Coptic SCRIPTORIUM Diplomatic Transcription Guidelines

Version: *1.0.2_2013.6.24*

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

1. Humboldt-Universität zu Berlin
2. University of the Pacific

1. Preamble

This document details guidelines for transcribing a diplomatic edition of a manuscript in Sahidic Coptic according to the Coptic SCRIPTORIUM project scheme. The diplomatic transcription currently requires extensive manual annotation, due to the complexities of processing a diplomatic text in which no word breaks exist in the original and yet words and even morphemes span across line, column, and page breaks.

The transcription procedure assumes familiarity with basic paleography and traditional manuscript transcription following the Leiden conventions. (http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden)

The diplomatic transcription also utilizes XML (eXtensible Markup Language) -like tagsets, including some of the TEI (Text Encoding Initiative) XML markup language, although the resulting document is **not** a valid XML document. Wherever possible, the EpiDoc subset of TEI XML is utilized for element nomenclature. EpiDoc TEI conventions were created by and for epigraphers and have come to be a standard in markup of ancient texts, epigraphic or otherwise. (http://sourceforge.net/p/epidoc/wiki/Home/) In contrast to TEI, SCRIPTORIUM utilizes no milestone XML tags (e.g., <cb/>). Instead, all tags are span annotations (e.g., <cb>This is a column of Coptic text.</cb>).

The aim is twofold: 1) to achieve a transcription that documents the text and visualization of the manuscript as closely as possible to the original; 2) to provide a text file that can be processed by various digital tools and software, such as a tokenizer, a part-of-speech tagger, or the ANNIS database infrastructure (http://www.sfb632.uni-potsdam.de/annis/; Zeldes et al. 2009). The resulting transcription itself does not resemble a traditional text of a diplomatic edition. The markup ensures optimization for processing and search using such tools and software. For examples of the diplomatic editions visualized in HTML generated from the post-ANNIS transformations, see the sample corpora at http://coptic.pacific.edu.

2. Character Encoding

Texts are encoded using the UTF-8 (Unicode) Coptic language character set. The freely available Antinoou font and Coptic-English keyboard created by Michael Everson in cooperation with the International Association of Coptic Studies is the standard (http://www.evertype.com/fonts/coptic/). Unicode characters in the private use area are not recommended.

2.1 Alphanumeric Characters

Characters follow the orthography of the manuscript.

Oversized characters typically are not marked with XML tagging. Instead the uppercase version of the character is used.

2.2 Punctuation and Decoration

Punctuation and decoration follows the manuscript as closely as possible within the Unicode character set. Not all decoration and punctuation can be encoded using characters; deviations or documentation that can’t be keyed in is instead typically indicated in a note element.

Notes on individual specific punctuation characters:

> For the character “`” that occasionally appears at the end of words in some manuscripts, use the keystroke option+` using the Coptic-English keyboard layout (for MacIntosh).
>
> Example: ⲡⲉⲙⲙⲟⲛ`ⲧⲉ

2.3 Accentuation and Supralinear Strokes

Accentuation and supralinear strokes follow the orthography of the manuscript. Some manuscripts have binding strokes between letters (e.g. ϩ︤ⲛ︥) whereas others in the case of the same word might only provide a stroke over a single letter (e.g., ϩⲛ̄). The diplomatic transcription follows the conventions of the manuscript, even if the manuscript is internally inconsistent or contains what seem to be errors.

Notes on encoding individual specific accents, strokes, etc, using the Coptic-English keyboard for Antinoou (for MacIntosh):

> ̄ (as in ⲙ̄) the supralinear stroke above only one letter: type the letter followed by ;
>
> ‾ (as in ⲙ︤ⲛ︥) the binding stroke between two letters: type first letter then U+FE24 (< in the Coptic-English keyboard) then second letter then U+FE25 (> in the Coptic-English keyboard), i.e. m<n> on a Mac using the Coptic-English keyboard
>
> ‾ (as in ⲙ︤ⲛ︦ⲧ︥) binding stroke over three letters: type the first letter then U+FE24 (< on a Mac using the Coptic-English keyboard) then second letter then U+FE26 (: [i.e. shift+;] on a Mac using the Coptic-English keyboard) then third letter then U+FE25 (> on a Mac using the Coptic-English keyboard), i.e. m<n:t>)
>
> ᷍ (as in ⲟ᷍ⲩ) circumflex combining two letters: U+1DCD (keystroke shift+option+/ on a Mac using the Coptic-English keyboard) typed between the letters, so o¿u (o then shift+option+/ then u)
>
> For curved or jagged strokes over etas, use a regular circumflex rather than a dot or line or trema (option+3)

Tremas (ï, ⲏ̈): type the letter followed by option+7.

3. Text Divisions

A diplomatic transcription aims to preserve the formatting of the original text. Line breaks, column breaks, and page breaks as they appear in the manuscript are all documented.

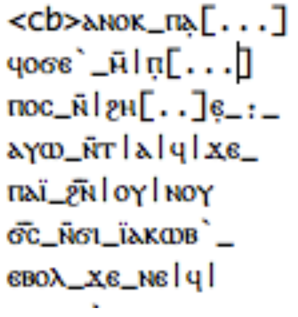
3.1 Line Breaks

All line breaks in the transcription should follow the line breaks of the manuscript. Use the “Enter” or “Return” key to produce a line break in the text file of the transcription.

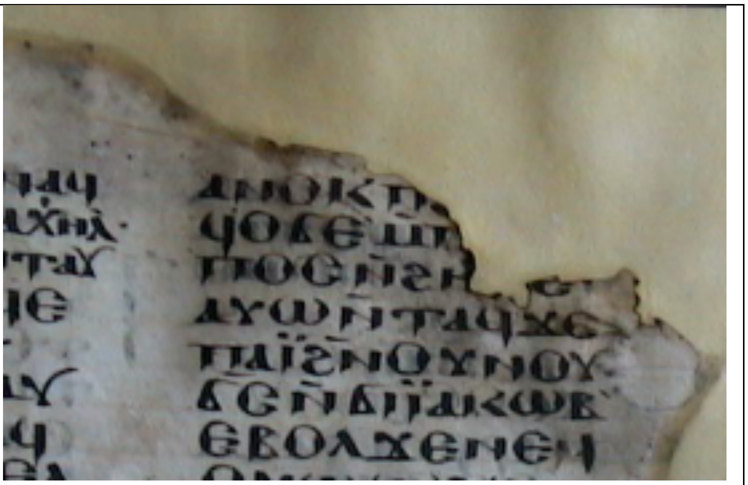
3.2 Column Breaks

All column breaks in the transcription should follow the column divisions in the manuscript. Columns are wrapped in span annotations using the <cb></cb> tagset.

Fig. 1: Opening column break tag, corresponding to beginning of manuscript column

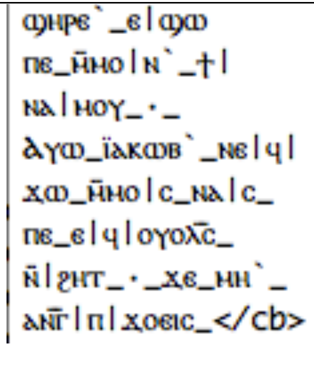


```
<cb>ⲁⲛⲟⲕ_ⲡⲁ̣[...]
ϥⲟϭⲉ`_ⲙ̄|ⲡ̣[...]
ⲡⲟⲥ_ⲛ̄|ϩⲏ[..]ⲉ̣_:_
ⲁⲩⲱ_ⲛ̄ⲧ|ⲁ|ϥ|ϫⲉ_
ⲡⲁï_ϩ̄ⲛ|ⲟⲩ|ⲛⲟⲩ
ϭ̄ⲥ_ⲛ̄ϭⲓ_ïⲁⲕⲱⲃ`_
ⲉⲃⲟⲗ_ϫⲉ_ⲛⲉ|ϥ|
```

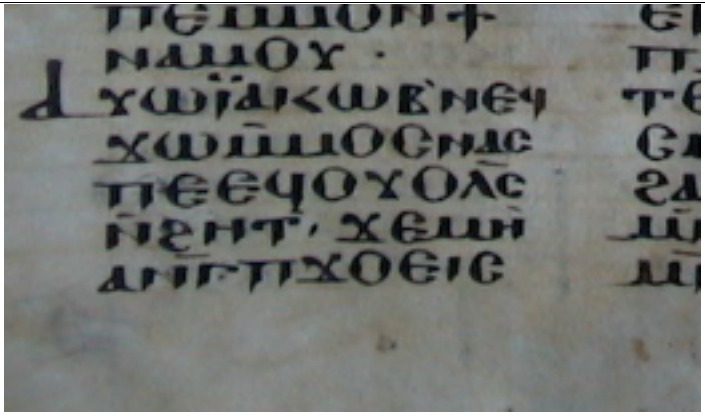


MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples, Biblioteca Vittorio Emanuele III

Fig. 2: Closing column break tag, corresponding to the end of a manuscript text column



```
ϣⲏⲣⲉ`_ⲉ|ϣⲱ
ⲡⲉ_ⲙ̄ⲙⲟ|ⲛ`_ϯ|
ⲛⲁ|ⲙⲟⲩ_·_
ⲁⲩⲱ_ïⲁⲕⲱⲃ`_ⲛⲉ|ϥ|
ϫⲱ_ⲙ̄ⲙⲟ|ⲥ_ⲛⲁ|ⲥ_
ⲡⲉ_ⲉ|ϥ|ⲟⲩⲟⲗ̄ⲥ̄_
ⲛ̄|ϩⲏⲧ_·_ϫⲉ_ⲙⲏ`_
ⲁⲛ̄ⲅ̄|ⲡ|ϫⲟⲉⲓⲥ_</cb>
```



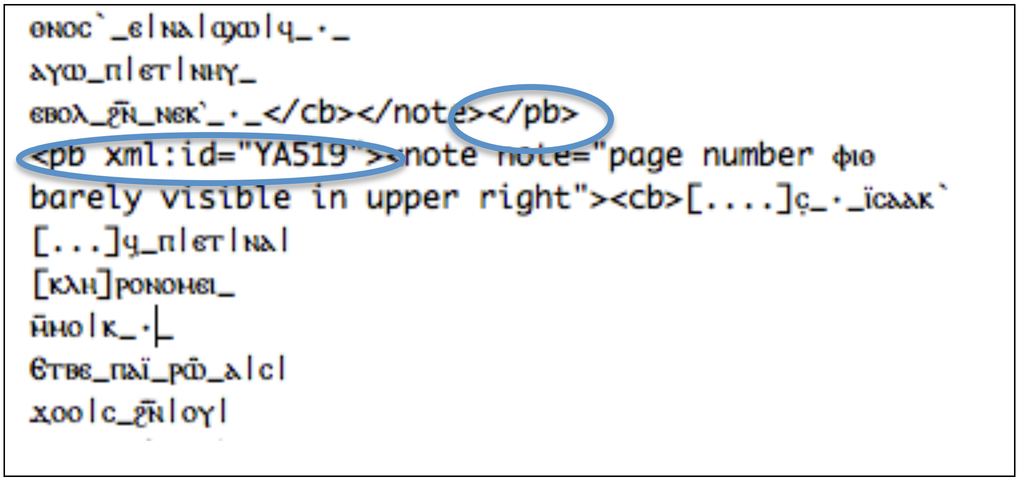
MONB.YA 520; Coptic Manuscript IB 2 f. 27v, Naples,

	Biblioteca Vittorio Emanuele III,”

3.3 Page Breaks and Numbering

All page breaks in the transcription should follow the page divisions in the manuscript. Page numbering in the transcription reflects the page numbering in the original manuscript codex. Codex sigla in the example below are two-letter codes following the White Monastery codex siglum list created by Tito Orlandi (Orlandi 2002; http://cmcl.aai.uni-hamburg.de/). Page breaks are wrapped in TEI compatible span annotations using the <pb></pb> tagset with the xml:id element. The entire page of text (including the relevant column tags) should be wrapped with these tags. Thus <pb xml:id="YA518"> is the opening tag for page 518 in White Monastery codex YA (MONB.YA). The xml:id should not contain spaces. (Thus, xml:id="YA518" not xml:id="YA 518">.)

Fig. 3: Closing and opening page break tags indicating the end of one page and beginning of the next. (Note: the opening tag for the first page and closing tag for the second page are not visible here but are required.)



```
ⲑⲛⲟⲥ`_ⲉ|ⲛⲁ|ϣⲱ|ϥ_·_
ⲁⲩⲱ_ⲡ|ⲉⲧ|ⲛⲏⲩ_
ⲉⲃⲟⲗ_ϩⲛ̄_ⲛⲉⲕ`_·_</cb></note></pb>
<pb xml:id="YA519"><note note="page number ⲫⲓⲑ
barely visible in upper right"><cb>[....]ⲥ̣_·_ⲓ̈ⲥⲁⲁⲕ`
[...]ϥ_ⲡ|ⲉⲧ|ⲛⲁ|
[ⲕⲗⲏ]ⲣⲟⲛⲟⲙⲉⲓ_
ⲙ̄ⲙⲟ|ⲕ_·⌊
ⲉⲧⲃⲉ_ⲡⲁⲓ̈_ⲣⲱ̄_ⲁ|ⲥ|
ϫⲟⲟ|ⲥ_ϩⲛ̄|ⲟⲩ|
```

The location and Coptic numeration of the page number is currently documented in a note element. (See Figure 3 above).

4. Word Segmentation, Spacing, and Tokenization

Sahidic Coptic words are formed by several morphemes attaching together. One word may include a preposition and object, or a verbal auxiliary + subject + infinitive, or even more morphemes strung together. In most manuscripts, no spaces between words are provided. Coptic SCRIPTORIUM marks both word and morpheme segmentation following the practices in Bentley Layton’s grammar (Layton 2011).

4.1 Word Segmentation

SCRIPTORIUM diplomatic transcription marks word segmentations according to Layton's conventions (Layton 2011). The transcriber inserts an underscore ("_") after each Coptic word, even when the end of the word falls at the end of a line.

Likewise, all punctuation is followed by an underscore.

(1) ⲉⲧⲉⲓ̈ⲥⲙⲁⲏⲗ_ (word ends at end of line)
(2) ⲡⲉ_ⲛϥⲛⲁⲕⲗⲏ (two words, in which the second word flows into line 3)
(3) ⲣⲟⲛⲟⲙⲉⲓ_ⲙ (the word continues from line 2, is followed by an underscore)
(4) ⲙⲟⲕ_ⲁⲛ_·_ (punctuation followed by an underscore)

These underscores are not and do not need to be visualized in HTML transformations of the diplomatic editions; they are nonetheless essential for processing the text, since they will enable searches and visualizations of a word-segmented text.

4.2 Spacing

Blank spaces in the transcription correspond to spaces in the manuscript. Consequently, if the manuscript provides no spaces between words or punctuation, the diplomatic transcription contains no spaces. Where there are significant spaces in the manuscript that the transcriber wishes to draw attention to, the transcription includes a space followed by an underscore ("_"), i.e. the space behaves like an additional independent word form.

4.3 Tokenization of Morphemes

Words are segmented into morphemes using the pipe character ("|").

(1) ϩⲓⲧⲙ|ⲡ|ⲛⲟⲩⲧⲉ_ (preposition|article|noun)
(2) ⲉⲧⲉ|ⲓ̈ⲥⲙⲁⲏⲗ_ (converter|noun)
(3) ⲡⲉ`_ⲛ|ϥ|ⲛⲁ|ⲕⲗⲏ (word_auxiliary|subject pronoun|future marker|verb (verb continues to line 4)
(4) ⲣⲟⲛⲟⲙⲉⲓ`_ⲙ̄

5. Rendering and Leiden Transcription Conventions

Coptic SCRIPTORIUM uses Leiden and Leiden+ conventions for transcribing manuscripts. The encoding follows the EpiDoc guidelines. Not all Leiden documentation is currently XML encoded as Leiden+, however.

5.1 Characters Highlighted, Raised, Lowered, or Set Apart in Some Way

Characters that are raised, lowered, or printed in different colors or styles are encoded using the TEI XML element <hi> with the rend attribute. Letters written above the line are encoded: <hi rend="superscript">. Characters written below the line are encoded: <hi rend="subscript">. Letters in a different color ink are encoded with the color ink, e.g., <hi rend="red">. It is possible to combine these annotations, e.g. <hi rend="red subscript">. Any additional information can be provided in a note element.

Example	*Diplomatic Visualization*
ϩⲓⲧⲙ̅|ⲡ|ⲛ<hi rend="superscript"><note note="o is directly above the ⲩ">o</note></hi>ⲩ	ϩⲓⲧⲙ̅ⲡⲛ^{o}ⲩ (ANNIS) or ϩⲓⲧⲙ̅ⲡⲛ\o/ⲩ (EpiDoc XSLT)

5.2 Damaged Characters

Characters that are damaged but restored based on context are marked with an underdot. Coptic SCRIPTORIUM uses the diacritical character ̣ (Unicode U+0323). These characters are not currently encoded in TEI XML using the EpiDoc tagset for Leiden+. Coptic SCRIPTORIUM uses the underdot character rather than annotation to designate this information.

5.3 Lacunae and Lost Characters

Lost lines and characters (lacunae) are indicated using square brackets, as in the Leiden conventions. They may be encoded using the EpiDoc tagset, but it is not required. See EpiDoc guidelines for more details (“EpiDoc Guidelines: Lost Characters, Quantity Unknown”; “EpiDoc Guidelines: Editorial Restoration: Characters Lost but Restored by Modern Editor”; “EpiDoc Guidelines: Lost Characters, Quantity Approximate”; “EpiDoc Guidelines: Lost Characters, Quantity Known”; “EpiDoc Guidelines: Erased and Lost”; “EpiDoc Guidelines: Lacunas, Other Units”).

(1) Example encoded using the gap element:
```
<gap reason="lost">
[ ]_
[ ]_
[ ]_
</gap>
```
(2) Unencoded gaps (no XML elements):
```
[.....]ⲡ̣ⲣ̣[..]
```

5.4 Other

Other rendering information is encoded either according to EpiDoc conventions or recorded as information within a note element.

6.0 File Format and Document Preferences

Documents are transcribed in a text editor such as TextEdit. Document preferences are set to UTF-8 encoding without byte-order Mark (BOM). (E.g., in TextEdit 1.7.1 for MacIntosh, in the File-->Preferences menu, click on “Open and Save,” and select “Unicode (UTF-8)” for Opening files and Saving files.)

Bibliography

An up-to-date bibliography can be found at the project’s Zotero page: https://www.zotero.org/groups/coptic_SCRIPTORIUM/items/collectionKey/8IHTW3NZ

"ANNIS2 - Search and Visualization." *ANNIS2 - Search and Visualization in Multilevel Linguistic Corpora with ANNIS*. Accessed May 29, 2013. http://www.sfb632.uni-potsdam.de/annis/.

Bodard, Gabriel. "EpiDoc Appendix: Glossary: Leiden, Leiden-plus." *Appendix: Glossary*. 18 Jun. 2013. <http://www.stoa.org/epidoc/gl/latest/app-glossary.html#leiden>.

"Corpus Dei Manoscritti Copti Letterari." *CMCL - Studies in Coptic Civilization*. 11 Sep. 2012. <http://cmcl.aai.uni-hamburg.de/>.

"EpiDoc Guidelines." *EpiDoc Guidelines*. 25 May 2013. <http://www.stoa.org/epidoc/gl/dev/>.

---. *EpiDoc: Epigraphic Documents in TEI XML*. 25 May 2013. <http://sourceforge.net/p/epidoc/wiki/Home/>.

"Evertype: Antinoou." *Evertype: Antinoou - A Standard Font for Coptic* 2012. 29 May 2013. <http://www.evertype.com/fonts/coptic/>.

Layton, Bentley. *A Coptic Grammar*. 3rd Edition, Revised. Wiesbaden: Harrassowitz, 2011. Print.

Orlandi, Tito. "The Library of the Monastery of Saint Shenute at Atripe." *Perspectives on Panopolis: An Egyptian Town from Alexander the Great to the Arab Conquest*. Leiden: Brill, 2002. 211–231. Print.

Zeldes, Amir et al. "ANNIS: A Search Tool for Multi-Layer Annotated Corpora." *Proceedings of Corpus Linguistics 2009* (2009) : n. pag. 10 Sep. 2012. <http://ucrel.lancs.ac.uk/publications/cl2009/>.