L2/18-013

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: Working Group Document
Title: Proposal to add standardized variation sequences for digits and various punctuation
Author: Ken Lunde (Adobe Systems Incorporated)
Status: Corporate Full Member Contribution
Action: For consideration by the UTC
Date: 2018-01-08

Background

This proposal is the second part of a split versions 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 this and the first part of the split proposal. The first part was previously submitted as L2/17-436.

Regional conventions affect how particular punctuation, to include digits, 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 proposal for adding 63 standardized variation sequences (SVSes) for 43 characters that use VS1 through VS3 (aka U+FE00 through U+FE02) 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 43 digit and 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 some 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 tends to use ASCII digits and non-fullwidth quotes, and Korean uses ASCII digits and punctuation, along with non-fullwidth quotes. The Western forms of these characters, in particular those that are within the scope of ASCII, are unambiguously narrow according to *East Asian Width* (see UAX #11) and therefore represent a reasonable default, meaning that SVSes are necessary only to specify non-Western usage.

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, such as the ones that are included in this proposal.

Standardized Variation Sequences

Standardized variation sequences offer a solution to this glyph-level alignment ambiguity by using variation selectors to specify these 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 63 proposed standardized variation sequences as they would appear in the UCD's *StandardizedVariants.txt* file (those lines that are highlighted in **bold** correspond to characters for which only a single SVS is proposed, because there is a reasonable default form whose use generally follows Western typographic conventions):

```
# Em-box, baseline, x-height, and cap-height aligned form variation sequences

0021 FE01; em-box centered form;                  # EXCLAMATION MARK
0022 FE01; em-box top aligned form;               # QUOTATION MARK
0027 FE01; em-box top aligned form;               # APOSTROPHE
0028 FE01; em-box centered form;                  # LEFT PARENTHESIS
0029 FE01; em-box centered form;                  # RIGHT PARENTHESIS
002C FE01; em-box bottom aligned form;            # COMMA
002D FE01; em-box centered form;                  # HYPHEN-MINUS
002E FE01; em-box bottom aligned form;            # FULL STOP
002F FE01; em-box centered form;                  # SOLIDUS
0030 FE01; em-box centered form;                  # DIGIT ZERO
0031 FE01; em-box centered form;                  # DIGIT ONE
0032 FE01; em-box centered form;                  # DIGIT TWO
0033 FE01; em-box centered form;                  # DIGIT THREE
0034 FE01; em-box centered form;                  # DIGIT FOUR
0035 FE01; em-box centered form;                  # DIGIT FIVE
0036 FE01; em-box centered form;                  # DIGIT SIX
0037 FE01; em-box centered form;                  # DIGIT SEVEN
0038 FE01; em-box centered form;                  # DIGIT EIGHT
0039 FE01; em-box centered form;                  # DIGIT NINE
003A FE01; em-box centered form;                  # COLON
003B FE01; em-box centered form;                  # SEMICOLON
003F FE01; em-box centered form;                  # QUESTION MARK
005B FE01; em-box centered form;                  # LEFT SQUARE BRACKET
005D FE01; em-box centered form;                  # RIGHT SQUARE BRACKET
007B FE01; em-box centered form;                  # LEFT CURLY BRACKET
007D FE01; em-box centered form;                  # RIGHT CURLY BRACKET
007E FE01; em-box centered form;                  # TILDE
00B7 FE00; x-height centered form;                # MIDDLE DOT
00B7 FE01; em-box centered form;                  # MIDDLE DOT
00B7 FE02; em-box centered fullwidth form;        # MIDDLE DOT
02C7 FE02; em-box centered fullwidth form;        # CARON
02C9 FE02; em-box centered fullwidth form;        # MODIFIER LETTER MACRON
02CA FE02; em-box centered fullwidth form;        # MODIFIER LETTER ACUTE ACCENT
02CB FE02; em-box centered fullwidth form;        # MODIFIER LETTER GRAVE ACCENT
02D9 FE02; em-box centered fullwidth form;        # DOT ABOVE
2012 FE01; em-box centered form;                  # FIGURE DASH
2013 FE00; x-height centered form;                # EN DASH
2013 FE01; em-box centered form;                  # EN DASH
2013 FE02; em-box centered fullwidth form;        # EN DASH
2014 FE00; x-height centered form;                # EM DASH
2014 FE01; em-box centered form;                  # EM DASH
2014 FE02; em-box centered fullwidth form;        # EM DASH
```

```
2018 FE00; cap-height aligned form;                 # LEFT SINGLE QUOTATION MARK
2018 FE01; em-box top aligned form;                 # LEFT SINGLE QUOTATION MARK
2018 FE02; em-box corner-justified fullwidth form; # LEFT SINGLE QUOTATION MARK
2019 FE00; cap-height aligned form;                 # RIGHT SINGLE QUOTATION MARK
2019 FE01; em-box top aligned form;                 # RIGHT SINGLE QUOTATION MARK
2019 FE02; em-box corner-justified fullwidth form; # RIGHT SINGLE QUOTATION MARK
201C FE00; cap-height aligned form;                 # LEFT DOUBLE QUOTATION MARK
201C FE01; em-box top aligned form;                 # LEFT DOUBLE QUOTATION MARK
201C FE02; em-box corner-justified fullwidth form; # LEFT DOUBLE QUOTATION MARK
201D FE00; cap-height aligned form;                 # RIGHT DOUBLE QUOTATION MARK
201D FE01; em-box top aligned form;                 # RIGHT DOUBLE QUOTATION MARK
201D FE02; em-box corner-justified fullwidth form; # RIGHT DOUBLE QUOTATION MARK
2026 FE00; baseline aligned form;                   # HORIZONTAL ELLIPSIS
2026 FE01; em-box bottom aligned form;              # HORIZONTAL ELLIPSIS
2026 FE02; em-box centered fullwidth form;          # HORIZONTAL ELLIPSIS
2E3A FE00; x-height centered form;                  # TWO-EM DASH
2E3A FE01; em-box centered form;                    # TWO-EM DASH
2E3A FE02; em-box centered fullwidth form;          # TWO-EM DASH
2E3B FE00; x-height centered form;                  # THREE-EM DASH
2E3B FE01; em-box centered form;                    # THREE-EM DASH
2E3B FE02; em-box centered fullwidth form;          # THREE-EM DASH
```

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 extracted and used. Although not shown in this document, vertical forms of fullwidth glyphs, if any, are supported via the 'vert' GSUB feature. The table uses VS1, VS2, and VS3 as described in this proposal. Note that characters in the “Default or VS1” column for which no SVS is proposed are shaded in green. Red registration marks are used to draw attention to how their glyphs are typically aligned within the em-box, with prototypical characters surrounding them: 永 to indicate Chinese, and あ to indicate Japanese, and 가 to indicate Korean.

Unicode	Default or VS1	VS2—Em-Box Aligned	VS3—Em-Box Aligned Fullwidth
U+0021	D!C	가!가	
U+0022	D"C	가あ"あ가	
U+0027	D'C	가あ'あ가	
U+0028	D(C	가(가	
U+0029	D)C	가)가	
U+002C	x,x	가,가	
U+002D	x-x	가-가	

Unicode	Default or VS1	VS2—Em-Box Aligned	VS3—Em-Box Aligned Fullwidth
U+002E	x.x	가.가	
U+002F	1/1	가1/1가	
U+0030	D0C	가あ0あ가	
U+0031	D1C	가あ1あ가	
U+0032	D2C	가あ2あ가	
U+0033	D3C	가あ3あ가	
U+0034	D4C	가あ4あ가	
U+0035	D5C	가あ5あ가	
U+0036	D6C	가あ6あ가	
U+0037	D7C	가あ7あ가	
U+0038	D8C	가あ8あ가	
U+0039	D9C	가あ9あ가	
U+003A	x:x	가:가	
U+003B	x;x	가;가	
U+003F	D?C	가?가	
U+005B	D[C	가[가	

Unicode	Default or VS1	VS2—Em-Box Aligned	VS3—Em-Box Aligned Fullwidth
U+005D	D]C	가]가	
U+007B	D{C	가{가	
U+007D	D}C	가}가	
U+007E	X~X	가~가	
U+00B7	X·X	가·가	永・永
U+02C7	XˇX		永ˇ永
U+02C9	X¯X		永¯永
U+02CA	XˊX		永ˊ永
U+02CB	XˋX		永ˋ永
U+02D9	X˙X		永・永
U+2012	x1‒1x	가1‒1가	
U+2013	x–x	가–가	あ永–永あ
U+2014	x—x	가—가	あ永—永あ
U+2018	D‘C	가あ‘あ가	永 ‘永
U+2019	D’C	가あ’あ가	永’ 永
U+201C	D“C	가あ“あ가	永 “永

Unicode	Default or VS1	VS2—Em-Box Aligned	VS3—Em-Box Aligned Fullwidth
U+201D	D”C	가あ”あ가	永” 永
U+2026	x.....x	가.....가	あ永…永あ
U+2E3A	x——x	가——가	あ永——永あ
U+2E3B	x———x	가———가	あ永———永あ

Rationale & Conclusion

This proposal addresses the varying regional conventions for digits and punctuation, 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 arises 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 also addresses the need to tailor ASCII punctuation for Korean use.

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, especially the ones that can be used for both Western and East Asian text.

That is all.