

Developing the Egyptian-UJaen Treebank

Roberto Antonio Díaz Hernández,¹ Marco Carlo Passarotti²

¹ University of Jaén (radiaz@ujaen.es)

² Università Cattolica del Sacro Cuore (marco.passarotti@unicatt.it)

Abstract

This paper presents preliminary results of the development of the Egyptian-UJaen treebank, the first dependency treebank created for pre-Coptic Egyptian in Universal Dependencies. It describes the current state of the treebank, explains the approach adopted for the morphosyntactic annotation and discusses some issues concerning the adoption of the CoNLL-U format for the annotation of Egyptian texts. This treebank will surely become a useful linguistic tool for understanding the synchronic and diachronic use of pre-Coptic Egyptian.

1 Introduction

Over the last decade, there has been a growing interest in less-resourced languages that has led to a boom in treebanks for such languages in Universal Dependencies, a useful framework that provides a systematic annotation of grammar across languages (de Marneffe *et al.*, 2023).¹ The creation of the Egyptian-UJaen treebank (henceforth EUJA treebank) aims to contribute to the development of UD by applying the universal inventory of categories developed therein to the morphosyntactic annotation of Egyptian texts. It is the first dependency treebank² created for pre-Coptic Egyptian. Texts are annotated morphosyntactically at the University of Jaén, according to the structuralist approach to Egyptian philology (see Polotsky’s key works, 1944, 1976 and Schenkel’s, 2012).

The EUJA treebank started as UD release 2.14 on 15 May 2024 with 5,515 words and 707 sentences from Old Egyptian texts. The data and results of the present paper are based on the current

state of the treebank consisting of 1,573 sentences and 14,650 words (UD release 2.15 to appear on 15 November 2024).

The aim of this paper is to describe the methodology used in the development of the EUJA treebank. It provides a brief overview of Egyptian and its scripts (2) and a description of the sources selected for the treebank (3). There follows a discussion on the annotation of Egyptian texts (4) and the evaluation of an NLP model trained on the treebank (5). Finally, the next stages of the development of the EUJA treebank are outlined in the conclusion (6).

2 Egyptian language and scripts

Egyptian is an Afroasiatic language that knew the following stages:

- 1) Old Egyptian (ca. 2700–2000 BC).
- 2) Middle Egyptian (ca. 2000–1550 BC).
- 3) Late Egyptian (ca. 1550–700 BC).
- 4) Demotic (7th century BC to 5th century AD).
- 5) Coptic (4th century to 14th century AD).

These stages can be classified into Earlier Egyptian, which includes Old Egyptian and Middle Egyptian, and Later Egyptian, which includes Late Egyptian, Demotic and Coptic. While the syntax of Earlier Egyptian is mainly synthetic, Later Egyptian is characterised by an analytic syntax. It should be noted that in the Middle Kingdom (ca. 1980–1760 BC) Old Egyptian was used as a sacred language for the transmission of the Pyramid Texts, even though Middle Egyptian was spoken, while Middle Egyptian became a standardised classical language from the 18th Dynasty (ca. 1539–

¹ <https://universaldependencies.org/>

² For the Coptic treebank in UD see Zeldes and Abrams (2018).

1292 BC) onwards, when other stages of Egyptian were spoken.

Different scripts were used for Egyptian. Hieroglyphs were usually the monumental script for Old Egyptian, Middle Egyptian and eventually Late Egyptian. Hieratic script was mainly used for documents, letters and copies of religious and literary texts in Old Egyptian, Middle Egyptian and Late Egyptian. This script was used exceptionally in monuments and steles. The hieroglyphic and hieratic scripts evolved throughout history, for example the Old Kingdom (ca. 2543–2436 BC) hieroglyphic and hieratic scripts are both different from those used in the New Kingdom (ca. 1539–1077 BC). Finally, Demotic and Coptic were written in Demotic and Coptic script respectively.

The EUJA treebank annotates Egyptian texts using the Tübingen transcription system (see 4.1, below). Hieroglyphs of Old Egyptian, Middle Egyptian and Late Egyptian texts are written in the MISC column (see 4.7, below). The same is planned for Demotic signs. Hieratic texts are transliterated into hieroglyphic script.

3 Sources

Egyptology or rather Egyptian philology is the discipline that studies Egyptian texts. Its official beginning dates back two centuries ago when Jean François Champollion deciphered hieroglyphs in 1822, not without the help of Thomas Young’s earlier attempts. Plenty of textual sources make Egyptian a well-documented ancient language, comparable to Akkadian, Ancient Greek or Latin. Considering such richness it is regrettable that only a handful of universities in the world offer the possibility of studying Egyptology as a fully official degree.

The amount of textual sources for Egyptian depends on their state of preservation. As a rule, the younger the linguistic stage, the more sources there are—Old Egyptian sources are scarcer than those of Middle Egyptian and Late Egyptian. An exception to this is the number of texts written in Classical Egyptian, for it is much larger than for Late Egyptian. Since the aim of the EUJA treebank is to provide a linguistic resource for the morphosyntac-

tic study of pre-Coptic Egyptian, it purports to contain the most representative texts of each stage, namely:

- 1) Old Egyptian: The Pyramid Texts (PT, Sethe, 1908–1922), Old Kingdom and First Intermediate biographical texts (Sethe, 1933 and Clère/Vandier, 1948).
- 2) Middle Egyptian: The Coffin Texts (CT, de Buck, 1935–1961), Middle Kingdom biographical texts (Lange/Schäfer, 1902–1925) and literary texts, such as Sinuhe (Koch, 1990) and the Eloquent Peasant (Parkinson, 1991).
- 3) Classical Egyptian: The Book of the Dead (BD, Naville, 1886), 18th Dynasty biographical texts (Sethe, 1906–1909 and Helck, 1955–1958) and literary texts, such as Neferti (Helck, 1970) and Ipuwer (Enmarch, 2008)
- 4) Late Egyptian: New Kingdom biographical texts (Kitchen, 1975–1990) and literary texts (Gardiner, 1932).
- 5) Demotic: Literary texts, such as the teaching of Onchsheshonqy (Glanville, 1955).

Several editions of these Egyptian texts were published in the first half of the twentieth century and are now available online as pdf files, such as the Coffin Texts.³ As the linguistic usage of Egyptian varies not only in sources from different stages, but also in some sources from the same period, each sentence in the EUJA treebank is assigned a bibliographic reference and an ID in order to identify and classify all sentences by source. The ID consists of the acronym EUJA, followed by a hyphen and a numeral, for example EUJA-1. Each sentence is also provided with a reference indicating the exact paragraph in the original text, its origin, date, the genre and source’s language stage, for example:

sent-id =EUJA-44

ref = PT § 1a T, Saqqara, 6th Dynasty, rel, OE⁴

EUJA-1 is a test sentence. EUJA-2 to 43 are multiword expressions taken from various Old Egyptian text corpora.

The systematic annotation of the Pyramid Texts begins with EUJA-44. 14,404 words, correspond-

³ <https://isac.uchicago.edu/research/publications/oriental-institute-publications-oip> (OIP 34, 49, 64, 67, 73, 81, 87 and 132).

⁴ The abbreviation “rel” stands for religious text, and “OE” for Old Egyptian.

ing to 1/5 of the whole corpus, have been annotated in the treebank, which means that the Pyramid Texts consist of 60,000–80,000 words.

From the beginning of the EUJA treebank the question whether to create a repository for each stage or for all stages of pre-Coptic Egyptian was discussed with Daniel Zeman.⁵ The latter option was chosen in order to have an overview of the evolution of Egyptian in a single CoNLL-U file.




If particular linguistic features of a text corpus or a stage are to be studied, the corpus name, e.g. Pyramid Texts, or the stage name, e.g. OE (Old Egyptian), should be mined to find all instances that match the search. The README file also contains a classification of the sentences according to the stage of Egyptian and the text corpus in order to facilitate searching:

sent_id = EUJA-	language and text corpus
1–1573	Old Egyptian
1, 3, 4, 11–15, 23, 25, 26, 30–34, 36–40, 43	biographical texts
2, 6, 7, 9, 10, 16–21, 24, 27–29, 41, 42, 44–1573	Pyramid Texts
5, 35	Letters to the Dead
8, 22	Captions to everyday life scenes

Table 1: Sentence classification in the treebank

4 Annotation

4.1 Transcription

Egyptian scripts consist of phonetic signs and classifiers. Phonetic signs are reproduced by means of transcription characters to make reading easier. Classifiers are signs that give information about a word (Goldwasser, 2022: 192), for example ⁶ is a classifier in the word   *ht(i)* “travel downstream”.

Some colleagues attending the 13th International Conference of Egyptologists held in Leiden from 6 to 11 August 2023 established the Leiden Unified Transliteration (LUT),⁷ and there has been constant pressure since then to adopt it for the transcription of Egyptian texts in digital resources, text editions and publications. However, the LUT is clearly a scientific regression, as it keeps traditional

signs, such as *t̄* and *q̄*, which were adopted in the 19th century only for typographical reasons.

The Tübingen transcription system (Schenkel, 2012: 19–25; Schneider, 2023: 4)⁸ has been followed for the annotation of Egyptian texts in the EUJA treebank, for its suitability for linguistic analysis; for example, as in the Slavic languages *č* stands for the sound /tʃ/, whereas the LUT’s *t̄* may confuse a linguist for it is used to transcribe /θ/ (ث) in Arabic. A table with both systems and the Unicode codes used for the transcription signs in the EUJA treebank is included in the appendix (see table 3, below).

As is usual with sources of extinct languages, Egyptian texts occasionally contain gaps and errors, which must be noted in their transcription. The Leiden system for editing ancient texts is consequently used in the EUJA treebank (Schenkel, 2012: 28–29). It includes the following critical signs:

1. Brackets () add a conventionally omitted element, for example the suffix pronoun of the first person singular *ʾt* is usually omitted in Old Egyptian as vowels or weak consonants such as *ʾt* were not written.
2. Square brackets [] enclose a restored text that was missing.
3. Curly brackets {} enclose typographical errors, for example the reduplication of a consonant (i.e. dittography). Such errors are labelled as Typo according to the CoNLL-U format in the EUJA treebank.
4. Angle brackets <> add an element that has been erroneously omitted from the text, for example a missing consonant due to haplography.

4.2 Sentence splitting

It is generally assumed that no punctuation marks are used in Old Egyptian and Middle Egyptian. However, the annotation of the Pyramid Texts in the EUJA treebank has revealed that a line is occasionally used as a punctuation mark (see fig. 1, below) to indicate the end of a spell (e.g. EUJA-1309) or to separate a recitation text from a ritual remark (e.g. EUJA-178). The line in the hieroglyphic text is transcribed by means of the vertical bar (|).

⁵ Thanks to Daniel Zeman for his support.

⁶ The hieroglyphs used in this paper are drawn from the hieroglyphic text processor JSesh.

⁷ <https://www.iae-egyptology.org/the-leiden-unified-transliteration>

⁸ See also Rössler, 1971: 263–326.

bank, the lemmata of derivatives, such as nisba adjectives, are the words from which they are derived, for example the lemma of the nisba adjective¹¹ *im.t* “one who is in” is the preposition *m* “in”. Likewise, participles, relative forms and infinitives are lemmatised after the verb stem, for example the passive participle *mr.y* “beloved” corresponds to the lemma *mr.i* “love”. Causative verbs are also lemmatised after the verb stem without the causative prefix, for example *š:w'b* “make pure” (i.e. “purify”, “cleanse”) corresponds to the lemma *w'b* “be pure”.

4.5 Universal Part-of-Speech tags and Morphological analysis

Fifteen Universal Part-of-Speech tags (cf. Petrov *et al.*, 2012) are documented in Old Egyptian according to the current state of the EUJA treebank.¹²

- 1) Adjective (ADJ; 528/3.60%): There are a few primary adjectives, for example *nb* “every”, “all”. Most of them are deverbal adjectives such as *nfr* “be good” and nisba adjectives such as *im(.t)* “one who is in”, derived from the preposition *m* “in”. In an attributive function, adjectives usually agree in gender and number with the noun they follow. The boundary between adjective and noun is occasionally diffuse in Old Egyptian, as it is unclear if a nisba is used as an adjective in an attributive function or as a noun in apposition.
- 2) Adverb (ADV; 46/0.31%): This part of speech is only used sporadically. Among the Old Egyptian adverbs, *im* “there” is common in the Pyramid Texts, although it is occasionally unclear whether it is the adverb *im* or the preposition *m* in *status pronominalis* with an omitted suffix pronoun. Instead of adverbs, adpositions (ADPs; 1,901/12.98%) are usually used in Old Egyptian, consisting of a preposition and a noun phrase. Nouns with an adverbial function, such as *č.t* “eternally” or *hrw* “day” are also found in Old Egyptian.
- 3) Interjection (INTJ; 66/0.45%): *hʹ* “O” and *i* “O” are interjections common in the Pyramid Texts. They precede a noun and have a vocative function, for example *hʹ Wnʹs* “O Unas”.
- 4) Noun (NOUN; 4,036/27.55%): There is no case distinction of nouns in Egyptian scripts.¹³ They have two genders, masculine and feminine. The ending *t* is used to mark the feminine gender and to form the neuter gender, especially in participles and relative forms, for example *nfr.t* “that which is good” i.e. “(the) good”. Nouns have three numbers, singular, plural, and dual.
- 5) Proper Noun (PROPN; 1,788/12.20%): Names of deities, kings and mythological places are common in the Pyramid Texts. All of them are annotated as PROPN.
- 6) Verb (VERB; 2,490/17.00%): The EUJA treebank follows the structuralist approach, reinforcing and developing Polotsky’s theoretical framework of the Egyptian verbal system. In Old Egyptian, there are two verb conjugations, the “suffix pronoun conjugation” (SPC) and the “Old Semitic suffix conjugation” (OSSC). The former needs a noun or a suffix pronoun as a subject in a similar way as non-pro-drop languages, such as English. Most of the exceptions to this rule are due to phonographic reasons. The OSSC consists of personal endings added to the verb stem similar to the verbs of pro-drop languages, such as Spanish. The SPC is based on a system of tenses:¹⁴ the past I *ščm ʹf* (SPC= Past-1), the past II *ščm.n ʹf* (SPC=Past-2), the present *ščm ʹf* (SPC=Pres), the future *ščm ʹf* (SPC=Fut), the bireferent future *ščm.t ʹf* (SPC=Bi-Fut)¹⁵ and the contingent tenses *ščm.in ʹf* (SPC=ContPast), *ščm.hr ʹf* (SPC=ContPres) and *ščm.kʹ ʹf* (SPC=ContFut).¹⁶ The SPC also has the subjunctive mood *ščm ʹf*

¹¹ In Semitic languages, such as Arabic, “nisba” is used to label an ending added to nouns, and rarely to prepositions and pronouns, to form (relative) adjectives and nouns (Schulz 2010, 86). The addition of the nisba ending to prepositions to form adjectives and nouns is a common feature in Egyptian.

¹² The absolute and relative frequency of each part of speech is given between brackets.

¹³ The genitive case is expressed by two consecutive nouns (“direct genitive”) or by the adjective nisba *n.i* “belonging to” (“indirect genitive”).

¹⁴ The keys in brackets are used in the XPOS column of the treebank.

¹⁵ The bireferent future has two reference points in time, one in the past and one in the future.

¹⁶ The contingent tenses are conditioned on the verbal action of the main clause.

(SPC=Sub). The identification of the present, future and subjunctive is usually circumstantial, depending on the context, since strong verbs in hieroglyphic writing have the same consonant spelling. The impersonal construction (Imprs=Man) corresponding to “one” in English is rendered by adding the noun *tt* / *tw* to the SPC verb form, for example:

n fh-tt n zk (EUJA-658)

“No one got rid of you.”

In addition, there are two passive verb forms in the SPC, the past passive *ščm.w šf* (SPC=PastPass) and the future passive *ščmm šf* (SPC=FutPass). The past II *ščm.n šf*, the present *ščm šf*, the future *ščm šf* and the passive forms can be used as abstract relative verb forms (Type=Abstrel), i.e. nominal finite verb forms used syntactically as nouns, especially in the emphatic construction, the Egyptian cleft sentence with an adverbial phrase as focus, for example:

pr.n šf hr ir.t Hr.w (EUJA-248)

“It is with the Eye of Horus that he came forth.”

The SPC may consist of adjective finite verb forms, known as “relative verb forms” (VerbForm=Relform), which match the gender and number of the antecedent, for example:

Wšr(.w) Wniš m n zk ir.t Hr.w šnm.tn šf (EUJA-222)

“Osiris Unas, take the Eye of Horus, which he rejoined.”

There are syntactic rules for the use of the OSSC in relation to SPC tenses. Thus, the tense, aspect and mood of the OSSC varies according to its syntactic function. The Early Egyptian verb system has an imperative (Imp) and infinite verb forms. The infinitive (Inf) is the nominal infinite verb form, as opposed to the nominal finite verb forms i.e. the abstract relative verb forms. In addition, there are two adverbial infinitives, the so-called negatival complement (NegCom) and the complementary infinitive (ComplInf). Participles (Part) are adjective infinitive verb forms as opposed to the adjective finite verb forms i.e. the relative forms. Both participles and relative forms are occasionally used as nouns.

- 7) Adposition (ADP; 1,901/12,98%): In Old Egyptian, adpositions are usually prepositions used before a noun. Prepositions occasionally show different spellings in status pronominalis (Status=Pron) and status constructus (Status=Cons), for example *im* (Status=Pron) and *m* (Status=Cons) “in”. Complex prepositions such as *m-ʿ* “in the hand” i.e. “from” are considered multiword expressions (MWEs). Old Egyptian also knows the use of postpositions, for example *is* “like”.
- 8) Auxiliary (AUX; 45/0.31%): The particle *tw* is considered an auxiliary as it is used to express the present perfect in combination with the past II *ščm.n šf* and the habitual aspect with the present *ščm šf*, for example: *tw rč.n (št) t' n hkr* (EUJA-1) “(I) have given bread to the hungry.” *tw phr n šf hʿ(.w)* (EUJA-1274) “Thousands (usually) serve him.”
- 9) Coordinating Conjunction (CCONJ; 8/0.05%): The use of CCONJs in Old Egyptian is exceptional. In the current state of the Egyptian-UJaen treebank, only *isč* “and” is attested as CCONJ (e.g. EUJA-548).
- 10) Determiner (DET; 369/2.52%): No articles are used in Old Egyptian. There are four types of demonstrative pro-adjectives (Dem) with three genders, masculine, feminine and neutral.
- 11) Numeral (NUM; 159/1.09%): There are ordinal and cardinal numbers in Egyptian. While “1” and “2” are adjectives, the remaining cardinals are nouns. Ordinal numbers usually follow a noun as attributives.
- 12) Particle (PART; 288/1.97%): Old Egyptian has many particles, three types of which are present so far in the EUJA treebank—negative particles (*n* and *ny*), emphatic particles (*tn*, *is*, *wnn.t*, *hm* and *mi*) and modal particles (*ʿ* and *my*).
- 13) Pronoun (PRON; 2,708/18.48%): There are three types of personal pronouns in Old Egyptian—the independent (IndPron), dependent (DepPron) and suffix (SFP). The keys to the three types are annotated in the XPOS column of the EUJA treebank.
- 14) Subordinating conjunction (SCONJ; 4/0.03%): Two SCONJs have been annotated so far in the EUJA treebank, *n-n.tt*

“because” (UDE-385) and *wn.t* “that” (UDE-1380).

- 15) Symbol (SYM): Although no symbols are found in the current state of the EUJA treebank, some signs may have been used as symbols in exceptional cases.

4.6 Universal Dependency Relations

Nominal core arguments (nominal and clause subject, object and indirect object), non-core arguments (oblique nominal, vocative, expletive and dislocated element) and nominal dependents (nominal modifiers, appositional modifier and numeric modifier) are widespread in Egyptian. It should be noted that the vocative is usually used in the Pyramid Texts (418/2.85%), as these are ritual texts addressed to the deceased king.

The dependency relation between verbal clauses is often established by “adordination” (Díaz Hernández, 2013: 5, footnote 20), i.e. the syntactic dependency relation caused by a temporal reference of the verb form in the “adordinate” clause:

mḥ-ib n(.t) nšw ḥnt ʃ (EUJA-32)

“One who earns the king’s trust (i.e. king’s confidant) (when) he sails upstream.”

Here the present tense *ḥnt ʃ* “he sails upstream” is syntactically dependent on the head of the preceding clause because of the temporal reference of the verb form.

The current state of the EUJA treebank also contains cases of modifier words and function words. The three types of universal modifier words are adverbial modifiers (172/1.17%, e.g. the negative particle *n* in EUJA-1072), discourse elements (151/1.03%, e.g. the particle *m ʃk* in EUJA-916) and adjectival modifiers (500/3.71%, e.g. the adjective *nfr.t* in EUJA-923). Among the function words, Old Egyptian has the particle *tw* used as an auxiliary (45/0.31%, e.g. EUJA-1), the demonstrative determiner *pt* or *pw* used as a copula (96/0.66%, e.g. EUJA-417), markers (54/0.37) such as the subordinating conjunction *wn.t* (EUJA-1380), determiners (246/1.68%) and prepositions usually used to mark a case of relation. In Egyptian, classifiers are not words, but signs that provide semantic information about the word they accompany.

Conjuncts (293/2.68%) are usually connected to other elements without coordinating conjunctions. The “fixed” relation is only used for com-

plex prepositions (111/0.76%), such as *m-ḥt* “behind” and the flat relation for names consisting of two or more elements, for example *Ḥr.w-nḥn(.y)* “Horus of Nekhen”. Egyptian multiword expressions are not annotated as elements in a “fixed” relation because they are expressions with an idiosyncratic meaning whose morphological and syntactic structure can change.


No list has yet been annotated in the EUJA treebank, although chains of items are found in Egyptian inventories.

Parataxis (303/2.07%) is a common relation, as it is used in reported speech (e.g. EUJA 973) and to link pairs of sentences in the so-called “balanced sentence” (e.g. EUJA-645).



The “orphan” relation used to indicate ellipsis is documented in the EUJA treebank (15/0.10%, e.g. EUJA-916) and combinations of lexemes considered morphosyntactically as single words are annotated as compounds (93/0.63%), for example *psč.t-nčš.t* “Little Ennead”. Unspecified dependency (dep; 48/0.33%) occurs when the relation between words cannot be determined due to the absence of vowels in the hieroglyphic script, for example in the offering formula (EUJA-168).

4.7 Hieroglyphs

Hieroglyphs have been annotated manually (over 15,000 signs) using Unicode characters in the MISC column. When hieroglyphs are omitted for phonographic or conventional reasons, the key “Hiero=No” is annotated.

It should be noted that the Unicode extended repertoire of Egyptian hieroglyphs and control characters are still not supported by computer systems (Suignard, 2023 and Glass et al., 2021). Thus, only hieroglyphs from Gardiner’s list are annotated (Gardiner, 1957: 438–548). The key “UC_No” means that there is no Unicode character for a given hieroglyph, whereas when a Unicode character for a hieroglyph of the extended repertoire (Suignard, 2023) cannot be used because still under development, its code is annotated with the key “UC_Code”, for example  is annotated “UC_1397B”.

As Unicode control characters cannot be used yet to arrange hieroglyphs, they are annotated using the following signs:

1. Colon (:) to indicate subordination of signs, for example  corresponds to  *pn* “this”.

2. Brackets () to segment groups of hieroglyphs.
3. Asterisk (*) to indicate the juxtaposition of hieroglyphs, for example (ⲡ*ⲟⲓ)ⲙ corresponds to ⲙⲡⲟⲓ *p.t* “sky”.

5 Training and Evaluating an NLP model

We used the CoNLL-U file containing the EUJA treebank to train a model of UDPipe 1 (Straka *et al.*)¹⁷ to automatically perform tokenisation, morphological analysis, part-of-speech tagging, lemmatisation and dependency parsing. The test set consisted of 160 sentences chosen randomly. These sentences were EUJA-181–200, 351–370, 491–510, 671–690, 811–830, 960–979, 1221–1240 and 1431–1450. The training set consisted of the remaining 1,413 sentences. Table 2 shows the results of the evaluation process.¹⁸

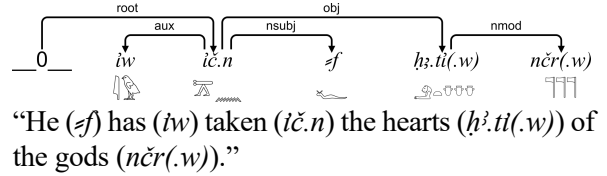
Metric	F1 Score
UPOS	90.30
XPOS	76.01
UFeats	75.87
AllTags	65.39
Lemmata	89.38
UAS	82.52
LAS	71.97
CLAS	69.13
MLAS	56.14
BLEX	63.27

Table 2: Evaluation of an NLP model trained on the treebank

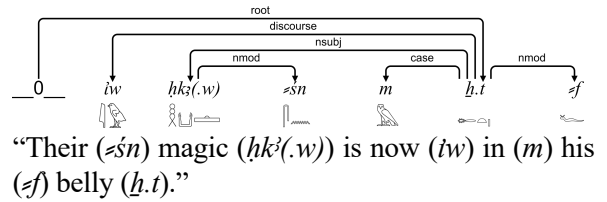
This table shows promising results as all categories get an F1 score over 50. The accuracy of lemmata (89.38) and Universal Part of Speech tags (UPOS: 90.30) is especially high. The Labeled Attachment Score (LAS), the Bilexical Dependency Score (BLEX), the Language Specific Part of Speech tags (XPOS) and the Morphological Features (UFeats) show an F1 score of between 60.00 and 80.00.¹⁹ The Morphology-Aware Labeled Attachment Score (MLAS) is the only category with a F1 score between 50.00 and 60.00.

The UDPipe 1 trained model usually provides a high accuracy rate on UPOS tags, especially nouns and nisba adjectives. As for parsing, it can automatically and accurately annotate short sentences, for example EUJA-1280 and 1287:

EUJA-1280:

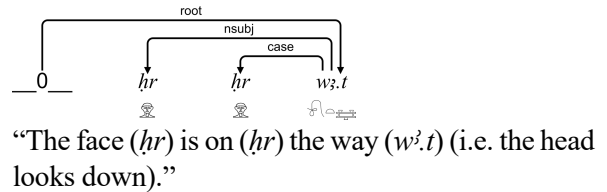


EUJA-1287:



The trained model reveals to be sufficiently good in assigning the correct morphological features and dependency relations to two words with the same spelling, for example:

EUJA-1324:



6 Conclusion

When Joris F. Borghouts published his Middle Egyptian grammar in 2010, with a large number of examples and references to Egyptian texts, Wolfgang Schenkel predicted that a digital database of syntactically analysed sentences would be the next step in Egyptian philology.²⁰ The EUJA treebank makes Prof. Schenkel’s prediction come true, as it contains morphosyntactically annotated sentences from the most representative texts of each pre-Coptic stage. It will be an auxiliary tool for the study of Egyptian grammar, facilitating the synchronic and diachronic parsing of structures and words.

The development of the EUJA treebank includes two further phases:

¹⁷ <https://ufal.mff.cuni.cz/udpipe/1>

¹⁸ The evaluation was performed using the `eval.py` script provided among other UD tools at <https://github.com/UniversalDependencies/tools>

¹⁹ For LAS and UAS see Buchholz and Marsi, 2006. For CLAS, MLAS, BLEX see Zeman *et al.*, 2018.

²⁰ Personal communication to R.A.D.H.

1) Annotation of the remaining part of the Pyramid Texts.

2) Annotation of the Old Kingdom and First Intermediate Period biographical texts.

Once these corpora are annotated, the treebank will certainly hold over 100,000 Old Egyptian words, and annotation of the Middle Egyptian corpus will begin.

Acknowledgments

This paper is the result of a three-month Short Term Scientific Mission (STSM) Grant awarded to Roberto Antonio Díaz Hernández by “UniDive” (COST Action 21167). The STSM was carried out under the supervision of Marco Carlo Passarotti at the *Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell’Espressione* (CIRCSE) at the *Università Cattolica del Sacro Cuore* in Milan from 1 May to 31 July 2024, with funding from the European Union. Thanks to the organisers of “UniDive” for this grant, and to Flavio Cecchini, Amir Zeldes and Daniel Zeman for discussing the use of Egyptian classifiers and Unicode characters in an issue published in June 2024 on the GitHub website of Universal Dependencies.

References

- Joris F. Borghouts. 2010. *Egyptian. An Introduction to the Writing and Language of the Middle Kingdom*. Peeters, Leuven.
- Sabine Buchholz, Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. New York City and Association for Computational Linguistics: 149–164.
- Adriaan de Buck. 1935–1961. *The Egyptian Coffin Texts*, (7 vols.). The University of Chicago Press, Chicago.
- Jacques J. Clère and Jacques Vandier. 1948. *Textes de la Première Période Intermédiaire et de la XI^{ème} Dynastie*. Bibliotheca Aegyptiaca X, Brussels.
- Roberto A. Díaz Hernández. 2013. *Tradition und Innovation in der offiziellen Sprache des Mittleren Reiches*. Wiesbaden.
- Roberto A. Díaz Hernández. 2021. The Man-impersonal *šçm.n-ti/tw(šf)* Form in Earlier Egyptian. *Lingua Aegyptia* 29: 37–59.
- Roberto A. Díaz Hernández. 2022. The Man-impersonal Verb Forms of the Suffix Pronoun Conjugation in Earlier Egyptian. *Lingua Aegyptia* 30: 25–90.
- Roland Enmarch. 2005. *The Dialogue of Ipuwer and the Lord of All*. Griffith Institute, Oxford.
- Adolf Erman and Hermann Grapow (eds.) 1926–1961. *Das Wörterbuch der Aegyptischen Sprache*, (5 vols.) Berlin.
- Hans-W. Fischer-Elfert. 2021. *Grundzüge einer Geschichte des Hieratischen*, (2 vols). Lit, Berlin.
- Alan H. Gardiner. 1932. *Late-Egyptian Stories*. Bibliotheca Aegyptiaca 1, Brussels.
- Alan H. Gardiner. 1957. *Egyptian Grammar. Being an Introduction to the Study of Hieroglyphs*. Griffith Institute, Oxford.
- Andrew Glass (et al.) 2021. Additional control characters for Ancient Egyptian hieroglyphic texts. <https://www.unicode.org/L2/L2021/21248-egyptian-controls.pdf>
- Orly Goldwasser. 2022. L’écriture énigmatique: distancée, cryptée, sportive. In Stéphane Polis (ed.) *Guide des écritures de l’Égypte ancienne*. Cairo: 192–199.
- Stephen R. Glanville. 1955. *The Instructions of ‘Onchsheshonqy* (British Museum Papyrus 10508), (vol. 2). London.
- Wolfgang Helck. 1955–1958. *Urkunden der 18. Dynastie*, (vols. 17–22). Berlin.
- Wolfgang Helck. 1970. *Die Prophezeiung des Nfr.tj. Otto Harrasowitz, Berlin*.
- Kenneth A. Kitchen. 1975–1990. *Ramesseid Inscriptions. Historical and Biographical*, (8 vols.) Oxford.
- Roland Koch. 1990. *Die Erzählung des Sinuhe*. Bibliotheca Aegyptiaca XVII, Brussels.
- Hans O. Lange and Heinrich Schäfer. 1902–1925. *Catalogue général des antiquités égyptiennes du Musée du Caire. Grab und Denksteine des Mittleren Reichs*, (4 vols.). Berlin.
- Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre and Daniel Zeman. 2021. Universal Dependencies. Association for Computational Linguistics, 47(2): 255–308. https://doi.org/10.1162/COLI_a_00402.
- Edouard Naville. 1886. *Das aegyptische Tottenbuch der XVIII. bis XX. Dynastie*. Berlin.
- Richard Parkinson. 1991. *The Tale of the Eloquent Peasant*. Oxford.
- Slav Petrov, Dipanjan Das and Ryan MacDonald. 2012. A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 12)*: 2089–2096.

Alexandre Piankoff. 1968. *The Pyramid of Unas*. New York.

Hans J. Polotsky. 1944. *Études de syntaxe copte*. Publications de la Société d'Archéologie Copte, Cairo.

Hans J. Polotsky. 1976. Les transpositions du verbe en égyptien classique. *Israel Oriental Studies* 6: 1–50.

Otto Rössler. 1971. Das Ägyptische als semitische Sprache. In Franz Altheim and Ruth Stiehl (eds.) *Christentum am Roten Meer*, vol. 1, de Gruyter, Berlin: 263–326-

Wolfgang Schenkel. 2012. *Tübinger Einführung in die klassisch-ägyptische Sprache und Schrift*. Pagina, Tübingen.

Thomas Schneider. 2023. *Language Contact in Ancient Egypt*, Lit, Berlin.

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

Kurt Sethe. 1908–1922. *Die altägyptischen Pyramidentexte nach den Papierabdrücken und Photographien des Berliner Museums*, (4 vols.). Leipzig.

Kurt Sethe. 1906–1909. *Urkunden der 18. Dynastie*, (vols. 1–16). Leipzig.

Kurt Sethe. 1933. *Urkunden des Alten Reichs*. Leipzig.

Milan Straka (et al.) 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 16)*: 4290–4297.

Michel Suignard. 2023. Encoding proposal for an extended Egyptian Hieroglyphs repertoire. <https://www.unicode.org/L2/L2023/23181-n5240-hieroglyphs.pdf>

Amir Zeldes and Mitchell Abrams. 2018. The Coptic Universal Dependency Treebank. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Brussels and Association for Computational Linguistics: 192–201.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels and Association for Computational Linguistics: 1–21.

A Appendix

	LUT	Tübingen	Unicode
	ʒ	ʒ	A723
	i	i	A7BD
	y	y	
	ĩ	ĩ	00EF
	‘	‘	A725
	w	w	
	b	b	
	p	p	
	f	f	
	m	m	
	n	n	
	r	r	
	h	h	
	ħ	ħ	1E25
	ḥ	ḥ	1E2B
	ḥ	ḥ	1E96
	z	s	
	s	ś	015B
	š	š	0161
	q	ḳ	1E33
	k	k	
	g	g	
	t	t	
	ṯ	č	010D
	d	ṯ	1E6D
	ḏ	č	010D+0323

Table 3: The LUT, the Tübingen transcription system and the Unicode signs used in the EUJA treebank