Data Sources for Automatic Classification and Analysis of Texts from Egyptian Antiquity

Tommi Jauhiainen, Heidi Jauhiainen, Marja Vierros
University of Helsinki, Finland
firstname.surname@helsinki.fi

Erik Henriksson
University of Turku, Finland
firstname.surname@utu.fi

Poster Abstract

In this poster, we present the aims and the current state of the research project "Automatic Classification and Analysis of Texts from Egyptian Antiquity", funded by the Kone Foundation.

In short, the project aims to develop new state-of-the-art language technological methods for automatically processing textual documents from Egypt dating from the 8th century BCE to the Arab conquest in the 7th century CE. The project investigates the extensive textual evidence from the region as a whole, including the texts in both the Greek and the Egyptian languages. We aim to develop new and improved methods for automatically identifying languages and dialects (Jauhiainen et al. 2019) and detecting the text's date and place of origin (Jauhiainen et al. 2023). Furthermore, one of our goals is to build automated methods for finding loan words between languages by creating converters between all the target languages' transliteration systems (Jauhiainen & Jauhiainen 2023).

A large part of the project is dedicated to collaboration between the project and various entities that own the copyright to the existing machine-readable texts within the focus of the research (Jauhiainen 2022). We will identify sources for machine-readable texts pertinent to our study and, if they are not openly available, negotiate with the rightsholders for suitable access to the texts to use in the project.

We will create a database of all sources where relevant machine-readable text collections are available. The listing will be openly available on the project's website and updated throughout the project's lifespan. We will contact the entities and persons behind the text collections and aim to get the data as exports from their system instead of reverting to methods like web scraping. We have already identified several sources for texts that are usable by the project.

For the texts primarily written in Greek, we use all the transcribed texts available through the Papyri.info project as our data. Currently, the papyri.info collection contains metadata for over 100,000 texts, of which more than 50,000 are transcribed. In addition to the document data from papyri.info, we already have access to several thousand inscriptional texts from the Packard Humanities Institute's collection.

Thesaurus Linguae Aegyptiae (TLA), a digital publication platform, includes machine-readable texts written in Egyptian using either Hieroglyphic, Hieratic, or Demotic scripts. The TLA is the largest ongoing project collecting and publishing machine-readable ancient Egyptian texts, and their collection is continuously increasing. We expect the latest

form of the logographic Egyptian writing, Demotic, to be most interesting regarding language contact, as it was used while the Greeks ruled Egypt.

Keywords: corpora, egyptology, language identification, multilinguality, papyrology

REFERENCES

Jauhiainen, H. (2022). Encoding Hieroglyphic Texts. In K. Berglund, M. La Mela, & I. Zwart (Eds.), *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)* (pp. 244-250). Article 22 (CEUR Workshop Proceedings ; Vol. 3232). CEUR-WS.org.

Jauhiainen, H., & Jauhiainen, T. (2023). Transliteration Model for Egyptian Words. In A. Rockenberger, J. Tiemann, & S. Gilbert (Eds.), *DHNB2023 : Conference Proceedings* (pp. 149-164). (Digital Humanities in the Nordic and Baltic Countries Publications; Vol. 5, No. 1). University of Oslo Library. https://doi.org/10.5617/dhnbpub.10659

Jauhiainen, T., Henriksson, E., Vierros, M., & Jauhiainen, H. (2023). *Automatic detection of place and time for Greek texts in Egypt*. Poster session presented at International Congress of Egyptologists, Leiden, Netherlands.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., & Lindén, K. (2019). *Automatic Language Identification in Texts: A Survey*. Journal of Artificial Intelligence Research, 65, 675-782. https://doi.org/10.1613/jair.1.11675