

# Neural Machine Translation for Coptic-French: Strategies for Low-Resource Ancient Languages

Nasma Chaoui, Richard Khoury

Université Laval, Department of Computer Science and Software Engineering  
Québec City, Canada  
nasma.chaoui.1@ulaval.ca, richard.khoury@ift.ulaval.ca

## Abstract

This paper presents the first systematic study of strategies for translating Coptic into French. Our comprehensive pipeline systematically evaluates: pivot versus direct translation, the impact of pre-training, the benefits of multi-version fine-tuning, and model robustness to noise. Utilizing aligned biblical corpora, we demonstrate that fine-tuning with a stylistically-varied and noise-aware training corpus significantly enhances translation quality. Our findings provide crucial practical insights for developing translation tools for historical languages in general.

**Keywords:** Machine Translation, Statistical and Machine Learning Methods, Low-resource languages

## 1. Introduction

Neural Machine Translations (NMT), with its steady growth, has profoundly impacted the field of computational linguistics. Transformer models and large-scale multilingual systems have enabled translation coverage across hundreds of language pairs. However, many historical languages — especially ancient ones with scarce data — have yet to benefit from these advancements. Coptic is one such language. As the last stage of the Egyptian language family, understanding and translating it is crucial for advancing fields such as Egyptology and Early-Christian studies. As a result, the language is studied in universities around the world; according to one listing<sup>1</sup>, over 50 universities in 13 countries offer Coptic courses. But despite this international interest, there are currently few NMT tools for Coptic, and only to translate to English or Arabic.

This paper aims to fill that gap by investigating a wide range of NMT strategies to translate Coptic into French. In the process, we will focus on four important questions: (1) should we translate directly to our target language or use English as a pivot language; (2) how should we pick a language model to fine-tune for this task; (3) should we provide the model with a diverse set of possible translations for a passage; and (4) how can we make the model robust to noise in the original manuscript. While our focus is solely on Coptic→French NMT, these questions are of general interest, and our results will help guide the development of NMT tools between any ancient-modern language pair.

The remainder of this paper is organized as follows: Section 2 reviews relevant work on natural language processing for Coptic. Section 3 details the methodology we use to explore our four ques-

tions. Section 4 presents our experimental setup: our dataset, NMT models and metrics, and Section 5 presents the results we obtain on the experiments about each of the four questions. Finally, Section 6 has some concluding remarks. Our dataset and code are available on our GitHub repository<sup>2</sup>, while the best Coptic→French NMT models we trained are distributed on HuggingFace<sup>3 4</sup>.

## 2. Background

Coptic occupies a unique position as the final stage of the Egyptian language family, bridging thousands of years from Old Egyptian through Demotic. It holds critical historical and cultural significance for Egyptology, biblical and early monasticism studies, as many Coptic manuscripts illuminate these fields.

In recent years, substantial progress has been made in building digital resources for Coptic. The Coptic SCRIPTORIUM project (Schroeder and Zeldes, 2016) stands out as a foundational effort, providing annotated corpora with multiple linguistic layers, including part-of-speech tagging, lemmatization, morphological analysis, and syntactic parsing (Zeldes and Schroeder, 2016). Similarly, tools such as the BabyLemmatizer have extended lemmatization and POS tagging capabilities across Coptic, Demotic, and Earlier Egyptian, demonstrating the shared linguistic complexities within the Egyptian language family (Sahala and Lincke, 2025).

Recent work has also explored NMT within the Egyptian language family, particularly focusing on

<sup>1</sup>Listing assembled by <https://www.cmcl.it/~iacs/course.htm>.

<sup>2</sup><https://github.com/chaouin/coptic-french-nmt>

<sup>3</sup><https://huggingface.co/chaouin/coptic-french-translation-helsinki>

<sup>4</sup><https://huggingface.co/chaouin/coptic-french-translation-hiero>

hieroglyphic texts. Notably, (Cao et al., 2024) developed a Hieroglyphic Transformer by adapting the multilingual M2M-100 model to translate ancient Egyptian hieroglyphs into both German and English. Their approach involved constructing a carefully curated dataset from the *Thesaurus Linguae Aegyptiae*, which includes hieroglyphic inscriptions paired with transliterations, POS tags, and translations. They relied on transliteration (romanization) to render hieroglyphic sequences compatible with modern NLP model vocabularies, and demonstrated that transfer learning from existing multilingual architectures could achieve promising results even on low-resource ancient scripts.

The authors of (Saeed et al., 2024) introduced the largest parallel corpus to date for Coptic, comprising approximately 24,000 Coptic-English sentence pairs and 8,000 Coptic-Arabic pairs, and used it to train the first NMT system from Coptic to Arabic and English. They used both basic Seq2Seq models and fine-tuned multilingual models such as mT5 and M2M-100. Their work emphasized the necessity of romanization due to Coptic characters being absent from standard vocabularies. Similarly, (Enis and Megalaa, 2024) implemented Coptic→English NMT by fine-tuning multilingual models on normalized and romanized texts from the Coptic SCRIPTORIUM. They employed back-translation and explored multiple data representations, demonstrating the critical impact of morphological normalization and tokenization for handling Coptic's agglutinative structure.

### 3. Methodology

While implementing our Coptic→French NMT model, we seek to study four fundamental design questions that are common to most if not all ancient-modern language translation projects. Thus, our methodology is divided into four stages, each building on the findings of the previous.

#### 3.1. Pertinence of Fine-Tuning

To begin, we seek to answer a fundamental question: is it even worthwhile to fine-tune a NMT model for Coptic→French translation? After all, while no such model exists, there are alternatives. The first one would be to use a more general multilingual NMT model that can work out of the box without any fine-tuning. Another alternative could use a NMT model that can translate the ancient language into another pivot language, and prompt it to answer in our target modern language. In our case, this will mean using a readily-available Coptic→English NMT model and prompting it to answer in French. Finally, since NMT models trained to translate between pairs of modern languages abound today,

we could again start from an NMT model that can translate the ancient language into a pivot language, and use a second NMT model to translate that content into our target modern language. In our case, we could daisy-chain a Coptic→English and an English→French NMT model to accomplish this. Consequently, the first stage of our research consists in comparing the translations produced by a fine-tuned Coptic→French NMT model and these three alternatives, to see the differences between these four options.

#### 3.2. Pre-Trained Model Choice

While there are not a lot of ancient→modern NMT models, there are a handful of them available, and therefore a question arises: how to pick which one to fine-tune? There are technical and linguistic considerations in this decision. From a technical standpoint, the different models available may have different numbers of parameters and be trained using different quantities of data (but given the scarcity of ancient-language data, they will typically not have different data sources). And linguistically, when no model exists for our ancient and modern language pair, three options are available. The first is to take a model trained to translate the ancient language to a different modern language and fine-tune it to our target language, or in our case fine-tune a Coptic-English model to translate to French. The second option is to take a model trained to translate into the modern language and fine-tune it to the ancient language. In our case this would mean taking any of the available multilingual NMT models trained for French and fine-tune it to translate from Coptic. Finally, a NMT translation model may exist for a different ancient language in the same family, giving us a third option. As mentioned previously, Coptic is the final stage of the Egyptian language, so a NMT model trained for an earlier stage of the language could be fine-tuned for Coptic.

#### 3.3. Fine-Tuning Dataset Choice

If we want to fine-tune an existing model, we need a corpus of parallel text in the ancient and modern languages. However, for a variety of reasons, an ancient-language text may be translated to a modern language in several different ways. We thus conduct an experiment to see whether or not fine-tuning can benefit from this stylistic diversity. Our hypothesis is that having multiple different translations of a same original text will make it possible to build a larger training corpus from the scarce ancient language resource, while also allowing the NMT model to better learn concepts of syntactic variation and word synonymy.

### 3.4. Training Robustness to Noise

Ancient-language texts often suffer from corruption due to the passage of time. In our final experiment, we consider the benefit of training a NMT model using noisy data as a way of building in robustness to noise. We consider two sources of noise. The first are lacunae that result from missing or damaged fragments of the text (Levine et al., 2024). We can introduce this problem in our training data by randomly replacing characters in the text with a “missing character” symbol. We acknowledge that this is a simplification of real manuscript damage, which typically affects entire lines or sections of a page rather than isolated characters; more realistic corruption models are a direction for future work. The second source of noise in our study is letter substitutions. This problem results from two different sources: on the one hand certain characters in a language can look similar and hand-writing can be ambiguous, and on the other optical character recognition (OCR) systems used to digitize manuscripts are prone to misrecognize characters, with one study finding that character recognition accuracy in ancient text OCR could be as low as 62% (Tzogka et al., 2021). We can simulate this problem by creating a list of similar-looking characters, and randomly replacing a character by a similar one from the list. A final source of noise is spelling variations and mistakes, which we simulate by randomly switching the order of two adjacent characters. By controlling the rate of these three types of random replacements, we can further experiment with the impact of different proportions of noise in the data.

## 4. Experimental Setup

### 4.1. Dataset and Preprocessing

To construct our parallel Coptic–French corpus, we first extracted Coptic texts from the *Coptic SCRIPTORIUM* project (Schroeder and Zeldes, 2016), which provides richly annotated PAULA XML files. We focused on the Sahidic New and Old Testament Bibles, the most complete and standardized Coptic corpus available. We automatically parsed the XML files to extract verse-level Coptic content along with their verse ID (the biblical book, chapter and verse number).

Since most pre-trained NMT models are not compatible with the original Coptic Unicode script, we romanized the Coptic data using the `uroman` tool<sup>5</sup>, an idea proposed by (Amrhein and Sennrich, 2020) and used for Coptic by (Enis and Megalaa, 2024). Although Coptic uses only a small number of characters beyond the Greek alphabet, romanization

<sup>5</sup><https://github.com/isi-nlp/uroman>

ensures full compatibility with the tokenizers and vocabularies of all four models used in our study, and is consistent with prior NMT work on Coptic and related ancient languages (Cao et al., 2024).

For the French text, we collected three public-domain Bible translations: the Louis Segond, Crampon, and Darby versions. These were chosen for their stylistic diversity, modern relevance, and free availability. Alignment with the Coptic text was performed automatically using verse ID.

Next, we performed a cleaning step to ensure a high-quality parallel corpus. Any verse pair missing either the Coptic or French translation was eliminated. We also removed structurally mismatched pairs by filtering out entries where the Coptic segment consisted only of ellipses (e.g., "[...]"), indicating incomplete or missing verses. Additionally, we discarded verses whose French translation was missing or blank, and stripped leading inline annotations (e.g., (1.2)) from the French verses to preserve only the canonical content. All removed entries were logged separately to ensure reproducibility. After cleaning, we retained 23,561 aligned verses for both the Segond and Darby Bibles, and 23,718 verses for the Crampon version, for a total of 70,840 aligned pairs from 63 biblical books.

We split this dataset into a training and testing corpus. We opted for a book-level split, keeping the books of *1 Corinthians*, *Mark*, *Galatians* and *Hebrews* as test data and the other 59 books as training data. This represents a total of 1,460 Coptic test verses, each with 3 French translations. This setup prevents data contamination, where similar verses from a same book could be found in the training and testing corpora. It also simulates a real-world use case, of a new never-before-seen Coptic document being discovered.

### 4.2. Pre-trained NMT Models

We selected four pre-trained NMT models for our study. They were the most related ones to our challenge of Coptic→French translation.

- **Enis**: The first-ever Coptic→English NMT model, by Enis and Megalaa (2024)<sup>6</sup>.
- **Helsinki**: A multilingual model from the University of Helsinki Tiedemann et al. (2023) trained on over 100 languages using a massive parallel Bible corpus, including Coptic (using a New Testament text different from the SCRIPTORIUM version) and French<sup>7</sup>.
- **T5**: A general-purpose encoder–decoder model trained by Google on the C4 corpus

<sup>6</sup><https://huggingface.co/megalaa/coptic-english-translator>

<sup>7</sup><https://huggingface.co/Helsinki-NLP/opus-mt-tc-bible-big-mul-mul>

(Raffel et al., 2020), with exposure to French. It does not include Coptic.<sup>8</sup>

- **Hiero**: A NMT model fine-tuned by (Cao et al., 2024) for Hieroglyphic→English and Hieroglyphic→German translation<sup>9</sup>.

Unless specified differently, all models were fine-tuned for 15 epochs on our training corpus.

### 4.3. Metrics

As our dataset clearly demonstrates, a single Coptic statement can have several different but equivalent French translations. Thus, a good model is one that correctly translates the intended meaning of the original statement, even if the wording differs from the gold-standard reference. To evaluate translation quality, we focus on metrics that measure semantic similarity rather than lexical overlap.

- **BERTScore** (Zhang et al., 2020) computes token-level similarity using contextual embeddings from a pre-trained BERT model. Scores range from 0 to 1, where higher values indicate greater semantic similarity between the candidate and reference sentences. It is well-suited to evaluate semantic equivalence, particularly when multiple valid versions of a statement exist.
- **BLEURT** (Sellam et al., 2020) is a learned evaluation metric that fine-tunes BERT-based models to predict human judgments of text quality. It combines semantic similarity, fluency, and grammaticality into a single score, and has been widely used to evaluate machine translation outputs (Pu et al., 2021). BLEURT scores are unbounded but typically center around [0, 1], with higher values indicating better quality.
- **COMET** (Rei et al., 2020) is a learned evaluation metric specifically designed for machine translation. It trains a multilingual BERT model to predict human judgment of translation quality. COMET scores are in the range of 0 to 1, with higher scores correlating with higher human appreciation of the translation.
- **METEOR** (Banerjee and Lavie, 2005) is a machine translation metric based on unigram matching, but that takes into account synonymy, stemming, and paraphrases, and that has been adapted to reward semantically-valid phrasings rather than exact lexical matches. METEOR scores are normalized between 0 and 1, with 1 indicating a perfect match.

<sup>8</sup><https://huggingface.co/t5-base>

<sup>9</sup><https://huggingface.co/mattiadc/hiero-transformer>

By considering these metrics together, we can obtain a more holistic evaluation of translation quality, one that accounts for meaning preservation and text quality and is tolerant of semantically-equivalent lexical changes.

## 5. Results and Analysis

### 5.1. Baseline

We begin by establishing a set of baselines using the models of section 4.2. To do this, each model is evaluated by translating our test corpus books in the most relevant language pair they are trained for: Coptic→English for Enis, Coptic→French for Helsinki, and English→French for T5. As no Biblical texts exist in hieroglyphic form to evaluate the Hiero model, we tested it using its own testing corpus sentences, which is composed of 50 Hieroglyph→English pairs and 50 Hieroglyph→German pairs.

Model	BERT Score	BLEURT	COMET	METEOR
Enis	0.632	0.344	0.328	0.119
Helsinki	0.595	0.136	0.248	0.082
T5	0.809	0.501	0.673	0.429
Hiero	0.833	0.555	0.664	0.496

Table 1: Baseline pre-trained model performances.

### 5.2. Pertinence of Fine-Tuning

Our first experiment compares the four different strategies to achieve Coptic→French translation described in Section 3.1. First is the strategy of fine-tuning an existing Coptic translation model, the Enis model, to a French target. The second strategy is to use a multilingual model that includes both Coptic and French out of the box, namely the Helsinki model. The third strategy is to prompt the Enis model to translate to French without fine-tuning, in a zero-shot manner. Finally, we will use English as a pivot language, by using Enis to translate from Coptic to English then translating that output to French using T5.

Approach	BERT Score	BLEURT	COMET	METEOR
Fine-tuning	<b>0.820</b>	<b>0.402</b>	<b>0.613</b>	<b>0.467</b>
Multilingual	0.595	0.136	0.248	0.082
Prompting	0.606	0.149	0.238	0.041
Pivot language	0.620	0.115	0.274	0.100

Table 2: Comparison of translation strategies.

The results in Table 2 clearly demonstrates that the benefits of fine-tuning justify the extra labor,

compared to the other three methods. The Enis model generates better translations in French after being fine-tuned than it did in English in our baseline results in Table 1, thanks to the extra training. Prompting Enis to translate to French without fine-tuning gives the worst results of the three Enis tests, due to mismatched decoding expectations. The model fine-tuned to a single pair of languages also outperforms the multilingual model, a result coherent with the literature, which finds that multilingual models perform suboptimally for low-resource languages (such as Coptic) (Joshi et al., 2025). Finally, it is interesting to see that the Enis-T5 pair achieves worse results on all metrics than either Enis or T5 did in the baseline test of Table 1. This indicates that translation errors in each step are compounded by the sequence of translations, reducing overall accuracy. This matches the observations of (Paul et al., 2013).

### 5.3. Pre-Trained Model Choice

Having confirmed that fine-tuning a model for Coptic→French is worth the effort, our second experiment explores the question of Section 3.2, namely which model should be fine-tuned. For this experiment, we fine-tune each model on our entire training corpus. This enables Enis to learn to translate to French, T5 to translate from Coptic, and Hiero to translate from Coptic to French. While Helsinki was already trained on Coptic and French among its 100 languages, it was also fine-tuned to become specialized to those two languages.

Model	BERT Score	BLEURT	COMET	METEOR
Enis	0.820	0.402	0.613	0.467
Helsinki	<b>0.850</b>	<b>0.564</b>	<b>0.731</b>	<b>0.563</b>
T5	0.763	0.240	0.474	0.325
Hiero	0.832	0.498	0.681	0.501

Table 3: Comparison of fine-tuned models.

The results translating our test corpus, presented in Table 3, show that Helsinki outperforms the other models on all four metrics. This performance can be attributed to two key advantages. First, its pre-training corpus includes both the source and target languages, unlike the other models that lacked one or both of the languages and had to learn them from scratch. Second, it’s the model whose pre-training corpus is most aligned with our task, being also a Bible corpus. The Helsinki model is also relatively large, with 248M parameters, which may contribute to its strong performance. However, we cannot exclude the possibility that it is becoming over-fitted to translating biblical texts; since we do not have non-biblical Coptic-French translations to

test on, we cannot check that possibility.

The next best model on all metrics is Hiero. Interestingly, this one has the opposite properties from Helsinki: it is trained on neither Coptic, French, nor biblical text. It is however trained for Hieroglyphics, a writing system related to Coptic. This seems to indicate that knowledge learned on a language transfers well to related languages. A similar observation was made by (Babych et al., 2007) in the case of NMT using pivot languages. The Enis model is a close third place. While that model benefits from having been pre-trained on Coptic, it is smaller than the Hiero model, with 77.1M against 484M parameters for Hiero, and thus doesn’t learn the new language as well. Finally, the T5 model, with 223M parameters, which does not have any Coptic nor related language in its pre-training data, achieves the worst fine-tuning results. It should be noted that given the significant differences in model size, not all models may have fully converged after 15 epochs; a systematic study of training hyperparameters such as the number of epochs and early stopping criteria is left for future work.

### 5.4. Fine-Tuning Dataset Choice

To test our hypothesis of Section 3.3, we fine-tune the two best models of Section 5.3, Helsinki and Hiero, on each of our three French translations individually. For this experiment, since each single-source model trains on roughly one-third of the complete training dataset, we train them for 45 epochs instead of 15 for the model that trains on the entire dataset. We tested each model on each translation of the test corpus separately, to observe the impact each different training has.

Test set	Training set	BERT Score	BLEURT	COMET	METEOR
Crampon	All datasets	0.836	<b>0.549</b>	<b>0.723</b>	0.513
	Crampon only	<b>0.849</b>	0.537	0.715	<b>0.583</b>
	Darby only	0.821	0.494	0.690	0.460
	Segond only	0.825	0.519	0.703	0.472
Darby	All datasets	0.858	<b>0.576</b>	<b>0.737</b>	0.587
	Crampon only	0.825	0.511	0.696	0.493
	Darby only	<b>0.868</b>	0.558	0.729	<b>0.615</b>
	Segond only	0.846	0.535	0.710	0.546
Segond	All datasets	0.856	<b>0.569</b>	<b>0.732</b>	0.587
	Crampon only	0.828	0.517	0.699	0.514
	Darby only	0.844	0.515	0.701	0.550
	Segond only	<b>0.865</b>	0.564	0.727	<b>0.611</b>

Table 4: Comparison of dataset choice on Helsinki.

The results, given in Tables 4 and 5, show that the models trained on a specific dataset translation perform best when tested on that same translation in two of our four metrics, BERTScore and METEOR. These are the two metrics that value seman-

Test set	Training set	BERT Score	BLEURT	COMET	METEOR
Crampon	All datasets	0.815	<b>0.480</b>	<b>0.672</b>	0.445
Crampon	Crampon only	<b>0.825</b>	0.458	0.655	<b>0.493</b>
Crampon	Darby only	0.805	0.433	0.642	0.410
Crampon	Segond only	0.806	0.448	0.646	0.413
Darby	All datasets	0.844	<b>0.517</b>	<b>0.692</b>	0.535
Darby	Crampon only	0.812	0.450	0.648	0.444
Darby	Darby only	<b>0.847</b>	0.493	0.678	<b>0.547</b>
Darby	Segond only	0.830	0.472	0.660	0.489
Segond	All datasets	0.837	<b>0.495</b>	<b>0.680</b>	0.519
Segond	Crampon only	0.810	0.442	0.642	0.443
Segond	Darby only	0.826	0.451	0.648	0.487
Segond	Segond only	<b>0.839</b>	0.479	0.663	<b>0.521</b>

Table 5: Comparison of dataset choice on Hiero.

tic similarity to the correct translation. By training and testing on the same translation, these models learn the correct choice of wording and translation decisions, and thus can generate a translation that is very similar to the expected one. By contrast, when training on one translation and testing on another, these systems all yield much weaker BERTScore and METEOR performances, showing they are generating translations worded differently from the expected result.

On the other hand, the model trained on all three dataset translations always achieves the top BLEURT and COMET scores, no matter which dataset translation it is tested on. These two metrics are tuned to replicate human appreciation of the generated text’s quality. By training on a variety of possible translations for each Coptic sentence, that model has learned to generate better translations overall. In addition, this model always achieves the second-best BERTScore and METEOR scores, behind the model trained on the same test corpus translation but ahead of the two trained on other translations. So while the wording generated by this model is not optimal compared to the correct translation, it is always closer than any other model not specialized to generate that specific translation. Overall, this shows that stylistic diversity in the training data improves generation quality at inference time, especially when multiple valid translations are acceptable.

## 5.5. Training Robustness to Noise

As explained in Section 3.4, we simulate real-world degradation of the manuscripts by injecting noise in our training and testing data. To simulate textual corruption, each character in a Coptic verse is subjected to three independent noise processes: a 2% probability of deletion (simulating lacunae), a 2% probability of switch with an adjacent character (simulating spelling errors), and a 10% probabil-

ity of replacement by a visually similar character (simulating OCR errors), using the visual confusion mapping of Figure 1, which we constructed manually based on the visual similarity of Coptic character shapes. These rates were chosen heuristically to introduce meaningful variation while preserving overall text readability, with the higher rate for visual substitution reflecting the greater prevalence of OCR errors compared to physical lacunae in digitized manuscripts (Tzogka et al., 2021). We do not alter the French translations of the verses. We experiment with different levels of simulated degradation by creating 5 versions of our corpora, setting the probability that a verse is corrupted to 0%, 10%, 30%, 50%, and 100%.

$\alpha \rightarrow o, \epsilon$      $o \rightarrow \alpha, \epsilon$      $\epsilon \rightarrow \alpha, o$      $i \rightarrow h, \gamma$   
 $h \rightarrow i, \epsilon$      $\gamma \rightarrow i, n$      $c \rightarrow i, h$      $\varpi \rightarrow q, x$   
 $q \rightarrow \varpi, x$      $x \rightarrow \varpi, q$      $\dagger \rightarrow \tau, \theta$      $\tau \rightarrow \dagger, \theta$   
 $\theta \rightarrow \tau, \dagger$      $n \rightarrow \gamma$      $m \rightarrow n$      $p \rightarrow \gamma, m$

Figure 1: Visual confusion of Coptic characters.

Test noise	Train noise	BERT Score	BLEURT	COMET	METEOR
0%	0%	0.850	0.564	0.731	0.563
0%	10%	<b>0.852</b>	<b>0.568</b>	<b>0.733</b>	<b>0.564</b>
0%	30%	0.849	0.560	0.727	0.560
0%	50%	0.848	0.553	0.723	0.559
0%	100%	0.846	0.547	0.719	0.546
10%	0%	0.844	0.545	0.717	0.549
10%	10%	<b>0.849</b>	<b>0.557</b>	<b>0.725</b>	<b>0.557</b>
10%	30%	0.847	0.553	0.722	0.555
10%	50%	0.846	0.545	0.718	0.554
10%	100%	0.844	0.544	0.717	0.543
30%	0%	0.829	0.492	0.679	0.511
30%	10%	0.842	0.534	0.708	0.538
30%	30%	0.842	<b>0.539</b>	0.711	0.545
30%	50%	<b>0.843</b>	0.536	<b>0.712</b>	<b>0.546</b>
30%	100%	0.841	0.533	0.709	0.536
50%	0%	0.817	0.457	0.650	0.478
50%	10%	0.837	0.514	0.695	0.522
50%	30%	<b>0.838</b>	0.522	0.701	<b>0.533</b>
50%	50%	<b>0.838</b>	0.519	0.698	<b>0.533</b>
50%	100%	<b>0.838</b>	<b>0.524</b>	<b>0.702</b>	0.529
100%	0%	0.788	0.366	0.579	0.411
100%	10%	0.822	0.471	0.665	0.488
100%	30%	0.826	0.491	0.675	0.502
100%	50%	0.829	0.496	0.683	0.509
100%	100%	<b>0.831</b>	<b>0.505</b>	<b>0.688</b>	<b>0.513</b>

Table 6: Comparison of robustness to noise for Helsinki.

The results in Tables 6 and 7 show several clear trends. First, both models trained with any level of noise perform better on clean test data, and their scores drop steadily as noise increases. This shows that texts with errors or gaps are naturally harder to translate than clean ones. Likewise, both the Helsinki and Hiero models trained with low noise do better on low-noise data, while those

Test noise	Train noise	BERT Score	BLEURT	COMET	METEOR
0%	0%	<b>0.832</b>	0.497	<b>0.682</b>	<b>0.500</b>
0%	10%	<b>0.832</b>	0.496	0.679	0.496
0%	30%	0.831	<b>0.498</b>	0.679	0.494
0%	50%	0.830	0.492	0.674	0.492
0%	100%	0.827	0.482	0.668	0.482
10%	0%	0.825	0.474	0.665	0.485
10%	10%	0.828	0.484	0.670	0.488
10%	30%	<b>0.829</b>	<b>0.489</b>	<b>0.674</b>	<b>0.489</b>
10%	50%	<b>0.829</b>	0.485	0.671	0.488
10%	100%	0.826	0.478	0.665	0.480
30%	0%	0.809	0.412	0.619	0.445
30%	10%	0.822	0.452	0.649	0.470
30%	30%	<b>0.825</b>	0.465	<b>0.658</b>	<b>0.476</b>
30%	50%	0.824	<b>0.466</b>	0.656	<b>0.476</b>
30%	100%	0.824	<b>0.466</b>	0.657	0.473
50%	0%	0.795	0.364	0.582	0.416
50%	10%	0.815	0.426	0.630	0.453
50%	30%	0.819	0.447	0.643	0.463
50%	50%	<b>0.821</b>	0.452	0.647	<b>0.466</b>
50%	100%	<b>0.821</b>	<b>0.456</b>	<b>0.650</b>	<b>0.466</b>
100%	0%	0.762	0.232	0.486	0.335
100%	10%	0.800	0.366	0.592	0.411
100%	30%	0.810	0.405	0.616	0.436
100%	50%	0.813	0.419	0.627	0.445
100%	100%	<b>0.815</b>	<b>0.426</b>	<b>0.629</b>	<b>0.451</b>

Table 7: Comparison of robustness to noise for Hiero.

trained with more noise perform better on noisy data. For Helsinki, the model trained with 10% noise leads early on, then the 50% model takes over in mid-noise settings, and the 100% model does best at the end. For Hiero, the 0% model performs best early, the 30% model leads in the middle, and the 100% model dominates in the last tests.

What’s really happening is that models trained with less noise do better on clean data, but they’re also more fragile: their scores drop sharply as noise rises. To quantify this fragility, Table 8 shows, for each model and training noise level, the relative drop in score between the *cleanest* test condition (0% test noise) and the *noisiest* one (100% test noise) for each metric. Each row corresponds to one model under one training noise setting, and the columns report the relative score drops for BERTScore, BLEURT, COMET, and METEOR. Models trained with 0% and 10% noise lose around 20% on average—Hiero’s BLEURT score even drops by 53%. In contrast, models trained with 50% or 100% noise lose less than 15%, and as little as 6% on average. This confirms that controlled noise injection creates NMT models more resilient to real-world degradation, such as OCR errors or fragmentary manuscripts, at the cost of lower performance on clean manuscripts. While the optimal choice thus depends on the characteristics of the manuscript being translated, a good general-case compromise value seems to be 50% training noise.

Model	Training noise	BERT Score	BLEURT	COMET	METEOR
Helsinki	0%	7.3%	35.1%	20.8%	27.0%
Helsinki	10%	3.5%	17.1%	9.3%	13.5%
Helsinki	30%	2.7%	12.3%	7.2%	10.4%
Helsinki	50%	2.2%	10.3%	5.5%	8.9%
Helsinki	100%	1.8%	7.7%	4.3%	6.0%
Hiero	0%	8.4%	53.3%	28.7%	33.0%
Hiero	10%	3.8%	26.2%	12.8%	17.1%
Hiero	30%	2.5%	18.7%	9.3%	11.7%
Hiero	50%	2.0%	14.8%	7.0%	9.6%
Hiero	100%	2.0%	11.6%	5.8%	6.4%

Table 8: Relative drop in score between 0% and 100% *test* noise, per model and training noise level. Lower values indicate greater robustness to noisy input.

## 6. Conclusion

This paper presents the first systematic study of neural machine translation for the Coptic→French language pair. Through four experiments, we outlined best practices for building a translation model for this new pair. First, we showed that task-specific fine-tuning clearly improves translation quality compared to using pre-trained multilingual models or translating through a pivot language. Next, we found that fine-tuning a model already trained on the ancient language or a related one in the same family (such as Hieroglyphics) yields better results. Third, we examined the effect of having multiple translations of the same text, and found that including them all enhanced translation quality. Finally, we tested the impact of noise in the dataset and found that the best trade-off between translation quality and robustness was achieved when training with 50% corrupted verses. This led to the release of two strong NMT models for this task, one based on the multilingual model of (Tiedemann et al., 2023) and the other on the Hieroglyphic translator of (Cao et al., 2024), both fine-tuned on three French versions of the training corpus with 50% noise.

An important direction for future work is the evaluation of our models on non-biblical Coptic texts. Our study is necessarily limited to biblical material, as no parallel Coptic→French corpus exists outside this domain. Evaluating generalization to other genres, could be achieved through backtranslation and constitutes a key objective for future research.

Together, our findings extend beyond Coptic→French translation, providing practical guidance for building robust NMT systems in other low-resource or ancient language settings.

## 7. Acknowledgements

This research was made possible by the support of the Natural Sciences and Engineering Research

Council of Canada (NSERC) USRA and Discovery Grant programs, and the Fonds de recherche du Québec (FRQ).

## 8. Bibliographical References

- Chantal Amrhein and Rico Sennrich. 2020. On romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469. Association for Computational Linguistics.
- Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: Comparing direct transfer against pivot translation. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.
- Satanjeev Banerjee and Alon Lavie. 2005. ME-TOR: An automatic metric for MT evaluation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures*, pages 65–72.
- Mattia Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: A hieroglyphic transformer for translating ancient egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages*, pages 71–86.
- Maxim Enis and Andrew Megalaa. 2024. Ancient voices, modern technology: Low-resource neural machine translation for coptic texts. *Coptic Translator*, pages 1–15.
- Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjana Wartikar, and Eileen Long. 2025. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus. In *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*, pages 50–57. Association for Computational Linguistics.
- Lauren Levine, Cindy Li, Lydia Bremer-McCollum, Nicholas Wagner, and Amir Zeldes. 2024. Lacuna language learning: Leveraging RNNs for ranked text completion in digitized coptic manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages*, pages 61–70. Association for Computational Linguistics.
- Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to choose the best pivot language for automatic translation of low-resource languages. *ACM Transactions on Asian Language Information Processing*, 12(4):1–17.
- Amy Pu, Hyung Won Chung, Ankur Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for machine translation. In *Proceedings of EMNLP*.
- 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 EMNLP*, pages 2685–2702. Association for Computational Linguistics.
- Muhammed Saeed, Asim Mohamed, Mukhtar Mohamed, Shady Shehata, and Muhammad Abdul-Mageed. 2024. From Nile sands to digital hands: Machine translation of coptic texts. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 298–308.
- Alexi Sahala and Eliese-Sophia Lincke. 2025. Neural models for lemmatization and POS-tagging of earlier and late egyptian. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 99–104. Association for Computational Linguistics.
- Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of ACL*. Association for Computational Linguistics.
- Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58:713–755.
- Christina Tzogka, Fotini Koidaki, Stavros Doropoulos, Ioannis Papastergiou, Efthymios Agrafiotis, Katerina Tiktopoulou, and Stavros Vologianidis. 2021. OCR workflow: Facing printed texts of ancient, medieval and modern greek literature. *Journal of Documentation*.
- Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for coptic. In *Proceedings of the*

*10th SIGHUM Workshop on Language Technology for Cultural Heritage*, pages 146–155.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.