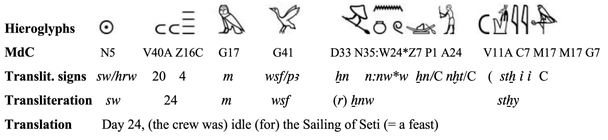
Automatic Word Segmentation for Egyptian Hieroglyphic Texts

Jauhiainen, Heidi
heidi.jauhiainen@helsinki.fi
University of Helsinki, Finland

Jauhiainen, Tommi
tommi.jauhiainen@helsinki.fi
University of Helsinki, Finland

In order to use digital methods for researching texts, machine-readable texts are needed. Assyriology, for example, has freely downloadable corpora of machine-readable texts, such as Open Richly Annotated Cuneiform Corpus (ORACC). The scarcity of similar corpora for Egyptian hieroglyphic texts hinders the digital study of ancient Egypt using textual sources.

There is a tradition in Egyptology of using encoding to produce properly positioned hieroglyphs for print (Fig. 1). Various versions of the encoding have been used when publishing texts in books, the so-called Manuel de Codage (MdC) being the most often used. The encoding uses letter-number combinations from the Gardiner list, a standard reference list for Ancient Egyptian hieroglyphs (Gardiner 1957), where letters refer to various categories of signs and numbers to signs within the category (Fig. 1). The encoded texts, which are used for producing pictures of hieroglyphic texts, are machine-readable. However, Egyptologists cannot usually “read” the encoding, which explains why these have not been considered important enough to publish or even to keep. When Egyptologists encode texts, the various hieroglyphic text editors offer them the possibility of writing many signs using the transliteration they are more familiar with. Transliteration is always an interpretation of the text, and producing it is a slow endeavor requiring checking dictionaries and sign lists. Computer-assisted transliteration of hieroglyphic texts will speed up producing such texts for digital methods, which is our future aim.



Hieroglyphs						
MdC	N5	V40A Z16C	G17	G41	D33 N35:W24*Z7 P1 A24	V11A C7 M17 M17 G7
Translit. signs	*sw/hrw*	20 4	*m*	*wsf/pꜣ*	*ẖn* *n:nw*w* *ẖn*/C *nẖt*/C	(*stẖ ỉ ỉ* C
Transliteration	*sw*	24	*m*	*wsf*	(*r*) *ẖnw*	*stẖy*
Translation	Day 24, (the crew was) idle (for) the Sailing of Seti (= a feast)					

Figure 1: A segmented hieroglyphic sentence with encoding (MdC), transliteration alternatives for each sign, transliteration of the sentence, and translation. C = classifier (without phonetic value).

As the word boundaries are not natively indicated in a hieroglyphic text, the first task is word segmentation. Word segmentation methods used for Chinese and Japanese texts have been tested on texts written with cuneiform signs (Homburg / Chiarcos 2016). The best-performing methods in that study were based on dictionaries. We have already generated language models of words from two available corpora of machine-readable hieroglyphic texts. Thesaurus Linguae Aegyptiae (TLA) includes a collection of texts where c. 280,000 Egyptian encoded words have been aligned with their transliteration counterparts (Richter et al. 2018; Schweitzer 2021). The second source is the Ramses Transliteration Corpus (RTC) published in 2021 with almost 500,000 encoded words (Rosmorduc 2021). These language models can be used with the aforementioned dictionary-based methods.

The RTC consists of encoded hieroglyphic sentences, each on its own line, and respective transliteration lines in another file. Since the RTC data was originally collected in the Ramses Project (Polis / Winand, 2013) and is available for word searches online, it contains, in addition to sentences without word boundaries, also separate versions where the encoded words have been separated with underscores.

We began by testing the best method Homburg and Chiarcos found: MaxMatch. MaxMatch provided an F-score of 65.05 in their Middle Babylonian cuneiform test set. Our implementation of MaxMatch gained an F-score of 65.95 on the RTC validation partition. Next, we created three versions of their prefix/suffix algorithm. Their algorithm only predicts separation after the character under scrutiny, and for that, we got an F-score of 36.28. We also implemented a version that predicted a break before the current character (37.60). Finally, we modified the algorithm so that instead of characters, we were focusing on between the characters and arrived at an F-score of 50.47.

Then we implemented the bigram method described by Homburg and Chiarcos and, to a small surprise, arrived at an F-score of 75.09, which is clearly higher than for the other algorithms and quite the opposite from Hamburg and Chiarcos's findings. As bigrams fared so well, we implemented a similar system that used sign 4-grams and arrived at an F-score of 48.06. These findings were in line with our previous experiences with character n-gram-based language models from our language identification experiments. Following our HeLI method (Jauhiainen et al. 2016), we created a back-off scheme to start from 4-grams with the ability to back off to bigrams, to our improved prefix/suffix method, and even to the simple probability of separation. This back-off scheme proved very efficient, providing an F-score of 86.25, which is clearly better than the 65.05 that was achieved by Homburg and Chiarcos for Middle Babylonian.

The source code of the segmentation method will be released with an open-source license. A segmentation tool will also be published for others to use on their texts.

Bibliography

Alstola, Tero / Zaia, Shana / Sahala, Aleksi / Jauhiainen, Heidi / Svärd, Saana / Lindén, Krister (2019): “Aššur and His Friends: A Statistical Analysis of Neo-Assyrian Texts”, in: *Journal of Cuneiform Studies*, Vol. 71: 159–180.

Gardiner, Alan H. (1957): *Egyptian Grammar: Being an Introduction to the Study of Hieroglyphs*. Oxford: Griffith Institute .

Homburg, Timo / Chiarcos, Christian (2016): “Word Segmentation for Akkadian Cuneiform”, in:

Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). https://aclanthology.org/L16-1642.pdf [27.04.2023].

Polis, Stephane / Winand, Jean (2013): “The Ramses Project: Methodology and Practices in the Annotation of Late Egyptian Texts”, in: Hafeman, Ingelore (ed.): *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften 81–108.

Rosmorduc, Serge (2021): “Ramses automated translitteration software”, in: *Lingua Aegyptia* (2021-06-15, Vol. 28, 233–257). Zenodo. https://doi.org/10.5281/zenodo.4954597 [27.04.2023].

Schweitzer, Simon (2021): *AES - Ancient Egyptian Sentences; Corpus of Ancient Egyptian sentences for corpus-linguistic research*. GitHub. https://github.com/simondschweitzer/aes [27.04.2023].

Svärd, Saana / Alstola, Tero / Jauhiainen, Heidi / Sahala, Aleksi / Lindén, Krister (2020): “Fear in Akkadian Texts: New Digital Perspectives on Lexical Semantics”, in: Hsu, Shih-Wei / Llop-Raduà, Jaume (eds.): *The Expression of Emotions in Ancient Egypt and Mesopotamia*. Leiden: Brill 470–502.

Richter, Tonio S. / Hafemann, Ingelore / Fischer-Elfert, Hans-Werner / Dils, Peter (eds.) (2018): *Teilauszug der Datenbank des Vorhabens ‘Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache’ vom Januar 2018*. Akademienvorhaben Strukturen und Transformationen des Wortschatzes der ägyptischen Sprache. Text- und Wissenskultur im alten Ägypten. [27.04.2023].