SCRIPTORIUM Part-of-Speech Tagsets for Sahidic Coptic

Amir Zeldes[1] & Caroline T. Schroeder[2]

1. Georgetown University
2. University of the Pacific

Version: *1.2.1_2019.06.05*

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/).

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/).

2. Tagsets

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

Tag	Name	Examples
A[*]	Auxiliary tripartite base	ⲁ[ⲩ], ⲙⲉ[ⲩ], ⲧⲣⲉ[ⲩ], ...
ADV	Adverb	ⲉⲃⲟⲗ, ⲟⲛ, ⲡⲱⲥ
ART	Article	ⲡ(ⲉ), ⲧ(ⲉ), ⲛ(ⲉ), ϩⲉⲛ, ⲕⲉ
C[*]	Converter	ⲉ, ⲉⲧⲉ, ⲛⲉ, ...
CONJ	Conjunction	ⲁⲩⲱ, ⲭⲉ, ⲏ, ⲙⲏ, ⲉⲓⲧⲉ, ...
COP	Copula	ⲡⲉ/ⲧⲉ/ⲛⲉ
EXIST	Existential/possessive	ⲟⲩⲛ/ⲙⲛ
FM	Foreign material	ⲡⲁⲣⲁ ⲧⲟⲩⲧⲟ
FUT	Future	ⲛⲁ
IMOD	Inflected modifier	ⲧⲏⲣ[ⲩ], ϩⲱⲱ[ⲧ], …
N[*]	Noun	ⲁⲑⲏⲧ, ⲣⲱⲙⲉ, ⲁⲣⲭⲏ, ...
NEG	Negation	ⲛ, ⲁⲛ, ⲧⲙ[ⲥⲱⲧⲙ]
NUM	Numeral	ⲟⲩⲁ, ⲥⲛⲁⲩ, …
PDEM	Pronoun, demonstrative	ⲡⲉⲓ/ⲡⲁⲓ, ⲧⲉⲓ/ⲧⲁⲓ, ⲛⲉⲓ/ⲛⲁⲓ
PINT	Pronoun, interrogative	ⲟⲩ, ⲛⲓⲙ
PPER[*]	Pronoun, personal	ⲩ,ⲥ,ⲓ,ⲧ,ⲛ,ⲁⲛⲟⲕ,ⲁⲛⲛ̄,...
PPOS	Pronoun, possessive	ⲡⲉⲩ,ⲧⲉⲧⲛ̄,ⲡⲟⲩ,ⲡⲁ,ⲡⲱⲓ,...
PREP	Preposition	ⲉⲧⲃⲉ, ϩⲛ̄, ⲛ, ⲙ̄ⲙⲟ[ⲩ], …
PTC	Particle	ⲇⲉ, ⲛ̄ⲑⲓ, …
PUNCT	Punctuation	. , · …
UNKNOWN	Unknown morph, lacuna	ⲃ_ _ _, _ _ⲟⲥ, _ _ _, …
V[*]	Verb	ⲥⲱⲧⲙ, ⲥⲱⲧⲡ, ⲥⲟⲧⲡ, ⲉⲓⲣⲉ, ⲟ, ⲁⲣⲓ, ...
VBD	Verboid	ⲛⲁⲛⲟⲩ[ⲩ], ⲡⲉⲭⲁ[ⲩ], ⲡⲉⲭⲉ,...

2.2 Fine-Grained Tagset

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

AAOR	*AJUS*	*ANY*
ACAUS	*ALIM*	*AOPT*
ACOND	*ANEGAOR*	*APREC*
ACONJ	*ANEGJUS*	*APST*
ADV	*ANEGOPT*	ART
AFUTCONJ	*ANEGPST*	*CCIRC*

CFOC	NEG	PREP
CPRET	*NPROP*	PTC
CONJ	NUM	PUNCT
COP	PDEM	UNKNOWN
CREL	PINT	V
EXIST	*PPERI*	VBD
FUT	*PPERO*	*VIMP*
IMOD	*PPERS*	*VSTAT*
N	PPOS	

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 2nd person singular feminine forms). Portmanteau tags currently supported by the SCRIPTORIUM tools are:

tag	example	notes
AOPT_PPERS	ϵⲩϵ(ⲥⲱⲧⲙ)	Personal pron. within optative ϵ_ϵ. Note that ϵⲣϵ(ⲥⲱⲧⲙ) for 2nd pers. sg. fem. is also AOPT_PPERS, but nominal ϵⲣϵ(ⲡⲣⲱⲙϵ ⲥⲱⲧⲙ) is only AOPT.
ACOND_PPERS	ϵⲩⲫⲁⲛ(ⲥⲱⲧⲙ)	Personal pron. within conditional ϵ_ⲫⲁⲛ. Note that ϵⲣⲫⲁⲛ(ⲥⲱⲧⲙ) for 2nd pers. sg. fem. is also ACOND_PPERS, but nominal ϵⲣⲫⲁⲛ(ⲡⲣⲱⲙϵ ⲥⲱⲧⲙ) is only ACOND.
ACONJ_PPERS	ⲧⲁ(ⲥⲱⲧⲙ)	Truncated conjunctive 1st person (instead of ⲛⲧⲁⲥⲱⲧⲙ)
ANEGPST_PPERS	ⲙⲡϵ(ⲥⲱⲧⲙ)	Fused negative past 2nd pers. sg. fem. form.
APST_PPERS	ⲁⲣ(ⲥⲱⲧⲙ)	Fused positive past 2nd pers. sg. fem. form.
CCIRC_PPERS	ϵⲣϵ(ⲥⲱⲧⲙ)	Fused circumstantial 2nd pers. sg. fem. form.
CFOC_PPERS	ϵⲣϵ(ⲥⲱⲧⲙ)	Fused focalized 2nd pers. sg. fem. form.
CPRET_PPERS	ⲛϵⲣϵ(ⲥⲱⲧⲙ)	Fused preterit 2nd pers. sg. fem. form.
CREL_PPERS	ϵⲧϵⲣϵ(ⲥⲱⲧⲙ)	Fused relative 2nd pers. sg. fem. form.
IMOD_PPERO	ⲙⲛⲓⲛⲙⲙⲟ	The 2nd pers. sg. fem. form of 'yourself' (not to be confused with ⲙⲛⲓⲛⲙⲙⲟ(ⲩ) etc.)
PREP_PPERO	ϵⲣⲟ	Any preposition where a 2nd pers. sg. fem. is realized as zero (not to be confused with ϵⲣⲟ(ⲩ) etc.)
V_PPERO	(ⲩ)ⲛⲧ	Verb forms with a fused 1st pers. sg. object, e.g. ⲛⲧ 'bring me' from ϵⲓⲛϵ 'bring', where the presuffixal form ⲛⲧ= is merged with the 1st pers. object marker -ⲧ)

Note that in all cases, coarse grained tags can be substituted for fine grained ones, e.g. CCIRC_PPERS and CFOC_PPERS 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 ⲉⲃⲟⲗ is tagged as an ADV, although in the sequence ⲟⲩ|ⲉⲃⲟⲗ ⲉⲙ|ⲡ|ⲥⲱⲙⲁ ‘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 ⲉⲃⲟⲗ 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 ⲡ|ⲧⲱⲉⲙ ‘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) ⲁ|ⲩ|ⲥⲱⲧⲙ̄ (3rd person masculine past tense)
(2) ⲁⲣⲉ|ⲥⲱⲧⲙ̄ (2nd person feminine past tense, with zero subject)
(3) ⲙ̄ⲡ|ⲓ|ⲥⲱⲧⲙ̄ (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) ⲉⲩⲉ/AOPT_PPERS ⲥⲱⲧⲙ̄ (optative and 3rd pers. masc. pronoun)
(5) ⲉⲩⲱⲁⲛ/ACOND_PPERS ⲥⲱⲧⲙ̄ (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:

APST	Auxiliary, past	ⲁ
ANEGPST	Auxiliary, negated past	ⲙ̄ⲡ(ⲉ)

ANY	Auxiliary, ‘not yet’	ⲙ̄ⲡⲁⲧ(ⲉ)
AAOR	Auxiliary, aorist	ⲱⲁ, ⲱⲁⲣⲉ
ANEGAOR	Auxiliary, negated aorist	ⲙⲉ(ⲣⲉ)
AOPT	Auxiliary, optative	ⲉ[ⲩ]ⲉ, ⲉⲣⲉ
ANEGOPT	Auxiliary, negated optative	ⲛ̄ⲛⲉ
AJUS	Auxiliary, jussive	ⲙⲁⲣ(ⲉ)
ANEGJUS	Auxiliary, negated jussive	ⲙ̄ⲡⲣ̄ⲧⲣⲉ
APREC	Auxiliary, precursive (‘after’)	ⲛ̄ⲧⲉⲣ(ⲉ)
ACOND	Auxiliary, conditional	ⲉ[ⲩ]ⲱⲁⲛ, ⲉⲣⲱⲁⲛ
ALIM	Auxiliary, limitative (‘until’)	ⲱⲁⲛⲧ(ⲉ)
ACONJ	Auxiliary, conjunctive	ⲛ̄(ⲧⲉ)
AFUTCONJ	Auxiliary, future conjunctive	ⲧⲁⲣ(ⲉ)
ACAUS	Auxiliary, causative	ⲧⲣⲉ

Note that the irregular negation ⲙⲉ in ⲙⲉ-ⲱⲱⲉ ‘it is not appropriate’ is also tagged NEG and not as ANEGAOR.

3.2 Adverbs (ADV)

Adverbs include indeclinable native Egyptian and Greek lexemes that modify verbs and other phrases as in the following examples.

(6) ⲧⲁⲁⲩϫⲁⲛⲉ ⲙ̄ⲙⲟⲩ ⲉⲙⲁⲧⲉ/ADV	‘I shall glorify him greatly’
(7) ⲡⲉⲧ|ⲙ̄ⲙⲁⲩ/ADV	‘the one (who is) there’
(8) ⲙ̄ⲡⲣⲙⲟⲩ ⲕⲁⲕⲱⲥ/ADV	‘don’t die badly’

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

(9)	ⲉⲃⲟⲗ/ADV ϩⲛ̄/PREP	‘from, out of’ (lit. ‘out in’)
(10)	ⲉϩⲟⲩⲛ/ADV ϩⲓ/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 ⲉ in words such as ⲉⲃⲟⲗ separated from the adverb (see segmentation guidelines).

3.3 Articles (ART)

Articles include definite articles, indefinite articles and article-like words such as ⲕⲉ/ⲟⲩⲉ ‘other’. The following examples illustrate some variants:

(11)	ⲡ/ART ⲣⲱⲙⲉ/N	‘the man’
(12)	ⲧⲉ/ART ⲕⲗⲏⲣⲟⲛⲟⲙⲓⲁ/N	‘the inheritence’
(13)	ⲟⲩ/ART ⲛⲟⲙⲟⲥ/N	‘a law’

(14) ϩⲉⲛ/ART ϩⲃⲏⲩⲉ/N ‘(some) deeds’
(15) ⲕⲉ/ART ⲡⲟⲛⲏⲣⲟⲥ/N ‘another wicked one’

Note that possessive pronouns like ⲡⲉⲩ are not tagged as articles (see PPOS) and relative articles like ⲡ|ⲉⲧ are segmented to contain a relative converter (see C and CREL).

Articles followed by a noun beginning with ϩ and consequently spelled ⲑ or ⲫ e.g. ⲑⲉ ‘the way’ are normalized and tokenized as ⲧ and ϩⲉ before part-of-speech tagging, so that ⲧ 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	ⲉ, ⲉ[ⲁ], ⲉⲣⲉ
CFOC	Converter, focalizing (a.k.a. 2nd tenses)	ⲉ, ⲉⲣⲉ, ⲉⲧⲉ, ⲛ̄ⲧ[ⲁ], ⲉⲛⲧ[ⲁ]
CPRET	Converter, preterite	ⲛⲉ, ⲛⲉⲣⲉ
CREL	Converter, relative	ⲉⲧⲉ, ⲉⲧ, ⲛ̄ⲧ[ⲁ], ⲉⲛⲧ[ⲁ], ⲉⲧⲉⲣⲉ

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

(16) ⲛⲧ/CREL ⲁ/APST ⲫ|ⲥⲙⲟⲩ ‘which he blessed’

The converter includes only ⲛⲧ, while ⲁ is a separate auxiliary base. The fused second person singular feminine form preceding a future marker, ⲛⲉⲣⲁ (‘you(F) would’), is tokenized into norms and tagged as follows: ⲛⲉⲣ/CPRET_PPERS ⲁ/FUT. Note that the normalized form of the future marker in this case remains ⲁ, but the lemma is ⲛⲁ.

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) ⲁⲩⲱ/CONJ ⲁⲓⲉⲓⲃⲉⲩ ‘and I became thirsty’
(18) ⲉⲓⲭⲱ ⲙ̄ⲙⲟⲥ ⲭⲉ/CONJ ⲙⲏⲡⲟⲧⲉ/CONJ ⲧⲁⲉⲓⲃⲉ ‘saying [that:] lest I become thirsty’

In the first example, the coordinating conjunction ⲁⲩⲱ ‘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: ϫⲉ ‘that, saying’ introduces the direct speech and the Greek origin ⲙⲏⲡⲟⲧⲉ ‘lest’ is a conjunction within the direct speech clause. Also note that the word ϫⲉ, originally derived from ϫⲱ ‘say’ is not considered a verb in this usage. The word ⲕⲁⲓ in the phrase ⲕⲁⲓ ⲅⲁⲣ is tagged as CONJ as well (but ⲅⲁⲣ is PTC).

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 ⲡⲉ/ⲧⲉ/ⲛⲉ. 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) ⲟⲩⲥⲁⲉⲓⲛ ⲡⲉ/COP ‘he is a doctor’

(20) ⲛⲉⲩⲧⲱⲃϩ ⲙⲡϫⲟⲉⲓⲥ ⲡⲉ/COP ‘(it is that) he prayed to God’

In the latter example, it is less obvious that ⲡⲉ is the copula, as its predicate is formally a clause and the form never changes its gender or number (i.e. as ⲧⲉ/ⲛⲉ; this is also referred to as ‘invariable ⲡⲉ’). 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 ⲟⲩⲛ̄ and ⲙⲛ̄ in both pure existential and possessive forms, positive and negative, illustrated in the following examples.

(21) ⲟⲩⲛ̄/EXIST ⲟⲩⲁ ⲉⲩⲉⲓⲛⲉ ⲙⲙⲟⲕ ‘there is one who is like you’

(22) ⲙⲛ̄/EXIST ϩⲙ̄ϩⲁⲗ ⲉⲩϫⲟⲥⲉ ⲉⲡⲉⲩϫⲟⲉⲓⲥ ‘there is no servant who is above his master’

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

(23) ⲟⲩⲛⲧⲁ/EXIST ⲛ/PPERO ⲙ̄ⲙⲁⲩ/ADV ⲙⲡⲉⲛⲉⲓⲱⲧ ⲁⲃⲣⲁϩⲁⲙ
‘we have Abraham our father’, lit. ‘exists to us … of Abraham…’

(24) ⲙ̄ⲙⲛ̄/EXIST ϭⲟⲙ ⲛⲧⲉ|ⲧⲉ|ⲅⲣⲁⲫⲏ ⲃⲱⲗ ⲉⲃⲟⲗ ‘scripture cannot be broken’

Note that the possessor pronoun is segmented apart from ⲟⲩⲛⲧⲁ and tagged as a pronoun, and the accompanying ⲙ̄ⲙⲁⲩ 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) ⲟⲩ ⲡⲁⲣⲁ ⲧⲟⲩⲧⲟ/FM ⲛⲟⲩ ⲉⲃⲟⲗ ⲁⲛ ⲥⲙⲡⲥⲱⲙⲁ ⲧⲉ ‘it is therefore not part of the body’

3.9 Future Marker (FUT)

The future marker ⲛⲁ, 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) ⲧ ⲛⲁ/FUT ⲥⲟⲧⲃⲉⲕ ‘I will kill you’

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

(27) ⲛⲉⲣ/CPRET_PPERS ⲁ/FUT ⲥⲱⲫ ‘you would despise’ (2nd pers. fem.)

Contractions of multiple ⲛ are usually restored in the normalization, so that a diplomatic sequence like ⲧⲉⲧⲛⲁⲣⲡⲓⲙⲉⲉⲩⲉ ‘you will think’ are usually normalized and only then tagged as follows:

(28) ⲧⲉⲧⲛ/PPERS ⲛⲁ/FUT ⲣ/V

3.10 Inflected modifiers (IMOD)

Inflected modifiers are a somewhat heterogeneous class of suffixally inflecting non-verboids, including the quantifier ⲧⲏⲣ= ‘all of’, the focus particle ⲟⲩⲁⲁ(ⲧ)= ‘only’ and the reflexive ⲙ̄ⲙⲓⲛⲙⲙⲟ̄= ‘oneself’ (see Layton 2004: 118-123 and contrast the tag VBD). The suffix itself is tokenized apart and tagged as PPERO. These items are tokenized apart even within larger phrases, as in the second examples below.

(29) ⲁⲛⲟⲕ ⲥⲱⲫ/IMOD ⲧ/PPERO ‘I, as for me / me too’
(30) ⲉ ⲡ ⲧⲏⲣ/IMOD ⲫ ‘in all of it, at all, wholly’

If the suffix is a 2nd pers. sg. fem. realized as zero, a portmanteau tag is assigned:

(31) ⲙⲙⲓⲛⲙⲙⲟ/IMOD_PPERO ‘yourself (2nd pers. sg. fem.)’

3.11 Nouns (N)

The tag N is used for all nouns, common and proper, though the fine-grained tagset offers the specific tag NPROP for proper nouns.

(32) ⲡⲉⲛ ⲉⲓⲱⲧ/N ‘our father’
(33) ⲁⲛⲧⲱⲛⲓⲟⲥ/NPROP ‘Antonius’

Note that verbal infinitives in the durative patterns and elsewhere, though technically and etymologically nominal in nature, are nevertheless tagged as verbs in order to facilitate the retrieval of verbal lexemes across constructions.

(34) ϯ πιστεγε/V επνογτε 'I trust in God'

Demonyms such as ⲣⲱⲙⲁⲓⲟⲥ 'Roman' are tagged as regular nouns (since they do not refer to a specific, named Roman).

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: ⲛ, ⲁⲛ, ⲧⲙ̄ and ⲙⲡⲣ (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 ⲉⲓⲣⲉ (this also applies to its lemmatization as an independent item, see lemmatization guidelines)

(35) ⲛ̄/NEG ⲭⲛⲁⲕⲗⲏⲣⲟⲛⲟⲙⲉⲓ ⲙ̄ⲙⲟⲕ ⲁⲛ/NEG 'he will not inherit you'
(36) ⲉⲩⲱⲁⲛ ⲧⲙ̄/NEG ⲥⲱⲧⲙ̄ 'if they do not listen'
(37) ⲙⲡⲣ/NEG ⲙⲟⲩ ⲕⲁⲕⲱⲥ 'don't die badly!'

Note that the irregular negation ⲙⲉ in ⲙⲉ-ⲱⲱⲉ 'it is not appropriate' is also tagged NEG and not as ANEGAOR.

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) ϯⲟⲩ/NUM ⲛ̄ⲟⲉⲓⲕ 'five (loaves) of bread'
(39) ⲭⲟⲩⲧ/NUM ⲇⲫⲧⲉ/NUM 'twenty-four'
(40) ⲛ̄|ⲥⲉⲡ ⲥⲛⲁⲩ/NUM 'two times, twice'

Note that the indefinite article ⲟⲩ '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) ⲛ ⲧⲉⲓ/PDEM ⲣⲉ 'in this way'
(42) ⲧⲁⲓ/PDEM ⲧⲉ ⲧ ⲣⲉ 'this is the way'

3.15 Interrogative pronouns (PINT)

This tag is used for the interrogative pronouns ⲟⲩ ‘what’, ⲛⲓⲙ ‘who’, ⲧⲱⲛ ‘where’, ⲁⲱ ‘which’, ⲟⲩⲏⲣ ‘how much’. This is also true when they are used in complex phrases, as in the examples below.

(43) ⲉⲧⲃⲉ/PREP ⲟⲩ/PINT ‘what for, why?’

(44) ⲉ/PREP ⲧⲱⲛ/PINT ‘where to?’

Note that the item ⲛⲓⲙ 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) ⲁ ⲩ/PPERS ⲥⲱⲧⲙ̄ ⲉⲣⲟ ⲕ/PPERO ‘he heard you’

(46) ⲉⲧⲃⲏⲏⲧ ⲥ/PPERO ‘for her’

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

(47) ⲡⲉⲭⲁ ⲩ/PPERS ‘he said’

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

(48) ⲁⲛⲟⲕ/PPERI ϩⲱⲱ ⲧ/PPERO ⲁⲛ̄/PPERI ⲡⲉⲩ ϩⲙ̄ϩⲁⲗ
‘I, as for me, I am his servant’

Also note that possessive pronouns like ⲡⲉⲩ ‘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 ⲡⲉⲩ ‘his’, ⲡⲁ ‘my’ and ‘the one that belongs to’, ⲡⲟⲩ ‘your (fem.)’, ⲡⲱⲓ ‘mine’ etc. The following example illustrates these different types of possessives:

(49) ⲧⲁ/PPOS ⲡⲁ/PPOS ⲥⲟⲛ ⲧⲱⲓ/PPOS ⲧⲉ ‘the one of my brother is mine’

This tag only applys to prefixal, article-like possessives. Suffix possessives, such as ⲣⲁⲧ ⲩ ‘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 ⲛ/ⲙ̄ⲙⲟ is regarded as a preposition. The following examples illustrate these principles.

(50)	ⲉⲧⲃⲉ/PREP ⲟⲩ	‘for what? why?’
(51)	ⲉⲃⲟⲗ/ADV ϩⲛ̄/PREP	‘from, out of’ (lit. ‘out in’)
(52)	ⲉϧⲛ̄/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)	ⲙⲙⲟ/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 ⲛⲑⲓ ‘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. ⲇⲉ, ⲅⲁⲣ).

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) єⲙ[...]/UNKNOWN ‘?’
(55) єⲙ[...]/UNKNOWN ⲡⲙ[...]/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) ⲁ ⲫ ⲥⲱⲧⲙ̄/V ⲉⲣⲟ ⲕ ‘he heard you’
(57) ϯ ⲟⲩⲉ/VSTAT ‘I am thirsty’
(58) ⲁ.ϫⲓ/VIMP ⲥ ‘say it!’
(59) ⲉⲙ̄ ⲡ ⲥⲟⲟⲩⲛ/N ⲙ̄ ⲡ ⲛⲟⲩⲧⲉ ‘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) ⲧⲉⲧⲛ/PPERS ⲛⲧ/V_PPERO ‘you bring me’

Since ⲛⲧ= as the presuffixal form of ⲉⲓⲛⲉ 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 compund 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 ϯ-, ⲣ-, and ϫⲓ-.

(61) ⲉⲧ/CREL ⲣⲛⲟⲃⲉ/V
(62) ⲉ/CCIRC ⲕ/PPERS ϯⲥⲃⲱ/V

The basic criterion for identifying compound verbs is the absence of an article: ⲣⲛⲟⲃⲉ ‘to sin’ is considered as single, compound verb (which can still be analyzed morphologically into two units ⲣ+ⲛⲟⲃⲉ, perhaps like English ‘sin-ify’, if there were such a word). However ⲣ ⲡ ⲙⲉⲉⲩⲉ ‘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 ⲛⲓⲙ ‘some, any’, ⲗⲁⲁⲩ ‘something’ and ⲉⲟⲓⲛⲉ ‘some (ones)’, number words, as well as verbal objects with a suffixal possessive pronouns, such as ⲕⲉ ⲣⲁⲧ ⲫ ‘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 ⲡⲉϫⲉ-/ⲡⲉϫⲁ= ‘say’, ⲛⲁⲛⲟⲩ= ‘be good’ etc., but not including possessive existentials of the type ⲟⲩⲛⲧⲉ- (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) ⲡⲉϫⲁ/VBD ⲫ/PPERS ‘he said’

(64) ⲛⲁⲛⲟⲩ/VBD ⲥ/PPERS ‘she/it is good’

For the form ⲙⲉⲫⲁ= note that two analyses exist. When it is declinable and literally means ‘X does not know’ (also prenominal ⲙⲉⲫⲉ-), then it is VBD. When it is the lexicalized adverb form ⲙⲉⲫⲁⲕ 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) ⲙⲉⲫⲁ/VBD ⲫ/PPERS ⲙⲡⲛⲁⲩ ⲉⲧⲫⲛⲁⲕⲱ ⲛⲥⲱⲫ ⲙⲡⲕⲟⲥⲙⲟⲥ
‘he does not know when he will leave the world’

(66) ⲙⲉⲫⲁⲕ/ADV ϯⲛⲁⲟⲱ ⲉⲁⲧⲉⲧⲏⲩⲧⲛ ‘maybe I’ll stay with you’

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

4. References

Layton, Bentley (2004), *A Coptic Grammar*. Second Edition, Revised and Expanded. (Porta linguarum orientalium 20.) Wiesbaden: Harrassowitz.

Schmid, Helmut (1994), Probabilistic part-of-speech tagging using decision trees. *Proceedings of the Conference on New Methods in Language Processing*. Manchester, UK, 44–49.

Shisha-Halevy, Ariel. 1988. *Coptic Grammatical Chrestomathy. A Course for Academic and Private Study*. (Orientalia Lovaniensia Analecta 30.) Leuven: Peeters.