L2/16-162

From: **Mark Davis**
To: **UTC**
Date: **2016-05-12**
Re: **Cleanup of constraints on variation sequences**
Draft: ***link***

The original purpose of the restrictions on the initial character of a variation sequence was so that canonically equivalent sequences would never have the VS apply to different characters because of reordering or decomposition, and to make sure that the character to which the VS applies is always unambiguous. As the text says: "to prevent problems in the interpretation of such sequences in normalized text"

However, it has become clear that the original formulation, with the restriction to *base* characters, is overly strict, and artificially imposes constraints on characters changing general categories, and forces special cases into the formation to accommodate recent additions to the standard. In addition, the structure of the text separates an important part of the definition ("The initial character in a variation sequence is never … a canonical decomposable character") from the main statement of the definition, placing it several paragraphs below.

For clean, reliable guidance to implementers, we should not hack up the definition with special cases: instead we should define it in terms of the actual constraints that are required to "to prevent problems in the interpretation of such sequences in normalized text". This document proposes the following changes to the draft 9.0 text (allowing of course for editorial rewording).

I believe it is better to allow ourselves some leeway in the conditions, to prevent problems where we need to change a character from GC=Mc to GC=Mn, but would be prevented by the existence of a standardized variation sequence. It doesn't mean that we need to take advantage of that for normal text!

23.4 Variation Selectors

Characters in the Unicode Standard can be represented by a wide variety of glyphs, as discussed in Chapter 2, General Structure. Occasionally the need arises in text processing to restrict or change the set of glyphs that are to be used to represent a character. Normally such changes are indicated by choice of font or style in rich text documents. In special circumstances, such a variation from the normal range of appearance needs to be expressed side-by-side in the same document in plain text contexts, where it is impossible or inconvenient to exchange formatted text. For example, in languages employing the Mongolian script, sometimes a specific variant range of glyphs is needed for a specific textual purpose for which the range of "generic" glyphs is considered inappropriate.

Variation selectors provide a mechanism for specifying a restriction on the set of glyphs that are used to represent a particular character. They also provide a mechanism for specifying variants, such as for CJK ideographs and Mongolian letters, that have essentially the same semantics but substantially different ranges of glyphs.

Variation Sequence. A variation sequence always consists of ~~a base character or a spacing mark (gc=Mc) followed by a single variation selector character~~ a sequence of two characters, where the final character is a variation selector character. There are some additional constraints on the initial character:

1. It must have a zero canonical combining class

2. It must not be a variation selector
3. It must not be a canonical decomposable character
4. It must not have a General Category value of *Other* (Cc | Cf | Cs | Co | Cn)

These constraints are required because it is important that variation sequences remain stable under normalization, and that the effects of variation selector can always be characterized as unambiguously applying to a single character. Versions of the Unicode Standard prior to version 9.0 had a more limited statement of constraints on variation sequences.

That two-element sequence is referred to as a variant of the ~~base character or spacing mark~~ initial character. For simplicity of exposition, the following discussion only mentions base characters; variation sequences involving ~~spacing marks~~ non-base characters are uncommon, but otherwise behave similarly.

[Ed note: instead of the final sentence, we could replace "base character" by "initial character" in the rest of the section.]

In a variation sequence the variation selector affects the appearance of the base character. Such changes in appearance may, in turn, have a visual impact on subsequent characters, particularly combining characters applied to that base character. For example, if the base character changes shape, that should result in a corresponding change in shape or position of applied combining marks. If the base character changes color, as can be the case for emoji style variation sequences, the color may also change for applied combining marks. If the base character changes in advance width, that would also change the positioning of subsequent spacing characters.

In particular, the emoji style variation sequences for digits and U+0023 "#" number sign are intended to affect the color, size, and positioning of U+20E3 0 combining enclosing keycap when applied to those base characters. For example, the variation sequence <0023, FE0F> selects the emoji style variant for "#". The sequence <0023, FE0F, 20E3> should show the enclosing keycap with an appropriate emoji style, matching the "#" in color, shape, and positioning. Shape changes for variation sequences, with or without additional combining marks, may also result in an increase of advance width; thus, each of the sequences <0023, FE0F>, <0023, 20E3>, and <0023, FE0F, 20E3> may have a distinct advance width, differing from U+0023 alone.

The variation selector is not used as a general code extension mechanism; only certain sequences are defined, as follows:

> Standardized variation sequences are defined in the file StandardizedVariants.txt in the Unicode Character Database. Ideographic variation sequences are defined by the registration process defined in Unicode Technical Standard #37, "Unicode Ideographic Variation Database," and are listed in the Ideographic Variation Database. Only those two types of variation sequences are sanctioned for use by conformant implementations. In all other cases, use of a variation selector character does not change the visual appearance of the preceding base character from what it would have had in the absence of the variation selector.

~~The initial character in a variation sequence is never a nonspacing combining mark (gc=Mn) or a canonical decomposable character. These restrictions on the initial character of a variation sequence are necessary to prevent problems in the interpretation of such sequences in normalized text.~~

The variation selectors themselves are combining marks of combining class 0 and are default ignorable. Thus, if the variation sequence is not supported, the variation selector should be invisible and ignored. This does not preclude modes or environments where the variation selectors should be given visible appearance.

For example, a “Show Hidden” mode could reveal the presence of such characters with specialized glyphs, or a particular environment could use or require a visual indication of a base character (such as a wavy underline) to show that it is part of a standardized variation sequence that cannot be supported by the current font.

The standardization or support of a particular variation sequence does not limit the set of glyphs that can be used to represent the base character alone. If a user requires a visual distinction between a character and a particular variant of that character, then fonts must be used to make that distinction. The existence of a variation sequence does not preclude the later encoding of a new character with distinct semantics and a similar or overlapping range of glyphs.