L2/14-006

Universal Multiple-Octet Coded Character Set
International Organization for Standardization

Doc Type: Working Group Document
Title: Proposal to add standardized variation sequences for nine characters
Author: Dr. Ken Lunde (Adobe Systems Incorporated)
Status: Corporate Full Member Contribution
Action: For consideration by the UTC
Date: 2014-01-09

Background

Due to the presence of both full-width and non–full-width (proportional- or half-width) glyphs for particular characters in mainstream CJK 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. Because the glyphs for these characters share the same code point, a font change or font-based features (such as via OpenType's 'GSUB' table) must be used to 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 in question. Such usage demonstrates a need to preserve the distinction in plain text situations. Although rich text environments are becoming more common, plain text environments persist, and are likely to continue to persist for a long time due to their robust nature.

Characters With Ambiguous Alignment

Although all nine of the characters in this proposal share ambiguity in terms of Western versus CJK usage, they can be grouped into the four 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 on the Western x-height. Typical CJK usage requires full-width glyphs that are aligned to the center or corner of the em-box.

U+2014 & U+2015—Center of x-height versus center of em-box alignment; OS & national standard mapping disagreement

When treated as a pair, these two characters—EM-DASH and HORIZONTAL BAR—are 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 JIS X 0213 itself and Apple's OS X, but maps to U+2015 according to Microsoft's Windows. Based on numerous discussions in the past, this disagreement is likely to never be resolved, and for the purposes of mixed Western/CJK usage, these two characters should be treated as equivalent characters for reasons of practicality. As equivalent characters in such usage, these characters appear differently depending on whether they are used for Western or CJK purposes, mainly in terms of alignment. In the former 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 proportional-width. In the latter case, the glyphs for these characters are typically designed to align with the center of the em-box, and are full-width. Note that the mapping disagreement is somewhat orthogonal to the Western versus CJK alignment ambiguity.

U+2018, U+2019, U+201C & U+201D—Cap-height versus em-box alignment

These four characters—LEFT SINGLE QUOTATION MARK, RIGHT SINGLE QUOTATION MARK, LEFT DOUBLE QUOTATION MARK, and RIGHT DOUBLE QUOTATION MARK—appear differently depending on whether they are used for Western or CJK purposes. In the former case, the glyphs for these characters are typically designed to align with the cap-height, such as U+0041 through U+005A for basic Latin, and are proportional-width. In the latter case, the glyphs for these characters are typically designed to align with the top corners of the em-box, and are full-width.

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

This character—HORIZONTAL ELLIPSIS—appears differently depending on whether it is used for Western or CJK purposes. In the former 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-width. In the latter case, the glyph for this character is centered within the em-box, and is full-width.

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

These two characters—TWO-EM DASH and THREE-EM DASH—appear differently depending on whether they are used for Western or CJK purposes, and are included here because they are extended forms of U+2014, and thus have the same issue. In the former 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 (EM DASH). In the latter case, the glyphs for these characters are centered within the em-box, and are composed of two or three connected instances of a full-width U+2014.

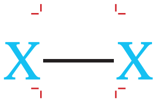
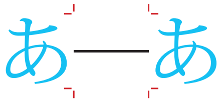
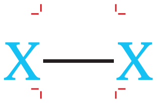
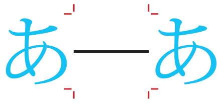
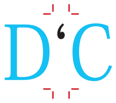
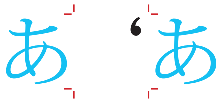
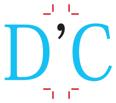
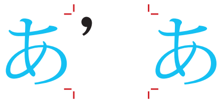
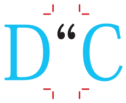
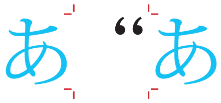
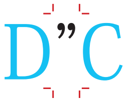
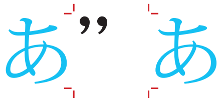
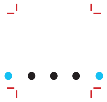
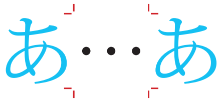
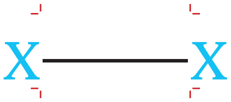
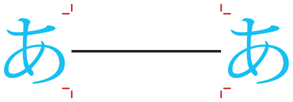
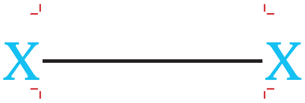
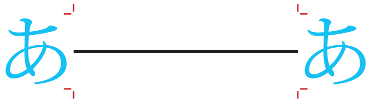
Proposal

Standardized variation sequences offer a solution to this glyph-level alignment ambiguity by using one variation selector, such as VS1 (U+FE00), to indicate Western conventions, and another, such as VS2 (U+FE01), to indicate CJK 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. Below is a complete list of the proposed sequences as they would appear in the *StandardizedVariants.txt* file.

```
# Western style versus CJK style variation sequences

2014 FE00; Western style; # EM DASH
2014 FE01; CJK style;     # EM DASH
2015 FE00; Western style; # HORIZONTAL BAR
2015 FE01; CJK style;     # HORIZONTAL BAR
2018 FE00; Western style; # LEFT SINGLE QUOTATION MARK
2018 FE01; CJK style;     # LEFT SINGLE QUOTATION MARK
2019 FE00; Western style; # RIGHT SINGLE QUOTATION MARK
2019 FE01; CJK style;     # RIGHT SINGLE QUOTATION MARK
201C FE00; Western style; # LEFT DOUBLE QUOTATION MARK
201C FE01; CJK style;     # LEFT DOUBLE QUOTATION MARK
201D FE00; Western style; # RIGHT DOUBLE QUOTATION MARK
201D FE01; CJK style;     # RIGHT DOUBLE QUOTATION MARK
2026 FE00; Western style; # HORIZONTAL ELLIPSIS
2026 FE01; CJK style;     # HORIZONTAL ELLIPSIS
2E3A FE00; Western style; # TWO-EM DASH
2E3A FE01; CJK style;     # TWO-EM DASH
2E3B FE00; Western style; # THREE-EM DASH
2E3B FE01; CJK style;     # THREE-EM DASH
```

The table below demonstrates an actual implementation—using an OpenType Japanese font with an appropriately-built Format 14 ‘cmap’ subtable—that uses VS1 and VS2 as described above for all of the nine characters in this proposal. Red registration marks are used to better illustrate how the glyphs are typically aligned for Western versus CJK usges.

UCS	VS1	VS2
U+2014		
U+2015		
U+2018		
U+2019		
U+201C		
U+201D		
U+2026		
U+2E3A		
U+2E3B		

It is also worthwhile to point out that the characters covered by this proposal have been problematic for developers for years.

Rationale

The issue that this proposal addresses arises when—or is exposed by—mainstream fonts that include both proportional-width (for Western use) and full-width (for CJK use) forms of the same character, and whereby the possibility of use in the same text is relatively high.

Mainstream Japanese fonts, most of which adhere to the public Adobe-Japan1-*x* glyph set, and which number well over one thousand, include a relatively rich set of proportional-width Latin glyphs, well beyond ASCII, that are intended for Western usage, along with an even richer set of full-width glyphs that are intended for Japanese (CJK) usage. Some of the proportional- and full-width glyphs correspond to the same character, but have different usage scenarios, specifically Western versus CJK.

Excluded Characters

The characters described in this section, which also have similar alignment differences in Western and CJK usages, were also considered, but were explicitly exluded for reasons that are explained below. I would have no objections to adding to this proposal, for good measure, the characters that have been excluded merely due to the lack of frequent use, or for not having both Western and CJK forms in mainstream fonts.

U+2010—Center of x-height versus center of em-box alignment

Mainstream Japanese fonts include a CJK (full-width) version of this character—HYPHEN—along with a Western (proportional-width) one, but the latter is typically encoded at U+002D—HYPHEN-MINUS—within the ASCII range. There are similar alignment differences, such as x-height versus em-box, but current conventions give both forms separate homes.

U+2022—Center of x-height versus center of em-box alignment

In lieu of this character—BULLET—it is common for Japanese texts to use U+30FB—KATAKANA MIDDLE DOT—as a CJK bullet. For that matter, U+00B7—MIDDLE DOT—also serves as an alternate home for the Western-style glyph.

U+2025—Baseline versus center of em-box alignment

Although this character—TWO DOT LEADER— is obviously related to U+2026, in terms of Western versus CJK alignment, and thus has the potential to require similar treatment, it is incredibly rare for a single font to include both a Western and CJK form. Mainstream CJK fonts include only a single form of this character, which is intended for CJK usage.

U+00B0, U+2032 & U+2033—Cap-height versus em-box alignment

These three characters—DEGREE SIGN, PRIME, and DOUBLE PRIME—are treated as a matching set, and share the same Western versus CJK alignment differences, but their use in a single text is infrequent.

That is all.