Neural Lemmatization and POS-tagging models for Coptic, Demotic and Earlier Egyptian

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 BabyLemmatizer models for lemmatizing and POS-tagging Earlier Egyptian, Coptic and Demotic to test the performance of our pipeline for the ancient languages of Egypt.[1] Of these languages, Demotic and Earlier Egyptian are known to be difficult to annotate due to their high extent of ambiguity. We report lemmatization accuracy of 86%, 91% and 99%, and XPOS-tagging accuracy of 89%, 95% and 98% for Earlier Egyptian, Demotic and Coptic, respectively.

1 Introduction

Lemmatization is an annotation task that aims to label word forms with their dictionary forms, known as lemmata. This is necessary for languages with complex writing systems or morphology that would otherwise preclude effective word searches using simple keywords. By enabling the location of all inflected forms and spelling variants of any searched word, lemmatization opens several interesting avenues for quantitatively studying historical texts and their language.

POS tagging is another annotation task that aims to label word forms with their part-of-speech tags. This can be useful for simple named entity recognition, syntactic parsing, and disambiguation of lemmatization results. The more fine-grained the POS tagging is, the more information it can provide about the words in the corpus.

In this paper, we present lemmatizer and POS-tagger models for Earlier Egyptian, Coptic, and Demotic. Earlier Egyptian and Demotic pose particular challenges for lemmatization due to their ambiguous word forms, which are often only one or two characters long. To our knowledge, neural lemmatization of these languages has not been attempted before. Our models are based on BabyLemmatizer, an OpenNMT-based neural lemmatizing and POS-tagging pipeline designed primarily for historical languages. Previously, BabyLemmatizer has been evaluated on Sumerian, Babylonian, Neo-Assyrian, Urartian, Latin, and Ancient Greek with promising results (Sahala and Lindén, 2023).

[1]The models are available at https://huggingface.co/asahala

2 Languages and Datasets

Egyptian-Coptic existed as a spoken language long before its first written records (Pre-Old Egyptian, (Kammerzell, 2005)). It is attested in writing from approximately 3000 BCE until around 1400 CE. For several millennia, it was the majority language of the lower Nile valley until it was gradually displaced by Arabic, leading to its eventual extinction. Today, only the Bohairic dialect of Coptic remains, serving as the liturgical language of the Coptic Orthodox Church. Egyptian-Coptic is classified as the only member of a now extinct branch of Afroasiatic, with its closest relatives being the Semitic and Berber languages (Schenkel, 1990; Grossman and Richter, 2015). Its placement within the Afroasiatic language family has recently become a topic of renewed debate (Almansa-Villatoro and Štubňová Nigrelli, 2023). The language history is generally divided into two major phases: Earlier Egyptian, which includes Old Egyptian (2700–2000 BCE) and Middle Egyptian (2000–1400 BCE), and Later Egyptian, which encompasses Late Egyptian (1350–600 BCE), Demotic (800 BCE-450 CE), and Coptic (300–1400 CE). Numerous comprehensive linguistic overviews discuss the phonology, morphology, and syntax of the language and its long-term developments (Allen, 2013; Haspelmath, 2015; Loprieno, 1995, 2004; Loprieno and Müller, 2012; McLaughlin, 2022; Müller, 2020; Schenkel, 1990; Stauder, 2020).

According to Egyptological conventions, Egyptian texts (including Demotic) are presented in several layers: (1) in the original script (e.g., as a facsimile, as a handcopy, or printed in a hieroglyphic

Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024), pages 87–97

font) or, in the case of hieratic, transliterated into hieroglyphs, (2) in Egyptological transcription (commonly referred to as transliteration in English), and (3) in translation. In linguistic studies, morphological analyses are often presented as interlinear glosses following the Leipzig Glossing Rules (Di Biase-Dyson et al., 2009). Coptic, using a Greek-based alphabetic script, is usually not transliterated unless it is presented to an audience not familiar with ancient languages (Grossman and Haspelmath, 2015).

Like the native writing systems that do not represent vowels—except for the Coptic script—Egyptological transcription focuses exclusively on consonants. It does not attempt to encode the spellings on a character level, but rather aims to represent the consonantal skeleton (roots). Consequently, distinctions made in the indigenous Egyptian scripts are not captured, leading to a high number of homographs in the scholarly representation of Egyptian, including Demotic (see Figure 1). In response to this, lexicographical projects have adopted lemma IDs in addition to lemma forms, and have established chronolect-specific lemma lists (Egyptian and Demotic: TLA = Thesaurus Linguae Aegyptiae, (Grallert et al., 2024); Coptic: CCL = Comprehensive Coptic Lexicon, (Burns et al., 2020)). As a result, a lemmatizer designed for scholarly purposes must be trained to map tokens to lemma IDs, not just to lemma forms, to effectively integrate with existing digital corpora.

For Coptic, which is typically not transliterated, the issue of homonymy is less pronounced but nonetheless present, often resulting from phonetic changes or only obvious when considering material from several different dialects (see Figure 2).

2.1 Earlier Egyptian

Earlier Egyptian encompasses the chronolects Old Egyptian (Allen, 2015) and Middle Egyptian (Schenkel, 2001). It is classified as a fusional language, characterized by root-and-pattern morphology (roots inflection). The word order is relatively fixed; in sentences with a verbal predicate, the structure follows a V-S-O schema (Loprieno, 1988). Additionally, there are three other sentence types with non-verbal predicates: nominal, adjectival, and adverbial (Loprieno et al., 2017).

Texts from these periods are written either in monumental hieroglyphic or in hieratic, a cursive script. Both scripts are mixed systems that utilize various sign function classes (Polis and Rosmorduc, 2015; Polis, 2023): logograms, mono- or multiconsonantal phonograms, classifiers (traditionally termed determinatives), and interpretants (also known as phonetic complements). Some researchers propose more nuanced categorizations of these sign functions, e.g. by including radicograms (Schenkel, 2003; Polis and Rosmorduc, 2015: pp. 166-167).

Although the Thesaurus Linguae Aegyptiae currently includes almost 1.16 million tokens, a significant number of corpora and texts, while published in print, remain unavailable in digital format. This includes important works such as the Coffin Texts, the Netherworld Books, and the Heqanakhte papyri (letters). Other materials still not digitized include most temple inscriptions or recently discovered texts like the letters from Balat and the Wadi al-Jarf papyri. Additionally, many inscriptions on objects located on-site, in collections and storerooms have yet to be cataloged and are neither available in print nor electronically.

The Earlier Egyptian dataset (TLA-Egy 2024) is derived from the Thesaurus Linguae Aegyptiae, corpus v18, 2023 (Richter et al., 2023). The TLA is the largest digital corpus of Egyptian texts, currently comprising approximately 1.16 million tokens (Grallert et al., 2023). This dataset includes texts from the 3rd to the early 2nd millennium BCE (Old Kingdom to the so-called Second Intermediate Period) across various genres: archival, historical-biographical (royal and non-royal), tomb inscriptions (non-royal), Letters to the Dead, religious texts (Pyramid Texts), literary works (narratives, dialogues, wisdom literature, hymns), magical and medical texts, votive labels and inscriptions, rock inscriptions, and stelae inscriptions (offering formulas). From this corpus, only sentences from the pre-New Kingdom era without emendations, lacunae, questionable readings or questionable translations were selected, ensuring the dataset consists solely of complete sentences from Old and Middle Egyptian. Sentences were further filtered to include only those with fully encoded hieroglyphic spellings and lemmatization. The final dataset comprises 12,773 sentences, totaling 70,267 tokens.

The data is organized in a spreadsheet format, with each sentence displayed on a separate row (tokens are separated by spaces) and various columns providing detailed annotations: hieroglyphic spelling (hieratic script is transliterated

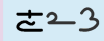
Egyptian (Old–Late Egyptian)			Demotic			Coptic (B = Bohairic, else: Sahidic)		Translation
	mn	69560		mn	d2418	ⲙⲛ-	C1890	there is no (non-existence)
			—	mn	dm733	ⲙⲙⲟⲛ	C1897	really
	mn	69590		mn	d2422	ⲙⲟⲩⲛ	C1913	remain, continue
	mn	69610		mn	d2419	ⲙⲁⲛ (B)	C1900	so and so (a certain person)
	mn	69630		mn	d2424	ⲁⲙⲓⲛ (B)	C89	jug, pot
	mn	70110		mn	d2429			establish, examine
	mn	69660	—	mn	d9203			be ill, suffer
	mn	69670						sick person
	mn	69640						[a kind of fabric]
				mn	dm5140	ⲙⲟⲟⲛⲉ	C1925	pasture, feed
			—	mn	dm7835	ⲙⲛ-	C1903	prefix of neg. imperative
				mn	d2420	ⲙⲁⲉⲓⲛ	C1904	dvine statue; sign, mark
			—	mn	d9297	ⲁⲙⲟⲛⲓ (B)	C91	seize, possess

Figure 1: Homonymy (homography) in Egyptological transcription illustrated by the lemma 'mn'. (Lemma forms and IDs from the Thesaurus Linguae Aegyptiae (TLA) for Earlier Egyptian and Demotic, and from the Comprehensive Coptic Lexicon (CCL) for Coptic. Demotic spellings—written right-to-left— are sourced from the variant list of the Demotic Palaeographical Database Project (Quack et al., 2024).

into hieroglyphs) presented in Unicode[2], Egyptological transcription (following the Leiden Unified Transliteration),[3] lemmatization (including both lemma-ID numbers from the TLA and lemma forms), Part-of-speech tags (UPOS),[4] morphological glossing of the word form (in the following treated as XPOS), and contextual translation into German (translating the entire sentence rather than word-by-word). The dataset also includes the dates (*post quem* and *ante quem*) of the manuscripts and credits to the editors/translators. All annotations have been made by trained Egyptologists. This dataset is published under the CC-BY-SA 4.0 International license.

[2]Currently, not all hieroglyphs are available as Unicode code points. Those not included in the Unicode standard are represented by alphanumeric codes (e.g., Gardiner numbers, JSesh numbers) and enclosed within a tag, e.g., <g>M134</g>.

[3]https://www.iae-egyptology.org/the-leiden-unified-transliteration/

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

2.2 Demotic

The term 'Demotic' refers to the chronolect predominantly used in the second half of the 1st millennium BCE and the early part of the current era, as well as to the cursive script used to write it. Following Alexander the Great's conquest (332 BCE), Greek emerged as the prestige and administrative language, significantly influencing the linguistic environment. Demotic, however, remained dominant in the literary and religious genres as well as for personal communication and in documentary texts. Demotic represents the stage of the language where the evolutionary trends initiated in (late) Middle Egyptian or Late Egyptian fully manifest, such as the shift from a V-S-O to an (AUX-)S-V-O word order (McLaughlin, 2022, pp. 274-275), the analyticization of constructions that were still synthetic in Middle and Late Egyptian, and the (re-)synthetication of Late Egyptian analytic constructions (McLaughlin, 2022). Thus, Demotic exhibits par-

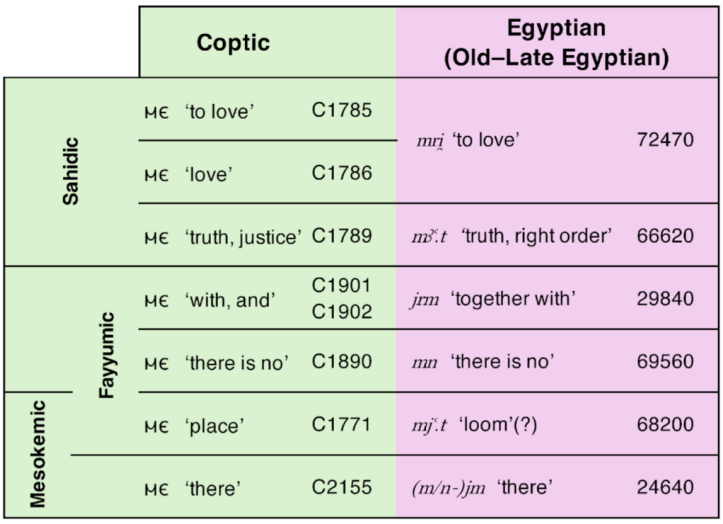


Figure 2: Homonymy in the Coptic dialects Sahidic, Fayyumic and Mesokemic illustrated by a selection of lemmata with the form ⲙⲉ. (Lemma forms and IDs for Earlier Egyptian from the Thesaurus Linguae Aegyptiae (TLA); lemma IDs for Coptic from the Comprehensive Coptic Lexicon (CCL); lemma forms for Sahidic also from the Comprehensive Coptic Lexicon, for Fayyumic and Mesokemic from (Westendorf, 1977)).

tial alignment with both Late Egyptian and Coptic. This dual alignment is reflected in linguistic overviews, where Demotic is often characterized by its similarities to or contrasts with Late Egyptian (Quack, 2006; Winand, 2018) and Coptic (Richter, 2023), respectively.

Despite its significance for understanding the Egyptian Late and Greco-Roman periods, and the substantial amount of material preserved, Demotic remains largely underrepresented in digital corpora. This underrepresentation is attributed to the challenging nature of the material—marked by fragmentation and extremely cursive script—and the limited number of experts capable of editing it. In 1998, Kim Ryholt estimated that since the 1930s, 'less than one per cent of the known material' in the literary corpus had been published (Ryholt, 1998, p. 151). Although many texts have been edited and are available in print since that time, the number of texts available in electronic form remains limited, both for literary and documentary texts.

The Demotic dataset (tla-demotic-v18-premium, TLA-Dem 2024) represents a well-balanced selection of genres, encompassing literary works (narratives, mythological texts, wisdom texts, etc.), religious texts, documentary/administrative records (priestly decrees, temple inventories, letters, receipts, ration lists, among others), legal documents (codes, marriage and divorce settlements, sales deeds, wills, guarantees), graffiti/dipinti, oracular, omen, dream, medical and magical texts, as well as school exercises. Similar to the Earlier Egyptian dataset, this dataset is derived from corpus v18 of the TLA from 2023. It comprises 13,383 sentences totaling 117,314 tokens. The selection, presentation, and licensing criteria mirror those of the Earlier Egyptian dataset, with the exceptions that (1) the tokens are represented exclusively in scholarly transcription ('transliteration'), not in any indigenous script, and (2) XPOS pertains to the lemma, not to the word form. The corpus has been annotated by trained Demotists.

2.3 Coptic

Coptic was the vernacular language during the Christian period in Egypt, while Greek continued to serve as the prestige and administrative language. Following the Arab conquest of Egypt, Arabic began to spread. By the 8th century CE, Greek had been replaced by Coptic in all domains, only to be gradually overtaken by Arabic. During the emergence of Coptic, indigenous writing systems were abandoned in favor of an alphabetic script that included vowels, primarily based on Greek with an addition of 6 or 7 characters borrowed from Demotic, varying by dialect. Coptic does not exhibit root inflection and displays polysynthetic features, including noun incorporation (Grossman, 2019; Miyagawa, 2023). Grammatical morphemes are typically affixed, which categorizes Coptic as an agglutinative language. Particularly in the early centuries CE, the linguistic landscape was marked by significant dialectal variation (Funk, 1988; Richter,

2023). The commonly preferred Coptic word order is (AUX-)S-V-O, and the adjectival sentence pattern has disappeared.

The Coptic data utilized in this study is sourced from the Coptic Scriptorium project (Schroeder and Zeldes, 2016). The corpus, spanning versions 4.2.0 to 4.5.0, primarily comprises Christian literary and biblical texts, along with some letters from a monastic setting in the Sahidic dialect. Available for download in various formats, including CoNLL-U, from the Coptic Scriptorium's GitHub repository, the CoNLL-U formatted data includes 515,142 tokens. The annotation layers in the CoNLL-U files, used for this paper, adhere to the standard CoNLL-U format specifications: ID, form, lemma, Universal POS (UPOS), project-specific POS (XPOS), morphological features, and, to some extent, syntactic head, Universal Dependencies Relation, along with other annotations not pertinent to our study. Unlike the Earlier Egyptian and Demotic corpora, the lemmatization in this corpus maps tokens to surface forms (strings) rather than to IDs, and does not disambiguate homonyms. The numerous Greek loanwords in Coptic are annotated in the same manner like the Egyptian-based vocabulary. The annotation quality varies across three levels: automatic (machine-only annotations), checked (verified for accuracy by a Coptic expert), and gold (extensively reviewed for accuracy). The data is licensed under CC-BY-SA 3.0 and 4.0, except for the 'Sahidica' New Testament sub-corpus, which is copyrighted (c) 2000-2006 by J Warren Wells.

3 Previous Work

Schroeder and Zeldes trained the TreeTagger for POS-tagging and lemmatization, achieving an average accuracy of 95.12% for POS-Tagging and of 96.78% for lemmatization (Zeldes and Schroeder, 2016, 2015), both in ten-fold cross-validation. The same authors implemented a look-up based lemmatizer for the Coptic Scriptorium in Python, which first POS tags the word forms and then assigns the wordform + POS combination to its most common lemma (Schroeder and Zeldes, 2016). As of now, this system does not do disambiguation in case multiple lemmatization options are possible. Smith and Hulden built the first finite-state grammar for Sahidic Coptic (Smith and Hulden, 2016). The lexicon of this implementation comprised 95 verbs, 50 nouns, 65 productive prefixes, 36 closed-class words such as demonstratives and conjuctions, and numerous proper names, all represented in Latin transliteration. The authors reported their system to achieve a recall of 94.6% (precision is not reported), every input word form having 2.9 analyses on average. This implementation does not feature lemma disambiguation either.

SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages had Coptic as one of its languages. The tasks included POS-tagging, lemmatization, prediction of morphological labels and gap filling. In the constrained track that disallowed the use of additional data the best POS-tagger model was reported to have an accuracy of 96.92% (predicting top-1 label) and the lemmatizer an weighted average accuracy of 95.07% over predicting top-1 and top-3 labels (Dereza et al., 2024, Table 5).

4 Preprocessing

For Earlier Egyptian and Demotic we converted the JSON into CoNLL□U. For Coptic, the data was already in the CoNLL□U format, and could be used for BabyLemmatizer as it was.

The Demotic and Earlier Egyptian lemmatization use identifiers to disambiguate between homophonic lemmata. This is necessary, because Demotic and Earlier Egyptian word forms are often ambiguous and short, as already demonstrated earlier in this paper. The identifiers are encoded as integer sequences up to six digits in length, separated from the lemma with a pipe, as in $550034|nfr$. In our initial tests, these sequences seemed to cause slight performance issues for the lemmatizer in terms of accuracy, as accidental incorrect prediction of a single identifier digit resulted into a wrong lemma even if the phonetic part of the lemma was predicted correctly. In addition, it turned out that prediction of long arbitrary integer sequences with no relation to the phonetic form for out-of-vocabulary (OOV) lemmata was very unreliable, rendering predictions for word forms with OOV forms nearly impossible.

To overcome this issue, we compressed the identifiers by replacing them with shorter number sequences tied to the phonological representations of the lemmata. For instance, in the case of a lemma *wr* having four different senses, we enumerated them as *0|wr*, *1|wr*, *2|wr* and *3|wr* instead of using arbitrarily long integer sequences. We based the compressed identifiers on the lemma frequency, 0 having the highest frequency. We hoped that this

decision would make leading zero the most likely prediction for OOV word forms, and therefore, the model would suggest the statistically most probable lemmata for word forms the model has not seen in the training data.[5]

Based on our experiments, identifier compression effectively doubles the accuracy of OOV lemmatization and increases the overall accuracy on average by 3%. After the lemmatization, the original identifiers can be restored by a simple dictionary mapping for all in-vocabulary words with known lemmata. For OOV word forms with previously unseen lemmata, the identifiers have to be defined manually. As Babylemmatizer marks the predictions for OOV word forms automatically in the output CoNLL-U, finding these instances is relatively easy.

Due to character encoding issues with the Egyptian hieroglyphs, we represented them as their Unicode code points in 8-character long sequences separated from each other with a dash symbol.[6] The input encoding will be discussed in a closer detail in the following section.

5 BabyLemmatizer

BabyLemmatizer is a lemmatization and POS-tagging pipeline designed especially for historical languages.[7] It has been optimized for the cuneiform writing system used in Mesopotamia from 3200 BCE to 100 CE, but its tokenizer has been recently extended to also support alphabetic scripts (Sahala and Lindén, 2023).

BabyLemmatizer 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 generated by a two layer unidirectional LSTM decoder with input feeding attention. In our models we use the default batch size of 64 and start the learning rate decay halfway through the training process.

The system is based on the Open Neural Machine Translation Toolkit (Klein et al., 2017) and it handles POS-tagging and lemmatization as machine translation tasks by mapping two sequences of symbols with each other and trying to learn their relation to each other. Examples are given in the following section.

BabyLemmatizer combines the strengths of neural and look-up based lemmatizers by first lemmatizing the input text using the neural network and then using a look-up to verify the labels predicted for all in-vocabulary words. The system also scores the lemmatizations by their confidence, which allows human annotators to first focus on the most likely incorrect lemmata instead of going through the whole dataset. This scoring system is designed for cuneiform languages and has a slightly less relevance for non-logosyllabic scripts, but it still labels the words with scores as shown in Table 1. These scores are included in the output CoNLL-U file.

5.1 Input Encoding

For all models except the Egyptian Hieroglyphic model, we use BabyLemmatizer's alphabetic tokenization, which splits the inputs into character sequences. We use the default context window sizes for POS and lemma prediction: two preceding and two following word forms for POS tagging, and the preceding and following POS tags for lemmatization. Examples of the source and target sequences are shown for the POS tagger in Table 2 and for the lemmatizer in Table 3, using Demotic transliteration.

We use transliteration as input for Demotic because the Demotic script is not supported by Unicode. For Coptic, we use the Unicode representation of the Coptic script. For Earlier Egyptian, which appeared to be the most difficult dataset to annotate, we use two different input formats: transliteration and a concatenation of hieroglyphs and transliteration. In our initial tests, using the hieroglyphic script alone yielded poor results, so we have not reported these results.

We represent hieroglyphs as their Unicode code points in hexadecimal format merged in pairs, the pairs separated from each other with dashes, as in D80CDEA2-D80CDC9D from \ud80c\udea2\ud80c\udc9d. We concatenated these representations in the beginning of the transliterations and used BabyLemmatizer's cuneiform tokenizer to treat the hieroglyphs as monolithic indivisible tokens, but preserving the transliterations as divisible character sequences to retain substring information.

Our motivation for concatenating hieroglyphs and transliteration came from the transliteration of the cuneiform script, where homophonic transliter-

[5] Alternative option would have been to handle the ID sequences as monolithic tokens, but this would have required modifications to the BabyLemmatizer source code.

[6] We had issues reading UTF-16 characters when converting the JSON data into ConLL-U on Windows and had to read them in binary to get the code points.

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

Score	Description of the word form
0 & 1	Reserved for cuneiform languages only (out-of-vocabulary logograms)
2	Out-of-vocabulary (does not occur in training data)
3	Ambiguous (distribution of lemmata assigned for this word form in training data is close to uniform)
4	Slightly ambiguous (of all lemmata given to this word form in training data one occurs 70% of the time.)
5	Likely unambiguous (as in score 4, and occurs in a known XPOS context)

Table 1: Confidence scoring.

Source	= y | (r) | « d y . t » | w y | = f
Target	V

Table 2: POS-tagger input and output label. The center word is enclosed in double angle brackets and the words are separated from each other with pipes.

Source	d y . t P0=PTCL P1=V P2=V
Target	0 | d y

Table 3: Lemmatizer input and output label. The input word form is given first, followed by its POS tag and the POS tags immediately before and after it.

ations are distinguished from each other by adding an index number to indicate which sign was used in the original text (for example, u_2 and u_3 are written using different cuneiform signs despite having the same phonetic value in Akkadian). Since Egyptological transliteration does not use indexing, we hypothesized that adding information about the hieroglyphs would alleviate some of the ambiguity in the transliterations. As reported in the evaluation section, this did not significantly impact the results, but it did improve the out-of-vocabulary (OOV) lemmatization accuracy.

We made various unsuccessful attempts to deal with the ambiguity, especially in the Earlier Egyptian texts, by altering the input and output strings. First, we attempted to use the UPOS tags instead of XPOS tags as context information for the lemmatizer, due to UPOS tags being easier to predict correctly and being simpler. Second, we predicted lemmata without the numeric identifiers alongside the XPOS tags and used these simplified lemmata as context information for predicting the final lemma. Third, we attempted to produce the lemmata with identifiers by using a concatenation of word forms as the input format, taking one or more preceding and following word forms into account.

Finally, we also modified the BabyLemmatizer source code to use a larger context window when predicting POS tags and lemmata for Earlier Egyptian, but this did not improve the results either. In fact, increasing the context window for lemmatization was generally detrimental to accuracy, possibly due to the small dataset, which rendered the model unable to make generalizations based on very long input sequences.

As none of these experiments consistently improved accuracy, we will report only the results for the default BabyLemmatizer settings in the evaluation section.

6 Evaluation

We make a 80/10/10 train/dev/test split of our datasets and evaluate our models using 10-fold cross-validation. We use accuracy as our evaluation metric, that is, the percentage word forms that were assigned the correct label (LEMMA, XPOS, UPOS) by the system. As out baseline, we use a dictionary-based lookup that assigns the word forms with their most common UPOS, XPOS and LEMMA labels (see Table 5). Our final results are summarized in Table 6, confidence intervals of the cross-validation shown in parentheses.

Category	Coptic	Demotic	E. Egy.
XPOS	61	46	234
UPOS	15	11	10
LEMMA	8 557	5 683	6 270
FORM	8 977	7 807	8 109
Tokens	515,142	117,314	70,267

Table 4: Number of unique labels and word forms in our datasets. Earlier Egyptian word form count is based on the number of unique Latin transliterations.

The performance for Coptic is high, but this is partly explainable due to the low number of out-of-vocabulary words, and as for lemmatization, due to the lack of lemma identifiers. Yet, even when the

	Coptic	Demotic	E. Egyptian T	E. Egyptian H+T
XPOS	83.74	87.06	71.52	68.09
UPOS	87.41	88.22	84.99	78.54
LEMMA	90.20	81.19	75.73	71.21

Table 5: Baseline results. Average labeling accuracy (%) over the test sets.

Whole dataset

	Coptic	Demotic	E. Egyptian T	E. Egyptian H+T
XPOS	97.98 (±0.05)	95.14 (±0.13)	88.43 (±0.18)	88.65 (±0.10)
UPOS	97.96 (±0.07)	96.83 (±0.31)	94.32 (±0.22)	94.70 (±0.21)
LEMMA	98.60 (±0.03)	91.40 (±0.20)	85.52 (±0.33)	85.42 (±0.33)
OOV-rate	0.91	3.90	5.90	14.59

OOV word forms only

	Coptic	Demotic	E. Egyptian T	E. Egyptian H+T
XPOS	77.60 (±1.15)	71.11 (±1.53)	59.14 (±1.99)	66.70 (±0.89)
UPOS	75.33 (±2.13)	82.51 (±2.05)	76.88 (±2.15)	82.92 (±1.11)
LEMMA	87.44 (±0.76)	48.16 (±1.57)	50.47 (±1.36)	61.38 (±2.16)

Table 6: Results of the 10-fold cross-validation. OOV-rate shows the average percentage of OOV word forms in the test set in respect to training corpus. E. Egyptian T stands for transliteration and H+T for concatenated hieroglyphs and transliteration. The upper table shows overall results and the lower table the results for OOV word forms only.

number of OOVs are taken into account, the labels seem to be easy to predict compared to our other two datasets. Coptic dataset is also likely easier due to it being almost five times larger than that of Demotic, for instance. The word form to corpus size ratio is thus significantly lower, allowing the system to better learn their relations to the labels in context (cf. Table 4). For bench marking purposes, we also evaluated our system on the SIGTYP 2024 Shared Task dataset for Coptic. Our POS-tagger achieved an accuracy of 94.76% and our lemmatizer an accuracy of 96.20%. Although our POS-tagger underperformed the winner by 2.16%, the performance of our lemmatizer was at least on par with the best implementation, taking into account our system predicted only one label, whereas the best SIGTYP 2024 model's accuracy of 95.07% was based on the average two scores: predicting the correct lemma among the top-3 predictions and predicting only the top-1 lemma (Dereza et al., 2024).

The results for Demotic are on par with those earlier reported for Akkadian, Greek and Latin (Sahala and Lindén, 2023), except for lemmatization that performs slightly worse than expected due to high degree of ambiguity.

Low performance on Earlier Egyptian XPOS tagging is partly explainable by the size of its XPOS label set that also encodes the morphological analysis of the word. This makes the set four times larger than that of Coptic and five times the size of that of Demotic (Table 4). Another factor is the ambiguity of Egyptian word forms, which makes predicting the morphological labels difficult. The ambiguity also affects lemmatization performance, which is untypically low compared to other languages lemmatized with BabyLemmatizer. For UPOS tagging the results are better, but still slightly lower than for our other two datasets.

It seems that using the concatenation of hieroglyphs and transliteration yields slightly better results, but as it increases the portion of OOV word forms, the overall accuracy remains same. Noticeable improvement takes place in OOV lemmatization and POS-tagging, where including information about the hieroglyphs increases the accuracy up to ca. 10% (compare the E. Egyptian T and E. Egyptian H+T results in the lower section of Table 6).

7 Conclusions

We presented models for predicting lemma, UPOS and XPOS labels for Earlier Egyptian, Demotic and Coptic. Our models achieved an accuracy of 88% to 98% for XPOS tagging and 85% to 99% for lemmatization, depending on the input format

and the language in question. We attempted various techniques to improve the accuracy of Earlier Egyptian lemmatization and POS tagging but were unable to achieve significantly better results. We hypothesized that the poor results are likely due to the small corpus size and the proportionally higher number of word form types compared to our other datasets.

Acknowledgments

We wish to 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). We are grateful to Daniel A. Werning (Berlin-Brandenburg Academy of Sciences and Humanities) for extracting the training data from the TLA corpus, discussing it with us, and providing comments on the manuscript of this paper. We also value the feedback from our reviewers which added valuable insights to the paper.

Sources

Coptic Scriptorium 2024: Coptic Scriptorium Corpora v4.2.0–v4.5.0 (downloaded 2023-11-20). Caroline T. Schroeder, Amir Zeldes, et al., Coptic SCRIPTORIUM, 2013-2024, urlhttp://copticscriptorium.org, https://github.com/CopticScriptorium/corpora

TLA-Dem 2024: Thesaurus Linguae Aegyptiae, Demotic sentences, corpus v18, premium, https://huggingface.co/datasets/thesaurus-linguae-aegyptiae/tla-demotic-v18-premium, v1.1, 2/16/2024 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-Egy 2024: Thesaurus Linguae Aegyptiae, Original Earlier Egyptian sentences, corpus v18, premium, https://huggingface.co/datasets/thesaurus-linguae-aegyptiae/tla-Earlier_Egyptian_original-v18-premium, v1.1, 2/16/2024 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.

References

James P. Allen. 2013. *The Ancient Egyptian Language: An Historical Study*. Cambridge University Press, Cambridge.

James P. Allen. 2015. Old Egyptian. In Julie Stauder-Porchet, Andréas Stauder, and Willeke Wendrich, editors, *UCLA Encyclopedia of Egyptology*. Los Angeles.

María Victoria Almansa-Villatoro and Silvia Štubňová Nigrelli, editors. 2023. *Ancient Egyptian and Afroasiatic: Rethinking the Origins*, volume 11 of *Languages of the Ancient Near East*. Eisenbrauns, University Park, PA.

Dylan Michael Burns, Frank Feder, Katrin John, and Maxim Kupreyev. 2020. Comprehensive Coptic Lexicon: Including loanwords from Ancient greek v 1.2.

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.

Camilla Di Biase-Dyson, Frank Kammerzell, and Daniel A. Werning. 2009. Glossing Ancient Egyptian: Suggestions for adapting the Leipzig Glossing Rules. *Lingua Aegyptia*, 17:343–366.

Wolf-Peter Funk. 1988. Dialects wanting homes: A numerical approach to the early varieties of Coptic. In Jacek Fisiak, editor, *Historical Dialectology, Part 1: Regional and Social*, volume 37 of *Trends in Linguistics. Studies and Monographs*, pages 149–192. Mouton de Gruyter, Berlin/New York/Amsterdam.

Silke Grallert, Tonio Sebastian Richter, Simon D. Schweitzer, and Daniel A. Werning. 2023. TLA Text Corpus, corpus issue 18. In Tonio Sebastian Richter, Daniel A. Werning, Hans-Werner Fischer-Elfert, and Peter Dils, editors, *Thesaurus Linguae Aegyptiae*. Berlin-Brandenburgische Akademie der Wissenschaften and Sächsische Akademie der Wissenschaften zu Leipzig. Web-App-Version 2.1.3, Accessed: 5/15/2024.

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

Eitan Grossman. 2019. Language-specific transitivities in contact: The case of Coptic. *Journal of Language Contact*, 12(1):89–115.

Eitan Grossman and Martin Haspelmath. 2015. The Leipzig-Jerusalem transliteration of Coptic. In (Grossman et al., 2015), pages 145–153.

Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors. 2015. *Egyptian-Coptic Linguistics in Typological Perspective*, volume 55 of *Empirical Approaches to Language Typology*. De Gruyter Mouton, Berlin/Munich/Boston.

Eitan Grossman and Tonio Sebastian Richter. 2015. The Egyptian-Coptic language: its setting in space, time and culture. In (Grossman et al., 2015), pages 69–101.

Martin Haspelmath. 2015. A grammatical overview of Egyptian and Coptic. In (Grossman et al., 2015), pages 104–143.

Frank Kammerzell. 2005. Old Egyptian and Pre-Old Egyptian: Tracing linguistic diversity in archaic Egypt and the creation of the Egyptian language. In Stephan J. Seidlmayer, editor, *Texte und Denkmäler des ägyptischen Alten Reiches*, number 3 in Thesaurus Linguae Aegyptiae, pages 165–246. Achet, Berlin.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*.

Antonio Loprieno. 1988. On the typological order of constituents in Egyptian. *Journal of Afroasiatic Languages*, 1:26–57.

Antonio Loprieno. 1995. *Ancient Egyptian: A Linguistic Introduction*. Cambridge University Press, Cambridge.

Antonio Loprieno. 2004. Ancient Egyptian and Coptic. In Roger D. Woodard, editor, *The Cambridge Encyclopedia of the World's Ancient Languages*, pages 160–217. Cambridge University Press, Cambridge.

Antonio Loprieno and Matthias Müller. 2012. Ancient Egyptian and Coptic. In Zygmunt Frajzyngier and Erin Shai, editors, *The Afroasiatic Languages*, pages 102–144. Cambridge University Press, Cambridge.

Antonio Loprieno, Matthias Müller, and Sami Uljas. 2017. *Non-Verbal Predication in Ancient Egyptian*, volume 2 of *The Mouton Companions to Ancient Egyptian*. De Gruyter Mouton, Berlin/Boston.

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

So Miyagawa. 2023. Noun incorporation in Coptic. In Diliana Atanassova, Frank Feder, and Heike Sternberg el Hotabi, editors, *Pharaonen, Mönche und Gelehrte: Auf dem Pilgerweg durch 5000 Jahre ägyptische Geschichte über drei Kontinente: Heike Behlmer zum 65. Geburtstag*, volume 4 of *Texte und Studien zur Koptischen Bibel*, pages 565–574. Harrassowitz, Wiesbaden. Festschrift for Heike Behlmer's 65th Birthday.

Matthias Müller. 2020. Egyptian. In Rebecca Hasselbach-Andee, editor, *A Companion to Ancient Near Eastern Languages*, pages 107–128. Wiley Blackwell, Hoboken, NJ.

Stéphane Polis, editor. 2023. *Guide to the Writing Systems of Ancient Egypt*, volume 4 of *Guides de l'Institut Français d'Archéologie Orientale*. IFAO, Cairo.

Stéphane Polis and Serge Rosmorduc. 2015. The hieroglyphic sign functions: Suggestions for a revised taxonomy. In Hans Amstutz, Andreas Dorn, Matthias Müller, et al., editors, *Fuzzy Boundaries: Festschrift Antonio Loprieno I*, pages 149–174. Kai Widmaier, Hamburg.

Joachim F. Quack, Claudia Maderna-Sieben, Jannik Korte, and Fabian Wespi. 2024. The Demotic Palaeographical Database Project. Accessed: 15 May 2024.

Joachim Friedrich Quack. 2006. En route vers le copte. notes sur l'évolution du démotique tardif. *Faits de langues, Les langues chamito-sémitiques (afro-asiatiques), Vol. 2*, 27:191–216.

Tonio Sebastian Richter. 2023. Coptic. In Andréas Stauder and Willeke Wendrich, editors, *UCLA Encyclopedia of Egyptology*. Los Angeles.

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: 5/13/2024.

Kim Ryholt. 1998. A parallel to the Inaros Story of P. Krall (P. Carlsberg 456+P CtYBR 4513): Demotic narratives from the Tebtunis temple library (I). *Journal of Egyptian Archaeology*, 84:151–169.

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.

Wolfgang Schenkel. 1990. *Einführung in die altägyptische Sprachwissenschaft*. Wissenschaftliche Buchgesellschaft, Darmstadt.

Wolfgang Schenkel. 2001. Middle Egyptian. In Donald B. Redford, editor, *The Oxford Encyclopedia of Ancient Egypt*, volume 2, pages 52–56. Oxford University Press, Oxford.

Wolfgang Schenkel. 2003. *Die hieroglyphische Schriftlehre und die Realität der hieroglyphischen Graphien*, volume 138 of *Sitzungsberichte der Sächsischen Akademie der Wissenschaften zu Leipzig. Philologisch-historische Klasse*. Hirzel, Stuttgart.

Caroline T. Schroeder and Amir Zeldes. 2016. Raiders of the lost corpus. *Digital Humanities Quarterly*, 10(2).

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.

Wolfhart Westendorf. 1977. *Koptisches Handwörterbuch*, 2nd edition. Universitätsverlag Winter, Heidelberg.

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.