Low-Resource Interlinear Translation: Morphology-Enhanced Neural Models for Ancient Greek

Maciej Rapacz
AGH University of Krakow
mrapacz@agh.edu.pl

Aleksander Smywiński-Pohl
AGH University of Krakow
apohllo@agh.edu.pl

Abstract

Contemporary machine translation systems prioritize fluent, natural-sounding output with flexible word ordering. In contrast, *interlinear translation* maintains the source text's syntactic structure by aligning target language words directly beneath their source counterparts. Despite its importance in classical scholarship, automated approaches to interlinear translation remain understudied.

We evaluated neural interlinear translation from Ancient Greek to English and Polish using four transformer-based models: two Ancient Greek-specialized (GreTa and PhilTa) and two general-purpose multilingual models (mT5-base and mT5-large). Our approach introduces novel morphological embedding layers and evaluates text preprocessing and tag set selection across 144 experimental configurations using a word-aligned parallel corpus of the Greek New Testament.

Results show that morphological features through dedicated embedding layers significantly enhance translation quality, improving BLEU scores by 35% (44.67 → 60.40) for English and 38% (42.92 → 59.33) for Polish compared to baseline models. PhilTa achieves state-of-the-art performance for English, while mT5-large does so for Polish. Notably, PhilTa maintains stable performance using only 10% of training data.

Our findings challenge the assumption that modern neural architectures cannot benefit from explicit morphological annotations. While preprocessing strategies and tag set selection show minimal impact, the substantial gains from morphological embeddings demonstrate their value in low-resource scenarios.[1]

[1]We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017156. The research presented in this paper was partially supported by the funds of Polish Ministry of Science and Higher Education assigned to the AGH University of Kraków.

1 Introduction

Machine translation (MT) is a well-established subfield in Natural Language Processing (NLP), 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. While these conventional MT systems prioritize natural and fluent translations, there exists a spectrum of translation approaches, ranging from free translation to extremely literal renderings.

At the far end of this spectrum lies *interlinear translation* (Shuttleworth and Cowie, 2014), a method that strictly preserves the source text's syntactic structure. This approach aligns target language words directly below or above their corresponding source text elements. Commonly applied to ancient (and oftentimes sacred) texts, this method allows readers unfamiliar with the source language to understand both the meaning and structure of the original text. Such alignment enables students to critically evaluate translations by observing how specific source words were translated, which is especially crucial for interpreting source texts in fields such as philosophy and religious studies. Figure 1 illustrates an example of interlinear translation.

Despite the significance of interlinear translation, which Benjamin (1923/2000) called "the archetype or ideal of all translation", there has been limited research on automating this process. This may be attributed to the pre-existing interlinear translations for many influential texts. However, we believe automating this process remains relevant, making these texts more accessible to those without expertise in ancient languages.

While prior research (Tenney et al., 2019) suggests that modern neural architectures like BERT inherently learn linguistic patterns without explicit

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 145–165
January 20, 2025.

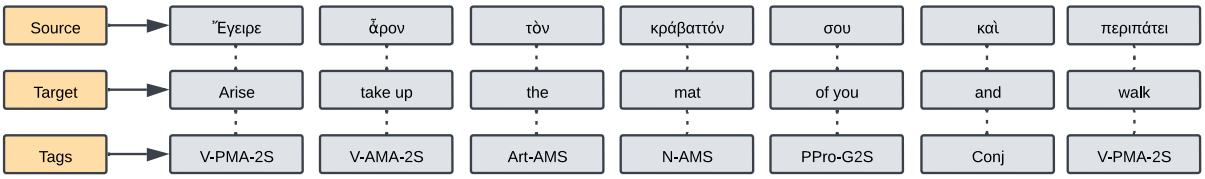


Figure 1: Interlinear translation example from John 5:8, showing Ancient Greek source text, English translations, and BibleHub morphological tags.

linguistic annotations, our findings challenge this assumption in low-resource scenarios. We demonstrate that for small datasets with limited sentence pairs, properly encoded morphosyntactic tags significantly enhance translation performance.

In the presented paper we aim to achieve the following objectives:

- Evaluate interlinear translation of Ancient Greek texts using modern MT models for both English and Polish targets,
- Study how linguistic features affect translation quality, focusing on morphological tags and text preprocessing methods,
- Compare specialized ancient language models (PhilTa, GreTa) with general multilingual transformers (mT5) in low-resource settings.

We focus on the Greek New Testament as our source corpus, given its international significance, original Ancient Greek text (Nestle et al., 2012), and abundant translations. Our analysis compares model performance between two syntactically distinct target languages: English (positional) and Polish (inflectional).

Our contributions This paper presents three main contributions. Firstly, we construct a novel word-level-aligned parallel corpus of the Greek New Testament with interlinear translations in English and Polish, based on data from BibleHub (BH) and Oblubienica (OB).

Secondly, we present the first systematic approach to automating interlinear translation using modern machine learning methods. We evaluate four base models – PhilTa, GreTa (Riemenschneider and Frank, 2023a) and mT5 (Xue et al., 2020) (in two sizes) – across 144 experimental scenarios, providing comprehensive insights into the task's feasibility.

Finally, our experiments demonstrate that incorporating morphological information in low-resource settings significantly improves translation quality, with proper morphological tag encoding yielding improvements of 38% for Polish and 35% for English over the baseline. We also find that the choice of normalization method and tag set has minimal impact on model performance.

We make the resources developed as part of this work (parallel corpus, training code, and fine-tuned models) publicly available.[2]

2 Related Work

Recent years have witnessed substantial advances in applying machine learning to ancient languages, particularly Ancient Greek (Sommerschield et al., 2023). While most research focuses on tasks like POS tagging and lemmatization, machine translation of ancient texts presents unique challenges that intersect multiple research areas. This section examines relevant work across these domains.

2.1 Current State of Machine Translation

Recent studies demonstrate significant progress in machine translation across different resource settings. For high-resource language pairs, state-of-the-art models achieve BLEU scores between 30-33 when translating into English, and 22-26 when translating from English (Zhang et al., 2020). More recent research (Xu et al., 2024) reports similar performance levels, with BLEU scores of 32.2 for translation into English and 27.8 for translation from English for Central and Eastern European languages.

For low-resource scenarios (less than 0.1M training pairs), performance varies significantly but remains surprisingly robust. Models trained on limited data consistently outperform zero-shot translation approaches, which typically achieve BLEU scores between 4 and 15 (Zhang et al., 2020).

[2] https://github.com/mrapacz/loreslm-interlinear-translation

2.2 Machine Translation for Ancient Greek

Recent research in Ancient Greek Natural Language Processing has primarily focused on encoder models from the BERT family (Devlin et al., 2019). These models have been successfully applied to foundational tasks like Part-of-speech tagging, lemmatization (Singh et al., 2021), translation alignment (Yousef et al., 2022; Keersmaekers et al., 2023) and dependency parsing (Nehrdich and Hellwig, 2022).

Despite this progress in encoder models, dedicated sequence-to-sequence models for Ancient Greek remain scarce. Only one notable effort exists: Riemenschneider and Frank (2023a) developed two T5-based models – *GreTa* (monolingual) and *PhilTa* (trilingual, trained on Ancient Greek, Latin, and English).

This scarcity of translation models is matched by limited parallel corpora. The OPUS project (Tiedemann, 2012), a major repository of parallel texts, contains just 635 sentence pairs for Ancient Greek-English and only 2 pairs for Ancient Greek-Polish. These numbers firmly place Ancient Greek translation in the low-resource category according to established benchmarks (Zhang et al., 2020), which classify language pairs with fewer than 0.1M training examples as low-resource.

2.3 Machine Translation for Biblical Texts

The exponential growth in Bible translations across languages (Gerner, 2018) has made it a valuable parallel corpus for machine translation research. However, most studies utilizing biblical texts focus on translation between modern language pairs, such as Navajo-English (Liu et al., 2021), Mizo-English (Devi et al., 2022), and other contemporary languages (Hurskainen, 2020), rather than working with the original ancient source texts.

While some research has explored ancient language processing of biblical texts, such as Latin-Spanish translation (Martínez Garcia and García Tejedor, 2020) and Greek-English corpus alignment (Riemenschneider and Frank, 2023b), these efforts primarily focus on intermediate translations or specific NLP tasks like embedding evaluation (Krahn et al., 2023). Direct translation from original Ancient Greek biblical manuscripts to modern languages remains largely unexplored, particularly in the context of structured translation approaches that preserve source text characteristics.

2.4 Interlinear Translation Approaches

While we have not found prior work directly addressing interlinear translation, the related field of interlinear glossing has been extensively studied, particularly in the context of language documentation and preservation. Morpheme-level glossing dominates research compared to word-level glossing, likely due to its applications in language preservation. Word-level glossing, while less common, serves primarily as a tool for readers to better understand source texts without necessarily knowing the source language (Carter, 2019).

Research has explored both using source language glosses to generate free translations (Zhou et al., 2020) and generating glosses as part of the output (Moeller and Hulden, 2018; McMillan-Major, 2020; Zhao et al., 2020). The field's significance is highlighted by SIGMORPHON's recent introduction of an interlinear glossing shared task, which focuses on producing morpheme-level grammatical descriptions of input sentences.

2.5 Role of Morphological Information

The impact of morphological features on neural models, especially in low-resource settings, is still under investigation. While Moeller et al. (2021) found mixed results for part-of-speech tags, Perera et al. (2022) reported improvements in specific language pairs. Overall, incorporating linguistic information, as shown in Chakrabarty et al. (2020, 2022, 2023), can enhance translation quality in resource-constrained scenarios.

Chakrabarty et al. (2020) introduced a neural model using linguistic features via self-relevance and word-relevance methods. Both involve projecting feature embeddings and applying a sigmoid non-linearity to combine with original embeddings. These methods improved BLEU scores by 0.67-3.09 points for English-to-Asian language translation. Chakrabarty et al. (2022) showed that simple feature embedding concatenation with a Transformer model pre-trained on span reconstruction also yields significant improvements.

For Ancient Greek, with its rich morphology and relatively free word order, the value of morphological information may be more significant. Beyond basic part-of-speech tags, detailed morphological features – including mood, tense, voice, person, case, gender, and number – could potentially enhance translation quality, though this hypothesis requires empirical validation.

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.

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 prepare a word-level-aligned corpus consisting of two interlinear translations available online – an Ancient Greek New Testament translated into English (sourced from BibleHub) and one into Polish (sourced from Oblubienica). Each translation contains source text, translation, and morphological tags, discussed in the following paragraphs.

Source Text The corpora include different critical editions of the Greek text. Specifically, the Greek text in the Oblubienica corpus follows Nestle Aland Novum Testamentum Graece 28 – NA28 (Nestle et al., 2012), while Bible Hub merges NA28's predecessor – NA27 (Aland, 1927) – with other critical editions (Robinson and Pierpont, 2005; Scrivener, 1881; Westcott and Hort, 1882; Holmes, 2010; Nestle, 1904), each marked using special quotes. Although the primary disparity between the two corpora lies in the textual edition used, there are additional distinctions, which include varying casing, usage of diacritics, and punctuation, as depicted in Figure 2.

Translations The Oblubienica corpus provides a Polish translation that combines three sources: Gdansk Bible (1632), Updated Gdansk Bible (2009) and Polish Interlinear Translation (1993). Bible Hub provides an English translation, though its source is not specified. Both translations are aligned word-by-word with the Greek text.

Tag Sets share common categories like Part of Speech, Pronoun (with subtypes), Person, Tense, Mood, Voice, Case, Number, Gender, and Degree (see Appendix A). The corpora differ in total unique tags (Oblubienica: 1068, Biblehub: 693), primarily due to verbs (Oblubienica: 743, Biblehub: 385), while other parts of speech have similar counts (Table 1).

Part of Speech	Bible Hub	Oblubienica
Verb	385	743
Pronoun	169	193
Adjective	68	56
Noun	31	39
Article	30	23
Adverb	3	5
Particle	3	4
Interjection	1	1
Preposition	1	1
Conjunction	1	1
Hebrew Word	1	1
Aramaic Word	0	1

Table 1: Comparison of the number of unique morphological tags per part of speech (including dedicated categories for Hebrew and Aramaic words) between Oblubienica and Bible Hub.

Oblubienica's detailed tagging system results in more unique verb tags. It distinguishes first and second aorist tenses (+100 forms), marks Attic dialect verbs (+100 forms), and notes uncertain participle genders (+50 forms) more often. Additionally, it employs more combinations of voice categories with tense and mood (+370 forms). This gap might narrow with a larger dataset, as Bible Hub's system allows for these distinctions but doesn't utilize them fully.

Both corpora use natural language tags (e.g., Article – Nominative Masculine Plural) and abbreviated forms (e.g., A-NMP). When encoding tags directly in text, we use the shorter forms due to model memory constraints.

The corpora occasionally differ in word classification – for example, δαυιδ (David) is tagged as *N-GMS* (Noun – Genitive, Masculine, Singular) in Bible Hub but as *ni proper* (Properly Indeclinable Noun) in Oblubienica.

Corpus Alignment To enable tag set comparison across models, we performed word-level alignment between the two corpora. First, we standardized the Bible Hub text by retaining only NA27 textual editions to match Oblubienica's NA28 version. We then implemented a hierarchical matching algo-

rithm that first attempted exact word matches, followed by within-verse matches, and finally nearest-neighbor matching for ambiguous cases. This approach successfully aligned over 99% of words between the corpora. We excluded the remaining unmatched words, to maintain consistent tag coverage across both datasets.

Word Forms Our corpus maintains two versions of each Greek word. The first version preserves diacritics, following Bible Hub's spelling which includes breathing marks, accents, and other diacritical signs. The second version is normalized: stripped of diacritics and converted to lowercase. Since our corpora are aligned, we use Bible Hub's spelling as the canonical form with diacritics, discarding the corresponding words in Oblubienica. This dual representation enables experiments with both diacritical and normalized text processing approaches, following two major schools of thought in Ancient Greek NLP: preservation of full orthographic information (Riemenschneider and Frank, 2023a) versus normalized processing (Yamshchikov et al., 2022).

Final Dataset The aligned corpus contains Greek words (with diacritics and normalized), paired with morphological tags (Oblubienica and Bible Hub) and translations (English and Polish). Table 2 summarizes the dataset.

Statistic	Count
Verses	7,940
Words (GR)	137,323
Words (PL / EN)	133,581 / 185,722
Unique Tags (OB / BH)	1,068 / 693

Table 2: Corpus statistics: verses, source words (Greek), target words (Polish/English), and unique morphological tags in the corpus (Oblubienica/BibleHub).

3.2 Base Models

We use four T5-based models (Chung et al., 2022): GreTa and PhilTa (Riemenschneider and Frank, 2023a) (both T5-base variants), and mT5-base/large (Xue et al., 2020). GreTa was trained on Ancient Greek texts, while PhilTa was trained on Ancient Greek, Latin and English. mT5 was trained on mC4 (Raffel et al., 2020), covering 101 languages including English and Polish. While mC4 includes Modern Greek, it does not contain Ancient Greek – these are distinct languages that differ significantly in vocabulary, grammar and syntax. We include mT5-base to match GreTa/PhilTa's size and mT5-large to test if more parameters help performance.

3.3 Tokenizer Efficiency

We evaluate tokenizer efficiency across our models using the average number of tokens per word metric (Yamshchikov et al., 2022), reported in Table 3. For Greek text with diacritics, mT5 requires approximately twice as many tokens per word compared to PhilTa or GreTa. However, this gap disappears when processing normalized text. For Polish, English, and morphological tags, mT5 generally achieves better tokenization efficiency.

The tag tokenization shows notable differences between corpora, with Oblubienica tags requiring significantly more tokens than Bible Hub tags. This stems from Oblubienica's more verbose tagging format – for example, where Bible Hub uses *N-DFS*, Oblubienica expresses the same information as *n_Dat Sg f*. It is worth noting that this distinction affects only the scenarios where morphological tags are encoded as part of the text input.

Tokenizer Dataset	GreTa	PhilTa	mT5
GR – diacritics	**1.49**	1.50	3.15
GR – normalized	2.45	**2.30**	2.31
PL	4.02	4.14	**2.31**
EN	3.45	**1.86**	1.94
Tags (OB)	7.20	6.89	**5.39**
Tags (BH)	5.00	5.20	**3.76**

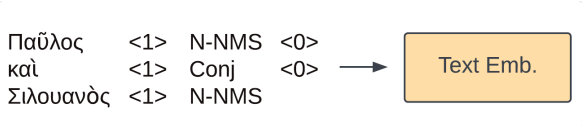
Table 3: Overview of tokenization metrics. The consecutive rows display the average number of tokens required by each tokenizer for: a Greek word with diacritics, a normalized Greek word, a Polish word, an English word, a tag from the Oblubienica (OB) tag set, and a tag from the Bible Hub (BH) tag set, respectively.

3.4 Model Inputs

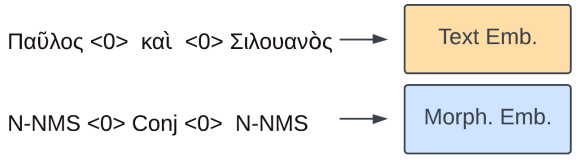
We evaluate the impact of morphological tags on interlinear translation performance through five scenarios, grouped into three categories. Each category is visualized below:

Baseline No morphological information; Greek words separated by sentinel tokens.

Tags Within Text (t-w-t) Tags encoded as part of the text input, using sentinel tokens to separate word-tag pairs and demarcate word-tag boundaries:



Morphological Embeddings (emb-*) Introduces a dedicated embedding layer trained during fine-tuning. Text is tokenized and tags are one-hot-encoded, maintaining alignment. For multi-token words, tags are replicated. The combined vector input maintains pre-training dimensions (768 for *-base*, 1024 for *-large*). This approach is visualized below:



We explore three variations of this embedding-based approach:

- *Embeddings – Sum (emb-sum)*: Sums embedded text and tag embeddings.
- *Embeddings – Autoencoder (emb-auto)*: Compresses and decompresses tag embeddings before summing with text embeddings.
- *Embeddings – Concatenation (emb-concat)*: Concatenates compressed text and tag embeddings.

These three solutions are visualized in detail in Figure 3.

3.5 Model Output Format

Models output translations in a format similar to text-only input, using distinct tokens to separate translated Greek words.

3.6 Experimental Setup

Dataset Splits The New Testament's 7940 verses were randomly shuffled and split into training (6352 verses, 80%), validation (794 verses, 10%), and test (794 verses, 10%) sets.

Experiment Configurations Our experiments covered 144 distinct configurations, as detailed in Table 4. This number is lower than the theoretical maximum of 160 combinations since text-only scenarios do not use morphological tags.

Factor	Options	#
Language	EN, PL	2
Tag Set	BH, OB	2
Preprocessing	Diacritics, Normalized	2
Base Model	mT5-base, mT5-large, GreTa, PhilTa	4
Input Encoding	baseline, t-w-t, emb-sum, emb-auto, emb-concat	5

Table 4: Experiment configuration factors and their options.

Training Configuration Each experiment used an A100 GPU with an effective batch size of 32 (achieved through gradient accumulation). For the morphological embedding layers, we used a dedicated optimizer and learning rate, as shown in Table 5.

Parameter	Value
Effective Batch Size	32
Morph. Emb. Optimizer	Adafactor
Morph. Emb. Learning Rate	3e-3
Morph. Emb. Size	64
Tokenizer Max Length	512

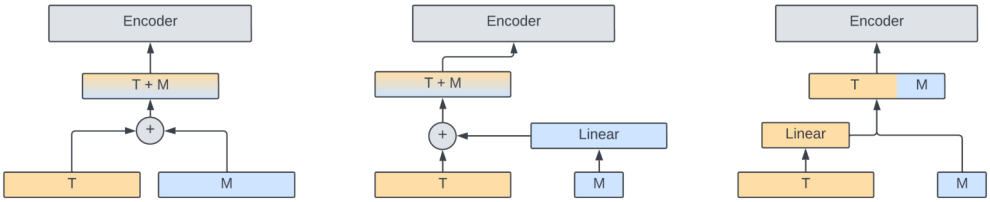
Table 5: Training hyperparameters.

Sequence Length Handling We set a maximum tokenizer length of 512 tokens per verse to match the models' pre-training configuration. To ensure fair comparison across all parameter combinations, we normalized verse lengths by trimming each verse to the number of words that could be encoded by the least efficient model configuration. This approach resulted in the removal of only 151 words (0.11%) from the dataset.

4 Evaluation

We evaluate model performance using *BLEU* (Papineni et al., 2002) and *SemScore* (Aynetdinov and Akbik, 2024) backed by `all-mpnet-base-v2`[3]. While modern metrics like COMET (Rei et al., 2020) could provide better assessment, they lack Ancient Greek support, so we could not apply them in these experiments. To ensure fair evaluation, separator tokens are removed from the output sequences before comparison with references,

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2



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

Figure 3: Three embedding-based strategies for incorporating morphological information: (a) positional sum of text (T) and morphological (M) embeddings, (b) compression and decompression of morphological embeddings before summation, and (c) compression and concatenation of both text and morphological embeddings.

preventing the metrics from artificially rewarding proper output formatting. Statistical significance of differences between configurations was assessed using two-sided Mann-Whitney U tests (Nachar et al., 2008).

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 Feasibility of Automated Interlinear Translation

BLEU and SemScore metrics for all experiment sets are presented in Figure 4 (see Appendix B for complete results).

Top results for both languages are very high, showing that the task is feasible – SemScore of 0.8 was surpassed and BLEU scores above 60 were achieved.

Both translation tasks received comparable top results, but in case of Polish there is a visible sample of results (roughly 40%) that never surpassed a BLEU score of 2. However, looking at how these results perform at SemScore, they're usually placed between 0.4 and 0.7. The plot allows for further analysis of discrepancies between the two metrics. While both metrics are strongly correlated, the correlation is not as strong for Polish (r=0.89) as for English (r=0.97). A brief, manual analysis of the unsuccessful experiments with BLEU < 2 shows that SemScore values of 0.7 can indeed be treated as very low.

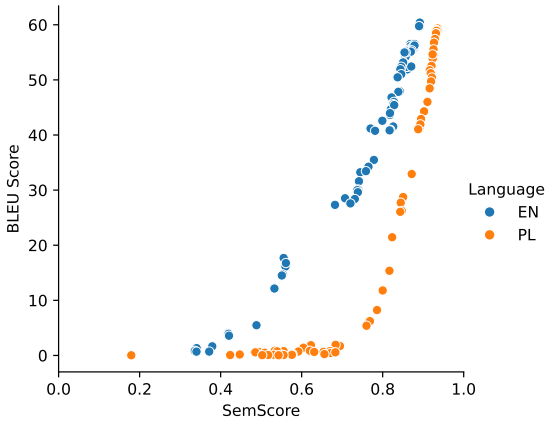


Figure 4: Distribution of BLEU and SemScore for English and Polish translations across 144 fine-tuned models.

The top results between languages suggest that interlinear translations' strict syntax may enable cross-language comparisons that normally are impossible in regular, free translation settings.

5.2 Impact of Linguistic Features

We examine the impact of morphological metadata on translation performance, focusing on encoding strategies and tag set selection.

Morphological Feature Integration Table 6 compares morphological feature encoding strategies (see Appendix C for more detailed results). Two embedding-based approaches significantly outperform the baseline model ($p < 0.05$), with improvements of 38% for Polish (59.33 vs 42.92) and 35% for English (60.40 vs 44.67) BLEU scores. This demonstrates that transformer models can effectively utilize dedicated morphological embeddings in low-resource settings.

Both *emb-auto* and *emb-sum* yield significant improvements ($p < 0.02$). In contrast, encoding tags directly in text (*t-w-t*) and *emb-concat* per-

form worse than baseline on average, even though the latter one achieves better results in the best case scenario (55.55 vs 42.92 for Polish and 55.93 vs 44.67 for English). The poor performance of this method likely stems from compression disrupting pre-trained representations, suggesting maintaining these representations is crucial for effective translation.

Regarding morphological tag sets, both Bible Hub and Oblubienica perform similarly across languages ($p > 0.07$, see Appendix D for detailed statistical analysis), suggesting that the encoding strategy has more impact on performance than tag set choice.

	PL		EN	
Encoding	Avg	Best	Avg	Best
baseline	17.57	42.92	32.40	44.67
t-w-t	12.73	41.93	30.86	46.00
emb-concat	10.74	55.55	26.33	55.93
emb-auto	**42.58**	**59.33**	**53.26**	**60.40**
emb-sum	36.75	58.92	48.04	60.10

Table 6: BLEU scores for different encoding strategies: baseline (text only), t-w-t (tags within text), emb-sum (embedding sum), emb-auto (embedding autoencoder), and emb-concat (embedding concatenation).

Text Preprocessing Strategies Analysis of preprocessing strategies (preserving vs. removing diacritics) showed no statistically significant differences in translation performance for either language ($p > 0.4$). Detailed results are presented in Appendix E.

5.3 Comparison of Model Architectures

Table 7 compares the base models (see Appendix F for more detailed results). For Polish translations, mT5-large significantly outperforms all other models ($p < 0.01$). For English, PhilTa achieves the highest scores, significantly outperforming GreTa and mT5-base ($p < 0.01$), though not mT5-large ($p = 0.46$). Larger models generally perform better – mT5-large outperforms mT5-base for both Polish ($p < 0.01$) and English ($p = 0.02$). Notably, PhilTa achieves the best English results despite being smaller than mT5-large, suggesting that targeted pre-training can compensate for model size. This raises the question of whether a model pretrained on both Ancient Greek and Polish could achieve similar gains for Polish translations.

	PL		EN	
Base Model	Avg	Best	Avg	Best
GreTa	21.69	51.30	29.94	55.22
PhilTa	3.12	15.37	**48.75**	**60.40**
mT5-base	27.75	54.63	32.46	52.43
mT5-large	**46.61**	**59.33**	44.13	56.51

Table 7: BLEU scores for base models on Polish (PL) and English (EN) translations.

We further compared learning efficiency between PhilTa and mT5-large models using varying amounts of training data (10%-80%). PhilTa demonstrated remarkable stability and efficiency, achieving a BLEU score in range [36.20 - 43.52] with just 10% of the dataset (794 verses), with performance improving monotonically as data increased. In contrast, mT5-large showed instability with smaller dataset samples, failing to achieve even a BLEU 1 with 10% data across all experiments, despite eventually matching PhilTa's performance with the full training split.

The results challenge the assumption that mT5-large's multilingual exposure offers an advantage in normalization. PhilTa's focused Ancient Greek pretraining proved more effective, excelling in low-resource settings with stable, efficient, and predictable performance. In contrast, mT5-large showed volatile scaling, making data-driven improvements uncertain.

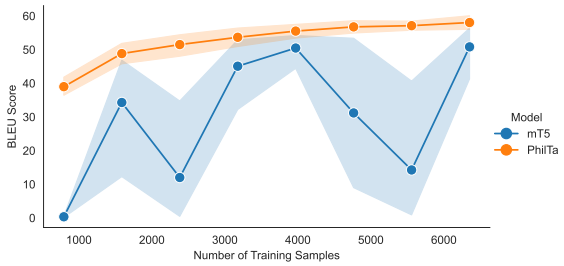


Figure 5: Mean learning efficiency with 95% confidence intervals comparing PhilTa and mT5-large models using varying training split sizes (10%-80%) on English translations.

6 Conclusions

We demonstrated the feasibility of automated interlinear translation from Ancient Greek, achieving BLEU scores above 60 and SemScore values exceeding 0.8 for both target languages. PhilTa outperformed larger models for English (60.40 BLEU),

while mT5-large performed best for Polish (59.33 BLEU).

Our novel morphological information encoding through dedicated embedding layers substantially improved translation quality, with gains of 38% for Polish (59.33 vs 42.92 BLEU) and 35% for English (60.40 vs 44.67 BLEU) over the baseline.

PhilTa showed remarkable stability in low-resource scenarios, maintaining consistent performance (BLEU 36.20-43.52) with just 10% of the dataset, while mT5-large struggled with smaller samples. This challenges the assumption that exposure to multiple languages necessarily provides an advantage in adaptation.

The interlinear translations' strict syntax enabled cross-language comparisons, revealing different metric correlations (BLEU-SemScore: r=0.97 English, r=0.89 Polish). While models trained on text with preserved diacritics achieved numerically better results, these differences were not statistically significant. Similarly, the choice between morphological tag sets showed minimal impact across both target languages.

Future work could explore targeted Polish pretraining, given PhilTa's English success.

7 Ethics

We acknowledge the use of GPT-4 and Claude 3.5 Sonnet for assistance with text editing and experimental code refinement.

8 Limitations

Limited Corpus Scope Our research focused solely on the New Testament due to its readily available interlinear format. While this ensured a consistent dataset, it may limit the broader applicability of our findings. Future work should explore other classical texts with interlinear translations, such as the Septuagint or Homeric epics, to test our findings across varied genres and styles.

Bias in Generative Language Models Models used for translating Bible text may have been trained on it, risking biased output. Instead of testing translation ability, we might be assessing memorization. Carlini et al. (2021) used methods like perplexity and model-to-model comparison to detect training data in LLM outputs, finding that 604 of 1800 GPT-2 samples, including 25 from religious texts, originated from its training set.

Limited Dataset Size Our dataset of 137,000 words is small compared to modern machine translation datasets with millions of parallel sentences. This low-resource setting limits the models' ability to learn complex patterns and generalize, especially for ancient languages with scarce parallel data.

Ancient Greek Interlinear translation is a valuable tool for studying ancient languages like Ancient Greek, Latin, Sanskrit, and Syriac. Our study focused on Ancient Greek as the source language of the New Testament, our chosen corpus. Challenges included obtaining high-quality interlinear translations and the limited availability of language models for ancient languages, especially Sanskrit and Syriac.

Inclusion of Two Target Languages Our study focused on two target languages: English and Polish. Alternatives like Turkish or Chinese could add linguistic and cultural diversity, requiring central texts like the Quran or Confucian works. However, this expansion would complicate the research beyond our current scope.

Morphological Tag Coverage The morphological tagging systems we used, while comprehensive with over 700 - 1100 unique tags, may not capture all nuances of Ancient Greek grammar. Some rare grammatical constructions or dialectal variations might be inadequately represented, potentially affecting translation quality for specific text segments.

Transformer Models Our study focused on neural networks, specifically the transformer architecture, which dominates NLP research. Emerging paradigms, like the S4 architecture in the Mamba model (Gu and Dao, 2023), show promise, but transformers offer a strong ecosystem of pre-trained models for languages and tasks like sequence-to-sequence MT. Pre-training new models to evaluate these paradigms is beyond our scope.

Model Size Constraints Our research compared Ancient Greek models (GreTa: 250M, PhilTa: 300M) with the multilingual MT5-base (580M). While all performed well, mT5-large (1.2B) showed notable improvements, especially for Polish translation, suggesting larger models may better handle languages without dedicated pre-trained models. Future work could test performance beyond 1.2B parameters.

Cross-Cultural Evaluation Our evaluation prioritized linguistic accuracy over cultural and theological considerations. This is a limitation when translating religious texts, where interpretative traditions influence translation. Future work could address these cross-cultural dimensions.

References

1632. *Biblia Święta to jest Księgi Starego y Nowego Przymierza z Żydowskiego y Greckiego Języka na Polski pilnie y wiernie przetłumaczone*. Self-published, Gdańsk.

2009. *Pismo Święte. Stary i Nowy Testament: pilnie i wiernie przetłumaczone w 1632 roku z języka greckiego i hebrajskiego na język polski, z uwspółcześnioną gramatyką i uaktualnionym słownictwem*. Fundacja Wrota Nadziei, Toruń.

Kurt Aland. 1927. *Novum testamentum graece*. Württembergische Bibelanstalt.

Ansar Aynetdinov and Alan Akbik. 2024. Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity. *arXiv preprint arXiv:2401.17072*.

Walter Benjamin. 1923/2000. The task of the translator. In Lawrence Venuti, editor, *The Translation Studies Reader*. Routledge.

BibleHub. Interlinear Bible. `https://biblehub.com/interlinear/`. Accessed: 2024-10-04.

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. *Preprint*, arXiv:2012.07805.

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

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Hideki Tanaka, Masao Utiyama, and Eiichiro Sumita. 2022. FeatureBART: Feature based sequence-to-sequence pre-training for low-resource NMT. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5014–5020, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2020. Improving low-resource NMT through relevance based linguistic features incorporation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4263–4274, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abhisek Chakrabarty, Raj Dabre, Chenchen Ding, Masao Utiyama, and Eiichiro Sumita. 2023. Low-resource multilingual neural translation using linguistic feature-based relevance mechanisms. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(7):1–36.

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. *arXiv preprint*.

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. *Preprint*, arXiv:2312.00752.

Michael W Holmes. 2010. Society of biblical literature. *Greek New Testament: SBL Edition*.

Arvi Hurskainen. 2020. Can machine translation assist in Bible translation? *Technical Reports on Language Technology Report 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.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The Usefulness of Bibles in Low-Resource Machine Translation. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. 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.

Nadim Nachar et al. 2008. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20.

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.

E. Nestle, B. Aland, K. Aland, H. Strutwolf, and Universität Münster Institut für Neutestamentliche Textforschung. 2012. *Novum Testamentum Graece (Na28): Nestle-Aland 28th Edition*. Deutsche Bibelgesellschaft.

Eberhard Nestle. 1904. *Hē Kainē Diathēkē: text with critical apparatus*. British and Foreign Bible Society.

Oblubienica. Ewangeliczny Przekład Interlinearny Biblii. `https://biblia.oblubienica.eu/`. Accessed: 2024-10-04.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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 preprint*. ArXiv:2308.12008 [cs].

Maurice A Robinson and William G Pierpont. 2005. *The New Testament in the Original Greek: Byzantine Textform, 2005*. Chilton Book Publishing.

Frederick Henry Ambrose Scrivener. 1881. *The New Testament in the original Greek: according to the text followed in the Authorized version, together with the variations adopted in the Revised version*. The University Press.

M. Shuttleworth and M. Cowie. 2014. *Dictionary of translation studies*. St. Jerome Publishing.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. 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.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Preprint*, arXiv:1905.05950.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Brooke Foss Westcott and Fenton John Anthony Hort. 1882. *The New Testament in the original greek*, volume 1. Harper.

Michał Wojciechowski and Remigiusz Popowski. 1993. *Grecko-polski Nowy Testament: wydanie interlinearne z kodami gramatycznymi*. Oficyna Wydawnicza "Vocatio".

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *Preprint*, arXiv:2410.03115.

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 preprint*. 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.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639, Online. Association for Computational Linguistics.

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 preprint*. ArXiv:1911.02709 [cs].

A Morphological Tag Set Description

This appendix presents the morphological annotation scheme found in the tag sets of our scraped datasets.

Grammatical Categories in the Corpora
Part of Speech: Verb, Noun, Adverb, Adjective, Article, Pronoun, Preposition, Conjunction, Interjection, Particle, Aramaic Word, Hebrew Word
Pronoun Subtype: Personal / Possessive, Demonstrative, Interrogative / Indefinite, Reciprocal, Relative and Reflexive
Person: 1st, 2nd, 3rd
Tense: Present, Imperfect, Future, Aorist, Perfect, Pluperfect
Mood: Indicative, Imperative, Subjunctive, Optative, Infinitive, Participle
Voice: Active, Middle, Passive, Middle or Passive
Case: Nominative, Vocative, Accusative, Genitive, Dative
Number: Singular, Plural
Gender: Masculine, Feminine, Neuter
Degree: Positive, Comparative, Superlative

Table 8: Morphological annotation scheme: grammatical categories and their possible values in the Bible Hub and Oblubienica corpora.

B Complete Experimental Results

This appendix presents the complete experimental results across all model configurations, tag sets, and preprocessing approaches. For both English (EN) and Polish (PL) translations, we evaluate using BLEU and SemScore metrics. Each metric is evaluated across four base models: GreTa, PhilTa, mT5-base, and mT5-large. Bold values indicate the best performance for each configuration.

Encoding	Tag Set	Language / Base Model / Preprocessing	EN GreTa	EN PhilTa	EN mT5-base	EN mT5-large
baseline	Unused	Diacritics	17.69	41.55	31.61	**44.67**
		Normalized	16.77	33.24	29.99	**43.64**
t-w-t	BH	Diacritics	14.70	40.95	30.11	**46.00**
		Normalized	16.13	34.25	27.59	**43.97**
	OB	Diacritics	14.51	40.84	29.62	**45.59**
		Normalized	12.14	33.44	28.39	**35.47**
emb-concat	BH	Diacritics	3.58	**55.93**	1.33	50.47
		Normalized	4.05	**46.82**	27.32	0.70
	OB	Diacritics	5.48	**45.43**	42.59	41.18
		Normalized	3.93	40.76	0.69	**51.04**
emb-sum	BH	Diacritics	55.22	**60.10**	52.34	56.03
		Normalized	51.93	**56.24**	1.66	55.61
	OB	Diacritics	54.98	**59.75**	51.90	0.83
		Normalized	52.39	55.49	47.95	**56.24**
emb-auto	BH	Diacritics	54.18	**60.40**	28.52	56.51
		Normalized	53.17	**56.16**	47.84	55.12
	OB	Diacritics	54.98	**59.66**	52.37	55.81
		Normalized	53.15	**56.51**	52.43	55.37

Table 9: BLEU Scores for English translations.

Encoding	Tag Set	Language Base Model Preprocessing	EN GreTa	PhilTa	mT5-base	mT5-large
baseline	Unused	Diacritics	0.56	**0.83**	0.74	0.82
		Normalized	0.56	0.74	0.74	**0.82**
t-w-t	BH	Diacritics	0.55	0.82	0.74	**0.83**
		Normalized	0.56	0.76	0.72	**0.82**
	OB	Diacritics	0.55	0.82	0.74	**0.83**
		Normalized	0.53	0.76	0.73	**0.78**
emb-concat	BH	Diacritics	0.42	**0.87**	0.34	0.84
		Normalized	0.42	**0.82**	0.68	0.37
	OB	Diacritics	0.49	**0.83**	0.80	0.77
		Normalized	0.42	0.78	0.34	**0.85**
emb-sum	BH	Diacritics	0.86	**0.89**	0.86	0.88
		Normalized	0.84	0.87	0.38	**0.88**
	OB	Diacritics	0.85	**0.89**	0.86	0.34
		Normalized	0.85	0.86	0.84	**0.88**
emb-auto	BH	Diacritics	0.86	**0.89**	0.71	0.88
		Normalized	0.85	0.87	0.84	**0.87**
	OB	Diacritics	0.86	**0.89**	0.86	0.87
		Normalized	0.85	0.87	0.87	**0.87**

Table 10: SemScore for English translations.

Encoding	Tag Set	Language Base Model Preprocessing	PL GreTa	PhilTa	mT5-base	mT5-large
baseline	Unused	Diacritics	0.86	0.03	28.75	**42.92**
		Normalized	0.63	0.07	26.21	**41.05**
t-w-t	BH	Diacritics	0.49	0.04	21.45	**41.93**
		Normalized	0.56	0.08	**26.07**	0.17
	OB	Diacritics	0.74	0.08	27.72	**41.62**
		Normalized	0.78	0.05	0.24	**41.58**
emb-concat	BH	Diacritics	0.71	0.11	**0.79**	0.57
		Normalized	1.86	0.26	1.93	**54.54**
	OB	Diacritics	0.84	0.13	0.63	**55.55**
		Normalized	1.41	0.26	0.45	**51.75**
emb-sum	BH	Diacritics	50.89	6.18	52.54	**56.75**
		Normalized	48.47	1.71	50.43	**58.46**
	OB	Diacritics	51.21	0.12	54.41	**58.90**
		Normalized	32.92	5.39	0.66	**58.92**
emb-auto	BH	Diacritics	51.30	11.79	54.63	**59.04**
		Normalized	46.01	15.37	54.47	**57.42**
	OB	Diacritics	51.06	8.24	53.87	**58.44**
		Normalized	49.72	6.23	44.29	**59.33**

Table 11: BLEU Scores for Polish translations.

Encoding	Tag Set	Language Base Model Preprocessing	PL GreTa	PL PhilTa	PL mT5-base	PL mT5-large
baseline	Unused	Diacritics	0.53	0.18	0.85	**0.89**
		Normalized	0.49	0.42	0.85	**0.89**
t-w-t	BH	Diacritics	0.51	0.54	0.82	**0.89**
		Normalized	0.49	0.52	**0.84**	0.45
	OB	Diacritics	0.54	0.56	0.84	**0.89**
		Normalized	0.56	0.50	0.66	**0.89**
emb-concat	BH	Diacritics	0.59	0.58	0.67	**0.68**
		Normalized	0.62	0.58	0.68	**0.92**
	OB	Diacritics	0.62	0.53	0.63	**0.93**
		Normalized	0.60	0.58	0.67	**0.92**
emb-sum	BH	Diacritics	0.92	0.77	0.92	**0.93**
		Normalized	0.92	0.69	0.92	**0.93**
	OB	Diacritics	0.92	0.55	0.93	**0.93**
		Normalized	0.87	0.76	0.65	**0.94**
emb-auto	BH	Diacritics	0.92	0.80	0.92	**0.93**
		Normalized	0.91	0.82	0.93	**0.93**
	OB	Diacritics	0.92	0.79	0.92	**0.93**
		Normalized	0.92	0.77	0.90	**0.94**

Table 12: SemScore for Polish translations.

C Morphological Encoding Strategies

This appendix examines the impact of different encoding strategies: baseline, tags-within-text (t-w-t), embedding concatenation (emb-concat), embedding sum (emb-sum), and embedding autoencoder (emb-auto). We present aggregated BLEU and SemScore metrics for both English and Polish translations, along with statistical significance tests between strategy pairs. For each metric, we report both average and best scores across all configurations. Mann-Whitney U tests were used to assess the statistical significance of differences between encoding strategies.

Language	Metric	Encoding	baseline	t-w-t	emb-concat	emb-sum	emb-auto
EN	BLEU Score	Avg	32.40	30.86	26.33	48.04	**53.26**
		Best	44.67	46.00	55.93	60.10	**60.40**
	SemScore	Avg	0.73	0.72	0.63	0.80	**0.86**
		Best	0.83	0.83	0.87	0.89	**0.89**
PL	BLEU Score	Avg	17.57	12.73	10.74	36.75	**42.58**
		Best	42.92	41.93	55.55	58.92	**59.33**
	SemScore	Avg	0.64	0.66	0.68	0.85	**0.89**
		Best	0.89	0.89	0.93	**0.94**	0.94

Table 13: Performance comparison of encoding strategies: average and best scores across configurations.

	baseline	t-w-t	emb-concat	emb-sum	emb-auto
baseline	-	0.569	0.787	0.016*	0.002**
t-w-t	0.569	-	0.462	0.001**	0.000***
emb-concat	0.787	0.462	-	0.006**	0.000***
emb-sum	0.016*	0.001**	0.006**	-	0.396
emb-auto	0.002**	0.000***	0.000***	0.396	-

Table 14: Statistical significance matrix: BLEU scores for Polish translations.

	baseline	t-w-t	emb-concat	emb-sum	emb-auto
baseline	-	0.697	0.742	0.002**	0.000***
t-w-t	0.697	-	0.749	0.000***	0.000***
emb-concat	0.742	0.749	-	0.001***	0.000***
emb-sum	0.002**	0.000***	0.001***	-	0.585
emb-auto	0.000***	0.000***	0.000***	0.585	-

Table 15: Statistical significance matrix: BLEU scores for English translations.

	baseline	t-w-t	emb-concat	emb-sum	emb-auto
baseline	-	0.928	0.528	0.009**	0.002**
t-w-t	0.928	-	0.169	0.001***	0.000***
emb-concat	0.528	0.169	-	0.002**	0.000***
emb-sum	0.009**	0.001***	0.002**	-	0.418
emb-auto	0.002**	0.000***	0.000***	0.418	-

Table 16: Statistical significance matrix: semantic similarity for Polish translations.

	baseline	t-w-t	emb-concat	emb-sum	emb-auto
baseline	-	0.787	0.653	0.002**	0.000***
t-w-t	0.787	-	0.611	0.000***	0.000***
emb-concat	0.653	0.611	-	0.001**	0.000***
emb-sum	0.002**	0.000***	0.001**	-	0.534
emb-auto	0.000***	0.000***	0.000***	0.534	-

Table 17: Statistical significance matrix: semantic similarity for English translations.

D Tag Set Selection Impact

This appendix evaluates the impact of different morphological tag sets on model performance, comparing the one collected from BibleHub (BH), to the one from Oblubienica (OB), and approaches where no tags were used (Unused). We present aggregated BLEU and SemScore metrics for both English and Polish translations. For each metric, we report both average and best scores across all configurations. Mann-Whitney U tests were used to assess the statistical significance of differences between tag sets.

Language	Metric	Tag Set	BH	OB	Unused
EN	BLEU Score	Avg	38.90	**40.34**	32.40
		Best	**60.40**	59.75	44.67
	SemScore	Avg	0.74	**0.76**	0.73
		Best	**0.89**	0.89	0.83
PL	BLEU Score	Avg	**25.84**	25.55	17.57
		Best	59.04	**59.33**	42.92
	SemScore	Avg	0.77	**0.77**	0.64
		Best	0.93	**0.94**	0.89

Table 18: Performance comparison of morphological tag sets: BibleHub (BH), Oblubienica (OB), and baseline.

Metric Language	BLEU Score	SemScore
EN	0.96	0.97
PL	0.89	0.99

Table 19: Statistical significance of differences between tag sets (p-values).

E Text Preprocessing Impact

This appendix evaluates the impact of preprocessing choices on model performance, comparing diacritic-preserved and normalized (stripped of diacritics, lowercased) text approaches. We present aggregated BLEU and SemScore metrics for both English and Polish translations, with results broken down by tokenizer type (GreTa, PhilTa, mT5). For each metric, we report both average and best scores across all configurations. Mann-Whitney U tests were used to assess the statistical significance of differences between preprocessing approaches.

Language	Metric	Preprocessing	Diacritics	Normalized
EN	BLEU Score	Avg	**60.40**	56.51
		Best	**40.48**	37.16
	SemScore	Avg	**0.89**	0.88
		Best	**0.76**	0.74
PL	BLEU Score	Avg	59.04	**59.33**
		Best	**26.26**	23.33
	SemScore	Avg	0.93	**0.94**
		Best	**0.76**	0.75

Table 20: Aggregated BLEU and SemScore results across preprocessing approaches.

			GreTa		PhilTa		mT5	
			Diacritics	Normalized	Diacritics	Normalized	Diacritics	Normalized
EN	BLEU	Avg	30.59	29.30	**51.62**	45.88	39.86	36.72
		Best	55.22	53.17	**60.40**	56.51	56.51	56.24
	SemScore	Avg	0.67	0.65	**0.86**	0.82	0.76	0.74
		Best	0.86	0.85	**0.89**	0.87	0.88	0.88
PL	BLEU	Avg	23.12	20.26	2.97	3.27	**39.47**	34.89
		Best	51.30	49.72	11.79	15.37	59.04	**59.33**
	SemScore	Avg	0.72	0.71	0.59	0.63	**0.86**	0.83
		Best	0.92	0.92	0.80	0.82	0.93	**0.94**

Table 21: Impact of preprocessing on model performance: breakdown by tokenizer and preprocessing approach.

Language	Tokenizer Metric	GreTa	PhilTa	mT5
EN	BLEU Score	0.48	0.13	0.60
	SemScore	0.48	0.05	0.81
PL	BLEU Score	0.66	0.60	0.54
	SemScore	0.54	0.93	0.65

Table 22: Statistical significance of preprocessing impact across tokenizers (p-values).

F Base Model Performance Analysis

This appendix analyzes the performance differences between the four base models: GreTa, PhilTa, mT5-base, and mT5-large. We present aggregated BLEU and SemScore metrics for both English and Polish translations, along with statistical significance tests between model pairs. For each metric, we report both average and best scores across all configurations. Mann-Whitney U tests were used to assess the statistical significance of differences between model pairs.

Language	Metric	Base Model	GreTa	PhilTa	mT5-base	mT5-large
EN	BLEU Score	Avg	29.94	**48.75**	32.46	44.13
		Best	55.22	**60.40**	52.43	56.51
	SemScore	Avg	0.66	**0.84**	0.71	0.79
		Best	0.86	**0.89**	0.87	0.88
PL	BLEU Score	Avg	21.69	3.12	27.75	**46.61**
		Best	51.30	15.37	54.63	**59.33**
	SemScore	Avg	0.71	0.61	0.81	**0.88**
		Best	0.92	0.82	0.93	**0.94**

Table 23: Performance comparison of base models: average and best scores across all configurations.

Language	PL				EN			
Model	GreTa	PhilTa	mT5-base	mT5-large	GreTa	PhilTa	mT5-base	mT5-large
GreTa	-	0.003**	0.457	0.003**	-	0.005**	0.812	0.097
PhilTa	0.003**	-	0.000***	0.000***	0.005**	-	0.001**	0.457
mT5-base	0.457	0.000***	-	0.006**	0.812	0.001**	-	0.017*
mT5-large	0.003**	0.000***	0.006**	-	0.097	0.457	0.017*	-

Table 24: Statistical significance of BLEU score differences between base models (p-values).

Language	PL				EN			
Model	GreTa	PhilTa	mT5-base	mT5-large	GreTa	PhilTa	mT5-base	mT5-large
GreTa	-	0.110	0.038*	0.003**	-	0.005**	0.602	0.079
PhilTa	0.110	-	0.000***	0.000***	0.005**	-	0.002**	0.740
mT5-base	0.038*	0.000***	-	0.006**	0.602	0.002**	-	0.022*
mT5-large	0.003**	0.000***	0.006**	-	0.079	0.740	0.022*	-

Table 25: Statistical significance of SemScore differences between base models (p-values).