L2/23-286

Unicode core spec improvements for variation selectors

Markus Scherer, Asmus Freytag, and other PAG members; 2023-dec-27

For

152-A5a	Ken Whistler, Mark Davis	EDC	Draft a new section for Chapter 3 on variation selectors and variation sequences, for Version 11.0. (retargeted to 13.0, 14.0, 15.0)			15.0 retargeted to 16.0	Unicode text

(which replaced 147-A137)

and

166-A61	Markus Scherer, Norbert Lindenberg	EDC	Propose changes to the specification of variation sequences in TUS chapter 23.4 and appropriate additions to chapter 3, based on document L2/21-012 item D2. The intent is to clarify the restrictions on initial characters in order to avoid issues under normalization. Include examples of characters and sequences that are excluded. See also action item 152-A5a.	L2/21-012		15.0 retargeted to 16.0	Unicode text

See

- L2/L2021/21012-utc166-properties-recs.pdf = UTC #166 properties feedback & recommendations
- Unicode15.0.0/ch03.pdf
- Unicode15.0.0/ch23.pdf

Changes in chapter 3

D56 *Combining character sequence*

Change

- Combining character sequences involving a variation selector (which is both default_ignorable and a combining mark), typically consist of only the base character followed by a single variation selector. (See *Section 23.4, Variation Selectors*.) In limited cases, a variation selector may also apply to a spacing combining mark (gc=Mc), which in turn is part of a longer combining character sequence. Placing a random variation selector inside a longer combining character sequence would create a sequence that is formally still a combining character sequence, but that would not contain a conformant variation sequence. However, occasionally a variation selector in a conformant variation sequence may be followed by another combining mark. For example, the sequence <0030, FE00, 20E3> represents a variant form of the digit zero, followed by an enclosing keycap.

to

- ~~Combining character sequences involving a variation selector (which is both default_ignorable and a combining mark), typically consist of only the base character followed by a single variation selector.~~ A two-character sequence consisting of an initial graphic character followed by a variation selector, and satisfying additional constraints, is a variation sequence. ~~(~~See *Section 23.4, Variation Selectors*.~~) In limited cases, a variation selector may also apply to a spacing combining mark (gc=Mc), which in turn is part~~ Because any variation selector is a combining character, a variation sequence is either a combining character sequence, or it is a subsequence of a longer combining character sequence. ~~Placing a random variation selector inside a longer combining character sequence would create a sequence that is formally still a combining character sequence, but that would not contain a conformant variation sequence. However, occasionally a variation selector in a conformant variation sequence may be followed by another combining mark.~~ For example, the sequence <0030, FE00, 20E3> represents a variant ~~form~~ of the digit zero, followed by an enclosing keycap. A variation sequence can be a non-initial subsequence within a combining mark sequence. For example, the sequence <1000, FE00, 1031, FE00> is a single combining mark sequence with two variation sequences representing variants of the base character MYANMAR LETTER KA and the combining mark MYANMAR VOWEL SIGN E.

Changes in chapter 23

(This subsumes the changes proposed in L2/16-162 “Cleanup of constraints on variation sequences” which triggered 147-A137.)

23.4 Variation Selectors

a) Change

> ***Variation Sequence***. A variation sequence always consists of a base character or a spacing mark (gc = Mc) followed by a single variation selector character. That two-element sequence is referred to as a *variant* of the base character or spacing mark. For simplicity of exposition, the following discussion only mentions base characters; variation sequences involving spacing marks are uncommon, but otherwise behave similarly.

to

Variation Sequence. A variation sequence ~~always consists of a base character or a spacing mark (gc = Mc) followed by a single variation selector character.~~ is a two-character sequence in which a variation selector follows an initial character. ~~That two-element sequence is referred to as a *variant* of the base character or spacing mark.~~ Each variation sequence defines a *variant* of the initial character. ~~For simplicity of exposition, the following discussion only mentions base characters; variation sequences involving spacing marks are uncommon, but otherwise behave similarly.~~ The initial character must have the following properties:

- ○ Graphic Character (TUS chapter 3 D50: gc=L, M, N, P, S, Zs)
- ○ Not a Variation_Selector
- ○ ccc=0 (does not reorder)
- ○ NFD_QC=Yes (does not decompose)
- ○ NFC_QC=Yes (does not get consumed in composition)

For example, the following types of characters cannot be initial characters of variation sequences: Control codes, format controls, most diacritics, Indic vowels, viramas, Hangul Jamo medial vowels, and canonical composite characters.
[Ed note: These are examples of excluded characters; not sure whether we need to write something about excluded sequences — seems obvious.]

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 16.0 had a more limited statement of constraints on variation sequences.
[Review note: This allows characters with a wider set of General_Category values than before (e.g., now including Mn) as long as they have the other properties (e.g., ccc=0). See L2/20-244R “the restriction against nonspacing combining marks is both too loose and too restrictive to meet its goal” etc.]

b) In this whole section 23.4, change “base character” to “initial character”.

c) Remove the later paragraph (whose contents have been pulled up):

~~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.~~