

# Coptic SCRIPTORIUM – Lemmatization Guidelines

Version 1.0.4 / 2017-03-29

Amir Zeldes

## Preamble

The purpose of lemmatization is to facilitate finding variant and inflected forms that are related to the same lexical entry, roughly equivalent to a dictionary entry. However in many cases, it may be unclear what the underlying, uninflected form of a word is: is the lemma of the pronoun ‘me’ defined as ‘I’ (i.e. the nominative form)? Should the lemma of ‘us’ then be ‘we’? Alternatively we could put all personal pronouns under one lemma: then ‘we’, ‘us’, ‘I’, and ‘me’ all belong to the same lemma, but which form should be taken for the common lemma?

There can be many arguments for and against certain practices. In these guidelines we attempt to give a set of instructions for Coptic which is: a. easy to apply consistently and b. useful for searching purposes.

## Guidelines by Part-of-Speech Class

### Articles and copulas

Articles are lemmatized according to the non-assimilated, simple short form of the corresponding masculine singular article (if distinct). This means that the lemma of π, πε, τ, τε, η, ηε and ι (assimilated form of η before a labial consonant) is for all of the above π. For indefinites ογ and γεν there is no special masculine form, but the singular lemma ογ is taken for the plural γεν and also for the variant spelling γ.

Copulas follow a similar rule: the lemma for all three number/gender forms (πε/τε/ηε) is πε.

### Pronouns

#### *Personal pronouns*

Lemmas are mainly helpful where they deliver added value over searching for plain strings. It is therefore useful to give common lemmas for each of the personal forms: first person, second M/F, ... Given that the SCRIPTORIUM part-of-speech guidelines already distinguish subject and object pronouns, it is considerably more useful to group subjects and objects of the same person together, while not distinguishing the different forms (e.g. ⲧ, ⲓ for first person) which can be found using a plain-text search anyway. We therefore annotate the following personal pronouns (SCRIPTORIUM tags in PPER\*, i.e. PPERs,

PPERO, PPERI) with the following lemmas, based on the independent stressed pronoun forms (note that lemmatization is based on normalized forms without supralinear strokes or other diacritics; cf. transcription and normalization guidelines):

Person	Lemma	Pronoun forms
1st sg.	ΔΝΟΚ	ΔΝΟΚ, ΔΝΓ, †, Ι, ΝΤ, Τ, Δ
2nd sg. masc.	ΝΤΟΚ	ΝΤΟΚ, ΝΤΚ, Κ, Γ, ΤΚ
2nd sg. fem.	ΝΤΟ	ΝΤΟ, ΝΤΕ, ΤΕ, ΤΡ, Ρ, Ε
3rd sg. masc.	ΝΤΟϞ	ΝΤΟϞ, Ϟ
3rd sg. fem.	ΝΤΟC	ΝΤΟC, C
1st pl.	ΔΝΟΝ	ΔΝΟΝ, ΔΝ, Ν, ΤΝ, CN
2nd pl.	ΝΤΩΤΝ	ΝΤΩΤΝ, ΝΤΕΤΝ, ΤΝ, ΤΗΥΤΝ
3rd pl.	ΝΤΟΟΥ	ΝΤΟΟΥ, Υ, ΟΥ, CΕ, CΟΥ

The pronoun lemmas alone therefore primarily give access to search by person (1<sup>st</sup>, 2<sup>nd</sup> ...); to cross-reference these with the form, e.g. independent pronoun, cross-reference the POS annotation (in ANNIS: pos="PPERI"). For a specific subform (e.g. ΔΝΓ not ΔΝΟΚ) use the form search norm="ΔΝΓ".

#### *Possessives, interrogatives and demonstratives*

Interrogative pronouns are each equivalent to their own lemma, i.e. οΥ is lemmatized οΥ and ΝΙΜ as ΝΙΜ.

Possessive, and demonstrative pronouns are lemmatized to their own normalized form, but with one modification: non-masculine singular determiners are given the masculine form, i.e. the lemma of πεϞ is πεϞ, the lemma of πα is πα etc., but the lemma of τεC and νεC is also πεC. Similarly, the lemma of πεΙ and νεΙ is πεΙ, and the lemma of παΙ and ταΙ is παΙ. This allows an easier search for all possessives (in ANNIS: pos="PPOS", finds πεΥ, τεΥ, νοΥ ...), all third person plural possessives (lemma="πεΥ", finds πεΥ, τεΥ and νεΥ) and all third person plural possessives of feminine objects (norm="τεΥ"), and similarly for demonstratives.

#### **Adverbs, particles and conjunctions**

Adverbs, particles and conjunctions are always given their own normalized form as a lemma. This includes Greek adverbs in -ωC, which are lemmatized as such, e.g. ζωλωC has the lemma ζωλωC.

#### **Nouns**

Nouns are given their dictionary form as a lemma. For most nouns, singular and plural forms are identical, meaning there is no dilemma. For nouns with irregular plural forms, the singular form is taken as a lemma, e.g. ζωΒ 'deed' is the lemma of both singular ζωΒ and plural ζβηγε, and similarly, possessed forms like τοοτ(Ϟ) are lemmatized to the absolute form, i.e. τωρε. In order to find irregular forms, one can then simply search for

nouns whose lemma is different from the noun form (in ANNIS: lemma != norm). The same rules apply to proper nouns, though these rarely occur in the plural.

For nouns which only occur in the possessed form, if both prenominal and presuffixal forms exist, the prenominal is taken as the lemma, e.g.  $\chi\eta\lambda\alpha\varsigma$  and  $\chi\eta\epsilon-$  ‘(one’s) will’ are lemmatized as  $\chi\eta\epsilon$ . If only a presuffixal form exists, it is taken as the lemma as well, e.g.  $\eta\lambda\iota\alpha\tau\varsigma$  ‘blessed is...’ has the lemma  $\eta\lambda\iota\alpha\tau$ .

Nouns that have related masculine and feminine forms are considered separate lemmas. For instance, the noun  $\omega\eta\rho\epsilon$  ‘son’ is its own lemma, and the separate noun  $\omega\epsilon\epsilon\rho\epsilon$  ‘daughter’ also has a separate lemma (which is  $\omega\epsilon\epsilon\rho\epsilon$ ). Similarly, Greek words in -ος are considered separate from related words in -ον, e.g.  $\pi\omicron\eta\eta\rho\omicron\varsigma$  ‘wicked person’ is its own lemma, and so is the separate  $\pi\omicron\eta\eta\rho\omicron\nu$  ‘wicked deed/thing’ an independent lemma.

### Verbs

Verbs are lemmatized to the form of the absolute infinitive. This means that special prenominal or presuffixal forms are lemmatized to their respective dictionary entries, e.g.  $\sigma\omicron\tau\tau\iota\varsigma$  and  $\sigma\epsilon\tau\tau\iota-$  are lemmatized as  $\sigma\omega\tau\tau\iota$  ‘choose’. The same applies to stative and imperative forms, which are lemmatized to the dictionary entry, e.g.  $\kappa\eta\tau^{\dagger}$  has  $\kappa\omega\tau$  as a lemma and  $\lambda\rho\iota$  has  $\epsilon\iota\pi\epsilon$ . Likewise for prenominal forms,  $\sigma\epsilon\tau\tau\iota$  and  $\rho$  are lemmatized as  $\sigma\omega\tau\tau\iota$  and  $\epsilon\iota\pi\epsilon$ .

Note that auxiliaries are not lemmatized to their etymological verbs, i.e. the lemma of the past tense  $\lambda-$  is not  $\epsilon\iota\pi\epsilon$  but  $\lambda$ . Additionally, the negative imperative marker  $\mu\eta\rho$  is lemmatized as  $\mu\eta\rho$  as well, as it is considered to be a form of negation independent from the verb  $\epsilon\iota\pi\epsilon$ . However, the negative imperative of  $\epsilon\iota\pi\epsilon$  itself,  $\mu\eta\omega\rho$  *IS* lemmatized as  $\epsilon\iota\pi\epsilon$  (since it is a morphological imperative of  $\epsilon\iota\pi\epsilon$  itself, and functions as part of its paradigm with the sense ‘to do’).

For fused verb-object forms like  $\eta\tau$  ‘bring me’, see Portmanteau Tags.

### Prepositions

Prepositions are lemmatized to their standard form **before noun phrases**. Therefore the lemma of  $\epsilon-$  and  $\epsilon\rho\omicron-$  is  $\epsilon$ . For preposition forms containing a second person singular feminine pronoun (realized as zero), e.g.  $\epsilon\chi\omega$  ‘on you (fem.)’,  $\eta\kappa\omega$  ‘behind you (fem.)’ etc. see Portmanteau Tags.

### Existential and possessive predicates

The existential predicates are lemmatized as  $\omicron\upsilon\eta$  ‘there is’ and  $\mu\eta$  ‘there isn’t’ (again note that lemmatization does not contain supralinear strokes). Like auxiliaries, the related

possessive predicates are lemmatized using their form before the third person masculine singular:  $\omicron\gamma\eta\tau\alpha$  and  $\mu\eta\tau\alpha$ .

### **Auxiliaries, negations and future marker**

Auxiliaries are generally lemmatized to their form when preceding a nominal subject. Attention should be paid to auxiliaries sometimes ending in  $-\epsilon$ : in normalized orthography, this is generally present before a nominal subject. The lemma of  $\mu\alpha\rho\epsilon$ - and  $\mu\alpha\rho$ - (jussive) is  $\mu\alpha\rho\epsilon$ , and the lemma of  $\omega\alpha\eta\tau$ - and  $\omega\alpha\eta\tau\epsilon$  is  $\omega\alpha\eta\tau\epsilon$ .

However, the lemmas of auxiliaries that sometimes contain an intermediate pronoun do not contain that pronoun when they occur uninterrupted: the lemmas of  $\epsilon\rho\omega\alpha\eta$  (conditional) and  $\epsilon\rho\epsilon$  (optative) remain  $\epsilon\rho\omega\alpha\eta$  and  $\epsilon\rho\epsilon$ . These receive the tags ACOND and AOPT respectively. For cases with an intervening pronoun, which receive different tags, see Portmanteau Tags.

Negative morphemes such as  $\eta$ ,  $\alpha\eta$  and  $\tau\eta$  are their own lemmas (the form  $\mu$  before a labial is also lemmatized as  $\eta$ ). The negative imperative marker  $\mu\eta\rho$  is lemmatized as itself ( $\mu\eta\rho$ ), and NOT as  $\epsilon\rho\epsilon$ .

The future marker is given its own lemma  $\eta\alpha$ . Note that the lemma remains so whenever a future marker is separately identified, even if the diplomatic realization is assimilated and reduced to  $\alpha$ , e.g. in complex forms like  $\tau\epsilon\tau\eta\alpha$  ‘you will... (pl.)’ or  $\eta\epsilon\rho\alpha$  ‘you would have (fem. sg.)’.

### **Converters**

Like auxiliaries, converters are lemmatized to their form before a nominal subject, viz.:

CCIRC/CFOC:	$\epsilon\rho\epsilon$
CREL:	$\epsilon\tau\epsilon\rho\epsilon$
CPRET:	$\eta\epsilon\rho\epsilon$

For second person singular feminine  $\epsilon\rho/\epsilon\rho\epsilon$  (lemma=" $\epsilon\rho\epsilon\_η\tau\omicron$ ") see Portmanteau Tags.

### **Inflected modifiers**

Modifiers of the type  $\gamma\omega\omega$ -,  $\mu\mu\iota\eta\mu\mu\omicron$ -,  $\mu\alpha\gamma\alpha\alpha$ -,  $\tau\eta\rho$ - are lemmatized to their form before the **third person masculine singular** pronoun  $\eta$ . Thus  $\mu\mu\iota\eta\mu\mu\omicron$ - and  $\mu\mu\iota\eta\mu\mu\omega$ - are lemmatized as  $\mu\mu\iota\eta\mu\mu\omicron$ . The portmanteau form  $\mu\mu\iota\eta\mu\mu\omicron$  (yourself, fem. sg.) is lemmatized  $\mu\mu\iota\eta\mu\mu\omicron\_η\tau\omicron$  (see Portmanteau Tags).

## Portmanteau Tags

Some fused items receive a so-called portmanteau tag representing two categories at once. For example, the form  $\epsilon\varphi\alpha\lambda\iota$  is considered to contain a conditional auxiliary and a subject pronoun:  $\text{pos}=\text{"ACOND\_PPERS"}$ . In order to facilitate finding such cases regardless of the pronoun in use, in tags containing a conjugation base and a personal pronoun the form is lemmatized using both lemmas, separated by an underscore. For example, the lemmas of  $\epsilon\iota\varphi\alpha\lambda\iota$ ,  $\epsilon\sigma\varphi\alpha\lambda\iota$  and  $\epsilon\varphi\alpha\lambda\iota$  are  $\epsilon\rho\varphi\alpha\lambda\iota\_\alpha\lambda\omicron\kappa$ ,  $\epsilon\rho\varphi\alpha\lambda\iota\_\nu\tau\omicron\varsigma$  and  $\epsilon\rho\varphi\alpha\lambda\iota\_\nu\tau\omicron\upsilon$  respectively. The lemma of  $\epsilon\rho\varphi\alpha\lambda\iota$  remains  $\epsilon\rho\varphi\alpha\lambda\iota$  ( $\text{pos}=\text{"ACOND"}$ ), unless it contains a second person feminine singular subject, in which case the lemma is  $\epsilon\rho\varphi\alpha\lambda\iota\_\nu\tau\omicron$  according to the rule above.

For the past tense second person singular feminine form  $\alpha\rho$  the lemma is similarly  $\alpha\_\nu\tau\omicron$  ( $\text{pos}=\text{"APST\_PPERS"}$ ). The form  $\mu\mu\iota\mu\iota\mu\iota\omicron$  (yourself, fem. sg.) is identical to the base of other personal forms, but is lemmatized  $\mu\mu\iota\mu\iota\mu\iota\omicron\_\nu\tau\omicron$  just like other forms containing a personal pronoun.

The same principle applies to prepositions: forms containing a second person singular feminine pronoun (realized as zero) are given portmanteau lemmas, e.g.  $\epsilon\chi\omega$  ‘on you (fem.)’ has  $\epsilon\chi\iota\_\nu\tau\omicron$ ,  $\nu\sigma\omega$  ‘behind you (fem.)’ has  $\nu\sigma\alpha\_\nu\tau\omicron$  etc.

For circumstantial or focalizing converter + second person feminine singular, the lemma  $\epsilon\rho\epsilon\_\nu\tau\omicron$  is used (and similarly preterit  $\nu\epsilon\rho\epsilon\_\nu\tau\omicron$  and relative  $\epsilon\tau\epsilon\rho\epsilon\_\nu\tau\omicron$ ).

Verbs containing an object pronoun, such as  $\nu\tau$  ‘bring me’ are lemmatized using the base form of the verb and the pronoun’s lemma:  $\epsilon\iota\omicron\epsilon\_\alpha\lambda\omicron\kappa$ .

## Confusing cases

### Nouns with variant spellings

Lemmatization is the highest level of lexical abstraction, and as such should unify variants, even if the underlying forms are alternative ‘canonical’ spellings (e.g. listed in Crum’s dictionary as subentries). Spelling variants that are canonical should be retained on the norm level in order to make them findable as well – this is distinguished from the orig level, which is the least abstract, and may additionally contain non-canonical spelling variation, as well as diacritics.

Examples:

orig form	norm	lemma	(in bound group)	translation
$\omicron\omicron\Upsilon$	$\omicron\omicron\Upsilon$	$\varrho\omicron\omicron\Upsilon$	$\bar{\mu}\rho\omicron\omicron\Upsilon$	today
$\bar{\mu}\mu\Lambda$	$\mu\mu\Lambda$	$\mu\Lambda$	$\bar{\mu}\mu\Lambda\rho\omega\mu\epsilon$	there isn't

In other words, because the form  $\pi\omicron\omicron\gamma$  is a standard spelling for  $\pi\zeta\omicron\omicron\gamma$  in ‘today’, we do not normalize with a hori, allowing users to find this variant in norm while abstracting away from potential diacritics (including damage to letters, etc.), which would be present in orig. Users wishing to find all cases of the word for ‘day’ can still do so by searching for the lemma  $\zeta\omicron\omicron\gamma$ .