L2/18-073

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: Working Group Document
Title: Proposal to add standardized variation sequences for various punctuation
Author: Ken Lunde (Adobe Systems Incorporated)
Status: Corporate Full Member Contribution
Action: For consideration by the UTC
Date: 2018-03-05

Background

This proposal is a stripped-down version of L2-18-013, which is the second part of a split version of L2/17-056 that was originally discussed during UTC #138 in early 2014 as L2/14-006. L2/17-056 itself was discussed during UTC #153 in late 2017 for the purpose of soliciting feedback that led to the split proposal. The first part was previously submitted as L2/17-436 that was discussed during UTC #154 in early 2018, and which resulted in 16 SVSes (*Standardized Variation Sequences*) being accepted for Unicode Version 12.0.

Regional conventions affect how particular characters, such as punctuation, should display, and for the characters within the scope of this proposal, the general difference is whether they are aligned to Western typographic attributes, such as the baseline, x-height, or cap-height, or to the em-box for East Asian use. The fundamental issue is that the glyphs for these characters share the same Unicode code point, meaning that an explicit font change or layout feature invocation (such as the OpenType 'locl' GSUB feature) must be used to specify or distinguish them, which is not possible in “plain text” environments.

Although “rich text” environments are becoming more common, including those that support language-tagging and the OpenType 'locl' GSUB feature, “plain text” environments persist, and are likely to continue to persist for a long time due to their robust nature. In addition, environments that support variation sequences outnumber those that support language-tagging.

Proposal Summary

This document is a modest proposal for adding 12 standardized variation sequences (SVSes) for six characters that use VS1 and VS2 (aka U+FE00 and U+FE01) to distinguish between the forms, whose usage varies according to well-established Western or East Asian conventions.

Characters With Ambiguous Alignment or Width

This proposal covers six punctuation characters whose shapes are generally the same regardless of regional conventions, but whose alignment or width can vary by region. Western typographic conventions require that these characters are aligned to or centered on the baseline, x-height, or cap-height. In contrast, East Asian typographic conventions require that these characters are aligned to or centered within the em-box, and in most cases should be fullwidth.

It is true that East Asian punctuation characters are generally fullwidth, though regional conventions may vary for some or most of them. For example, Japanese and Korean tend to use non-fullwidth smart quotes. Furthermore, all of the characters referenced in this proposal have the *East Asian Width* (see UAX #11) property value “A” (*East Asian Ambiguous*), which means that there is no universal or reasonable default form, and therefore require SVSes for both specified use cases.

While Pan-CJK fonts, such as those of the open source *Source Han* and *Noto CJK* typeface families, tend to include glyphs for Western and multiple East Asian regional conventions for particular characters, single-region East Asian fonts are beginning to include both Western and East Asian glyphs for the same characters, including the ones that are included in this proposal.

Standardized Variation Sequences

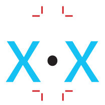
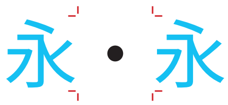
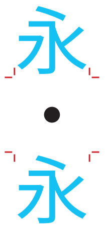
Standardized variation sequences offer a solution to this ambiguity by using variation selectors to specify alignment or glyph-width conventions on a per-character basis. A font with appropriate entries in its Format 14 (*Unicode Variation Sequences*) 'cmap' subtable can enable these distinctions to be shown and preserved in “plain text” environments.

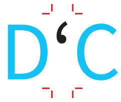
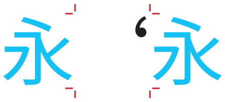
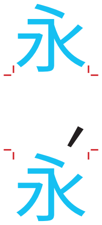
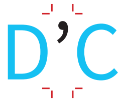
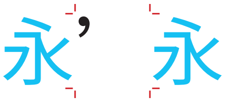
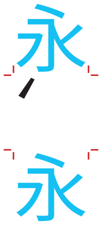
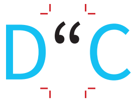
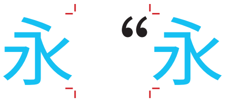
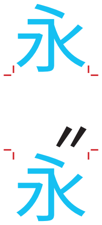
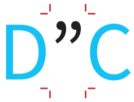
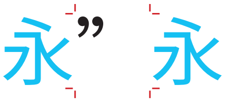
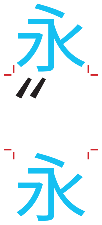
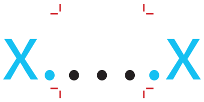
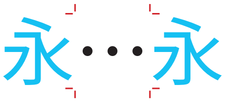
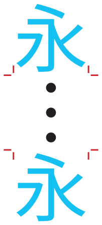
Below is a complete list of the 12 proposed standardized variation sequences as they would appear in the UCD’s *StandardizedVariants.txt* file:

```
# Non-fullwidth, centered fullwidth, corner-justified fullwidth, or baseline-aligned form
variation sequences

00B7 FE00; non-fullwidth form;                # MIDDLE DOT
00B7 FE01; centered fullwidth form;           # MIDDLE DOT
2018 FE00; non-fullwidth form;                # LEFT SINGLE QUOTATION MARK
2018 FE01; corner-justified fullwidth form;   # LEFT SINGLE QUOTATION MARK
2019 FE00; non-fullwidth form;                # RIGHT SINGLE QUOTATION MARK
2019 FE01; corner-justified fullwidth form;   # RIGHT SINGLE QUOTATION MARK
201C FE00; non-fullwidth form;                # LEFT DOUBLE QUOTATION MARK
201C FE01; corner-justified fullwidth form;   # LEFT DOUBLE QUOTATION MARK
201D FE00; non-fullwidth form;                # RIGHT DOUBLE QUOTATION MARK
201D FE01; corner-justified fullwidth form;   # RIGHT DOUBLE QUOTATION MARK
2026 FE00; baseline-aligned form;             # HORIZONTAL ELLIPSIS
2026 FE01; centered fullwidth form;           # HORIZONTAL ELLIPSIS
```

The table below demonstrates an actual implementation—using a fully-functional OpenType/CFF font with an appropriately-built Format 14 'cmap' subtable that specifies the UVSes (*Unicode Variation Sequences*) that correspond to the proposed standardized variation sequences. This OpenType/CFF font is also attached to this proposal, and can be easily extracted and used. Example vertical forms of the fullwidth glyphs that follow Japanese conventions are shown in the last column of the table. The second and third columns of the table use VS1 and VS2 as described in this proposal. Red registration marks are used to draw attention to how their glyphs are typically aligned within the em-box, with prototypical characters surrounding them, with 永 indicating typical East Asian usage.

Unicode	VS1	VS2—Horizontal	VS2—Vertical—Japanese
U+00B7			

Unicode	VS1	VS2—Horizontal	VS2—Vertical—Japanese
U+2018			
U+2019			
U+201C			
U+201D			
U+2026			

Rationale & Conclusion

This proposal addresses the varying regional conventions for a modest number of punctuation characters, which is a real-world issue for Pan-CJK fonts that support multiple East Asian languages and regions, especially in “plain text” environments with limited font-selection capability, or in environments that lack support for per-character language-tagging. Issues arise when mainstream fonts that include both proportional (for Western or East Asian use) and fullwidth (for East Asian use) forms of the same character, and whereby the possibility of use in the same document is relatively high.

It is worthwhile to point out that many of the characters covered by this proposal have been problematic for both developers and their customers for years.

That is all.