L2/01-309
8 Aug. 2001

Variation Selectors and Han

John H. Jenkins
Apple Computer, Inc.

The UTC has been talking about adding variation selectors to the standard for some time to help with the proliferation of glyphs in Unihan, and now that variation selectors are actually being added to the standard, this matter should really be finalized.

The impetus is coming from a number of directions:

There are a number of characters in Unihan which can have different glyphic shapes which are generally recognized. When one of these is used in a personal name, the person very frequently will insist that one shape be used and not another for their name. (This is analagous to insisting on the spelling of names like "Jon", "Marc", "Ric", and "Cooke".) It is generally desired that such a distinction be possible within a plain-text context, such as program source code, email, or database fields.

There are also a number of standard glyph sets used in East Asia (such as AJ-14 or the Morisawa collection). People wishing to map data using these fonts may desire round-trip compatibility without the overhead of having to maintain synonym lists for Unihan characters which should be treated identically under most circumstances. (We already have a number of such characters in Unihan, such as U+8AAA and U+8AAC. We don't want to add to them.)

Finally, as research is continuing on new characters that need to be standardized, a large collection numbering in the tens of thousands is being produced by South Korea, most of which are variants of standard characters. These are needed to represent standard religious texts, such as the Tripitaka, a collection of sacred Buddhist texts. Whether or not such a distinction needs to be made in plain text is debatable; what is *not* debatable is that these characters/glyphs are coming down the pike.

What we're proposing is the following:

1) A block of variation selectors (256) be allocated in Plane 14.

2) The IRG be charged with maintaining data on which base character/variation marker maps to which character. This needs to be matched not only with a specific glyph, but also (if possible) with a specific source reference, such as a glyph ID within an AJ-14 font.

3) Unicode create and maintain a database of information on variations for public use. At the minium, this would mean a Web page that people can go to to see what's what.

The basic criterion of when something should be encoded using a variation marker is whether or not it is a Z-variant of another character already in Unihan. This seems clean and straightforward and is consistent

with the unification model the IRG already uses.

So far as the semantics are concerned, this is already covered in L2/01-268. To a font designer, the presence of the variation marker in text indicates a preferred choice for a glyph, if that glyph is available. An intelligent font model like AAT or OpenType can be used to make the selection.

This is just a preliminary sketch of what needs to be done. It's presented as a starting point for discussions within the UTC.