A Letter in Support of N5266 “Proposal to Encode a Set of 128 User-Defined Variation Selectors”

Daniel Yacob, 2024-07-22

Dear Members of JTC1/SC2/WG2 and the UTC,

This letter is written in support of a mechanism to be made available for end-users to define private variation sequences.

The N5266 authors have aptly described the use cases for individual users and research groups to manage community registries of variation sequences. Their reasoning aligns perfectly with my own experience in managing nearly seven hundred historic sequences under the EMUFI [1] project. Accordingly, I can emphasize the need for a formalized capability for defining non-standard variation sequences.

The fifth answered question in the Unicode Consortium’s “Variation Sequences” FAQ[2] explains that user defined sequences *cannot* be created and that *“Private use characters should be used instead.”* While the private use area (PUA) provides an excellent solution for defining isolated private symbols as-needed, it is problematic when working with symbols where a *variant* is the appropriate character model. Variants are different from other symbol types in that they are by definition lexically equivalent to a reference (or base) glyph. Variation selectors are a means to maintain the identity and character properties of the reference glyph allowing text search, collation, indexing, and other text analytics to continue working as expected.

This association with the reference symbol is lost when a new glyph is given a separate code point under the PUA and the textual tools that researchers depend upon become unusable without new and specialized programming. Which in turn may not be a viable undertaking for the user community.

The recommendation put forth in N5266 to introduce 128 additional variation selectors seems unfortunate but is also avoidable. The primary intent of the PUA is to allow end-users to define their own symbols while avoiding any impact on the standardized inventory. I believe this objective is maintained when a private sequence (e.g. <1362, FE06>) does not collide with a definition in a standard registry. Such sequences are undefined and could *only* be interpreted under a privately defined context. Thus, as an alternative to the proposed solution, I would favor allowing the existing inventory of variation selectors to be leveraged by the end-user community. Albeit with the restriction that standard sequences not be duplicated. This restriction is in keeping with forbidding end-users from overriding the standard-encoded character code points and simply extends the notion to an encoded tuple.

Thank you for considering this alternative approach.

Daniel Yacob

[1] Ethiopic Manuscript Unicode Font Initiative
[2] https://www.unicode.org/faq/vs.html