THE BASIC COPTOLOGICAL CHARACTER SET IN UNICODE'S UCS
Some Comments by Stephen Emmel, 27 January 2004

Here I want to acknowledge the valued assistance of Deborah W. Anderson, Researcher, Deptartment of Linguistics, University of California at Berkeley, who helped me in refining my memo ("expert contribution") to the Unicode Technical Committee (UTC), "Changes to ISO/IEC JTC1/SC2/WG2 N2636 and N2676 recommended by the International Association for Coptic Studies," dated 26 January 2004.

What follows is a list of all the characters that I regard as essential to have in the Unicode Universal Character Set (UCS) if we are to have a firm basis for an optimal Coptologist's (set of) Coptic font(s) sometime in the future. Some of these characters exist in the UCS already (as indicated by codes of the form "U+xxxx"), although not necessarily as specifically Coptic characters (e.g. general punctuation marks, which may or may not require a specifically "Coptic" shape ["glyph"] in a particular font); others are already on their way into the UCS via N2676 (as indicated by codes of the form "xxxx"); others are proposed in my comments to the UTC (UNDERLINED CHARACTER NAMES). **The present memo could be used as the basis for a set of recommendations/guidelines for creating a Coptic font using Unicode. Obviously, what follows is not sufficient in and of itself. But *one* task is dealt with here, namely determining which Unicode characters (and I mean *precisely* which ones) need to be included in a minimally complete Coptologist's Coptic font.**

1. Alphabetic characters. I count 81 essential characters, 32 of them in cased paired (*alpha* through *ti*, including *stigma*), the remaining 17 being non-Bohairic characters (which therefore have no special upper-case shapes). 14 of these characters were in the UCS from the beginning (U+03E2 – U+03EF); 51 more are added in N2676 (2C80–2C98 and 2CA0–2CB9); the remaining 16 are added in my memo to the UTC (§ II.2d).[1]

2. Non-alphabetic numerals. I count 3, two of which appear in N2676 (2CBA COPTIC SMALL LETTER RO WITH STROKE and 2CBF COPTIC FRACTION ONE HALF); as for a diacritic to be used to mark any numeral as a multiple of 1000, I assume that the existing character U+033F COMBINING DOUBLE OVERLINE will be sufficient for our needs. (Documentary papyrologists wanting more, should consult proposals submitted by the TLG; see: http://www.tlg.uci.edu/Uni.prop.html.)

3. Cryptograms. I consider 2 as essential: the COPTIC VERTICAL CRYPTOGRAM and the COPTIC HORIZONTAL CRYPTOGRAM, both of which are added in my memo (§ II.2e).

4. Symbols (compendia, monograms). I consider 6 as essential, two of which are added in N2676 (2CBB COPTIC SYMBOL MI RO and 2CBB COPTIC SYMBOL KHI RO), and four of which are in my memo (§ II.2f: COPTIC SYMBOL TAU RO, COPTIC SYMBOL PI RO, COPTIC SYMBOL KAI, and COPTIC SYMBOL O UA). N2676 adds a compendium for *stauros* (2CBD ⳧ COPTIC SYMBOL STAUROS [as corrected in my memo, § II.2b]), perhaps taken over from rather arbitrary lists such as Mallon 1956 (4th ed.), 234 top. While I saw no reason to try to be exhaustive in cataloging such symbols, it did seem to me to make sense to add the typical Bohairic ligatured abbreviation for *cōis*, **ϭ̅ⲥ̅** (§ II.2f: COPTIC SYMBOL COIS).

5. "Markers of syllabicity" ("combining" characters because they occur above, or "combined with," another character). In principle, only 3 characters must be registered: the superlinear stroke and the jinkim in two forms (dot and grave accent mark). But in § II.2g of my comments I tried to explain something about the complexity of these characters' behavior, in order to indicate that a font designer might have a hard time working with just three characters. But then again, since I myself have no good idea how a professional font designer might go about solving this problem, I refrained from proposing a specific number of characters beyond the three "Platonic forms" COPTIC COMBINING SUPERLINEAR STROKE, COPTIC COMBINING JINKIM, and COPTIC COMBINING GRAVE ACCENT. Experience has taught me that Unicode is most unlikely to see these three characters as unique to Coptic, and so I took the pragmatic step of conceding that the Coptic superlinear stroke could probably be handled more or less effectively with the existing characters U+0304 COMBINING MACRON, U+0305 COMBINING OVERLINE, and U+035E COMBINING DOUBLE MACRON. Beyond that, it will be up to the font designer to determine how the special behavior of the Coptic superlinear stroke is to be dealt with, especially when it occurs, in various functions, over three letters or more. For the jinkim, clearly the existing characters U+0307 COMBINING DOT ABOVE and U+0300 COMBINING GRAVE ACCENT are equivalent to the basic Coptic character. For left- and right-shifted jinkims, I proposed two new "general punctuation" characters: COMBINING DOT ABOVE LEFT and COMBINING DOT ABOVE RIGHT.

[1] I should perhaps note here, with regard to the Old Coptic characters discussed by Iain Gardner, "An Old Coptic Ostracon from Ismant el-Kharab?" *Zeitschrift für Papyrologie und Epigraphik* 125 (1999) 195–200, pl. 10, at p. 198, that I finally decided that all six of these characters can be identified as ("unified with," in Unicode-speak) other, known characters. 1 = ȝ; 2 = ϥ; 3 = ϧ; 4 = ϩ; 5 = ϣ; and 6 = ϭ or perhaps some kind of a compendium (perhaps of ϭⲁ or something else).

6. Additional combining characters. The obvious list is: circumflex, trema (“diaeresis”), spiritus (asper and lenis), acute accent mark, and the Trenner (“division mark” or “apostrophe,” of which I distinguish three shapes: straight apostrophe, hook, and point). See § II.2h of my memo. Here again it is unlikly that Unicode would agree to register these as Coptic characters, because in most cases there exist already equivalent characters in the UCS. Where they do not exist already, I have recommended adding them, as follows:

one-letter circumflex: U+0302 COMBINING CIRCUMFLEX ACCENT
  or: U+0311 COMBINING INVERTED BREVE
two-letter circumflex: U+0361 COMBINING DOUBLE INVERTED BREVE
trema: U+0308 COMBINING DIAERESIS
spiritus asper: U+0485 COMBINING CYRILLIC DASIA PNEUMATA
spiritus lenis: U+0486 COMBINING CYRILLIC PSILI PNEUMATA
acute accent mark: U+0301 COMBINING ACUTE ACCENT
straight apostrophe Trenner (combining): COMBINING GRAVE ACCENT ABOVE RIGHT
hook Trenner (combining): COMBINING COMMA ABOVE RIGHT
point Trenner (combining): COMBINING DOT ABOVE RIGHT (same as for right-shifted dot jinkim)

Additionally, and in imitation of what has already been done for Greek, I proposed registering two pre-combined characters: ï COPTIC SMALL LETTER IAUDA WITH TREMA and ϔ COPTIC SMALL LETTER UA WITH TREMA.

7. Punctuation marks. One punctuation mark is added in N2767 (2CBE COPTIC FULL STOP). Here I finally decided that I could request only the one mark that we need that really seems to be missing from the UCS, namely the MIDDLE COMMA). Beyond this, all other essential marks are already represented in the UCS in at least one form, as follows:

U+002C	COMMA
U+002E	PERIOD
U+00B7	MIDDLE DOT (= U+0387 GREEK ANO TELEIA) (= “raised point”)
U+003B	SEMICOLON
U+003A	COLON
U+0020	SPACE
U+2053	SWUNG DASH (= “swash divider”)
U+0060	GRAVE ACCENT (= U+02CB MODIFIER LETTER GRAVE ACCENT) (=“straight apostrophe Trenner” [non-combining])
U+0027	APOSTROPHE (= U+02BC MODIFIER LETTER APOSTROPHE) (= “hook Trenner” [non-combining])
U+02D9	DOT ABOVE (= “point Trenner” [non-combining])

Or they are added in N2676:

2E0E	EDITORIAL CORONIS (= “decorative coronis”)
2E0F	PARAGRAPHOS
2E10	FORKED PARAGRAPHOS
2E13	DOTTED OBELOS
2E16	DOTTED RIGHT-POINTING ANGLE (= “diple”)
2056	THREE DOT PUNCTUATION
2058	FOUR DOT PUNCTUATION
2059	FIVE DOT PUNCTUATION
205A	TWO DOT PUNCTUATION
205B	FOUR DOT MARK
etc.	

8. Grammatical symbols. I have heard that our beloved (but still “etymologically” mysterious) “slanted equals sign” (or whatever one wishes to call it) for marking the status pronominalis of infinitives etc. caused a good deal of discussion among the Unicoders, and it is now added in N2676 as 2E17 ⸗ DOUBLE OBLIQUE HYPHEN (“used in ancient Near-Eastern linguistics”). Of course the hyphen that we use to mark the status nominalis (and so forth) must be unified with U+2010 HYPHEN (= U+002D HYPHEN-MINUS; cf. U+2011 NON-BREAKING HYPHEN). And our small raised dagger for marking the stative of verbs will have to be U+2020 DAGGER rendered small and superscript by our word processors and rendering devices.

9. Editorial symbols. As far as I can tell, all the essential editorial symbols are already either in the UCS or are added in N2676, as follows:

( )	U+0028 and U+0029 LEFT/RIGHT PARENTHESIS
[ ]	U+005B and U+005D LEFT/RIGHT SQUARE BRACKET
< >	U+003C and U+003E LESS-THAN/GREATER-THAN SIGN
{ }	U+007B and U+007C LEFT/RIGHT CURLY BRACKET
⟦ ⟧	U+27E6 and U+27E7 MATHEMATICAL LEFT/RIGHT WHITE SQUARE BRACKET
|	U+007C VERTICAL LINE (= “line break”)
‖	U+2016 DOUBLE VERTICAL LINE (= “fifth-line break” or the like)
*	U+002A ASTERISK (= “page break,” for example)
.	U+0323 COMBINING DOT BELOW (= “subliteral dot”)

paired grave and acute accents: 2E0C and 2E0D RAISED SMALL DIAGONAL UPPER LEFT TO LOWER RIGHT/LOWER LEFT TO UPPER RIGHT in N2676 (to mark extralinear insertions)
underline: U+0332 COMBINING LOW LINE (to mark documented, but lost, text, for example)
dagger: U+2020 DAGGER (to mark corrupt text)

etcetera etcetera (note, for example, in N2676 the section of “New Testament editorial symbols,” 2E00–2E0D, and so forth).