Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: Working Group Document
Title: Proposal to add standardized variation sequences
Author: Ken Lunde (Adobe Systems Incorporated)
Status: Corporate Full Member Contribution
Action: For consideration by the UTC
Date: 2017-02-13

Background

This proposal is a more comprehensive version of L2/14-006 that was originally discussed during UTC #138 in early 2014, and was resurrected for continued discussion during UTC #150 in early 2017. The scope of the proposal has been expanded to include ASCII punctuation, ASCII digits, and fullwidth punctuation.

Due to the presence of—or need for—both fullwidth and non-fullwidth (which are generally proportional, but may be halfwidth) glyphs for particular characters in mainstream East Asian fonts, along with a long-standing disagreement among OSes and national standards with regard to how particular characters map between legacy encodings and Unicode, various ambiguities persist in today's environments. Regional conventions can also affect how particular fullwidth punctuation should display. 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.

Using Japanese as an example, it is not uncommon for a Japanese-language document, or even a single Japanese-language paragraph, to include full runs of English-language text, which may include one or more of the characters that are covered by this proposal. Another example is Korean, which uses ASCII punctuation by default, but whose Western forms are inappropriate for Korean use due to the alignment of their glyphs to Western typographic features, such as the baseline, x-height, and cap-height. Such usage demonstrates a need to preserve the distinction in "plain text" situations. Although "rich text" environments are becoming more common, including those that support language-tagging, "plain text" environments persist, and are likely to continue to persist for a long time due to their robust nature. Also, environments that support variation sequences outnumber those that support language-tagging.

Proposal Summary

This document consists of two separate proposals.

The first proposal is to add standardized variation sequences for 45 punctuation and digit characters that use VS1 through VS3 (U+FE00 through U+FE02) to distinguish *Western*, *East Asian*, and *East Asian fullwidth* forms, respectively. The vast majority of these characters correspond to ASCII, and it is worth repeating that Korean typography makes extensive use of ASCII punctuation and digits.

The second proposal is to add standardized variation sequences for eight (8) fullwidth punctuation characters that use VS1 and VS2 (U+FE00 and U+FE01) to distinguish *left-justified* and *centered* forms, respectively, which vary according to East Asian regional conventions.

First Proposal—Characters With Ambiguous Alignment or Width

Although all 45 characters in the first proposal share ambiguity in terms of Western versus East Asian usage, they can be grouped into the classifications as described in this section. In a nutshell, typical Western usage requires

proportional-width glyphs that are aligned to the Western baseline or cap-height, or are centered within the Western x-height. Typical East Asian usage requires glyphs that are aligned to the em-box. Depending on the typeface design, the difference can be somewhat subtle or more pronounced.

ASCII Punctuation—Baseline, cap-height, or center of x-height versus em-box alignment

With the exception of U+0022 QUOTATION MARK and U+0027 APOSTROPHE that are also used for Japanese typography, ASCII punctuation is preferred for Korean typography, and their glyphs need to be aligned according to the em-box. Western fonts include glyphs that are aligned to the baseline, cap-height, or x-height, but East Asian use, particularly for Korean, requires glyphs that are instead aligned to the em-box.

Also included within this scope of this category is U+00B7 MIDDLE DOT.

ASCII Digits—Height

The ASCII digits—U+0030 DIGIT ZERO through U+0039 DIGIT NINE—are commonly used in Japanese and Korean typography. Typical Western fonts include glyphs that are slightly shorter than the cap-height, but East Asian fonts benefit from slightly taller glyphs that align better with the glyphs for Japanese kana, Korean hangul, and ideographs.

U+2011 through U+2014—Center of x-height versus center of em-box alignment; mapping disagreement

These four characters—U+2011 NON-BREAKING HYPHEN, U+2012 FIGURE DASH, U+2013 EN DASH and U+2014 EM DASH—require different alignment for Western versus East Asian usage, with the difference being center of x-height for Western use versus center of em-box alignment for East Asian use.

Also, due to a pre-existing—and in all likelihood persisting—issue that conflates U+2014 EN DASH with U+2015 HORIZONTAL BAR, the East Asian fullwidth form of U+2014 should be rendered the same as U+2015. These two characters are somewhat ambiguous in terms of how they are transcoded between legacy encodings and Unicode, according to OSes and national standards. In the case of Japanese, JIS X 0213 1-01-29 maps to U+2014 according to the JIS X 0213 standard itself and Apple's macOS, but it maps to U+2015 according to Microsoft's Windows OS. Note that this mapping disagreement is somewhat orthogonal to the Western versus East Asian alignment ambiguity.

U+2018 through U+201A, & U+201C through U+201E—Cap-height or baseline versus em-box alignment, and fullwidth

These six characters—U+2018 LEFT SINGLE QUOTATION MARK, U+2019 RIGHT SINGLE QUOTATION MARK, U+201A SINGLE LOW-9 QUOTATION MARK, U+201C LEFT DOUBLE QUOTATION MARK, U+201D RIGHT DOUBLE QUOTATION MARK, and U+201E DOUBLE LOW-9 QUOTATION MARK—appear differently depending on whether they are used for Western or East Asian purposes. In the former case, the glyphs for these characters are typically designed to align with the cap-height, such as U+0041 LATIN CAPITAL LETTER A through U+005A LATIN CAPITAL LETTER Z for basic Latin, or the baseline, and are proportional-width. In the latter case, the glyphs for these characters are typically designed to align with the top and bottom of the em-box. Four of these characters—U+2018, U+2019, U+201C, and U+201D—may also be aligned to the top corners of the em-box and have fullwidth metrics.

U+2026—Baseline versus em-box versus center of em-box alignment, and fullwidth

This character—U+2026 HORIZONTAL ELLIPSIS—appears differently depending on whether it is used for Western or East Asian purposes. In the Western case, the glyph for this character is aligned to the Western baseline, and is typically composed of three evenly-spaced instances of U+002E FULL STOP, and is proportional. In the East Asian case, the glyph for this character is aligned close to the baseline, but its actual alignment should be to the East Asian form of U+002E FULL STOP. In the East Asian fullwidth case, the glyph for this character is centered within the em-box, and is fullwidth.

U+203C & U+2047 through U+2049—Baseline versus em-box alignment, and fullwidth

These four digraph characters—U+203C DOUBLE EXCLAMATION MARK, U+2047 DOUBLE QUESTION MARK, U+2048 QUESTION EXCLAMATION MARK, and U+2049 EXCLAMATION QUESTION MARK—appear differently depending on whether they are used for Western or East Asian purposes. In the Western case, the glyphs for these characters are typically designed to align with the baseline and cap-height, and are proportional. In the East Asian case, the glyphs for these characters are typically designed to align with the top and bottom of the em-box, and can be proportional or fullwidth. Korean typography prefers proportional forms.

U+2E3A & U+2E3B—Center of x-height versus center of em-box alignment, and fullwidth

These two characters—U+2E3A TWO-EM DASH and U+2E3B THREE-EM DASH—appear differently depending on whether they are used for Western or East Asian purposes, and are included here because they are extended forms of U+2014 EM DASH, and thus have the same issues. In the Western case, the glyphs for these characters are typically designed to align with the center of the x-height, such as U+0041 through U+005A for basic Latin, and are composed of two or three connected instances of a proportional-width U+2014. In the East Asian case, the glyphs for these characters are centered within the em-box, and are composed of two or three connected instances of an East Asian U+2014. Finally, in the East Asian fullwidth case, the glyphs are still centered within the em-box, but are composed of two or three connected instances of an East Asian fullwidth U+2014, which is typically rendered the same as U+2015.

Second Proposal—Fullwidth Characters With Ambiguous Alignment

The second proposal covers eight (8) East Asian punctuation characters that are Wide or Fullwidth, and are ambiguous in terms of whether the glyphs are left-justified or centered. While single-language or single-region East Asian fonts need to include only one of these forms, Pan-CJK fonts that support multiple East Asian languages and regions need to include both forms to accommodate multiple typographic conventions.

China (PRC) uses left-justified forms, Taiwan (ROC) and Hong Kong SAR use centered forms, and Japan uses left-justified periods and commas, but uses centered exclamation points, question marks, colons, and semicolons.

Standardized Variation Sequences

Standardized variation sequences offer a solution to this glyph-level alignment ambiguity by using variation selectors to indicate these conventions. 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 proposed standardized variation sequences as they would appear in the *StandardizedVariants.txt* file, separated into two sections that correspond to each of the two proposals:

```
# Western form, East Asian form, and East Asian fullwidth form variation sequences

0021 FE00; Western form;        # EXCLAMATION MARK
0021 FE01; East Asian form;     # EXCLAMATION MARK
0022 FE00; Western form;        # QUOTATION MARK
0022 FE01; East Asian form;     # QUOTATION MARK
0027 FE00; Western form;        # APOSTROPHE
0027 FE01; East Asian form;     # APOSTROPHE
0028 FE00; Western form;        # LEFT PARENTHESIS
0028 FE01; East Asian form;     # LEFT PARENTHESIS
0029 FE00; Western form;        # RIGHT PARENTHESIS
0029 FE01; East Asian form;     # RIGHT PARENTHESIS
002C FE00; Western form;        # COMMA
002C FE01; East Asian form;     # COMMA
002D FE00; Western form;        # HYPHEN-MINUS
002D FE01; East Asian form;     # HYPHEN-MINUS
002E FE00; Western form;        # FULL STOP
002E FE01; East Asian form;     # FULL STOP
```

```
002F FE00; Western form;            # SOLIDUS
002F FE01; East Asian form;         # SOLIDUS
0030 FE00; Western form;            # DIGIT ZERO
0030 FE01; East Asian form;         # DIGIT ZERO
0031 FE00; Western form;            # DIGIT ONE
0031 FE01; East Asian form;         # DIGIT ONE
0032 FE00; Western form;            # DIGIT TWO
0032 FE01; East Asian form;         # DIGIT TWO
0033 FE00; Western form;            # DIGIT THREE
0033 FE01; East Asian form;         # DIGIT THREE
0034 FE00; Western form;            # DIGIT FOUR
0034 FE01; East Asian form;         # DIGIT FOUR
0035 FE00; Western form;            # DIGIT FIVE
0035 FE01; East Asian form;         # DIGIT FIVE
0036 FE00; Western form;            # DIGIT SIX
0036 FE01; East Asian form;         # DIGIT SIX
0037 FE00; Western form;            # DIGIT SEVEN
0037 FE01; East Asian form;         # DIGIT SEVEN
0038 FE00; Western form;            # DIGIT EIGHT
0038 FE01; East Asian form;         # DIGIT EIGHT
0039 FE00; Western form;            # DIGIT NINE
0039 FE01; East Asian form;         # DIGIT NINE
003A FE00; Western form;            # COLON
003A FE01; East Asian form;         # COLON
003B FE00; Western form;            # SEMICOLON
003B FE01; East Asian form;         # SEMICOLON
003F FE00; Western form;            # QUESTION MARK
003F FE01; East Asian form;         # QUESTION MARK
005B FE00; Western form;            # LEFT SQUARE BRACKET
005B FE01; East Asian form;         # LEFT SQUARE BRACKET
005D FE00; Western form;            # RIGHT SQUARE BRACKET
005D FE01; East Asian form;         # RIGHT SQUARE BRACKET
007B FE00; Western form;            # LEFT CURLY BRACKET
007B FE01; East Asian form;         # LEFT CURLY BRACKET
007D FE00; Western form;            # RIGHT CURLY BRACKET
007D FE01; East Asian form;         # RIGHT CURLY BRACKET
007E FE00; Western form;            # TILDE
007E FE01; East Asian form;         # TILDE
00B7 FE00; Western form;            # MIDDLE DOT
00B7 FE01; East Asian form;         # MIDDLE DOT
2011 FE00; Western form;            # NON-BREAKING HYPHEN
2011 FE01; East Asian form;         # NON-BREAKING HYPHEN
2012 FE00; Western form;            # FIGURE DASH
2012 FE01; East Asian form;         # FIGURE DASH
2013 FE00; Western form;            # EN DASH
2013 FE01; East Asian form;         # EN DASH
2014 FE00; Western form;            # EM DASH
2014 FE01; East Asian form;         # EM DASH
2014 FE02; East Asian fullwidth form; # EM DASH
2018 FE00; Western form;            # LEFT SINGLE QUOTATION MARK
2018 FE01; East Asian form;         # LEFT SINGLE QUOTATION MARK
2018 FE02; East Asian fullwidth form; # LEFT SINGLE QUOTATION MARK
2019 FE00; Western form;            # RIGHT SINGLE QUOTATION MARK
2019 FE01; East Asian form;         # RIGHT SINGLE QUOTATION MARK
2019 FE02; East Asian fullwidth form; # RIGHT SINGLE QUOTATION MARK
201A FE00; Western form;            # SINGLE LOW-9 QUOTATION MARK
201A FE01; East Asian form;         # SINGLE LOW-9 QUOTATION MARK
201C FE00; Western form;            # LEFT DOUBLE QUOTATION MARK
201C FE01; East Asian form;         # LEFT DOUBLE QUOTATION MARK
201C FE02; East Asian fullwidth form; # LEFT DOUBLE QUOTATION MARK
201D FE00; Western form;            # RIGHT DOUBLE QUOTATION MARK
201D FE01; East Asian form;         # RIGHT DOUBLE QUOTATION MARK
201D FE02; East Asian fullwidth form; # RIGHT DOUBLE QUOTATION MARK
201E FE00; Western form;            # DOUBLE LOW-9 QUOTATION MARK
```

```
201E FE01; East Asian form;          # DOUBLE LOW-9 QUOTATION MARK
2026 FE00; Western form;             # HORIZONTAL ELLIPSIS
2026 FE01; East Asian form;          # HORIZONTAL ELLIPSIS
2026 FE02; East Asian fullwidth form; # HORIZONTAL ELLIPSIS
203C FE00; Western form;             # DOUBLE EXCLAMATION MARK
203C FE01; East Asian form;          # DOUBLE EXCLAMATION MARK
203C FE02; East Asian fullwidth form; # DOUBLE EXCLAMATION MARK
2047 FE00; Western form;             # DOUBLE QUESTION MARK
2047 FE01; East Asian form;          # DOUBLE QUESTION MARK
2047 FE02; East Asian fullwidth form; # DOUBLE QUESTION MARK
2048 FE00; Western form;             # QUESTION EXCLAMATION MARK
2048 FE01; East Asian form;          # QUESTION EXCLAMATION MARK
2048 FE02; East Asian fullwidth form; # QUESTION EXCLAMATION MARK
2049 FE00; Western form;             # EXCLAMATION QUESTION MARK
2049 FE01; East Asian form;          # EXCLAMATION QUESTION MARK
2049 FE02; East Asian fullwidth form; # EXCLAMATION QUESTION MARK
2E3A FE00; Western form;             # TWO-EM DASH
2E3A FE01; East Asian form;          # TWO-EM DASH
2E3A FE02; East Asian fullwidth form; # TWO-EM DASH
2E3B FE00; Western form;             # THREE-EM DASH
2E3B FE01; East Asian form;          # THREE-EM DASH
2E3B FE02; East Asian fullwidth form; # THREE-EM DASH


# Left-justified form and centered form variation sequences

3001 FE00; left-justified form;      # IDEOGRAPHIC COMMA
3001 FE01; centered form;            # IDEOGRAPHIC COMMA
3002 FE00; left-justified form;      # IDEOGRAPHIC FULL STOP
3002 FE01; centered form;            # IDEOGRAPHIC FULL STOP
FF01 FE00; left-justified form;      # FULLWIDTH EXCLAMATION MARK
FF01 FE01; centered form;            # FULLWIDTH EXCLAMATION MARK
FF0C FE00; left-justified form;      # FULLWIDTH COMMA
FF0C FE01; centered form;            # FULLWIDTH COMMA
FF0E FE00; left-justified form;      # FULLWIDTH FULL STOP
FF0E FE01; centered form;            # FULLWIDTH FULL STOP
FF1A FE00; left-justified form;      # FULLWIDTH COLON
FF1A FE01; centered form;            # FULLWIDTH COLON
FF1B FE00; left-justified form;      # FULLWIDTH SEMICOLON
FF1B FE01; centered form;            # FULLWIDTH SEMICOLON
FF1F FE00; left-justified form;      # FULLWIDTH QUESTION MARK
FF1F FE01; centered form;            # FULLWIDTH QUESTION MARK
```

The two tables below demonstrate an actual implementation—using a fully-functional OpenType/CFF font with an appropriately-built Format 14 'cmap' subtable that specify the UVSes (*Unicode Variation Sequences*) that correspond to the proposed standardized variation sequences. This OpenType/CFF font is also attached to this PDF, 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 first table uses VS1 through VS3 as described in the first proposal, and the second table uses VS1 and VS2 as described in the second 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.

Unicode	VS1—Western	VS2—East Asian	VS3—East Asian Fullwidth
U+0021	D!C	가!가	
U+0022	D"C	가あ"あ가	

Unicode	VS1—Western	VS2—East Asian	VS3—East Asian Fullwidth
U+0027	D'C	가아˙아가	
U+0028	D(C	가(가	
U+0029	D)C	가)가	
U+002C	X,X	가,가	
U+002D	X–X	가–가	
U+002E	X.X	가.가	
U+002F	X/X	가/가	
U+0030	D0C	가아0아가	
U+0031	D1C	가아1아가	
U+0032	D2C	가아2아가	
U+0033	D3C	가아3아가	
U+0034	D4C	가아4아가	
U+0035	D5C	가아5아가	
U+0036	D6C	가아6아가	
U+0037	D7C	가아7아가	
U+0038	D8C	가아8아가	

Unicode	VS1—Western	VS2—East Asian	VS3—East Asian Fullwidth
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	
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	
U+2011	X‑X	가‑가	
U+2012	X‒X	가‒가	
U+2013	X–X	가–가	
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

Unicode	VS1—Western	VS2—East Asian	VS3—East Asian Fullwidth
U+201A	D‚C	가あ‚あ가	
U+201C	D"C	가あ"あ가	永 " 永
U+201D	D"C	가あ"あ가	永 " 永
U+201E	D„C	가あ„あ가	
U+2026	….	가….가	あ永…永あ
U+203C	D.‼.C	가あ‼あ가	あ永‼永あ
U+2047	D.⁇.C	가あ⁇あ가	あ永⁇永あ
U+2048	D.⁈.C	가あ⁈あ가	あ永⁈永あ
U+2049	D.⁉.C	가あ⁉あ가	あ永⁉永あ
U+2E3A	X——X	가——가	あ——あ
U+2E3B	X———X	가———가	あ———あ

Unicode	VS1—Left-Justified		VS2—Centered
U+3001	あ汉、 汉あ		永 、 永
U+3002	あ汉。 汉あ		永 。 永
U+FF01	汉！ 汉		あ永 ！ 永あ
U+FF0C	あ汉，汉あ		永 ，永

Unicode	VS1—Left-Justified	VS2—Centered
U+FF0E	あ汉．汉あ	永・永
U+FF1A	汉：汉	あ永：永あ
U+FF1B	汉；汉	あ永；永あ
U+FF1F	汉？汉	あ永？永あ

Rationale & Conclusion

The issue that this proposal addresses arises when—or is exposed by—mainstream fonts that include both proportional (for Western 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 different regional conventions for fullwidth East Asian punctuation, which is an issue for Pan-CJK fonts that support multiple East Asian languages and regions, along with the need to tailor ASCII punctuation for Korean use.

It is worthwhile to point out that many of the characters covered by these two proposals 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.