The Coptic Universal Dependency Treebank

Amir Zeldes and **Mitchell Abrams**
Department of Linguistics, Georgetown University
{amir.zeldes,mja284}@georgetown.edu

Abstract

This paper presents the Coptic Universal Dependency Treebank, the first dependency treebank within the Egyptian subfamily of the Afro-Asiatic languages. We discuss the composition of the corpus, challenges in adapting the UD annotation scheme to existing conventions for annotating Coptic, and evaluate inter-annotator agreement on UD annotation for the language. Some specific constructions are taken as a starting point for discussing several more general UD annotation guidelines, in particular for appositions, ambiguous passivization, incorporation and object-doubling.

1 Introduction

The Coptic language represents the last phase of the Ancient Egyptian phylum of the Afro-Asiatic language family, forming part of the longest continuously documented human language on Earth. Despite its high value for historical, comparative and typological linguistics, as well as its cultural importance as the heritage language of Copts in Egypt and in the diaspora, digital resources for the study of Coptic have only recently become available, while syntactically annotated data did not exist until the beginning of the present project. This paper presents the first treebank of Coptic, constructed within the UD framework and currently encompassing over 20,000 tokens. In this section we give a brief overview of some pertinent facts of Coptic grammar, before moving on to describing how these are encoded in our corpus.

Unlike earlier forms of Ancient Egyptian, which were written in hieroglyphs or hieratic script throughout the first three millennia BCE, Coptic was written starting in the early first millenium CE using a variant of the Greek alphabet, with several added letters for Egyptian sounds absent from Greek. Figure 1 shows the script, which was originally written without spaces (the Greek loan word ⲯⲩⲭⲏ 'psyche' is visible at the top left). Manuscript damage, also shown in the figure, represents a frequent challenge to annotation efforts (see Section 7).

Figure 1: Excerpt from a papyrus letter by Besa, Abbot of the White Monastery in the 5th century, showing text without spaces and a lacuna. Image: Österreichische Nationalbibliothek, http://digital.onb.ac.at/rep/access/open/10099409.

Modern conventions separate Coptic text into multi-word units known as bound groups (Layton, 2011, 19-20) using spaces, based on the presence of one stressed lexical item in each group. This leads to multiple units being spelled together which would normally receive separate tokens and part of speech tags in annotated corpora. Similarly to languages such as Arabic, Amharic, or Hebrew, simple examples include noun phrases or prepositional phrases spelled together, as in (1), or clitic possessors spelled together with nouns, as in (2).[1]

(1) ϩⲙ̀ⲡⲣⲁⲛ hm-p-ran 'in-the-name'

(2) ⲣⲛⲧⲉⲕ rnt=k 'name-your (SG.M)'

However, Coptic fusional morphology can be much more complex than in Semitic languages, for several reasons. Developing from a morphologically rich synthetic language through an analytic phase in Late Egyptian, Coptic has fusional morphology and is usually seen as an agglutinative

[1] We follow common Egyptological practice in separating lexical items within bound groups by '-' and clitic pronouns by a '='.

or even polysynthetic language (Loprieno, 1995, 51). Similarly to inflection in Hausa, auxiliaries and clitics attach to verbs as in (3), and unlike in Semitic languages, compounds are spelled together and do not allow intervening articles. The language also exhibits frequent verb-object incorporation, complicating word segmentation for tokenization (see Grossman 2014), as in the complex verb shown in (4). Such complex verbs can be embedded in word formation processes, leading to nominalizations such as (5).

(3) ⲁ-ϥ-ϩⲱⲧⲃ ⲙ̀ⲡ-ⲣⲙ̀ⲛ̀ⲕⲏⲙⲉ
a-f-hōtb m̀-p-rmnkēme
PST-3.SG.M-kill ACC-the-Egyptian
'he killed the Egyptian'

(4) ϩⲉⲧⲃ-ⲯⲩⲭⲏ
hetb-psychē
kill-soul
'(to) soul-kill' (incorporated)

(5) ⲙ̀ⲛ̀ⲧ-ⲣⲉϥ-ϩⲉⲧⲃ-ⲯⲩⲭⲏ
mnt-ref-hetb-psychē
ness-er-kill-soul
'soul-killing' (lit. 'soul-kill-er-ness')

Finally, some auxiliaries, such as the optative in (6) may either fuse with and even circumfix adjacent pronouns as in (7), or in some cases exhibit 'zero' forms for pronouns, as in (8).

(6) ⲉⲣⲉ-ⲡ-ⲣⲱⲙⲉ ⲥⲱⲧⲙ̀ ⲉⲣⲟ=ⲕ
ere-p-rōme sōtm ero=k
OPT-the-man hear to-you.2SG.M
'may the man hear you'

(7) ⲉ-ϥ-ⲉ-ⲥⲱⲧⲙ̀ ⲉⲣⲟ=ⲕ
e-f-e-sōtm ero=k
OPT-3.SG.M-OPT-hear to-you.2SG.M
'may he hear you' (circumfix auxiliary)

(8) ⲉⲣⲉ-ⲥⲱⲧⲙ̀ ⲉⲣⲟ=ϥ
ere-sōtm ero=f
OPT+2.SG.F-hear to-him.3.SG.M
'may you hear him' (SG.F subj, fused)

Representing these discontinuous and null phenomena within the UD framework is difficult in the first instance because of their intrinsic complexity (for example, UD prohibits null pronoun nodes, even in enhanced dependencies), but is further complicated by the use of existing standards in Coptic tokenization and tagging, which we present next.

2 Previous work

Of the vast literary, documentary and epigraphic material available in Coptic, print editions have focused on a small subset of early literature in the Sahidic dialect of Upper Egypt, the most prominent of six major dialects (see Shisha-Halevy 1986), which is also considered to be the classical form of the language. While all examples in this paper come from Sahidic sources, we believe that the analyses will generalize well to other dialects, which we intend to approach in the future.

Sizable digital corpora, which have only recently become available in machine readable formats (see Schroeder and Zeldes 2016 on the Coptic Scriptorium project and http://marcion.sourceforge.net/, which provides transcriptions of multiple out of copyright editions) have generally followed the same path of starting with classic Sahidic authors. Other targeted projects have focused on translations from Greek, and especially the Bible, e.g. the Digital Edition of the Coptic Old Testament in Göttingen (Behlmer and Feder, 2017), but also tracking Greek influence in Coptic in general (Almond et al., 2013). Finally Some other projects are advancing the availability of documentary, mostly papyrus materials as well (notably http://papyri.info/), which are as yet only digitized in small quantities.

Although there is a plan to build a constituent treebank of hieroglyphic Ancient Egyptian (Polis and Rosmorduc, 2013), it is as yet unavailable. The UD Coptic Dependency Treebank represents the first dependency treebank for the entire Egyptian language family as well as the only publicly available treebank for Coptic in particular, and for any phase of Egyptian in general.

As a basis for the Coptic Treebank, we selected data from Coptic Scriptorium (available at http://copticscriptorium.org/; see the next section for the specific genres and texts), for two main reasons: 1. the data is freely available under a Creative Commons license, facilitating its re-annotation and distribution; and 2. the data is already tokenized and POS tagged, using a native Coptic POS tagging scheme. Using the Coptic Scriptorium (CS) corpora therefore substantially reduces the required annotation effort, but imposes certain constraints on the segmentation and tagging schemes chosen, which will be presented in Section 4.

source	genre	documents	tokens	sents
translated				
Apophthegmata Patrum	hagiography	1–6, 18–19, 23–26	1,318	62
Gospel of Mark	Bible (narrative)	Chapters 1–6	7,087	248
1 Corinthians	Bible (epistle)	Chapters 1–6	3,571	124
original				
Shenoute, Discourses 4	sermons	Not Because a Fox Barks	2,553	97
Shenoute, Canons 3	sermons	Abraham our Father (XL93-94)	579	26
		Acephalous 22 (YA421-28)	1,703	43
Letters of Besa	letters	Letters 13, 15, 25	1,981	93
Martyrdom of Victor	martyrdom	Chapters 1–6	1,985	88
total			20,777	781

Table 1: Texts and genres in UD Coptic.

3 Texts

The selection of texts for the Coptic Treebank was meant to satisfy four criteria:

1. Data should be freely available
2. A range of different genres should be covered
3. Text types should be chosen which are interesting to users
4. Data should resemble likely targets for automatic parsing using the treebank for training

A dilemma in realizing 3. is that typical UD users interested in computational linguistics, corpus linguistics and language typology may have different interests than Coptologists: the former may prefer texts which resemble other treebank texts or are even available in other languages, such as the Bible, while the latter may be most interested in classic Coptic literature by prominent authors such as Shenoute of Artipe, archmandrite of the White Monastery in the 3^{rd}–4^{th} centuries.

To balance these needs, we decided to include both translated Biblical material and original Coptic works, with a view to allowing comparisons with other languages for which Bible treebanks are available, as well as studies of untranslated Coptic syntax. Table 1 shows the selection of texts currently available in the corpus.

4 Segmentation

While all digital corpora of Coptic referenced in Section 2 separate bound groups, for treebanking purposes we require a more fine grained tokenization. The only tokenization for which NLP tools are available is the one used in the Coptic Scriptorium project, though automatic segmentation accuracy is currently around 94.5% (Feder et al., 2018), meaning that working with data that is already gold-segmented is highly desirable. As a result, the Coptic Treebank inherits some segmentation guidelines, which will be discussed below.[2]

To represent Coptic segmentation correctly, at least three levels of granularity are required: at the highest level, bound groups, which are spelled together, can be regarded as a purely orthographic device, similar to fused spellings of clitics in English, but much more common. To represent these in the CoNLL-U format, we use multi-tokens and the property SpaceAfter=No on non final tokens, as shown in Table 2 for the two bound groups 'in|his|deeds of|soul-killing', which contains the deverbal incorporated noun from (5). This practice corresponds to the same guideline used in Semitic languages, such as Arabic or Hebrew, which use multi-tokens to represent multiword units with a single lexical stress. The second level of granularity corresponds to POS-tag bearing units, which correspond to CoNLL-U tokens.

Finally, for units below the POS tag level, such as components of incorporated 'soul-killing', we

[2]Compatibility with existing resources will motivate several annotation guidelines below; following reviewer comments we suggest this is in keeping with Manning's Law: it offers satisfactory linguistic analysis (rule 1, evidenced by use in existing linguistic studies), allows for consistent human annotation (rule 3, see Section 4 on agreement), and forms a standard comprehensible to and used by non-linguist annotators (rule 5). We also attempt to follow rule 2 in adhering to decisions in other languages to allow for typological comparison where possible. Finally, we have reason to believe the present scheme works well for parsing and downstream NLP tasks (rule 6), though evaluating these is outside the scope of this paper.

text= ... ϩⲛⲛⲉϥϩⲃⲏⲧⲉ ⲙⲙⲛⲧⲯⲧⲣⲉϥϩⲉⲧⲃⲱⲧϫⲏ
transc=... hn|nef|hbēue m|mnt-ref-hetb-psuxē
gloss= ... in|his|deeds of|ness-er-kill-soul
...

12-14	ϩⲛⲛⲉϥϩⲃⲏⲧⲉ	_	_	_	_	_	_	_	_
12	ϩⲛ	in	ADP	PREP	_	14	case	_	Orig=ϩⲛ̄|SpaceAfter=No
13	ⲛⲉϥ	his	DET	PPOS	...	14	det	_	SpaceAfter=No
14	ϩⲃⲏⲧⲉ	deeds	NOUN	N	_	9	obl	_	_
15-16	ⲙⲙⲛⲧⲣⲉϥϩⲉⲧⲃⲱⲧϫⲏ	_	_	_	_	_	_	_	_
15	ⲙⲙ	of	ADP	PREP	_	16	case	_	Orig=ⲙ̄ⲙ̄|SpaceAfter=No
16	ⲙⲙⲛⲧⲣⲉϥϩⲉⲧⲃⲱⲧϫⲏ	soul-killing	NOUN	N	_	14	nmod	_	Morphs=ⲙⲙⲛⲧ-ⲣⲉϥ-ϩⲉⲧⲃ-ⲱⲧϫⲏ

Table 2: Segmentation in CoNLL-U format for a sentence fragment. The lemma column has been filled with glosses for convenience, and features in column 6 have been omitted for space.

use the MISC column to reproduce the morphological segmentation of complex items, as shown in the final column in the example, using hyphens as morpheme separators. Although we considered using sub-tokens to represent incorporation, and using the *compound* relation, we decided against this in order to maintain parity with CS tokens and segmentation practices, and to match up with the practice in Hebrew and Arabic, which use subtokens for constituents of bound groups (and not for smaller units, e.g. portmanteau compounds in both languages[3]). This also allows us to benefit from existing POS tagging software to feed automatic parsing. At the same time, because we have a morphological analysis of complex tokens in the tagged source corpora, we retain this information in the MISC column, and a version of the data instantiating the components as tokens could be produced fully automatically if needed. The MISC column is also used to hold an attribute *Orig* with original forms of tokens as spelled in the source manuscripts, which often deviate from standard spellings or contain added optional diacritics (the word form column is always normalized). As a result the data can be used to train automatic normalization tools.

A further complication arises in the case of fused auxiliaries and pronouns, as in the cases from examples (7) and (8). Here too, a solution splitting the fused form into three tokens would be conceivable, in order to represent the circumfix auxiliaries. However, CS guidelines do not tokenize such units apart, instead using portmanteau tags such as AOPT_PPER (optative auxiliary, fused with personal pronoun), and a lemma joining the lemmas of both units via an underscore. A potential pitfall of splitting these units is that, if we consider a form such as *e-f-e* to consist of three tokens, there is a chance that automatic taggers and parsers will tag one of the two 'e' vowels correctly as an auxiliary, but not the other, leading to an incoherent analysis.[4] The token *efe*, by contrast, will always receive a single tag, and since the form is unambiguous, it will always be correct. While we would not prioritize ease of tagging over an adequate linguistic analysis, we feel that, coupled with the desire to maintain parity with larger corpora, Manning's Law favors this analysis, which is unambiguous, deterministic and easy to convert into a different form if necessary using the native XPOS tags.

We therefore decided to retain CS tokenization practices with regard to fused forms, both in order to benefit from existing NLP tools and to retain parity with the un-treebanked source corpora, which contain a variety of additional non-linguistic annotations. In order to adhere to strict UPOS and UD dependency relations, we have opted to always tag such cases by reference to the argument pronoun, i.e. a form such as 'efe' is tagged as *PRON* and labeled *nsubj*, not *AUX/aux*. The native CS XPOS tag nevertheless uses the portmanteau notation, and the MISC field includes a segmented form, which can be converted into a subtoken representation if desired.

[3] e.g. Hebrew רמזור *ramzor* 'stop-light' (a portmanteau, lit. 'light-cue'), which is left unsegmented as a single token. We thank an anonymous reviewer for providing this example.

[4] The form *e* in Coptic is highly polysemous: it can stand for the preposition meaning 'to', a relativizer, an adverbial subordinating conjunction, a focus marker, the second person singular feminine (in some inflections), and more. One reviewer has asked whether contemporary taggers are actually susceptible to such errors, and the answer in our experience has been positive, probably because 'e' and 'f' are among the most common Coptic tokens. Additionally, due to null forms associated with the 2.SG.F subject (cf. (8) for example) and UD's policy against null subject nodes, fused forms become unavoidable.

5 POS tags

Coptic Scriptorium offers two tagsets with different levels of granularity: CS Fine and CS Coarse, distinguishing 44 and 23 tags respectively. Due to the possibility of a number of portmanteau tags in fusional cases, the CS Fine tagset effectively included 15 additional distinct labels arising from the cross-product of fusable parts-of-speech.

Table 3 gives the mapping between CS tags and UPOS, but excluding portmanteau tags. In all cases of portmanteau tags, we adopt the strategy outlined in the previous section, of giving content words priority over function words, and more specifically, of preferring arguments over fused auxiliaries.

Coptic auxiliaries fall into two main syntactic classes: main clause auxiliaries (e.g. past tense, CS *APST*) and subordinating auxiliaries (e.g. precursive, *APREC*, which roughly means 'after [VERB]ing, ...'. The tag A* in Table 3 stands for any main clause auxiliary (12 CS Fine tags), while subordinating auxiliary tags are listed separately, all corresponding to *SCONJ* in UPOS. The entry P* stands for four pronoun tags mapped to *PRON*, and V* stands for all CS verbal tags.

CS	UPOS	CS	UPOS
A*	AUX	FUT	AUX
ACAUS	VERB	IMOD	ADV
ACOND	SCONJ	N	NOUN
ADV	ADV		ADJ
ALIM	SCONJ	NEG	ADV
APREC	SCONJ	NPROP	PROPN
ART	DET	NUM	NUM
CCIRC	SCONJ	PDEM	DET
CCOND	SCONJ	P*	PRON
CFOC	PART	PPOS	DET
CONJ	CCONJ	PREP	ADP
COP	PRON	PTC	PART
CPRET	AUX	PUNCT	PUNCT
CREL	SCONJ	UNKNOWN	X
EXIST	VERB	V*	VERB
FM	X		

Table 3: Mapping of CS Fine tags to UPOS.

A point worth noting is that although the CS tags are generally more fine grained than UPOS, no CS tag maps unambiguously to UPOS *ADJ*. This is because true adjectives are extremely rare in Coptic, limited to about a dozen items, which can appear immediately following a noun they describe. For almost all attributive modification, Coptic uses an 'of'-PP, i.e. a 'wise man' is simply a 'man of wisdom'. Due to the fact that true adjectives are so rare in Coptic (all are archaisms left over from Late Egyptian), and the fact that some can also be used in the 'of' construction as though they were nouns, the CS tagset does not reserve a POS tag for them. However for the handful of items that do occur as adjectival modifiers (postnominal, not mediated by 'of'), we use the *amod* relation and UPOS *ADJ* based on the relation.

Additionally, some CS tags provide morphological information that would otherwise be lost in UPOS, but can be represented in UD features (CoNLL-U column 6), which are outlined in the next section.

6 Morphological features

Morphological features are automatically added to the corpus using DepEdit,[5] a freely available Python library for manipulating dependency data in the CoNLL-U format (see Peng and Zeldes 2008). Some of the morphological feature categories are trivial to assign based on word forms, such as gendered and numbered article forms, or pronoun types.

However there are also some features that can be derived from native POS tags, such as mood and polarity: the imperative CS tag *VIMP* can be used to feed the UD `Mood=Imp` feature, and some auxiliaries are inherently negative, feeding the `Polarity=Neg` feature. For example, Coptic distinguishes some tenses with paired negative and positive auxiliaries (e.g. CS tags *APST* and *ANEGPST* for positive and negative past tense). Some tensed auxiliaries are exclusively negative, such as the perfective negative conjugation (CS *ANY*, cf. Loprieno 1995, 221), which roughly translates into a clause modified by 'not yet' which has no morphologically positive counterpart. All forms of such auxiliaries are automatically flagged as `Polarity=Neg` based on CS tags.

Finally, Coptic possessive determiners indicate gender and number for both the possessor and possessed, as in languages such as French or German, and therefore we use the 'layered feature' facility in the CoNLL-U format, distinguishing `Gender` and `Number` from `Gender[psor]` and `Number[psor]` for possessor features, as in (9), which shows a masculine singular noun possessed by an article agreeing with these features, but also marking a third person singular feminine possessor.

[5] https://corpling.uis.georgetown.edu/depedit/

(9) ⲡⲉⲥ-ϩⲓ
pes-ēi
her-house (house = Masc. Sg.)
Gender=Masc|Gender[psor]=Fem|
Number=Sing|Number[psor]=Sing|
Person=3|Poss=Yes|PronType=Prs
'her house'

7 Dependencies

7.1 Absent relations

UD Coptic uses all UD relations, with the exception of *expl* and *clf*, since the language does not have expletive pronouns or classifiers. Among the recommended and frequently used subtypes, we do not use the *:pass* subtypes (i.e. *nsubj:pass* and *aux:pass*) due to the ambiguous nature of Coptic passives. While there is a morphological form, the 'stative' (CS tag *VSTAT*) which can express a stative passive for transitive verbs, as in (10), the same form simply means persisting in a state for intransitive verbs, as in (11).

(10) ⲡ-ϩⲓ ⲕϩⲧ
p-ēi kēt
the-house build.VSTAT
'the house is built'

(11) ⲡ-ⲙⲙⲟⲟⲩ ϫⲟⲗⲃ
p-moou holk^{j}
the-water sweet.VSTAT
'the water is sweet[6]'

In both cases, the sense is not actional. For the actional passive more directly translating the English passive, Coptic uses an ambiguous 3rd person plural, as in (12). When an oblique agent is supplied which conflicts in agreement with the non-referential 3rd person plural, it is possible to distinguish active plural from the passive, as shown in (13).

(12) ⲁ-ⲩ-ϫⲟⲧⲃ-ⲫ
a-u-hotb-f
PST-3.PL-kill-3.SG.M
'they killed him/he was killed'

(13) ⲁ-ⲩ-ϫⲟⲧⲃ-ⲫ ϫⲓⲧⲛ-ⲧⲉ-ⲥϩⲓⲙⲙⲉ
a-u-hotb-f hitn-te-shime
PST-3.PL-kill-3.SG.M by-the-woman
'he was killed by the woman'
(lit. 'they killed him by the woman')

However since cases like (13) are rare, we have opted not to distinguish passives, annotating 3rd person plural verbs uniformly with regular dependent *nsubj* and *aux* children (i.e. active syntax).

7.2 Other problematic constructions

During the annotation process, we encountered several problems and special constructions highlighting the complications of adapting the UD annotation scheme to Coptic. One difficulty was handling lacunae in the data: since we wanted to include some major literary texts in their entirety which are only attested in damaged manuscripts, we were not able to select only texts with complete sentences, and we also expect parsers trained on our data to be applied to damaged text. In cases where the damaged words can be reconstructed with high confidence (usually meaning that at least their POS tag can be assigned), words are attached as usual. For more incomprehensible or very fragmentary phrases, especially those tagged as CS *UNKNOWN* (UPOS: *X*), we attach all tokens to the root as *dep*. For linguistically interpretable scribal errors, by contrast, we use the *reparandum* label, using the general UD guidelines for disfluency annotation.

As an example of a more linguistic issue with Coptic annotation, we consider the case of appositions that are non-adjacent, as the current UD guidelines define appositional modifiers as "immediately following the first noun that serves to define, modify, name, or describe that noun".[7] This definition assumes that appositions are adjacent, with nothing intervening between two nominals. However, this is problematic for some Coptic constructions where enclitic particles, mostly borrowed from Greek such as ⲇⲉ 'but, and', must appear in the second position in the sentence (immediately following the first stressed word), breaking up two appositional nominals, as shown in (14).

(14) ⲡ-ⲣⲣⲟ ⲇⲉ ⲇⲓⲟⲕⲗⲏⲧⲓⲁⲛⲟⲥ ⲁ-ⲫ-ϩⲣⲟⲕⲉⲩⲉ
p-rro **de** Dioklētianos a-f-hrokeue
the-king **but** Diocletian PST-3SGM-amble
'but the Emperor Diocletian went about'

Since the very same two nominals would be considered an apposition if the particle did not occur, and since the particle is always a clause-level dependent that invariably appears in second position, we decided to analyze this construction as *appos*.[8]

[6] Many words translated as adjectives in English are verbs in Coptic: the intransitive infinitive *hlok*^{j} means 'become sweet', and the corresponding stative *holk*^{j} means 'be sweet'. Morphologically both are verbal forms in Coptic.

[7] http://universaldependencies.org/u/dep/appos.html, accessed 2018-07-10.

[8] An anonymous reviewer has suggested creating a sub-

Further difficulties in applying UD guidelines to Coptic arise in handling direct objects. Coptic exhibits a regular alternation or differential object marking depending on tense/aspect distinctions. In the durative tenses (Layton, 2011, 233–250), including indicative present, future and imperfect, objects are usually mediated by the preposition ⲛ- *n-* 'of' (or before pronouns, taking the form ⲙⲙⲟ= *mmo=*), as in (15), whereas in other tenses featuring an auxiliary before the subject, objects are enclitic, appearing directly after the verb without a preposition (this is known as Stern-Jernstedt's Rule, Jernstedt 1927), as shown earlier in (12).

(15) ⲥⲉ-ϩⲱⲧⲃ ⲙⲙⲟ-ϥ
se-hōtb mmo=f
3.PL-kill ACC-3.SG.M
'they are killing him'

The fact that these object positions are semantically identical has led us to analyze both constructions as *obj*. This has the uncomfortable result of the same preposition *n-* sometimes acting as an adnominal modifier marker (*nmod*, in a literal 'of'-PP), and sometimes as an accusative case marker, similarly to the analysis of the differential object marking preposition *et* in the UD Hebrew treebank (only used with definite objects). The advantage is that it is easier to use the corpus to extract all object arguments of a certain verb, or to identify all cases of transitive verbs in general. As a criterion for objecthood, we use the possibility of the Stern-Jernstedt alternation: this criterion is more easily decidable than other tests which have been advocated, such as passivization (Zeman, 2017), since passives are not always reliably identifiable in Coptic (see above), though if passivizability is taken as a criterion (cf. Przepiórkowski and Patejuk 2018) then objects mediated by the prepositional case marker are in fact equally passivizable as well.

A further complication in Coptic direct objects arises from the fact that object clauses can co-occur with correlate pronouns in the main clause, as shown in Figure 2. In adopting the analysis in the figure we followed the practice found in most UD treebanks, tolerating *obj* and coreferential *ccomp* for one verb, despite some misgivings.[9]

Although this analysis conforms to the practice in other treebanks, we are still considering alternatives, such as marking the pronoun in the matrix clause as *expl*, or using *dislocated* for the clause. However these solutions also lead to odd splits, whereby a pronoun could be expletive if the object clause was mentioned, but an object if the clause is fully pronominalized (i.e. when only a pronoun is used). Using *dislocated* is also counter-intuitive, since the clause is not actually out of place: it is in its expected position (not topicalized or unusually postponed). Finally some have proposed marking either the nominal argument or the clause as oblique (Przepiórkowski and Patejuk, 2018), but this seems odd too, since each construction in isolation looks like a core object.

8 Evaluation

In this section we evaluate the application of the UD annotation scheme to Coptic by conducting an inter-annotator agreement experiment using three pairs of annotators. We report label scores (LS) using Cohen's Kappa and % unlabeled attachment score (UAS) with and without punctuation.

The annotators include two pairs of BA students with three semesters of Coptic but no experience with corpus annotation or dependencies, and a third pair consisting of one MA student with two semesters of Coptic but substantial experience annotating English (and some Coptic) dependencies, and one professor proficient in Coptic and dependency annotation (these are also the co-authors of the present paper, and will be referred to as the 'Expert' group below).[10] For the undergraduate students, labeled group A and B, we conducted

type for these cases, e.g. *appos:disjoint*. While this would certainly be possible, such cases are overall rare, making such a label potentially very sparse. Conversely, it is fairly easy to locate such cases based on the dependency graph if needed, and from a linguistic perspective, there is nothing unusual about such appositions – the unusual construction is more properly the particle invariably appearing in second position.

[9] We take this to be a still open point, which we are looking forward to discussing: The current UD guidelines explicitly rule out multiple *obj* relations, but do not specifically refer to *obj* + *ccomp*, which Przepiórkowski and Patejuk (2018) take to be equivalent. Other UD literature has been ambivalent about ruling out multiple *obj* dependents in general (Zeman, 2017, 290). In practice, we have seen UD treebanks in multiple languages allow *obj* + *ccomp*, such as UD German-GSD, UD English-EWT, the UD French treebanks and others. German cases in particular seem to mirror the construction above, e.g. *Ich finde* ***es*** *wirklich toll, dass es Euch jetzt* ***gibt****!*, lit. "I find $\textbf{it}_{obj}$ really cool, that you $\textbf{exist}_{ccomp}$ now!".

[10] An anonymous reviewer has inquired whether the developers of the annotation scheme also taught the annotators Coptic, thereby facilitating higher than expected agreement. This was actually not the case: the BA students studied Coptic at the Hebrew University of Jerusalem, apart from the authors, and the MA student studied Coptic independently using a textbook.

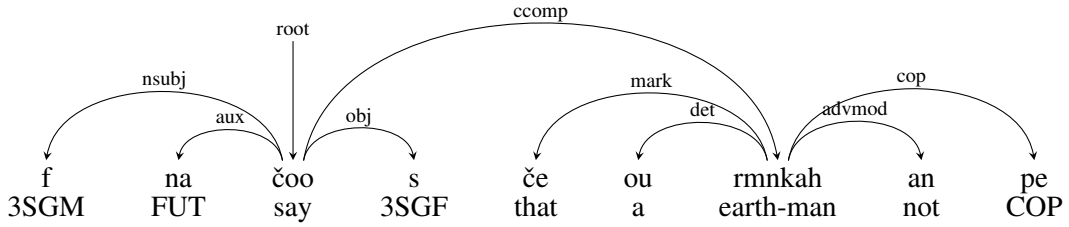

Figure 2: Analysis of a doubled object clause construction: *He would say (it) that he is not an earthly man.*

annotators	tokens	UAS (% agreement)		LS (kappa)	
		punctuation	no punctuation	punctuation	no punctuation
Group A: Pre-Adjud.	276	81.1%	79.0%	0.78	0.75
Group A: Post-Adjud.	319	87.7%	86.5%	0.88	0.86
Group B: Pre-Adjud.	287	84.3%	82.9%	0.79	0.76
Group B: Post-Adjud.	297	86.5%	84.6%	0.81	0.79
Expert	703	96.0%	95.8%	0.93	0.92

Table 4: Agreement Scores. 'no punctuation' denotes scores with punctuation removed from evaluation

two experiments: a pre-adjudication round and a post-adjudication round. In pre-adjudication, annotators only read the online UD Coptic guidelines without any prior annotation experience. Afterwards, student annotators discussed points of disagreement with the professor and adjudicated their sentences, before proceeding to the post-adjudication round, in which we expected annotators to fare better. Annotators had unlimited time to complete the task and the text in all rounds was a portion of *the Martyrdom of St. Victor*, which was presented together with a standard literary translation. As an annotation interface, we used the Arborator (Gerdes, 2013).

Table 4 compares the results of the three pairs of annotators. All results are divided into two sections: with and without punctuation.[11] Results are further separated into pre-adjudication and post-adjudication for the two undergraduate groups.

As shown, the expert annotator scores and the student annotator scores after post-adjudication exhibit relatively high levels of agreement. Within the label score (LS) category, expert annotators scored $k = 0.92$ without punctuation and 0.93 with punctuation, both of which can be considered very good agreement. Post-adjudication, group B produced a label score (LS) of 0.81, while group A scored 0.88. Both of these scores can be interpreted as strong agreement, and noticeably higher than scores between 0.75–0.79, which were achieved solely by reading the guidelines and without previous annotation experience.

Unlabeled attachment scores (UAS) also shows good results. Expert annotators achieve 95.8% without punctuation and 96.0% with, and the student groups have reasonable post-adjudication agreement scores as high as 86.5% and 87.7%, respectively. We observed notable improvements from pre-adjudication to post-adjudication from the student groups. This shows that annotation accuracy on this task can improve after experience and discussing common annotation errors.

The fact that annotators are non-native speakers with limited experience with the language likely affects the inter-annotator agreement results and makes this a challenging task relative to evaluations in other languages, such as English. Berzak et al. (2016) report an agreement experiment on English dependencies with a UAS score of 97.16% and an LS score of 96.3%, conducted on section 23 of the Wall Street Journal corpus (Marcus et al., 1993). Although the labeled score is evaluated as % agreement rather than kappa, these results likely outperform our scores. However in a more challenging task of annotating English tweets, Liu et al. (2018) report a UAS score of 88.8% and LS score of 84.3%, showing that quality can vary substantially across text types.[12]

[11] Scores that include punctuation are based on punctuation attachment to the root, but Udapi (Popel et al., 2017) is used to automatically attach punctuation according to UD guidelines for the final adjudicated gold version.

[12] We do not mean to imply that Coptic data is similar to

Bamman et al. (2009) report results from a dependency annotation experiment on Ancient Greek with an attachment score of 87.4% and a label score of 85.3%. While this experiment wasn't within the UD framework, it offers comparable agreement scores with respect to non-native speaker annotation. The scores presented in their study are close to the attachment scores from our undergraduate student annotator pairs, though admittedly Coptic and Greek are typologically very distant. Scores from other African languages are scarce, but Seyoum et al. (2018) report a kappa score of 0.488 for agreement on UD relations for the morphologically rich language Amharic. This score is interpreted as moderate agreement and is substantially lower than our label scores.

We conducted an error analysis to find common areas of disagreement. While some errors can be attributed to simple, non-systematic mistakes, many high frequency errors are the result of complicated constructions or alternative interpretations of the text, which is at times not trivial to translate. The majority of disagreements for the expert annotators pertained to coordination scope (which is often ambiguous in the translation); confusion over labeling objects (*obj*) and obliques (*obl*), often due to annotating more closely to the source language or the available translation's interpretation; and whether an item has an (*obl*) relation to a verb or an (*nmod*) relation to its dependent noun in constructions that are close to light-verb constructions, but not entirely lexicalized. Coordination proved challenging for longer ambiguous sentences where, as non-native speakers, we relied on our own interpretation of the text for parsing. Confusion over labeling items as *obj* and *obl* can also be attributed to similar syntactic environments where objects and obliques are both mediated by the preposition ⲛ- *n-* 'of'.

9 Conclusion

In this paper we presented the Coptic Universal Dependency Treebank, the first treebank in the UD project from the Egyptian phylum of the Afro-Asiatic language family, and the first Coptic treebank in general. Our evaluation shows that UD guidelines can be applied to Coptic consistently, with rising accuracy based on annotator experience. We are currently expanding the treebank and aim to reach a size allowing for the training of robust parsers and evaluating parsing results on Coptic in future shared tasks.

The discussion has also shown that there are a number of challenges in adapting the UD scheme for Coptic, some of which are shared with other languages: in particular, we advocate a less strict interpretation of adjacency constraints for the *appos* relation, which would also be needed for languages such as Classical Greek, and raise issues with the consistent encoding of pronominal/clausal double object constructions, as well as differential object marking and the handling of ambiguous passivization. We look forward to discussing these issues with the UD community.

tweets, but rather point out the variability in UD agreement scores depending on context.

Acknowledgments

This work is funded by the National Endowment for the Humanities (NEH) and the German Research Foundation (DFG) (grants HG-229371 and HAA-261271). Special thanks are due to Elizabeth Davidson for work on annotating multiple documents in the treebank, as well as to Israel Avrahamy, Asael Benyami, Yinon Kahan and Oran Szachter for annotating sections of the Martyrdom of Victor. We also thank the anonymous reviewers for helpful comments on previous versions of this paper.

References

Mathew Almond, Joost Hagen, Katrin John, Tonio Sebastian Richter, and Vincent Walter. 2013. Kontaktinduzierter Sprachwandel des Ägyptisch-Koptischen: Lehnwort-Lexikographie im Projekt Database and Dictionary of Greek Loanwords in Coptic (DDGLC). In *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pages 283–315, Berlin. BBAW.

David Bamman, Francesco Mambrini, and Gregory Crane. 2009. An ownership model of annotation: The Ancient Greek dependency treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, pages 5–15, Groningen.

Heike Behlmer and Frank Feder. 2017. The complete digital edition and translation of the Coptic Sahidic Old Testament. A new research project at the Göttingen Academy of Sciences and Humanities. *Early Christianity*, 8:97–107.

Yevgeni Berzak, Yan Huang, Andrei Barbu, Anna Korhonen, and Boris Katz. 2016. Anchoring and agreement in syntactic annotations. In *Proceedings of EMNLP 2016*, pages 2215–2224, Austin, TX.

Frank Feder, Maxim Kupreyev, Emma Manning, Caroline T. Schroeder, and Amir Zeldes. 2018. A linked Coptic dictionary online. In *Proceedings of LaTeCH 2018 - The 11th SIGHUM Workshop at COLING2018*, pages 12–21, Santa Fe, NM.

Kim Gerdes. 2013. Collaborative dependency annotation. In *Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013)*, pages 88–97, Prague.

Eitan Grossman. 2014. Transitivity and valency in contact: The case of Coptic. In *47th Annual Meeting of the Societas Linguistica Europaea*, Poznań, Poland.

Peter V. Jernstedt. 1927. Das koptische Praesens und die Anknüpfungsarten des näheren Objekts. *Doklady Akademii Nauk SSSR*, 1927:69–74.

Bentley Layton. 2011. *A Coptic Grammar*, third edition, revised and expanded edition. Porta linguarum orientalium 20. Harrassowitz, Wiesbaden.

Yijia Liu, Yi Zhu, Wanxiang Che, Bing Qin, Nathan Schneider, and Noah A. Smith. 2018. Parsing tweets into Universal Dependencies. In *Proceedings of NAACL-HLT 2018*, pages 965–975.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Special Issue on Using Large Corpora, Computational Linguistics*, 19(2):313–330.

Siyao Peng and Amir Zeldes. 2008. All roads lead to UD: Converting Stanford and Penn parses to English Universal Dependencies with multilayer annotations. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 167–177, Santa Fe, NM.

Stéphane Polis and Serge Rosmorduc. 2013. Building a construction-based treebank of Late Egyptian: The syntactic layer in Ramses. In Stéphane Polis & Jean Winand, editor, *Texts, Languages & Information Technology in Egyptology. Selected papers from the meeting of the Computer Working Group of the International Association of Egyptologists*, pages 45–59. Presses Universitaires de Liège.

Martin Popel, Zdenek Zabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Universal Dependencies Workshop at NoDaLiDa 2017*, pages 96–101.

Adam Przepiórkowski and Agnieszka Patejuk. 2018. Arguments and adjuncts in Universal Dependencies. In *Proceedings of COLING2018*, pages 3837–3852, Santa Fe, NM.

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

Binyam Ephrem Seyoum, Yusuke Miyao, and Baye Yimam Mekonnen. 2018. Universal Dependencies for Amharic. In *Proceedings of LREC 2018*, pages 2216–2222, Miyazaki, Japan.

Ariel Shisha-Halevy. 1986. *Coptic Grammatical Categories. Structural Studies in the Syntax of Shenoutean Sahidic*. Pontificum Institutum Biblicum, Rome.

Dan Zeman. 2017. Core arguments in Universal Dependencies. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 287–296, Pisa, Italy.