ETG Use of the Plane 0 PUA Area

A proposal for management of the entity portion of the PUA—Draft version 0.22

Peter Constable,
SIL Non-Roman Script Initiative (NRSI)

The Private Use Area (PUA) in plane 0 of Unicode is a somewhat limited resource. In order to manage it to the best benefit for all SIL entities, the 1998 CTC passed a motion requesting the NRSI to develop a plan for entities to follow. A draft of the NRSI's recommendations was presented at that conference [1] and is available on the IPub Resource Collection 98 CD-ROM [2]. That proposal allows entities to make free use of the lower portion of the PUA range while NRSI manages the upper portion for corporation-wide use.

This document describes a proposal for use of the entity-specific portion of the PUA by Ethiopia Group (ETG). We assume that this proposal would also be adopted by projects under other entities that make use of Ethiopic script, particularly by projects in Eritrea.

This proposal intends to allow for current work on international standardisation and existing known needs within ETG as well as past experiments in orthography, potential future needs, and the need to build real implementations using limited technologies.

Some references are made to Ethiopic characters that are already part of the Unicode standard. These statements are made in relation to version 3.0 of the standard. [5]

1. Entity-specific versus corporation-wide PUA sub-areas

The original NRSI proposal for overall management of the plane 0 PUA was to allow entities free use of the range U+E000–U+EFFF and for NRSI to manage U+F000–U+F8FF for corporation-wide use. This would provide 4,096 codepoints for field-entity-specific use and 2,304 for corporation-wide use.

In fact, a portion of the upper range may not be safe for us to use due to use by major software developers, such as Microsoft, Apple and Adobe; i.e. codepoints may exhibit undesirable behaviour in software developed by those companies due to proprietary use of those codepoints. (This is perfectly legal use of the PUA, though in the interest of end users one hopes that it will be kept to a minimum.) So, for example, NRSI already plans to avoid using U+F000–U+F0FF since these codepoints are used for “symbol-encoded” fonts in Microsoft software.

In addition to the U+F000–U+F0FF range, Apple has documented assignments that they have made in the range U+F800 through U+F8FF. Adobe has likewise documented assignments in the range U+F600 through U+F7FF. We also know that Microsoft has made use of other codepoints in the upper PUA range in some of their fonts for presentation forms, though these are probably less of a concern. (Actually, they've also used some of the U+Exxx range in some fonts, but again I don't think this use is a particular concern.) This is not a particular concern because the only software that would be aware of these assignments is Microsoft's Uniscribe rendering engine, and it would only use these codepoints as output, never as input. So, text that contains characters that use one of these codepoints would be unaffected. Furthermore, such use by Microsoft is likely only to be temporary as they implement OpenType fonts, and

so there is little likelihood that software will begin to appear that assumes definitions for those codepoints based upon their use in Microsoft fonts.

What is of greater concern is that Adobe has not only documented their use of codepoints but that they have also suggested that other major vendors cooperate with them in making assignments in the upper PUA range, and they have had some vendors such as HP interact with them in this way. (Such quasi-standardisation is *not* condoned by the Unicode standard.) This is a serious concern in that it could lead to commercial software that makes some assumptions about the meaning of these PUA characters, which would mean that other uses of those codepoints could meet with problems.

In addition, we know that software vendors including Microsoft have used some upper PUA codepoints for internal processing purposes within their code. Such uses make those codepoints particularly unreliable for encoding text.

As a result, the range that NRSI has to work with for corporation-wide characters is actually somewhat smaller than 2,304 codepoints. To be safe, we should probably try to avoid the ranges U+F000–U+F0FF and U+F700–U+F8FF. The remainder leaves us with 1,536 codepoints. This may be entirely adequate to meet the pan-corporation PUA needs, but it is hard to make any attempt at guessing whether that will be the case or not.

Because of this restriction, I would like to encourage entities to avoid using the range U+EC00–U+EFFF if they can at all avoid it. That would leave 1,024 codepoints available to NRSI for future expansion for corporation-wide characters should the need arise. Restricting use to U+E000–U+EBFF still provides 3,072 codepoints. This should be more than adequate for most entities' needs.

In accordance with this, I have limited the proposed use of the PUA by ETG to the range U+E000–U+EBFF.

2. Overall organisation of ETG's PUA space

2.1 Categories of need

The intention in this proposal is to manage a variety of past, present and potential future character needs in an organised manner. These needs fall into several categories:

- proposed extensions to Unicode and ISO/IEC 10646

There is a current proposal to add new characters to Unicode and ISO/IEC 10646. [3] Most of these are likely to end up being given standard (non-PUA) assignments within Unicode at some point. In the mean time, however, users need a way to work with them. Even for characters in the proposal that may prove to be unnecessary, there is a need to be able to encode these in documents for purposes of discussing the proposal.

In addition to the current proposal, there will likely be other proposals in the future. Characters in future proposals may already have been given assignments in the ETG PUA space. It will always be convenient to be able to arrange them as they have been presented in a proposal document. For these reasons, a number of codepoints should be set aside for representing characters proposed as additions to Unicode.

- characters that have been used or have been considered for some languages which, while not in Unicode or in any proposed extensions to Unicode, have reasonable potential for becoming candidates for addition to Unicode

Apart from what is already in Unicode or what has already been proposed for addition to Unicode, there are characters that we know are currently being used or have a good likelihood of being used in the future. These should be collected together. In addition, to allow for the likelihood that new characters will be

devised for new orthographies in the future, a number of codepoints should be set aside for such characters.

- glyphs that are known to be merely variants of other characters

There are glyphs that are known to be used but that are never distinguished from other existing characters; i.e. they are considered mere variants of existing characters. Eventually, font technologies should allow feature mechanisms that permit display of such variants without using explicit, distinct encoding. Until such technologies become commonly available, however, it is necessary to provide some means for displaying such glyphs. A block of codepoints should be reserved for this purpose.

- glyphs that may have been considered at one time for some language but which are not being promoted

There are some fidels that may have been considered as possibilities at some time for use in an orthography or that may have been mentioned in proposal documents on the basis of unclear evidence, but that are spurious and unlikely to actually be used in any orthographies. There is a need to encode such characters in documents simply for the purpose of discussing them. For this reason, a block should be set aside for such characters.

This category could also include some variants that may have marginal attestation in manuscripts and older books. (The degree of marginality would be the basis for deciding between this category and the previous category.)

- non-Ethiopic presentation forms

There are some non-Ethiopic characters, such as punctuation, for which Ethiopic-stylized glyphs may be desired in a font. At the same time, there may also be reasons for wanting to include regular (non-stylised) glyphs in the same font. As with fidel variants, font technologies may eventually provide feature mechanisms that handle such needs quite well. In the mean time, however, inclusion of alternate glyphs in a single font would require distinct codepoints, and a block should perhaps be set aside for this purpose.

(This need may also arise for other scripts used by other entities, and for that reason it may make better sense for these to be handled using corporation-wide PUA characters. This issue needs to be resolved as quickly as possible.)

2.2 Proposed organisation

I propose the following organisation of the ETG PUA space:

- U+E000–U+E1FF: proposed extensions to Unicode
- U+E200–U+E37F: characters with some potential for future addition to Unicode
- U+E380–U+E47F: fidel variant glyphs
- U+E480–U+E4FF: marginal or unattested fidel glyphs
- U+E500–U+E57F: miscellaneous presentation forms and non-fidel characters
- U+E580–U+E9FF: reserved
- U+EA00–U+EB7F: Latin presentation forms

Note that the categories associated with various blocks are intended to be guidelines for managing characters and glyphs in an organised manner and not as a rigid rule. In other words, we shouldn't be overly concerned if something is assigned in one block but later it is decided that it really should have been in another block. Of course, we should try to assign characters as carefully as we can. (In particular, the initial assignments described later in this document are a draft proposal only, which I expect to be

reviewed before being accepted.) But we should also limit the burden we place on ourselves to get everything perfect ahead of time. This organisation of the PUA is intended to help us and not to become a hindrance.

It will be noted that most or all of these ranges provide far more codepoints than are likely ever to be needed. This was done to ensure that any unforeseen future needs will have adequate space and because, given the nature of ETG's character needs, there is no particular concern for ETG running out of PUA codepoints. As it is, this proposal still leaves an undesignated block of 1280 codepoints reserved for future use.

The following sections consider the proposed blocks in further detail.

2.2.1 Proposed extensions to Unicode (U+E000– U+E1FF)

This block is initially subdivided as follows:

- U+E000–U+E07F: currently proposed extensions to Unicode (128 codepoints)

This corresponds to the actual number of columns used in the current proposal [3].

- U+E080–U+E1FF: reserved for future proposed extensions to Unicode (384 codepoints)

This allows for as many as 48 complete series.

If and when future proposals are submitted, columns from the latter range will be assigned to those proposals.

Note: it could be possible to reuse codepoints after characters have either been approved or rejected as additions to Unicode. This practice could lead to confusion, however, with the interpretation of older text becoming unclear. It would be better to allow more than enough room for potential needs and then to deprecate characters after a proposal has been either approved or rejected.

2.2.2 Characters with potential for future addition to Unicode (U+E200– U+E37F)

There is no clear measure provided by which to decide whether or not a character has potential for future addition to Unicode. Rather, such decisions must be made with some measure of subjectivity with the goal being that this area be used for characters that have some reasonable likelihood of ending up in actual use for some language and that are not understood as mere variants of other fidels.

This can include characters that have been or are being tentatively considered for use in some language, even if they do not end up being adopted for long-term use. If shapes are considered but are ultimately rejected in favour of different shapes, a judgement has to be made as to whether the former shapes are considered variants of the latter. They should perhaps only be considered variants for purposes of PUA assignment if they are commonly recognized as variants of the latter shapes. Otherwise, they should perhaps be assigned in this block as potential candidates for future assignment in Unicode. Again, though, no hard and fast rules are offered for making such judgements.

This block is divided into two sub-blocks:

- U+E200–U+E2FF: fidel series that are currently used or that are deemed to have some likelihood of being used in the future and are therefore potential candidates for addition to Unicode (256 codepoints)

This allows for up to 32 complete series.

- U+E300–U+E37F: individual characters and fidels that are currently used or that are deemed to have some likelihood of being used in the future and are therefore potential candidates for addition to Unicode (128 codepoints)

If fidels assigned in this block do come to be accepted within a user community, they could likely become the focus of a proposal, and in turn be added to the standard. This may lead to characters in this block

being duplicated first in the U+E080–U+E1FF range, and then later into Unicode. In such cases, the codepoints in this block would be deprecated and retired, but with the assignments left intact to avoid any potential confusion.

2.2.3 Fidel variant glyphs (U+E380– U+E47F)

This block should be the default for shapes that are commonly recognised to be mere variants of other fidel shapes. In some cases, however, this block could also be used for shapes that are variants of other fidels but which are not commonly recognised as such.

This block is divided into two sublocks:

- U+E380–U+E3FF: fidel series that are known variants of other existing series (128 codepoints)
- U+E400–U+E47F: individual fidels that are known variants of other existing fidels (128 codepoints)

These codepoints should eventually be made obsolete as font technologies evolve to provide widespread support for feature mechanisms and language-specific glyph variants.

2.2.4 Marginal or unattested fidel glyphs (U+E480–U+E4FF)

This block is not subdivided and contains 128 codepoints. As for other blocks, no rules are suggested for making judgements as to whether a shape is to be considered marginal or not.

2.2.5 Miscellaneous presentation forms and non-fidel characters (U+E500–U+E57F)

This block of 128 codepoints is effectively for any ETG-specific characters and glyphs that do not fit in any of the other designated blocks. It is assumed that a fidel will always fit into one of the other categories, and if cases arise in which a fidel shape requires an assignment but none of the other blocks is appropriate, then it is probably preferable to define a new block from the 1152 reserved codepoints in the range U+E580–U+E9FF.

2.2.6 Latin glyph-variant presentation forms (U+EA00–U+EB7F)

This block of 384 codepoints should eventually be made obsolete as font technologies evolve to provide widespread support for feature mechanisms and language-specific glyph variants.

The size of this block should easily allow for all of the Latin characters for which glyph variants are likely to be needed. These will generally be limited to just punctuation characters. The initial proposal for assignments (see below) allows for copies of most or all of the Basic Latin, Latin-1, General Punctuation and the Superscripts and Subscripts blocks, as well as 64 additional Latin glyph variants. These are organised as follows:

- U+EA00–U+EA5F for variants of the printable portion of the Basic Latin block (i.e. U+0020–U+007F)
- U+EA60– U+EABF for variants of the printable portion of the Latin-1 block (i.e. U+00A0–U+00FF)
- U+EAC0– U+EB2F for variants of the General Punctuation block (U+2000–U+206F)
- U+EB30– U+EB5F for variants of the Superscripts and Subscripts block (U+2070–U+209F)
- U+EB60–U+EB7F for other Latin glyph variants

As mentioned earlier, it is possible that this need should actually be handled within a range intended for corporate-wide assignments, and that this issue needs to be resolved as soon as possible.

3. Initial assignments to ETG's PUA space

For the purpose of creating a Unicode-conformant Ethiopic-script font, we have needed to begin assigning ETG-specific characters to the PUA space. The proposed assignments are shown in tables on the following pages. Comments on some of the proposed characters are made here.

There is one issue that has not been addressed in this initial proposal for PUA assignments: defining character semantics. This has been done for the proposed extensions to Unicode, but not for the remaining characters. For every character assigned to the ETG PUA block, there should be information regarding the Unicode semantics of that character. Typically, most of the semantics of new fidels will be the same and will match those for fidels already in Unicode. A unique name must be assigned to each character, however. These names should be conformant to the naming rules employed in Unicode and ISO/IEC 10646.

The short, labialised series at U+E200 and U+E498 have been organised in a manner like similar series such as those at U+1248 and U+1258, and not like the short series for Gurage that are found in the proposed extensions at U+E020, U+E028, U+E030 and U+E038.

It is my understanding that the series with ሁ at U+E200 has been used for Gurage. This series is located in the block for potential candidate series on the assumption that these fidels are in use and are not variants of other fidels already in Unicode or in the proposed extensions.

The fidel at U+E300, ፞, was at one time in the initial proposal for adding Ethiopic characters to Unicode, but for some reason was rejected for version 3.0. It was not known to us until we saw the Unicode proposals. I have given it an assignment as a potential candidate for addition to Unicode based on the assumption that someone must have had a reason for putting it in the initial proposal, and since it is an appropriate shape for a labialised addition to the series at U+12E8–U+12EE and so would be what would most likely be used should such a form be needed. On the other hand, it may be that there was no factual basis for its inclusion in early Unicode proposals and that it was rejected for lack of attestation. It might, therefore, equally make sense to treat this as a marginal/unattested character.

It is unclear to me why the character at U+E308 for 1000, ፸, was not included in Unicode since I have been told that it is attested in manuscripts and books. If that is indeed the case, then it is certainly a potential candidate for addition to Unicode.

Tsigie et al [4] report that the combining "gemination mark" at U+E309, ◌̤, is in modern usage in linguistic works such as dictionaries. It is, therefore, considered a potential candidate for future acceptance in Unicode.

The fidel at U+E30A, "yeneged meleket", is reported by Tsigie et al [4] to have been recently introduced by the Ethiopian government as an Ethiopic counterpart to our copyright and trademark symbols. Assuming its continued acceptance within Ethiopia, it is a likely future candidate for acceptance in Unicode, and I have assigned it a PUA codepoint as such.

It is my understanding that the series at U+E380 with ጇ is a known variant for that at U+12F8 with ኗ and that it would not ever be used contrastively with that series. On this basis I assigned it to the block for known variants.

The fidels at U+E400, U+E402 and U+E403 have been in our inventory for some time, and are listed in our glyph database as fourth form alternates. There are some questions about each of these, however.

First, the labialised nua, ኌ, does appear to be a variant, but of an existing labialised form, U+1297 ኗ, and not of a fourth form. Since it's actual usage is not clear to me, I'm not sure if I've got it in the right block.

Secondly, with regard to the fidel at U+E402, ጟ, both Pete Unseth and Daniel Yacob remarked that this glyph seems like a strange sixth form variant. Neither was aware of any languages that use it. I am wondering if it should be treated as marginal.

Thirdly, with regard to U+E403 ᐓ, Pete suggested that this glyph should be what is used as the fourth form in the series at U+12C0 (i.e. U+12C3) rather than what we have now, viz. ᐔ. If that is the case, the question would remain as to whether the other glyph should be considered as a normal variant or as marginal.

It is my understanding that the fidel at U+E401, ᐡ, is a variant for U+1313, ሓ, or that it is superseded by the series at U+1310–U+1316. For this reason, I have assigned it a PUA codepoint among known variants.

The fidel at at E404, ᐕ, was at one time in the initial proposal for adding Ethiopic characters to Unicode. Information was later available (e.g. Tsigie et al [4]) indicating that this should be considered a variant of and is superseded by the xx78 series with ᐦ in the currently proposed extensions (at U+E078–U+E07D in our proposed PUA assignments). I have therefore considered this a variant fidel. There may be other reasons for treating this as a marginal/unattested character, however.

The glyph at U+E405, ᐣ, was first encountered by us in the current draft extensions proposal for the character at xx34—U+E034 in these proposed PUA assignments. Since the draft was completed, Daniel Yacob indicated to me that this glyph is the traditional form but that currently the glyph we have shown at U+E034, ᐤ, is preferred for that character. The glyph at U+E406, ᐥ, was another variant for this character that has been in our collection for some time, but apparently the glyph at U+E034 is very much preferred over this.

I assigned the character at U+E408, ᐧ, on the assumption that it is a common variant for U+1368 ፨.

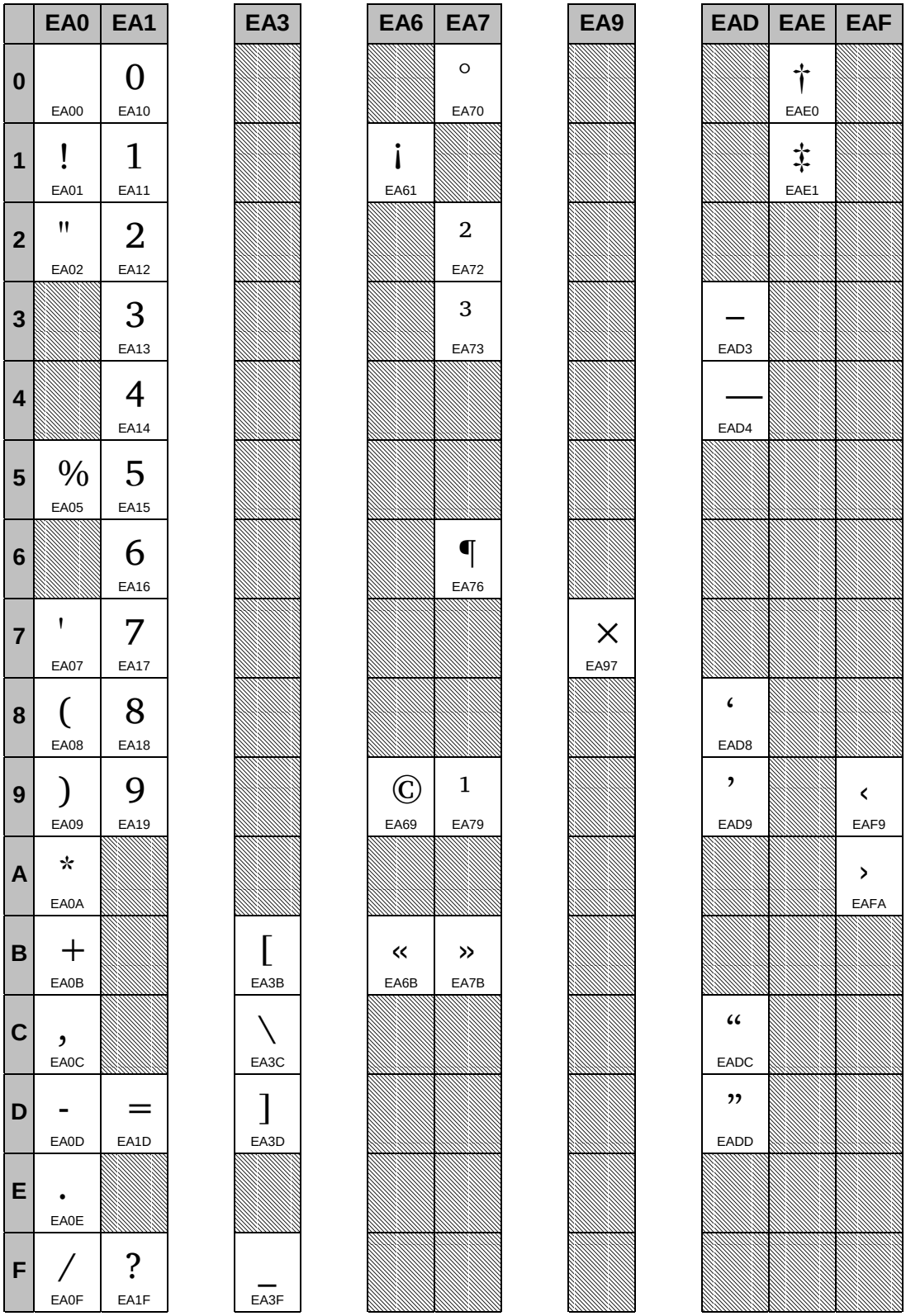
I have not considered the individual fidel variants at U+E400–U+E408 in relation to other variants I have seen discussed to evaluate whether these individual fidel variants need more careful organisation than I have applied here.

The series with ᐬ at U+E480 was first known to us in an early version of the first proposal for adding Ethiopic characters to Unicode. We don't know on what basis they were included in that proposal, but they were rejected. Pete Unseth indicated to us that these fidels seem unlikely ("How do you labialise a /w/?"). For these reasons, I have considered these to be marginal characters.

The five glyphs at U+E490–U+E49A were put there simply because the status of each of these is completely unknown to me.

	E00	E01	E02	E03	E04	E05	E06	E07
0	E000	E010	E020	E030	E040	E050	E060	E070
1	E001	E011	E021	E031	E041	E051	E061	E071
2	E002	E012	E022	E032	E042	E052	E062	E072
3	E003	E013			E043	E053	E063	
4	E004	E014	E024	E034		E054	E064	
5	E005	E015	E025	E035		E055	E065	
6	E006	E016				E056	E066	
7								
8	E008	E018	E028	E038	E048	E058	E068	E078
9	E009	E019	E029	E039	E049	E059	E069	
A	E00A	E01A	E02A	E03A	E04A	E05A	E06A	E07A
B	E00B	E01B			E04B	E05B	E06B	E07B
C	E00C	E01C	E02C	E03C	E04C	E05C	E06C	E07C
D	E00D	E01D	E02D	E03D	E04D	E05D	E06D	E07D
E	E00E	E01E			E04E	E05E	E06E	
F								

	E20	E30	E38	E40	E48	E49	E50
0	 E200	 E300	 E380	 E400	 E480	 E490	. E500
1			 E381	 E401		 E491	· E501
2	 E202		 E382	 E402	 E482		, E502
3	 E203		 E383	 E403	 E483		/ E503
4	 E204		 E384	 E404	 E484		- E504
5	 E205		 E385	 E405	 E485		– E505
6			 E386	 E406			E506
7							E507
8		 E308		 E408		 E498	
9		 E309				 E499	
A		 E30A				 E49A	
B							
C							
D							
E							
F							


EA0
EA1
EA3
EA6
EA7
EA9
EAD
EAE
EAF
0
1
2
3
4
5
6
7
8
9
A
B
C
D
E
F
EA00
EA10
EA70
EAE0
EA01
EA11
EA61
EAE1
EA02
EA12
EA72
EAD3
EA13
EA73
EAD4
EA14
EA05
EA15
EA76
EA97
EA07
EA17
EAD8
EA08
EA18
EA69
EA79
EAD9
EAF9
EA09
EA19
EAFA
EA0A
EA6B
EA7B
EADC
EA0B
EA3B
EADD
EA0C
EA3C
EA0D
EA1D
EA3D
EA0E
EA0F
EA1F
EA3F

4. References

[1] Hosken, Martin. 1998. "PUA corporate strategy: A discussion on the organization of the PUA." *SIL IPub Resource Collection 98* (CD-ROM). Dallas: SIL International.

[2] International Publishing Services, 1998. *Resource Collection 98 CD*. Dallas: SIL International.

[3] "Proposal to encode Ethiopic extensions in the BMP of ISO/IEC 10646." (ISO/IEC JTC1/SC2/WG2 document N1846.) Available at http://anubis.dkuug.dk/jtc1/sc2/wg2/docs/n1846.pdf.

[4] Tsigie, Asteraye; Berhanu Beyene; Daniel Aberra; and Daniel Yacob. 1999. "A roadmap to the extension of the Ethiopic writing system standard under Unicode and ISO-10646." Paper presented at the Fifteenth International Unicode Conference, San Jose, California, September 1–2, 1999. Available in the conference proceedings.

[5] The Unicode Consortium. 2000. The Unicode standard. Version 3.0. Reading, MA: Addison-Wesley. (Information on the Ethiopic characters included in Unicode version 3 are also available at http://charts.unicode.org/PDF/U1200.pdf.)