Copy

UFEthiopianProposal -- Proposal for Ethiopian Encoding in Unicode/10646

Joe Becker

January 10, 1993

The Ethiopian proposal consists of a list of questions/issues, a chart, a character names list, and a block introduction. The content is based on UTC/1991-026 "On the Extended Ethiopic Alphabet" of February 26, 1991 and its later adjustments by Lloyd Anderson, unioned with features of the Xerox Amharic implementation by Joe Becker. The character names are based on those in DP 10646, which came from WG2/N459 "Ethiopian character sets" by Michael Mann.

==========

QUESTIONS FOR REVIEWERS:

> Is this collection missing any important, well-established "extension" letters for writing less-common languages?

> Are the glyphs in the charts appropriate?

> Can you supply documentation to support the specification of the following two characters?

121D ETHIOPIAN CONSONANT GG
Bilen

1237 ETHIOPIAN VOWEL PHONETIC AE
used primarily with U+1211 ETHIOPIAN CONSONANT GLOTTAL

In particular, does U+1237 occur (as a vowel, not as a mark of "w" rounding) on any consonant other than U+1211? Should the combination of U+1237 with U+1211 simply be encoded as a distinct consonant (to be added between current U+1211 and U+1212)?

> Are the following characters specified correctly?

1256 ETHIOPIAN COMMA
modern usage like colon

1257 ETHIOPIAN COLON
modern usage like semicolon

1259 ETHIOPIAN NEW COMMA
modern usage

> Do syllable glyph variants ever occur distinctively within the same text, or are they merely font design choices like the glyph variants of Latin "a" or "g"?

ISSUES:

> In this design, no provision is made for coding the syllable glyphs; it is intended that they be *excluded* from Unicode/10646 BMP. If we learn that glyph variants may occur distinctively, then we may need to define some additional means for specifying glyph variants within plain text.

> Should we define an Ethiopian White Space character which can be easily guaranteed to have the same (minimum) width as U+1255 ETHIOPIAN WORDSPACE? Currently opinion is that this is unnecessary.

==========

==========

CHARACTER NAMES LIST

```
@               Consonant phonetic letters
1200    ETHIOPIAN CONSONANT H
1201    ETHIOPIAN CONSONANT L
1202    ETHIOPIAN CONSONANT HH
1203    ETHIOPIAN CONSONANT M
1204    ETHIOPIAN CONSONANT SZ
1205    ETHIOPIAN CONSONANT R
1206    ETHIOPIAN CONSONANT S
1207    ETHIOPIAN CONSONANT SH
1208    ETHIOPIAN CONSONANT Q
1209    ETHIOPIAN CONSONANT QH
120A    ETHIOPIAN CONSONANT B
120B    ETHIOPIAN CONSONANT V
120C    ETHIOPIAN CONSONANT T
120D    ETHIOPIAN CONSONANT C
120E    ETHIOPIAN CONSONANT X
120F    ETHIOPIAN CONSONANT N
1210    ETHIOPIAN CONSONANT NY
1211    ETHIOPIAN CONSONANT GLOTTAL
1212    ETHIOPIAN CONSONANT K
1213    ETHIOPIAN CONSONANT XX
1214    ETHIOPIAN CONSONANT W
1215    ETHIOPIAN CONSONANT NULL
1216    ETHIOPIAN CONSONANT Z
1217    ETHIOPIAN CONSONANT ZH
1218    ETHIOPIAN CONSONANT Y
1219    ETHIOPIAN CONSONANT D
121A    ETHIOPIAN CONSONANT DD
        Oromo
121B    ETHIOPIAN CONSONANT J
121C    ETHIOPIAN CONSONANT G
121D    ETHIOPIAN CONSONANT GG
        Bilen
121E    ETHIOPIAN CONSONANT TH
121F    ETHIOPIAN CONSONANT CH
1220    ETHIOPIAN CONSONANT PH
1221    ETHIOPIAN CONSONANT TS
1222    ETHIOPIAN CONSONANT TZ
1223    ETHIOPIAN CONSONANT F
1224    ETHIOPIAN CONSONANT P
1225
1226
1227
1228
1229
122A
122B
122C
122D
122E
122F

@               Vowel phonetic letters
1230    ETHIOPIAN VOWEL AE
1231    ETHIOPIAN VOWEL U
1232    ETHIOPIAN VOWEL I
1233    ETHIOPIAN VOWEL A
1234    ETHIOPIAN VOWEL E
1235    ETHIOPIAN VOWEL SCHWA
1236    ETHIOPIAN VOWEL O
```

```
1237    ETHIOPIAN VOWEL PHONETIC AE
        used primarily with U+1211 ETHIOPIAN CONSONANT GLOTTAL
1238    ETHIOPIAN VOWEL WAE
1239
123A    ETHIOPIAN VOWEL WI
123B    ETHIOPIAN VOWEL WA
123C    ETHIOPIAN VOWEL WE
123D    ETHIOPIAN VOWEL W
123E
123F

@               Numbers
1240
1241    ETHIOPIAN NUMBER ONE
1242    ETHIOPIAN NUMBER TWO
1243    ETHIOPIAN NUMBER THREE
1244    ETHIOPIAN NUMBER FOUR
1245    ETHIOPIAN NUMBER FIVE
1246    ETHIOPIAN NUMBER SIX
1247    ETHIOPIAN NUMBER SEVEN
1248    ETHIOPIAN NUMBER EIGHT
1249    ETHIOPIAN NUMBER NINE
124A    ETHIOPIAN NUMBER TEN
124B    ETHIOPIAN NUMBER TWENTY
124C    ETHIOPIAN NUMBER THIRTY
124D    ETHIOPIAN NUMBER FORTY
124E    ETHIOPIAN NUMBER FIFTY
124F    ETHIOPIAN NUMBER SIXTY
1250    ETHIOPIAN NUMBER SEVENTY
1251    ETHIOPIAN NUMBER EIGHTY
1252    ETHIOPIAN NUMBER NINETY
1253    ETHIOPIAN NUMBER HUNDRED
1254    ETHIOPIAN NUMBER TEN THOUSAND

@               Punctuation
1255    ETHIOPIAN WORDSPACE
1256    ETHIOPIAN COMMA
        modern usage like colon
1257    ETHIOPIAN COLON
        modern usage like semicolon
1258    ETHIOPIAN PERIOD
1259    ETHIOPIAN NEW COMMA
        modern usage
125A    ETHIOPIAN QUESTION MARK
        archaic
125B    ETHIOPIAN PARAGRAPH SEPARATOR
        archaic
================================================================


================================================================
BLOCK INTRODUCTION

Ethiopian  U+1200 -> U+125F


The Ethiopian script, which originally evolved for the archaic language Ge'ez, is
currently used to write several languages of Eastern Africa, including Amharic,
Tigre, and Oromo.  The script continues to be extended for writing languages that
have little tradition of printed typography; new characters to cover such
extensions may added to the standard later as definitive information about them
becomes available.
```

Encoding Principles. The visible glyphs of the Ethiopian script are not the objects shown in the encoding chart. The elements of the encoding are the alphabet *underlying* the script, thus the encoding is (roughly) phonetic rather than glyphic. These alphabetic letters are expected to be the units of keyboard input and all text representation short of rendering.

Rendering. Each visible glyph of the Ethiopian script represents a syllable rather than a single letter. The syllables can all be treated as simple (consonant + vowel) pairs, so that each glyph can be thought of as a *ligature* of two underlying letters. Thus the syllable "MA" would be represented in the encoding as U+1203 ETHIOPIAN CONSONANT M plus U+1233 ETHIOPIAN VOWEL A. The syllable glyphs themselves are not intended to be incorporated in this encoding. The individual consonant or vowel codes should not be isolated (i.e. unpaired) in normal final text, and their rendering in such circumstances is an option of the implementation. One possibility is to use special symbols for the individual letters, as is done in the code charts here.

Chart Symbols Individual Letters. Since the Ethiopian glyphs are normally syllabic, the script provides no unambiguous way of representing the underlying individual letters. Therefore in the code charts and names list, a convention has been adopted in which consonant letters are represented by their "first" form surrounded by a dotted circle, and vowel letters are represented by a typical glyph fragment attached to a dotted circle. This is not intended to imply direct glyphic composition of those forms, but merely to signify the underlying letters.

Encoding/Rendering of "First Form" Syllables. The circled consonants in the charts U+1200 -> U+1224 are underlying letters, they should not be confused with rendered full first form syllable glyphs. As with all glyphs in the script, the first form syllables are encoded as simple (consonant + vowel) pairs. Thus the glyph "MAE" would be represented in the encoding as U+1203 ETHIOPIAN CONSONANT M plus U+1230 ETHIOPIAN VOWEL AE. This pair would then be rendered via a "ligature" MAE whose appearance would resemble the chart symbol for U+1203 ETHIOPIAN CONSONANT M without the circle.

Encoding/Rendering of Lone Consonants ("Sixth Form" Syllables). The sixth form syllable glyphs are sometimes pronounced as though they were lone consonants (i.e. the vowel is dropped in speech), but this does not change their encoding. As with all glyphs in the script, the sixth form syllables are encoded as simple (consonant + vowel) pairs. Thus the spoken lone consonant "M" would be represented in the encoding as U+1203 ETHIOPIAN CONSONANT M plus U+1235 ETHIOPIAN VOWEL SCHWA.

Variant Glyph Forms. The script sometimes provides different glyph forms to represent the same syllables. It is assumed that these alternatives do not vary freely, in other words that is appropriate for a given font to contain only one selected glyph form for each syllable. Therefore no mechanism is provided for specifying glyph variants within a plain text stream of characters. The situation is analogous to that of the glyph variants of Latin "a" or "g".

Letter Names. The Ethiopian script often has multiple letters corresponding to the same Latin letter, making it difficult to assign unique Latin names. Therefore the names list makes use of certain devices (such as doubling a Latin letter in the name) merely to create uniqueness; this has no relation to the phonetics of the Ethiopian letters.

Encoding Order and Sorting. The order of the letters in the encoding is based on the traditional alphabetical order. This order differs from the sort order used for one or another language, if only because in many languages various pairs or triplets of letters are treated as equivalent in the first sorting pass. For example, an Amharic dictionary is likely to start out with a section headed by *three* letters:

U+1200 ETHIOPIAN CONSONANT H
U+1202 ETHIOPIAN CONSONANT HH
U+120E ETHIOPIAN CONSONANT X

Thus the encoding order cannot and does not implement a collation procedure for any particular language using this script.

Space Characters. The traditional word separator is U+1255 ETHIOPIAN WORDSPACE (:), but in modern usage a plain white wordspace is becoming common. The ASCII character U+0020 SPACE is suitable for the latter usage, although its (minimum) width is not guaranteed to be the same as that of the traditional wordspace.

Diacritical Marks. The mark U+030E NON-SPACING DOUBLE VERTICAL LINE ABOVE may occasionally be used to indicate emphasis or gemination. If this or other diacritical marks are used, they follow the vowel letter of the syllable to which they apply.

Encoding Structure. The Unicode block for the Ethiopian script is divided into the following ranges:

U+1200 -> U+1224	Consonant phonetic letters
U+1225 -> U+122F	Currently unassigned
U+1230 -> U+123D	Vowel phonetic letters (U+1239 is an intentional gap)
U+123E -> U+123F	Currently unassigned
U+1240 -> U+1254	Numbers (U+1240 is an intentional gap)
U+1255 -> U+125B	Punctuation
U+125C -> U+125F	Currently unassigned

==========