-hot-encode the POS tags, maintaining alignment between the two sequences. Whenever a Greek word is tokenized into multiple tokens, the corresponding tag is replicated the same number of times. The three scenarios differ in how the vectors are processed and transformed, but in all cases, the combined vector constitutes input to the encoder stack, retaining the same number of dimensions as during the model's pre-training phase (768 dimensions for *-base* and 1024 for *-large*). We visualize the three approaches in Figure 4 and discuss them in the subsequent paragraphs.

In *Embeddings – Sum (emb-sum)*, morphological tags are embedded in a vector space of the same size as the one used by the base model. The embedded text and POS tags are then summed, and the result is passed to the encoder stack.

Embeddings – Autoencoder (emb-auto) also sums the two sequences positionally, but first, the tags sequence is embedded in a smaller space. Given the small number of unique tags in the tag sets (roughly 1000), the tag embedding layer may essentially one-hot-encode the tags. Thus, this approach aims to force the model to synthesize information carried within the tags by compressing them and then decompressing them back to the expected number of dimensions. The compressed embedding size is a hyperparameter to be tuned.

In the last approach – *Embeddings – Concatenation (emb-concat)* – the two sequences are concatenated, but to ensure that the output vector is of the desired size, an extra linear layer is introduced to reduce the number of dimensions in the text embedding. The dimensions in the text embedding and morphological embedding sum up to the desired size, and the text-to-tag ratio in the output vector is a tunable hyperparameter.

Trimming Our experiments aim to evaluate the



(a) emb-sum (b) emb-auto (c) emb-concat

Figure 4: The figures depict three embedding-based strategies for incorporating morphological information into the model's input. The first strategy (emb-sum) combines text (T) and morphological (M) embeddings into a single vector using positional sum. The second compresses morphological embeddings before decompression and summation with the text counterpart. The last scenario compresses both text and morphological embeddings and concatenates the resulting vectors.

performance of proposed methods, regardless of their efficiency in encoding inputs. With a sequence limit of 256, some models may appear to perform worse simply because they fit fewer words from the source text into the input. To ensure fair comparisons, we trim each verse to the number of words that can be encoded by the least efficient setup among all 144 parameter combinations.

3.5 Training Details

The dataset was split into three subsets: training (7543 verses, 95%), validation (198 verses, 2.5%), and test (199 verses, 2.5%). Fine-tuning experiments encompassed 144 combinations, varying the target language, tag set, text preprocessing strategy, base model, and morphological information encoding strategy. Training employed an NVIDIA A100 GPU, with batch sizes ranging from 4 to 16 for training and 1 to 8 for validation, adjusted based on memory constraints. For *emb-**, the learning rate for new neural network layers was increased from the default 1e-3 to 3e-1, 1e-2, or 3e-3, with similar performance observed across these values. The optimizer remained Adafactor. In *emb-auto* and *emb-concat*, morphological embedding size was set to 64 dimensions. A token limit of 256 was enforced across all scenarios.

4 Evaluation

Model Output The output sequence contains translations for each Greek word separated by sentinel tokens, similar to the input formatting. We employ *BLEU* (Post, 2018) to measure model performance. However, prior to comparing predictions and references, we remove the separator tokens from the output sequences to prevent the metric from rewarding a model solely for structuring the output correctly. Sequences are trimmed as during training (see **Trimming** in Section 3.4), and predictions and references are further trimmed to the same number of 'translation blocks.'

5 Results

We address each research question in the subsequent sections, beginning with an examination of the overall performance of the models. We then compare the performance of each base model used for fine-tuning. Finally, we investigate the impact of morphological metadata and text preprocessing on the final results. All scores presented in this section represent the BLEU score obtained on the test split.

5.1 Overall Performance

Given the large number of parameter sets (144), this section discusses general trends observed in the data on an aggregate level before delving into more detailed breakdowns in subsequent sections.

Empirical cumulative distribution functions (eCDFs) for Ancient Greek to English and to Polish translation are presented in Figure 5. Despite visible differences favoring experiments in English, approximately the top 40% of scenarios for both groups hover around the same values of BLEU (45-55), with slightly higher scores for the English subset. The discrepancy might possibly stem from the fact that English was part of pre-training corpus for both mT5 and PhilTa, while Polish was only seen by mT5.

The close results in the two groups suggest that the strict syntactical regime of interlinear translations may allow for cross-language comparisons, which are normally impossible due to differences in syntax and morphology.

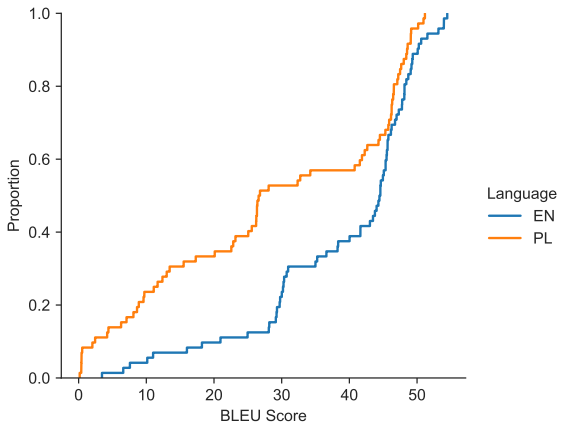


Figure 5: Empirical Cumulative Distribution Function (eCDF) for BLEU scores secured by the 144 fine-tuned models. The results are divided into two categories according to the solved task — English (EN) and Polish (PL) translation.

	PL		EN	
Base Model	Avg	Best	Avg	Best
GreTa	27.51	49.08	33.22	49.36
PhilTa	13.21	26.79	42.68	**54.46**
mT5-base	28.83	47.22	34.07	45.65
mT5-large	**45.00**	**51.16**	**45.74**	48.41

Table 3: Aggregated performance of each of the base models. The rows showcase the average and best results of a given base model on translation into Polish (PL) and English (EN), respectively.

5.2 Base Model

Table 3 provides a general comparison of the base models. On average, mT5-large outperformed all other base models and achieved the best result in Polish translation. However, in English translation, its best score fell behind those achieved by GreTa and PhilTa, with PhilTa as the winner. Despite not being pre-trained on English or Polish, GreTa performed similarly to mT5-base in both tasks, slightly outperforming it both on average and in maximum score, even though mT5-base was pre-trained on both English and Polish. Surprisingly, PhilTa struggled to match the results of other models in Polish translation, failing to surpass a BLEU score of 30 in its best run.

These results suggest that, for translations of classic texts, models pre-trained on both the source and target languages offer the best performance. If such models are not available, the next best options appear to be selecting a model pre-trained on the source language or opting for a larger multi-lingual model.

Based on these results, fine-tuning a model pre-trained on both Ancient Greek and Polish, similar to how PhilTa was utilized for Ancient Greek and English, could render the best results. However, as of now, such a model does not exist.

5.3 Impact of Morphological Tags

This section answers two questions. Firstly, we address the question of whether the inclusion of morphological metadata leads to improved performance on the interlinear translation task and if so, how the metadata should be encoded. Secondly, we assess the impact of the chosen tag set.

Impact of Encoding Strategy Table 4 compares different morphological feature encoding strategies. The results consistently show improved performance on the interlinear translation task when morphological metadata is included, regardless of the chosen encoding strategy. The best-performing morphologically-enhanced models outperform the baseline (text-only) by approximately 20% for Polish (51.16 vs 42.65) and 21% for English (54.46 vs 44.95). This challenges the common belief that pre-trained transformer cannot utilize such information to perform better on NLP tasks, particularly for the translation tasks.

Furthermore, the results suggest that using a separate embedding representation is preferable to encoding the morphological information directly within the text to be translated. Across all strategies, embedding-based solutions outperform the approach that encodes tags within the text. Additionally, it's worth noting that the latter is also the least efficient in terms of both memory and time complexity among all tested scenarios.

When comparing the three embedding-based strategies, *emb-concat* yields the lowest scores and is most prone to convergence issues. *emb-auto* and *emb-sum* provide similar results, with *emb-sum* leading for Polish (BLEU 51.16) and *emb-auto* for English (54.46). The superior stability of the two sum-based methods over the concatenation-based one may stem from the lack of an additional compression layer for text embeddings. This extra layer can potentially disrupt the semantic representation that the model learned during pre-training, thereby hindering its performance.

Tag Set Comparison Table 5 presents results for each of the two tag sets. While both sets

	PL		EN	
Encoding	Avg	Best	Avg	Best
t-o	18.42	42.65	33.54	44.95
t-w-t	21.26	48.04	35.38	48.14
emb-sum	37.96	**51.16**	**47.84**	53.98
emb-auto	**40.42**	49.13	46.52	**54.46**
emb-concat	20.02	46.56	28.66	51.53

Table 4: Performance comparison of encoding strategies. Rows display average and best results of each encoding strategy for translation into Polish and English. The strategies are: text only (t-o), text with tags (t-w-t), embedding – sum (emb-sum), embedding – autoencoder (emb-auto) and embedding – concatenation (emb-concat), respectively.

yielded strong results, the average score and top-performing models favored the tag set from Bible Hub, outperforming Oblubienica in both translation tasks. Notably, the top-performing tag set (refer to Table 1) had roughly 50% fewer forms. This difference in score might be attributed to insufficient training data for the model to learn to represent less frequent forms, or it might stem from a difference in tagging quality. Further investigation would be necessary to determine the cause.

	PL		EN	
Tag Set	Avg	Best	Avg	Best
BH	**30.19**	**51.16**	**40.39**	**54.46**
OB	29.64	50.16	38.81	53.95

Table 5: Effect of tag set selection on translation performance. Rows show average and best results of models using Bible Hub (BH) and Oblubienica (OB) tag sets for translation into Polish (PL) and English (EN).

5.4 Impact of Preprocessing

The results analyzing the impact of preprocessing strategy are presented in Table 6. In the vast majority of cases, regardless of the chosen tokenizer, runs with diacritics achieved better results both on average and in the best-case scenario. We find it interesting that the inclusion of diacritics in the translation task generally improves the results, especially considering that in many experiments found in the literature for the analysis of Ancient Greek, diacritics are often removed. We postulate that more attention should be paid to the preservation of this additional information, given its value for the models' performance, as shown in our experiments.

	PL		EN	
Preprocessing	Avg	Best	Avg	Best
Diacritics	**30.03**	**51.16**	**41.33**	**54.46**
Normalized	27.24	50.95	36.53	50.61

Table 6: Impact of preprocessing strategy (with diacritics or with normalization) on final results. Rows display average and best results on the test dataset for each approach on the two translation tasks.

6 Conclusions

We have presented research addressing interlinear translation from Ancient Greek, offering a dataset for assessing multiple sequence-to-sequence models. Among our findings, PhilTa emerges as the top performer for English, while mT5-large excels for Polish. Surprisingly, GreTa, pretrained solely on Ancient Greek, yields comparable results.

Our proposed methods for encoding morphological information via dedicated embedding layers consistently improve translations, particularly when the model sums morphological embeddings with text embeddings. This approach leads to an 8.5 percentage point improvement for Polish and a 9.5 percentage point improvement for English compared to models without morphological data.

Additionally, we have observed a positive impact on model scores when preserving original diacritics, a practice often overlooked in NLP studies focusing on Ancient Greek.

7 Ethics

While our research involves the translation of theologically significant documents, it is important to note that our primary focus is on evaluating machine translation methodologies. We acknowledge the potential for bias in both the models and the underlying data, particularly in texts of religious significance. Therefore, we caution against drawing any theological conclusions from our translations, as our study does not investigate or account for potential biases. Our aim is to contribute to the advancement of machine translation technology while maintaining a neutral stance on theological interpretations.

Additionally, we acknowledge the usage of ChatGPT for assistance with text editing and refining code for experiments.

8 Limitations

Limited Text Scope We focused only on the New Testament for our research. This decision was influenced by several factors. Firstly, interlinear translations are less common in the public domain compared to standard parallel corpora for training MT systems. Additionally, aligning the source translation for two distinct target languages requires substantial resources and good quality of data, both being scarce. While our study only looked at the New Testament, future research could include texts from Ancient philosophers (like Plato) and writers (such as Homer) to better assess the impact of tested features on model's performance.

Ancient Greek Interlinear translation serves as a valuable educational tool in the study of ancient languages such as Ancient Greek, Latin or Sanskrit. Our study focused exclusively on Ancient Greek, primarily because it is the source language of the New Testament – the corpus of our choice – making it a logical choice for our research. In addition to the issues from the previous section, such as obtaining high-quality interlinear translations for other languages, a significant limitation was the scarcity of language models specifically trained for these ancient languages. While some models exist for Latin (e.g. Ströbel (2022)), the availability of models for Sanskrit is limited.

Transformer Models In our study, we focused exclusively on neural networks, specifically the transformer architecture, which has dominated recent NLP research. However, new paradigms are emerging, such as the S4 (Gu et al., 2022) architecture implemented in the Mamba language model (Gu and Dao, 2023). Despite this, transformers benefit from a robust ecosystem of pre-trained models available for many languages (including Polish and Ancient Greek) and tasks (such as sequence-to-sequence, essential for MT). Evaluating these new paradigms would require pre-training new models, which is beyond the scope of our current research.

Inclusion of Two Target Languages Our study focused on only two target languages: English and Polish. Potential alternatives could include Turkish, an agglutinative language, and languages from the Chinese family, which feature a distinct writing system that could significantly impact interlinear translation. However, these languages not only differ linguistically but also culturally. To conduct comparative studies effectively, we would need to include central texts from these cultures, such as the Quran and the works of Confucius. This expansion would significantly complicate our research and exceed our current objectives.

Bias in Generative Language Models There is a potential risk that the models used for translating the Bible text may have been previously trained on its parts, biasing the output. Instead of measuring their translation ability, we might simply be assessing their capacity to regenerate memorized Bible text. Carlini et al. (2021) explored methods to detect whether samples generated by large language models (LLMs) come from their training data, using techniques like perplexity measurement and model-to-model comparison. Their findings revealed that 604 out of 1800 samples generated by GPT-2 (Radford et al., 2019), including 25 from religious texts such as the Bible and the Quran, were identified as originating from the training data, suggesting a tendency of these models to reproduce text from their training datasets.

References

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models.

David Carter. 2019. Using translation-based CI to read Latin literature. *Journal of Classics Teaching*, 20(39):90–94. Publisher: Cambridge University Press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Chanambam Sveta Devi, Bipul Syam Purkayastha, and Loitongbam Sanayai Meetei. 2022. An empirical study on English-Mizo Statistical Machine Translation with Bible Corpus. *International journal of electrical and computer engineering systems*, 13(9):759–765. Publisher: Elektrotehnički fakultet Sveučilišta J.J. Strossmayera u Osijeku.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthias Gerner. 2018. Why Worldwide Bible Translation Grows Exponentially. *Journal of Religious History*, 42(2):145–180. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9809.12443.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces.

Albert Gu, Karan Goel, and Christopher Ré. 2022. Efficiently modeling long sequences with structured state spaces.

Arvi Hurskainen. 2018. Can machine translation assist in Bible translation? (62).

Alek Keersmaekers, Wouter Mercelis, and Toon Van Hal. 2023. Word Sense Disambiguation for Ancient Greek: Sourcing a training corpus through translation alignment. In *Proceedings of the Ancient Language Processing Workshop*, pages 148–159, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kevin Krahn, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence Embedding Models for Ancient Greek Using Multilingual Knowledge Distillation. In *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Zixuan Ling, Xiaoqing Zheng, Jianhan Xu, Jinshu Lin, Kai-Wei Chang, Cho-Jui Hsieh, and Xuanjing Huang. 2023. Enhancing unsupervised semantic parsing with distributed contextual representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11454–11465, Toronto, Canada. Association for Computational Linguistics.

Eva Martínez Garcia and Álvaro García Tejedor. 2020. Latin-Spanish Neural Machine Translation: from the Bible to Saint Augustine. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 94–99, Marseille, France. European Language Resources Association (ELRA).

Angelina McMillan-Major. 2020. Automating Gloss Generation in Interlinear Glossed Text. Publisher: University of Mass Amherst.

Sarah Moeller and Mans Hulden. 2018. Automatic Glossing in a Low-Resource Setting for Language Documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sarah Moeller, Ling Liu, and Mans Hulden. 2021. To POS tag or not to POS tag: The impact of POS tags on morphological learning in low-resource settings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 966–978, Online. Association for Computational Linguistics.

Sebastian Nehrdich and Oliver Hellwig. 2022. Accurate dependency parsing and tagging of latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 20–25.

Ravinga Perera, Thilakshi Fonseka, Rashmini Naranpanawa, and Uthayasanker Thayasivam. 2022. Improving English to Sinhala Neural Machine Translation using Part-of-Speech Tag. ArXiv:2202.08882 [cs].

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Frederick Riemenschneider and Anette Frank. 2023a. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023b. Graecia capta ferum victorem cepit. Detecting Latin Allusions to Ancient Greek Literature. ArXiv:2308.12008 [cs].

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021a. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2021)*, pages 128–137. Association for Computational Linguistics.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021b. A pilot study for BERT language modelling and morphological analysis for ancient and medieval Greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137, Punta Cana, Dominican Republic (online). Association for Computational Linguistics.

Thea Sommerschield, Yannis Assael, John Pavlopoulos, Vanessa Stefanak, Andrew Senior, Chris Dyer, John Bodel, Jonathan Prag, Ion Androutsopoulos, and Nando de Freitas. 2023. Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, 49(3):703–747.

Phillip Benjamin Ströbel. 2022. Roberta base latin cased v1.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Ivan P. Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in Plutarch's Shadows. ArXiv:2211.05673 [cs].

Tariq Yousef, Chiara Palladino, David J. Wright, and Monica Berti. 2022. Automatic Translation Alignment for Ancient Greek and Latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107, Marseille, France. European Language Resources Association.

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhong Zhou, Lori Levin, David R. Mortensen, and Alex Waibel. 2020. Using Interlinear Glosses as Pivot in Low-Resource Multilingual Machine Translation. ArXiv:1911.02709 [cs].