Coptic SCRIPTORIUM – Lemmatization Guidelines

Version 1.0.5 / 2017-09-15

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 ⲡ, ⲡⲉ, ⲧ, ⲧⲉ, ⲛ, ⲛⲉ and ⲙ (assimilated form of ⲛ before a labial consonant) is for all of the above ⲡ. For indefinites ⲟⲩ and ϩⲉⲛ there is no special masculine form, but the singular lemma ⲟⲩ is taken for the plural ϩⲉⲛ and also for the variant spelling ⲩ.

Copulas follow a similar rule: the lemma for all three number/gender forms (ⲡⲉ/ⲧⲉ/ⲛⲉ) is ⲡⲉ.

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. ϯ, 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.	ⲁⲛⲟⲕ	ⲁⲛⲟⲕ, ⲁⲛⲅ, ϯ, ⲓ, ⲛⲧ, ⲧ, ⲁ
2nd sg. masc.	ⲛⲧⲟⲕ	ⲛⲧⲟⲕ, ⲛⲧⲕ, ⲕ, ⲅ, ⲧⲕ
2nd sg. fem.	ⲛⲧⲟ	ⲛⲧⲟ, ⲛⲧⲉ, ⲧⲉ, ⲧⲣ, ⲣ, ⲉ
3rd sg. masc.	ⲛⲧⲟϥ	ⲛⲧⲟϥ, ϥ
3rd sg. fem.	ⲛⲧⲟⲥ	ⲛⲧⲟⲥ, ⲥ
1st pl.	ⲁⲛⲟⲛ	ⲁⲛⲟⲛ, ⲁⲛ, ⲛ, ⲧⲛ, ⲥⲛ
2nd pl.	ⲛⲧⲱⲧⲛ	ⲛⲧⲱⲧⲛ, ⲛⲧⲉⲧⲛ, ⲧⲛ, ⲧⲏⲩⲧⲛ
3rd pl.	ⲛⲧⲟⲟⲩ	ⲛⲧⲟⲟⲩ, ⲩ, ⲟⲩ, ⲥⲉ, ⲥⲟⲩ

The pronoun lemmas alone therefore primarily give access to search by person (1st, 2nd…); 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. ⲁⲛⲅ not ⲁⲛⲟⲕ) use the form search norm="ⲁⲛⲅ".

Possessives, interrogatives and demonstratives
Interrogative pronouns are each quivalent to their own lemma, i.e. ⲟⲩ is lemmatized ⲟⲩ and ⲛⲓⲙ as ⲛⲓⲙ.

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 ⲡⲉϥ is ⲡⲉϥ, the lemma of ⲡⲁ is ⲡⲁ etc., but the lemma of ⲧⲉⲥ and ⲛⲉⲥ is also ⲡⲉⲥ. Similarly, the lemma of ⲡⲉⲓ and ⲛⲉⲓ is ⲡⲉⲓ, and the lemma of ⲡⲁⲓ and ⲧⲁⲓ is ⲡⲁⲓ. This allows an easier search for all possessives (in ANNIS: pos="PPOS", finds ⲡⲉⲩ, ⲧⲉϥ, ⲛⲟⲩ …), all third person plural possessives (lemma="ⲡⲉⲩ", finds ⲡⲉⲩ, ⲧⲉⲩ and ⲛⲉⲩ) and all third person plural possessives of feminine objects (norm="ⲧⲉⲩ"), 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 -ως, which are lemmatized as such, e.g. ϩωλως has the lemma ϩωλως.

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. ϩωⲃ ‘deed’ is the lemma of both singular ϩωⲃ and plural ϩⲃⲏⲩⲉ, and similarly, possessed forms like ⲧⲟⲟⲧ(ϥ) are lemmatized to the absolute form, i.e. ⲧⲱⲣⲉ. 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. ϩⲛⲁ= and ϩⲛⲉ- '(one's) will' are lemmatized as ϩⲛⲉ. If only a presuffixal form exists, it is taken as the lemma as well, e.g. ⲛⲁⲓⲁⲧ= 'blessed is…' has the lemma ⲛⲁⲓⲁⲧ.

Nouns that have related masculine and feminine forms are considered separate lemmas. For instance, the noun ϣⲛⲣⲉ 'son' is its own lemma, and the separate noun ϣⲉⲉⲣⲉ 'daughter' also has a separate lemma (which is ϣⲉⲉⲣⲉ). Similarly, Greek words in -ος are considered separate from related words in -ον, e.g. ⲡⲟⲛⲏⲣⲟⲥ 'wicked person' is its own lemma, and so is the separate ⲡⲟⲛⲏⲣⲟⲛ '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. ⲥⲟⲧⲡ= and ⲥⲉⲧⲡ- are lemmatized as ⲥⲱⲧⲡ 'choose'. The same applies to stative and imperative forms, which are lemmatized to the dictionary entry, e.g. ⲕⲛⲧ† has ⲕⲱⲧ as a lemma and ⲁⲣⲓ has ⲉⲓⲣⲉ. Likewise for prenominal forms, ⲥⲉⲧⲡ and ⲣ are lemmatized as ⲥⲱⲧⲡ and ⲉⲓⲣⲉ. Compound imperatives receive compound infinitive forms, i.e. for ⲁⲣⲓⲑⲩⲥⲓⲁϫⲉ 'sacrifice!', the lemma is ⲣⲑⲩⲥⲓⲁϫⲉ.

Note that auxiliaries are not lemmatized to their etymological verbs, i.e. the lemma of the past tense ⲁ- is not ⲉⲓⲣⲉ but ⲁ. Additionally, the negative imperative marker ⲙⲡⲣ is lemmatized as ⲙⲡⲣ as well, as it is considered to be a form of negation independent from the verb ⲉⲣⲓⲉ. However, the negative imperative of ⲉⲓⲣⲉ itself, ⲙⲡⲱⲣ *IS* lemmatized as ⲉⲓⲣⲉ (since it is a morphological imperative of ⲉⲓⲣⲉ itself, and functions as part of its paradigm with the sense 'to do').

For fused verb-object forms like ⲛⲧ 'bring me', see Portmanteau Tags.

Prepositions

Prepositions are lemmatized to their standard form **before noun phrases**. Therefore the lemma of ⲉ- and ⲉⲣⲟ- is ⲉ. For preposition forms containing a second person singular feminine pronoun (realized as zero), e.g. ⲉϫⲱ 'on you (fem.)', ⲛⲥⲱ 'behind you (fem.)' etc. see Portmanteau Tags.

Existential and possessive predicates

The existential predicates are lemmatized as ⲟⲩⲛ 'there is' and ⲙⲛ '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: ⲟⲩⲛⲧⲁ and ⲙⲛⲧⲁ.

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 -ⲉ: in normalized orthography, this is generally present before a nominal subject. The lemma of ⲙⲁⲣⲉ- and ⲙⲁⲣ- (jussive) is ⲙⲁⲣⲉ, and the lemma of ⲱⲁⲛⲧ- and ⲱⲁⲛⲧⲉ is ⲱⲁⲛⲧⲉ.

However, the lemmas of auxiliaries that sometimes contain an intermediate pronoun do not contain that pronoun when they occur uninterrupted: the lemmas of ⲉⲣⲱⲁⲛ (conditional) and ⲉⲣⲉ (optative) remain ⲉⲣⲱⲁⲛ and ⲉⲣⲉ. 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 ⲛ, ⲁⲛ and ⲧⲙ are their own lemmas (the form ⲙ before a labial is also lemmatized as ⲛ). The negative imperative marker ⲙⲡⲣ is lemmatized as itself (ⲙⲡⲣ), and NOT as ⲉⲓⲣⲉ.

The future marker is given its own lemma ⲛⲁ. Note that the lemma remains so whenever a future marker is separately identified, even if the diplomatic realization is assimilated and reduced to ⲁ, e.g. in complex forms like ⲧⲉⲧⲛⲁ ‘you will… (pl.)’ or ⲛⲉⲣⲁ ‘you would have (fem. sg.)’.

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

CCIRC/CFOC:	ⲉⲣⲉ
CREL:	ⲉⲧⲉⲣⲉ
CPRET:	ⲛⲉⲣⲉ

For second person singular feminine ⲉⲣ/ⲉⲣⲉ (lemma="ⲉⲣⲉ_ⲛⲧⲟ") see Portmanteau Tags.

Inflected modifiers
Modifiers of the type ⲣⲱⲱ-, ⲙⲙⲓⲛⲙⲙⲟ-, ⲙⲁⲩⲁⲁ-, ⲧⲏⲣ- are lemmatized to their form before the **third person masculine singular** pronoun ⲩ. Thus ⲙⲙⲓⲛⲙⲙⲟ- and ⲙⲙⲓⲛⲙⲙⲱ- are lemmatized as ⲙⲙⲓⲛⲙⲙⲟ. The portmanteau form ⲙⲙⲓⲛⲙⲙⲟ (yourself, fem. sg.) is lemmatized ⲙⲙⲓⲛⲙⲙⲟ_ⲛⲧⲟ (see Portmanteau Tags).

Numerals
Feminine and masculine numerals take the masculine form as the lemma in order to facilitate searches based on the quantity itself (specific searches for either gender can be done using the literal form). For example the lemma of ⲥⲁϣϥⲉ ‘seven (fem.)’ is ⲥⲁϣϥ ‘seven (masc.)’. Note that compound numbers receive a complex lemma, therefore the lemma of ϫⲟⲩⲱⲧⲥⲁϣϥⲉ ‘twenty-seven (fem.)’ is ϫⲟⲩⲱⲧⲥⲁϣϥ ‘twenty-seven (masc.)’ (the individual parts cann still be annotated at the morph level, which is not lemmatized).

Portmanteau Tags
Some fused items receive a so-called portmanteau tag representing two categories at once. For example, the form ⲉⲩϣⲁⲛ is considered to contain a conditional auxiliary and a subject pronoun: pos="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 ⲉⲓϣⲁⲛ, ⲉⲥϣⲁⲛ and ⲉⲩϣⲁⲛ are ⲉⲣϣⲁⲛ_ⲁⲛⲟⲕ, ⲉⲣϣⲁⲛ_ⲛⲧⲟⲥ and ⲉⲣϣⲁⲛ_ⲛⲧⲟϥ respectively. The lemma of ⲉⲣϣⲁⲛ remains ⲉⲣϣⲁⲛ (pos="ACOND"), unless it contains a second person feminine singular subject, in which case the lemma is ⲉⲣϣⲁⲛ_ⲛⲧⲟ according to the rule above.

For the past tense second person singular feminine form ⲁⲣ the lemma is similarly ⲁ_ⲛⲧⲟ (pos="APST_PPERS"). The form ⲙⲏⲓⲛⲏⲙⲟ (yourself, fem. sg.) is identical to the base of other personal forms, but is lemmatized ⲙⲏⲓⲛⲏⲙⲟ_ⲛⲧⲟ 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. ⲉϫⲱ ‘on you (fem.)’ has ⲉϫⲛ_ⲛⲧⲟ, ⲛⲥⲱ ‘behind you (fem.)’ has ⲛⲥⲁ_ⲛⲧⲟ etc.

For circumstantial or focalizing converter + second person feminine singular, the lemma ⲉⲣⲉ_ⲛⲧⲟ is used (and similarly preterit ⲛⲉⲣⲉ_ⲛⲧⲟ and relative ⲉⲧⲉⲣⲉ_ⲛⲧⲟ).

Verbs containing an object pronoun, such as ⲛⲧ ‘bring me’ are lemmatized using the base form of the verb and the pronoun’s lemma: ⲉⲓⲛⲉ_ⲁⲛⲟⲕ.

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
ⲟⲟⲩ	ⲟⲟⲩ	ⲉⲟⲟⲩ	ⲙ̄ⲡⲟⲟⲩ	today
ⲙ̄ⲙⲛ	ⲙⲙⲛ	ⲙⲛ	ⲙ̄ⲙⲛⲣⲱⲙⲉ	there isn't

In other words, because the form ⲡⲟⲟⲩ is a standard spelling for ⲡⲉⲟⲟⲩ 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 ⲉⲟⲟⲩ.