

Title: Proposal to add standardized variation sequences for four quotation marks

Author: Ken Lunde

Date: 2023-10-14

This document is a proposal for adding eight standardized variation sequences (SVSes) for the following four quotation marks that use VS1 (aka U+FE00) and VS2 (aka U+FE01) to distinguish between the forms whose usage varies according to well-established Western versus East Asian conventions:

U+2018 ‘ LEFT SINGLE QUOTATION MARK

U+2019 ’ RIGHT SINGLE QUOTATION MARK

U+201C “ LEFT DOUBLE QUOTATION MARK

U+201D ” RIGHT DOUBLE QUOTATION MARK

Background

This proposal is a stripped down version of document [L2/18-073](#) that was discussed during the UTC #155 meeting, which itself is a stripped down version of document [L2/18-013](#) that was never discussed, which is the second part of a split version of document [L2/17-056](#) that was originally discussed during the UTC #138 meeting as document [L2/14-006](#). Document L2/17-056 itself was discussed during the UTC #153 meeting for the purpose of soliciting feedback that led to the split proposal. The first part was previously submitted as document [L2/17-436](#) that was discussed during the UTC #154 meeting, and which resulted in 16 SVSes being accepted for Unicode Version 12.0.

Regional conventions affect how particular characters, such as quotation marks, should display, and for the characters within the scope of this proposal, the general difference is whether they are non–full-width (aka proportional) or full-width and aligned to the left or right side of the em-box. 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](#)' (*Localized Forms*) layout feature, must be used to specify or distinguish them, which is not possible in “plain text” environments. Some have even claimed that these four characters should have been disunified, like the non–full-width and full-width ASCII character pairs that can be found in the [C0 Controls and Basic Latin](#) and [Halfwidth and Fullwidth Forms](#) blocks, respectively. SVSes provide an encoded representation for these important typographic differences.

Although “rich text” environments are becoming more and more common, including those that support language-tagging and the OpenType '[locl](#)' layout 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 vastly outnumber those that support language-tagging.

Quotation Marks

This proposal covers four quotation marks whose shapes are generally the same regardless of regional conventions, but whose width or alignment can vary by regional conventions. West-

ern typographic conventions use quotation marks that are proportional. Modern Japanese and Korean practice generally follows Western typographic conventions, but in some contexts may require quotation marks that are full-width and aligned to the left or right side of the em-box. In contrast, Chinese typographic conventions use quotation marks that are full-width and aligned to the left or right side of the em-box.

It is true that East Asian punctuation characters are generally full-width, though regional conventions may vary for some of them. For example, and as mentioned in the previous paragraph, Japanese and Korean tend to use non-full-width quotation marks. Furthermore, these four quotation marks 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 benefit from SVSes for the described use cases.

It is worth pointing out that these four quotation marks have the *Vertical_Orientation* (see [UAX #50](#)) property value “R” (*Rotated*), but are transformed when displayed as full-width forms. In other words, tailoring is required to accommodate their full-width forms. UAX #50 should be updated to document the behavior of these four quotation marks.

While Pan-CJK fonts, such as those of the open source *Source Han* (Sans, Serif, and Mono) and *Noto CJK* (Sans and Serif) typeface families, tend to include glyphs for Western and multiple East Asian regional conventions for particular characters, single-region East Asian fonts may also include both Western and East Asian glyphs for the same characters, including the four quotation marks that are covered by this proposal.

Standardized Variation Sequences

Standardized variation sequences offer a solution to this ambiguity by using variation selectors to specify typographic 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 eight proposed standardized variation sequences as they would appear in the UCD’s *StandardizedVariants.txt* file:

```
# Quotation mark width variants
2018 FE00; non-fullwidth form; # LEFT SINGLE QUOTATION MARK
2018 FE01; right-justified fullwidth form; # LEFT SINGLE QUOTATION MARK
2019 FE00; non-fullwidth form; # RIGHT SINGLE QUOTATION MARK
2019 FE01; left-justified fullwidth form; # RIGHT SINGLE QUOTATION MARK
201C FE00; non-fullwidth form; # LEFT DOUBLE QUOTATION MARK
201C FE01; right-justified fullwidth form; # LEFT DOUBLE QUOTATION MARK
201D FE00; non-fullwidth form; # RIGHT DOUBLE QUOTATION MARK
201D FE01; left-justified fullwidth form; # RIGHT DOUBLE QUOTATION MARK
```

The table below demonstrates an actual implementation, specifically a fully-functional OpenType/CFF font with a 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 as a PDF attachment, and can be easily extracted and used. 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 both the glyph metrics and how the glyphs are aligned within the em-box, with prototypical characters surrounding them. In the VS1 column in particular, a *space* (U+0020) is present between the first “X” and the opening quotes, and between the closing quotes and the second “X.”

Unicode	VS1	VS2
U+2018	X ‘ X	永 ‘ 永
U+2019	X ’ X	永 ’ 永
U+201C	X “ X	永 “ 永
U+201D	X ” X	永 ” 永

Vertical Layout Considerations

These four quotation marks are expected to transform in regional-specific ways in vertical layout when the full-width form is used, regardless of whether their full-width glyphs are directly encoded at these four code points in a font, accessed via language-tagging and the 'locl' layout feature, or as proposed in this document (VS2). The 'vert' (*Vertical Alternates*) layout feature operates on GIDs (*Glyph IDs*), not code points or sequences, meaning that the information in this section is purely informative, and the use of SVSes is therefore out-of-scope.

The table below illustrates how the full-width forms of these four quotation marks are expected to transform when in vertical layout, which would be equivalent to the *Vertical_Orientation* property value “Tr” (*Transformed and Rotated*), with “Hans” (Simplified Chinese) being an exceptional case:

Unicode	Horizontal	Vertical	Vertical—Hans
U+2018	永 ‘ 永	永 永	永 永
U+2019	永 ’ 永	永 永	永 永

Unicode	Horizontal	Vertical	Vertical—Hans
U+201C	永 “ 永	永 永	永 永
U+201D	永 ” 永	永 永	永 永

The *Hans* vertical substitutions are unique in that the quotation marks are transformed into the vertical forms of the following *corner brackets* whose vertical forms are separately encoded in the [CJK Compatibility Forms](#) block:

U+300C 〔 LEFT CORNER BRACKET
 U+300D 〕 RIGHT CORNER BRACKET
 U+300E 〔 LEFT WHITE CORNER BRACKET
 U+300F 〕 RIGHT WHITE CORNER BRACKET
 U+FE41 ㄐ PRESENTATION FORM FOR VERTICAL LEFT CORNER BRACKET
 U+FE42 ㄌ PRESENTATION FORM FOR VERTICAL RIGHT CORNER BRACKET
 U+FE43 ㄐ PRESENTATION FORM FOR VERTICAL LEFT WHITE CORNER BRACKET
 U+FE44 ㄌ PRESENTATION FORM FOR VERTICAL RIGHT WHITE CORNER BRACKET

In addition, the standard (non-*Hans*) vertical substitutions of full-width U+201C and U+201D are also unique in that they also the vertical forms of the following two separately-encoded characters:

U+301D ˆ REVERSED DOUBLE PRIME QUOTATION MARK
 U+301F ˘ LOW DOUBLE PRIME QUOTATION MARK

The table below illustrates how the *corner brackets* (aka the horizontal forms of the *Hans* vertical forms of the four quotation marks) and *double prime quotation marks* are expected to transform in vertical layout, which is consistent across East Asian regions:

Unicode	Horizontal	Vertical
U+300C	永 〔 永	永 永

Unicode	Horizontal	Vertical
U+300D	永」 永	永 └ 永
U+300E	永『 永	永 ┐ 永
U+300F	永』 永	永 └ 永
U+301D	永＝ 永	永 〃 永
U+301F	永≡ 永	永 〃 永

Rationale & Conclusion

This proposal addresses the varying regional conventions for four quotation marks, which is a real-world issue for Pan-CJK fonts that support multiple East Asian languages and regions, particularly 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 full-width (for East Asian use) glyphs of the same character, and whereby the possibility of use in the same document is relatively high.

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

That is all.