# Coptic Universal Dependency Guidelines

*Version 1.0.0-2016-05-23*

Amir Zeldes
Georgetown University
amir.zeldes@georgetown.edu

## 1. Preamble

These guidelines describe the application of the Stanford Universal Dependency scheme (de Marneffe et al. 2014) to Sahidic Coptic sentences. Where cases are left unspecified, the principles governing the Universal Dependency guidelines for languages other than Coptic may be consulted (see http://universaldependencies.org/).

In general, the attempt has been made to remain 'lexico-centric' in preferring lexical items as the heads of sub-trees, i.e. Coptic prepositions are analyzed as 'case markers', and the lexical infinitive is seen as the verbal root, with conjugation bases forming auxiliaries.

Some functions from the Universal Dependencies are not used, most notably labels for passives. Although Coptic has some form of passive constructions, they are not easily distinguishable and often ambiguous (actional passive identical with third person active sentence in the absence of a singular agent phrase; stative passive depending on transitivity of the verb for interpretation; see the *nsubj* label below). In these, and other cases, the realistic demands that accurate automatic parsing must satisfy have played a role in the decision in favor of a simpler analysis.

The dependency guidelines assume part of speech tagging based on the Coptic Scriptorium Guidelines (Zeldes & Schroeder 2016). Please consult the tagging guidelines for background on tagging, as well as tokenization decisions (e.g. portmanteau tokens and tags, such as fused ⲉⲣⲉ for 2nd person feminine singular, etc.; see Section 3). In general, these guidelines have been formed to complement the POS tag's expressivity, e.g. favoring distinctions that cannot be easily obtained from the POS tags (for example the label *amod* for the archaic attributive adjectives, which have no distinct POS tag).

## 2. List of dependency labels

acl – adjunct clause                ⲡⲣⲱⲙⲉ ⲉⲧⲥⲱⲧⲙ

advcl – adverbial clause         ⲉϥ<ins>ⲭⲱ</ins>, ϫⲉⲕⲁⲁⲥ ϥ<ins>ⲟⲩⲱϣ</ins>

advmod – adverbial modifier     <ins>ⲟⲛ</ins>, <ins>ⲕⲁⲗⲱⲥ</ins>

amod – adjective modifier        ϣⲏⲣⲉ <ins>ϣⲏⲙ</ins>

appos – appositions            ⲡⲉⲕⲉⲓⲱⲧ <ins>ⲙⲁⲙⲙⲱⲛⲁⲥ</ins>

aux – auxiliary               ⲁ, ⲙⲡⲁⲧ, ϣⲁⲣⲉ

case – case marking/preposition    ⲥⲱⲧⲡ <ins>ⲙ</ins>ⲡⲣⲱⲙⲉ, <ins>ⲙⲙⲟ</ins>ϥ, <ins>ϩⲙ</ins>ⲡⲏⲓ

cc – coordination            <ins>ⲁⲩⲱ</ins>, <ins>ⲙⲛ</ins>, <ins>ϩⲓ</ins>, <ins>ⲏ</ins>

ccomp – complement clause      ⲡⲉⲭⲁ ϥ ϫⲉ ⲁ ⲓ <ins>ⲥⲱⲧⲙ</ins>

compound – nominal or verbal    ϣⲃⲣ <ins>ⲣϩⲱⲃ</ins>

conj – conjunct in coordination   ϣⲏⲣⲉ ⲙⲛ<ins>ϣⲉⲉⲣⲉ</ins>, ⲁϥⲛⲁⲩ ⲁⲩⲱ ⲁϥ<ins>ⲥⲱⲧⲙ</ins>

cop – the copula              <ins>ⲡⲉ</ins>, <ins>ⲧⲉ</ins>, <ins>ⲛⲉ</ins>

csubj – clausal subject (fin./inf.)   ⲡ ⲉⲧ ϣϣⲉ ⲉ <ins>ⲁⲛⲉⲭⲉ</ins> ⲙⲙⲟ ⲕ

det – article or other determiner  <ins>ⲡ</ins>ⲣⲱⲙⲉ, <ins>ⲕⲉ</ins>ⲥⲟⲡ, ⲥⲟⲛ <ins>ⲛⲓⲙ</ins>

discourse – interjections etc.   <ins>ⲉϩⲉ</ins>, <ins>ⲟⲩⲟⲉⲓ</ins>!

dislocated – extraposed argument ϥ ... ⲛϭⲓⲡ<ins>ϩⲗⲗⲟ</ins>, ⲡ<ins>ⲣⲱⲙⲉ</ins> ⲁϥⲥⲱⲧⲙ

dobj – direct object           ⲥⲱⲧⲡ ⲙⲡ<ins>ⲣⲱⲙⲉ</ins>, ϯⲛⲁϩⲟⲧⲃⲉ<ins>ⲕ</ins>

iobj – indirect object (possessor)  ⲟⲩⲛⲧⲁ<ins>ⲩ</ins> ⲁⲡⲟⲧ

mark – clause marker or converter ⲡⲉⲭⲁ ϥ <ins>ϫⲉ</ins> ⲉⲧⲃⲉ ⲟⲩ, <ins>ⲉ</ins> ⲓ ⲥⲱⲧⲙ ⲉⲣⲟ ⲕ

mwe – multiword expression     ⲉⲃⲟⲗ <ins>ϩⲛ</ins>

name – multi-word name       ⲁⲡⲁ <ins>ⲡⲁⲡⲛⲟⲩⲧⲉ</ins>

neg – negation             <ins>ⲧⲙ</ins> ⲥⲱⲧⲙ, ⲟⲩ ⲣⲱⲙⲉ <ins>ⲁⲛ</ins>

nmod – nominal modifier       ⲟⲩ ⲣⲱⲙⲉ ⲛ <ins>ⲁⲡⲓⲥⲧⲟⲥ</ins>

nummod – number modifier     <ins>ϩⲙⲉ</ins> ⲛ ϩⲟⲟⲩ

nsubj – nominal subject       ⲡ<ins>ⲣⲱⲙⲉ</ins> ⲥⲱⲧⲙ, ⲁⲓⲣⲛⲟⲃⲉ

parataxis – loose clausal joint  ⲛⲁ ⲉⲡⲓⲥⲧⲟⲗⲏ ⲛⲉ ⲏ ϯ <ins>ⲥⲟⲟⲩⲛ</ins> ⲁⲛ ϫⲉ ⲕ ⲛⲁ ⲡⲁϩ ⲟⲩ

punct – punctuation           <ins>⁚</ins> <ins>.</ins> <ins>,</ins>

root – root/predicate of utterance <ins>ⲥⲱⲧⲙ</ins>, <ins>ⲛⲁⲛⲟⲩ</ins> ϥ, ⲡ <ins>ⲣⲣⲟ</ins> ⲡⲉ

vocative – appellation        ⲡ <ins>ⲡⲟⲛⲏⲣⲟⲥ</ins>!

xcomp – subjectless clause    ⲙⲡ ϥ ⲟⲩⲱϣ ⲉ <ins>ⲱⲛϩ</ins>


dep – undefined dependency – used when no other definition applies

## 3. Dealing with portmanteau tags

There are two modes of annotation: using portmanteau dependency tags, and pure universal dependencies.

**Portmanteau functions**
For portmanteau tags, tokens which carry a fused portmanteau POS tag receive both functions associated with those components, separated by an underscore. For example, the form ⲉⲕϣⲁⲛ, which is tokenized as one fused token with the POS tag ACOND_PPERS, is both an auxiliary and a subject pronoun. Therefore, its associated verb dominates it as *aux_nsubj*, by analogy to the POS tag.

**Pure universal dependencies**
When using pure dependencies, more 'lexical' functions trump more 'grammatical' ones, so that examples like ACOND_PPERS are still labeled *nsubj*, omitting the *aux* label entirely. This preserves the pure universal dependency tag set.
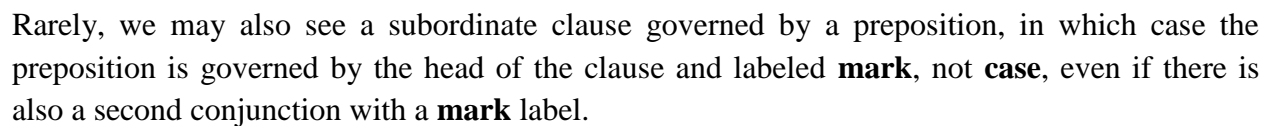
Alternatively, if the intended application of the annotation project supports sub-tokenization, the CoNLL-U format can be used as follows, specifying subtokens/supertokens for fused units:
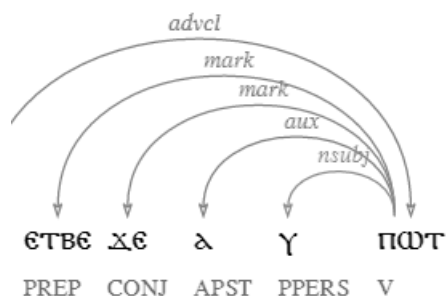
```
1-2    ⲉⲕϣⲁⲛ      _       _       _       _       _       _       _
1      ⲉⲣϣⲁⲛ      _       ACOND   _       _       3       aux     _
2      ⲕ          _       PPERS   _       _       3       nsubj   _
3      ⲥⲱⲧⲙ       _       V       _       _       0       root    _
```

## 4. Function labels in detail

**acl**
Head of an 'adjectival' or relative clause, usually the verb or nominal predicate, when the clause is marked by a relative converter (the POS tag CREL, which is given the function **mark**, regardless of which form is used, such as ∈ⲧ, ∈ⲧ∈, ∈ⲧ∈ⲣ∈ or ∈; for circumstantial conversion see **advcl**). The arrow points from the modified element (usually a noun) to the predicate of the relative clause:



*Your wine which is foul*

**advcl**
An adverbial subordinate clause (e.g. a clause answering the question why? How? Where? When? or a conditional, etc.), usually introduced by a subordinating conjunction or auxiliary (∈ⲡ∈ⲓⲇ�
ⲏ, ∈ⲣϣⲁⲛ), or a circumstantial converter (∈/∈ⲣ∈).



*if he were really a monk, would he have committed three murders?*

Rarely, we may also see a subordinate clause governed by a preposition, in which case the preposition is governed by the head of the clause and labeled **mark**, not **case**, even if there is also a second conjunction with a **mark** label.

| ЄТВЄ | ХЄ | ⲁ | Ⲩ | ⲠⲰⲦ |
|------|-----|------|-------|-----|
| PREP | CONJ | APST | PPERS | V |

*because they have fled* (lit. *for that they fled* – both 'for' and 'that' are labeled **mark**)

This analysis keeps a parallel structure with a similar clause without the preposition (e.g. only with ⲭⲉ to mean 'because').
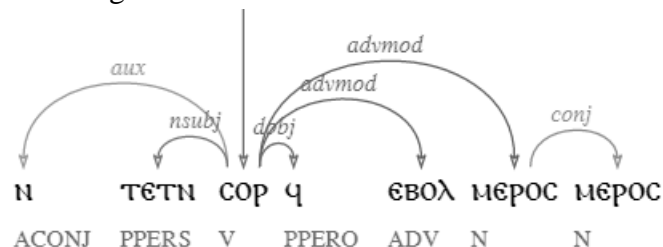
**advmod**

An adverbial modification, usually modifying a verb or a noun. This can be an adverb like ⲉⲙⲁⲧⲉ 'very much', ⲙⲙⲁⲩ 'there', a sentence particle (modifying the main predicate) like ⲅⲁⲣ 'after all', or a directional adverbial such as ⲉⲃⲟⲗ 'out', as well as Greek adverbs in -ⲱⲥ.



| ⲘⲚ | ⲗⲁⲁⲩ | Ⲛ̄Ⲉ̄ⲎⲦ | ⲞⲨ | ⲅⲟⲗⲱⲥ | ⲘⲘⲀⲨ |
|-----|-------|--------|------|--------|--------|
| EXIST | N | PREP | PPERO | ADV | ADV |
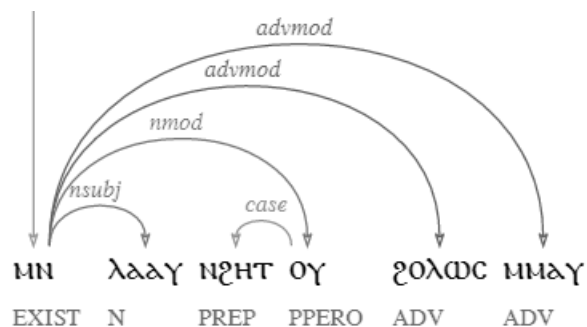
*There isn't a thing inside them <u>at all</u> <u>there</u>*

Occasionally nouns will be used as adverbial modifiers depicting manner or time, as in the following:



| Ⲛ | ⲦⲈⲦⲚ | ⲤⲞⲢ | ϥ | ⲈⲂⲞⲖ | ⲘⲈⲢⲞⲤ | ⲘⲈⲢⲞⲤ |
|-----|-------|------|------|--------|--------|--------|
| ACONJ | PPERS | V | PPERO | ADV | N | N |

*As you divide it <u>out</u>, <u>limb</u> for limb* (lit. *limb-limb*, i.e. *limb-wise*)

Inflected modifiers (Scriptorium tag IMOD, cf. Layton 2011: 118-123) are also seen as adverbial. For example, ⲅⲱⲱ⸗ is used together with an object pronoun to mean 'also X' or 'X for X's part. Because of its basic modifier semantics, meaning 'also', the combination is seen as

adverbial, so that the function of the phrase is again **advmod**. Note that ϩⲱⲱ is not a preposition, and the analysis treats it similarly to a possessed noun, so that the pronoun is seen as a determiner **det**:
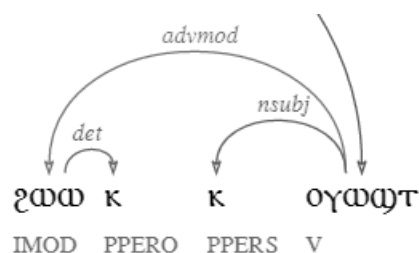


*You also worship / for your part you worship*

The same applies to other IMODs: the inflected modifier ⲧⲏⲣ= 'all of X' is also seen as **advmod**, i.e. as syntactically more similar to 'completely' than a determiner 'all'. Like all inflected modifiers, the pronoun is seen as a determiner in this case, similar to a possessive. In the following example, we can think of the meaning as 'their entirety(-wise)', or 'by way of their entirety'.
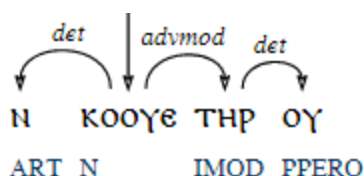


*all the others* (lit. 'the others, [in] their entirety)

**amod**
This function is reserved to the small closed class of Egyptian adjectives which may follow a noun without mediating ⲛ. The label is only given if the construction is actually noun+adjective: if an ⲛ appears, then the label should be **nmod**.
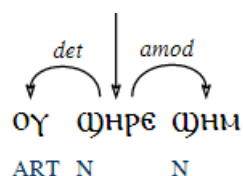


*A little boy*

Note that such adjectives are still tagged with the POS tag N, following the Scriptorium POS tagging guidelines.

**appos**

Marks appositions in all free apposition constructions, as well as the special construction with ⲛϭⲓ 'namely'. Appositions are preferred to be marked from left to right, including in sentences with nominal subject after the verb (apposition from pronoun to noun). In cases where both a nominal and pronominal realization are given to the same argument with the same function, the tightly bound argument of the verb takes precedence in receiving subject or object marking, with subsequent realizations connected as appositions. Typical cases include normal apposition (two consecutive nominal expressions with same reference and grammatical function) and the ⲛϭⲓ construction.
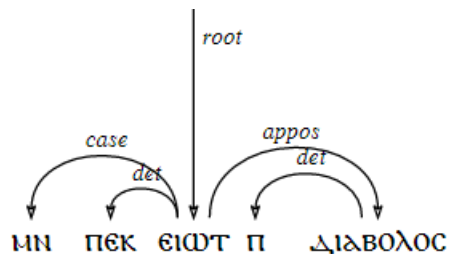


*With your father, the <u>devil</u>*
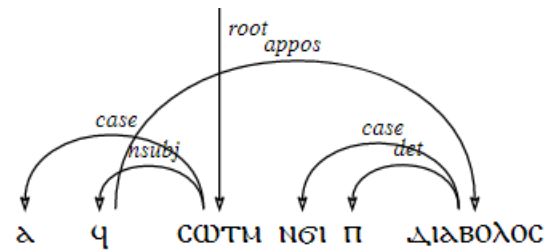


*He heard, namely the <u>devil</u>*

In the first example, *your father* and *the devil*, refer to the same thing, and have the same function (both relate to *with*: *with your father*, i.e. *with the devil*). In the second example, an apposition to the pronoun ϥ 'he' is mediated by ⲛϭⲓ 'namely, that is', which is considered to be a case marker, like a preposition (but with nominative case). As usual, the apposition goes from left to right, to the lexical head 'devil' (he = devil).

Unusually, if a nominal subject or object is mentioned **before** an auxiliary and is then referred to by a pronoun in the verbal complex (e.g. ⲡⲣⲱⲙⲉ ⲁϥⲥⲱⲧⲙ 'the man, he heard'), the pronoun is treated as the subject or object, and the preceding noun is labeled **dislocated**.

**aux**

The relation between a lexical verb and a conjugation base marks the base as **aux** to the verb, as shown below. This applies to all tripartite conjugation bases, but NOT to converters, which receive the **mark** label.

*They chose her* (or: *she was chosen*)
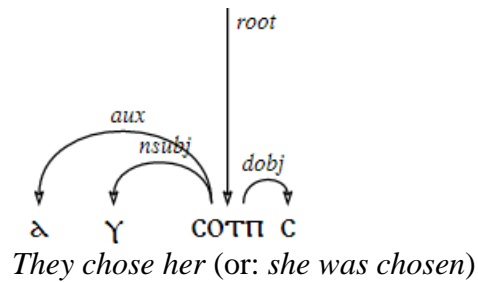
Some other elements that are marked as aux include the future auxiliary ⲛⲁ (tagged FUT) and the potential verb ⳟ 'be able to', which is marked as an auxiliary to the lexical verb that follows it. This should not be confused with the impersonal verb ⳟⳟⲉ 'it is appropriate', which is treated as a main verb governing an infinitive.

**case**
Used for all prepositions, including the marker of accusative case ⲛ and all other prepositions, which are understood as 'oblique' cases. Nouns govern their prepositions as in all other Universal Dependency guidelines, and not the other way around.



*the sun __in__ the sky*
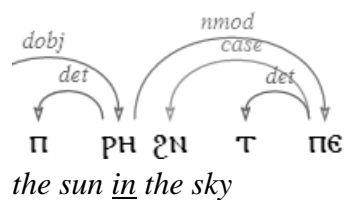
The hermeneutic particle ⲛϭⲓ, roughly 'namely', is also considered a case marker, assigning nominative case (it is only compatible with subject appositions, never objects or obliques, cf. Grossman 2014); see **dislocated** for more guidelines on ⲛϭⲓ.

**cc**
The label for coordinating conjunctions. These are usually ⲁⲩⲱ 'and' between clauses and ⲙⲛ 'and, with' between phrases, but could also be ϩⲓ in the sense 'X upon Y' or ⲏ 'or'. If the sense is not coordinating (e.g. ⲙⲛ to mean 'with'), cc should not be used, but nmod as with a regular preposition.

Coordination is marked left to right, with the first coordinate dominating all subsequent coordinations and receiving the incoming grammatical relations. The reasoning is that to retrieve the function of any coordinate, we can check its parent's function, even without knowing how many coordinates there are (*X and Y and Z and …*). For example:
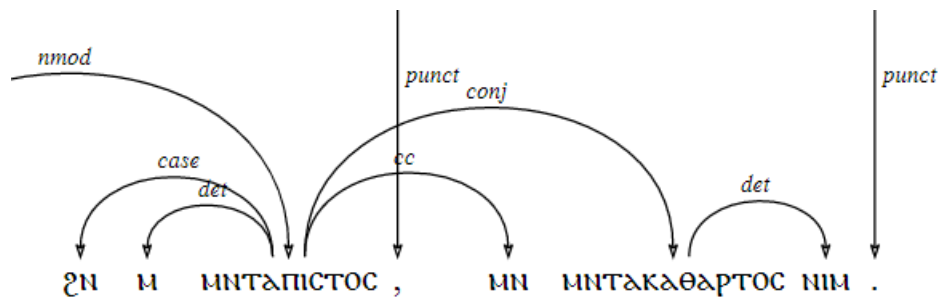
*In faithlessness <u>and</u> every impurity*

Note that the function of ⲙⲛⲧⲁⲡⲓⲥⲧⲟⲥ 'faithlessness' is nmod. To recover the function of ⲙⲛⲧⲁⲕⲁⲑⲁⲣⲧⲟⲥ 'impurity', we can look at 'faithlessness', its immediate parent and establish that 'impurity' is also nmod. Also note that the word ⲙⲛ 'and' is ambiguous with the meaning of comitative 'with' (e.g. go somewhere **with someone**). When used in the latter way, it is not labeled **cc** + **conj** but rather **nmod** + **case**, as with all other prepositions.

Exceptionally, clause initial 'and' or 'but' is connected to the root of the clause, pointing **backwards**, as in this example:
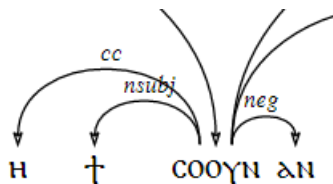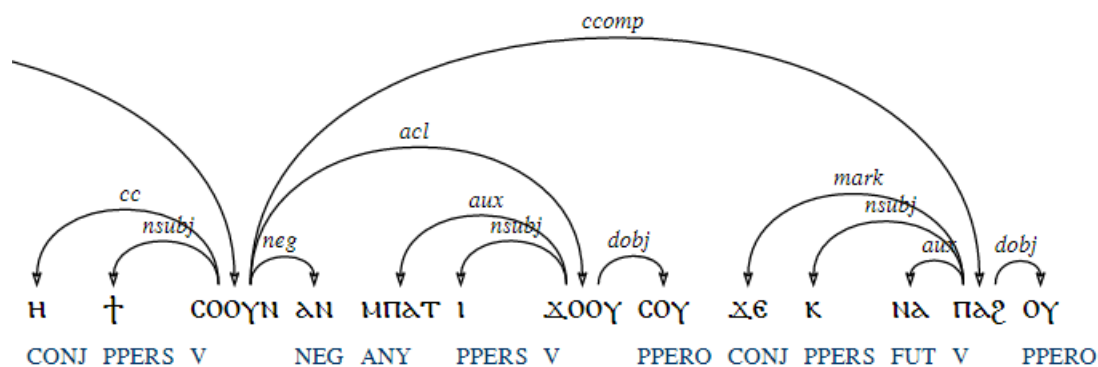


*<u>Or</u> don't I know?*

In this case, the word ⲏ 'or' cannot be attached to a preceding word, so it is pointed to from the following conjunct 'know' with the usual function, **cc**.

**ccomp**
Marks a complement clause, e.g. an object clause to a verb of saying (said that: …., saw that: …). The dependency goes from the main clause predicate to the subordinate clause predicate. Markers like ϫⲉ are governed by the **mark** function (see **mark**).

ccomp · acl · cc · nsubj · neg · aux · nsubj · dobj · mark · nsubj · aux · dobj

Н  †  СООΥΝ  ⲀΝ  ⲘⲡⲀⲦ  Ⲓ  ϪⲞⲞΥ  ⲤⲞΥ  ϪⲈ  Ⲕ  ΝⲀ  ⲡⲀϨ  ОΥ

CONJ  PPERS  V  NEG  ANY  PPERS  V  PPERO  CONJ  PPERS  FUT  V  PPERO

*Don't I know, before I have sent them, that you will tear them up?*

Note how the verb *know* is the source of **ccomp**, and the subordinate clause main verb, *tear*, is the target: [I] know [that you] tear – X knows about tearing – know -> tear.

**compound**
Used to connect compound noun heads to their modifier. For example the compound 'accomplice' is comprised of 'friend' and 'doing' (a 'doing-friend'), which is a type of 'friend' (not a type of 'doing'). Therefore 'friend' is the head, and 'doing' is attached to it via the function **compound**:



vocative · det · compound

ⲡⲈ  ϢⲂⲢ  ⲢϨⲰⲂ

ART  N  N
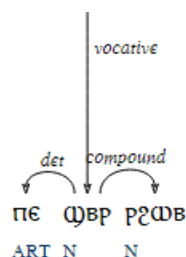
*accomplice!* (lit. 'the friend-doing', or 'doing-friend')

Note: This label is only used for cases in which tokenization has left parts of a compound as separate units. Generally speaking, Scriptorium guidelines specify that compound constituents are only annotated at the morpheme level, and do not constitute independent normalized units which are assigned a part of speech. As a result, this label should almost never be needed in corpora following Scriptorium segmentation practices; but for exceptional cases or corpora not following these practices, the **compound** label is the alternative.

**conj**
This function marks the coordinate (not the coordinating word 'and' etc.) and attaches it to the first member of the chain of coordinates. To retrieve the function of the daughter of *conj*, only its parent needs to be examined, cf. the guidelines for cc. The same example applies:

ⲘⲚⲦⲁⲠⲓⲥⲦⲟⲥ , ⲘⲚ ⲘⲚⲦⲁⲕⲁⲑⲁⲢⲧⲟⲥ ⲚⲓⲘ .
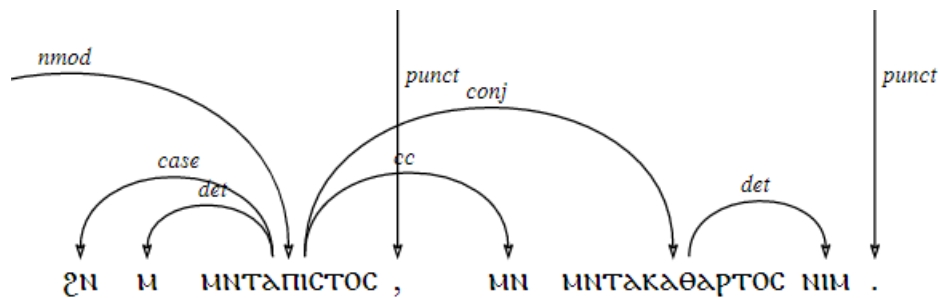
*In faithlessness and every impurity*

The first coordinate points right to the second.

**cop**

This label marks the copula forms ⲡⲉ, ⲧⲉ, ⲛⲉ. Note that the copula is not the root of a copula predication, but rather a dependent of the lexical predicate, usually a noun preceding the copula, but sometimes a verb (especially with preterit conversion, followed by the 'optional' copula). Both cases are illustrated below.
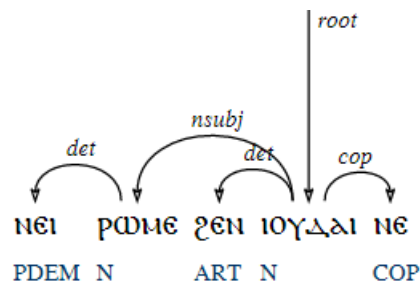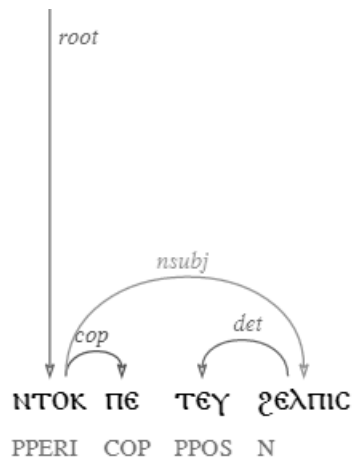


*These men <u>are</u> Jews.*
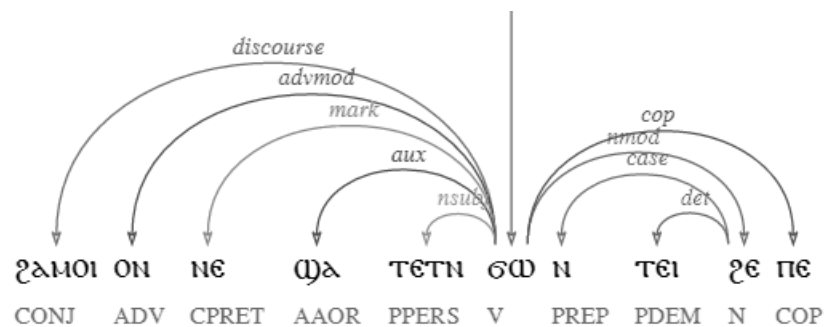


*Oh, would that you would stop in this way!*

Also note that in nominal sentences, the subject is **nsubj** to the predicate, not to the copula. This is also true when the subject follows the copula, as shown below.

*Their hope is you* (lit. '[it]'s you, their hope')

**csubj**
Used to mark the head of a clausal subject. The dependency goes from the predicate that governs the subject, to the local root (i.e. the predicate) of the subject clause. This construction is fairly rare, and the subject clause is often an infinitive clause, as shown below.
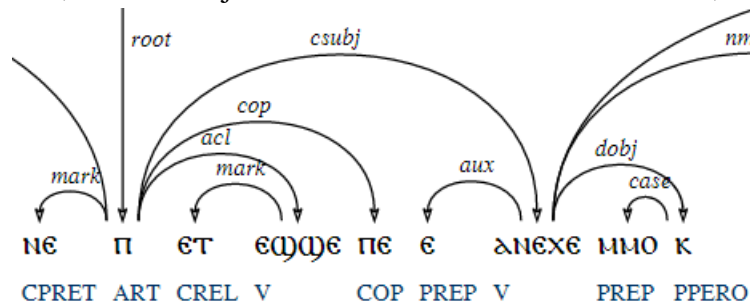


*it would be fitting to tolerate you*

Note that this is just like a nominal subject (**nsubj**): *what would be fitting? To tolerate you. To tolerate you is fitting.*

**det**
Function label for determiners, including definite and indefinite articles governed by their noun, but also the postposed ⲚⲓⲘ 'any' and the determiner ⲕⲉ 'other', which is unique in being allowed to stand for an article. If a noun has both a normal article and ⲕⲉ, both are marked as **det**:



*some other matters*

**discourse**

Function label for discourse particles, such as negative/affirmative answers (yes, …) or interjections (oh! Hah! etc.). The discourse particle is connected to the root of its clause using the **discourse** function.
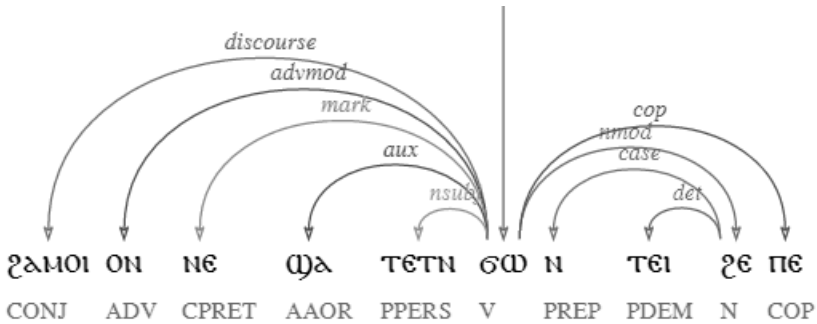


*Oh, would that you would stop in this way!*


**dislocated**

This label is used for arguments or 'hanging topics' that are preposed before the verbal complex, and which are referred to again as pronouns governed by the verbal complex.
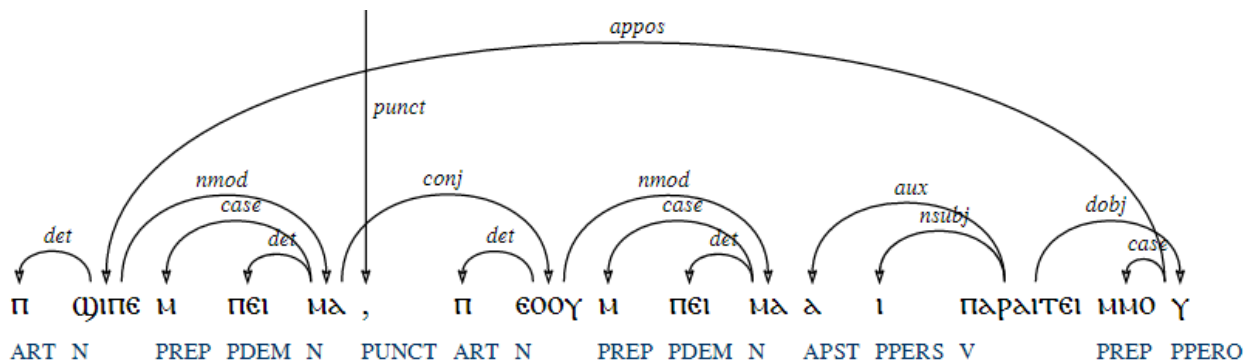


*The shame of this place, the glory of this place, I have forsaken them*

Note how the lexical arguments 'the shame … the glory' act as *dislocated* and are pre-posed, to be referred to again as pronouns by the pronoun object 'them', which acts as *dobj* (*forsaken* what? *them*. What is *them*? *shame + glory*). The dislocation is linked to the word governing the 'duplicate pronoun', usually the verb (here ΠΑΡΑΙΤΕΙ 'forsake', which also governs the object pronoun).

**dobj**

Designates the direct object of the verb. This can be one of the following:

- A nominal object standing directly after a verb


- A pronominal object directly after a verb

13
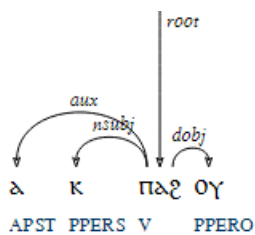
*You have torn them.*

- A nominal or pronominal object mediated by the accusative marker (ⲛ/ⲙ/ⲙⲙⲟ⸗), usually in the durative patterns according to Jernstedt's Law. The marker itself is linked to the noun with the **case** function.
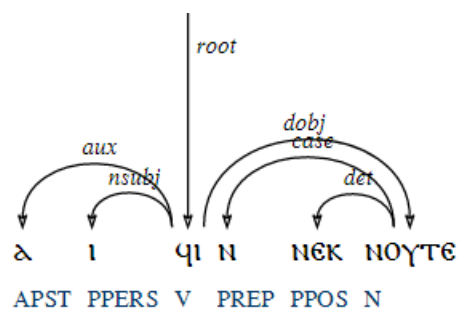


*I have carried your gods*

**iobj**

Only used to mark the possessor in the possessive existential construction:
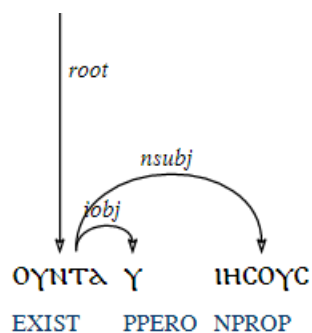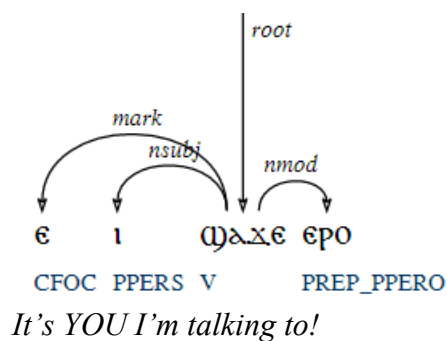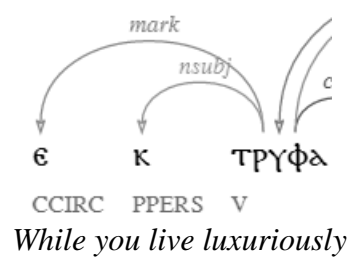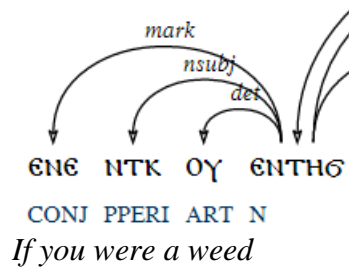


*They have Jesus*

(can be interpreted as 'exists to-them Jesus', etymologically ⲟⲩⲛⲧⲁ-ⲩ < 'exists in their hand')

**mark**

Marker indicating clause status. For subordinate clauses, this is the subordinating conjunction or particle, such as ⲝⲉ introducing the object clause of direct or indirect speech, a relative pronoun such as ⲉⲧ, or an adverbial subordination, such as the circumstantial converter ⲉ/ⲉⲣⲉ. In Coptic, some main clauses also have marker elements determining clause status, which are morphologically and paradigmatically comparable to subordinators: the focalizing and preterit

converters. As a result, all converters (POS tags CCIRC, CREL, CPRET and CFOC) are treated as **mark**, modifying their clause's predicate, but also non-converter conjunctions with similar functions (e.g. ⲉⲛⲉ 'if') as shown in the examples below.



ⲈⲚⲈ  ⲚⲦⲔ  ⲞⲨ  ⲈⲚⲦⲎϬ
CONJ  PPERI  ART  N

*If you were a weed*



Ⲉ  Ⲕ  ⲦⲢⲨⲫⲁ
CCIRC  PPERS  V

*While you live luxuriously*



Ⲉ  Ⲓ  ϢⲀⲘⲈ  ⲈⲢⲞ
CFOC  PPERS  V  PREP_PPERO

*It's YOU I'm talking to!*


**mwe**
Multi-word expressions are sequences of tokens which form a fixed expression, for which internal grammatical relations are not represented. For Coptic, this often corresponds to multi-token complex prepositions, often with a frozen adverbial modifier such as ⲉⲃⲟⲗ 'out'. For example in the following example, the complex sequence ⲉⲃⲟⲗ ϩⲛ 'out of', has individual tokens which literally mean 'out in'.



ⲁ  Ⲩ  ⲚⲦ  ⲞⲨ  ⲈⲂⲞⲖ  ϨⲘ  ⲠⲈⲔ  ϨⲒ
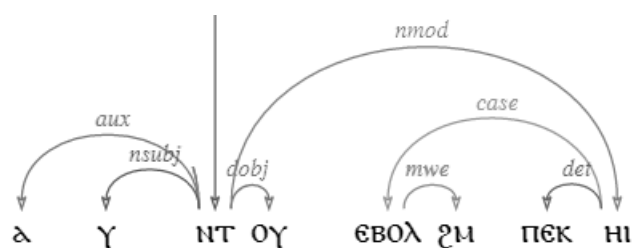
*they took them out of your house*

While an interpretation connecting ⲉⲃⲟⲗ to the verb to mean 'take out' is possible, that would leave the sense of 'in' to mean 'of' unexplained. Rather, the combination ⲉⲃⲟⲗ ⲍⲛ 'out of' is lexicalized as a multiword expression, or **mwe**. By convention, mwe's point in a chain from left to right, whereas the first token in the chain carries the external function of the expression – in this case a preposition (the label **case**).

**name**
This label is used to connect parts of multi-word names, pointing from left to right in a chain. The most typical case is titles such as ⲁⲡⲁ, but the guideline applies to all complex names.



*Apa Papnoute*

**neg**
The label for negations such as ⲁⲛ, ⲛ, ⲧⲙ etc. which receive the POS tag NEG. The attachment is to the negated element, often the predicate or verb. Copula sentence negation is attached to the predicate, not to the copula. In circum-negation (ⲛ...ⲁⲛ), both elements are attached to the same element with the **neg** label.
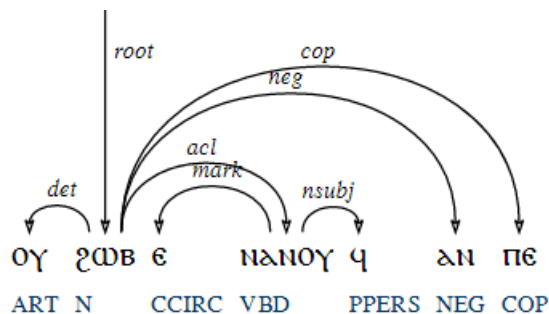


*it is not a good deed.*

**nmod**
A nominal modifier. This is the label given to prepositional objects and other types of nominal dependents which are non-core arguments (i.e. neither subject nor object). Note that in keeping with Universal Dependencies for other languages, the **nmod** noun attaches directly to the lexeme it modifies (usually a noun or verb), while the preposition is seen as a **case** dependent of the modifying noun.

*the sun in the sky*
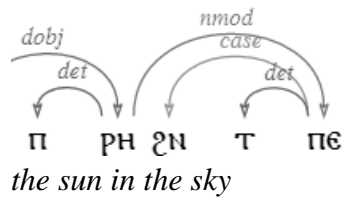
Note also that prepositional arguments of verbs are also marked as **nmod**, including prepositional objects of verbs of perception:
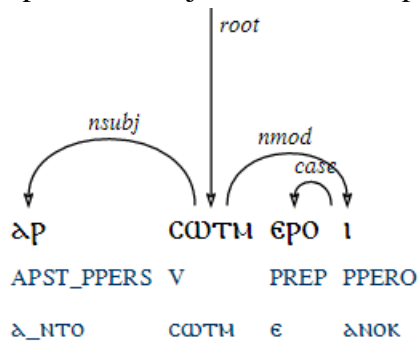


*You have heard me*

### nummod

This label is used for number modifiers, counting the modified nouns. Note that for numbers taking a preposition ⲛ, the preposition is attached via the case label to the counted noun, but the article (and any further prepositions preceding the article) are attached to the nominal head as well, as shown below.
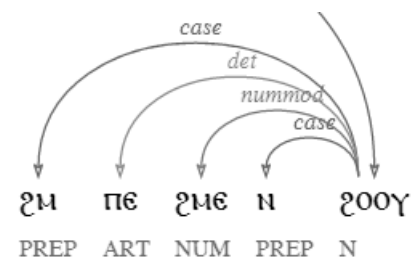


*in the forty days*

For the number ⲥⲛⲁⲩ 'two', which follows the counted noun, the noun is still the head, i.e. the dependency arrow points forward from the noun to the number.
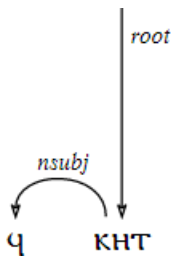
### nsubj

Designates the subject of a verb. The relation is from the lexical verb to the subject pronoun or noun, not from the auxiliary, as shown below. The annotation guidelines do not distinguish passive subjects, since the two forms of passive-like expressions in Coptic may be seen as syntactically indistinct from active:

- Actional passive – formed by a third person plural with non-plural reference, but if there is no singular 'by' phrase, it is formally indistinguishable from the active equivalent:

root
aux
nsubj
dobj

ⲁ   ⲩ   ⲥⲟⲧⲡ   ⲥ
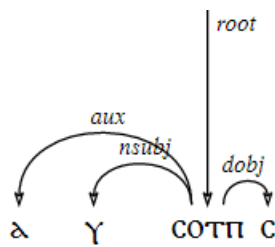
*She was chosen* (or: *they chose her*)

- Stative passive: formed by the morphological stative form (pos=VSTAT) with a transitive verb, however syntactically same as statal reading with intransitive verb:

root
nsubj

ϥ   ⲕⲏⲧ

*It is built* (transitive ⲕⲱⲧ 'build')

root
nsubj

ϥ   ϩⲟⲗϭ
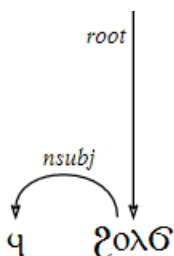
*It is sweet* (intransitive ϩⲗⲟϭ 'become sweet')

Note also that the existential predicates (pos=EXIST) take **nsubj** for the existing entity, even when they are used in the possessive construction:

aux
nsubj

ⲙⲛ   ⲙⲛⲧⲗⲏⲥⲧⲏⲥ   ϣⲟⲟⲡ
EXIST   N                    VSTAT
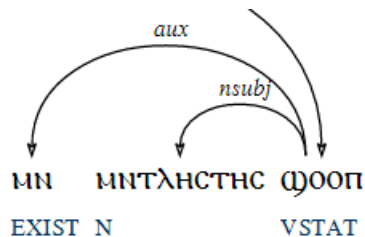
*there is no robbery*

18

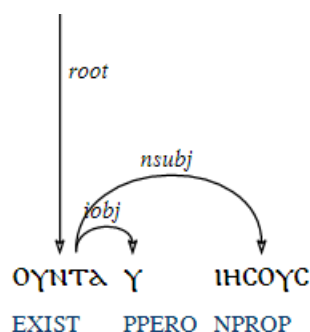If the possessor is indicated in the EXIST construction, the possessed is still annotated as nsubj, and the possessor is annotated as the indirect object, see **iobj** (i.e. the construction is interpreted as 'there exists X to Y').



OYNTA Y IHCOYC
EXIST PPERO NPROP
*They have Jesus* ('exists to them Jesus')

In pure existence predication, the existing entity is the subject, and the EXIST predicate is the local root:
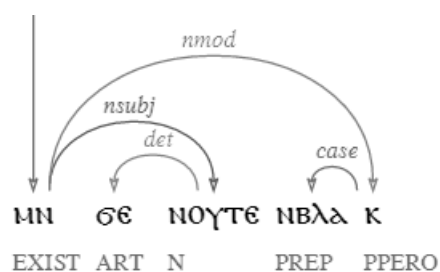


MN GE NOYTE NBλλ K
EXIST ART N PREP PPERO
*There is no God but you* ('not existing' is the main predicate, 'God' is the subject)

**parataxis**
This label is used to link two main clauses that are listed together as one sentence (either by accident, or because they are not quite independent sentences). It is also used for parenthetical clauses in the middle of other clauses. The dependency goes from the root of the first clause, or in parenthetical cases, the non-parenthetical clause, to the other one:
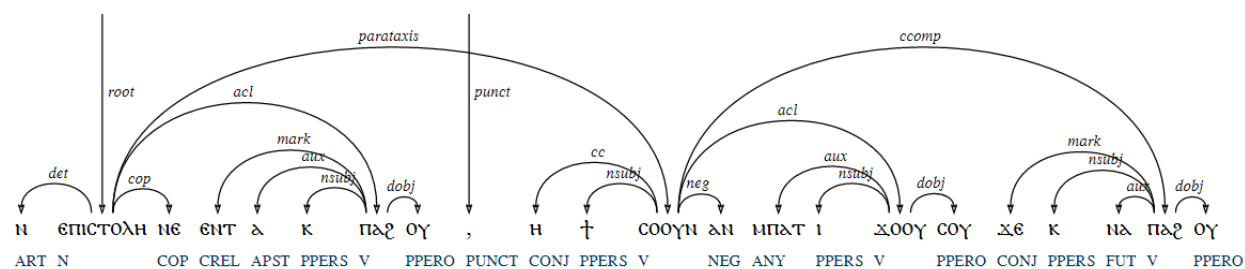


*It's the letters you tore, don't I know before I sent them that you'll tear them?*

If both of these clauses are seen as one sentence, there is no other relation to call the connection between the first clause and the second one. Note that this is distinct from two coordinated clauses, e.g. with ayw 'and', for which **cc** and **conj** should be used.

19

**punct**
The function for punctuation. It is seen as a root function, i.e. punctuation does not depend on any other word in the sentence. If a single tree form is desired, it is also possible to attach punctuation to the local root of the graph.



ⲀⲚⲞⲔ .
PPERI   PUNCT

*I.*

**root**
The root of the utterance, that word, which depends on no other word. Usually this is the predicate, a verb if available, otherwise the nominal predicate of a nominal sentence. In fragments, such as a plain nominal phrase or a single interjection, the local root (i.e. the noun head, or the interjection) is the root.

**vocative**

Used to mark direct forms of address, usually introduced by a definite article: ⲡⲉϣⲃⲣⲣⲅⲱⲃ! 'oh you accomplice!'. If the vocative forms the entire utterance, it is labeled vocative and functions as the root of the utterance. If there is a further proposition in the sentence (usually an impreative), then the vocative is attached to the root of the predication:



Π     ⲚⲞⲨⲦⲈ  ⲀⲢⲒ   Ⲧ    ⲀⲄⲀⲠⲎ
ART  N               VIMP  ART  N
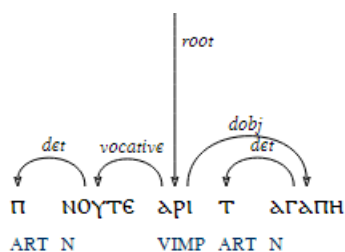
*Oh God, have mercy!*

**xcomp**
This label is used to mark dependent clauses that do not contain their own subject, most often infinitive object clauses.

ТЄТΝ  ΚШ  ΜΜΟ  ΟΥ  ΑΝ  Є  СРЦЄ  Є  ΑΑ  Ц
PPERS  V  PREP  PPERO  NEG  PREP  V  PREP  V  PPERO
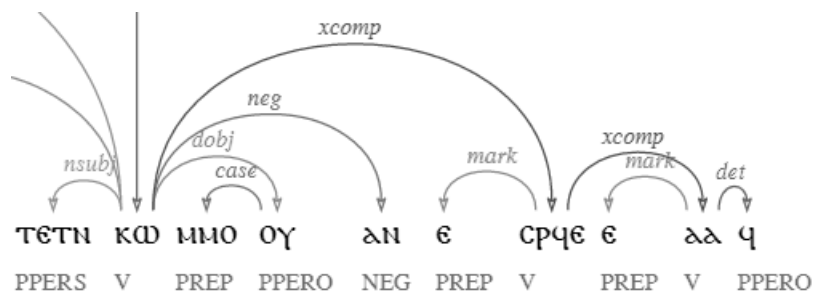
*you do not let them engage in making it*

Note how in the example above, the subject of both infinitives marked by є is the object of the main verb κω, which is external to the infinitive clauses. The infinitive clauses marked by **xcomp** are inheriting their subject argument from the main clause argument.
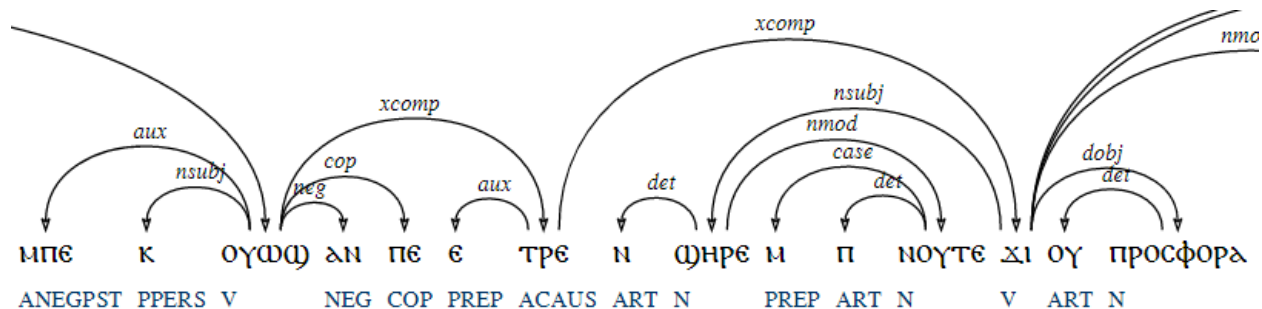
Additionally, the subordinate infinitive of the causative construction with τρє is also analyzed as **xcomp**, although the etymological subject of the auxiliary τρє is attached to the lexical infinitive as a subject. This facilitates syntactic recognition of the construction next to semantic argument structure extraction:



ΜΠЄ  Κ  ΟΥШΩ  ΑΝ  ПЄ  Є  ΤΡЄ  Ν  ШΗΡЄ  Μ  П  ΝΟΥΤЄ  ϪΙ  ΟΥ  ΠΡΟСΦΟΡΑ
ANEGPST  PPERS  V  NEG  COP  PREP  ACAUS  ART  N  PREP  ART  N  V  ART  N

*you did not wish to make the sons of God carry an offering*

Note that in the example above, the first **xcomp** is the normal infinitive case, with no explicit subject, but the second **xcomp** illustrates the causative construction: 'sons' are both the object of 'making' and subject of 'carrying'.

**dep**
For all cases not covered by these guidelines, the label **dep** may be used to denote the unusual dependency. Ideally, these cases should be re-examined and integrated into the guidelines at a later date.

## 5. Special and Confusing Cases

**Non-coordinating Greek conjunctions and particles**
Greek conjunctions and particles that are non-coordinating (i.e. not meaning 'and/or') are labeled as **advmod** to their associated predicate, as in the following example:
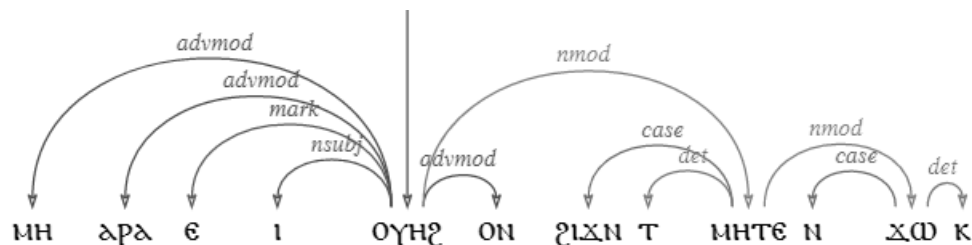


*After all do I still sit upon the middle of your head?*

**Inverted modifying construction - ⲚⲞϬ ⲚϬⲞⲘ**
Inverted modifiers of the type ⲚⲞϬ ⲚϬⲞⲘ 'great power' (lit. a 'great of a power'), are analyzed in purely syntactic terms, such that the semantic modifier (ⲚⲞϬ, ⲔⲞⲨⲒ etc.) is the head, as shown below. The initial article also attaches to the syntactic head. The reason for this is primarily to allow for better parser performance, since making the second noun the head would be a very unusual exception. To find this construction we can look for the set of lexemes appearing in this configuration, most often ⲚⲞϬ and ⲔⲞⲨⲒ.
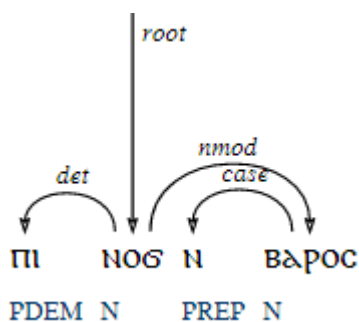


*this great burden*

**Independent possessive pronoun construction – ⲡⲁ/ⲧⲁ/ⲛⲁ + noun phrase**
The independent possessive pronoun 'that, which is of X, belongs to X' is analyzed as the head of the phrase, and the possessor is attached as **nmod** to this:
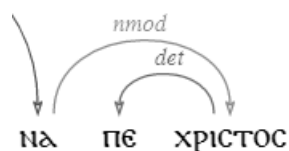


*That, which is Christ's*

# 6. Universal Part of Speech Tags

In keeping with other Universal Dependency treebanks, the Coptic dependency treebank also offers a mapping to the Universal Part of Speech tag set proposed in Petrov et al. (2012) and developed further into the version currently adopted by the UD project. However, unlike the syntactic annotation scheme, which we offer as the primary annotation guidelines for Coptic syntax, universal POS tags are very limited in their coarseness, and in many cases do not map well onto Coptic grammar. We therefore recommend the use of the Coptic specific Scriptorium tag set whenever possible. Below we list the mapping from Coptic Scriptorium to Universal POS tags, with some additional notes at the end.

| Coptic Scriptorium | Universal Tags |
| --- | --- |
| AAOR | AUX |
| ACAUS | AUX |
| ACOND | SCONJ |
| ACONJ | AUX |
| ADV | ADV |
| AFUTCONJ | AUX |
| AJUS | AUX |
| ALIM | SCONJ |
| ANEGAOR | AUX |
| ANEGJUS | AUX |
| ANEGOPT | AUX |
| ANEGPST | AUX |
| ANY | AUX |
| AOPT | AUX |
| APREC | SCONJ |
| APST | AUX |
| ART | DET |
| CCIRC | SCONJ |
| CCOND | SCONJ |
| CFOC | PART |
| CONJ | CONJ |
| COP | PART |
| CPRET | AUX |
| CREL | SCONJ |
| EXIST | VERB |
| FM | X |
| FUT | AUX |
| IMOD | ADV |
| N | NOUN |
| NEG | ADV |

| | |
|---|---|
| NOUN | NOUN |
| NPROP | PROPN |
| NUM | NUM |
| PDEM | DET |
| PINT | PRON |
| PPERI | PRON |
| PPERO | PRON |
| PPERS | PRON |
| PPOS | DET |
| PREP | ADP |
| PTC | PART |
| PUNCT | PUNCT |
| UNKNOWN | X |
| V | VERB |
| VBD | VERB |
| VIMP | VERB |
| VSTAT | VERB |

**Notes**

The Universal POS tags do not map well onto Coptic tags in several cases; in all instances, the attempt has been made to choose the nearest category, especially with syntactic function in mind. The objective is to create dependency trees that connect similar categories to those of other languages.

Most tripartite conjugation bases have been mapped to either auxiliaries (AUX), if they are main clause conjugations (past auxiliary APST, aorist AAOR, etc.) or not the main conjugation morpheme (e.g. future marker FUT, which may join a durative conjugation or irrealis preterit). For the subordinate conjugations (APREC, ALIM), the universal tag SCONJ (subordinating conjunction) is used.

The category IMOD is cast as a form of ADV. While the alternatives of ADP (adposition) or PART (particle) are semantically appealing, the mapping to ADV best represents their sentential function and parallels the dependency label advmod. Note that this results in some adverbs carrying determiners, which is rather odd in terms of underlying categories for the syntax trees. It is perhaps similar to some extent to situations with the Stanford Typed label npadvmod, with the distinction that Coptic IMODs only attach to pronouns, never nouns.

The existential predicates (EXIST) have been mapped as VERB, whereas the copula (COP) is mapped to PTC, since unlike in the case of existence, it does not contain the actual predicate, and is also absent in the interlocutive patterns.

Finally the converters have been treated similarly to conjugation bases, although they co-occur with the bases. Subordinate converters (CCIRC, CREL) are treated as SCONJ, while (potentially) main clause converters (CFOC, CPRET) are tagged as AUX. In all cases, we stress that these are not ideal tag assignments, but ones that aim to stay closest to the limited universal

tag set's behavior. For all new projects we recommend using Scriptorium tags and converting automatically to universal tags if necessary.

## References

Grossman, E. (2014). Transitivity and Valency in Contact: The Case of Coptic. In *47th Annual Meeting of the Societas Linguistica Europaea*. Poznań, Poland.

Layton, B. (2011). *A Coptic Grammar*. Third Edition, Revised and Expanded. (Porta linguarum orientalium 20.) Wiesbaden: Harrassowitz.

de Marneffe, M.-C., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J. & Manning, C. D. (2014). Universal Stanford Dependencies: A Cross-linguistic Typology. In *Proceedings of 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 4585–4592). Reykjavík, Iceland.

Petrov, S., Das, D. & McDonald, R. (2012). A Universal Part-of-speech Tagset. In *Proceedings of LREC 2012*. Istanbul, Turkey.

Zeldes, A. & Schroeder, C. T. (2016). *SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic. Version: 1.1.7_2016.03.14*. Georgetown University and University of the Pacific, Technical Report. Available at: https://github.com/CopticScriptorium/tagger-part-of-speech/raw/master/scriptorium_tagset_documentation.pdf.