Developing Natural Language Processing Tools for Egyptian

Roberto Antonio Díaz Hernández

University of Jaén (radiaz@ujaen.es)

Relevant UniDive working groups: WG1, WG2, WG3, WG4

1 Introduction

Egyptian comprises the following phases: Old Egyptian (ca. 2700–2000 BC), Middle Egyptian (ca. 2000–1400 BC), Late Egyptian (ca. 1300–700 BC), Demotic (7th century BC to 5th century CE) and Coptic (4th to 14th century CE). Before UniDive, NLP tools for Egyptian were scarce. Some lexical databases existed, such as the Thesaurus Linguae Aegyptiae[1] and Ramses Online,[2] as well as Coptic NLP tools developed by the Scriptorium team.[3] However, this situation changed during UniDive (2022–2026), when support was provided for the creation of NLP tools for pre-Coptic Egyptian, including the Universal Dependencies Egyptian_Pre-Coptic treebank, the PARSEME corpus of Egyptian multi-word expressions, GrewPT, and the EPC parser for pre-Coptic sentences.

2 Contents

This paper provides an overview of the NLP tools developed for pre-Coptic Egyptian during the COST Action UniDive. Section 3 describes the Universal Dependencies Egyptian_Pre-Coptic treebank (hereafter referred to as the EPC treebank), the first dependency treebank created for the morphosyntactic annotation of pre-Coptic Egyptian. Section 4 discusses the PARSEME corpus of Egyptian MWEs, while Section 5 describes GrewPT, a web application for researching hieroglyphic spellings and grammatical features in the Pyramid Texts. Section 6 introduces the EPC parser, created to parse pre-Coptic Egyptian texts. Section 7 outlines future work and Section 8 provides references.

3 The Egyptian_Pre-Coptic treebank

The EPC treebank was created in March 2024 after the 2nd UniDive general Meeting at the University of Naples L'Orientale.[4] It currently contains 3,089 sentences and 34,234 tokens from the Pyramid Texts,[5] written in Old Egyptian. The Pyramid Texts are the oldest corpus of Egyptian religious texts. They comprise a collection of spells recorded on the walls of the pyramids of Old Kingdom kings and queens (ca. 2543–2120). The Pyramid Texts were edited by Sethe in two volumes (1908–22).

The EPC treebank contains the main witnesses of the Pyramid Texts[6] in Sethe's first volume of the Pyramid Texts and the first 100 pages of his second volume.[7] Each sentence is provided with 16 metadata sections in the EPC treebank: sent_id, title, edition, spell, section, king, date, language, genre, place, TLA,[8] text, literal translation, free translation, sentence type and comments. Words are annotated in ten columns according to the CoNLL-U format. Column 5 (XPOS) is left blank because the specific part-of-speech tags are annotated using universal part-of-speech tags in column 4 and morphological features in column 6, for example the so-called "pseudoparticiple" is annotated using VERB in column 4 and Conjug=AfroAsia (i.e. Conjugation=Afroasiatic) in column 6. Column 10 is used for hieroglyphic spellings, classifiers and word IDs,[9] for example:

Hiero=  |Hierocl=  |ID=314 (*ʾgb(.w)* "abundance").

[1] https://thesaurus-linguae-aegyptiae.de/home (last updated on 13.05.26).
[2] https://ramses.ulg.ac.be/ (last updated on 13.05.26).
[3] https://copticscriptorium.org/ (last updated on 13.05.26).
[4] See Díaz Hernández, Passarotti 2024, see also: https://web.ujaen.es/investiga/nile-in-contact/EPC_en.html and https://github.com/UniversalDependencies/UD_Egyptian-PC/tree/dev (last updated on 13.05.26).
[5] UD release 2.18 (May 2026).
[6] In textual criticism, the term "main witness" is used to refer to the most reliable version of a text compared to other variants.
[7] Sethe (1908–1922).
[8] This section contains a link to the Thesaurus Linguae Aegyptiae (TLA) transcription and translation of the sentence.
[9] The word ID is the number assigned to each Egyptian word by the Thesaurus Linguae Aegyptiae.

The EPC treebank has a wide range of applications in computational linguistics. It has been used to develop the following natural language processing tools: the PARSEME repository for Egyptian multi-word expressions, GrewPT and the EPC parser for pre-Coptic sentences.

4 PARSEME corpus of Egyptian MWEs

PARSEME guidelines 2.0 for the identification and analysis of multi-word expressions has been adapted for Egyptian.[10] All categories of multi-word expressions can be found in Egyptian except for idiomatic verb-particle constructions, which are typical of Germanic languages, and deverbal adverbial multi-word expressions. In addition, two categories of multi-word expressions are specific of Egyptian: adjectival idioms consisting of nisba adjectives (example 1) and nominal idioms from nisba adjectives (example 2).

1) Pyramid Texts, 644e, Teti:

Ḥr.w	*ḥr(.i)-*	*tp*	*rḫ.(w)t*	*=f*
Horus-GN	who is on-M.SG	head-M.SG	subject-F.PL	=3SG.M

LT: “Horus who is on the head of his subjects.”
FT: “Horus who rules over his subjects.”

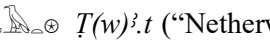
2) Pyramid Texts, 211a, Unas:

nỉ	*ỉm(.ỉ)-*	*rṭ*	*=k*
not-NEG	one who is in-M.SG	foot-M.SG	=2SG.M

LT: “There is no one who is in your foot.”
FT: “You have no enemy.”

In Semitic languages, such as Arabic, “nisba” designates an ending added to nouns, and rarely to prepositions and pronouns, to form relative adjectives and nouns,[11] for example the nisba ending in Arabic is ي (*ī*) and in Egyptian is  *ỉ/ỉ̈*, cf.:

3) لبنان (“Lebanon”,) > لبناني (“a Lebanese”);

4)  *Ṯ(w)ʾ.t* (“Netherworld”) > *ṯ(w)ʾ.tỉ* (“A dweller of the Netherworld”).

In Egyptian, nisba adjectives are usually formed adding the *ỉ* ending to a preposition, for example the nisba adjective *ḥr(.ỉ)* “one who is on” derives from the preposition *ḥr* “on”. A nisba adjective can be used in adjectival and nominal idioms if it is fixed to another word with an idiosyncratic meaning, as in examples 1 and 2 (see above).

PARSEME guidelines 2.0 are used by the PARSEME Egyptian research group at the University of Jaén for identifying Egyptian MWEs.[12] The group participated in the last PARSEME campaign for the multilingual annotation of MWEs in the summer of 2025. The Egyptian research group contributed to the PARSEME 2.0 corpus with the identification of over 600 MWEs in the Pyramid Texts, as annotated in the EPC treebank. The results of this campaign are included in the paper entitled “PARSEME 2.0 multilingual corpus of multiword expressions”,[13] which has been selected for an oral presentation at the next Language Resources and Evaluation Conference, to be held in Palma de Majorca on 11–16 May 2026. The corpus of Egyptian MWEs can be explored and studied in Grew-match PARSEME.[14]

5 GrewPT

The EPC treebank was used by the author and Bruno Guillaume to develop GrewPT during a Short-Term Scientific Mission grant at the French Institute for Computational Research in Nancy in spring 2025. GrewPT is a web application for analysing the language and writing of the Pyramid Texts. It enables the graphic and morphosyntactic analysis of each word in the Pyramid Texts and it can be used to study the linguistic variations between different versions of these texts, for example if the following pattern is entered in GrewPT:

```
pattern { X[VerbForm="Part", Prefix="Yodh"] }
```
[15]

it will yield 74 occurrences of participles with a *ỉ* prefix in the Pyramid Texts, showing that distribution differs depending on the king’s pyramid: 34 occurrences in Unas’s pyramid, 18 in Pepi’s pyramid, 11 in Teti’s pyramid, 10 in Neferkare’s pyramid and one in Neith’s Pyramid.

[10] Díaz Hernández 2026.
[11] Schulz 2010, 86.
[12] https://web.ujaen.es/investiga/nile-in-contact/PARSEME_en.html (last updated on 13.05.26).
[13] Savary et al. 2026.
[14] https://parseme.grew.fr/?corpus=PARSEME-EGY@dev# (last updated on 13.05.26).
[15] https://pt.grew.fr/?custom=69b803cfbd740 (last updated on 13.05.26).

6 The EPC Parser

The EPC treebank was processed using Stanza's neural pipeline in order to develop an online morphosyntactic parser for pre-Coptic Egyptian.[16] This parser can already be used to analyse Old Egyptian sentences. If the sentence *Ppy pw mr.y =f* "Pepi is his beloved" is entered, the morphosyntactic analysis is obtained in the CoNLL-U format alongside a tree diagram. The parser also automatically generates the hieroglyphic spelling for each word in the MISC column. As the EPC treebank currently relies on the Pyramid Texts, the parser can only analyse Old Egyptian texts. Once texts from later stages of Egyptian are annotated in the EPC treebank, the parser will be able to analyse sentences from these stages of Egyptian too.

7 Conclusion and future work

Developing these NLP tools for Egyptian will improve our understanding of the evolution undergone by this language throughout its history, for they can afford a large number of examples previously inaccessible.

Future work will involve expanding the EPC treebank by annotating the most significant texts from each stage of the Egyptian language.[17] This will allow the PARSEME Egyptian research group to carry out a diachronic study of the use of MWEs in Egyptian texts. Expanding the EPC treebank will also benefit GrewPT, as it will enable comparisons to be made between the morphosyntactic constructions of the Pyramid Texts and other Egyptian texts. Furthermore, the accuracy of the EPC parser when analysing sentences will increase significantly once the pre-Coptic Egyptian text corpus has fully been morphosyntactically annotated in the EPC treebank.

8 References

a) Bibliography:

Díaz Hernández, Roberto A. 2026. "PARSEME-Ansatz 2.0 für Mehrwortausdrücke im Ägyptischen", in *Zeitschrift für Ägyptische Sprache*, 153/1, 1–12.

Díaz Hernández, Roberto A., Passarotti, Marco C. 2024. "Developing the Egyptian-UJaen Treebank", in: Daniel Dakota (*et. al.*) *Proceedings of the 22nd International Workshop on Treebanks and Linguistic Theories (TLT 2024)*, 1–10.

Savary et al. 2026. "PARSEME 2.0 Multilingual Corpus of Multiword Expressions", in: *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC 2026), 4819–4834.

Schulz, E. 2010. *A Student Grammar of Modern Standard Arabic*. Cambridge.

Sethe, K. 1908–1922. *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums.* 4 volumes, Heinrich'sche Buchhandlung, Leipzig.

b) Digital resources:

Egyptian Pre-Copic treebank in UD: https://github.com/UniversalDependencies/UD_Egyptian-PC/tree/dev

Grew-match PARSEME: https://parseme.grew.fr/?corpus=PARSEME-EGY@dev#

GrewPT: https://pt.grew.fr/?corpus=PT_all

PARSEME Egyptian research group: https://web.ujaen.es/investiga/nile-in-contact/PARSEME_en.html

Parser for pre-Coptic Egyptian: https://web.ujaen.es/investiga/nile-in-contact/Analizador_es.html

Ramses Online: https://ramses.ulg.ac.be/

Scriptorium: https://copticscriptorium.org/

Thesaurus Linguae Aegyptiae: https://thesaurus-linguae-aegyptiae.de/home

[16] https://web.ujaen.es/investiga/nile-in-contact/Analizador_en.html (last updated on 13.05.26).

[17] Díaz Hernández, Passarotti (2024, 2).