A UD Treebank for Bohairic Coptic

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

Nina Speransky
Hebrew University of Jerusalem
gitchinama@gmail.com

Nicholas Wagner
Duke University
nicholas.wagner@duke.edu

Caroline T. Schroeder
University of Oklahoma
ctschroeder@ou.edu

Abstract

Despite recent advances in digital resources for other Coptic dialects, especially Sahidic, Bohairic Coptic, the main Coptic dialect for pre-Mamluk, late Byzantine Egypt, and the contemporary language of the Coptic Church, remains critically under-resourced. This paper presents and evaluates the first syntactically annotated corpus of Bohairic Coptic, sampling data from a range of works, including Biblical text, saints' lives and Christian ascetic writing. We also explore some of the main differences we observe compared to the existing UD treebank of Sahidic Coptic, the classical dialect of the language, and conduct joint and cross-dialect parsing experiments, revealing the unique nature of Bohairic as a related, but distinct variety from the more often studied Sahidic.

1 Introduction

1.1 Coptic

Coptic was the indigenous spoken and written language of Egypt during the Late Roman, Byzantine, and early Islamic periods. As the final stage of the Ancient Egyptian branch of the Afro-Asiatic language family, Coptic concludes a linguistic tradition with the longest continuous written record in human history, which includes three millennia of Hieroglyphic, Hieratic and Demotic Egyptian writing, as well as over a millennium of writing in Coptic itself, a form of the same language written mainly in Greek letters.

Initially a very low resource language, recent efforts to digitize and annotate data for Coptic have resulted in now substantial resources for Sahidic Coptic, the classical dialect of the language, with corpora (Schroeder and Zeldes, 2020), a machine readable dictionary (Feder et al., 2018) and a UD dependency treebank (Zeldes and Abrams, 2018). At the same time, other forms of Coptic, namely dialects beyond Sahidic, continue to have little or no annotated resources. In this paper, we aim to address this gap by introducing a UD treebank for a second, very significant dialect of the language: Bohairic Coptic. In this section we offer a brief summary of Coptic and its dialects, highlighting especially some of the main differences between the classical Sahidic, and Bohairic Coptic, which we explore in more detail later on using our data.

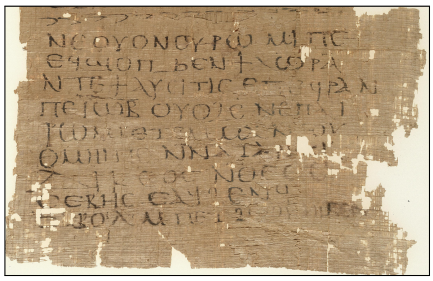

Figure 1: Excerpt from a papyrus containing Bohairic Job 1:1 (P.Mich.inv. 926 recto, TM no. 107875). Image: University of Michigan Library Digital Collections.

Geographic diversity in late ancient Egypt resulted in a range of regional dialects rather than one standardized form of Coptic. Scholars identify six principal dialects, but two of these, Sahidic and Bohairic, were the most influential (Kasser, 1991). Sahidic dominated as the literary language from the third to the ninth century CE before being gradually replaced by Bohairic, a northern variety that continues to serve today as a heritage and liturgical language among Coptic communities in Egypt and the diaspora. Despite the diversity of surviving evidence for Coptic dialects, print and digital resources for the language have remained limited and have focused almost exclusively on Sahidic. The textual evidence for the dialect of Bohairic, though sizeable (and primarily literary), is accessible mostly in facsimiles (see Figure 1 for a papyrus manuscript example) or in older print publications, and many works remain unpublished (Shisha-Halevy, 2007). Using terms from Joshi et al. (2020), Bohairic is at rank 0 of the language technology hierarchy, a 'left-behind' language.

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 59–69

1.2 Bohairic and Sahidic

Expanding digital corpora to include Bohairic texts presents substantial challenges, as existing tools for even basic preprocessing operations, such as word segmentation and lemmatization, let alone syntactic parsing, require tools trained on dialect-specific data (see Section 5 for some evaluation). Although Sahidic and Bohairic are dialectal manifestations of a single language – sharing a broadly consistent grammatical architecture and much of their lexical inventory, their phonological, orthographic and morphosyntactic systems diverge in ways that make NLP tools trained on Sahidic unsuitable for processing Bohairic texts.

At the most elementary level, the two dialects diverge already in their orthographic systems. Unlike earlier Egyptian – written in hieroglyphic and hieratic scripts over the preceding three millennia – Coptic adopted a modified version of the Greek alphabet, with additional characters of Demotic origin (ultimately derived from hieroglyphs) to represent sounds absent from Greek. Both dialects make use of six such additional letters, including the letter hore (ϩ, Unicode U+03E9 lowercase / U+03E8 uppercase) for the voiceless glottal fricative /h/, but Bohairic distinguishes hore /h/ from khei /x/ (ϧ , Unicode U+2CC9 lowercase / U+2CC8 uppercase). This distinction is semantically consequential – while the Sahidic word ϵϩⲣⲁⲓ (ehrai) can confusingly mean both 'up' or 'down', and can only be disambiguated in context, in Bohairic we see distinct forms that had merged in Sahidic: ϵϧⲣⲏⲓ (exrēi) 'down' and ϵϩⲣⲏⲓ (ehrēi) 'up'.

As an agglutinative language, Coptic combines multiple morphosyntactic elements into units known as bound groups. Following modern editorial conventions, these groups are defined by the presence of a single stressed lexical item at their core (Layton, 2011). Material that would normally be tokenized into separate words in annotated corpora often appears conflated into a single space-delimited string in Coptic. Such fusion is common, for example, in noun phrases or prepositional phrases, as in (1), or in auxiliaries and clitics attaching to verbs, as in (2), both examples in the Bohairic dialect.

(1) ϩⲓⲧⲉⲛ-ⲫ-ⲟⲩⲱⲥ ⲙ-ⲫ-ⲛⲟⲩⲧ
hiten-p^{h}-wōš m-p^{h}-nuti
by-the-will of-the-god
"by the will of God"

(2) ⲁ-ⲥ-ⲑⲣⲉ-ϥ-ⲛⲁϩⲙ-ⲟⲩ
a-s-t^{h}re-f-nahm-u
PAST-3.SG.F-CAUS-3.SG.M-hear-3.PL
"She made him save them"

In this paper we use hyphens to split such space-delimited word forms into tokens, which correspond to units that can receive independent parts of speech, such as nouns and verbs, articles and prepositions, etc. Note that such separators do not exist in source texts and are represented in the treebank directly through the tokenization. For words containing derivational affixes as in (3), and compound nouns or verbs (4), we additionally provide a segmentation into component parts in an additional annotation called MSEG in the conllu format MISC field, following existing practices in UD treebanks.[1]

(3) ⲙⲉⲧⲁⲧϫⲱⲛⲧ
MSeg=ⲙⲉⲧ-ⲁⲧ-ϫⲱⲛⲧ
met-at-čōnt
less-ness-anger
"angerlessness"

(4) ϵⲣϩⲁⲗ
MSeg=ϵⲣ-ϩⲁⲗ
er-hal
do-service
"serve"

Such words are easy to recognize since they have distinct, recurring forms (known affixes, special reduced forms of verbs with incorporated objects), lack internal syntactic markers (e.g. incorporated objects like *hal* in (4) appear without articles or other modifiers, the verb 'ire' meaning "do" is reduced to 'er') and are considered single nouns or verbs in terms of parts of speech, as well as single dictionary entries in terms of lemmatization. Due to these multiple levels of complexity, bound groups must first be analyzed and segmented before we can digitize Coptic texts in a way that allows for searchability. Each token can then be lemmatized and tagged to enable structured querying and lexical lookup, or linking to resources such as the Coptic Dictionary Online (CDO, Feder et al. 2018).

2 Previous work

The Bohairic UD treebank joins a growing body of typologically diverse languages analyzed using the UD framework (de Marneffe et al., 2021), in-

[1] See https://universaldependencies.org/misc.html#mseg

cluding recent treebanks of related Afro-Asiatic languages such as Biblical and Modern Hebrew (Swanson and Tyers, 2022; Zeldes et al., 2022), Arabic (Taji et al., 2017), and very recently, Ancient Egyptian (Díaz Hernández and Carlo Passarotti, 2024), resources we consider in the development of comparable annotation guidelines (and to a lesser extent, treebanks converted to UD, e.g. for Hausa, Caron 2015).

The most important previous resource we model our work on is the existing UD treebank for Sahidic Coptic (Zeldes and Abrams, 2018), which contains around 60K words from a range of works in a number of genres. In particular, the Sahidic treebank contains some Biblical material which is in part also available in Bohairic (see Section 3 below). By selecting the same Biblical books and chapters where possible, we are able to conduct some direct comparisons between the dialects which only target parallel passages (see Section 4). At the same time, for texts in the Sahidic treebank that are unavailable in Bohairic, we select substitutes from similar genres, offering a similar range of language usage.

In terms of annotation scheme, we closely follow and adapt to Bohairic the Coptic Scriptorium guidelines for annotating parts of speech, lemmatization, sentence splitting and UD dependency relations. In a recent paper, Crellin (2025) criticizes the choice of UD as a treebanking framework, among other languages for Coptic, as unmotivated, stating "no overt discussion of the rationale for choosing [...] Universal Dependencies" could be found (Crellin, 2025, 100). We would therefore like to explicitly motivate the choice of UD for the Bohairic Treebank beyond the obvious benefit of comparability with other resources (see Section 4), and outline some of the key decisions in treebanking Coptic.

The most crucial decision we follow is treatment of lexical verbs as heads of their clauses, despite their etymology in many tenses in earlier Egyptian as subordinated, nominalized infinitives. For example, the verb "hear" in (5) is the only possible head for the clause, which otherwise contains only the subject. Meanwhile in (6), the past auxiliary 'a' in 'a-f-sōtem' would have been analyzed as the head in earlier Late Egyptian, as the construction is derived from a periphrasis, Late Egyptian *jr=f sḏm*, lit. "did-he hear", or more freely, "he did hearing".

(5) ϥ-ⲥⲱⲧⲉ
f-sōtem
he-hear
"he hears"

(6) ⲁ-ϥ-ⲥⲱⲧⲉ
a-f-sōtem
PST-he-hear
"he heard"

We motivate the choice of the lexical verb as the head by the basic UD principle of lexico-centrism, which also provides a more parallel analysis for subjects in the durative and past tenses shown above. Since Coptic, synchronically an agglutinative language, has a broad range of tenses and constructions, but only a handful of etymological sources for agglutinative morphemes (almost always forms of the Late Egyptian verbs for "do", "give" or "know"), choosing a non-lexico-centric scheme would result in an analysis in which Coptic has only a handful of distinct verbs. The choice of UD is therefore quite conscious, and especially given the benefits of comparability, without obvious superior alternatives.

We also match our native Coptic part of speech tagging scheme and its mapping to Universal POS tags to those used in the Sahidic treebank, as well as using Multiword Tokens (MWTs) to represent bound groups in the conllu format (i.e. examples (5) and (6) would be one MWT each, containing two and three word forms respectively).

3 Data

The textual data in the Bohairic UD corpus consists of selections from works in multiple genres. We include selections from hagiography (saints' lives) of two prominent figures in early Christian Egypt: Shenoute, the leader of a federation of male and female monasteries in the fourth-fifth centuries; and Isaac, Patriarch (or Coptic Pope) of Alexandria from 686 to 689 CE. The Life of Shenoute is a compilation of panegyric works about Shenoute written in Bohairic Coptic over decades after his death and compiled into the genre of a saint's life (see Lubomierski 2008 and Berno 2019c). The Life of Isaac is a saint's life written in Bohairic Coptic possibly in the seventh century after Isaac's death, focusing mostly on his adult life as a monk, priest, and patriarch during the early Islamic period (Berno, 2019b). The Lausiac History is a narrative text originally written in Greek consisting of anecdotes about fourth-century monks and so has hagiographical elements as well as generic elements from travel narratives (Berno, 2019a). Three biblical texts are also translations from Greek: the Gospel of Mark and 1 Corinthians from the New

source	genre	chapters	tokens	sentences
1 Corinthians	Biblical Epistle	1–7	4,789	164
Gospel of Mark	Biblical Narrative	1–9	11,091	373
Book of Habbakuk	Poetic	1–3	1,988	56
Life of Shenoute	Hagiography	1–26	4,970	148
Life of Isaac	Hagiography	1–19	5,433	143
Lausiac History	Hagiography	1–16	4,453	117
Total:			32,724	1,001

Table 1: Data in the Bohahiric Treebank.

Testament, and the Christian Old Testament book of Habakkuk. Bohairic Habakkuk is likely a translation from the Septuagint (the Greek version of the Hebrew Bible). Mark is a gospel or ancient biography, and 1 Corinthians is a letter by Paul the apostle, meaning our Biblical data also spans multiple genres.

The text of each of these works is derived from previous digital or print editions. The *Life of Shenoute* comes from a digital edition created by Hany Takla of the St. Shenouda the Archimandrite Coptic Society, based primarily on the edition published by Leipoldt (1906). Dr. Lydia Bremer-McCollum of the Coptic Scriptorium project[2] extracted digitized text using optical character recognition (OCR) from the public domain edition of the *Life of Isaac* and the public domain edition of the Bohairic *Lausiac History*, both edited by Amélineau (1887, 1890), followed by manual correction of OCR errors. The Gospel of Mark and 1 Corinthians digital texts come from the Marcion Project;[3] both ultimately derive from the print edition of Horner (1905). The text of Habakkuk is from the working digital edition in-progress created by the Göttingen Coptic Old Testament project.[4] All of the digital versions were processed by the Coptic Scriptorium project's natural language processing tools (Zeldes and Schroeder, 2016), which have normalized the text, including normalizing variant spellings and expanding any abbreviations; for this paper, all NLP processing has been manually checked by one or more of the authors before treebanking.

3.1 Inter-annotator agreement

To evaluate the quality of our annotations, we double-annotated 166 sentences (6,207 tokens) from two different texts, the Life of Shenoute and the Lausiac History. Table 2 shows the results for Cohen's κ and mutual F1 score, for both dependency relations (labels) and dependency heads. In order to avoid inflating κ due to a large range of possible labels for heads (i.e. all numbers attested as dependency heads), we represented heads as offsets from the position of the child (i.e. if token 37 has token 35 as its parent, we tally the value as -2). This increases the probability of chance agreement and prevents an inflated metric due to label proliferation. For dependency relations, we use the full label including subtypes (see Appendix A for details).

	Labels		Heads	
	Kappa	**F1**	**Kappa**	**F1**
Life of Shenoute	92.91	93.49	93.84	94.77
Lausiac History	95.53	95.78	94.61	95.53
Macro average	94.22	94.64	94.23	95.15
Micro average	94.79	95.12	94.39	95.29

Table 2: Cohen's Kappa (κ) and mutual F1 score in two texts (166 sentences) for dependency relation labels and heads, the latter represented as offsets relative to child tokens.

The results show very high agreement, substantially in excess of initial (pre-adjudication) scores in the 80s for the original version of the Sahidic treebank (Zeldes and Abrams, 2018, 199). This is likely due to the fact that annotators in this case were post-graduate researchers with substantial Coptic annotation experience, as opposed to the novice student scores reported in the Sahidic paper.

The data above constitutes the first openly available morphosyntactically annotated corpus of Bohairic, and allows for a number of quantitative comparisons with the Sahidic, to which we now turn below.

[2] https://copticscriptorium.org/
[3] https://marcion.sourceforge.net/
[4] https://coptot.manuscriptroom.com/

4 Comparing UD Bohairic and Sahidic

Like other projects adding UD treebanks in low-resource languages for which closely related languages already have treebanks (Jobanputra et al., 2024), one of our goals is to explore the ways in which Bohairic diverges from other varieties of Coptic – ideally, we would like to have treebanks of all Coptic dialects, but for the present we must focus on the comparison with Sahidic.

Dialects and closely related languages can differ in two different ways: they can have categorically distinct constructions, such as different auxiliaries, distinct argument structures for equivalent verbs etc., or they can use similar constructions but in quantitatively different usage patterns. While categorical differences between Bohairic Coptic and the better studied Sahidic are relatively well understood, quantitative differences are more elusive, but can show up in a corpus analysis.

On the lexical level, we can note that of the approximately 2,800 unique words attested in the Sahidic treebank, and around 2100 unique words in the Bohairic treebank, only about 600 are shared, and even among these, identical forms do not always translate to identical meanings. For example while ⲉⲧ- 'et-' can be a form of the relative marker in both Sahidic and Bohairic, it can also represent the precursive form meaning 'after' in Bohairic. Among the disjoint, dialect-specific vocabulary, many words have corresponding words in the other dialect which differ only due to pronunciation, but some words are totally unique, such as ⲃⲁⲕⲓ 'baki', which means 'town' in Bohairic, but does not exist in Sahidic.

On the syntactic level, we cannot find differences and commonalities as easily, but thanks to the existence of UD trees in both dialects, we can still leverage annotated data in a straightforward way. To find some of the clearest differences that our data reveals, we extract relative proportions of the dependency relations attested in the Bohairic and Sahidic Coptic treebanks, an excerpt of which is presented in Table 3, which is sorted by the ratio of frequencies.

As the table shows, striking differences are present for example in the frequencies of `dislocated` arguments (much more common in Bohairic) and `iobj` (indirect objects, much more common in Sahidic). We also note that some control labels, which we would expect to align across datasets, are quite comparable, such as `cop` for copulas, or `nsubj` for subjects.

	Sahidic		Bohairic		
	count	**per 1K**	**count**	**per 1K**	**ratio**
iobj	84	1.471	36	1.100	0.747
...					
cop	500	8.757	291	8.892	1.015
nsubj	5,549	97.185	3,275	100.073	1.029
...					
dislocated	889	15.569	670	20.473	1.314

Table 3: Frequencies and their ratios for some Bohairic and Sahidic dependency relations.

4.1 Subject dislocation

While all Coptic dialects use a basic SVO word order for tensed clauses with lexical verbs (Loprieno, 2000), both left and right dislocations are well attested, with left-dislocation of any argument (e.g. subject or object) bearing no special marking, as in (7).[5] Here again, the choice of UD (contra Crellin 2025) means we can easily find these, thanks to the UD label `dislocated`.

(7) ⲡⲓ-ϫⲁⲓ ⲅⲁⲣ ϥ-ϫⲱⲧⲉⲃ
pi-sxai gar f-xōteb
the-scripture for it-kill
"For scripture, it kills" (i.e. scripture kills)

By contrast, right dislocation or extraposition is obligatorily marked for subjects using the particle ⲛⲭⲉ 'nče' in Bohairic, paralleled in Sahidic by the form ⲛⲑⲓ 'nkʲi'. This particle has been analyzed as a post-verbal nominative case marker by Grossman (2015), who notes that "postverbal subjects are more frequently new referents" in Sahidic, but rarely so in Bohairic. Examples (8)–(9) demonstrate the use of the marker for right dislocation in Sahidic and Bohairic respectively:

(8) ⲁ-ϥ-ⲃⲱⲕ ⲉϩⲟⲩⲛ ⲉ-ⲡⲉϥ-ⲏⲓ ⲛⲑⲓ-ⲓⲱϩⲁⲛⲛⲏⲥ
a-f-bōk ehoun e-pef-ēi nkʲi-iōhannēs
PST-he-go inside to-his-house PTC-John
"He went into his house, (that is) John"

(9) ⲁ-ϥ-ⲑⲓⲥⲓ ⲛⲭⲉ-ⲫ-ⲣⲏ
a-f-kʲisi nče-ph-rē
PST-it-exalt PTC-the-sun
"It was exalted, (that is) the sun"

One question we can immediately explore using our data and the existing UD Sahidic treebank is whether the constructions are used comparably often in the two dialects. Since segmentation, tagging

[5] In the following, some examples from the treebank are abbreviated for space and clarity.

and parsing guidelines match across the treebanks, we can be confident that all relevant cases can be found using equivalent searches – the results are shown in Table 4.

	matches	total words	freq per 1K
	All data		
Boh. ⲛⲭⲉ	258	32,724	7.88
Sah. ⲛⲃⲓ	206	57,097	3.61
	Parallel only		
Boh. ⲛⲭⲉ	122	15,880	7.68
Sah. ⲛⲃⲓ	66	15,619	4.22

Table 4: Bohairic and Sahidic frequencies for right dislocated subjects.

The table shows that right subject dislocation is much more common in Bohairic, with more hits in total in the smaller Bohairic treebank, and more than double the relative frequency. The bottom of the table shows results for the same queries, restricted to only the parallel data available in both corpora – these numbers are slightly less reliable since they are based on less material, but should be expected to be much closer since they derive from the same chapters of the Bible. Here too, the gap remains very substantial, despite the fact that if anything Biblical style should represent a conservative and less free or colloquial style.[6]

Looking at the Sahidic data in more detail, it becomes clear that the construction responsible for the discrepancy is nominal subjects in the canonical position between auxiliaries and verbs, as in (10), which is paralleled in Bohairic by (11) – both examples render 1 Corinthians 2:10.

(10)
ⲁ-ⲡ-ⲛⲟⲩⲧⲉ ⲃⲟⲗⲡ-ⲟⲩ ⲛⲁ-ⲛ ⲉⲃⲟⲗ
a-p-noute k^{j}olp-u na-n ebol
PST-the-god reveal-3Pl to-1Pl out
"God revealed them to us".

(11)
ⲁ-ϥ-ⲃⲟⲣⲡ-ⲟⲩ ⲛⲁ-ⲛ ⲉⲃⲟⲗ ⲛⲭⲉ-ⲫ-ⲛⲟⲩⲧ
a-f-k^{j}orp-u na-n ebol nče-p^{h}-nuti
PST-3Sg-reveal-3Pl to-1Pl out PTC-the-god
"He revealed them to us, that is God".

This construction is much rarer in Bohairic, and indicates that the Bohairic data represents a further step in the grammaticalization of the pronominal subject + auxiliary group, which forces subjects to be realized before or after the verbal complex. This tendency is well known from other languages in the Afro-Asiatic family, such as Hausa, which has similar subject + auxiliary structures, but can only realize a nominal subject outside of the verbal complex, with a 'duplicate' pronoun mirroring the subject – a pronominal TAM marker within the verbal complex is mandatory (see Crysmann 2012, 331, and Hartmann 2006 on fronting in Hausa).

[6] An anonymous reviewer has also asked about comparable breakdowns by genre, but we feel that for some of the texts, comparisons would be hard to establish, and these would also split the already small treebank to the point where counts would be substantially less reliable.

4.2 Indirect objects

Following the Sahidic treebank, we use the `iobj` label primarily to indicate the possessor in the predicative possession constructions with the predicates ⲟⲩⲟⲛ(ⲧⲉ) "there exists" and ⲙⲙⲛⲟⲛ(ⲧⲉ) "does not exist", which are used with a subject expressive the possessum. Thus the Sahidic construction in (12) with the corresponding Sahidic form ⲟⲩⲛⲧⲉ represents an oblique predication "exists to your father something except sins" (or etymologically more precisely, "to the hands of your father").

(12)
ⲟⲩⲛⲧⲉ-ⲡⲉⲕ-ⲉⲓⲱⲧ ⲗⲁⲁⲩ ⲛⲁ-ϩⲉⲛ-ⲛⲟⲃⲉ
wnte-pek-iōt la'u nsa-hen-nobe
EXIST-your-father thing beyond-some-sin
"Does your father have anything except sins?"

Although Sahidic also prefers pronouns to nouns in the possessor position, this is more extremely the case in Bohairic. What is more, although the same constructions as in Sahidic are possible in Bohairic, the Bohairic data shows a tendency to postpone possessors to a later prepositional phrase, leaving the indirect object slot immediately after the existence predication empty. For example, we can contrast the constructions from Mark 4:9 in Bohairic (13) and Sahidic (14):

(13)
ⲫⲏ ⲉⲧⲉ-ⲟⲩⲟⲛ ⲟⲩ-ⲙⲁⲥϫ ⲙⲙⲟ-ϥ
phē ete-won u-mašč mmo-f
DEM REL-EXIST ear of-him
"he who has ears (to listen, let him listen)".

(14)
ⲡ-ⲉⲧⲉ-ⲩⲛⲧ-ϥ ⲙⲁⲁⲭⲉ ⲙⲙⲁⲩ
p-ete-wnt-f ma'če mmau
the-REL-EXIST-him ear there
"he who has ears (to listen, let him listen)".

The postponed cases of possessors mediated by prepositions thus appear to be the main driver of the lower frequency of `iobj` dependencies in the Bohairic data. As far as we are aware, this finding has not been published on to date.

4.3 Focus and preterit marking

Coptic belongs to the group of languages that employ morphological devices known as 'converters' (Layton, 2011, 319-366), which are applied to entire clauses, converting them for example into information structurally marked focalized clauses (focus conversion), signal anterior past tense (the preterit conversion, creating imperfect readings from present clauses or pluperfects from perfect clauses), among others.

For focus marking, two competing strategies are found: Cleft Sentences as in (15) and the morphological focus converter (sometimes referred to as the Second Tense, where the focalized present tense is called the 'Second Present' etc.) as in (16).

(15) ⲟⲩ ⲡ-ⲉⲧ-ⲟⲩ-ⲓⲣⲓ ⲙⲙⲟ-ⲩ ϫⲉⲛ-ⲛⲓ-ⲥⲁⲃⲃⲁⲧⲟⲛ
ou p-et-ou-iri mmo-f xen-ni-sabbaton
what COP-REL-3PL-do ACC.it on-the.PL-sabbath
"What is it that they do on Sabbaths?"

(16) ⲉⲧ-ⲁ-ⲓ-ⲧ̀-ⲱⲙⲥ ⲛⲱ-ⲧⲉⲛ ϫⲉⲛ-ⲟⲩ-ⲙⲟⲩ
et-a-i-ti-ōms nō-ten xen-u-mōw
FOC-PST-1SG-give-baptism to-you in-a-water
"I christened you with WATER"

The presence of the morpheme glossed as FOC in (16) indicates that a constituent is focalized in the sentence, in this case 'water' ('I christened you with WATER' rather than something else).

It has been observed that these strategies are represented unequally in Sahidic and Bohairic, with the preference for the focalizer in the former dialect, and for the cleft sentence in the latter (Müller, 2021). However, up until now such observations have not been backed up with precise and reproducible quantitative data. The UD treebanks make it possible to find the Sahidic-to-Bohairic ratio of focus markers in various parts of the New Testament texts. Since the use of these markers is heavily context and content dependent, we restrict our search to just the Biblical sources available in both treebanks: For the Gospel of Mark, this ratio roughly equals 2.44, and for 1 Corinthians, it is 3.58, both clearly favoring the prevalence of the focalizer in Sahidic.

A less stark, but more surprising result can be found for the preterit marker in the two dialects: as the only formal category denoting anterior past, we could logically expect its equal representation in identical texts in Sahidic and Bohairic. Yet, the UD treebank statistics show that the preterit marker occurs in Bohairic almost twice as often as in Sahidic (the ratios for Mark and 1 Corinthians are 2.21 and 1.76 favoring Bohairic, respectively). These numbers show that there is a substantial difference in how the tense systems of the two dialects are constructed, though we leave a more detailed study of what stands in place of the preterit in Sahidic to future work.

5 Parsing

5.1 Cross-corpus experiments

To evaluate the possibility of using the treebank to train an effective parser for Bohairic Coptic, we train and test several models in different scenarios:

1. Training on just the Bohairic train-set
2. Training on just the Sahidic train-set
3. Joint training on the Bohairic and Sahidic UD treebank train-sets
4. Balanced training on the smaller Bohairic train-set, and an equal amount of Sahidic data

The balanced setting is meant to evaluate whether joint training is more feasible if we ensure Bohairic examples are not overwhelmed by the larger amount of data available for Sahidic. In this case we take care not to include the same document (e.g. the same Bible chapter) from both dialects, and otherwise randomly select Sahidic documents from the appropriate partition, until the Bohairic data size has been reached.

To run the experiments, we use DiaParser (Attardi et al., 2021), a neural biaffine dependency parser, using the default hyperparameters (see the Appendix for exact values). As input embeddings we use the MicroBERT architecture and Sahidic Coptic embeddings made available by Gessler and Zeldes (2022), and train corresponding MicroBERT embeddings for Bohairic using all available Bohairic Coptic data from Coptic Scriptorium's online repository;[7] the new Bohairic transformer embeddings will be released publicly via Hugging Face.

Table 5 shows the results for labeled and unlabeled attachment scores on the respective test sets in each setting, along with train and test partition sizes in thousands of tokens. The results initially reveal that, unsurprisingly, training and testing across dialects (red numbers) produces very poor results, with LAS and UAS scores around 73 and 62 respectively – the scores are rather comparable in both

[7] https://github.com/CopticScriptorium/corpora

train (tokens)	test Bohairic (11K tokens) LAS	UAS	Sahidic (10.3K tokens) LAS	UAS
Bohairic (16.5K)	86.349	89.486	62.205	74.633
Sahidic (35.8K)	62.683	73.178	**89.760**	**92.489**
Joint (52.3K)	**89.929**	**92.677**	88.449	91.602
Balanced (36K)	88.989	91.927	86.858	90.628

Table 5: Labeled (LAS) and Unlabeled (UAS) Attachment Scores in each setting when testing on each dialect. Within and across dialect scores are in green and red respectively. The best settings for each dialect are bolded.

directions, despite the availability of almost double the data when training on Sahidic. This indicates that the parser is unable to generalize when surface forms vary, since as we noted above, even auxiliaries and prepositions look quite different across dialects.

Single dialect models (green numbers) reveal a gap between the smaller Bohairic data and its larger Sahidic counterpart: while Sahidic obtains a LAS of 89.76, Bohairic lags behind with 86.349 (2.5 point difference), with an even starker difference in UAS (92.489 vs. 89.486, about 3 points). Given that the dialects and texts available in them are rather similar, this suggests that more Bohairic data is likely to have an impact in a single-dialect setting.

Moving to the joint models, both the balanced and full-joint scenario improve the score on Bohairic, suggesting that although word forms are different, syntactic structures are similar enough to generalize across the datasets. In fact, the JOINT setting outperforms BALANCED, suggesting that simply having more data is better, as long as there is a core of Bohairic examples to inform the parser about pivotal Bohairic word forms. In absolute terms, the joint Bohairic scores even slightly surpass the Sahidic single-dialect model scores, though these numbers are not strictly comparable, since the test sets contain different documents. We suspect this means the Bohairic test set may be slightly easier overall than the Sahidic one, but we also take it to be an indication that our annotations match the Sahidic guidelines closely.

By contrast to the Bohairic benefit from joint training, both joint scenarios perform worse on Sahidic than the Sahidic-only model. This suggests that given the amount of data available in Sahidic, the infusion of the smaller Bohairic data is not helpful. In the Sahidic case BALANCED is unsurprisingly the worse setting, since there is less total Sahidic data involved.

5.2 Error analysis

To better understand what is challenging about Bohairic parsing, we perform quantitative and qualitative error analysis. Figure 2 shows the confusion matrix for dependency labels in the Bohairic test set for the Bohairic-only model (merging subtypes of the same major relations and omitting labels with fewer than 10 occurrences in the test set).

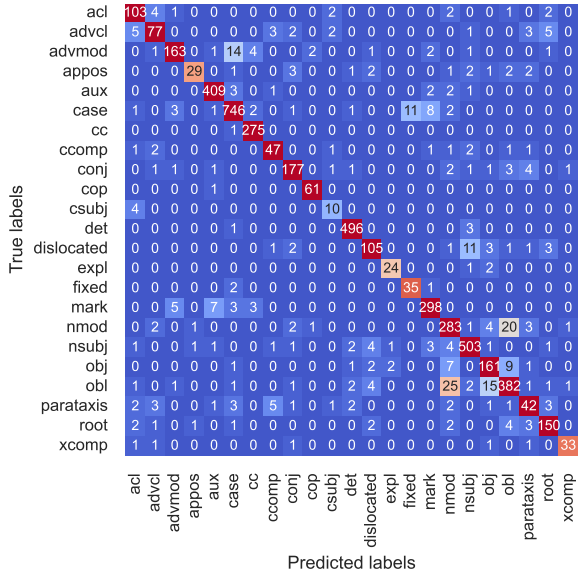

Figure 2: Confusion matrix for collapsed major dependency relations for the Bohairic-only model on the Bohairic test set (labels with <10 occurrences are omitted).

As with most parsers, the most common confusion is between the `obl` and `nmod` labels, indicating problems with classic PP-attachment ambiguities. Overprediction of low adnominal attachment (`nmod` for true `obl`) is slightly more common than the opposite. Additional relatively common confusions occur for `dislocated` subjects versus regular subjects (`nsubj`), which is not surprising, and for `advmod` and `case` being confused with `case` and `fixed` respectively. The latter two are due to ambiguous phrasal verbs, illustrated in (17)–(18):

(17) ⲁ-ϥ-ⲱⲩ ⲉⲃⲟⲗ ϫⲉⲛ-ⲟⲩ-ⲥⲙⲏ ⲙ-ⲡⲣⲟⲫⲏⲧⲓⲕⲟⲛ
a-f-ō-š ebol xen-u-smē m-prophētikon
PST-he-cry out in-a-voice of-prophetic
'He called out in a prophetic voice'

(18) ⲉⲧ-ⲁ-ⲡⲓ-ⲁⲗⲟⲩ ⲓ ⲉⲃⲟⲗ ϫⲉⲛ-ⲡ-ⲏⲓ
et-a-pi-alu i ebol xen-p-ēi
PRC-PST-the-boy came out in-the-house
'after the boy came out of the house'

Like English phrasal verbs, some Coptic verbs

combine with postposed adverbs to form a unique meaning – for example in (17), cry + out means 'cry out, call out' much like in English. In the example, 'out' is coincidentally followed by the preposition 'in', for the phrase "cry out in a prophetic voice". However Coptic also has several frequent fixed combination of adverbs with following prepositions, such as ⲉϧⲟⲗ ϩⲉⲛ, lit. 'out + in' but actually meaning 'out of' (similar to the English fixed expression 'out of' sometimes spelled as 'outta'). In these cases, our guidelines annotate the second token as `fixed`, and the first token takes on the expected `case` label – this ambiguity of adverbs next to prepositions, in which adverbs may belong together with a verb or be part of a multi-word preposition, causes the relatively frequent label confusion in Figure 2. We note that in the interest of comparability with other languages and following UD guidelines (see Ahrenberg 2024 for discussion), the list of fixed expressions of this kind is kept small and is meant to be exhaustive, covering as of writing 25 unique combinations of lemma pairs.

6 Conclusion

In this paper we presented the first morphosyntactically annotated corpus of Bohairic Coptic, containing over 30K word forms from a range of texts. By adopting the same guidelines as the existing UD Sahidic Coptic Treebank, we have been able to perform some first studies of more subtle quantitative differences between the dialects, complementing the better known categorical differences between them. We also ran parsing experiments which indicated that models trained on both dialects jointly were able to boost performance on the lower resource Bohairic dialect, but not on Sahidic.

We are hopeful that this corpus will represent a starting point for further expansion of annotated data for Bohairic Coptic in particular, and Coptic dialects in general. We are confident there is much room for both improving NLP tools for Coptic using such data, and for studying Coptic dialects individually and comparatively.

Limitations

By its nature, this study is based on specific texts which lead to specific results. Although the attempt has been made to select somewhat diverse texts for the corpus, it is always possible that a different selection would have led to different results. In particular, the inclusion of translated texts, such as material from the Bible, is not ideal for some types of research, but as is often the case in resources for historical languages with limited attestation, this is somewhat inevitable. We are hopeful that as new data becomes available, additional studies may revisit some of our findings and either validate or relativize these results.

Acknowledgments

This work was made possible by a grant from the National Endowment for the Humanities Preservation and Access Humanities Collections Reference and Resources Program (PW-290519-23). We thank our postdoctoral fellow, Lydia Bremer-McCollum, for digitizing some of the text used in this paper via OCR, and Wolfgang Jentner and the Data Institute for Societal Challenges at the University of Oklahoma for their server support. We would also like to acknowledge work done by other members of the Coptic Scriptorium project, the Marcion project and the Coptic Old Testament project, as well as by Hany Takla of the St. Shenouda the Archimandrite Coptic Society, which was instrumental in making the data we annotated in this paper available in digital formats. All resources developed in this project will be released under an open license in the hope that we can contribute and match their generosity.

References

Lars Ahrenberg. 2024. Fitting fixed expressions into the UD mould: Swedish as a use case. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 33–42, Torino, Italia. ELRA and ICCL.

Emile Amélineau, editor. 1887. *De Historia Lausiaca: adjecta sunt quaedam hujus Historiae coptica fragmenta inedita*. Ernest Leroux, Paris.

Emile Amélineau, editor. 1890. *Histoire du patriarche copte Isaac: étude critique, texte et traduction*. Ernest Leroux, Paris.

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for EnhancedUniversal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online. Association for Computational Linguistics.

Francesco Berno. 2019a. Historia lausiaca. In *Archaeological Atlas of Coptic Literature*, volume paths.works.245.

Francesco Berno. 2019b. Vita isaac ep. alexandriae. In *Archaeological Atlas of Coptic Literature*, volume paths.works.225.

Francesco Berno. 2019c. Vita Sinuthii. In *Archaeological Atlas of Coptic Literature*, volume paths.works.461.

Bernard Caron. 2015. Hausa grammatical sketch. In Amina Mettouchi, Martine Vanhove, and Dominique Caube, editors, *Corpus-based Studies of Lesser-described Languages. The CorpAfroAs corpus of spoken Afro-Asiatic languages*. John Benjamins, Amsterdam & Philadelphia.

Robert S. D. Crellin. 2025. Considerations for the design of dependency treebanks for linguistic research in Biblical Hebrew. In Aaron D. Hornkohl, Nadia Vidro, Janet C. E. Watson, Eleanor Coghill, Magdalen M. Connolly, and Benjamin M. Outhwaite, editors, *Interconnected Traditions: Semitic Languages, Literatures, Cultures – A Festschrift for Geoffrey Khan*, pages 99–130. Open Book Publishers.

Berthold Crysmann. 2012. HaG: A computational grammar of Hausa. In *Selected proceedings of the 42nd annual conference on African linguistics (ACAL 42)*, pages 321–337.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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

Frank 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.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Eitan Grossman. 2015. No case before the verb, obligatory caseafter the verb in Coptic. In Eitan Grossman, Martin Haspelmath, and Tonio Sebastian Richter, editors, *Egyptian-Coptic Linguistics in Typological Perspective*, Empirical Approaches to Language Typology [EALT] 55, pages 203–225. De Gruyter Mouton, Berlin.

Katharina Hartmann. 2006. Focus constructions in Hausa. In Valéria Molnár and Susanne Winkler, editors, *The Architecture of Focus*, pages 579–607. De Gruyter Mouton, Berlin.

George W. Horner. 1905. *The Coptic version of the New Testament in the northern dialect*, volume 3. Clarendon Press.

Mayank Jobanputra, Maitrey Mehta, and Çağrı Çöltekin. 2024. A Universal Dependencies treebank for Gujarati. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 56–62, Torino, Italia. ELRA and ICCL.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Rodolphe Kasser. 1991. Dialects: Grouping and major groups of. In Aziz S. Atiya, editor, *The Coptic Encyclopedia*, volume 8, pages 97–101. Macmillan, New York.

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

Johannes Leipoldt, editor. 1906. *Sinuthii Archimandritae Vita et Opera Omnia. Vol. 1*. Corpus Scriptorum Christianorum Orientalium 41. Imprimerie nationale, Paris.

Antonio Loprieno. 2000. From VSO to SVO? Word order and rear extraposition in Coptic. In Rosanna Sornicola, Erich Poppe, and Ariel Shisha-Halevy, editors, *Stability, Variation and Change of Word-Order Patterns over Time*, Current Issues in Linguistic Theory 213, pages 23–40. John Benjamins, Amsterdam/Philadelphia.

Nina Lubomierski. 2008. The Coptic life of Shenoute. In Gawdat Gabra and Hany N. Takla, editors, *Christianity and Monasticism in Upper Egypt: Volume 1, Akhmim and Sohag*, pages 91–98. American University in Cairo Press, Cairo and New York.

Matthias Müller. 2021. *Grammatik des Bohairischen*. Lingua Aegyptia 24. Widmaier, Hamburg.

Caroline T. Schroeder and Amir Zeldes. 2020. A collaborative ecosystem for digital Coptic studies. *Journal of Data Mining and Digital Humanities. Special Issue on Collecting, Preserving, and Disseminating Endangered Cultural Heritage for New Understandings through Multilingual Approaches*, pages 1–9.

Ariel Shisha-Halevy. 2007. *Topics in Coptic Syntax: Structural Studies in the Bohairic Dialect*. Orientalia Lovaniensia analecta 160. Peeters, Leuven.

Daniel G. Swanson and Francis M. Tyers. 2022. A Universal Dependencies treebank of Ancient Hebrew. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2353–2361, Marseille, France. European Language Resources Association.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

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

Amir Zeldes, Nick Howell, Noam Ordan, and Yifat Ben Moshe. 2022. A second wave of UD Hebrew treebanking and cross-domain parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4331–4344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

A Dependency relations

We use the entire inventory of Universal Dependency relations with the exception of the `clf` relation, since Coptic has no classifiers, and no cases of an underspecified `dep` relation, for a total of 32 basic relations. In addition, we use the following four subtypes, as used in the Sahidic treebank:

- `acl:relcl` - to distinguish relative clauses from adnominal infinitives and other adnominal clauses
- `nmod:poss` - for adnominal possessive pronouns, including both enclitic pronoun possessors and prenominal possessive pronouns
- `nmod:unmarked` - for adnominal, adverbially used noun phrases, not mediated by a preposition
- `nmod:unmarked` - for adverbially used noun phrases, not mediated by a preposition, when modifying a verbal head

We do not use the subtype `nsubj:pass` since Coptic has no unambiguous actional passive, instead using impersonal third person active syntax ("they built it" = "it was built"). The total distinct labels in the corpus therefore number 36.

B Hyperparameters

The following hyperparameters were used for DiaParser, based on the default parameters combined with the embeddings size of the MicroBERT transformer model:

- BertEmbedding
  - n_layers=4
  - n_out=100
  - max_len=512
- embed_dropout: p=0.33
- LSTM
  - dimensions: 200 x 400 x 3 layers
  - bidirection=True
  - dropout=0.33
- MLP dropouts (arc_d/h, rel_d/h): 0.33
- criterion=CrossEntropyLoss