Constructing Word Similarities in Meroitic as an Aid to Decipherment

Reginald Smith

Constructing Word Similarities in Meroitic as an Aid to Decipherment

Reginald Smith

This paper addresses a technique using a combination of known words, and methods from information theory, to attempt decipherment of additional words in the extinct and as yet undeciphered phonetic script, Meroitic. A short history of the script, and the problems translating it, will be followed by a description of the statistical techniques, their results and implications.

A Short History of Meroitic

Meroitic was the written phonetic script of the ancient civilization of Kush, located for centuries in what is now Northern Sudan. The word ‘Meroitic’ derives from the name of the city Meroë, which was located on the east bank of the Nile, south of where the Atbara River flows off to the east. It is the oldest written script in Africa other than Egyptian hieroglyphs and the related hieratic and demotic scripts. It has a hieroglyphic form using some adopted Egyptian signs and a cursive form similar to demotic. The script had one innovation uncommon in ancient written scripts, such as Egyptian hieroglyphics or Greek, in that there was a word separator, similar in function to spaces in modern scripts, that looks similar to a colon (fig.1). Meroitic is first attested in the 2nd century BC and was continuously used until the fall of Meroë in the mid-4th century AD.

The script was rediscovered in the 19th and 20th centuries as Western archaeologists began investigating the ancient ruins of northern Sudan. The first substantial progress in deciphering Meroitic came around 1909 when British archaeologist Francis Llewellyn Griffith was able to use a barque stand bearing the names of Meroitic rulers in both Meroitic and Egyptian hieroglyphs. The Meroitic hieroglyphs were then corresponded to the Meroitic cursive script, allowing the transliteration of Meroitic (fig.1). Some vocabulary was later deciphered by scholars including loan words from Egyptian, gods, names, honorific phrases, and other common words. However, the script remains largely undeciphered. The greatest hope for decipherment, an inscription similar to the Rosetta Stone, containing writing in Meroitic and a known language such as Egyptian, Greek, Latin, or Axumite, has yet to be found. Further confounding research is the debate regarding which language family Meroitic belongs to. Cognate analysis has proceeded extremely slowly due to the dispute as to which language family Meroitic properly belongs.

There is the possibility of a recent breakthrough in the categorization of Meroitic. Rilly has used evidence from some Meroitic script remnants, as well as linguistic comparisons, to show that the language likely belongs within the northern branch of the Eastern Sudanic language family (fig.3). In particular, though no existing Nubian language is a direct descendent of

1 Török, The Kingdom of Kush; Lobban, Historical Dictionary.
2 Rilly, ARKAMANI; Rilly, Journal des Africanistes 76, 63; Rilly, La langue du royaume de Meroë.
Meroitic, by constructing a proto-Eastern Sudanic (Northern) vocabulary, we may be able to then correlate and decipher some unknown words. For example, in the text REM 1165, the image of a dog and his work has shown that *wle* is likely the Meroitic term for dog, similar to the Dongola word *wel* for dog.

**Past Statistical and Mathematical Work on Meroitic**

Meroitic was one of the earliest ancient scripts to be investigated using computers. Much of this work was dedicated to creating an alphabetical index of Meroitic and also comparing Meroitic words to possible cognates in Nubian or other known ancient and modern languages from the region. Smith analyzed many of the longest texts by ranking words according to frequencies, to verify whether the current texts we have follow the mathematical relation known as Zipf’s Law, where the word frequencies \( f \) vary with the rank \( z \) according to the relation:

\[
 f_z = \frac{C}{z^\alpha}, \quad z = 1, 2, 3 \ldots n
\]

where \( \alpha \approx 1 \). In analyzing the Meroitic texts, though many did not fit the strict criterion of \( \alpha \approx 1 \), the frequency-rank distribution followed the behaviour of a truncated power law distribution whose exact parameters varied by text. Some texts, such as the long stela REM 1003, more closely fit Zipf’s Law. From these results, without knowing the meaning of the text it is clear that the statistical variations and occurrences of words in the Meroitic texts in our possession are not surprising and mirror those of other human languages. Though this may seem a trivial property at first glance, it gives us the hope of using more advanced statistical techniques to help tease meaning from the unknown portions of the script.

**Introduction to Statistical Techniques**

It is evident that no language has ever been fully deciphered using purely statistical or mathematical techniques, and Meroitic will of course never be completely understood using these tools alone. In particular, many of the subtleties of human semantics and syntax are irregular or do not follow consistent patterns, which statistics would be excellent at analyzing. This paper will attempt to find words which are used very similarly in the text, rather than seek to derive the meaning of those words (a loaded concept in the study of linguistics). When two words are used very similarly with one of the words being known, we can hope to possibly infer what the other word means. In linguistics, the hypothesis that words that appear in similar contexts have similar semantics is known as the Distributional Hypothesis.

Similarity, which will be explained in more technical detail below, will be defined by looking at whether two different words share similar word neighbors within a distance of one or two

---

3 REM refers to the *Répertoire d'épigraphie méroïtique*, the most comprehensive catalogue of Meroitic texts.


5 Smith, *Glottometrics* 15, 53.

words away. The steps in analyzing the similarity are five-fold. First, I combined several long Meroitic texts into one giant corpus. I separated out some common bound morphemes to help better identify particular words. Second, I used a computer program in Python to create three matrices: one showing the relative frequency of each word, one showing the frequency of a given word pair (\textsc{word1}:\textsc{word2} for any combination of the distinct words in the text for a word distance of one), and a final array with word pair frequencies for a word distance of two. Third, for all possible pairs of different words in the texts, I used the frequency arrays to find the mutual information between every distinct word pair. I created separate arrays of the mutual information metric for the mutual information based on word distance one and mutual information based on word distance two. A blended mutual information was then calculated, based on weightings of the one and two word distance mutual information. Fourth, using the blended mutual information array, I used a similarity metric to find similarity between words on the basis that they had similar mutual information for the other words in the texts. Finally, I compared the results for high similarity word pairs to what is known about Meroitic words. A minimum spanning tree graphically showing the relationship between words was also aided to clarify the similarity relationships.

Step 1

The long stelae texts REM 1001, REM 1003, and REM 1044A-D were combined into one corpus separated by a character XXXX between the beginning and end of each text. The XXXX made sure that the last word of one text and the first word of another are not accidentally matched for either a distance one or two word pair. In addition, several common and recognized bound morphemes were separated from the words by the word separator character so they would be treated as separate words. Many Meroitic verbs, as well as some nouns, have suffixes which contain grammatical meaning. For example, it is known that the suffix \textsc{telowi} or \textsc{tel} is appended to the name of a place, such as a city, to indicate that the subject of the sentence was affiliated with this place. There is also an extremely common suffix \textsc{lowi} ("he/she/it is") or \textsc{li} ("the") that is appended to nouns that may denote the noun as an indirect object in the sentence. Though their definitions are still tenuous, however, these bound morphemes are very common and were separated into independent words for the second Zipf plot. The six bound morphemes separated out were \textsc{qo}, \textsc{lo}, \textsc{li}, \textsc{te}, \textsc{lebkwi}, \textsc{mbe}. They were separated in the manner:

\begin{align*}
\text{qo} & \rightarrow \text{separated out to 'qo'} \\
\text{atomb} & \rightarrow \text{ato and mbe} \\
\text{telowi} & \rightarrow \text{te and lo and wi} \\
\text{li} & \rightarrow \text{separated out to 'li'} \\
\text{qowi} & \rightarrow \text{qo and wi} \\
\text{lebkwi} & \rightarrow \text{lebk and wi} \\
\text{lw} & \rightarrow \text{separated out to 'lw'} \\
\text{lo} & \rightarrow \text{separated out to 'lo'} \\
\text{atmb} & \rightarrow \text{at and mbe} \\
\text{tel} & \rightarrow \text{te and li} \\
\text{lowi} & \rightarrow \text{lo and wi}
\end{align*}

\footnote{As in Smith, \textit{Glottometrics} 15, 53.}

http://www.britishmuseum.org/research/online_journals/bmsaes/issue_12/smith.aspx
Step 2

The word frequency arrays were created as follows. First, a normalized frequency of each different word in the text was calculated ranging between 0 and 1, where the total frequency of a word divided by the total number of words in a text defines the word frequency. To understand word pair frequency, imagine a string of words separated by the colon-like word separator character, A:B:C. B/C and A/B are distance one neighbors and A/C are distance two neighbors. This is repeated for all words throughout the text. The frequency of a word pair is the number of occurrences of that pair divided by the total number of word pairs in the text.

Step 3

Here the procedure becomes more complicated and theoretical so the appropriate background is necessary. Many statistical natural language methods for analyzing corpora, such as hidden Markov models (HMM) or neural networks, require ‘training’ with a tagged corpus that emphasizes parts of speech and grammar. Since these are mostly unknown for Meroitic, we are forced to rely on techniques that make no a priori assumptions about the language syntax or word relationships.

Two relatively similar approaches relying on the Distributional Hypothesis were employed in several other works, using genetic algorithms\(^8\) and similarity measures\(^9\) to find relationships between words based on their distributions within a text. In Lankhorst, a fixed number of categories is created and each word is randomly assigned a category. The mutual information among words in each category is measured and the categories are altered using a genetic algorithm with mutual information as the fitness. A maximum mutual information is asymptotically approached after a certain number of generations, and the word/categories at this point typically reflect known grammatical categories. Word synonyms are discovered in a text by taking the similarity among words based on the mutual information between the two words and other words in the text.\(^{10}\) Those words who have the highest similarity are often semantically similar.

The approach in this paper most closely follows that of Lin et al. in finding the mutual information amongst words in the corpus and then computing a similarity between the words based off of this. The mutual information between two words in the text, \(x\) and \(y\), is termed \(I(x,y)\) and is defined as

\[
I(x, y) = \sum_x \sum_y p_{xy} \log \frac{p_{xy}}{p_x p_y}
\]

where \(p_{xy}\) is the frequency of word pair \((x, y)\) and \(p_x\) and \(p_y\) are the frequencies of words \(x\) and \(y\) in the texts. Two different arrays of mutual information were calculated for the word distance one and two pair frequencies. Finally, a blended mutual information is calculated

---

\(^8\) Lankhorst, Automatic Word Categorization with Genetic Algorithms.


using different weightings of the one and two distance mutual information.

The blended mutual information, $I_B$, is

$$I_B = \sqrt{I_1^2 + (WI_2)^2}$$

where $I_1$ and $I_2$ are the mutual information for distance one and two word pairs, respectively and the weight ($W$) takes a value between 0 and 1. It is difficult to find an objective value for $W$. The method used in the paper, which will be explained more in the next section is that different values of $W$ were tested until many known words with similar meanings had high measures of similarity. Though this could be accused of affirming the consequent, it can be considered a method of calibration based on our small current knowledge.

**Step 4**

For the blended mutual information the cosine similarity measure, $S$, was calculated where $S$ is defined as

$$S_{xy} = \frac{2I_B(x,z)I_B(y,z)}{I_B(x,z)^2 + I_B(y,z)^2}$$

where $z$ is all words in the corpus where $z \neq x, y$.

**Step 5**

The word pairs are ranked by descending similarity and the results analyzed. Since relatively infrequent words will likely give spurious or insignificant results, only word pairs where both words appeared at least three times were used in the final analysis for comparison. In table 1, the top word pairs by descending similarity are shown. A similarity cutoff of 0.95 was used given the clustering of words above 0.95 and the poor matching of known words and wider spread of similarity scores for word pairs with a score under 0.95. The value of $W$ used is 0.75. This value was chosen because of the excellent and high similarity match of the first two word pairs, which consist entirely of known words with similar meanings. The following words in the ranking also show promise. The word *kek* is still undeciphered but may likely have a religious meaning, given its tight similarity with *mk* and its appearance before *wosqol* (*wos* being Isis) in REM 0075. However, it is difficult to tell whether it is a noun or adjective, since it never occurs with article or adjective suffixes. This could reinforce the opinion of some, such as Hoffman, that it is a conjunction of some sort, but since it appears only in three documents (REM 0075, REM 1044A, REM 1044B) we may not have a large enough sample. If a noun, it could be a word such as ‘soul’ (Egyptian *ka*) or the name of a Meroitic deity. The word *seb* is well-known among Meroitic scholars to have a religious meaning, possibly the name of a deity, but the exact meaning is still unknown. The word *abresel* means ‘every man’. Though *wwikewi* isn’t specifically understood, it is known from comparison and inference
from uses in other texts\textsuperscript{11} that \textit{wui-} is a stem, which together with the particles \textit{-ke-wi} may indicate directional movement or reference. One appearance of \textit{wukewi} in REM 1003 states \textit{teneke}\\textit{wukewi} where \textit{teneke} means ‘west’.

In order to more clearly see how the words relate to each other, I graphically visualized the similarity relationships using the distance metric derived by Gower\textsuperscript{12}. This distance metric is used to convert comparison metrics such as correlation or similarity among variables to metric distances:

\[ d_{ij} = \sqrt{2(1-s_{ij})} \]

where \(d_{ij}\) is the similarity between words \(i\) and \(j\). These distances can then be plotted onto a minimum spanning tree (fig. 2).

**Problems and Issues**

As stated before, I cannot claim to solve the issues related to Meroitic solely through statistical analysis. In particular, though the information such an analysis can provide is directional, it is sensitive to interpretation. The choice of the weight, \(W\), though not completely arbitrary, uses \textit{a priori} knowledge to set its value. While the results it returns are consistent with closely related known words, this may introduce bias. The cutoff for the similarity measurement, at a value of 0.95, is also arbitrary and based on a subjective analysis of the data. Therefore, despite the equations, much of this technique requires knowledge of the script and subjective interpretation to extract useful knowledge. In the end, however, I believe this technique will help shed a light on many previously intractable problems in Meroitic and could become a valuable tool in the eventual decipherment of the script.

**Acknowledgements**

The author would like to thank Claude Rilly, Laurance Doyle and Richard Lobban for helpful comments on a draft of this paper.

**Cover image.** Sandstone offering table with inscription in cursive Meroitic. From Faras. British Museum EA 1576.

**Bibliography**


Heyler, A., ‘Essai de Transcription Analytique des Textes Meroitiques Isoles’, \textit{Meroitic Newsletter}

\textsuperscript{11} Hoffmann, \textit{Material für eine meritiische Grammatik}, 310.

\textsuperscript{12} Gower, \textit{Biometrika} 53, 325.
Fig 1: Meroitic Cursive and Hieroglyphic words and their transliterations. Taken from the latest font set for Meroitic Hieroglyphic and Cursive characters developed by the Meroitic scholars Claude Carrier, Claude Rilly, Aminata Sackho-Autissier, and Olivier Cabon. www.egypt.edu/etausi/informatique/meroitique/meroitique01.htm

Fig 2: Graphic representation of the minimum spanning tree of the data represented in table 2 as well as some lower similarity word pairs found in the study. Size of nodes is only for displaying text and has no other significance.

http://www.britishmuseum.org/research/online_journals/bmsaes/issue_12/smith.aspx
Fig. 3: Eastern Sudanic (Northern) language family with Meroitic inserted according to the research of Rilly. (LVBN* = a possible vestigial language of Lower Nubia.).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 1 Meaning</th>
<th>Word 2 Meaning</th>
<th>Similarity</th>
<th>Word 1 Count</th>
<th>Word 2 Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>kdi</td>
<td>abr</td>
<td>women</td>
<td>man</td>
<td>1.000</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>mk</td>
<td>amnp</td>
<td>god</td>
<td>Aman of Napata</td>
<td>0.999</td>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>mk</td>
<td>kek</td>
<td>god</td>
<td>?</td>
<td>0.998</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>abrsel</td>
<td>wwikwi</td>
<td>every man</td>
<td>?</td>
<td>0.996</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>qorte</td>
<td>agro</td>
<td>in the king?</td>
<td>?</td>
<td>0.986</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>amnp</td>
<td>seb</td>
<td>Amun of Napata</td>
<td>?</td>
<td>0.979</td>
<td>17</td>
<td>15</td>
</tr>
<tr>
<td>7</td>
<td>qes</td>
<td>qor</td>
<td>Kush</td>
<td>king</td>
<td>0.978</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>ne</td>
<td>pqr</td>
<td>?</td>
<td>prince</td>
<td>0.976</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>mk</td>
<td>seb</td>
<td>god</td>
<td>?</td>
<td>0.970</td>
<td>7</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 1: Top word pair similarities with meanings where known.