Towards Ancient Meroitic Decipherment: A Computational Approach

Joshua Otten, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{jotten4,antonis}@gmu.edu

Abstract

The discovery of the Rosetta Stone was one of the keys that helped unlock the secrets of Ancient Egypt and its hieroglyphic language. But what about languages with no such "Rosetta Stone?" Meroitic is an ancient language from what is now present-day Sudan, but even though it is connected to Egyptian in many ways, much of its grammar and vocabulary remains undeciphered. In this work, we introduce the challenge of Meroitic decipherment as a computational task, and present the first Meroitic machine-readable corpus. We then train embeddings and perform intrinsic evaluations, as well as cross-lingual alignment experiments between Meroitic and Late-Egyptian. We conclude by outlining open problems and potential research directions.[1]

1 Introduction

Perhaps one of the most critical elements to deciphering an unknown language is a collection of bilingual texts. From a known language, one can make conclusions about phonetic, morphological, and lexical aspects of the target language, hopefully leading to eventual decipherment. Without such a text, translation of a lost language is practically inconceivable. Only in this day and age, where computer technological applications appear to nearly reach the limits of human imagination, is decipherment with a monolingual corpus potentially feasible, and Meroitic is a great candidate for such work.

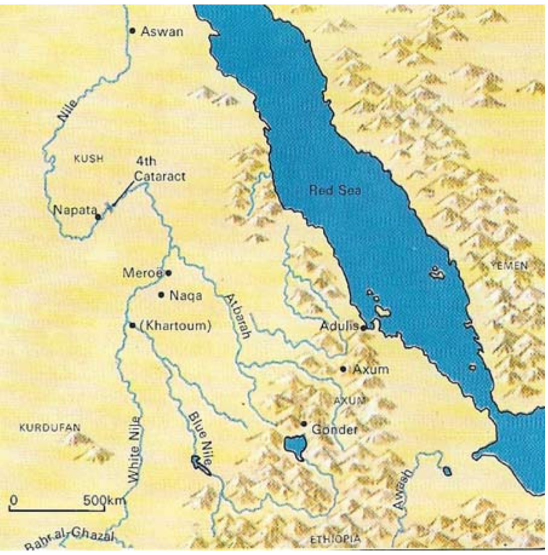


Figure 1: Ancient Meroë (Kush) between approximately 100 BC - 300 AD.

Meroitic is the language of the ancient state of Meroë, a Kushite-ethnic group living in approximately 270 BC - 330 AD of what is now present-day Sudan (see Figure 1). Partly due to its geographic location, the Meroë civilization has been studied relatively little, despite its significant presence in the ancient and classical world (Shinnie, 1967; Rilly and de Voogt, 2012). One of the largest obstacles to understanding the Meroitic state, however, is that its language is not well understood, and we currently possess no bilingual texts large enough to illicit an attempt at decipherment. As stated by Shinnie (1967), a British africanist and archaeologist, *"... until this language has been successfully read and the inscriptions translated, much of the story of Meroë will remain unknown."*

While there have been past attempts to understand the language, few have been made by Computer Scientists. Our hope is that by leveraging machine translation techniques, one could bridge the gap that has hindered progress in this language for decades. In the encouraging words of Griffith: "If new eyes, whether of trained decipherers or of scholars expert in North African philology, will exert themselves upon it, the secrets of Meroitic should soon be yielded up" (Griffith, 1911).

[1] The corpus, along with data and code necessary to replicate our experiments: https://github.com/Joshua-Otten/Meroitic-Corpus

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 208–219
May 4, 2025. ©2025 Association for Computational Linguistics

To our knowledge, this is the first work to use modern NLP techniques towards Meroitic decipherment. Our contributions include the following:

- First, we introduce the task of Meroitic decipherment to the NLP community, and provide an overview of the language and its unique challenges.
- Additionally, we present the first machine-readable Meroitic corpus.
- Then, we train embeddings on this and Late-Egyptian data, and provide intrinsic evaluations of each.
- Finally, we perform alignment experiments between the Meroitic and Late-Egyptian embeddings, and lay groundwork for future research in this area.

Meroitic decipherment would allow us to read one of Africa's oldest written languages as well as to better understand the Meroitic civilization and its historical and cultural role across the ancient world.

2 The Meroitic Language

A great deal of what we currently know of Meroitic vocabulary and grammar comes from funerary inscriptions, which represent about one third of the available corpus and contain formulas that have been extensively analyzed by Griffith, Hintze, and Rilly (Rilly and de Voogt, 2012). Fortunately, (aside from a few vowel uncertainties) the writing system has already been understood, which allows us to successfully transliterate Meroitic hieroglyphic texts. Scholars have already been able to uncover a number of grammatical elements, allowing them to identify such features as determinants, genitival constructions, and appositions[2] (Rilly and de Voogt, 2012).

The grammar of Meroitic appears to be agglutinative (Rilly and de Voogt, 2012), minimizing the complexity of analyzing roots and grammatical structure. The writing system utilizes an alphasyllabary (Rilly and de Voogt, 2012), which allows for nearly one-to-one phonetic mapping. This removes many of the challenges present in MT for Ancient Egyptian or Cuneiform languages, where signs are neither consistently phonetic or logographic (Sahala and Lindén, 2023). Finally, Meroitic's separator character, ' *:* ', although not consistently used (Rilly and de Voogt, 2012), greatly improves our ability to identify roots, suffixes, and postpositions.

Scholars have also proposed linguistic affiliations, and we are by now confident that Meroitic is Nilo-Saharan of the Eastern Sudanic group's Northern branch, making it 'North East Sudanic.' The closest language group to Meroitic is Nubian, followed by Nara, whereas Taman and Nyima are separate branches within the same family (Rilly, 2008).

One of the most critical goals for Meroitic decipherment will be expanding our limited vocabulary (Lobban Jr, 1994), the hope being that once we have identified more words, a better understanding of the grammar should be forthcoming. Cognate detection has presented one of the most promising avenues for this, especially with regard to prior scholarly efforts. To this end, we present a cognate investigation by hand for two common Meroitic words in Appendix C. However, since scholars have already been searching the cognate space since the writing system was deciphered by Griffith (Rilly and de Voogt, 2012), we consider it more fruitful to first focus on new computational methods that have never been tried before.

2.1 Challenges

Data Scarcity Over time, scholars have aggregated approximately 2,200 Meroitic texts. While it is a sizeable amount of material for a lost language, it still would of course be considered a drop in the bucket for standard computational linguistics tasks.

Additionally, collecting data for comparison will become an important task in the future. In particular all close language relatives to Meroitic are also extremely low-resource languages. Although some dictionaries (e.g. Nubian, Old Nubian, Nara) are available, there exist hardly any complete corpora for these relatives in machine readable format. Ideally, analysts would perform experiments on not merely one, but many languages, and use those results cumulatively to better understand Meroitic.

Orthographic Variation An additional challenge for those hoping to decipher Meroitic is the orthographic variation across the language. "All researchers since Griffith who have worked on Meroitic have observed and sometimes complained about the great variability of the writing. ... [T]here are frequent examples of different spellings at the same site and from the same era for the more commonly used terms" (Rilly and de Voogt, 2012). It is possible that these may partly con-

[2] Where two adjacent noun phrases refer to the same object; for instance when Meroitic titles precede personal names.

sist of dialectal differences, but the fact that we find examples from the same place and time undermines dialect as a primary suspect. Of course, variations also include scribal mistakes (Rilly and de Voogt, 2012), as well as differences in region and time period (Rilly, 2007). For instance, Osiris and Isis epithets often began with an initial /q/ in places near Meroë and the third cataract, whereas they primarily began with /w/ around the second cataract; /$qetneyineqeli$/ and /$wetneyineqeli$/ are two valid writings for Isis epithets. Also, the word for "sister" was written /$kdise$/ as well as /$kdite$/.

3 Related Work

Schenkel (1972) used computational systems to search Meroitic texts, identifying verbs and common suffixes in three long royal narratives, and comparing them to verbal suffixes in the Barya language. Later, Ouellette and Longpre (1999) used a computer program called "Thoth: Language Cognate Program" to search for cognates in Meroitic, and concluded that "From these word lists it may be possible to continue the work of deciphering the Meroitic writing system until such time as a bilingual text becomes available."

More recently, several works have used machine translation, statistical techniques, and Bayesian probability to decipher foreign scripts (Knight et al., 2006; Snyder et al., 2010; Luo et al., 2019, 2021). These include methods to determine probable phonetic mappings (Knight et al., 2006), morphological segmentation, cognates (Snyder et al., 2010), and language relatedness (Luo et al., 2021).

A foundational experiment deciphered Ugaritic with machine translation techniques.[3] Comparing the "unknown" Ugaritic texts with a closely related language, Hebrew, the computers iteratively theorized alphabetic mappings based on character frequency. They then searched for cognates in the roots and particles using assumptions about morphology, "correctly translat[ing] over 60% of all distinct Ugaritic word-forms with Hebrew cognates and over 71% of the individual morphemes that compose them, outperforming the baseline by significant margins" (Snyder et al., 2010).

Luo et al. (2021) built on this work by generalizing it for other lost languages using a neural approach, which additionally improved the Ugaritic decipherment by 5.5%. This work is particularly relevant since it extracts cognates in undersegmented texts between a known and a lost language, even when the two languages are not particularly related.

Another statistical experiment was performed by Smith (2008), who tested whether Meroitic's word frequency distribution followed Zipf's law, concluding that, like all other human languages, it does indeed adhere to a Zipfian distribution.

4 A Machine-Readable Meroitic Corpus

As part of this project, we present the first machine-readable transcribed corpus by manually converting pre-transcribed Meroitic examples into machine-readable format, using examples from three main works: the vocabulary list of Lobban Jr (2021), as well as example phrases from Rilly (2007) and Millet (1968). We have also refitted three lengthy royal narratives from previous word-frequency experiments: Tañyidamani, the Hamadab Stela of Amanirenas and Akinidad (Hofmann, 1998), and the Kalabsha Inscription of Kharamadoye (Hägg, 2000). These data will be made publicly available on Github. Some corpus statistics are listed in Table 1, and examples of this data can be found in Appendix A. Some data instances include proposed translations; however, these translations are often constrained to titularies, toponyms, and anthroponyms (Lobban Jr, 2021), so they offer limited use for full decipherment.

Despite the existence of a Unicode font for Meroitic cursive and hieroglyphs,[4] we opt to use an ASCII-mapping of transcription characters already in use by scholars, both for ease of compatibility (e.g. users might not possess this font) and because our data sources usually provided examples as transcriptions rather than hieroglyphs. The mappings are specified in the corpus, and could certainly be changed to the Unicode if necessary.

In the past, no one transcription standard for Meroitic has been consistently used by scholars. Since we use solely pre-transcribed text, it is important to ensure that differing conventions are not inter-mixed. Thus, we create separate files of Meroitic examples designated by scholar. For a corpus, we combine the data from all files, but first convert to one standard; in this paper, we conform to Millet's paradigm; however, we provide information on our mapping scheme for each file, and characters can easily be replaced by others, so this

[3] Ugaritic had already been deciphered prior to this, but not using computers.

[4] Link to Meroitic font

Type	Statistics
Translated Meroitic words	193
Meroitic Phrases	897
Late-Egyptian complete texts	302
Scanned Nubian pages	708

Table 1: Data-collection statistics; does not include the Meroitic royal narratives.

should not pose an issue for reproducibility.

Additional data for Meroitic may be taken from REM (Le Répertoire d'Épigraphie Méroïtique), a corpus with over 1,000 Meroitic digitized inscriptions[5] (Leclant et al., 2000); however, many of these texts still require transcription[6] (Rilly and de Voogt, 2012) if they are to be analyzed through use of computer technology.[7]

As for data from other relevant languages, we scrape the Ramses Online Corpus of Late-Egyptian texts into JSON files, and use a cleaned version of the corpus for our experiments. Additionally, we are currently in the process of scanning and organizing materials from Old Nubian, Dongolese Nubian, and a few other modern Nubian varieties, in order to broaden the set of possible cognate candidates. We hope to soon develop a large enough sample set to conduct further experiments that may hopefully lead to an increased understanding of the Meroitic language. Note that all these languages are severely under-resourced, and almost all materials come in the form of books that require digitization and/or optical character recognition to be rendered useful. One challenge will be fine-tuning the OCR; for instance, we are currently unaware of any OCR developed for the Nubian or Old Nubian script, and up until now, our OCR attempts have yielded less-than ideal results. Eventually we will need an OCR model for the Meroitic REM texts as well.

5 Experiments

In this paper, we use our Meroitic corpus to train word embeddings. We evaluate their quality intrinsically with semantic similarity tests.

Afterwards, we attempt to align them with embeddings from Late-Egyptian. Creating cross-lingual representations is a method for lexicon induction, where embeddings can be aligned on a small dictionary of translation pairs (Mikolov et al., 2013b; Anastasopoulos and Neubig, 2020). We try this here with Meroitic and Egyptian, inducing lexemes of known words for evaluation. Through alignment to Egyptian, we hope to gain an understanding of the meaning (or grammatical function) of unknown words.

5.1 Why Egyptian?

Even though Ancient Egyptian is not phylogenetically related to Meroitic,[8] there are good reasons to believe the content of some Egyptian texts may be very relevant, both topically and chronologically, due to geographic and cultural similarities of the neighboring entities.

Napatan texts, Egyptian writings from the Napatan period (circa 800-300 BC), could be especially useful for translating words in the long royal narratives. Unfortunately, the number of long royal narratives and corresponding Napatan texts is not nearly large enough alone for comparison, and even what is available is not in ready machine-readable format. Additionally, unlike many Napatan texts, it is likely that the royal narratives came from oral tradition, since there are no dates, coronations, etc. apparent in the texts; this minimizes our ability to find similarities in format or structure.

Therefore, as a preliminary investigation, we choose to use Late-Egyptian (written between approximately 1550-700 BC (Hoch, 2023)) texts and stories for comparison, as they are openly accessible on the Ramses Online annotated corpus.

5.2 Data and Cleanup

For Late-Egyptian data, we scrape 302 texts, ranging from a few sentences to many paragraphs, into JSON format from the Ramses Online Corpus[9] (Polis et al., 27 August 2015).

Note that we use the phonological transcribed version of the Egyptian texts, rather than representations for the specific hieroglyphs used. This is in part because we did not see Gardiner Code representations (alphanumerical codes for individual hieroglyphs) in the Ramses Online texts. Egyptian Hieroglyphic writing makes use of non-phonetic

[5] Many of these can be found at https://ancientworldonline.blogspot.com/2017/11/repertoire-depigraphie-meroitique.html

[6] They include photographs of the physical carvings/documents, along with drawings of scholars' reconstruction of the hieroglyphs, but they have not been converted to an alphanumeric script.

[7] At this point, the texts from REM are image files, and hence not amenable for text-based language technologies.

[8] Egyptian is classified as Afro-Asiatic, Meroitic is a Nilo-Saharan language.

[9] http://ramses.ulg.ac.be/

features, such as determinatives, in order to contribute semantic and sometimes grammatical (eg. plurality) information of words (Allen, 2000). Using only the phonetic representation of texts leaves open the possibility of losing linguistic information, and may even result in ambiguity over certain lexical items. On the other hand, it may make sense to compare the words phonetically, considering that Meroitic hieroglyphs are purely alphasyllabic and do not use determinatives.

We create machine-readable corpora for both Meroitic and Egyptian by eliminating translations, metadata, dashes, colons, etc. and separate each example or text by a new line. Many of the Meroitic words are pre-segmented, so eliminating certain punctuation helps to separate words by morpheme in each language. This Meroitic corpus contains 871 example texts or phrases, and the Egyptian contains 1,729 unique types for 99,338 total tokens.

Data Augmentation Additionally, since scholars have been able to detect many words that are anthroponyms (people names), we augment the Meroitic data by swapping out royal names with each other, and then non-royal names with other non-royal names, thereby creating additional synthetic (yet valid) Meroitic examples. The resulting Meroitic corpus contains 1,868 unique word forms from 17,257 sentences or phrases: 782,761 words in all.

Evaluation Dictionaries We also compile seven small dictionaries (statistics in Table 2), pairing known Meroitic word forms with Egyptian counterparts that appear in the corpora; these act as our training and evaluation sets. The combined sets include over 90 pairs, with our largest dictionary (of nouns) containing 26. Known orthographic variants are present as distinct entries. Words are grouped by categories, such as part of speech, and these serve as training and test sets. We note that certain Egyptian words can be written with multiple independent morphemes yet have a distinct meaning. For instance, the word for "priest" is written as a genitival construction with two words: *ḥm-nṯr*, literally meaning "servant of god." The dash in transcription is important because it implies the single meaning in the presence of two independent morphemes. Therefore, to account for this kind of issue, we include certain punctuation, such as periods, and we also add dashes back into the Egyptian corpus for specific word pairs, just as they were written in the pre-cleaned version of the corpus.

5.3 Methods

We train `Word2Vec` embeddings (Mikolov et al., 2013a) on the Meroitic and Egyptian data. Although we considered using `fastText` which is good for learning subword information (Bojanowski et al., 2017), our cleaning process separated words into their constituent morphemes, so this would not be as helpful here. In order to consider how the small size of the corpora may affect the embedding space, we test with varying word vector dimensions: 20, 50, 100, and 120.

Next, we perform intrinsic evaluation on both embedding spaces (of dimension 100) with respect to semantic similarity, including both nearest-neighbors and word analogy tests. We carefully select known (or hypothesized) words and observe the top 10 most similar lexemes, with the hope that other known words that appear will be semantically related in some way. We also do this for numerals, expecting numerals to align with other numerals. Since most Meroitic words are unknown, our results may not include many known words; in these cases it is difficult to tell how semantically similar the words are. Therefore, we also compare cosine similarity scores to determine how close the words are in the embedding space.

Finally, we attempt embedding space alignment between the Egyptian and Meroitic in three settings: unsupervised, aligning on numbers (mostly shared numerals), and on our dictionary of nouns. We use `VecMap` (Artetxe et al., 2018b,a), since Anastasopoulos and Neubig (2020) found that it can perform better than other methods (`MUSE` (Conneau et al., 2017) and `UMWE` (Chen and Cardie, 2018)) for lexicon induction when the languages or writing-systems are distant.

We then evaluate with a lexicon induction task on each of our dictionaries, using a neighborhood of 10 words (reporting precision@10). Additionally, we perform a similar experiment with French and English Wikipedia-trained embeddings, using a hand-crafted alignment dictionary of 26 pairs, and testing on nearly 5,000 pairs from Anastasopoulos and Neubig (2020). This serves as a skyline to demonstrate the level of accuracy we might expect using higher-quality embeddings but still using a minimal amount of training word pairs. If we receive low accuracy on Meroitic/Egyptian but high

Type	# entries
nouns	26
names	18
numbers	15
verbs	14
titularies	9
adj/adv	6
prepositions	3

Table 2: Alignment dictionaries

accuracy on French/English, this suggests our problem lies in the sparsity/quality of embeddings.

6 Results

Overall, our intrinsic evaluation for both Meroitic and Egyptian embeddings shows promise. However, our lexicon induction experiments are found lacking. None of our models could correctly translate terms that had not been seen before, and of the terms that had been seen, only a maximum of 20% were correctly aligned.

6.1 Intrinsic Evaluation

We evaluate our embeddings by calculating the cosine similarity between known words.

Egyptian Overall, the Egyptian embeddings do very well on our tests considering the limited nature of the dataset. To begin with, lower numerals tend to be paired with low numerals (ex. 1, 5, 3, 6, 8), while high numerals match higher numerals (ex. 1000, 500, 800, 2000).

We find that words associated with kingship or gods often have high cosine similarities and often appear in the top 10 nearest neighbors of each other. For instance, /$nswt$/ ("king"), /r'/ ("Ra"), and /jmn/ ("Amun") all have over 92% cosine similarities with each other.

In addition, we perform several word analogy tests (similar to the famous "man" is to "woman" as "king" is to "queen" paradigm). Not all these are successful, but we do obtain certain interesting results:

- `nswt`→`wr` as `ms`→`b3ḳ`, meaning "'king' is to 'great' as 'child' is to 'servant,'" which is exactly something we might expect.
- `rmṯ`→`hm.t` as `nswt`→`mry`, which means "'man' is to 'woman' as 'king' is to 'beloved.'" Ideally, the result would be /$nswy.t$/, meaning 'queen,' but 'beloved' may still contain a relevant connotation; it should also be noted that /mry/ was often used in the context of a king's relationship to a god.

We provide cosine similarity scores for some selected Egyptian word comparisons in Table 3, and note that all scores are high.

Meroitic The Meroitic embeddings do not perform quite as well on the intrinsic evaluations, but we do find that they capture some semantic information. For instance, the embeddings of the numerals 2, 12, and 6 are all very near to each other, and the gods Isis and Osiris are similar—this in particular is expected since in mythology Isis is the wife of Osiris. Testing with /qor/[10] for "ruler" returned /$abrse$/ (a nominal group, meaning "every man" when containing an article: /*abr-se-l*/), /$qorte$/ (literally "in the king's," probably meaning "palace"), and /$amnp$/ ("the God Amun of Napata"). We also find certain titularies grouped with titularies, for example: /$perite$/ ("local official"), /$ttnylkh$/[11] (some official title), and a word seemingly related to /$pelmo\hat{s}$/[12], which has to do with regional military administration (Millet, 1968).

Additionally, variant word forms appear as nearest neighbors, for example /mni/ and /$mnpte$/ for "Amun" and "Amun of Napata," and /$(a)\hat{s}or(i)$/ and /$(a)\hat{s}oreyi$/ (vocative form) for Osiris. This gives hope to future orthographic variation detection efforts. Note that this is despite the fact that we use a method that does *not* take into account character n-grams (like `fasttext` would, much more suitable for modeling orthographic variation) and hence this confirms that these are indeed variants of the same word, as opposed to them being two distinct words with very similar forms.

Word analogy results prove difficult to analyze, since it is first more complicated to construct them with our limited vocabulary, and most of the words that are returned are unknown. However, one very good result within the top-10 turns out to be `qor`→`pqr` as `abr`→`yetmdelo`, which means "'ruler' is to 'crown prince' as 'man' is to 'nephew[13].'"

It should also be noted that unlike the Egyptian embeddings (whose nearest neighbors often had cosine similarity scores greater than 95%), many of

[10] written as 'qEr' in our corpus; all *o*'s are written as 'E', since some of Millet's publications transcribed as /*ê*/. However, it should be noted that *o* is the standard convention.

[11] written 'ttNlX' in our corpus

[12] written 'pelmES' in our corpus. The actual word returned was /$pelmo\hat{s}lis\underset{.}{p}\underset{.}{q}ebete.$/

[13] Note that technically /*yetmde-l-o*/ is a nominal clause meaning "he is the nephew" or "she is the niece"

Word 1		Word 2		Cosine
egy	en	egy	en	Similarity
r'	Ra	$n\underline{t}r$	god	**0.96**
r'	Ra	wsr	power	0.93
r'	Ra	$\underset{.}{h}m - n\underline{t}r$	priest	0.88
jmn	Amun	$\underset{.}{h}m - n\underline{t}r$	priest	**0.98**
$nswt$	king	stp	choice/elite	**0.96**
$\underset{.}{h}m.t$	woman	$rm\underline{t}.t$	woman	0.95
$\underset{.}{h}m.t$	woman	s	man	0.86
ms	child	$s3$	son	0.91
$\underset{.}{h}m.t$	woman	$sn.t$	sister	**0.98**
$'3$	large	wr	great	0.93

Table 3: Cosine similarity scores between Egyptian words. egy is the Egyptian word; en is its English translation. We choose words that we feel are related, so we get high similarity for the majority of tests. Notice that woman/man is slightly lower than the rest, which may be expected due to the difference in gender.

Word 1		Word 2		Cosine
xmr	en	xmr	en	Similarity
$qor(e)$	ruler	qr	ruler	**0.91**
$qor(e)$	ruler	pqr	prince	-0.04
$qor(e)$	ruler	mlo	head	**0.62**
kdi	woman	$kdileb$	women	0.42
kdi	woman	$sem(l)$	wife	0.44
kdi	woman	abr	man	0.522
kdi	woman	$kdis$	sister	0.46
dd	infant/son[14]	as	child	**0.69**
$kdis(e)$	sister	$wi(de)$	brother	-0.11
tr	big	lx	large/high	**0.73**

Table 4: Cosine similarity scores between Meroitic words. xmr is the Meroitic word; en is its hypothesized English translation. We choose words that we feel are related, so we would expect similarity to be high. However, while we do get some high scores, results are somewhat inconsistent.

these Meroitic "nearest neighbors" display cosine similarities below 60 or even 50%, indicating that related words are not as near to each other in the embeddings space. We believe this can be attributed to the extremely low training data. Nonetheless, we still present the cosine similarity scores between several known Meroitic words in Table 4. Some pairs have reasonably high scores, but the results are inconsistent.

6.2 Alignment Results

Our alignment results (Table 6 in Appendix B) are far from ideal. None of our Meroitic-Egyptian cross-lingual embeddings were able to do lexicon induction for a dictionary they had not seen before. Our best setting appears to be on numerals with 100-dimension vectors; however, even for the training dictionary they were not able to achieve more than 20% accuracy. In contrast, our French-English cross-lingual embeddings performed 70% on the training dictionary, and close to 68% on the test set.

The most obvious explanation for this poor performance is twofold. Firstly, our Meroitic-Egyptian test sets are so small that we cannot expect our models to correctly pair the specific words we have chosen. We should remember that the accuracy on these few words is not an indication of complete failure. However, the fact that we could not achieve better than 20% on the very words we aligned on is an indication that these embeddings are insufficient for proper alignment and lexicon induction. This is likely due to the extreme low-resource nature of the training sets, although it is possible that we may be able to achieve better accuracy when aligning Meroitic to a different language, such as Old Nubian or Coptic, despite the differences in content. One might also try with modern, higher-resourced languages, such as Hebrew or Egyptian Arabic; however, we could hardly expect these to bear any meaningful resemblance to the language in question.

7 Discussion and Future Work

Despite the extremely limited nature of our corpora, our embeddings are still able to capture semantic information. This is especially true in our Egyptian embeddings, but Meroitic also shows promise, suggesting that our corpus and embeddings can be useful for future experiments to further understand Meroitic. We believe the Egyptian embeddings were better due to the difference in example length; many Egyptian texts were equivalent to several paragraphs, but most of the Meroitic examples were short sentences or fragments, and heavily augmented using anthroponyms. Regardless, there is still a long way to go before achieving results that may be useful for scholars in any major decipherment effort, which is clear when considering the abysmal performance of our lexicon induction tasks. Future work should attempt the same alignment but with other languages, such as Coptic or Old Nubian. However, we believe the prime reasons for this is simply the lack of quality training data. If more Meroitic examples could be gathered and made machine-readable, then we could expand our corpus and obtain more reasonable results.

Other avenues for future work, now made possible with our new corpus, include cognate detection, orthographic variant recognition, NER tasks, and

POS-tagging. Additionally, Meroitic inscriptions tend to use substantially different vocabulary in different contexts. Thus, performing a study of lexical elements common to various genres would also be useful.

7.1 Cognate Detection

One important direction for Meroitic research includes attempting to find cognates in related languages. We hope to first benchmark methods similar to Snyder et al. (2010), Luo et al. (2019), and Luo et al. (2021) (see Section 3). The idea is to search for cognates in related languages by comparing their high-frequency word roots and particles with Meroitic's, based on phonetic values and overall frequency.

We present an initial attempt on cognate detection (by hand, as not all resources are digitized) for two common but unknown words in Appendix C.

7.2 Leveraging Related Languages

In contrast to previous machine translation attempts for language decipherment (Snyder et al., 2010), we currently know of no language that serves as a very close relative to Meroitic. However, because we can find similarities between Meroitic and other languages, such as Old Nubian, which shares with it both lexical and grammatical features (van Gerven Oei, 2020), the hope is that we could perform experiments using multiple semi-relatives, perhaps using methods established in Luo et al. (2021) (see Section 3), and combine the data to build a comprehensive understanding of the Meroitic language. At this stage, Old Nubian presents one of the most likely candidates for comparison, although unfortunately its content is primarily Christian-oriented (van Gerven Oei, 2020), contrasting sharply with Meroitic's Kushite Pantheon of gods, and its lexicon is more limited compared to modern Nubian dictionaries[15]. Hopefully there exists enough of a connection to use in computer analyses, but other Nilo-Saharan languages, such as Nara, Tama, and Dinka, may also be useful for comparison. Ideally, analysts would perform experiments on not merely one, but many languages, and use those results cumulatively to better understand Meroitic.

7.3 Handling Orthographic Variation

We suspect that orthographic variation may play a significant role in the quality of our embeddings, since each distinct form would erroneously appear to have an entirely new meaning. We plan to attempt the same experiments after modifying the training data to eliminate all known variants, similarly to methods used in Sahala and Lindén (2023). However, it is quite possible that many more variants exist than scholars have previously been able to uncover. One solution would be to compare the words in Meroitic texts of related genres with each other, either considering cosine similarity, or word frequency and phonemes, perhaps taking region and time-period into account as well; in this way we may guess which words are orthographic variants of each other. Seeing how words were written and therefore pronounced by different people might also give insight into Meroitic phonology and where language variations occurred, which would be important not only for knowledge of Meroitic, but for linguistics and history as well. Regardless, this test should improve our ability to read the Meroitic language, as it minimizes the number of terms that are truly unknown, and could lead to higher-quality embeddings.

Our current corpus provides the raw texts as they currently appear, i.e., including all the above-mentioned variations. But we hope to release a "normalized" version in the near future.

8 Conclusion

The use of computational methods to decipher Meroitic looks hopeful. Large-scale programs can search for cognates much more effectively than any human, and statistical brute-force comparisons can help to identify word roots and grammatical particles. Meroitic is an ideal language on which to attempt translation, as we already have some knowledge of vocabulary and grammar (albeit limited). The primary challenges will be finding the right language to effectively map word and particle meanings (Lobban Jr, 2003), paired with acquiring enough machine-readable data on both ends.

The corpus, embeddings, and analyses we present here constitute a step in that direction. Despite the disappointing results of our lexical induction tests, our embeddings appear to have the capacity to capture non-trivial semantic information. With additional attempts with other languages, as well as methods to handle orthographic vari-

[15] Modern Nubian dictionaries (e.g. Khalil (1996) and Armbruster (1965)) have many more words than the Old Nubian dictionary Browne (1996).

ation, perhaps we may achieve more promising results. Ultimately, decipherment of Meroitic–or any untranslated language–will require computer efficiency and persistence, paired with human ingenuity and intuition.

Limitations

Creating a machine-readable Meroitic corpus is not a trivial task. Firstly, the language is so obscure that it is difficult to obtain access to Meroitic materials, and putting them into machine-readable format requires extensive care and some expertise. Thus, we had to use some materials that were fairly old and may contain outdated transcriptions and translation hypotheses. However, we believe that even a possibly outdated machine-readable corpus is better than no corpus at all, and given some of our positive results for the intrinsic evaluation, it seems that what we do have is still worthwhile. We hope to eventually curate an up-to-date machine-readable corpus, perhaps based on the recent publication of Hallof (2024). Note, however, that this book is not currently available in any digital format, and our attempts at contacting the author have been unsuccessful. Should we manage to eventually obtain access to this book, it may also lead to substantial improvements in results.

Acknowledgments

This work was generously supported by the Presidential Scholarship awarded by the George Mason University Graduate Division. It was also partially supported by NEH Award PR-276810-21 and NSF Award IIS-2327143. Computational resources for experiments were provided by the Office of Research Computing at George Mason University (URL: `https://orc.gmu.edu`).

Special thanks to Dr. Richard Lobban Jr., Professor Emeritus at Rhode Island College, for his invaluable mentorship, access to his extensive library, and the sharing of his expertise on Meroë and ancient Nubia; and to Dr. Jeremy Pope from the College of William and Mary, for his guidance and inspiration throughout this project; also to Dr. Claude Rilly, director of research at the CNRS, the Sedeinga Excavations, and professor at École pratique des hautes études in Paris, France, for his thorough review and advice for this paper. Their expertise has proved invaluable for our research thus far.

References

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

Antonios Anastasopoulos and Graham Neubig. 2020. Should all cross-lingual embeddings speak English? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8658–8679, Online. Association for Computational Linguistics.

Charles Hubert Armbruster. 1965. *Dongolese Nubian: A Lexicon*. Cambridge University Press.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gerald Browne. 1996. *Old Nubian dictionary*. Peeters Publishers.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 261–270, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. arXiv:1710.04087.

Francis LLewellyn Griffith. 1911. *The Meroitic Inscriptions of Shablul and Karanòg*. Pennsylvania University.

Tomas Hägg. 2000. T. eide, t. hägg, rh pierce, l. török (edd.): Fontes historiae nubiorum, vol. iii. textual sources for the history of the middle nile region between the eighth century bc and the sixth century ad: From the first to the sixth century ad. pp. 751–1216. bergen: University of bergen, 1998. paper, nok 220. isbn: 82-91626-07-3. *The Classical Review*, 50(1):1103–1107.

Jochen Hallof. 2024. *Analytic Meroitic Dictionary*. J H Roll Verlag; Bilingual edition.

James Hoch. 2023. Egyptian language.

Inge Hofmann. 1998. Fontes historiae nubiorum: Textual sources for the history of the middle nile region between the eighth century bc and the sixth century ad. vol. ii: From the mid-fifth to the first century bc.

Mokhtar M. Khalil. 1996. *Wörterbuch der nubischen Sprache (Fadidja/MaḥasDialekt): Arbeitfassung/Vorabdruck*.

Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 499–506.

Jean Leclant, Claude Rilly, Catherine Berger-El-Naggar, Claude Carrier, and André Heyler. 2000. Répertoire d'épigraphie méroïtique.

Richard Lobban Jr. 1994. Problems and strategies in the decipherment of meroitic. *Northeast African Studies*, 1(2):159–164.

Richard A Lobban Jr. 2003. *Historical dictionary of ancient and medieval Nubia*, volume 10. Scarecrow Press.

Richard A Lobban Jr. 2021. *Historical Dictionary of Ancient Nubia*. Rowman & Littlefield Publishers.

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From Ugaritic to Linear B. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155, Florence, Italy. Association for Computational Linguistics.

Jiaming Luo, Frederik Hartmann, Enrico Santus, Regina Barzilay, and Yuan Cao. 2021. Deciphering undersegmented ancient scripts using phonetic prior. *Transactions of the Association for Computational Linguistics*, 9:69–81.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. arXiv:1301.3781.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting similarities among languages for machine translation. arXiv:1309.4168.

Nicholas Byram Millet. 1968. *Meroitic Nubia*. Yale University.

Monica Ouellette and Helene Longpre. 1999. *Thoth: Language Cognate Program*. Ph.D. thesis.

Stéphane Polis, Serge Rosmorduc, and Jean Winand. 27 August 2015. Ramses goes online. an annotated corpus of late egyptian texts in interaction with the egyptological community. F.R.S.-FNRS - Fonds de la Recherche Scientifique [BE].

Claude Rilly. 2007. *La langue du royaume de Méroé: un panorama de la plus ancienne culture écrite d'Afrique subsaharienne*. Honoré Champion.

Claude Rilly. 2008. Linguistic position of meroitic. new perspectives for understanding the texts. *The Sudan Archaeological Research Society*.

Claude Rilly and Alex de Voogt. 2012. *The Meroitic language and writing system*. Cambridge University Press.

Aleksi Sahala and Krister Lindén. 2023. A neural pipeline for pos-tagging and lemmatizing cuneiform languages. In *Proceedings of the Ancient Language Processing Workshop*, pages 203–212.

Wolfgang Schenkel. 1972. Meroitisches und baryaverb: Versuch einer bestimmung der tempusbildung des meroitischen. *Meroitic Newsletter*, 11:1–16.

Peter Lewis Shinnie. 1967. *Meroe: A civilization of the Sudan*, volume 55. FA Praeger.

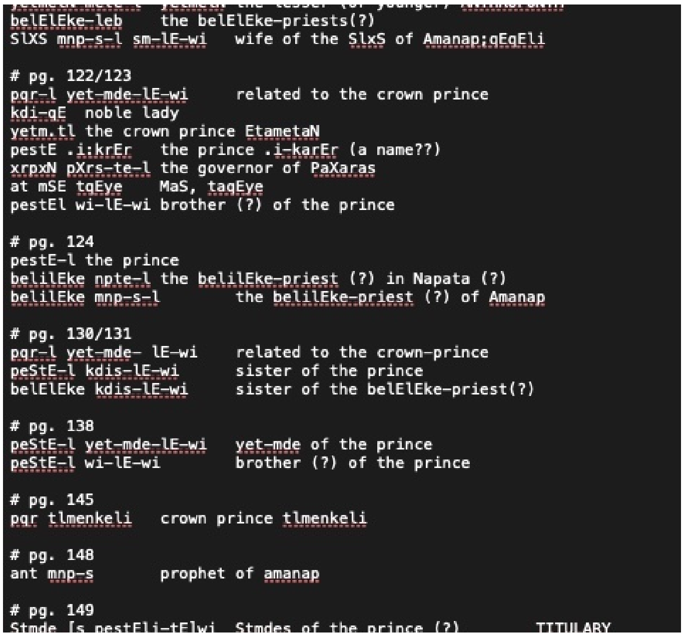
Reginald D Smith. 2008. Investigation of the zipf-plot of the extinct meroitic language. arXiv:0808.2904.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057.

Vincent van Gerven Oei. 2020. Old nubian crash course – day 1.

Meroitic Corpus Examples
plSn aqmks penn 5 ni ye teke lE
xbxN wES qer qE sskemxr qE wESi yntke pipl pxilX pli ptrEti pipn pbx
wErEteliye krErE
t dxe mlEqErebr qEre s l xrws
pestE aberEtemte pestE n. yetmde betewi

Table 5: Example lines from our corpus, where each line is a unique example.



```
belElEke-leb    the belElEke-priests(?)
SlXS mnp-s-l sm-lE-wi   wife of the SlxS of Amanap;qEqEli

# pg. 122/123
pqr-l yet-mde-lE-wi     related to the crown prince
kdi-qE  noble lady
yetm.tl the crown prince EtametaN
pestE .i:krEr   the prince .i-karEr (a name??)
xrpxN pXrs-te-l the governor of PaXaras
at mSE tqEye    MaS, taqEye
pestEl wi-lE-wi brother (?) of the prince

# pg. 124
pestE-l the prince
belilEke npte-l the belilEke-priest (?) in Napata (?)
belilEke mnp-s-l        the belilEke-priest (?) of Amanap

# pg. 130/131
pqr-l yet-mde- lE-wi    related to the crown-prince
peStE-l kdis-lE-wi      sister of the prince
belElEke kdis-lE-wi     sister of the belElEke-priest(?)

# pg. 138
peStE-l yet-mde-lE-wi   yet-mde of the prince
peStE-l wi-lE-wi        brother (?) of the prince

# pg. 145
pqr tlmenkeli   crown prince tlmenkeli

# pg. 148
ant mnp-s       prophet of amanap

# pg. 149
Stmde [s pestEli-tElwi  Stmdes of the prince (?)        TITULARY
```

Figure 2: Screenshot of lines from the Millet examples.

A Corpus Example

Table 5 displays example lines taken from our Meroitic corpus, and Figure 2 shows a screenshot image of the examples from Millet.

B Alignment Results

Table 6 shows the results from aligning Meroitic to Egyptian embeddings on numerals and nouns. The columns represent the cross-lingual embeddings, while the rows are the test dictionaries.

Dict	unsup	num-20	num-50	num-100	num-120	nn-20	nn-50	nn-100	nn-120
`numer`	-	6.67	13.33	**20**	13.33	-	-	-	-
`nouns`	-	-	-	-	-	8	4	**16**	12
`other`	-	-	-	-	-	-	-	-	-

Table 6: Results from Meroitic-Egyptian cross-lingual embeddings. The `-numbers` are the dimensions of the word vectors. `unsup` stands for an unsupervised model on 100-dimension vectors; `num-` models are aligned on numbers, and `nn-` are aligned on nouns. The results are abysmal, suggesting that we cannot reliably perform lexicon induction between the Meroitic and Late-Egyptian corpora.

C Preliminary Cognate Study

Once we had compiled the three long royal narratives in machine-readable format, we calculated overall word frequency within the texts. Then, consulting with an expert in Nubian and Meroitic history and languages, we focused on two of the most frequent words, and hand-identified possible cognates from related languages. Tables 7 and 8 show our results for the words /*seb*/ and /*kek*/, respectively. Current theories suggest that /*seb*/ is a noun related to kingship and that /*kek*/ may possibly be a coordinating conjunction (although this is fragile).

Word	**Meaning**	**Language**
seb	unknown	Meroitic
sab	cat	Nubian Kenzi/Dongolawi
esbyni	villager	Nubian Kenzi/Dongolawi
sablo	waterfall	Nubian Kenzi/Dongolawi/Fadija/Mahas
sablo	obstruction to the flow, irrigation canal	Nubian Kenzi
sablo	trough (especially for a waterwheel)	Nubian Dongolawi
sib	to fly	Nubian Kenzi/Dongolawi
sab	clouds	Nubian Dongolawi/Mahas
sabe	wall	Nubian Kenzi/Dongolawi
saab	downstream end	Nubian (17th century)
asab	sinew/muscle	Nubian
seb	intelligent	Coptic
sabat	basket	Old Nubian

Table 7: Hand-identified possible cognates/borrowings for the Meroitic word /*seb*/.

Word	**Meaning**	**Language**
kek	Unknown	Meroitic
kakke	small scorpion	Nubian Kenzi
kok	hammer	Nubian Kenzi/Dongolawi/Fadija/Mahas
kuk	to hatch	Nubian Kenzi/Dongolawi/Fadija/Mahas
kk	darkness	Egyptian
ukk	wean	Nubian Kenzi/Dongolawi/Fadija/Mahas
kkki	lineage name of island land cultivators	Nubian
kak	room	Nubian
kikko	chop	Nubian
Kuk/*Kek*	God of darkness	Egyptian
Keket	Goddess of darkness	Egyptian

Table 8: Hand-identified possible cognates/borrowings for the Meroitic word /*kek*/.

Interestingly, the cognates found for these two words do not appear to directly support the current scholarly theories. Based purely on these results, any hard translation for /*seb*/ or /*kek*/ would be speculative. However, there appears to be somewhat of a theme regarding “earth,” “wall,” “blockage,” “water,” or “cataract” relating to the word /*seb*/. Therefore, considering the importance of the Nile in Meroë geography and culture, one possibility is that this Meroitic word means or is related to a cataract. For /*kek*/, many of the Nubian/Egyptian words appear to have a dark or destructive connotation, so one possibility is that /*kek*/ means “to cut” or perhaps “hurt,” “hit,” or “break.” This also makes sense in context of conquest, which is likely a prevailing theme in the royal narratives.

The hope is that once we acquire enough data from languages in addition to Meroitic, we will be able to automate the process of cognate detection. In future work, we also expect to take into account word clusters and *n*-grams.