**SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic**

Amir Zeldes\(^1\) & Caroline T. Schroeder\(^2\)

1. Georgetown University  
2. University of the Pacific

**Version:** 1.1.7_2016.03.14

1. **Preamble**  
   This document details guidelines for part-of-speech tagging Sahidic Coptic according to the SCRIPTORIUM project scheme. The tagging procedure assumes the text has already been normalized to the orthography and morpheme based segmentation described in the SCRIPTORIUM tokenization guidelines, which are closely related to the conventions found in Layton’s (2004) grammar. In case of doubt we refer to Layton (2004) as well as Shisha-Halevy (1988).

   As in all tagging projects, the aim is to achieve a practicable compromise between linguistic accuracy/usefulness, speed and reliability of human tagging, and performance of automatic tagging software. This means that in many cases concepts that are linguistically distinct are not distinguished since they are difficult to tell apart in practice in many cases, or determining some distinctions is too costly in terms of annotation time. Additionally, the project is using the CMCL lexicon, kindly provided by Prof. Tito Orlandi, which has its own, much more detailed scheme, so that in some cases the categories used here are chosen to be derivable from the CMCL scheme (see [http://cmcl.let.uniroma1.it/](http://cmcl.let.uniroma1.it/)).

   There are two proposed tagsets, a coarse tagset with fewer tags for projects wishing to save annotation time, and a finer tagset with more detailed subcategories for some of the coarse grained tags, which is also expected to yield lower accuracy in automatic tagging. Links to the latest training models are provided from the SCRIPTORIUM website and have been tested and developed using the freely available TreeTagger (Schmid 1994, see [http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/](http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/)).

2. **Tags**  
   The two tagsets described below are compatible with each other in that the fine-grained tagset uses the same overarching categories of the coarse one, but with further categories distinguished. The tag names are built ‘hierarchically’, so that additional letters in the name of a tag specify a special type of the superordinate category, e.g. all pronoun tags being with P, though not all tags with P are pronouns, as in PREP for prepositions.
In the coarse-grained list below, tags that have multiple fine-grained variants are followed by [*] (this is not part of the tag within the course-grained tagset).

Additionally, both tagsets admit certain cases where a single form contains two categories and must therefore be assigned two tags. This results in special underscore separated portmanteau tags, which are described in Section 2.3.

### 2.1 Coarse-Grained Tagset

<table>
<thead>
<tr>
<th>Tag</th>
<th>Name</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A[*]</td>
<td>Auxiliary tripartite base</td>
<td>δ[θ], ῳ[θ], τρ[θ], ...</td>
</tr>
<tr>
<td>ADV</td>
<td>Adverb</td>
<td>εβολ, ον, πως</td>
</tr>
<tr>
<td>ART</td>
<td>Article</td>
<td>π(ε), τ(ε), η(ε), ρην, κε</td>
</tr>
<tr>
<td>C[*]</td>
<td>Converter</td>
<td>ε, ετε, κε, ...</td>
</tr>
<tr>
<td>CONJ</td>
<td>Conjunction</td>
<td>λψυ, η, κε, καλ, ετε, ...</td>
</tr>
<tr>
<td>COP</td>
<td>Copula</td>
<td>πε/τε/κε</td>
</tr>
<tr>
<td>EXIST</td>
<td>Existential/possessive</td>
<td>οψη/θη</td>
</tr>
<tr>
<td>FM</td>
<td>Foreign material</td>
<td>παρα τογτο</td>
</tr>
<tr>
<td>FUT</td>
<td>Future</td>
<td>ηα</td>
</tr>
<tr>
<td>IMOD</td>
<td>Inflected modifier</td>
<td>τνη[θ], ῳαρ[τ], ...</td>
</tr>
<tr>
<td>N[*]</td>
<td>Noun</td>
<td>λψυτ, ρωνε, λρη, ...</td>
</tr>
<tr>
<td>NEG</td>
<td>Negation</td>
<td>η, αι, ηε[σωθη]</td>
</tr>
<tr>
<td>NUM</td>
<td>Numeral</td>
<td>οψα, σαγ, ...</td>
</tr>
<tr>
<td>PDEM</td>
<td>Pronoun, demonstrative</td>
<td>πευ/παι, τευ/ται, ηε/ναι</td>
</tr>
<tr>
<td>PINT</td>
<td>Pronoun, interrogative</td>
<td>οψ, κη</td>
</tr>
<tr>
<td>PPER[*]</td>
<td>Pronoun, personal</td>
<td>θι, ι, ϊ, ...</td>
</tr>
<tr>
<td>PPOS</td>
<td>Pronoun, possessive</td>
<td>πευ, τετη, που, ια, ιαι, ...</td>
</tr>
<tr>
<td>PREP</td>
<td>Preposition</td>
<td>ετβε, ρη, η, ληνο[κ], ...</td>
</tr>
<tr>
<td>PTC</td>
<td>Particle</td>
<td>Δε, ιδι, λε, ...</td>
</tr>
<tr>
<td>PUNCT</td>
<td>Punctuation</td>
<td>. . . . . .</td>
</tr>
<tr>
<td>UNKNOWN</td>
<td>Unknown morph, lacuna</td>
<td>ου_ _ _ , _ _ _ , _ _ _ , . . .</td>
</tr>
<tr>
<td>V[*]</td>
<td>Verb</td>
<td>σωθη, σωττ, σοττ, ερε, ο, αρι, ...</td>
</tr>
<tr>
<td>VBD</td>
<td>Verboid</td>
<td>νανου[κ], πεξα[κ], πεξε, ...</td>
</tr>
</tbody>
</table>

### 2.2 Fine-Grained Tagset

For descriptions of the added fine-grained tags, marked in cursive type, see the coarse tag descriptions below.

<table>
<thead>
<tr>
<th>Tag</th>
<th>Name</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACAUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACOND</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACONJ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADV</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFUTCONJ</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tag</th>
<th>Name</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>AJUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALIM</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANEGAOR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANEGJUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANEGOPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ART</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANEGPST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AOPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APST</td>
<td></td>
<td></td>
</tr>
<tr>
<td>APREC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCIRC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.3 Portmanteau tags

In certain cases, one indivisible form corresponds to what normally constitutes two categories. This can happen either because of a phonological merger of two units, or because the formal marker of one category can be ‘zero’, i.e. have no form at all (usually in the case of 2\textsuperscript{nd} person singular feminine forms). Portmanteau tags currently supported by the SCRIPTORIUM tools are:

<table>
<thead>
<tr>
<th>tag</th>
<th>example</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOPT_PPERS</td>
<td>(\epsilon ϕ)(\sigma \tau \mu)(\tau)</td>
<td>Personal pron. within optative (\epsilon___). Note that (\epsilon ϕ)(\sigma \tau \mu)(\tau) for 2nd pers. sg. fem. is also AOPT_PPERS, but nominal (\epsilon ϕ)(\pi ρ\omega \epsilon \sigma \tau \mu) is only AOPT.</td>
</tr>
<tr>
<td>ACOND_PPERS</td>
<td>(\epsilon ϕ\alpha\alpha)(\sigma \tau \mu)(\tau)</td>
<td>Personal pron. within conditional (\epsilon_\alpha\alpha). Note that (\epsilon ϕ\alpha\alpha)(\sigma \tau \mu)(\tau) for 2nd pers. sg. fem. is also ACOND_PPERS, but nominal (\epsilon ϕ\alpha\alpha)(\pi ρ\omega \epsilon \sigma \tau \mu) is only ACOND.</td>
</tr>
<tr>
<td>ACONJ_PPERS</td>
<td>τ(\alpha)(\sigma \tau \mu)(\tau)</td>
<td>Truncated conjunctive 1st person (instead of (\eta\tau\alpha\sigma \tau \mu))</td>
</tr>
<tr>
<td>ANEGPST_PPERS</td>
<td>(\mu\pi\epsilon)(\sigma \tau \mu)(\tau)</td>
<td>Fused negative past 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>APST_PPERS</td>
<td>(\alpha\rho)(\sigma \tau \mu)(\tau)</td>
<td>Fused positive past 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>CCIRC_PPERS</td>
<td>(\epsilon ϕ)(\sigma \tau \mu)(\tau)</td>
<td>Fused circumstantial 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>CFOC_PPERS</td>
<td>(\epsilon ϕ)(\sigma \tau \mu)(\tau)</td>
<td>Fused focalized 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>CPRET_PPERS</td>
<td>(\eta\rho\epsilon)(\sigma \tau \mu)(\tau)</td>
<td>Fused preterit 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>CREL_PPERS</td>
<td>(\epsilon\theta\epsilon ϕ)(\sigma \tau \mu)(\tau)</td>
<td>Fused relative 2nd pers. sg. fem. form.</td>
</tr>
<tr>
<td>IMOD_PPERO</td>
<td>(\mu\mu\mu\mu\mu)</td>
<td>The 2nd pers. sg. fem. form of ‘yourself’ (not to be confused with (\mu\mu\mu\mu\mu)(\rho) etc.)</td>
</tr>
<tr>
<td>PREP_PPERO</td>
<td>(\epsilon \rho)</td>
<td>Any preposition where a 2nd pers. sg. fem. is realized as zero (not to be confused with (\epsilon \rho)(\rho) etc.)</td>
</tr>
<tr>
<td>V_PPERO</td>
<td>(\eta\tau) (\mu)</td>
<td>Verb forms with a fused 1st pers. sg. object, e.g. (\eta\tau) ‘bring me’ from (\epsilon\omega) ‘bring’, where the presuffixal form (\eta\tau) is merged with the 1st pers. object marker (-\tau)</td>
</tr>
</tbody>
</table>

Note that in all cases, coarse grained tags can be substituted for fine grained ones, e.g. CCIRC_PPERO and CFOC_PPERO both become C_PPER. Further combination tags are not ruled out and new ones will therefore be added if they are determined to be necessary.
2.4 Part of speech in conversion
In rare cases, a part of speech may appear in a syntactically unusual position. For example, an adverb or preposition may follow an article if they begin a phrase that is treated as a nominal phrase syntactically: the word ἐβολ is tagged as an ADV, although in the sequence ἀγ'ἐβολ ἐμὲ ἐκῶς ἐχον ‘one (which is) out of the body’, it appears to behave like a noun. We consider such cases of ‘conversion’ between categories to be a syntactic phenomenon, and we therefore continue to tag ἐβολ morphologically as an adverb.

An exception to this rule is the tagging of verbal infinitives following an article. In essence, almost any Coptic infinitive may be used as a noun, for example ποταμ ἡ ‘the call’. Cases such as these are widespread and are tagged as nouns, not as verbs, when the infinitive is used in this way.

3. Guidelines
The following guidelines describe the recommended assignment of part of speech tags to segmented morphemes. Fine-grained tags are given in the section describing the corresponding coarse-grained tag. In each example, the area corresponding to the tag under discussion is underlined. Vertical lines (‘pipes’) are used to segment morphemes for added clarity only.

3.1 Auxiliaries (A)
Auxiliaries include all conjugation bases in the tripartite patterns described in Layton (2004:251-290). These include both negative and positive variants and cover all lexical material preceding the subject noun or pronoun, e.g.:

(1) ἀ|ψ|σατ | (3rd person masculine past tense)
(2) ἀ|κ|σατ ὑ | (2nd person feminine past tense, with zero subject)
(3) Ῥ|ὴ|σατ | (negative past tense)

Note that when used with pronominal subjects, the optative and conditional conjugation encompass the subject pronoun, leading to a portmanteau tag like AOPT_PPERS (or A_PPER in the coarse grained tagset):

(4) ἐψ̣/AOPT_PPERS ἁστ | (optative and 3rd pers. masc. pronoun)
(5) ἐψ̣/ACOND_PPERS σατ | (conditional and 3rd pers. masc. pronoun)

Fine-Grained Tags
The different individual fine-grained tags cover all distinct conjugation bases, making auxiliaries the largest fine-grained tag group. They are divided as follows:

<table>
<thead>
<tr>
<th>Tag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APST</td>
<td>Auxiliary, past</td>
</tr>
<tr>
<td>ANEGPST</td>
<td>Auxiliary, negated past</td>
</tr>
</tbody>
</table>
3.2 Adverbs (ADV)
Adverbs include indeclinable native Egyptian and Greek lexemes that modify verbs and other phrases as in the following examples.

(6) ταργαλείον ἀνθός εἰσι/ADV ‘I shall glorify him greatly’
(7) πέτω/ADV ‘the one (who is) there’
(8) ἵππος κακος/ADV ‘don’t die badly’

The first part of ‘complex prepositions’ is also tagged as an adverb, as in the following examples:

(9) εβολ/ADV ἃν/PREP ‘from, out of’ (lit. ‘out in’)
(10) εργον/ADV ὢν/PREP ‘in towards’ (lit. ‘inside at’)

This does not apply to etymologically complex one-word prepositions derived e.g. from nouns for body parts (see the tag PREP for details), nor is the initial ε in words such as εβολ separated from the adverb (see segmentation guidelines).

3.3 Articles (ART)
Articles include definite articles, indefinite articles and article-like words such as κε/σε ‘other’. The following examples illustrate some variants:

(11) το/ART ῥα/Ν ‘the man’
(12) τε/ART κληρονομία/Ν ‘the inheritance’
(13) ου/ART νομος/Ν ‘a law’
(14) ἐν/ART ἔννοια/Ν ‘(some) deeds’
(15) κε/ART ποιμεσ/Ν ‘another wicked one’
Note that possessive pronouns like Ṣε are not tagged as articles (see PPOS) and relative articles like τε are segmented to contain a relative converter (see C and CREL).

Articles followed by a noun beginning with τ and consequently spelled ο or φ e.g. Ṣε ‘the way’ are normalized and tokenized as τ and φe before part-of-speech tagging, so that τ etc. can be tagged as an article alone (see segmentation guidelines).

### 3.4 Converters (C)

The class of converters, which is syntactically heterogeneous, is described in Layton (2004: 319-366). It includes four types of converters which have several realizations depending on their syntactic environment. In the coarse tagset, all converters are tagged as C, allowing for lower error rates in automatic tagging (especially by removing the distinction between circumstantial and relative conversions, which can be ambiguous).

The examples below are for the four fine grained classes:

- **CCIRC**: Converter, circumstantial
  - ι, e[α], Ṣε

- **CFOC**: Converter, focalizing (a.k.a. 2nd tenses)
  - ι, e, e, e, τ[α], Ṣε[α]

- **CPRET**: Converter, preterite
  - ι, Ṣε, ṢΕ

- **CREL**: Converter, relative
  - ι, ṢΕ, ι, ṢE[α], ṢΕ[α], ṢΕ

Note that a following conjugation base is segmented separately from the converter (cf. segmentation guidelines), e.g.:

(16) ṢΕ/CREL ι/APST ṢΙΟΥ ‘which he blessed’

The converter includes only ι, while ι is a separate auxiliary base.

### 3.5 Conjunctions (CONJ)

Conjunctions are indeclinable words of Greek and Egyptian origin which link phrases and clauses. No distinction is made between subordinating conjunctions which introduce clauses (‘because’, ‘lest’) and coordinating conjunctions which connect phrases (e.g. ‘and’, ‘or’).

(17) ιΑΟΦ/CONJ ιΕΙΒΕΥ ‘and I became thirsty’

(18) ιΣΟΦ ιΗΟΙΟ ιΣ/CONJ ιΗΠΟΤΕ/CONJ ιΑΕΙΒΕ ‘saying [that:] lest I become thirsty’

In the first example, the coordinating conjunction ιΑΟ ‘and’ appears. Note that it is still tagged as a conjunction even if the first coordinated phrase is missing. In the second example, two consecutive conjunctions appear: ιΣ ‘that, saying’ introduces the direct speech and the Greek origin ιΗΠΟΤΕ ‘lest’ is a conjunction within the direct speech clause. Also note that the word ιΣ, originally derived from ιΟ ‘say’ is not considered a verb in this usage.
3.6 Copulas (COP)

Copulas are markers in so-called nominal sentences which express predications of the sort A is B. The copula forms are ὑ/τε/ὑ. The tag COP is given also to copulas following a verbal clause for focalizing emphasis (i.e. ‘it is the case that…’), as illustrated below.

(19) οὐσαειν ὑ/COP ‘he is a doctor’
(20) οὐετῶβε ὑπὰχοειν ὑ/COP ‘(it is that) he prayed to God’

In the latter example, it is less obvious that ὑ is the copula, as its predicate is formally a clause and the form never changes its gender or number (i.e. as τε/ὑ; this is also referred to as ‘invariable ὑ’). Though the English translation cannot convey the presence of the copula adequately, these types of cases are still tagged as COP (see Layton 2004:223).

3.7 Existentials (EXIST)

Existentials include the unique lexemes οὐ and μὴ in both pure existential and possessive forms, positive and negative, illustrated in the following examples.

(21) οὐ/EXIST ὑ χεινε μὴνοκ ‘there is one who is like you’
(22) μὴ/EXIST ρύβαλ εὐχοει επεμχοει ‘there is no servant who is above his master’

The same tag is also used for the indefinite durative present and the fixed phrase οὐ ΣΩΜ ‘be able’ literally ‘there is power’.

(23) οὐτὰ/EXIST οὐχ εὐθεία μὴνοκ ἡμὲνωτ αὐράγαμ ‘we have Abraham our father’, lit. ‘exists to us … of Abraham…’
(24) μὴ/EXIST ΣΩΜ ητε/τεὑράφων βολ εβολ ‘scripture cannot be broken’

Note that the possessor pronoun is segmented apart from οὐτὰ and tagged as a pronoun, and the accompanying μὴνοκ is an adverb.

3.8 Foreign Material (FM)

Foreign material includes text that is lexically and syntactically from a foreign language. It is distinct from loan words. Loan words are lexical entries that originate in another language (e.g., Greek, Latin) but are used in Coptic with Coptic syntax. Foreign material consists of words, especially multiword expressions, with foreign syntax. The writer has momentarily switched languages rather than embedded a loan word into a Coptic construction.

(25) οὐ παρὰ τοῦτο/FM οὐ έβολ ἄν ρηπτομα τε ‘it is therefore not part of the body’
3.9 Future Marker (FUT)
The future marker ṭⲧⲁ, derived from the verb ‘go’ is not considered an independent verb form when introducing a second verb and marking future tense. The following example illustrates the construction.

\[(26) \quad † \text{ⲧⲧⲁ/FUT ḫⲧⲡⲧⲡⲐⲡ} \quad \text{‘I will kill you’}\]

In rare cases, forms other than ṭⲧⲁ can be considered for the future marker, e.g. Ⲁ in:

\[(27) \quad \text{ⲧⲥⲣ/CPRET_PERS} \text{ⲧ/FUT} \text{ⲧⲟⲧⲃⲉⲕ} \quad \text{‘you would despise’ (2nd pers. fem.)}\]

Contractions of multiple Ⲁ are usually restored in the normalization, so that a diplomatic sequence like ⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧⲧ McD
3.12 Negations (NEG)
The tag NEG is used for independent negative items that are not part of an auxiliary base. The following lexemes are given the tag NEG: ⲙⲧ, Ⲗⲧ and Ⲩⲣⲓ (negative imperative marker). The first two can occur in the same sentence, in which case one NEG tag is used for each. The third negates infinitives and is tokenized separately from the verb and surrounding auxiliaries. The fourth is also a separate token and is not considered a verb form or part of the verb Ⲟⲣⲓ (this also applies to its lemmatization as an independent item, see lemmatization guidelines)

(35) ⲉⲧ/NEG ⲡⲟⲧⲱⲝⲛⲟⲩⲧⲉⲧⲔ Ⲩⲟⲧ/NEG ‘he will not inherit you’
(36) Ⲡⲧⲉⲩⲧⲁⲧ/NEG ⲧⲟⲧⲱⲝⲛ/NEG ‘if they do not listen’
(37) Ⲩⲣⲓ/NEG Ⲩⲟⲧ ⲡⲓⲧⲓⲧⲓⲧⲓ ‘don’t die badly!’

3.13 Numerals (NUM)
The tag NUM is given to numerals and numerical constituents of complex numerals, as well as suffixed numerals as in the last example below.

(38) Ⲩⲧⲧⲕⲧ ⲧⲧⲏ Ⲩⲟⲧ/NUM ‘five (loaves) of bread’
(39) ⲧⲧⲕⲧⲧⲧⲧ ⲧⲧⲧⲧ ⲧⲧⲧⲧ ⲧⲧⲧⲧ/NUM ‘twenty-four’
(40) Ⲩⲟⲧⲧⲧ ⲧⲧⲧⲧ ⲧⲧⲧⲧ/NUM ‘two times, twice’

Note that the indefinite article Ⲩⲧ ‘a, one’ preceding a noun is tagged as ART, not NUM. Letters being used as numbers are considered NUM (including an alpha preceding a noun for the quantity ‘one’)

3.14 Demonstrative pronouns (PDEM)
The demonstrative pronouns, both attributive to the noun and substituting for a noun are tagged as PDEM.

(41) Ⲩⲧⲧⲃⲧ/NUM Ⲩⲧ/NUM ‘in this way’
(42) Ⲩⲧⲧⲧⲧⲧ/NUM Ⲩⲧ/NUM ‘this is the way’

3.15 Interrogative pronouns (PINT)
This tag is used for the interrogative pronouns Ⲩⲧⲧⲧ ‘what’, Ⲩⲧⲧ ‘who’, ⲧⲧⲧⲧ ‘where’, Ⲩⲧⲧ ‘which’, ⲧⲧⲧⲧ ‘how much’. This is also true when they are used in complex phrases, as in the examples below.

(43) ⲧⲧⲧⲧⲧⲧⲧⲧ ⲧⲧⲧⲧⲧⲧⲧ ⲧⲧⲧⲧⲧⲧⲧ/NUM ‘what for, why?’
(44) Ⲩⲧⲧⲧⲧⲧ Ⲩⲧⲧⲧⲧ Ⲩⲧⲧⲧⲧ ⲧⲧⲧⲧ/NUM ‘where to?’

Note that the item Ⲩⲧⲧ is tagged PINT even when used after a noun to mean ‘some, any’. 
3.16 Personal pronouns (PPER[*])

Personal pronouns generally receive the tag PPER, with three subtypes in the fine-grained subset for subject pronouns (PPERS), object pronouns (PPERO) and independent pronouns (PPERI).

(45) ⲁⲃⲃⲕ/PPERS ⲑⲧⲓ ⲉⲣⲟ ⲁ/PPERO ‘he heard you’
(46) ⲉⲧⲡⲺⲧ/PPERO ‘for her’

Note that ‘object’ pronouns include objects of prepositions and all suffixed pronouns except the subject markers of verboids of the type [ⲧⲛⲟⲩ]ⲗ, [ⲧⲧⲧ]ⲗ etc., which are tagged as PPERS.

(47) ⲧⲧⲧⲧ/PPERS ‘he said’

The independent pronouns are reserved for emphatic uses and nominal sentences, including nominal sentence subject forms like ⲧⲧⲧ Ⲣ ‘I’ and the full forms of the type ⲧⲧⲧ Ⲣ ‘I’.

(48) ⲧⲧⲧⲧ/PPERS ⲧⲧⲧⲧ/PPERO ⲧⲧⲧⲧ/PPERI ⲧⲧⲧ ⲧⲧⲧ ⲧⲧⲧ/PPERI ⲧⲧⲧ ‘I, as for me, I am his servant’

Also note that possessive pronouns like ⲧⲧⲧ ‘his’ are not segmented and receive a separate tag, PPOS.

3.17 Possessive pronouns (PPOS)

Much like demonstratives, all possessive pronouns, both attributive and standing in for a noun are tagged as PPOS. The personal suffix at the end of the pronoun is not separated, rather the entire forms, including ⲧⲧⲧ ‘his’, ⲧⲧⲧ ‘my’ and ‘the one that belongs to’, ⲧⲧⲧ ‘your (fem.)’, ⲧⲧⲧ ‘mine’ etc. The following example illustrates these different types of possessives:

(49) ⲧⲧⲧ ⲧⲧⲧ/PPOS ⲧⲧⲧ ⲧⲧⲧ/PPOS ⲧⲧⲧ/PPOS ⲧⲧⲧ ‘the one of my brother is mine’

This tag only applys to prefixal, article-like possessives. Suffix possessives, such as ⲧⲧⲧ ⲧⲧⲧ ‘his foot’ are not tagged PPOS, but rather PPERO.

3.18 Prepositions (PREP)

This tag is used for all prepositions in both independent, prenominal states and presuffixal forms (which are tokenized apart from following suffixes). Note that prepositions that are historically derived from univerbized phrases but are now unsegmentable are tagged as one preposition, but complex prepositions involving a separable adverb are given two tags, ADV and PREP (cf. the tag ADV). Additionally, the
*nota relationis* and accusative marker ι/ιο is regarded as a preposition. The following examples illustrate these principles.

(50) ἐτέβα/PREP ογ ‘for what? why?’
(51) ἐβον/ADV ἐι/PREP ‘from, out of’ (lit. ‘out in’)
(52) ἐκι̣/PREP ‘upon, on account of’ (from ‘to head of’)

Also note that 2nd pers. sg. fem. objects often lead to portmanteau tags, e.g.:

(53) Ἰι̣/PREP_PPERO ‘you (2nd pers. sg. fem. accusative)’

If in doubt as to whether a lexicalized combination is considered a single preposition, please refer to the formatted CMCL lexicon supplied with the project’s tokenization module. This lexicon will be updated with future versions of the guidelines to accommodate dubious cases as they arise.

3.19 Particles (PTC)
The class of particles contains all indeclinable words that do not belong to one of the other classes, most notably and frequently the apposition marker ἐστι ‘that is…’ and a large number of, mostly Greek origin, sentence modifying particles that tend to appear in the second, Wackernagel position as they do in Greek as well (e.g. ἄε, ἐπ).

3.20 Punctuation (PUNCT)
All punctuation marks, including periods at any height in the line, commas (including punctuation added in editions when annotating edited texts) or even question marks, colons etc. if they are used, are all given the uniform tag PUNCT. If decorations are tokenized (tildes, clusters of dots etc.), they may also be tagged as PUNCT, though refer to the tokenization guidelines for recommendations on normalizing text before tagging.

3.21 Unknown, damaged and lost items (UNKNOWN)
The tag UNKNOWN is given to fragmentary word forms damaged or missing beyond the ability to reach a reliable part-of-speech assignment. It is understood in the case of larger lacunae that the string used to encode the visible part of a word may in fact contain several words. In cases where it is clear where word divisions occur, multiple tokens with corresponding UNKNOWN tags are given.

(54) ἐ[...]UNKNOWN ‘?’
(55) ἐ[...]UNKNOWN ἐ[...]UNKNOWN ‘?’

Generally UNKNOWN tags are given even if the range of possibility is limited, i.e. even if we are certain a damaged morpheme is either an article or a possessive pronoun, an uncertain case is usually tagged as UNKNOWN.
3.22 Verbs (V[*])

The coarse tag V is given to all lexical verb forms that are not conjugation bases, also not including verboids, which receive a separate tag even in the coarse tagset due to their distinct syntax (see the tag VBD). In the fine-grained tagset, normal verb forms (V) are distinguished from stative verb forms (VSTAT) and imperatives (VIMP) as shown in the examples below. Note that verbal infinitives in the durative present are still tagged as verbs, although they are historically nominalized in this position, whereas nominalized infinitives following an article are understood as nouns, as in the last example. Verbs are tagged as VIMP only when they appear in the specific imperative form.

(56) ⲉⲟⲝⲇⲏⲡⲧⲙⲁⲧ/ⲏⲣⲟ ‘he heard you’
(57) ⲉⲟⲥⲗⲉ/ⲏⲣⲟ ‘I am thirsty’
(58) ⲉⲃⲓ/ⲏⲣⲟ ⲁ ‘say it!’
(59) ⲉⲃⲕ ⲙⲟⲩⲛⲃⲧⲧⲢ/ⲕⲧ ⲡⲩⲧⲉ ‘in the knowledge of God, the knowing of God’

Also note that in rare cases, object pronouns that are realized as zero will lead to portmanteau tags, e.g.:

(60) Ⲙⲧⲉ Ⲙⲧⲉ/ⲧⲉⲧⲛ Ⲝⲧ/ⲧⲉ ‘you bring me’

Since ⲧⲦⲉ as the presuffixal form of ⲧⲦⲉ ends in ⲧ, the object pronoun ⲧ ‘me’ is subsequently dropped. However the portmanteau tag reflects the presence of a grammatical object.

For compound verbs (see §180 in Layton), the entire compound is considered “a single unit in boundness, syntax, and meaning.” Therefore, the entire compound is tagged V. The components of the compound may be annotated further on a morph level annotation. (See Transcription guidelines for more information on bound groups, morphemes, and word segmentation.) Common examples include compound verbs formed with Ⲡ-, Ⲝ-, and ⲧ-.

(61) Ⲣⲧ/ⲧⲉⲧⲛ Ⲣⲧⲉ/ⲧⲉⲧⲛ ⲧⲩⲧⲉ ‘to bring me’
(62) Ⲣⲧ/ⲧⲉⲧⲛ ⲧⲩⲧⲉ/ⲧⲉⲧⲛ ⲧⲩⲧⲉ ‘to think’

The basic criterion for identifying compound verbs is the absence of an article: ⲧⲩⲧⲉ ‘to sin’ is considered as single, compound verb (which can still be analyzed morphologically into two units ⲧⲩⲧⲉ, perhaps like English ‘sin-ify’, if there were such a word). However ⲧⲧⲉ ⲧⲩⲧⲉ ‘to think’ looks exactly like any verb + definite noun phrase combination, and is therefore tagged as three units despite being a common lexicalized combination: it comprises a verb, an article and a noun.

Exceptions: Some object nouns cannot appear as definite, or are made definite other than by an article. These include objects with ⲧⲧⲉ ‘some, any’, ⲧⲧⲉ ‘something’ and ⲧⲧⲉ
‘some (ones)’, number words, as well as verbal objects with a suffixal possessive pronouns, such as κЄ ράτ η ‘set one’s foot’ (the foot is definite). Even though they may appear next to a verb without an article, these are tokenized and tagged apart from the verb (for possessed objects, the possessive is its own token, tagged PPERO, not PPOS).

3.23 Verboids (VBD)

The category VBD is given to a small class of suffixally inflected predicates described in Layton (2004: 297-304), including the common πεξε/-πεξα ‘say’, νανογ ‘be good’ etc., but not including possessive existentials of the type ογιτε- (see the tag EXIST). The personal suffix following a VBD is tagged as its subject, i.e. PPERS (or simply PPER in the coarse tagset).

(63) πεξα/VBD υ/PPERS ‘he said’
(64) νανογ/VBD σ/PPERS ‘she is good’

For the form μωαξ note that two analyses exist. When it is declinable and literally means ‘X does not know’ (also prenominal μωαξ-), then it is VBD. When it is the lexicalized adverb form μωακ meaning ‘maybe’ (etymologically from ‘you never know’), it is a single unit, tagged ADV. Note that the latter form does not agree with the addressee if they are not masculine singular. Contrast the following examples from Layton (2004:303):

(65) μωαξ/VBD υ/PPERS ηπιαγ ετημακω μωαξ ηπικοκιος ‘he does not know when he will leave the world’
(66) μωακ/ADV διασω ρατειντου ‘maybe I’ll stay with you’

In the latter example, the addressee is plural (τιντυν), but the form remains μωακ, indicating that it is an unanalyzed adverb.

4. References

