

# Coptic Scriptorium – Entity Annotation Guidelines

Amir Zeldes<sup>1</sup> & Lance Martin<sup>2</sup>

1 Georgetown University

2 The Catholic University of America

Version: 1.1.0\_2020.07.20

## 1 Introduction

### 1.1 Preamble

Entity annotation concerns the annotation of **referring expressions** in a text, i.e. spans of text that refer to things in the world, and their classification into **entity types**. The purpose of entity annotation in Coptic Scriptorium is to facilitate searches which include specific entity types (e.g. finding a certain epithet using linguistic annotations, such as οὔλαβ ‘holy’, but only when applied to a PERSON), to inventorize entities (find all cases of e.g. places mentioned in the Apophthegmata Patrum), and to function as a gateway for entity linking, enabling searches for specific persons (“John the Baptist”), regardless of the exact expression used to mention them. The latter task of entity linking is left outside of the scope of the current guidelines.

Entity annotation can be applied to three types of referring expressions:

- Named entities, which are headed by a proper noun (e.g. “Apa Papnoute”)
- Non-named entities, headed by a common noun (e.g. “the angel”)
- Pronouns – these are currently not annotated by our schema (e.g. “she” is a person)

### 1.2 Referring expressions

Almost all nouns and proper nouns correspond to referring expressions, with the exception of non-referring nouns, such as:

- ἀρε πατ.. - “stand, set foot” - does not actually refer to the foot of a person
- εν ογ με - “truly” - does not actually introduce a referenceable ‘truth’

One test for referentiality is whether a pronominal or nominal subsequent mention is possible/plausible. For example, the following sounds odd:

- ?? ἀφἀρεπατϥ ἀϥω περατ ... “he stood on foot, and this foot...”

For more examples, also see the section “Non-referring expressions” below.

## 2 Entity Types

We distinguish 10 entity types:

- ABSTRACT - intangible entities not covered by other classes (incl. ideas, emotions)
- ANIMAL – dog, fish, ...
- EVENT – an occurrence, e.g. “the death of the king”, “the arrival of a monk”
- OBJECT – concrete inanimate object not belonging to other categories
- ORGANIZATION – organized body of people, e.g. ΤΕΚΚΛΗΣΙΑ, ΤΕΣΤΡΑΤΕΙΑ
- PERSON – references to humans, loose groups of humans (ΜΗΜΥΕ ‘crowd’), deities
- PLACE – towns, countries, but also ad-hoc places (behind the house, outside)
- SUBSTANCE – mass noun indicating a material, e.g. sand, water, wine
- TIME – date terms, durations like ‘year’, ‘day’, terms like ‘moment’

## 3 Markable selection guidelines

### 3.1 Appositions

Repeated mentions of the same entity in apposition are considered a single span, and do not contain more mentions of the same entity:

- [ΙΩΘΑΝΝΕΣ Π ΒΑΠΤΙΣΤΗΣ] “[John the Baptist]”
- [Π ΡΡΟ ΖΗΝΩΝ] “[King Zeno]”
- [ΠΕΝ Π ΕΤ ΟΥΑΔΒ ΚΑΤΑ ΣΜΟΤ ΝΙΜ ΑΠΑ ΚΥΡΟΣ Π ΕΝΤ Δ Ψ ...] “[Our Holy One in every way, Αρα Cyrus, who has ...]”

Although outwardly similar, appositions must be distinguished from dislocations, in which a pronominal subject or object is repeated separately. For personal pronouns, the pronoun is simply left out of the nominal span:

- [ΠΕΨ ΕΙΩΤ] Ψ ΝΑΨ ΕΡΟ ΟΨ - “[his father], he sees them”
- Ψ ΝΑΨ ΕΡΟ ΟΨ ΝΒΙ [ΠΕΨ ΕΙΩΤ] - “he sees them, that is [his father]”

If the pronoun is a substitutive demonstrative (ΠΑΙ, ΤΑΙ, ΝΑΙ), then two spans are annotated:

- [ΠΕΨ ΕΙΩΤ] [ΠΑΙ] ΝΑΨ ΕΡΟ ΟΨ - “[his father], [this one] sees them”
- [ΠΑΙ] ΝΑΨ ΕΡΟ ΟΨ ΝΒΙ [ΠΕΨ ΕΙΩΤ] - “[this one] sees them, [that is his father]”

But note that it is also possible for a substitutive demonstrative to stand in true apposition to a noun without dislocation, in which case a single span is annotated as for any apposition:

- Δ Ι ΝΑΨ Ε [ΠΕΨ ΕΙΩΤ , ΠΑΙ ΕΤ ΜΕΡΙΤ Ψ] - “I saw [their father, the one who loves her]”

See the UD Coptic guidelines for more information on identifying dislocation vs. apposition.

### 3.2 Expanded Relative Constructions

The relative construction expanding an article is annotated as an entity:

- [π ετ ου σωτη ερο ς] “the one they listened to” (person)

However, if the π is tagged as a copula, that part of the construction is not part of the entity span, since it is part of a predication. In these instances, we view the predicate noun phrase as an entity, and the relative clause as a subject clause (compare the Universal Dependencies annotation guidelines):

- [π νουτε] π/COP εντ α ς αγζανε “It is God who made them grow”

In this example, “God” receives a span, but “who made them grow” is considered a subject clause (i.e. ‘who made them grow is God’), which is not nominal and hence not annotated. Note that according to the tagging guidelines, the second π should be tagged as COP and lemmatized πε in this sentence. Therefore there is only a single entity mention in this case, [π νουτε].

Nominal sentences with a relative construction expanding the article should not be mistaken for the copula construction although similar in form. Compare the following, in which the π in the second phrase is tagged ART:

- [π νουτε] [π/ART εντ α ς αγζανε] πε/COP “[God] is [the one who made them grow]”
- [πεν δωεις] [π/ART νουτε ετ σωτη ερο ν] πε/COP “Our Lord is the God who hears us”
- [πεν δωεις] πε [π νουτε ετ σωτη ερο ν] “It is our Lord, the God who hears us”

In the first example, both “God” and “the one who made them grow” receive spans. “God” is the subject and “the one who made them grow” is the predicate, forming a nominal sentence with the copula πε. The π after νουτε is part of the predicate noun phrase in this instance: an article heading the relative clause which is the predicate for the copula at the end. It functions similarly to the relative clause expanding the nouns, i.e., “the God who hears us” in the subsequent examples.

The expansion of an article can stand in apposition if it repeats an entity mention. The repeated mention is annotated with one span, like all appositions, whereas subject and predicate spans are considered separate mentions:

- [π νουτε , π εντ α ς αγζανε] [π αγαθος] πε “God, the one who made them grow, he is the good one”

Here, “the one who made them grow” is again headed by an article and stands in apposition to “God.” The final *πε* is the copula in the sentence and marks the noun phrase before it as a predicate, “the good one.”

The position of the copula does not change the meaning and is not annotated. Note also that in first and second person, there is no copula:

- *αηΓ [ΟΥ ΧΡΙΣΤΙΑΝΟΣ]* “I am [a Christian]”

Occasionally, however, the copula can interrupt a referring expression as in the example below where the head noun *πνογτε* is separate from the relative clause by the copula, *πε*. In such cases, it should be included in the span (see 3.3 for more detail).

- *[ΠΕΝ ΔΟΕΙΣ] [Π ΝΟΥΤΕ ΠΕ ΕΤ ΣΩΤΜ ΕΡΟ Ν]* “Our Lord is the God who hears us”

### 3.3 Interrupted spans

Entity expressions interrupted e.g. by a copula or particle are spanned to **contain** the copula or particle. For example, the following span includes the intervening copula:

- *[ΝΕΓ ΑΠΟΣΤΟΛΟΣ ΝΕ ΕΤΟΥΔΑΒ]<sub>PERSON</sub>* ‘it is his holy apostles’ (literally [*his apostles are which holy*], with intruding ‘are’)

Similarly:

- *[ΥΤΟΥ ΔΕ Ν ΖΟΥ]<sub>TIME</sub>* – but four days (lit. ‘[four but days]’)
- *[ΟΥ ΩΒΗΡ · ΖΩΩ Κ ΟΝ ΝΤΕ Π ΝΟΥΤΕ]<sub>PERSON</sub>* – but also for your part a friend of God

Non-adjacent relative clauses are included, **unless the interruption contains the verb controlling the head noun** (this prevents some possibly very long ‘hermeneutical’ relatives inside mentions):

- *[ΡΩΜΕ ΝΙΜ ΟΝ ΕΤ ΣΩΤΜ]* - and also [any man who hears] (note the interruption ‘ον’, and inclusion of the relative clause)

But do not include a clause past the verb controlling the head of the span:

- *ΕΡΩΔΗΝ [Τ ΒΑΩΟΡ]<sub>ANIMAL</sub> ΔΥΚΑΚ ΕΒΟΛ ΑΝ **ΕΤΕ ΝΤΟΚ ΠΕ** ...* - it is not when [the fox] barks, which is you, ... (postponed hermeneutic relative clause in bold is not included, because it appears after the verb *δυκακ*, which controls ‘fox’ as a subject)

In this case the interruption by the verb *δυκακ* ‘bark’ which is the predicate of ‘fox’ triggers the guideline to omit the relative clause. Otherwise, the mention could potentially cover the entire

clause, in this case: Τ ΒΑΨΟΡ ΔΨΚΑΚ ΕΒΟΛ ΔΝ ΕΤΕ ΝΤΟΚ ΠΕ Π ΖΜΞΔΛ Η Π ΗΑΜΜΩΝΔΣ ΖΝ ΖΕΝ ΖΡΟΟΥ  
Ε Υ ΟΩ . . . .

### 3.4 Possessive Constructions

The possessive article construction, e.g. πα παγλος, forms two spans, as follows, with the entity type being decided based on meaning:

- [πα [παγλος]<sub>PERSON</sub>]<sub>PERSON</sub> - the ones (=people) belonging to Paul

But note that regular possessive articles are not annotated with spans, just as other pronouns are not annotated:

- [πεφ η]<sub>PLACE</sub> – his house

### 3.5 Groups and other quantity constructions

Semantically ‘empty’ heads such as quantity nouns (compare English ‘a number of people’, which is not both ‘a number’ and ‘people’; similarly ‘a lot of’, ‘the majority of’ etc.) are only given one span, for example:

- [ζεε Ν κοπ]<sub>TIME</sub> – a lot of times

Groups of entities are generally interpreted as the entity type of their constituents, for example, a herd of animals is of the type animal:

- [ογ αγελη Ν ψοψ]<sub>ANIMAL</sub> - a herd of buffaloes

Note that there is no nested entity for 'buffalo' in this case, since there is no distinct entity being mentioned (the herd consists exactly of all buffaloes being discussed). This is different in cases where the nested entity is not identical in reference, e.g. ‘[the houses of [the city]]’, where ‘the city’ can be said to contain more than just the houses.

An exception to the guideline that groups are classified as their constituent type is cases of people who form an organization, e.g. σϋναγωγη, στρατεγμα etc. are 'organization', not 'person'.

### 3.6 Idiomatic Expressions

Idiomatic expressions should be annotated as fully as possible even when certain components have low referentiality.

- [Ν Ζαλατε Ν [Τ ΠΕ]<sub>PLACE</sub>]<sub>ANIMAL</sub> “the birds of the sky (Mark 4:32)

In this example, the entire phrase may be a way of referring to birds in general, and “sky” therefore is unlikely to be referred to again. However, since it is possible that it could be referred to again and since it passes the referentiality test (1.2), it should be annotated.

### 3.7 No reference inside compounds

In morphologically complex items containing a verb inside a larger token, that noun cannot be annotated:

- α ϣ ΔΙΒΑΠΤΙΣΜΑ - he received-baptism

In this case baptism cannot be annotated as an entity, since it’s part of an incorporated verb ‘to baptize’, and receives the part of speech V in Coptic Scriptorium guidelines.

### 3.8 Coordination

Do not mark coordinate entities in addition to their constituents:

- [ΙΩΖΑΝΝΗΣ] ΜΝ [ΑΝΤΩΝΙΟΣ]

In this case we do not also annotate [ΙΩΖΑΝΝΗΣ ΜΝ ΑΝΤΩΝΙΟΣ] as a third mentioned entity.

### 3.9 Container and substance

Container and substance form two entities, for example:

- [ΟΥ ΠΥΓΗ Μ [ΜΟΟΥ]<sub>SUBSTANCE</sub>]<sub>PLACE</sub> - a fountain of water

The fountain can be a PLACE or OBJECT in context, but the water is SUBSTANCE, and both can be referred to separately later on.

### 3.10 Numeral entities

Numbers standing in for a phrase can be entities, most commonly:

- [ΟΥΑ]<sub>PERSON</sub> - ‘one (person)’

### 3.11 Distributive entities

Repeated distributive noun constructions are interpreted as single entity mentions:

- [Π ΟΥΑ Π ΟΥΑ]<sub>PERSON</sub> - ‘one by one’, ‘each man’

The rationale is that these are like a plural reference, rather than two mentions of individuals (in this case there can be more than two people, and they do not map neatly onto the two numerals).

## 4 Entity classification guidelines

### 4.1 Body Parts

Most body parts are marked as objects, since they are tangible:

- [ΟΥ ΓΙΧ]<sub>OBJECT</sub> – “a hand”
- [ΠΕΦ ΒΑΛ]<sub>OBJECT</sub> – “his eye”

However some referential body parts are usually considered abstract, notably ζΗΤ ‘heart’, which usually refers to one’s spirit, emotions, etc., and not the physical organ:

- † ΝΑ ΤΡΕ [ΠΟΥ ΖΗΤ]<sub>ABSTRACT</sub> ΜΚΑΖ – “I will make your [heart] suffer”

Other uses of body parts may be totally figurative or idiomatic (i.e. not referring to anything), in which case they are not annotated – see ‘Non-Referring Expressions’ above.

### 4.2 Parts of Plants

Parts of plants are marked as plants.

- [Π ΚΛΑΔΟΣ]<sub>PLANT</sub> “the branch”
- [Π ΣΡΟΣ]<sub>PLANT</sub> “the seed”

### 4.3 Peoples and demonyms

Pluralized demonyms indicating members of a people are labeled person:

- [Ν ΖΕΛΛΗΝ]<sub>PERSON</sub>

However peoples mentioned as a people (not as a group of individuals) are labeled organization:

- [ΠΕΚ ΛΑΟΣ ΙΣΡΑΗΛ]<sub>ORGANIZATION</sub>

These cases are usually singular and involve a named people. This guideline does not apply to ad-hoc groups of people who do not form an organized entity, e.g. ΜΗΝΩΕ ‘crowd’ is still annotated as PERSON.

## 4.4 Substance and Object

Inanimate entities that can be counted should be marked as object:

- [ἡ πέτρα]<sub>OBJECT</sub> “rocks”
- [τρεῖς ἄρτοι]<sub>OBJECT</sub> “Three breads (i.e., three loaves of bread)”

Inanimate entities that cannot be counted should be marked as substance:

- [ἡ γῆ]<sub>SUBSTANCE</sub> “the soil”
- [ἡ ὕδατος]<sub>SUBSTANCE</sub> “water”

## 4.5 Event and Abstract

Nominalized infinitives are often events:

- [ἡ ἀποκτείνωσις]<sub>EVENT</sub> “the murder”

## 4.6 Multiple entity types in one chain

It is possible for an entity to be referred to in multiple ways, underscoring different aspects of the entity. In such cases, where it seems clear that the entity types are distinct (e.g. due to metonymy, metaphorical extensions, etc.), it is possible to have different entity types

- [ὁ Χριστός]<sub>PERSON</sub> πᾶς [ἡ θεμελίωσις]<sub>ABSTRACT</sub> - ‘this foundation is Christ’ (where ‘foundation’ refers to ‘Christ’ metaphorically)

Mark each entity with its own type, i.e., ‘Christ’ as PERSON and ‘foundation’ as ABSTRACT.

# 5 Non-referring expressions

## 5.1 Interrogatives

No annotations are needed for plain interrogatives (τίς, τίς), but complex interrogatives including a noun are annotated:

- [τίς ἀνθρώπος]<sub>PERSON</sub> οἶσθαι τὰ ἄνθρωπων ἢ [ὁ ἀνθρώπος]<sub>PERSON</sub> [ὁ ἀνθρώπος]<sub>PERSON</sub> “[what human] is he who knows [those things which are of humans]?”
- [τίς ἀποκτείνωσις]<sub>EVENT</sub> “[what kind of trial]?”

## 5.2 Common figurative and other fixed expressions to ignore

The following are considered idiomatic or functional expressions, in which the constituent nouns are not construed as referential and no annotation is needed:



- ἀρε πατ ς – ‘stand, set foot’ - “foot” is not an entity mention
- ε πι εχτη - ‘down’, lit. ‘to the ground’
- ε πι τηρς meaning ‘at all’ is not referential
- νεϿ ερηϿ “themselves”
- μη μηνε - ‘daily’
- ν οϿ κοϿι - ‘a little’ (manner adverbial; note κοϿι \*can\* be referential if referring to a person or thing, e.g. ντκ [οϿ κοϿι]<sub>PERSON</sub> ‘you are a little one’)
- ν οϿ ζοϿο - ‘more so’
- ν οϿωτ - ‘together’
- ν ωορπ - ‘first(ly)’
- ν τε ϿνοϿ - ‘then’
- ν τ ζε - meaning ‘like’
- παζοϿ μη
- ρ ζνα ς – ‘want, do one’s will’ - the word ζνα / ζνε ‘will’ is figurative, as this is a fixed expression for ‘desire’
- ζα εοοϿ - ‘glorious’ - the εοοϿ is not referential, as the expression is only used adjectivally
- ζι οϿ σοπ - ‘at once’
- σομ - meaning ‘capable’ in constructions like ντκ σομ αν ‘you are not capable’, but σομ can be referential in, e.g. [τ σομ μη [πι νοϿτε]]<sub>ABSTRACT</sub>
- (ζωβ) ν σιϿ – ‘handywork’ - the whole phrase (handywork) is ABSTRACT or OBJECT in context, but ‘hand’ is not a referent
- † τοοτ ς – ‘help, give a hand’

### 5.3 Entities with low referentiality

If it is unclear if a span is referring or non-referring, it should be annotated if it has an article and is not on the list in 5.2. Otherwise, it should not receive annotation:

- ἀϿντοϿ ε [οϿ ςα]<sub>PLACE</sub> - “he took them to [a side]”

However, “behind” should not be annotated since it lacks an article:

- ε παζοϿ μη ... - “behind...”