[Neural Models for Lemmatization and POS-Tagging of Earlier and Late Egyptian (Supporting Hieroglyphic Input) and Demotic

Aleksi Sahala
University of Helsinki
Helsinki, Finland
aleksi.sahala@helsinki.fi

Eliese-Sophia Lincke
Freie Universität Berlin & Berlin-Brandenburg
Academy of Sciences and Humanities
Berlin, Germany
e.lincke@fu-berlin.de

Abstract

We present updated models for *BabyLemmatizer* for lemmatizing and POS-tagging Demotic, Late Egyptian and Earlier Egyptian with a support for using hieroglyphs as an input. In this paper, we also use data that has not been cleaned from breakages. We achieve consistent UPOS tagging accuracy of 94% or higher and an XPOS tagging accuracy of 93% and higher for all languages. For lemmatization, which is challenging in all of our test languages due to extensive ambiguity, we demonstrate accuracies from 77% up to 92% depending on the language and the input script.

1 Introduction

Since several ancient languages feature complex morphology and a high degree of spelling variation, lemmatization is an essential step for making large text collections of these languages searchable and usable for further computational analysis.

In this paper we present models for lemmatizing and part-of-speech tagging Earlier and Late Egyptian, as well as Demotic, which complements our earlier research on the topic by using larger datasets with *lacunae* (breakages), as well as text represented in Unicode hieroglyphs. Our models are available on https://huggingface.co/asahala.

2 Egyptian-Coptic

2.1 Diachronic Overview

Egyptian-Coptic was the indigenous language of the lower Nile valley, attested in written form from around 3000 BCE to 1400 CE. It belongs to the Afroasiatic language family and is generally divided into two major phases: Earlier Egyptian, which includes Old and Middle Egyptian, and Later Egyptian, comprising Late Egyptian, Demotic, and Coptic. The transition from Earlier to Later Egyptian is marked by significant linguistic changes in morphology and syntax. While Earlier Egyptian retained a more synthetic structure with root-and-pattern morphology, Later Egyptian initially exhibits increased analytic tendencies, particularly in its verbal system. However, this trend is later followed by a phase of re-synthetization. Another major difference between Earlier and Later Egyptian is the shift from marking main clauses to marking subordinate clauses (Kammerzell, 1998; Winand, 2018). Basic information about the Earlier Egyptian and Demotic language stages has been given elsewhere (Sahala and Lincke, 2024) and will not be repeated here. However, this study also addresses the chronolect Late Egyptian, which was not included in previous work and will briefly be introduced in the following section.

The language phases are represented in distinct corpora and scripts, necessitating different approaches to transcription, lemmatization, and other text processing techniques.

2.2 Late Egyptian

The chronolect referred to as ‘Late Egyptian’ (or French ‘Néo-Egyptien’) surfaces in the written record in the 14th century BCE although some features can be observed in considerably earlier texts (Kroeber, 1970). Late Egyptian is characterized by an analytical tendency as compared to Earlier Egyptian (fusional) and the later Demotic and Coptic (agglutinative) language stages (McLaughlin, 2022; Stauder, 2020), e.g. by employing periphrastic verb phrases. The word order pattern (AUX-)S-V-O becomes more prominent although it is only fully fledged in Coptic. With respect to the attested sentence types it can be stated that sentences with an adjectival predicate are receding and are being replaced by alternative constructions following the adverbial pattern (Winand, 2018).

As with pre-Demotic Egyptian in general, Late Egyptian texts are recorded in two native Egyptian scripts: monumental hieroglyphs, which were used for inscriptions on stone and, in cursive form,

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 77–82
May 4, 2025. ©2025 Association for Computational Linguistics

for certain texts on papyrus (e.g., the Book of the Dead) or wood; and hieratic, a cursive script written mostly on papyrus and ostraca (pottery and limestone sherds).

3 Datasets

The datasets for all language stages discussed here were exported and made available to us by Daniel A. Werning from the database that feeds the Thesaurus Linguae Aegyptiae (TLA), corpus v18 (Richter et al., 2023).[1] The export format is JSONL with each sentence (as defined by the TLA's data model and editors) stored as a separate JSON object and the tokens separated by blanks (Fig. 1). Each sentence is represented both in Unicode hieroglyphs (without quadrat placement) and in Egyptological transcription (i.e., Leiden Unified Transliteration), provided that hieroglyphs have been encoded for the respective text. It is annotated with TLA lemma IDs and POS-tagged using the UPOS tag set[2] and a simplified version of the project-specific subclass tag set of the TLA as the XPOS tag set (Werning, 2024). This XPOS tag set is fine-grained with respect to proper nouns, using different tags for divine names, royal names, personal names, animal personal names, names of institutions, names of artifacts, and place names. It also distinguishes epithets and titles from other types of nouns.

```
{"hieroglyphs": "𓋴𓆷𓂋 <g>G175</g> 𓎛𓀜 𓄂𓏏𓏤 𓌃𓂧𓏭𓀁",
"transliteration": "sḏr r-ḥꜣ.t mdwi̯",
"lemmatization": "150740|sḏr 500053|r-ḥꜣ.t/2 78140|mdwi̯",
"UPOS": "VERB ADP VERB",
"XPOS": "verb preposition verb"}
```

Figure 1: JSON object from the Late Egyptian dataset, *The Teaching of Amenemope* 5,13, pBM EA 10474, TLA ID: IBUBd2RAxJagbkako4lYd0WxDc8.

The lemmatization of the *Thesaurus Linguae Aegyptiae* is fine-grained and tailored to Egyptologists' needs, allowing them to distinguish and search for the individual meanings and functions of a lemma. Consequently, a single lemma may be divided into multiple sub-lemmata, each assigned its own TLA lemma ID, as illustrated in Table 1 for the preposition *m*, the most frequent word in Egyptian. Another reason why lemma IDs are necessary for Egyptian is the high number of homonyms (or more precisely, homographs) in the Egyptological transcription (”transliteration”), which we have described in more detail in Sahala and Lincke (2024).

	Lemma ID	Meaning / Function
1.	64360	[preposition]
2.	400007	in; to; on; from (spatial)
3.	64365	in; on (temporal)
4.	64362	in (condition, state)
5.	400082	(consisting) of (partitive)
6.	64364	by means of (instrumental)
7.	400080	together with (comitative)
8.	500292	like; as (predication)
9.	854625	[connector of the direct object]
10.	64369	[with infinitive]
11.	64370	when; if [as conjunction]

Table 1: Sub-lemmata of the preposition *m* in the *Thesaurus Linguae Aegyptiae* lemma list.

The Earlier Egyptian dataset consists of all sentences that predate the Egyptian New Kingdom (c. 1550–1070 BC). The Demotic dataset comprises the entire Demotic text corpus in the TLA. Defining a Late Egyptian dataset is more challenging, as texts in the TLA are not consistently tagged by language phase. Therefore, our Late Egyptian dataset includes only those texts explicitly labeled as Late Egyptian in the TLA metadata.[3]

Other than the material used in Sahala and Lincke (2024), our datasets are not filtered for “premium” sentences that are “fully intact” and “unambiguously readable” (TLA-Dem 2024, TLA-Egy 2024)[4]. The datasets include damaged text, i.e. broken or destroyed individual hieroglyphs or entire word forms that could not be reconstructed by the editors. The respective sizes of the datasets can be found in Table 2.

Language stage	Sentences	Tokens
Earlier Egyptian	43,447	∼286,000
Late Egyptian	9,005	∼86,100
Demotic	25,822	∼292,450

Table 2: Sentence and token counts for the Earlier Egyptian, Late Egyptian, and Demotic datasets.

Our aim is to train models that can handle two different types of input: (1) Unicode-encoded hiero-

[1] For detailed information on the datasets, see Section *Sources*.

[2] https://universaldependencies.org/u/pos/

[3] TLA ID: J3SNMB4AF5ERPDGE4VPMBZSYRE, Thot Thesauri & Ontology ID: thot-12.

[4] https://huggingface.co/datasets/thesaurus-linguae-aegyptiae

glyphs (e.g. as the output of a successful future hieroglyphic OCR) and (2) transcription, which remains the default digital representation of Egyptian, since many projects still render hieroglyphs only as images. Depending on the availability in the database, not all sentences of Earlier and Late Egyptian in our datasets contain hieroglyphic spellings, some texts were only encoded by means of transcription. Demotic is represented in transcription only, since there is no encoding for the Demotic script itself.

Challenges lie in the complexity of the input data. Currently, not all hieroglyphs are available in Unicode. In such cases, they are encoded using the alphanumerical system known as *Gardiner numbers* enclosed in the tag <g> (Fig. 1, first line, in purple). We test how well our lemmatizer can predict the lemma string plus a numerical index (Fig. 1, third line, in red) replacing the arbitrary TLA lemma IDs (in blue), instead of simply representing a lemma as a string (see Section 6). Effectively, this means training the lemmatizer to disambiguate homonyms caused by the simplified rendering of Egyptian in transcription and by the subdivision of lemmata.

4 Previous Work

In their paper on Neural Machine Translation for Egyptian, using the TLA data dump from 2018, De Cao et al. (2024) incorporated the prediction of lemma IDs (lemmatization) and POS tags into the training of some of their models. Their results look promising but cannot be directly compared to ours, as they use SacreBLEU and RougeL as their evaluation metrics, which cannot be converted into accuracy rates, our primary evaluation metric. Díaz Hernández and Carlo Passarotti (2024) manually annotated a dataset of 14,650 tokens from the Old Egyptian Pyramid Texts for the first Egyptian treebank, including lemmatization and POS-tagging. They trained a UDPipe model and evaluated their results with F1 scores of 89.38 (lemma), 90.30 (UPOS), and 76.01 (XPOS). However, their lemmatization approach was string-based and did not account for homonymy by using lemma IDs, making the task significantly simpler than ours, which requires the disambiguation between homonyms and/or multiple sub-lemmata.

Other than that and apart from our own effort (Sahala and Lincke, 2024), models have been created only to lemmatize and POS-tag Coptic. (Zeldes and Schroeder, 2016, 2015; Smith and Hulden, 2016; Dereza et al., 2024).

5 BabyLemmatizer

BabyLemmatizer[5] is a lemmatization and POS-tagging pipeline originally designed for the cuneiform languages of Mesopotamia, but is also capable of handling other transliteration and writing systems (Sahala and Lindén, 2023).

The system is based on the Open Neural Machine Translation Toolkit (Klein et al., 2017) and handles POS-tagging and lemmatization as machine translation tasks by mapping character or symbol sequences to each other. It uses a deep attentional encoder-decoder network with a two-layer BiLSTM encoder that reads the input as a character sequence. The output sequence is produced by a two-layer unidirectional LSTM decoder with input feeding attention. We use the default batch size of 64 and start the learning rate decay halfway through the training process.

The neural lemmatizer is followed by a dictionary-based post-corrector to verify the in-vocabulary lemmatizations for better accuracy. The post-corrector also labels lemmatizations with confidence scores that enable easier location of potentially incorrect lemmata.

6 Preprocessing and Training

We converted the datasets from the original JSONL format into CoNLL-U to make it usable by BabyLemmatizer. Our CoNLL-U lacks dependency labels and morphology, and uses a simplified lemma notation by representing the disambiguation identifiers in a shorter form (*r-ḥꜣ.t/2* instead of *500053|r-ḥꜣ.t*, see Fig. 1), since our previous experiments proved that the long identifiers are detrimental for OOV word lemmatization.

We use BabyLemmatizer's alphabetic tokenizer for all our models that splits the input strings into character sequences represented as Unicode hieroglyphs or transcribed Latin characters. The POS-tagger input sequence is encoded as a 5-gram of concatenated word forms. The lemmatizer is run after the POS-tagging, and its input sequences are encoded as concatenations of four strings, where the first one represents the input word form (in transliteration or hieroglyphs) and the three following its

[5] The tool is available at https://github.com/asahala/BabyLemmatizer

predicted XPOS tag, as well as the predicted XPOS tags of the preceding and the following words.

7 Evaluation

We generate a 80/10/10 train/dev/test split of our datasets and evaluate our models using 10-fold cross-validation. We estimate the performance of our models by using accuracy as our evaluation metric, since we only predict one lemma for each input word (instead of, for example, the most likely three candidates). Our predicted labels are LEMMA, XPOS and UPOS. Due to high lemmatization ambiguity, we do not predict the lemma alone, but also its index, which separates it from other homonymous lemmata. This makes the task significantly more challenging in comparison to typical lemmatization tasks, where only the dictionary forms are predicted. Our final results are summarized in Table 3 with confidence intervals of the cross-validation shown in parentheses.

Our results for Demotic and transcribed Earlier Egyptian show a moderate improvement in comparison to our previous paper albeit the used data contain breakages; for instance, the lemmatization for Earlier Egyptian in transcription improves by 2.04%.

It seems that the hieroglyphic input produces less accurate results than using the transcription. This is due to the increased vocabulary size, and hence a larger number of OOV-vocabulary words, which result from spelling variation that is normalized in transcription (for an example see Sahala and Lincke, 2024, p. 89, Fig. 1).

7.1 Data Augmentation and Model Corrector Experiments

We attempted to improve Egyptian lemmatization results by augmenting Late Egyptian training data with Earlier Egyptian data and vice versa, but this did not yield consistently better results for transcription or hieroglyphs.

In addition, we experimented with training a secondary model for post-correcting the lemma identifiers. This process involved first predicting the POS tags and simplified lemmata without identifiers, which can be predicted with an accuracy of ca. 94% for transcribed Earlier Egyptian. The post-corrector attempted to map varying length sequences of simplified lemmata and their POS-tags to the lemmata with identifiers, but we were unable to improve the results.

7.2 Error Analysis

In the test set for Earlier Egyptian, 2,960 tokens were erroneously lemmatized from the hieroglyphic input. Of these, 323 (10.91%) correspond to tokens with the hieroglyphic form 𓅓 (*m*), and 313 of these 323 specifically are instances of the preposition *m* ‘in’, which is divided into multiple sub-lemmata in our corpus (see Table 1). If all these sub-lemmata were assigned to a single lemma—e.g. the hypernym for the preposition *m* (TLA lemma ID 64360, see no. 1 in Table 1)—the total error count could be reduced by 313 (10.57%) solely by addressing this one hieroglyphic input form. The same is true for other frequently used prepositions, such as 𓈖 (*n*) ‘for, to’ and 𓂋 (*r*) ‘to, at’.

In an additional 192 errors (7.14%) in Earlier Egyptian lemmatization with hieroglyphic input, the tokens contain hieroglyphic characters not represented as Unicode points, but rather using the <g> tag and Gardiner numbers (see Fig. 1). This indicates that BabyLemmatizer struggles to effectively learn these non-Unicode representations from the given input data.

With an effective token count of 13.8k in the test set (out of a total size of 28.6k), the 505 instances of two mentioned error types alone account for 3.66%. This means that the accuracy—specifically for lemmatization based on hieroglyphic input—could be significantly improved by simplifying the data, e.g. by avoiding lemmatization at the sub-lemma level and by filtering out tokens with non-Unicode-compliant hieroglyphs.

8 Conclusions and Future Work

We presented lemmatization and POS-tagging models for Earlier Egyptian, Late Egyptian, and Demotic with varying results. Whereas the accuracy for Demotic is fairly good (tagger 97%, lemmatizer 92%), the Earlier and Late Egyptian yielded adequate results only for POS tagging (93-96%).

Disambiguating the highly ambiguous Egyptian lemmata is beyond the capabilities of BabyLemmatizer’s current model architecture. Therefore, we plan to tackle this issue in the future using more context-aware approaches, including transformers and LLMs, which could perhaps be fine-tuned for disambiguation tasks. Moreover, additional annotation layers, such as dependency parsing, could possibly improve the quality of the lemmatization, as syntactic and morphological labels have previously been used successfully in lemma disambiguation

Whole dataset

	Demotic	EarlierE T	EarlierE H	LateE T	LateE H
XPOS	97.13 (±0.09)	96.20 (±0.08)	92.97 (±0.16)	93.98 (±0.08)	93.13 (±0.26)
UPOS	97.45 (±0.09)	96.62 (±0.15)	93.64 (±0.04)	94.48 (±0.16)	93.52 (±0.23)
LEMMA	92.15 (±0.18)	87.56 (±0.19)	80.15 (±0.20)	79.98 (±0.26)	76.59 (±0.48)
OOV-rate	2.51	2.62	13.54	5.54	16.75

OOV word forms only

	Demotic	EarlierE T	EarlierE H	LateE T	LateE H
XPOS	82.12 (±1.45)	78.00 (±1.04)	81.39 (±0.69)	76.09 (±2.02)	82.19 (±1.67)
UPOS	85.70 (±1.94)	82.45 (±1.28)	83.89 (±0.48)	78.28 (±1.20)	83.35 (±1.43)
LEMMA	50.96 (±1.25)	53.85 (±1.18)	51.16 (±0.91)	43.63 (±1.80)	50.87 (±1.92)

Table 3: Evaluation results. OOV-rate shows the average percentage of OOV word forms in the test set with respect to training corpus. H = hieroglyphic input and T = transcription.

(Kanerva et al., 2021). We also plan to organize a shared task for Egyptian lemmatization, since the issues are rather unique and are likely to be more easily solved with input from a larger NLP community.

Acknowledgments

We sincerely thank Daniel A. Werning (Berlin-Brandenburg Academy of Sciences and Humanities) for creating tailor-made exports for all language stages from the TLA, for engaging in discussions on the lemma lists, data model and annotations, and for providing valuable feedback on our work. We also thank the Academy of Finland for funding the project Origins of Emesal (PI Krister Lindén) and the Centre of Excellence in Ancient Near Eastern Empires (PI Saana Svärd).

Sources

All datasets are taken from **Thesaurus Linguae Aegyptiae, corpus v18, 2023**, ed. by Tonio Sebastian Richter & Daniel A. Werning on behalf of the Berlin-Brandenburgische Akademie der Wissenschaften and Hans-Werner Fischer-Elfert & Peter Dils on behalf of the Sächsische Akademie der Wissenschaften zu Leipzig:

- **TLA-Dem 2025**: Thesaurus Linguae Aegyptiae, Demotic sentences, corpus v18, with destroyed tokens, v1.0, 1/28/2025. Rights reserved.
- **TLA-Egy-L 2025**: Thesaurus Linguae Aegyptiae, Late Egyptian sentences, corpus v18, with destroyed tokens, v1.0, 1/28/2025. Rights reserved.
- **TLA-Egy-E 2025**: Thesaurus Linguae Aegyptiae, Earlier Egyptian sentences, corpus v18, with destroyed tokens, v1.0, 1/28/2025. Rights reserved.

Please refer to the *TLA authors* website for a detailed account of annotators per text/corpus (Werning, 2025).

References

Mattia De Cao, Nicola De Cao, Angelo Colonna, and Alessandro Lenci. 2024. Deep learning meets egyptology: a hieroglyphic transformer for translating Ancient Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 71–86, Bangkok, Thailand. Association for Computational Linguistics.

Oksana Dereza, Adrian Doyle, Priya Rani, Atul Kr. Ojha, Pádraic Moran, and John McCrae. 2024. Findings of the SIGTYP 2024 shared task on word embedding evaluation for ancient and historical languages. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 160–172, St. Julian's, Malta. Association for Computational Linguistics.

Roberto Antonio Díaz Hernández and Marco Carlo Passarotti. 2024. Developing the Egyptian-UJaen treebank. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 1–10, Hamburg, Germany. Association for Computational Linguistics.

Frank Kammerzell. 1998. *Sprachkontakte und Sprachwandel im Alten Ägypten*. Habilitation thesis, University of Göttingen, Göttingen.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Burkhart Kroeber. 1970. *Die Neuägyptizismen vor der Amarnazeit. Studien zur Entwicklung der ägyptischen Sprache vom Mittleren zum Neuen Reich*. Dissertation, Universität Tübingen, aku Fotodruck, Bamberg.

Rachael Hannah McLaughlin. 2022. *The Linguistic Cycle in Ancient Egyptian Verbal Constructions*. Phd thesis, University of Liverpool.

Tonio Sebastian Richter, Daniel A. Werning, Hans-Werner Fischer-Elfert, and Peter Dils, editors. 2023. *Thesaurus Linguae Aegyptiae, Corpus issue 18*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. Web-App-Version 2.1.3, Accessed: 5/16/2024.

A. J. Aleksi Sahala and Krister Lindén. 2023. A neural pipeline for lemmatizing and POS-tagging cuneiform languages. In *Proceedings of the Ancient Language Processing Workshop at the 14th International Conference on Recent Advances in Natural Language Processing RANLP 2023*, pages 203–212.

Aleksi Sahala and Eliese-Sophia Lincke. 2024. Neural lemmatization and POS-tagging models for Coptic, Demotic and Earlier Egyptian. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 87–97, Bangkok, Thailand. Association for Computational Linguistics.

Daniel Smith and Mans Hulden. 2016. Morphological analysis of Sahidic Coptic for automatic glossing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2584–2588.

Andréas Stauder. 2020. History of the Egyptian Language. In Ian Shaw and Elizabeth Bloxam, editors, *Oxford Handbook of Egyptology*, pages 930–956. Oxford University Press, Oxford.

Daniel A. Werning. 2024. TLA Parts of Speech. In Tonio Sebastian Richter, Daniel A. Werning, Hans-Werner Fischer-Elfert, and Peter Dils, editors, *Thesaurus Linguae Aegyptiae, Corpus issue 19, Web app version 2.2.0, 11/5/2024*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. Accessed: 2/8/2025.

Daniel A. Werning. 2025. TLA authors. In Tonio Sebastian Richter, Daniel A. Werning, Hans-Werner Fischer-Elfert, and Peter Dils, editors, *Thesaurus Linguae Aegyptiae, Corpus issue 19, Web app version 2.2.1.1, 3/6/2025*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. Accessed: 3/16/2025.

Jean Winand. 2018. Late Egyptian. In Julie Stauder-Porchet, Andréas Stauder, and Willeke Wendrich, editors, *UCLA Encyclopedia of Egyptology*. Los Angeles.

Amir Zeldes and Caroline T. Schroeder. 2015. Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. *Digital Scholarship in the Humanities*, 30(suppl1):i164–i176.

Amir Zeldes and Caroline T. Schroeder. 2016. An NLP pipeline for Coptic. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 146–155, Berlin, Germany. Association for Computational Linguistics.