Coptic Scriptorium – Entity Annotation Guidelines

Amir Zeldes
Georgetown University

Version: 1.0.0_2020.05.20

1 Introduction

1.1 Premable

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 ⲟⲩⲁⲁⲃ '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:

- ⲁϩⲉ ⲣⲁⲧ.. - "stand, set foot" - does not actually refer to the foot of a person

- ϩɴ ογ ⲙⲉ - “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:

- ?? ⲁⲫⲁϩⲉⲣⲁⲧϥ ⲁⲩⲱ ⲡⲉⲓⲣⲁⲧ …“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. ⲧⲉⲕⲕⲗⲏⲥⲓⲁ, ⲧⲉⲥⲧⲣⲁⲧⲉⲓⲁ
- PERSON – references to humans, loose groups of humans (ⲙⲛⲏϣⲉ ‘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:

- [ⲓⲱϩⲁⲛⲛⲉⲥ ⲡ ⲃⲁⲡⲧⲓⲥⲧⲏⲥ] “[John the Baptist]”
- [ⲡ ⲣⲣⲟ ⲍⲏⲛⲱⲛ] “[King Zeno]”
- [ⲡⲉⲛ ⲡ ⲉⲧ ⲟⲩⲁⲁⲃ ⲕⲁⲧⲁ ⲥⲙⲟⲧ ⲛⲓⲙ ⲁⲡⲁ ⲕⲩⲣⲟⲥ ⲡ ⲉⲛⲧ ⲁ ⲫ …] “[Our Holy One in every way, Apa 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:

- [ⲡⲉϥ ⲉⲓⲱⲧ] ϥ ⲛⲁⲩ ⲉⲣⲟ ⲟⲩ - “[his father], he sees them”
- ϥ ⲛⲁⲩ ⲉⲣⲟ ⲟⲩ ⲛⲑⲓ [ⲡⲉϥ ⲉⲓⲱⲧ] - “he sees them, that is [his father]”

If the pronoun is a substitutive demonstrative (ⲡⲁⲓ, ⲧⲁⲓ, ⲛⲁⲓ), then two spans are annotated:

- [ⲡⲉϥ ⲉⲓⲱⲧ] [ⲡⲁⲓ] ⲛⲁⲩ ⲉⲣⲟ ⲟⲩ - “[his father], [this one] sees them”
- [ⲡⲁⲓ] ⲛⲁⲩ ⲉⲣⲟ ⲟⲩ ⲛⲑⲓ [ⲡⲉϥ ⲉⲓⲱⲧ] - “[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:

- ⲁ ⲓ ⲛⲁⲩ ⲉ [ⲡⲉϥ ⲉⲓⲱⲧ , ⲡⲁⲓ ⲉⲧ ⲙⲉⲣⲓⲧ ⲥ] - “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:

- [ⲡ ⲉⲧ ⲟⲩ ⲥⲱⲧⲙ ⲉⲣⲟ ϥ] “the one they listened to” (person)

However, if the ⲡ 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 Dependency annotation guidelines):

- [ⲡ ⲛⲟⲩⲧⲉ] ⲡ/COP ⲉⲛⲧ ⲁ ϥ ⲁⲩϫⲁⲛⲉ “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 ⲡ should be tagged as COP and lemmatized ⲡⲉ in this sentence.

3.3 Possessive Constructions

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

- [ⲡⲁ [ⲡⲁⲩⲗⲟⲥ]$_{\text{PERSON}}$]$_{\text{PERSON}}$ - the ones (=people) belonging to Paul

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

- [ⲡⲉϥ ⲏⲓ]$_{\text{PLACE}}$ – his houst

3.4 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:

- [ⲛⲉϥ ⲁⲡⲟⲥⲧⲟⲗⲟⲥ **ⲛⲉ** ⲉⲧⲟⲩⲁⲁⲃ]$_{\text{PERSON}}$ ‘it is his holy apostles’ (literally [*his apostles are which holy*], with intruding ‘are’)

Similarly:

- [ϥⲧⲟⲟⲩ ⲇⲉ ⲛ ϩⲟⲟⲩ]$_{\text{TIME}}$ – but four days (lit. ‘[four but days]’)
- [ⲟⲩ ⲱⲃⲏⲣ · **ϩⲱⲱ ⲕ ⲟⲛ** ⲛⲧⲉ ⲡ ⲛⲟⲩⲧⲉ]$_{\text{PERSON}}$ – 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):

- [ⲣⲱⲙⲉ ⲛⲓⲙ ⲟⲛ ⲉⲧ ⲥⲱⲧⲙ] - and also [any man who hears] (note the interruption 'ⲟⲛ', and inclusion of the relative clause)

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

- ⲉⲣⲱⲁⲛ [ⲧ ⲃⲁⲱⲟⲣ]$_{\text{ANIMAL}}$ ⲁⲱⲕⲁⲕ ⲉⲃⲟⲗ ⲁⲛ **ⲉⲧⲉ ⲛⲧⲟⲕ ⲡⲉ** … - 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 ⲁⲱⲕⲁⲕ, which controls ‘fox’ as a subject)

In this case the interruption by the verb ⲁⲱⲕⲁⲕ ‘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: ⲧ ⲃⲁϣⲟⲣ ⲁϣⲕⲁⲕ ⲉⲃⲟⲗ ⲁⲛ ⲉⲧⲉ ⲛⲧⲟⲕ ⲡⲉ ⲡ ⲯⲩϩⲁⲗ ⲙ ⲡ ⲙⲁⲙⲙⲱⲛⲁⲥ ϩⲛ ϩⲉⲛ ϩⲣⲟⲟⲩ ⲉ ⲩ ⲟϣ ….

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:

- [ϩⲁϩ ⲛ ⲥⲟⲡ]$_{\text{TIME}}$ – 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:

- [ⲟⲩ ⲁⲣⲉⲗⲏ ⲛ ϣⲟϣ]$_{\text{ANIMAL}}$ - 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. ⲥⲩⲛⲁⲣⲱⲣⲏ, ⲥⲧⲣⲁⲧⲉⲩⲙⲁ etc. are 'organization', not 'person'.

3.6 No reference inside compounds

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

- ⲁ ⲫ ϫⲓⲃⲁⲡⲧⲓⲥⲙⲁ - 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.

3.7 Coordination

Do not mark coordinate entities in addition to their constituents:

- [ⲓⲱϩⲁⲛⲛⲏⲥ] ⲙⲛ [ⲁⲛⲧⲱⲛⲓⲟⲥ]

In this case we do not also annotate [ⲓⲱϩⲁⲛⲛⲏⲥ ⲙⲛ ⲁⲛⲧⲱⲛⲓⲟⲥ] as a third mentioned entity.

3.8 Container and substance

Container and substance form two entities, for example:

- [ⲟⲩ ⲡⲩⲧⲏ ⲙ [ⲙⲟⲟⲩ]$_{\text{SUBSTANCE}}$]$_{\text{PLACE}}$ - 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.

3.9 Distributive entities

Repeated distributive noun constructions are interpreted as single entity mentions:

- [ⲡ ⲟⲩⲁ ⲡ ⲟⲩⲁ]$_{\text{PERSON}}$ - '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:

- [ⲟⲩ ϭⲓⲭ]$_{\text{OBJECT}}$ – "a hand"
- [ⲡⲉϥ ⲃⲁⲗ]$_{\text{OBJECT}}$ – "his eye"

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

- ϯ ⲛⲁ ⲧⲣⲉ [ⲡⲟⲩ ϩⲏⲧ]$_{\text{ABSTRACT}}$ ⲙⲕⲁϩ – "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 Peoples and demonyms

Pluralized demonyms indicating members of a people are labeled person:

- [ⲛ ⲯⲉⲗⲗⲏⲛ]$_{\text{PERSON}}$

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

- [ⲡⲉⲕ ⲗⲁⲟⲥ ⲓⲥⲣⲁⲏⲗ]$_{\text{ORGANIZATION}}$

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. ⲙⲛⲛϣⲉ 'crowd' is still annotated as PERSON.

4.3 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

- [ⲡⲉ ⲭⲣⲓⲥⲧⲟⲥ]$_{\text{PERSON}}$ ⲡⲉ [ⲧⲉⲓ ⲥⲛⲧⲉ]$_{\text{ABSTRACT}}$ - '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 (ⲛⲓⲙ, ⲟⲩ), but complex interrogatives including a noun are annotated:

- [ⲛⲓⲙ ⲅⲁⲣ ⲛ ⲣⲱⲙⲉ]$_{\text{PERSON}}$ ⲡ ⲉⲧ ⲥⲟⲟⲩⲛ ⲛ [ⲛⲁ [ⲛ ⲣⲱⲙⲉ]$_{\text{PERSON}}$]$_{\text{PERSON}}$ "[what human] is he who knows [those things which are of humans]]?"
- [ⲁϣ ⲙ ⲡⲓⲣⲁⲥⲙⲟⲥ]$_{\text{EVENT}}$ "[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:

- ⲁϩⲉ ⲣⲁⲧ ⲫ – 'stand, set foot' - "foot" is not an entity mention
- ⲧ ⲧⲟⲟⲧ ⲫ – 'help, give a hand'
- ⲉ ⲡ ⲉⲥⲏⲧ - 'down', lit. 'to the ground'

- ⲛ ⲟⲩ ⲉⲟⲩⲟ - ‘more so’
- (ⲉⲱⲃ) ⲛ ⲥⲓⲭ – ‘handywork’ - the whole phrase (handywork) is ABSTRACT or OBJECT in context, but ‘hand’ is not a referent
- ⲣ ⲉⲛⲁ ⲫ – ‘want, do one’s will’ - the word ⲉⲛⲁ / ⲉⲛⲉ ‘will’ is figurative, as this is a fixed expression for ‘desire’
- ⲉ ⲡ ⲧⲏⲣⲫ meaning ‘at all’ is not referential
- ⲉⲁ ⲉⲟⲟⲩ - ‘glorious’ - the ⲉⲟⲟⲩ is not referential, as the expression is only used adjectivally
- ⲛ ⲟⲩ ⲕⲟⲩⲓ - ‘a little’ (manner adverbial; note ⲕⲟⲩⲓ *can* be referential if referring to a person or thing, e.g. ⲛⲧⲕ [ⲟⲩ ⲕⲟⲩⲓ]$_{\text{PERSON}}$ ‘you are a little one)
- ⲛ ⲧ ⲉⲉ - meaning ‘like’
- ⲛ ϣⲟⲣⲡ - ‘first(ly)’
- ⲥⲟⲙ - meaning ‘capable’ in constructions like ⲛⲧⲕ ⲥⲟⲙ ⲁⲛ ‘you are not capable’, but ⲥⲟⲙ can be referential in, e.g. [ⲧ ⲥⲟⲙ ⲙ [ⲡ ⲛⲟⲩⲧⲉ]]$_{\text{ABSTRACT}}$
- ⲛ ⲟⲩⲱⲧ - ‘together’
- ⲉⲓ ⲟⲩ ⲥⲟⲡ - ‘at once’
- ⲙ ⲙⲏⲛⲉ - ‘daily’