Coptic Scriptorium – Guidelines Overview / Cheat Sheet

Segmentation

We use Layton's segmentation: each stressed unit is spelled together, with text split into stress-bearing bound groups (_), norm units (|), and morphemes (-). Bound groups are units which are uninterruptable (e.g. by ⲇⲉ) and are the maximum domain for phonotactic changes: ⲉⲙ|ⲡ|ⲏⲓ is one bound group, otherwise ⲛ → ⲙ before ⲡ would be word-external assimilation ('sandhi').

- ⲁ|ⲫ|ⲥⲱⲧⲙ_ⲇⲉ_ⲛⲟⲓ|ⲡ|ⲣⲱⲙⲉ_ⲭⲉ|ⲟⲩⲛⲧⲁ|ⲓ_ⲧⲁ|ⲭⲣⲉⲓⲁ (ⲭⲉ, ⲛⲟⲓ are spelled together!)
- ⲏⲛ|ⲧⲉⲓ|ⲙⲛⲧ-ⲁⲧ-ⲥⲱⲧⲙ (hyphens are used to separate constituent morphemes)
- ϯ|ⲣ-ⲉⲟⲧⲉ_ⲁⲛ (NB we can insert particles before ⲁⲛ; ⲣ-ⲉⲟⲧⲉ is a compound verb)

Coptic pausal marks (ˋ) also offer supporting evidence for stress-based bound groups – where they appear, they generally correspond to the bound group splits assumed here:

- ⲉⲣⲉ|ⲡⲉⲩ|ⲉⲟⲩⲟ̂_ⲣ̂-ⲙⲛ̄ⲧⲣⲉ̀_ⲭⲉ|ⲙⲉ|ⲛ|ⲭⲱⲕ̄ⲙ_ⲉ̄ⲛ|ⲥⲓⲟⲟⲩⲛⲉ̀

Compounds

Compound nouns are segmented into morphs (hyphenated) and identified by potentially having a single article at most. Compound verbs are identified by having no determiner (article, possessive or numeral) for the object:

- ⲡ|ⲙⲁⲓ-ⲛⲟⲩⲧⲉ (ⲙⲁⲓ-ⲛⲟⲩⲧⲉ is a single noun, since only one article is possible)
- ⲣ-ⲙⲉⲉⲩⲉ (no article on object → compound)
- ⲣ|ⲡ|ⲙⲉⲉⲩⲉ (article on object → separate norm units)
- ⲉ|ⲧⲣⲉ|ⲩ|ⲉⲟⲧⲡ|ⲥⲛⲁⲩ (number counts as a determiner → separate norm units)
- ⲁ|ⲫ|ⲁⲉⲉ|ⲣⲁⲧ|ⲫ (possessive counts as a determiner → separate norm units)

Handling theta/phi

Fused theta/phi is followed by a separator during transcription to indicate the correct number of segments. Later in spreadsheet mode, the norm and norm_group layers split the hori off:

- Orig/transcrition: ⲛ|ⲑ|ⲉ (norms: ⲛ+ⲧ+ⲉⲉ)

More examples

- ⲟⲩ|ⲣⲱⲙⲉ_ⲡⲉ (cf. ⲟⲩ|ⲣⲱⲙⲉ_ⲅⲁⲣ_ⲡⲉ)
- ⲁ|ⲓ|ⲥⲉⲧⲡ|ⲧⲉ|ⲥⲉⲓⲙⲉ (for prenominal form, BUT: ⲁ|ⲓ|ⲥⲱⲧⲡ_ⲧⲉ|ⲥⲉⲓⲙⲉ)
- ⲡ|ⲉⲧ|ⲛⲁⲛⲟⲩ|ⲫ (normal article, relative converter, etc.)
- ⲡ|ⲡ-ⲉⲧ-ⲛⲁⲛⲟⲩ-ⲫ (first article is normal, rest is a noun with morphological structure)

- ⲙⲡ|ⲟⲩ|ⲱ|ⲃⲱⲕ (the verb ⲱ is considered an independent auxiliary verb)
- ⲫ|ⲥⲙⲥⲟⲙ (NB quadrilateral verb stems are considered a single unit)
- ⲕⲁⲧⲁⲃⲁⲗⲗⲉ *but* ⲕⲁⲧⲁ|ⲫⲩⲥⲓⲛ (Greek verbs are not analyzed; but PREP+N is separate)

Markup

Markup in XML editor mode includes any structural annotations about the text which do not constitute Coptic text themselves. All Coptic text should be outside markup tags. Use TEI tags listed here. Markup totally ignores word segmentation, applying to any range of characters. Running examples look like this:

<pb xml:id="EG109"><cb n="1"><lb n="1"><hi rend="ekthetic">ⲁ|</hi>ⲫ|ϫⲟⲟ|ⲥ_ⲛⲥⲓ</lb>...

Common tags include: (note the use of opening AND closing tags throughout)

- <hi rend="ekthetic">...</hi> (or other rendering: tall, red, illuminated...)
- <pb xml:id="EG109">..</pb> (page break, i.e. the span of the page)
- <cb n="1">..</cb> (column break, i.e. the span of the column)
- <lb n="1">..</lb> (line break, i.e. the span of the line)
- <note note="barely visible">..</note>

Always surround "attributes" with straight double quotes; ***never*** use double quotes in attribute values.

Part of speech tagging

In spreadsheet mode, tags are given to each norm unit (but not to morphs, or bound groups). Compounds receive a single tag; for full guidelines see the documentation.

norm_group	orig_group	pos	lemma	orig	norm
ⲧⲉϯⲩⲭⲏ	ⲧⲉϯⲩⲭⲏ́	ART	ⲡ	ⲧⲉ	ⲧⲉ
		N	ϯⲩⲭⲏ	ϯⲩⲭⲏ́	ϯⲩⲭⲏ
ⲣⲱⲱⲥ	ⲣⲱ̀ⲱⲥ	IMOD	ⲣⲱⲱ	ⲣⲱ̀ⲱ	ⲣⲱⲱ
		PPERO	ⲛⲧⲟⲥ	ⲥ	ⲥ

Content words (tags in N, V*, ADV)*

- N for nouns, NPROP for proper names. Note that there are no adjective tags – these are interpreted as nouns (ⲱⲛⲙ) or verboids (ⲛⲁⲛⲟⲩ):
  - ⲓⲏⲥⲟⲩⲥ/NPROP ⲡ/ART ⲙⲁⲓⲣⲱⲙⲉ/N
- V for verbs, with VIMP for unique morphological imperative forms and VBD for verboids:
  - ⲛⲉ/ART ⲥⲛⲏ/N ⲣⲙⲉⲉⲩⲉ/V
  - ⲁⲣⲓ/V ⲧ/ART ⲁⲅⲁⲡⲏ/N
  - ⲡⲉⲭⲉ/VBD ⲡ/ART ⲝⲗⲗⲟ/N
- ADV for proper adverbs: ⲙⲛⲁⲩ/ADV , ⲕⲁⲗⲱⲥ/ADV (cf. PTC below for particles)

Auxiliaries and converters (tags in A, C*, FUT)*

APST	Aux., past	ⲁ	ACOND	Aux., conditional	ⲉⲣⲫⲁⲛ
ANEGPST	Aux., negated past	ⲙ̄ⲡ(ⲉ)	ALIM	Aux., limitative (‘until’)	ⲫⲁⲛⲧ(ⲉ)
ANY	Aux., ‘not yet’	ⲙ̄ⲡⲁⲧ(ⲉ)	ACONJ	Aux., conjunctive	ⲛ̄(ⲧⲉ)
AAOR	Aux., aorist	ⲫⲁ, ⲫⲁⲣⲉ	AFUTCONJ	Aux., future conjunctive	ⲧⲁⲣ(ⲉ)
ANEGAOR	Aux., negated aorist	ⲙⲉ(ⲣⲉ)	ACAUS	Aux., causative	ⲧⲣⲉ
AOPT	Aux., optative	ⲉ[ⲩ]ⲉ, ⲉⲣⲉ	CCIRC	Conv., circumstantial	ⲉ(ⲣⲉ)
ANEGOPT	Aux., neg. optative	ⲙ̄ⲙⲉ	CFOC	Conv., focalizing	ⲉ(ⲣⲉ)/(ⲉ)ⲛⲧ
AJUS	Aux., jussive	ⲙⲁⲣ(ⲉ)	CREL	Conv., relative	ⲉⲧ(ⲧ)/ⲉⲛⲧ
ANEGJUS	Aux., negated jussive	ⲙ̄ⲙ̄ⲧⲣⲉ	CPRET	Conv., preterite	ⲛⲉ(ⲣⲉ)
APREC	Aux., precursive	ⲛ̄ⲧⲉⲣ(ⲉ)	FUT	Future marker	ⲛⲁ

Pronouns

Tags distinguish personal subject (PPERS), object (PPERO) and independent pronouns (PPERI), as well as possessives (PPOS):

- ⲧ/PPERS ⲥⲟⲟⲩⲛ/V
- ⲛⲧⲟⲕ/PPERI
- ⲁ/APST ⲓ/PPERS ⲥⲟⲧⲙ/V ⲥ/PPERO
- ⲡⲁ/PPOS ⲙⲓ/N

Demonstratives are tagged PDEM, and interrogatives are PINT:

- ⲛⲧⲕ/PPERI ⲛⲓⲙ/PINT
- ⲡⲉⲓ/PDEM ⲙⲓ/N

Other function words

Tag	Name	Examples
ART	Article	ⲡ(ⲉ), ⲧ(ⲉ), ⲛ(ⲉ), ϩⲉⲛ, ⲕⲉ
CONJ	Conjunction	ⲁⲩⲱ, ⲏ, ⲙⲏ, ⲕⲁⲓ, ⲉⲓⲧⲉ, ...
COP	Copula	ⲡⲉ/ⲧⲉ/ⲛⲉ
EXIST	Existential/possessive	ⲟⲩⲛ/ⲙⲛ
FM	Foreign material	ⲡⲁⲣⲁ ⲧⲟⲩⲧⲟ
IMOD	Inflected modifier	ⲧⲏⲣ[ⲩ], ϩⲱⲱ[ⲧ], …
NEG	Negation	ⲛ, ⲁⲛ, ⲧⲙ[ⲥⲱⲧⲙ]
NUM	Numeral	ⲟⲩⲁ, ⲥⲛⲁⲩ, …
PREP	Preposition	ⲉⲧⲃⲉ, ϩⲛ̄, ⲛ, ⲙ̄ⲙⲟ[ⲩ], …
PTC	Particle	ⲇⲉ, ⲛ̄ⲟⲓ, ⲭⲉ, …
PUNCT	Punctuation	. , · …
UNKNOWN	Unknown morph, lacuna	ⲃ_ _ _, _ _ⲟⲥ, _ _ _, …

Portmanteau tags

Unsegmentable, fused forms, receive portmanteau tags, separated by underscore:

- ϩⲱ IMOD_PPERO (=ϩⲱⲱ+ⲁⲛⲟⲕ)
- ⲉⲣⲟ PREP_PPERO (=ⲉ + ⲛⲧⲟ, 2nd person singular feminine)
- ⲁ/APST ⲕ/PPRERS ⲛⲧ/V_PPERO (“you have brought me”)

Lemmatization

The guiding principle is to lemmatize to the most independent form possible, while ignoring number, but not gender or person (see details). Thus pronouns are lemmatized to the independent form:

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

Nouns are lemmatized to the independent form, if it exists, otherwise to the prenominal, or if it is not attested, to the presuffixal. Plurals are lemmatized to singular, but feminine forms receive their own lemma, and the same applies to Greek neuters:

- ⲥⲛⲏⲩ ⟶ ⲥⲟⲛ
- ⲱϩⲣⲉ ⟶ ⲱϩⲣⲉ, ⲱⲉⲉⲣⲉ ⟶ ⲱⲉⲉⲣⲉ
- ⲧⲟⲟⲧ[ⲥ] ⟶ ⲧⲱⲣⲉ (has independent form)
- ϩⲛⲁⲁ= ⟶ ϩⲛⲉ (prenominal form exists)
- ⲛⲁⲓⲁⲧ= ⟶ ⲛⲁⲓⲁⲧ (only presuffixal form)
- ⲡⲟⲛϩⲣⲟⲥ⟶ⲡⲟⲛϩⲣⲟⲥ, ⲡⲟⲛϩⲣⲟⲛ⟶ⲡⲟⲛϩⲣⲟⲛ

Verbs are lemmatized to independent forms and verboids to the prenominal form:

- ⲥⲱⲧⲡ, ⲥⲟⲧⲡ, ⲥⲉⲧⲡ ⟶ ⲥⲱⲧⲡ
- ⲡⲉⲭⲁ, ⲡⲉⲭⲉ ⟶ ⲡⲉⲭⲉ

Prepositions, converters and auxiliaries to prenominal forms:

- ⲉⲣⲟ[ϥ] ⟶ ⲉ, ⲛϩⲏⲧ[ϥ] ⟶ ϩⲛ
- ⲉ, ⲉⲣⲉ ⟶ ⲉⲣⲉ
- ⲉⲧ, ⲉⲧⲉ, ⲉⲧⲉⲣⲉ ⟶ ⲉⲧⲉⲣⲉ
- ⲙⲉ[ϥ], ⲙⲉⲣⲉ ⟶ ⲙⲉⲣⲉ
- Note past relative: ⲉⲛⲧ[ⲁ] ⟶ ⲉⲧⲉⲣⲉ
- But past focalizer: (ⲉ)ⲛⲧ[ⲁ] ⟶ ⲉⲣⲉ

Portmanteau tokens receive portmanteau lemmas (word/pos/lemma):

- ⲉⲣⲟ/PREP_PPERO/ⲉⲣⲉ_ⲛⲧⲟ
- ϩⲱ/IMOD_PPERO/ϩⲱⲱ_ⲁⲛⲟⲕ

Also see online guidelines on diplomatic transcription, entity tagging, and syntactic analysis