From Sherds of Pottery to Open Egyptological Data

Jauhiainen, Heidi

2022-12-02

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI
HUMANISTINEN TIEDEKUNTA
HUMANISTISKA FAKULTETEN
FACULTY OF ARTS

KONE FOUNDATION

FROM SHERDS OF POTTERY TO OPEN EGYPTOLOGICAL DATA

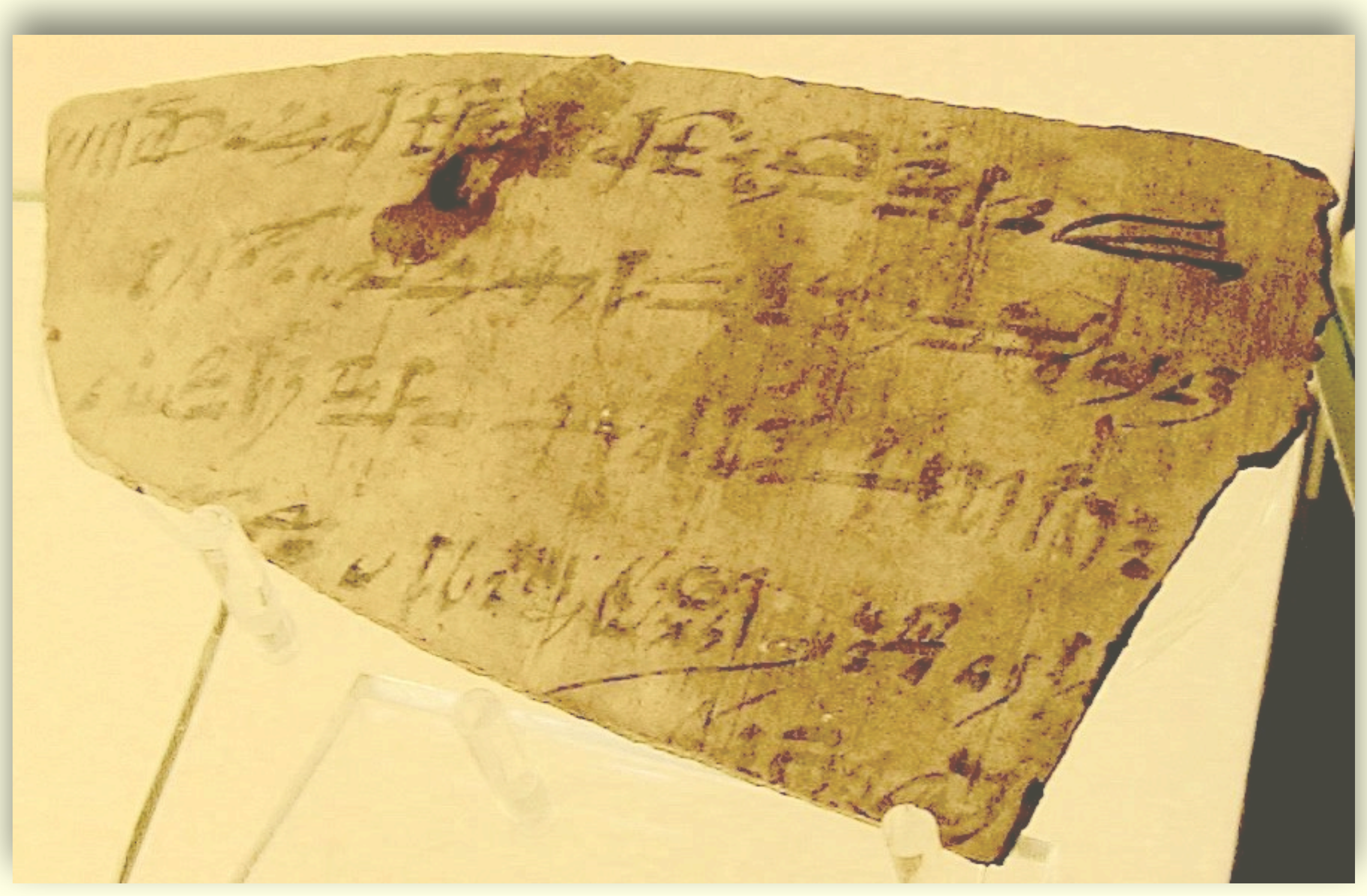
Heidi Jauhiainen
Department of Digital Humanities
heidi.jauhiainen@helsinki.fi

MaReTE

In Egyptology, there is no established way of publishing hieroglyphic texts in machine-readable form, and there are only a few ancient Egyptian text corpora. **Machine-Readable Texts for Egyptologists project** aims to publish openly available machine-readable ancient Egyptian texts, the absence of which hinders the development of digital Egyptology.

In 2020, the Finnish Cultural Foundation granted funding for the first preparatory year of the project, and work began at the beginning of 2021. The second part of the project is called **From Sherds of Pottery to Open Egyptological Data**, and it started in 2022 with funding from the Kone Foundation (2022–2024).

The research is carried out at the Department of Digital Humanities at the University of Helsinki and in connection with the Academy of Finland Centre of Excellence in Ancient Near Eastern Empires (2018–2025).

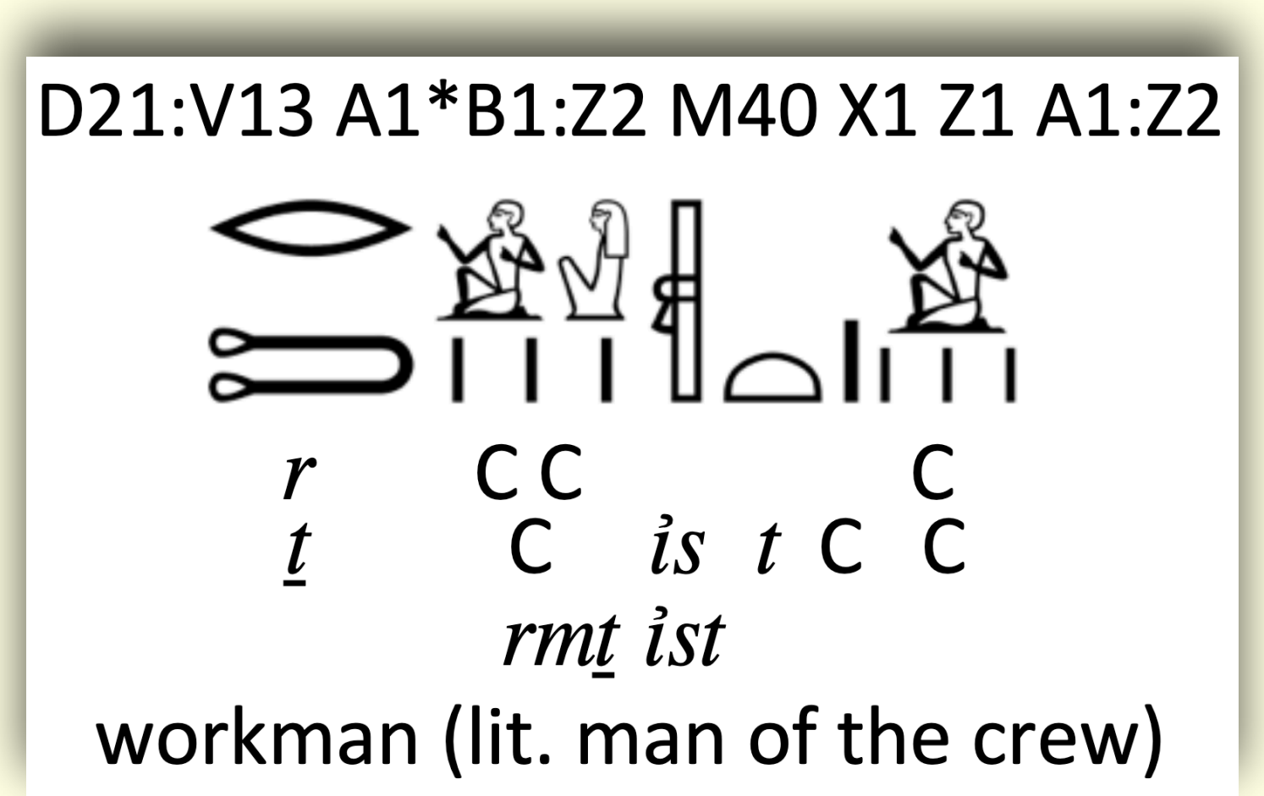


A hieroglyphic text written with cursive handwriting called Hieratic on a sherd of pottery. O. Turin N 57458, from Deir el-Medina, New Kingdom (c. 1550-1069 BCE). In Museo Egizio, Turin.

HIEROGLYPHIC TEXTS

In a hieroglyphic text, the signs can be next to, above, or over another, or even nested. There is, therefore, a tradition of using hieroglyphic text editors, where the signs are encoded, to maintain the information on the signs themselves and their places relative to each other when preparing hieroglyphic texts for printed publications. The encoding uses letter-number combinations from the Gardiner list, a standard reference list for Ancient Egyptian hieroglyphs, where letters refer to various categories of signs and numbers to signs within the category.

When Egyptologists study the texts written with hieroglyphs, they transliterate them with Latin letters and diacritics. Egyptologists cannot usually "read" the encoding, which explains why these have not been considered important enough to publish. When Egyptologists encode texts, the various hieroglyphic text editors offer them the possibility of writing many signs using the transliteration that they are more familiar with.


D21:V13 A1*B1:Z2 M40 X1 Z1 A1:Z2
r
C C
C
t̲
C
ỉs
t
C
C
rmt̲ ỉst
workman (lit. man of the crew)

An example of a hieroglyphic phrase produced with encoding (top row). After the hieroglyphic writing, there are the transliterations of individual signs (C = classifier), transliteration, and translation.

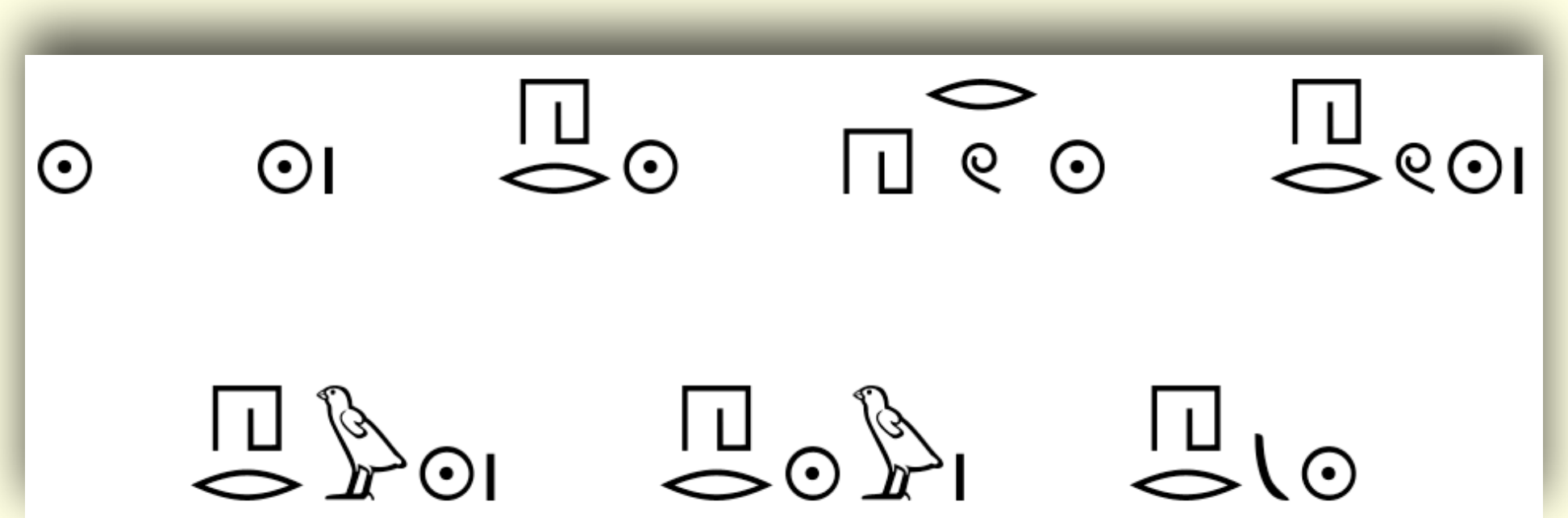
Transliteration is always an interpretation of the text, and producing it is a slow endeavor that requires checking dictionaries and sign lists. A method of **computer-assisted transliteration** would speed up the process of producing machine-readable hieroglyphic texts and make them more accessible to Egyptologists. Various tools to turn the encoded texts into a structured, machine-readable form will also be published (Jauhiainen, 2022).

SEMI-AUTOMATIC TRANSLITERATION

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.

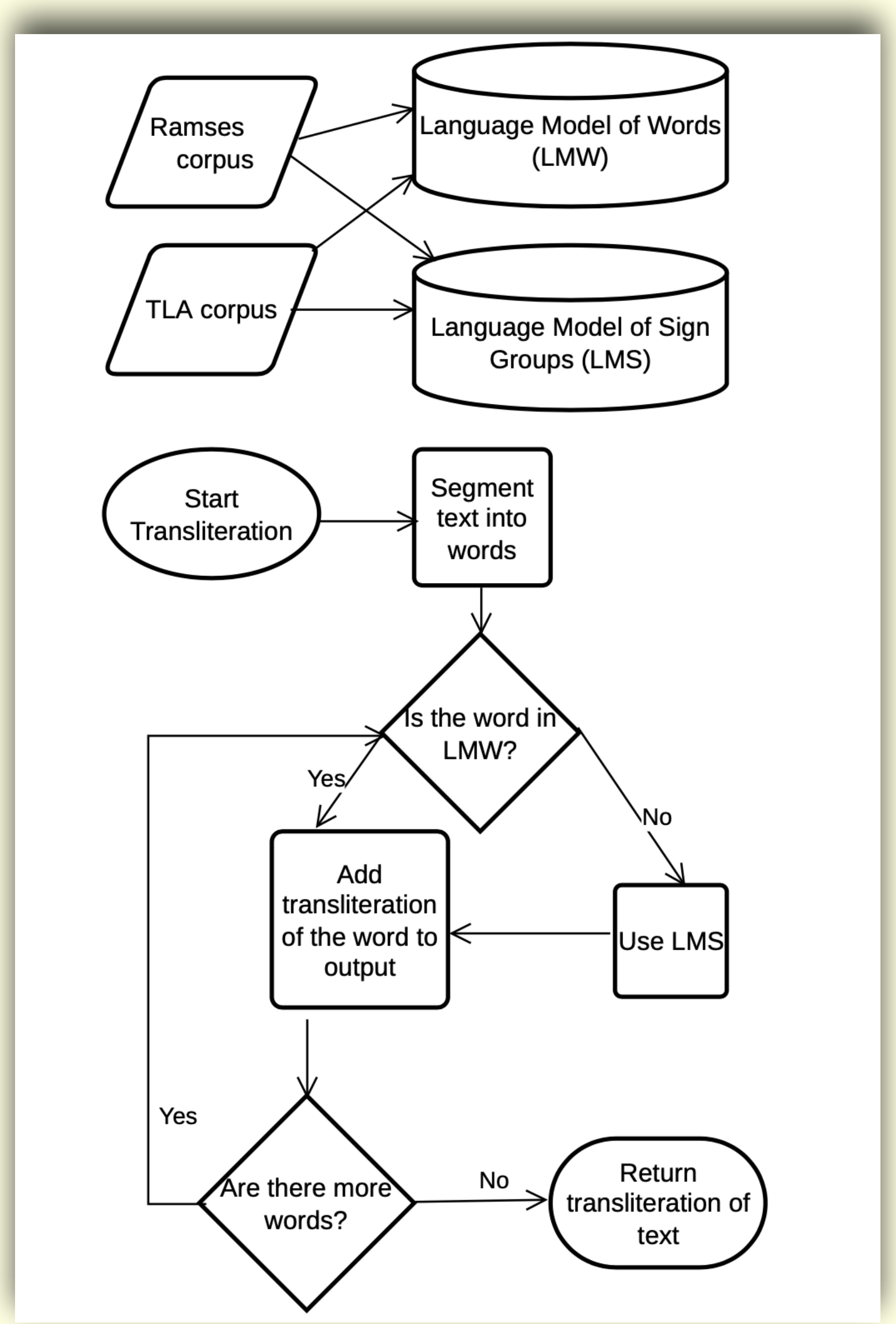
I have already generated **language models of words** from two available corpora of machine-readable hieroglyphic texts. Thesaurus Linguae Aegyptiae (TLA) includes a collection of texts where c. 280,000 words have been annotated with both transliteration and encoded hieroglyphs (Schweitzer, 2021). The second source is the Ramses Transliteration Corpus (Ramses) published in 2021 with almost 500,000 encoded words (Rosmorduc, 2021). These language models can be used with the aforementioned dictionary-based methods.

The transliteration method will, in its basic form, use the language model of words. However, the word forms found in the text to be transliterated might not be in the language model as Egyptian words tend to be written in multiple ways. Therefore, language models of sign groups will also be used.



Various different ways of writing the word *ḥrw*, **'day', with hieroglyphs.**

It cannot be expected for any automatic transliteration method to produce perfect transliteration as even human experts disagree. The proposed method will be sufficient for releasing large corpora of transliterated hieroglyphic text in a raw format, but for publishing transliterated texts they must be checked and corrected.


Ramses corpus
Language Model of Words (LMW)
TLA corpus
Language Model of Sign Groups (LMS)
Start Transliteration
Segment text into words
Is the word in LMW?
Yes
No
Add transliteration of the word to output
Use LMS
Yes
Are there more words?
No
Return transliteration of text

The method of computational transliteration visualized as a flowchart.

A platform for manually evaluating and correcting the automatically produced transliterations is planned. Using the platform, the language models and the segmentation algorithm can be automatically updated with information, such as new words, from the corrected transliterations. The method will be tested and improved by using new texts that have been encoded using JSesh, an open-source hieroglyphic text editor. These texts will be published for others to use with digital or traditional methods. All the tools and codes will also be released with an open license.

Homburg, T. & C. Chiarcos. 2016. Word Segmentation for Akkadian Cuneiform. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).*

Jauhiainen, H. 2022. Encoding Hieroglyphic Texts. *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022*.

Rosmorduc, S.. 2021. Ramses automated translitteration software. In Lingua Aegyptia (2021-06- 15, Vol. 28, pp. 233–257). Zenodo.

Schweitzer, Simon. 2021. AES - Ancient Egyptian Sentences; Corpus of Ancient Egyptian sentences for corpus-linguistic research. GitHub.