LogogramNLP: Comparing Visual and Textual Representations of Ancient Logographic Writing Systems for NLP

Danlu Chen[1], **Freda Shi**[2], **Aditi Agarwal**[1], **Jacobo Myerston**[1], **Taylor Berg-Kirkpatrick**[1]
UC San Diego[1], University of Waterloo[2]
danlu@ucsd.edu

Abstract

Standard natural language processing (NLP) pipelines operate on symbolic representations of language, which typically consist of sequences of discrete tokens. However, creating an analogous representation for ancient logographic writing systems is an extremely labor-intensive process that requires expert knowledge. At present, a large portion of logographic data persists in a purely visual form due to the absence of transcription—this issue poses a bottleneck for researchers seeking to apply NLP toolkits to study ancient logographic languages: most of the relevant data are *images of writing*. This paper investigates whether direct processing of visual representations of language offers a potential solution. We introduce **LogogramNLP**, the first benchmark enabling NLP analysis of ancient logographic languages, featuring both transcribed and visual datasets for four writing systems along with annotations for tasks like classification, translation, and parsing. Our experiments compare systems that employ recent visual and text encoding strategies as backbones. The results demonstrate that visual representations outperform textual representations for some investigated tasks, suggesting that visual processing pipelines may unlock a large amount of cultural heritage data of logographic languages for NLP-based analyses. Data and code are available at https://logogramNLP.github.io/.

1 Introduction

The application of computational techniques to the study of ancient language artifacts has yielded exciting results that would have been difficult to uncover with manual analysis alone (Assael et al., 2022). Unsurprisingly, one of the biggest challenges in this domain is data scarcity, which, in turn, means that transferring from pre-trained systems on well-resourced languages is paramount. However, it is more challenging to adopt similar techniques for ancient logographic writing systems,

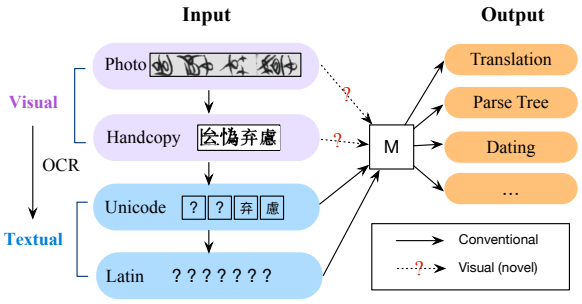

Figure 1: Illustration of the processing flow of Old Chinese (in Bamboo Script), an ancient logographic language, best viewed in color. M denotes the pre-trained model used in the pipeline. Vision-based models directly process visual representations (violet; dashed lines). Conventional NLP pipelines (blue; solid lines) first convert visual representations into symbolic text—either automatically, which is quite noisy, or manually, which is labor-intensive. However, as shown, some ancient logographic writing systems have symbol inventories that have not yet been fully mapped into Unicode. Even when Unicode codepoints exist, they are often mutually exclusive with the symbol inventories of high-resource languages, reducing the effectiveness of transferring from pre-trained models. Finally, *latinization* (a potential solution for finding common ground with pre-training languages) loses information from the original input, is not fully standardized, and is difficult to automate.

in which individual symbols represent entire semantic units like morphemes or words.

The challenges associated with NLP for ancient logographic languages mainly come from two aspects. First, for many ancient languages, most available data sources are in visual forms, consisting of untranscribed photographs or hand-drawn copies (i.e., *lineart*). Adopting the conventional NLP pipeline, which requires converting visual representations into symbolic text, is therefore not straightforward: automatic transcriptions are often noisy due to data scarcity, while manual transcriptions are labor-intensive and require domain expertise. Some logographic writing systems, such as

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14238–14254
August 11-16, 2024 ©2024 Association for Computational Linguistics

Old Chinese, even include symbol inventories that remain not fully mapped to Unicode (depicted in Figure 1).

Second, even when perfect Unicode transcriptions are available, their symbol inventories are often mutually exclusive with those of high-resource languages, which can substantially reduce the effectiveness of transfer from pre-trained multilingual encoders, such as mBERT (Devlin et al., 2018). One processing step that might be used to mitigate this issue is *latinization* of the Unicode transcripts (Rust et al., 2021; Muller et al., 2020). However, it is challenging to Latinize logographic languages due to uncertain pronunciations (Sproat and Gutkin, 2021) and the resulting inconsistent latinization schemes across artifacts from the same language and writing system. Such a process is laborious—humanists may devote months or even years to determine the correct transliteration. In contrast, once a correct transliteration is determined, translation into another language may only take minutes.

Fortunately, advances in visual encoding strategies for NLP tasks offer an alternative solution. Recent studies have investigated NLP systems that model text in the pixel space (Rust et al., 2023; Tschannen et al., 2023; Salesky et al., 2023), thereby opening new possibilities for the direct use of visual representations of ancient logographic writing systems. These approaches, to date, have primarily been applied to digitally rendered texts. They have not yet been extensively evaluated on handwritten texts, such as *lineart*, i.e., neatly hand-copied versions of texts by scholars.

In this paper, we attempt to answer the following questions: *(1) Can we effectively apply NLP toolkits, such as classifiers, machine translation systems, and syntactic parsers, to visual representations of logographic writing systems? (2) Does this strategy allow for better transfer from pre-trained models and lead to better performance?* Additionally, as shown in Figure 1, many logographic languages have multiple partially processed representations, including artifact photographs, hand-copied lineart, Unicode, Latin transliteration, and normalization—we also aim to empirically investigate the extent to which various representations at each stage, including textual and visual modalities, facilitate effective fine-tuning of downstream NLP systems.

We have curated **LogogramNLP**, a benchmark consisting of four representative ancient logographic writing systems (Linear A, Egyptian hieroglyphic, Cuneiform, and Bamboo Script), along with annotations for fine-tuning and evaluating downstream NLP systems on three tasks, including three attribute classification tasks, machine translation, and dependency parsing.

We conduct experiments on these languages and tasks with a suite of popular textual and visual encoding strategies. Surprisingly, visual representations perform better than conventional text representations for some tasks (including machine translation), likely due to visual encoding allowing for better transfer from cross-lingual pre-training. These results highlight the potential of visual representation processing, a novel approach to ancient language processing, which can be directly applied to a larger portion of existing data.

2 Dataset: Languages, Tasks and Challenges

Our benchmark consists of four representative ancient languages—Linear A, Egyptian hieroglyphic, Cuneiform, and Bamboo script (§2.1).[1] Each language is associated with a unique writing system and unique challenges. We refer the readers to Appendix A for data collection and cleaning details. Our benchmark covers three tasks: machine translation, dependency parsing, and attribute classification (§2.2).

2.1 Logographic Languages

A major characteristic of logographic languages is that the size of symbol inventories is significantly larger than that in alphabetic languages such as Ancient Greek (24 letters) or Modern English (26 letters). A summary of different representations of the languages of our interest is shown in Figure 2, and Table 2 summarizes the current status of each language.

Linear A. Linear A is an undeciphered language used by the Minoan at Crete and is believed to be not related to ancient Greek. Scholars have differentiated the glyphs and carefully hand-copied them into linearts. We collected a dataset of 772 tablets (i.e., manually drawing) from SigLA.[2] Each tablet also has a separable glyph with annotated Unicode.

[1] Bamboo scripts usually combine Seal scripts and Clerical scripts.

[2] https://sigla.phis.me/browse.html

Writing system	Language	abbr.	Visual Feature		Textual Feature		Task		
			Full Doc	Textline	Unicode	latinization	Translation	UD Parsing	Attribute
Linear A	Unknown	LNA	Y	**Y**		**Y**			**Y**
Egyptian hieroglyph	Ancient Egyptian	EGY		**Y**		**Y**	**Y**		**Y**
Cuneiform	Akkadian & Sumerian	AKK	Y		Y	Y	Y	Y*	Y
Bamboo script	Ancient Chinese	ZHO		**Y**	**Y**	**Y**	**Y***		**Y***

Table 1: A summary of the task availability across four ancient languages with unique writing systems. The **underlined** Y indicates that the data has not previously been used in a machine learning setup, which demonstrates the novelty of our benchmark; and asterisks (*) indicate that we conducted extra manual labeling.

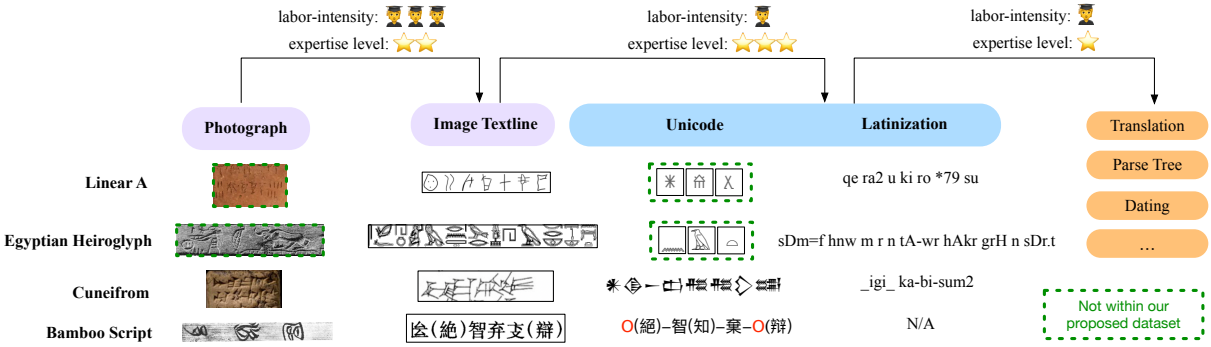

Figure 2: Example of four logographic languages with different representation formats. The arrow shows the typical processing flow of ancient languages by humanists. The workload and expertise required to transcribe the text from images is even greater than that of downstream tasks such as machine translation. The red circle O (in Bamboo Script) indicates the character is not digitized as Unicode yet. Green dashed boxes note that Unicode exists for Egyptian hieroglyphics and Linear A, but the alignment to documents is unavailable; the same goes for Egyptian and Linear A photographs.

status	LNA	AKK	EGY	ZHO	GRC
deciphered	None	Most	Most	Most	All
differentiated	Most	Most	Most	Most	All
encoded	Most	Most	Some	Some	All
Latinized	All	All	All	None	All

Table 2: Summary of the status of the ancient logographic languages presented in our paper. The status is measured from the perspective of paleography. We put Ancient Greek (GRC), a well-known ancient non-logographic language, here for comparison.

Akkadian (Cuneiform). CuneiML (Chen et al., 2023) is a dataset that contains 36k entries of cuneiform tablets. Each tablet consists of Unicode Cuneiform, lineart, and transliteration. We also use the Akkadian Universal Dependencies (UD) dataset (Luukko et al., 2020), which contains 1,845 sentences with dependency annotations. Since the UD annotation of Akkadian only keeps the normalization form of the language, we obtain the Unicode by running a dynamic programming-based matching algorithm.

Ancient Egyptian (Hieroglyph). We segmented the St Andrews Corpus (Nederhof and Berti, 2015)[3] using a rule-based segmenter, and obtained 891 examples of parallel data. Additionally, we collected data from the Thot Sign List (TSL; English translation)[4] and BBAW (German translation)[5] for 2,337 and 100,736 samples of parallel data, respectively. However, the transliteration standards differ among these three sources of data, and BBAW does not include hieroglyph image features. Therefore, we only used TSL's data.

Old Chinese (Bamboo script). We collected 13,770 pieces of bamboo slips from Kaom,[6] which come with the photograph of each line of the text. The Baoshan collection covers three genres: Wenshu (Document), Zhanbu (Divine), and Book. The Guodian collection contains parallel data translated into modern Chinese. The vocabulary size is 1,303. Notably, about 40% of the characters do not have a Unicode codepoint and are, therefore, represented as place-holder triangles or circles. This dataset

[3] https://mjn.host.cs.st-andrews.ac.uk/egyptian/texts/corpus/pdf/
[4] https://thotsignlist.org/
[5] https://aaew.bbaw.de/tla/servlet/TlaLogin
[6] http://www.kaom.net

does not come with human-labeled latinization due to the lack of transliteration standards.

2.1.1 Visual Representations

Since ancient scripts did not consistently adhere to a left-to-right writing order, breaking down multi-line documents into images of single-line text is nontrivial. These historical data, therefore, need additional processing to be machine-readable. Figure 3 shows examples of different processing strategies. We summarize the approaches we used in building the dataset as follows:

1. **Raw image (no processing)**: the raw images are already manually labeled and cut into text lines of images, and no extra processing is required.
2. **Montage**: we generate a row of thumbnails of each glyph using the montage tool in ImageMagick.[7] This strategy is used for Linear A, as the original texts are written on a stone tablet, and scholars have not determined the reading ordering of this unknown script.
3. **Digital rendering**: we digitally render the text using computer fonts when the language is already encoded in Unicode. Given that most ancient logographic scripts are still undergoing the digitization process, this option is currently unavailable except for Cuneiform.

2.1.2 Textual Representations

The processing of textual features for ancient logographic scripts also requires special attention. Unlike modern languages, ancient logographic writing systems can have multiple latinization standards or lack universally agreed-upon transcription standards. For example, the cuneiform parsing data is not in standard transliteration (ATF)[8] form, but rather, in the UD normalized form. This mismatch introduces extra difficulty to downstream tasks, especially in low-resource settings.

A similar issue also exists for Old Chinese: most ancient characters do not even exist in the current Unicode alphabet. While we may find some modern Chinese characters that look similar to the ancient glyphs, they are usually not identical, and such a representation loses information from the original text.

For Egyptian hieroglyphs, most characters are encoded in Unicode, but there is no standard encoding for "stacking" multiple glyphs vertically (Figure 3). Therefore, we do not include the Unicode text for our ancient Egyptian data as they are not available.

2.2 Tasks

Our benchmark covers three tasks (Table 1): translation, dependency parsing, and attribute classification. The model performance on these tasks reflects various aspects of ancient language understanding. To better understand the information loss when using a pipeline approach, we also report performance using this method: predicting the transliteration first and using the noisy predicted transliteration for downstream tasks.

Machine translation. The task is to translate the ancient languages, represented by either text or images, into modern languages, such as English. In all of our experiments, we translate ancient languages into English.

Dependency parsing. Given a sentence in the ancient language, the task is to predict the dependency parse tree (Tesnière, 1959) of the sentence. In the dependency parse tree, the parent of each word is its grammatical head.

Attribute classification. The task is to predict an attribute of the given artifact, for example, provenience (found place), time period, or genre.

3 Methods

In this section, we will describe feature encoding methods (§3.1) for both visual and textual inputs, as well as task-specific layers (§3.2) for each task we consider.

3.1 Feature Encoding

NLP for Low-resource languages has benefitted a lot from pre-trained models. However, modern pre-trained models do not cover the character inventories of the considered ancient logographic languages. To overcome this shortage, we summarize solutions to the problem into four categories and describe them as follows.

Extending vocabulary. In this line of approach (Wang et al., 2020; Imamura and Sumita, 2022), the vocabulary is extended by adding the unseen tokens. The embeddings of new tokens can be either initialized randomly or calculated by a function.

[7] https://imagemagick.org

[8] ATF is a format used to represent cuneiform text. More details can be found at http://oracc.ub.uni-muenchen.de/doc/help/

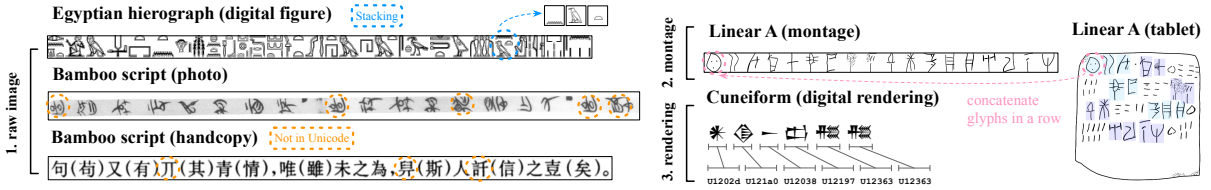

Figure 3: Image features of four ancient writing systems. (1) Egyptian hieroglyphs and Bamboo scripts are already manually segmented into images of lines. In the handcopy version of the Bamboo script, the word within parentheses indicates the corresponding modern Chinese glyph. Although both the Egyptian and Bamboo script images appear to be in a digital font, they are only accessible as images without underlying codepoint mappings to Unicode. (2) Linear A tablets are believed to be written in horizontal lines running from left to right (Salgarella, 2020); therefore, we use the montage concatenation of each glyph as the representation. (3) We digitally render Cuneiform Unicode using computer font as the visual representation.

In the fine-tuning stage, the embeddings of new tokens are updated together with the rest of the model.

Latin transliteration as a proxy. The majority of past work on cross-lingual transfer has focused on using Latin transliteration as the proxy to transfer knowledge from high-resource to low-resource languages (Pires et al., 2019; Fang et al., 2020). Following this line of work, we input latinization representations to mBERT (Devlin et al., 2018) to obtain the embeddings of the ancient languages.

Tokenization-free. The idea of the tokenization-free approach is to view tokens as a sequence of bytes and directly operate on UTF-8 codepoints without an extra mapping step. As representative models, ByT5 (Xue et al., 2022) and CANINE (Clark et al., 2022) use Unicode encoding of a string to resolve the cross-lingual out-of-vocabulary issues. This work uses ByT5 for machine translation and CANINE for classification.

Pixel Encoder for Text. Recently, there has been a novel approach (Rust et al., 2023) that aims to resolve the disjoint-character-set problem by rendering text into images and then applying a standard image encoder, such as the Vision Transformer with Masked Autoencoder (ViT-MAE) (He et al., 2022), to encode the features. In this work, we use PIXEL (Rust et al., 2023), a pixel-based language model pre-trained on the Common Crawl dataset with a masked image modeling objective, to encode the visual text lines for ancient languages. Additionally, we use PIXEL-MT (Salesky et al., 2023), a pixel-based machine translation model pre-trained on 59 languages, for the machine translation task.

Full Document Image Encoding. When the images of ancient artifacts are available (e.g., for Linear A and Cuneiform), we can encode the full-document images directly. We use ResNet-50 (He et al., 2016) as the backbone model for full-document image inputs.

3.2 Task-Specific Layers

Machine translation. After encoding the input to vectors, machine translation requires a decoder to generate sequential outputs. Encoder-decoder models, such as T5 (Raffel et al., 2020), ByT5, PIXEL-MT, and BPE-MT (Salesky et al., 2023), use 3/6/12 layers of Transformer blocks as the decoders. For Encoder-only models, such as (m)BERT or PIXEL, we attach a GPT2 model (Radford et al., 2019) as the decoder to produce sequential output. Among the aforementioned models, T5, ByT5, and PIXEL are pre-trained on large-scale text corpora such as the Common Crawl; PIXEL-MT and BPE-MT are pre-trained on 1.5M pairs of sentences of 59 modern languages; PIXEL-MT is an encoder-decoder model with a 6-layer Transformer encoder and a 4-layer Transformer decoder.

Classification. We attach a two-layer ReLU-activated perceptron (MLP) with a hidden size of 512 to the encoder for all classification tasks. The MLP outputs the predicted distribution over the candidate classes.

Dependency Parsing. After encoding, we use the deep bi-affine parser (Dozat and Manning, 2017) for dependency parsing, which assigns a score to each possible dependency arc between two words. We use the minimum spanning tree (MST) algorithm during inference to find the best dependency tree for each sentence.

Task	Model	BSZ	Steps	LR
translation	visual	64	30,000	5e-4
translation	textual	56	30,000	5e-4
translation	byT5	64	100,000	1e-3
classification	visual/textual	256	30,000	5e-4
parsing	visual/textual	256	1,000	8e-5

Table 3: Hyperparameter configuration. Note that, byT5 is particularly hard to converge compared to other transformer-based models. For the parsing task, due to the low-resource nature of the parsing data, 1,000 steps are sufficient to achieve model convergence.

4 Experiments and Analysis

We describe our general model fine-tuning approach in §4.1 and analyze model performance on the aforementioned tasks in the succeeding subsections.

4.1 General Experimental Setup

We use the Huggingface Transformers library (Wolf et al., 2020) in all experiments, except for machine translation, where we use the PIXEL-MT and BPE-MT models.[9] We modified code and model checkpoints provided by Salesky et al. (2023) based on fairseq (Ott et al., 2019) for the two exceptions.

We use Adam (Kingma and Ba, 2015) as the optimizer for all models, with an initial learning rate specified in Table 3. We use early stopping when the validation loss fails to improve for ten evaluation intervals (1000 iteration per interval). For data without a standard test set, we run a fixed number of training iterations and report the performance on the validation set after the last iteration. All experiments are conducted on an NVIDIA-RTX A6000 GPU, and the training time ranges from 2 minutes to 50 hours, depending on the nature of the task and the size of the datasets. Unless otherwise specified, all parameters, including those in pre-trained models, are trainable without freezing. We summarize other configurations in Table 3.

4.2 Machine Translation

We compare the performance of the models on machine translation, where we translate ancient Egyptian (EGY), Akkadian (AKK), and Old Chinese (ZHO) into English (Table 4a). We find that the PIXEL-MT model consistently achieves the best BLEU score across the three languages, outperforming the second-best method by a large margin.

Models with pre-training do not always outperform those trained from scratch (Gutherz et al., 2023). We find that all models that take textual (Unicode or latinized) input achieve worse performance than models trained from scratch with the same type of textual input, suggesting that the lack of overlap in symbol inventories poses a serious problem for cross-lingual transfer learning. Our results indicate that choosing the correct input format is crucial to achieving the full advantage of pre-training.

In addition, the PIXEL-MT model, pre-trained on paired data in modern languages (TED59), significantly outperforms PIXEL + GPT2 (pre-trained with masked language modeling) across the board. Another model, BERT-MT, which is further pre-trained on the same parallel text (TED59) with BERT initialization, also achieves comparable performance. These results emphasize the importance of pre-training on modern paired data, empirically suggesting that the PIXEL encoder with parallel text pretraining is an effective combination for ancient logographic language translation.

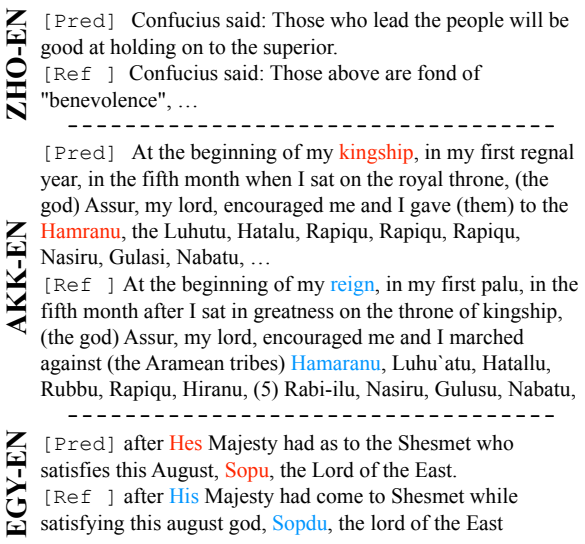

Figure 4: Case study for machine translation using the PIXEL-MT model. Notably, there are many spelling errors in the predictions, particularly with uncommon named entities.

Qualitative analysis. As shown in Figure 4, the low BLEU scores for ZHO-EN translation is a result of the translation model failing to capture the meaning of the input, instead focusing on repeated formatting queues: e.g., "`Confucius said:`

[9]The prefix PIXEL or BPE also indicates the type of input representation the model uses.

(a) Machine translation (BLEU score)

Modality	Tokenization	Input	Model	Pre-trained?		Source Language		
				MLM	MT	EGY	AKK	ZHO
			Dataset size (# lines)			2,337	8,056	500
Visual	token-free	textline	PIXEL + GPT2[1]	✓	✗	2.83	7.51	1.14
Visual	token-free	textline	PIXEL-MT	✗	✓	**29.16**	**44.15**	**5.45**
Textual	BPE *w/ ext vocab*[2]	Unicode	T5	✓	✗	n/a	12.42	0.28
Textual	byte-level	Unicode	ByT5	✓	✗	n/a	4.51	0.53
Textual	char-level	Unicode	Conv-s2s	✗	✗	-	36.52*	-
Textual	BPE	Unicode	BPE-MT	✗	✓	23.26	36.18	1.32
Textual	BPE	Latin	T5	✓	✗	21.18	10.67	n/a
Textual	char-level	Latin	Conv-s2s	✗	✗	-	37.47*	-

(b) Attribute prediction (F_1 accuracy)

Modality	Tokenization	Input	Model	LNA	AKK			EGY	ZHO
				geo	time	genre	geo	time	genre
			Number of classes	7	16	12	24	14	3
			Dataset size (# examples)	772	36,454	36,454	36,454	1,320	302
			Majority	14.28	6.25	8.33	4.17	7.14	33.33
Visual	token-free	photo	ResNet	8.24	75.02	45.45	62.99	n/a	n/a
Visual	token-free	textline	PIXEL	16.56	72.91	50.84	61.44	16.24	52.17
Textual	BPE *w/ ext vocab*[2]	Unicode	BERT	n/a	0**	0**	0**	n/a	74.85
Textual	BPE	Unicode	BERT	n/a	72.40	50.85	63.70	n/a	90.30
Textual	byte-level	Unicode	CANINE	n/a	82.83	47.88	56.42	n/a	**96.43**
Textual	BPE	Latin	BERT	32.92	80.91	53.45	65.10	34.71	n/a
Textual	BPE	Latin	mBERT	**50.52**	**83.08**	**56.71**	**66.33**	**36.25**	n/a

Table 4: (a) Results on machine translation (from each of the source languages to English), in terms of BLEU scores. MLM denotes models pretrained on unsupervised data with the masked language model (MLM) loss, while MT denotes models pretrained with supervised parallel data (TED59). (b) Macro F_1 scores for attribute prediction. *: numbers taken from Gutherz et al. (2023), where their models are trained from scratch, i.e., without pretraining. **: The character set is 100% disjointed without extending the vocabulary of the model, resulting in zero F_1 scores. [1]: This model is trained using PIXEL as the encoder and GPT2 as the decoder, with linear projection layers to convert the final layer of PIXEL into a prefix input for GPT2. [2]: This model is the only one experiencing out-of-vocabulary (OOV) issues with Unicode input. To address this, we extended the vocabulary with random initialization. **n/a**: indicates the representation of a specific language does not exist in our benchmark.

Those who...” Indeed, given that the topical domain of the ZHO-EN translation data is philosophical writing, achieving an accurate translation would be challenging even with a much larger set of parallel translations. For AKK-EN, we found that the overall quality to be quite good, despite the fact that errors in translating named entities appear more often than in standard MT tasks. This case study suggests that translation performance could improve further if we training using a custom target language (English) vocabulary. We also show more generated examples from the PIXEL-MT model in Appendix C.

4.3 Attribute Classification

Table 4b summarizes the performance of attribute classification with different features and models. As expected, the image features can work fairly well for some of these attribute classification tasks as many of the relevant features are visual (e.g., for time and location); but, are not generally as effective as textual input representations. By comparing BERT with latinized input and CANINE on Unicode, we find that when both accurate latinization and Unicode representations are available, latinization is the most informative feature—with the exception of time period classification for Akkadian. This exception is aligned with our understanding of Akkadian, as different Cuneiform characters are used across different time periods. Thus, in this case, Unicode can provide more clues for determining the time period of a sample. Note that the label distribution is not balanced for most ancient language attribution tasks. For more details, refer to Chen et al. (2024).

Modality	Model	Input	RIAO	MCONG
Dataset Size (# tokens)			5k	130k
Visual	PIXEL	Image	**92.74**	**85.22**
Textual	BERT	Latin	92.13	83.88

Table 5: Dependency parsing result on Akkadian (evaluated on the UD corpora RIAO and MCONG), in terms of labeled attachment scores (LAS). Note that the number of tokens are reported.

4.4 Dependency Parsing

We compare the dependency parsing performance of models with visual and textual encoders (Table 5).[10] While all models achieve quite high parsing accuracy, we find that models with visual encoders perform the best on both investigated corpora (RIAO and MCONG). During training, models taking visual input generally converge faster than their textual counterparts, which is in line with prior work (Salesky et al., 2023) that uses visual features for machine translation.

5 Ablation Study on OCR and Image Quality

As mentioned earlier, the majority of data from ancient times remain in the form of photographs. We first closely examine two different visual input representations for the ZHO-EN translation task, *handcopied figure* and *photograph* (§5.1). Next, we examine OCR performance on ancient logographic languages to gain better understanding of this bottleneck for current NLP pipelines (§5.2).

5.1 Handcopy v.s. Raw Image

Input representation	BLEU
photograph	2.09
handcopied figure	5.45

Table 6: Performance on ZHO-EN translation using the PIXEL-MT model with different visual input features.

For the ZHO-EN translation data, we have access to both photographs of the bamboo slips and handcopied textline figures (see the Bamboo script example in Figure 3 for reference). As shown in Table 6, the quality of the visual features significantly influences the translation accuracy—translations derived from photographs yield a low BLEU score of 2.09, whereas handcopied figures, which typically provide clearer and more consistent visual data, result in a higher BLEU score of 5.45. This result suggests that for models that perform implicit OCR as part of the translation process, the clarity of the source material is paramount.

[10]We only conduct experiments on Akkadian since it is the only language with off-the-shelf dependency annotations.

5.2 Text Recognition Study

We simplify the task of transcribing ancient texts by starting with lines of text that have been accurately segmented. For datasets that include glyph-level annotations, we employ glyph classification to recognize the text. Details on models and configuration of *line-level OCR* and *glyph classification* can be found in Appendix B.

Method	Output	LNA	EGY	AKK	ZHO
OCR	Unicode	57.17	N/A	5.72	71.85
OCR	Latin	63.44	65.88	21.98	N/A

Table 7: Line-level OCR results with the best validation character error rate (CER) reported. The study includes various writing systems using Kraken trained from scratch on segmented text lines. N/A: either the Unicode or Latin version of the text is not available.

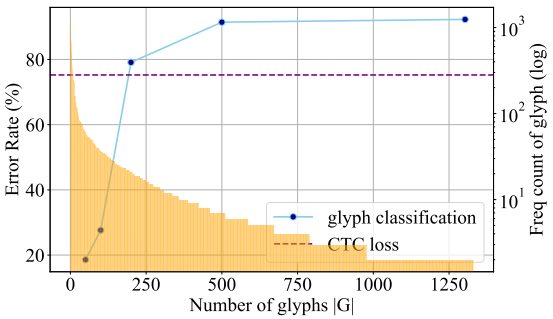

Figure 5: Glyph classification on Old Chinese (ZHO). **Left axis**: we plot the error rate of glyph classification. The data point at |G| = 50 shows the classification error calculated using the top 50 most frequent glyphs in the dataset. The purple horizontal line (71.85%) represents the line-level text recognition CER for ZHO, provided for reference. **Right axis**: The frequency count (in orange bars) of each glyph in the dataset. Note that the counts are in logarithmic scale, illustrating the long tail distribution of glyph counts.

Results. The line-level OCR performance for the four languages is presented in Table 7. When comparing digital renderings of text to handwritten samples, it is evident that Old Chinese (ZHO) achieves a CER of 71.85, while Linear A has a CER of 57.17. As shown in Figure 5, glyph classification

for ZHO is approximately 20% less accurate than line-level OCR, indicating that contextual features significantly aid in recognizing glyphs. Furthermore, there is a rapid increase in error rate as the number of glyphs increases, highlighting the intrinsic challenge of processing logographic languages, which typically have a large symbol inventories, and their frequency distribution often follows a long-tail pattern (see the orange bars in Figure 5). Therefore, developing robust visual models that can effectively leverage visual features is crucial for improving NLP on ancient logographic languages.

6 Related work

Because ancient languages are often low-resource, they present challenges that are closely related to other domains of NLP, such as low-resource machine learning and multi-lingual transfer learning. Recent work has explored the application of NLP techniques to ancient languages from the following perspectives:

Multilingual transfer learning and disjoint character sets. Muller et al. (2020) studied hard-to-process living languages such as Uyghur, and reported that a non-contextual baseline outperforms all pre-trained LM-based methods. Ancient languages also face the same problem, with even less data available. A major challenge that is mostly specific to ancient logographic languages, however, is the almost non-existent overlap of their symbol inventories with those of high-resource languages.

Visual representation of languages. Recently, several works have studied language processing based on images of text. Rust et al. (2023) pre-trained a masked language model on digitally rendered text and achieved comparable performance with text-based pre-training strategies on downstream tasks. Salesky et al. (2023) found that a multi-lingual translation system with pixel inputs was able to outperform its textual counterpart.

Machine learning for ancient languages. Sommerschield et al. (2023) surveyed the status of pipelines for ancient language processing. Notably, the study concludes that applying machine learning methods to ancient languages is bottlenecked by the cost of digitization and transcription. According to the Missing Scripts Project,[11] only 73 of 136 dead writing systems are encoded in Unicode. Ancient languages, such as Ancient Greek or Latin (Bamman and Burns, 2020), benefit greatly from multilingual pre-training techniques, such as mBERT, XLM-R (Conneau et al., 2020), and BLOOM (Scao et al., 2022). The applicability of these techniques is limited when it comes to languages that were historically written in obsolete or extinct writing systems—for instance, languages like Sumerian and Elamite were recorded in Cuneiform script and ancient Chinese was inscribed on oracle bone or bamboo. However, observations by existing work support the potential utility of visual processing pipelines for ancient languages.

Logographic writing systems. Logography typically denotes a writing system in which each glyph represents a semantic value rather than a phonetic one, however, all the languages studied in our paper have at least some phonetic component based on the rebus principle. This paper emphasizes *ancient* logographies that (i) possess extensive glyph inventories; (ii) feature glyphs with multiple transliterations or functional uses; and (iii) are low-resource and/or remain in photo format (Caplice et al., 1991; Allen, 2000; Woodard, 2004). Existing research on logographic languages has predominantly focused on those that are well-resourced and still in use, such as Modern Chinese (Zhang and Komachi, 2018; Si et al., 2023), or on data that has already been carefully transcribed into Latin or annotated with extra semantic information (Wiesenbach and Riezler, 2019; Gutherz et al., 2023; Jiang et al., 2024). Our paper aims to address the gap in resources (new data) and methodologies (visual-only approaches) for encoding and analyzing ancient logographic languages, which are crucial for a more comprehensive understanding of historical linguistic landscapes.

7 Conclusion

By comparing the results on four representative languages on three downstream tasks, we demonstrated the challenges faced in applying natural language processing techniques to ancient logographic writing systems. Our experiments demonstrate, however, that encoding more readily available visual representations of language artifacts can be a successful strategy for downstream processing tasks.

[11] https://worldswritingsystems.org/

8 Limitations

More discussion on ancient logographic languages. Due to page limits, we do not discuss ancient logographic languages in a critical way. Technically, there are no logographic languages, only languages written in logographic writing systems (aka *logography*) (Gorman and Sproat, 2023). In this paper, we use the term "logographic languages" to denote languages that are quite different from those with alphabetic writing systems especially when we tried to apply NLP toolkits for computational paleography. As mentioned in the related work section, these languages feature glyphs that have multiple transliterations or functional uses. In other words, these languages are homophonous or a glyph can be used as a phonetic value or semantic value. Therefore, the boundaries between logographic and phonographic is not sharply separated.

Including more logographic writing systems. We selected the four languages because we would like to include at least one language from early civilization in Ancient China, Ancient Egypt, Indus Valley Civilization, Mesoamerica, Mesopotamia and Minoan Civilization (Woodard, 2004). However, we fail to include Mayan hieroglyphs (Mesoamerica) and Oracle Bone script. However, Mayan is excluded because the collection[12] is still working in process. Oracle Bone script is primarily omitted due to copyright issues.

Textline images. Most ancient languages remain as full-document images. In this paper, we use digitally rendered text as a surrogate visual feature for Akkadian. In reality, much of Cuneiform data is still in hand copies or in photo format. In the future, we look to conduct apples-to-apples comparisons for all languages once the line segmentation annotations become available.

Annotation quality and quantity. The study of ancient languages is constantly evolving; humanities scholars have not agreed on explanations, transliterations, or even the distinctions between certain glyphs or periods. We try our best to carefully annotated the data without bias; however, future editions of the benchmark are needed as things change all the time. A collective platform to correct errors and make more data available should be considered for future development.

[12]The Maya Hieroglyphic Text and Image Archive: https://digitale-sammlungen.ulb.uni-bonn.de/maya

Label imbalance. The classification task in our benchmark is label imbalanced. This is known to be a major issue for all machine learning tasks related to the ancient world (Sommerschield et al., 2023; Chen et al., 2024).

Acknowledgements

We thank Professor Wenbo Chen from the Department of Humanities at Central South University, China, for his advice on Old Chinese data collection and explanation. We thank Professor Edward Kelting from the Department of Literature at UC San Diego for his advice on Ancient Egyptian data collection and explanation. We thank Jerry You from http://www.ccamc.org/ for his help on Unicode and data processing for ancient languages.

We thank Elizabeth Salesky for her guidance in setting up cross-lingual machine translation experiments for ancient languages using both PIXEL and BPE encoders. We thank Chenghao Xiao for the help setting up the PIXEL + GPT 2 experiment.

We thank Kyle Gorman, Alexander Gutkin and Richard Sproat for their inspiring work (Sproat and Gutkin, 2021; Gorman and Sproat, 2023), which has significantly contributed to our understanding of logographic writing systems from a computational perspective.

We thank Nikita Srivatsan, Nikolai Vogler, Ke Chen, Daniel Spokoyny, David Smith, and the anonymous ARR reviewers for their insightful feedback on the paper draft. This work was partially supported by the NSF under grants 2146151 and 2200333.

References

James P Allen. 2000. *Middle Egyptian: An introduction to the language and culture of hieroglyphs*. Cambridge University Press.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283.

David Bamman and Patrick J Burns. 2020. Latin bert: A contextual language model for classical philology. *arXiv preprint arXiv:2009.10053*.

Richard Caplice et al. 1991. Introduction to akkadian.

Danlu Chen, Aditi Agarwal, Taylor Berg-Kirkpatrick, and Jacobo Myerston. 2023. Cuneiml: A cuneiform

dataset for machine learning. *Journal of Open Humanities Data*, 9(1).

Danlu Chen, Jiahe Tian, Yufei Weng, and Taylor Berg-Kirkpatrick andJacobo Myerston. 2024. Classification of paleographic artifacts at scale: Mitigating confounds and distribution shift in cuneiform tablet dating. In *Proceedings of the 1st International Workshop on Machine Learning for Ancient Languages*. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *International Conference on Learning Representations*.

Mengjie Fang, Linlin Xu, and Pascale Fung. 2020. Unsupervised cross-lingual transfer learning for contextualized word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2616–2629.

Kyle Gorman and Richard Sproat. 2023. Myths about writing systems in speech & language technology. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 1–5, Toronto, Canada. Association for Computational Linguistics.

Gai Gutherz, Shai Gordin, Luis Sáenz, Omer Levy, and Jonathan Berant. 2023. Translating akkadian to english with neural machine translation. *PNAS nexus*, 2(5):pgad096.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Kenji Imamura and Eiichiro Sumita. 2022. Extending the subwording model of multilingual pretrained models for new languages. *arXiv preprint arXiv:2211.15965*.

Guangyuan Jiang, Matthias Hofer, Jiayuan Mao, Lionel Wong, Joshua B. Tenenbaum, and Roger P. Levy. 2024. Finding structure in logographic writing with library learning. 48. Proceedings of the 46th Annual Conference of the CognitiveScience Society (CogSci 2024).

Benjamin Kiessling. 2022. The Kraken OCR system.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations (ICLR)*.

Zhao Liu. 2003. *Annotation and Translation for Guodian Bamboo Slips*. Fujian Renmin Publisher.

Mikko Luukko, Aleksi Sahala, Sam Hardwick, and Krister Lindén. 2020. Akkadian treebank for early Neo-Assyrian royal inscriptions. In *Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories*, pages 124–134, Düsseldorf, Germany. Association for Computational Linguistics.

Benjamin Muller, Antonis Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2020. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models.

Mark-Jan Nederhof and M Berti. 2015. Ocr of handwritten transcriptions of ancient egyptian hieroglyphic text. *Altertumswissenschaften in a Digital Age: Egyptology, Papyrology and beyond, Leipzig*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Elizabeth Salesky, Neha Verma, Philipp Koehn, and Matt Post. 2023. Multilingual pixel representations for translation and effective cross-lingual transfer. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13845–13861, Singapore. Association for Computational Linguistics.

Ester Salgarella. 2020. *Aegean Linear Script (s): Rethinking the Relationship between Linear A and Linear B*. Cambridge University Press.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chenglei Si, Zhengyan Zhang, Yingfa Chen, Fanchao Qi, Xiaozhi Wang, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2023. Sub-character tokenization for Chinese pretrained language models. *Transactions of the Association for Computational Linguistics*, 11:469–487.

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*, pages 1–44.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528.

Lucien Tesnière. 1959. Éléments de syntaxe structurale.

Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2023. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11006–11017.

Zihan Wang, Stephen Mayhew, Dan Roth, et al. 2020. Extending multilingual bert to low-resource languages. *arXiv preprint arXiv:2004.13640*.

Philipp Wiesenbach and Stefan Riezler. 2019. Multitask modeling of phonographic languages: Translating middle egyptian hieroglyphs. In *Proceedings of the 16th International Conference on Spoken Language Translation*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Roger D Woodard. 2004. *The Cambridge encyclopedia of the world's ancient languages*. Cambridge Univ. Press.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Longtu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 17–25, Brussels, Belgium. Association for Computational Linguistics.

A Dataset collection

A.1 Linear A

A.1.1 Attribute classification

Linear A is a logo-syllabic, undeciphered writing system that was used in Bronze Age Greece (ca. 1800-1450BCE). We crawled a dataset of 772 tablets from the publicly available SigLA database [13]. The metadata includes find-place, tablet dimensions, total number of signs and possible transliteration. For the attribute classification task, we use find-place as the target class, with the following classes `['Arkhanes', 'Gournia', 'Haghia Triada', 'Haghios Stephanos', 'Kea', 'Khania', 'Knossos', 'Kythera', 'Malia', 'Mallia', 'Melos', 'Mycenae', 'Papoura', 'Phaistos', 'Psykhro', 'Pyrgos', 'Syme', 'Tylissos', 'Zakros']`. We only keep classes whose tablets are no less than 10, which results in only keeping 7 classes: `['Haghia Triada', 'Khania', 'Phaistos', 'Zakros', 'Knossos', 'Malia', 'Arkhanes']`.

[13] https://sigla.phis.me/browse.html

A.1.2 Machine Translation

A.2 Ancient Egyptian

A.2.1 Attribute classification

We use time periods whose data more than 50 to conduct this task, resulting in 1,230 samples of 12 classes.

The classes are as below: `['Unas', 'Senwosret I Kheperkare', 'Pepi I Merire', 'Ramesses II Usermaatre-Setepenre', 'Sety I Menmaatre', 'Tuthmosis III Menkheperre (complete reign)', 'Pepi II Neferkare', 'Hatshepsut Maatkare', '18th Dynasty', 'Mentuhotep II Nebhepetre (complete reign)', 'Amenemhat II Nebukaure', 'Tutankhamun Nebkheperure', 'Cleopatra VII Philopator', '12th Dynasty']`

A.3 Akkadian (Cuneiform)

A.3.1 Attribute prediction

We use the dataset from Chen et al. (2023), containing 34,562 photographs of Akkadian cuneiform tablets with their transcriptions in cuneiform unicode and Latin transliterations. The metadata contains attributes such as time period, genre and geographical location (found place) that we have used for attribute classification.

A.3.2 Dependency parsing

The data is from Luukko et al. (2020) and comes in as normalization form of Akkadian. We implemented a rule-based method to convert the normalization form back to standard transliteration for consistency. When the auto-conversion fails, we simply just keep the normalization form.

A.3.3 Machine translation

The data is from Gutherz et al. (2023), a collection of 8,056 lines of Akkadian-English pair. The Akkadian representation is available in both ATF transliteration and Cuneiform Unicode.

A.4 Bamboo script

There are actually more than one scripts used in bamboo slips, for example, Seal script and Clerical script. For simplicity, our dataset generally call it Bamboo script. We pick two famous collections of the bamboo script, BaoShan and GuoDian bamboo slip collections for attribute prediction and machine translation.

Figure 6: Sample parallel data of GuoDian Bamboo script dataset.

A.4.1 Attribute prediction

The BaoShan bamboo slips collection consists of 723 slips, which are in three genres.

A.4.2 Machine Translation

We used the GuoDian bamboo slips collection for machine translation, whose major genre is philosophical essay. By referring to (Liu, 2003), We manually labeled each complete sentences with their transliteration and translation in modern Chinese. After filtering out incomplete sentences or removing the sentences with high interpretation difficulty. We extracted and labelled 489 lines of parallel data. The Bamboo script was labeled by a trained interdisciplinary Ph.D. student under the guidance of Professor Wenbo Chen specializing in Ancient Chinese Philology.

B Ablation Study on Text Recognition

Line-level OCR. We use Kraken (Kiessling, 2022), a state-of-the-art OCR library for historical documents, to transcribe the image into Latin or Unicode. To handle unseen Unicode codepoints, we simply extend the vocabulary of the decoder. For Latin transliteration, we predict outputs at the character level. Similar to most OCR pipelines, the default OCR model of Kraken is a 2-layer bidirectional LSTM with a connectionist temporal classification (CTC) loss.

Glyph classification. We also applied Convolution Neural Networks (CNNs) to classify segmented glyphs. We use ResNet-50 as the backbone model, followed by a linear classification layer. We trained the model using cross-entropy loss with Adam optimizer. We resize the image of each glyph to 64×64 and apply 20% cropping. In total, there are $21,687$ segmented characters in the ZHO dataset.

C More Translation Case Study

We sample 10 examples from validation set for each language pairs.

C.1 AKK - EN

Akkadian translation samples from the PIXEL-MT models is shown in Figure 7. We showcase the first 10 examples from the validation set, we find that some annotation error present for example #2 and #6.

C.2 EGY - EN

Ancient Egyptian translation samples from the PIXEL-MT models is shown in Figure 8.

C.3 ZHO - EN

Old Chinese translation samples from the PIXEL-MT models is shown in Figure 9.

```
BLEU = 39.84 73.1/51.8/40.6/32.6 (BP = 0.842 ratio = 0.853 hyp_len = 1462 ref_len = 1713)
36 2811
0 BLEU = 34.30 71.3/45.3/30.6/20.2 (BP = 0.912 ratio = 0.916 hyp_len = 87 ref_len = 95)
[[pred]] At the beginning of my kingship, in my first regnal year, in the fifth month when I sat on the
royal throne, (the god) Assur, my lord, encouraged me and I gave (them) to the Hamranu, the Luhutu, Hatalu,
Rapiqu, Rapiqu, Rapiqu, Nasiru, Gulasi, Nabatu, Li<unk>ta<unk>u, Kaparu, Malitu, Adadu, Gibre, Gurumu,
Gibre, Gurumu,
[[ref ]] At the beginning of my reign, in my first palu, in the fifth month after I sat in greatness on the
throne of kingship, (the god) Assur, my lord, encouraged me and I marched against (the Aramean tribes)
Hamaranu, Luhu`atu, Hatallu, Rubbu, Rapiqu, Hiranu, (5) Rabi-ilu, Nasiru, Gulusu, Nabatu, Li`ta`u, Rahiqu,
Kapiru, Rummulitu (Rummulutu), Adile, Gibre, Ubudu, Gurumu,
 ====================
1 BLEU = 26.01 60.4/34.0/21.1/10.6 (BP = 1.000 ratio = 1.085 hyp_len = 154 ref_len = 142)
[[pred]] In the midst of the sea, just as Sarridu (and) NN seized ... in my hand ... I captured 7,999 ...
their ... ; Sarriduri fled to save his life, and as many as the sun had flowed ... with arrow(s), (and) ...
they threw his bed with ... s, as far as the back of the Euphrates, and his bed ... of his royal path,
together with stones, his royal chariot, ... , his ... , many as there were, (and) his .... s, as far as
the border of the Euphrates, his royal chariot(s), his ... , everything that was without number.
[[ref ]] In the midst of that battle, I captured Sarduri's ... I ... 72,950 of their ... from ... (10') ...
In order to save his life, Sarduri fled at night and (thus) escaped very quickly before sunrise... With an
arrow that cuts off lives, I drove him back to the bridge (crossing over) the Euphrates River, on the
border of his land. I took away from him his bed, ... , his royal processional chariot, the cylinder seal
(that hung around) his neck, together with his necklace, his royal chariot, ... , their ... , (and) many
other things, without number.
 ====================
2 BLEU = 0.00 0.0/0.0/0.0/0.0 (BP = 0.717 ratio = 0.750 hyp_len = 3 ref_len = 4)
[[pred]] ...
[[ref ]] (Completely destroyed)
 ====================
3 BLEU = 46.04 84.6/76.0/58.3/47.8 (BP = 0.707 ratio = 0.743 hyp_len = 26 ref_len = 35)
[[pred]] I received tribute from the Medes, the land Ellipi, and the city rulers of the mountain regions,
as far as Mount Bikni.
[[ref ]] I received the payment of the Medes, the people of the land Ellipu, and the city rulers of all of
the mountain regions, as far as Mount Bikni: (10)
 ====================
4 BLEU = 2.07 33.3/6.2/3.6/2.1 (BP = 0.329 ratio = 0.474 hyp_len = 9 ref_len = 19)
[[pred]] , Da'qa-il of the city Saakka, Ilili-Arbail.
[[ref ]] I captured (and) defeated the cities Daiqansa, Sakka, Ippa, Elizansu, (5)
 ====================
5 BLEU = 43.94 71.0/47.5/38.3/28.8 (BP = 1.000 ratio = 1.107 hyp_len = 62 ref_len = 56)
[[pred]] (As for) every royal treasure, live male sheep whose wool is sweet, the bird of the sky, the
winged bird of the sky whose wings are dyed with red-purples, which are dyed with red-purple wool, I
received horses, mules, oxen, and sheep and goats, camels, together with their camels.
[[ref ]] all kinds of precious things from the royal treasure, live sheep whose wool is dyed red-purple,
flying birds of the sky whose wings are dyed blue-purple, horses, mules, oxen, and sheep and goats, camels,
she-camels, together with their young, I received (from them).
 ====================
6 BLEU = 3.39 8.3/4.5/2.5/1.4 (BP = 1.000 ratio = 12.000 hyp_len = 12 ref_len = 1)
[[pred]] whose reign is unalterable, whose reign is Samas is firm.
[[ref ]] ,
 ====================
7 BLEU = 44.74 88.1/73.2/57.5/41.0 (BP = 0.717 ratio = 0.750 hyp_len = 42 ref_len = 56)
[[pred]] ... ni, the Alaya, ... the land Ursalma and ... Dalta of the land Ellipi, ... horses, mules,
donkeys, oxen, and sheep and goats ...
[[ref ]] ... the land ... ni, the land Ayyalaya, ... , the land Niksamma, ... I received the payment of
Dalta of the land Ellipu: ... horses, mules, Bactrian camels, oxen, and sheep and goats, ...
 ====================
8 BLEU = 52.45 80.4/63.6/55.6/47.2 (BP = 0.867 ratio = 0.875 hyp_len = 56 ref_len = 64)
[[pred]] (As for) Sardardari of the land Urartu, revolted against me and conspired with Mati'-ilu. In the
lands Kistan and Halpi, districts of the city Kummuhu, he became frightened of the terrifying radiance of
my weapons, and (then) fled to him in order to save his life.
[[ref ]] Sarduri of the land Urartu revolted against me and conspired with Mati`-il (against me). In the
lands Kistan and Halpi, districts of the city Kummuhu, I defeated him and took his entire camp away from
him. He became frightened of the terrifying radiance of my weapons and fled alone in order to save his
life.
 ====================
9 BLEU = 49.33 83.0/59.6/47.1/40.0 (BP = 0.893 ratio = 0.898 hyp_len = 53 ref_len = 59)
[[pred]] I fashioned image(s) of the gods, my lords, and my royal image made of gold (and) erected (it) in
the palace of the city Gaza. I regarded (them) as gods of their lands and (them) their regular offerings.
[[ref ]] I fashioned (a statue bearing) image(s) of the gods, my lords, and my royal image out of gold,
erected (it) in the palace of the city Gaza, (and) I reckoned (it) among the gods of their land; I
established their sattukku offerings.
 ====================
10 BLEU = 0.00 0.0/0.0/0.0/0.0 (BP = 0.368 ratio = 0.500 hyp_len = 3 ref_len = 6)
[[pred]] the city Labbe
[[ref ]] The cities Lab`u,
```

Figure 7: Akkadian translation samples from the PIXEL-MT models.

```
0 BLEU = 6.90 25.0/9.1/5.0/2.8 (BP = 0.920 ratio = 0.923 hyp_len = 12 ref_len = 13)
[[pred]] for the Great Heroon who is in front of the curtain.
[[ref ]] before the Great Heron which went forth (from) the the garden
 ====================
1 BLEU = 35.10 69.6/45.5/28.6/20.0 (BP = 0.957 ratio = 0.958 hyp_len = 23 ref_len = 24)
[[pred]] The king of Upper and Lower Egypt gave me two, or gold: two rings, two necropolis, one bracelet.
[[ref ]] The king of Upper and Lower Egypt has given to me, of gold: 2 rings, 2 necklaces, one bracelet,
 ====================
2 BLEU = 22.03 66.7/35.3/18.8/6.7 (BP = 0.946 ratio = 0.947 hyp_len = 18 ref_len = 19)
[[pred]] the butchers, the mother of the mother, the Ba of Re of the prime time.
[[ref ]] father of the fathers, mother of the mothers, the ba of Re of the primival times.
 ====================
3 BLEU = 66.87 80.0/75.0/66.7/50.0 (BP = 1.000 ratio = 1.250 hyp_len = 5 ref_len = 4)
[[pred]] The messenger of Ra.
[[ref ]] The messenger of Ra
 ====================
4 BLEU = 51.93 72.7/60.0/44.4/37.5 (BP = 1.000 ratio = 1.000 hyp_len = 11 ref_len = 11)
[[pred]] like the brewer or one who is under the knife.
[[ref ]] like the burden of one who is under the knife,
 ====================
5 BLEU = 11.37 57.1/20.0/5.3/2.8 (BP = 1.000 ratio = 1.050 hyp_len = 21 ref_len = 20)
[[pred]] after Hes Majesty had as to the Shesmet who satisfies this August, Sopu, the Lord of the East.
[[ref ]] after His Majesty had come to Shesmet while satisfying this august god, Sopdu, the lord of the East
 ====================
6 BLEU = 22.25 55.6/35.3/18.8/6.7 (BP = 1.000 ratio = 1.125 hyp_len = 18 ref_len = 16)
[[pred]] This Pepi Neferkare is on the right side of this island of the land of Pepi Neferkare.
[[ref ]] This Pepi Neferkare is one righteous near this island of land that Pepi Neferkare swum to
 ====================
7 BLEU = 7.59 25.0/14.3/8.3/5.0 (BP = 0.687 ratio = 0.727 hyp_len = 8 ref_len = 11)
[[pred]] Pepi Neferkare has brought down the aura.
[[ref ]] Pepy Neferkare has requested his need as ruler, 2 arouras
 ====================
8 BLEU = 19.51 83.3/60.0/25.0/16.7 (BP = 0.513 ratio = 0.600 hyp_len = 6 ref_len = 10)
[[pred]] Taking the rope of the sky
[[ref ]] Taking the run for the lady of the sky,
 ====================
9 BLEU = 20.39 66.7/45.5/20.0/5.6 (BP = 0.846 ratio = 0.857 hyp_len = 12 ref_len = 14)
[[pred]] for the one who speaks the monuments for Amun in Karnakhnak.
[[ref ]] for the ka of the one who calls the monuments for Amon in Karnak
```

Figure 8: Ancient Egyptian translation samples from the PIXEL-MT models.

```
0 BLEU = 0.83 8.8/0.7/0.4/0.2 (BP = 1.000 ratio = 1.030 hyp_len = 68 ref_len = 66)
[[pred]] Associating with those above is close to serving the king, associating with those below is close
to serving the king, and self-cultivation through self-reflection through self-reflection through self-
reflection through self-reflection through self-reflection through self-reflection through self-reflection
through self-reflection through self-reflection through self-reflection through self-reflection through
self-reflection through self-rection through self-rection through self-reflection through self-rection
through self-rection through self-rection through self-rection through self-rection through self-rection
through self-reflection
[[ref ]] If you are engaged in learning, your knowledge will increase day by day; if you are engaged in
"Tao", your "doing" will decrease day by day. Reduce it again and again until it reaches "inaction". Don’t
take the initiative to do something, but obey the natural “behavior” of all things in the world.
 ====================
1 BLEU = 2.44 18.2/2.4/1.2/0.7 (BP = 1.000 ratio = 1.294 hyp_len = 22 ref_len = 17)
[[pred]] The gentleman hears "Shang Shu Jun Chen" says: "Be careful and no longer tell me good."
[[ref ]] When the inferior people heard "Tao", they laughed and did not believe it.
 ====================
2 BLEU = 30.90 57.1/50.0/40.0/25.0 (BP = 0.751 ratio = 0.778 hyp_len = 7 ref_len = 9)
[[pred]] Wisdom is the source of virtue.
[[ref ]] Saint wisdom is the source of ritual and music
 ====================
3 BLEU = 4.72 31.8/4.8/2.5/1.3 (BP = 1.000 ratio = 1.222 hyp_len = 22 ref_len = 18)
[[pred]] There are thousands of people waiting before the birth of things, and they can preside over the
affairs of things.
[[ref ]] We must do something before things happen, and we must manage things before they become chaotic.
 ====================
4 BLEU = 2.83 20.0/3.6/1.9/1.0 (BP = 0.819 ratio = 0.833 hyp_len = 15 ref_len = 18)
[[pred]] Take the rule of thumb, and the rule of thumb will change accordingly.
[[ref ]] If this principle is implemented in a country, the country's "virtue" will be rich;
 ====================
5 BLEU = 3.13 13.3/3.6/1.9/1.0 (BP = 1.000 ratio = 1.500 hyp_len = 15 ref_len = 10)
[[pred]] Being able to hold on to the superior is not satisfied with the superior.
[[ref ]] Those who maintain this path should not worship abundance.
 ====================
6 BLEU = 2.37 40.0/25.0/16.7/12.5 (BP = 0.111 ratio = 0.312 hyp_len = 5 ref_len = 16)
[[pred]] Simplicity cannot last long.
[[ref ]] If you are not smart, you will not be able to think for long.
 ====================
7 BLEU = 2.81 57.1/16.7/10.0/6.2 (BP = 0.180 ratio = 0.368 hyp_len = 7 ref_len = 19)
[[pred]] Respect is the principle of self-reflecting.
[[ref ]] Being rich and arrogant will bring disaster to oneself. Retiring after success is the way of
heaven.
 ====================
8 BLEU = 1.42 10.5/2.8/1.5/0.8 (BP = 0.591 ratio = 0.655 hyp_len = 19 ref_len = 29)
[[pred]] Those who are happy after hearing the "Tao" can't do things without speaking because they can't.
[[ref ]] Confucius said: A gentleman does not say what he can say but cannot do; he does not do what he can
do but cannot tell others.
 ====================
9 BLEU = 9.03 35.7/15.4/8.3/4.5 (BP = 0.751 ratio = 0.778 hyp_len = 14 ref_len = 18)
[[pred]] If serious crimes are not committed, people will be punished without faults.
[[ref ]] If he acts with true feelings, people will trust him even before he starts taking action.
 ====================
10 BLEU = 9.52 40.0/14.3/7.7/4.2 (BP = 0.819 ratio = 0.833 hyp_len = 15 ref_len = 18)
[[pred]] There are seven types of poisonous things happening because they have something to do.
[[ref ]] There are seven kinds of love, and only natural love is close to "benevolence".
```

Figure 9: Old Chinese translation samples from the PIXEL-MT models.