

Feedback on L2/23-286, “Unicode core spec improvements for variation selectors”

Norbert Lindenberg, 2024-01-19

The changes proposed in L2/23-286 significantly improve the specification of variation sequences, both in their technical substance and in clarity.

Two comments though.

① The use of the properties `NFD_QC` and `NFC_QC` to define the restrictions on initial characters does not help. These properties are not discussed at all in the Core Specification. They are described in UAX 44 as “defined to enable various optimizations for implementations of normalization”, and discussed in UAX 15 in the section “Detecting Normalization Forms”. Neither UAX 44 nor UAX 15 nor the data file `DerivedNormalizationProps.txt` specify how they are derived from the properties that are discussed in the Core Specification. This makes it unnecessarily hard to understand why these restrictions on initial characters exist, and whether these restrictions are the right ones.

I suggest using properties that are defined in the Core Specification to specify the restrictions on initial characters.

② Some of the examples for types of characters that “cannot be initial characters of variation sequences” need to be qualified:

- “Indic vowels” should be “some Indic vowels”. Of the 1227 characters with `InSC` values `Vowel_Independent`, `Vowel_Dependent`, or `Vowel` in Unicode 15.1, 1129 are eligible as initial characters based on the specified criteria. Even when just looking at the 697 dependent vowels, the vast majority, 607, are eligible. Among them is U+1031 MYANMAR VOWEL SIGN E, which is used as an example in the proposed new text for D56.
- “Viramas” should be “most viramas”. The Kirat Rai script, accepted for Unicode 16, includes two viramas, U+16D6B KIRAT RAI SIGN VIRAMA and U+16D6C KIRAT RAI SIGN SAAT, that have `ccc=0` and therefore appear to be eligible (unless the as-yet unavailable `NFD_QC` and `NFC_QC` values decide otherwise).

၅၅၀၅၅