Proposed Draft Unicode Technical Report #50

UNICODE PROPERTIES FOR VERTICAL TEXT LAYOUT

Editor	Eric Muller (emuller@adobe.com)
Date	2012-04-20
This Version	http://www.unicode.org/reports/tr50/tr50-4.html
Previous Version	n/a
Latest Version	http://www.unicode.org/reports/tr50/
Latest Proposed Update	http://www.unicode.org/reports/tr50/proposed.html
Revision	4

Summary

When text is presented in vertical lines, there are various conventions for the orientation of the characters with respect to the line. In many parts of the world, most characters are upright. In East Asia, Kanji and Kana characters are upright, Latin letters of acronyms are upright, while words and sentences in the Latin script are typically sideways.

This report describes two Unicode character properties which can be used to determine a default orientation of characters in those two scenarios.

Status

This is a draft document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.

A Unicode Technical Report (UTR) contains informative material.

Conformance to the Unicode Standard does not imply conformance to any UTR. Other specifications, however, are free to make normative references to a UTR.

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this document is found in References. For the latest version of the Unicode Standard see [Unicode]. For a list of current Unicode Technical Reports see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

---

1 Editorial warnings.

The draft is currently structured around two properties, with values in the set V={R, U, T}. This is entirely equivalent to a single property with values in the set VxV. Going one step further, we can also give names to the values in VxV and use those names as the property values. Those are syntactic details which are easy to change as this draft progresses.

Another change that can be introduced is to have a level of indirection between the property values and the actual classes, and to bridge that indirection either via a simple mapping, or via rules (e.g. in the style of linebreak) or by some other machinery.

The motivation for the current choice is mostly to make the resulting orientation as clear as possible, and to delay the introduction of more complex machinery until a rationale is provided for doing so.

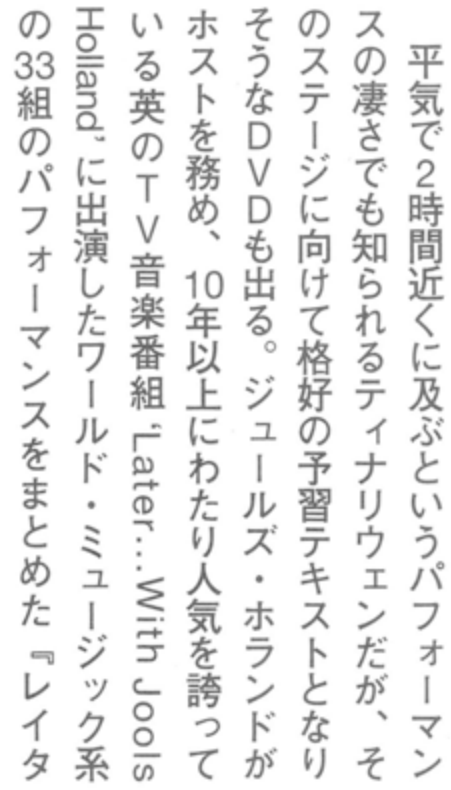
2 Introduction

When text is displayed in vertical lines, there are various conventions for the orientation of the characters with respect to the line. In many parts of the world, most characters are upright, that is appear with the same orientation as in the code charts.

Figure 1. Western vertical text

In East Asia, Kanji and Kana characters are upright, Latin letters of acronyms are upright, while words and sentences in the Latin script are typically sideways.

Figure 2. Japanese vertical text



This report describes two Unicode character properties which can be used to determine a default orientation of characters in those two scenarios.

If and when other scenarios are understood, they will be accommodated by additional properties or by some modification of the existing properties (e.g. to account for differences between Japanese and Chinese uses).

3 Conformance

The properties and algorithms presented in this report are informative. The intent is to provide a reasonable determination of the orientation of characters which can be used in the absence of other information, but can be overridden by the context, such as markup in a document or preferences in a layout application. This default determination is based on the most common use of a character, but in no way implies that that character is used only in that way.

For more information on the conformance implications, see [Unicode], section 3.5, Properties, in particular the definition (D35) of an informative property.

4 Property values

The two properties share the same set of values, which are given in table 1.

Table 1. Property Values

U	characters which are displayed upright, with the same orientation as they appears in the code charts.
R	characters which are displayed sideways, rotated 90 degrees clockwise compared to the code charts.
T, Tu, Tr	characters which are not just upright or sideways, but require a different glyph than in the code charts when used in vertical texts. In addition, Tu indicates that as a fallback, the character can be displayed with the code chart glyph upright; similarly, Tr indicates a possible fallback using the code chart glyph rotated.

Note that the orientation is described with respect to the appearance in the code charts. A number of scripts, such as Mongolian or Phags-pa, are used primarily in vertical lines, and have not developed a tradition of usage in horizontal lines. Similarly, some characters such as U+3031 VERTICAL KANA REPEAT MARK or the characers of the Vertical Forms block are intended for use

primarily in vertical lines. For those scripts and characters, the Unicode code charts show the characters in the orientation and shape they have in vertical lines. It is beyond the scope of this report to describe how those scripts and characters are displayed in horizontal lines (for example, in discursive texts).

5 Properties

The Stacked Vertical Orientation (short name svo) property is intended to be used for vertical lines in those parts of the world where characters are mostly upright.

The Mixed Vertical Orientation (short name mvo) property is intended to be used for vertical lines in East Asia, and more specifically in Japan, China and Korea.

The scope of these properties is limited by the scope of Unicode itself. For example, Unicode does not support directly the representation of texts and inscriptions using Egyptian Hieroglyphs. Instead, Unicode provides characters intended for use when writing about such texts or inscriptions, or for use in conjunction with a markup system such as the Manuel de Codage. While the properties are defined for Egyptian Hieroglyphs, they are meaningfull only for occurrences of these characters in discursive texts; when the characters are used with markup, the markup controls the orientation. See [Unicode], section 14.8 for a more complete discussion of the scope of Egyptian Hieroglyph characters.

5.1 Grapheme Clusters

As in all matters of typography, the interesting unit of text is not the character, but something of the order of a grapheme cluster: it does not make sense to use a base character upright and a combining mark attached to it sideways.

It is expected that the client of the two properties defined here will select a notion of grapheme cluster, and is interested in obtaining an orientation for the cluster as a whole.

A possible choice for the notion of grapheme cluster is either that of legacy grapheme cluster or that of extended grapheme cluster, as defined in [UAX29].

The orientation for a grapheme cluster as a whole is then determined by taking the orientation of the first character in the cluster, with the following

exceptions:

- if the cluster contains an enclosing combining mark (general category Me), then the whole cluster has SVO and MVO orientation U.
- if the cluster is made of U+0020 SPACE and some combining mark(s), then the whole cluster has SVO orientation U and MVO orientation R.
- if the cluster is made of U+3000 IDEOGRAPHIC SPACE and some combining mark(s), then the whole cluster has SVO and MVO orientation U.

5.2 Resulting orientation

The properties are intended to provide only a default orientation, rather than to handle correctly all situations. It is expected by when used in the context of a markup system, the user will be able to 1) have some control over which property is used and 2) specify an explicit orientation. For example, one could have an attribute orientation with possible values auto, 0, 90, 180 and 270; when the value of the attribute is not auto, the explicit orientation is used; when the value is auto, the property values are used.

The property values, if used, are intended to be used directly, with the value SB interpreted as equivalent to S.

There is actually one character for which a contextual determination would be useful and reliable: U+00AE ® REGISTERED SIGN, which can occur both following terms in kanji/kana and following terms in Latin. An occurrence of ® should be assigned the same class as the character it follows. Others? Enough to warrant the complexity of contextual rules?

There are other cases where the character is used routinely in both Japanese and Western contexts: the quotes are a good example. While contextual determination would be useful, it's probably the case that it's not going to be reliable.

6 Tailorings

To facilitate tailorings, this reports identifies sets of characters which behave similarly, and for each it can useful to tailor the orientation as a group.

6.1 The brackets

This set contains brackets, which while they appear rotated, are commonly implemented as if they were transformed.

Table 2. The backets set

00AB
00BB
201C..201F
2039..203A
2045–2046
3008–3011
3014–301B
FE59–FE5E
FF08–FF09
FF3B
FF3D
FF5B
FF5D
FF5F–FF60
FF62–FF63

7 Glyphs Changes for Vertical Orientation

Table 3 provides representative glyphs for the horizontal and vertical appearance of characters with the property value T.

Add glyphs for all the entries: 301F, 332C, FF61, FF64, 1F200, 1F201, halfwidth small kanas. Some glyphs (2018, 2019) may not be correct.

Table 3. Glyph Changes for Vertical Orientation

character	H	V
U+2018 LEFT SINGLE QUOTATION MARK	‘	‘
U+2019 RIGHT SINGLE QUOTATION MARK	’	’

character	H	V
U+3001 IDEOGRAPHIC COMMA		
U+3002 IDEOGRAPHIC STOP		
U+301C WAVE DASH		
U+301D REVERSED DOUBLE PRIME QUOTATION MARK		
U+301E DOUBLE PRIME QUOTATION MARK		
U+301F LOW DOUBLE PRIME QUOTATION MARK		
U+30FC KATAKANA-HIRAGANA PROLONGED SOUND MARK		
U+3041 HIRAGANA LETTER SMALL A		
U+3043 HIRAGANA LETTER SMALL I		
U+3045 HIRAGANA LETTER SMALL U		
U+3047 HIRAGANA LETTER SMALL E		
U+3049 HIRAGANA LETTER SMALL O		
U+3063 HIRAGANA LETTER SMALL TU		
U+3083 HIRAGANA LETTER SMALL YA		
U+3085 HIRAGANA LETTER SMALL YU		
U+3087 HIRAGANA LETTER SMALL YO		
U+308E HIRAGANA LETTER SMALL WA		
U+3095 HIRAGANA LETTER SMALL KA		
U+3096 HIRAGANA LETTER SMALL KE		

character	H	V
U+30A1 KATAKANA LETTER SMALL A		
U+30A3 KATAKANA LETTER SMALL I		
U+30A5 KATAKANA LETTER SMALL U		
U+30A7 KATAKANA LETTER SMALL E		
U+30A9 KATAKANA LETTER SMALL O		
U+30C3 KATAKANA LETTER SMALL TU		
U+30E3 KATAKANA LETTER SMALL YA		
U+30E5 KATAKANA LETTER SMALL YU		
U+30E7 KATAKANA LETTER SMALL YO		
U+30EE KATAKANA LETTER SMALL WA		
U+30F5 KATAKANA LETTER SMALL KA		
U+30F6 KATAKANA LETTER SMALL KE		
U+31F0 KATAKANA LETTER SMALL KU		
U+31F1 KATAKANA LETTER SMALL SI		
U+31F2 KATAKANA LETTER SMALL SU		
U+31F3 KATAKANA LETTER SMALL TO		
U+31F4 KATAKANA LETTER SMALL NU		
U+31F5 KATAKANA LETTER SMALL HA		
U+31F6 KATAKANA LETTER SMALL HI		

character	H	V
U+31F7 KATAKANA LETTER SMALL HU		
U+31F8 KATAKANA LETTER SMALL HE		
U+31F9 KATAKANA LETTER SMALL HO		
U+31FA KATAKANA LETTER SMALL MU		
U+31FB KATAKANA LETTER SMALL RA		
U+31FC KATAKANA LETTER SMALL RI		
U+31FD KATAKANA LETTER SMALL RU		
U+31FE KATAKANA LETTER SMALL RE		
U+31FF KATAKANA LETTER SMALL RO		
U+3300 SQUARE APAATO		
U+3301 SQUARE ARUHUA		
U+3302 SQUARE ANPEA		
U+3303 SQUARE AARU		
U+3304 SQUARE ININGU		
U+3305 SQUARE INTI		
U+3306 SQUARE UON		
U+3307 SQUARE ESUKUUDO		
U+3308 SQUARE EEKAA		
U+3309 SQUARE ONSU		

character	H	V
U+330A SQUARE OOMU		
U+330B SQUARE KAIRI		
U+330C SQUARE KARATTO		
U+330D SQUARE KARORII		
U+330E SQUARE GARON		
U+330F SQUARE GANMA		
U+3310 SQUARE GIGA		
U+3311 SQUARE GINII		
U+3312 SQUARE KYURII		
U+3313 SQUARE GIRUDAA		
U+3314 SQUARE KIRO		
U+3315 SQUARE KIROGURAMU		
U+3316 SQUARE KIROMEETORU		
U+3317 SQUARE KIROWATTO		
U+3318 SQUARE GURAMU		
U+3319 SQUARE GURAMUTON		
U+331A SQUARE KURUZEIRO		
U+331B SQUARE KUROONE		
U+331C SQUARE KEESU		

character	H	V
U+331D SQUARE KORUNA		
U+331E SQUARE KOOPO		
U+331F SQUARE SAIKURU		
U+3320 SQUARE SANTIIMU		
U+3321 SQUARE SIRINGU		
U+3322 SQUARE SENTI		
U+3323 SQUARE SENTO		
U+3324 SQUARE DAASU		
U+3325 SQUARE DESI		
U+3326 SQUARE DORU		
U+3327 SQUARE TON		
U+3328 SQUARE NANO		
U+3329 SQUARE NOTTO		
U+332A SQUARE HAITU		
U+332B SQUARE PAASENTO		
U+332C SQUARE PAATU		
U+332D SQUARE BAARERU		
U+332E SQUARE PIASUTORU		
U+332F SQUARE PIKURU		

character	H	V
U+3330 SQUARE PIKO		
U+3331 SQUARE BIRU		
U+3332 SQUARE HUARADDO		
U+3333 SQUARE HUIITO		
U+3334 SQUARE BUSSYERU		
U+3335 SQUARE HURAN		
U+3336 SQUARE HEKUTAARU		
U+3337 SQUARE PESO		
U+3338 SQUARE PENIHI		
U+3339 SQUARE HERUTU		
U+333A SQUARE PENSU		
U+333B SQUARE PEEZI		
U+333C SQUARE BEETA		
U+333D SQUARE POINTO		
U+333E SQUARE BORUTO		
U+333F SQUARE HON		
U+3340 SQUARE PONDO		
U+3341 SQUARE HOORU		
U+3342 SQUARE HOON		

character	H	V
U+3343 SQUARE MAIKURO		
U+3344 SQUARE MAIRU		
U+3345 SQUARE MAHHA		
U+3346 SQUARE MARUKU		
U+3347 SQUARE MANSYON		
U+3348 SQUARE MIKURON		
U+3349 SQUARE MIRI		
U+334A SQUARE MIRIBAARU		
U+334B SQUARE MEGA		
U+334C SQUARE MEGATON		
U+334D SQUARE MEETORU		
U+334E SQUARE YAADO		
U+334F SQUARE YAARU		
U+3350 SQUARE YUAN		
U+3351 SQUARE RITTORU		
U+3352 SQUARE RIRA		
U+3353 SQUARE RUPII		
U+3354 SQUARE RUUBURU		
U+3355 SQUARE REMU		

character	H	V
U+3356 SQUARE RENTOGEN		
U+3357 SQUARE WATTO		
U+337B SQUARE ERA NAME HEISEI		
U+337C SQUARE ERA NAME SYOUWA		
U+337D SQUARE ERA NAME TAISYOU		
U+337E SQUARE ERA NAME MEIZI		
U+337F SQUARE CORPORATION		
U+FF61 HALFWIDTH IDEOGRAPHIC FULL STOP		
U+FF64 HALFWIDTH IDEOGRAPHIC COMMA		
U+FF67 HALFWIDTH KATAKANA LETTER SMALL A		
U+FF68 HALFWIDTH KATAKANA LETTER SMALL I		
U+FF69 HALFWIDTH KATAKANA LETTER SMALL U		
U+FF6A HALFWIDTH KATAKANA LETTER SMALL E		
U+FF6B HALFWIDTH KATAKANA LETTER SMALL O		
U+FF6C HALFWIDTH KATAKANA LETTER SMALL YA		
U+FF6D HALFWIDTH KATAKANA LETTER SMALL YU		
U+FF6E HALFWIDTH KATAKANA LETTER SMALL YO		
U+FF6F HALFWIDTH KATAKANA LETTER SMALL TU		
U+1F200 SQUARE HIRAGANA HOKA		
U+1F201 SQUARED KATAKANA KOKO		

8 Data File

For the MVO property, there are two approaches for characters which are more symbolic than alphabetic. In approach "A", all symbolic characters have orientation U. In approach "B", arrows, math symbols, box drawing characters, and bracket pieces have orientation R; the remaining symbolic characters have orientation U.

A possibility to reconcile both approaches is to have a specific class and orientation for the characters which differ; this would let users of the properties resolve those values to either class/orientation combination.

Reviewers are encouraged to express a preference for one of the approaches, or for the combined approach.

The data file, in UCD syntax.

To help during the review, a slighlty more readable version, where differences between the A and B proposal are highlighted in red.

U+2016 ‖ DOUBLE VERTICAL LINE; JRLEQ classifies this character as cl-19 ideographic; typically, this is a clue that it is upright; also, JIS 0213:2000 does not give a vertical variant. On the other hand, it seems that 'vert' often presents it sideways. Which is right? Could it be that font vendors have been influenced by U+30A0 = KATAKANA-HIRAGANA DOUBLE HYPHEN?

Acknowledgments

Please let me know if I forgot your name or you prefer a different spelling/etc.

Thanks to the reviewers: Julie Allen, Ken Lunde, Nat McCully, Ken Whistler, Taro Yamamoto, htakenaka, John Cowan, Fantasai, Asmus Freytag, Van Anderson, Ishi Koji, sikeda, Shinyu Murakami, Tokushige Kobayashi, Addison Phillips, Martin Dürst, the W3C Internationalization Core Working Group, the W3C I18N Interest group, the W3C CSS Working group, Michael Everson, John Daggett.

References

[JLREQ] Requirements for Japanese Text layout, W3C Working Group Note 4 June 2009

[Errata] Updates and Errata

http://www.unicode.org/errata

[Feedback] http://www.unicode.org/reporting.html

For reporting errors and requesting information online.

[ISO 10646] International Organization for Standardization. Information Technology – Universal Multiple-Octet Coded Character Set (UCS). (ISO/IEC 10646:2011).

For availability, see: http://www/iso.org

[Reports] Unicode Technical Reports

http://www.unicode.org/reports/

For information on the status and development process for technical reports, and for a list of technical reports.

[UAX29] UAX #29: Unicode Text Segmentation

http://www.unicode.org/reports/tr29/

[Unicode] The Unicode Standard, Version 6.1.0, defined by: The Unicode Standard, Version 6.1.0 (Mountain View, CA: The Unicode Consortium, 2012. ISBN 978-1-936213-02-3)

http://www.unicode.org/versions/Unicode6.1.0

[Versions] Versions of the Unicode Standard

http://www.unicode.org/versions/

For details on the precise contents of each version of the Unicode Standard, and how to cite them.

Modifications

This section indicates the changes introduced by each revision.

Revision 4

- Properties renamed to Stacked Vertical Orientation (previously Default Vertical Orientation) and Mixed Vertical Orientation (previously East Asian Vertical Orientation)
- Introduced sets of characters for tailoring.
- Property value S renamed to R.

- Property value Sb merged with R; set created for brackets.

Revision 3

- Mongolian and Egyptian Hieroglyphs changed to U.
- Implementation of the UTC decisions made during meeting #130, February 2012.
    - Removal of the East Asian Class property
    - East Asian Orientation renamed East Asian Vertical Orientation
    - New property, Default Vertical Orientation. The initial assignment is: T if EAVO=T, SB if EAVO=SB and the bracket is specific to CJK, S to align with CSS Sv value except for vertical presentation forms, Tibetan, Mongolian, sup/sub parens, sup punctuation, FD3E, FD3F, which remain U.

Revision 2

- Clarification of the status of the properties (end of section 1)
- Clarification of the handling of grapheme clusters
- Removed the "comments" column in table 3.
- Hangul characters: new class cl-19.4, hangul, orientation U
- Yijing Hexagram symbosl are now cl-19-3, symbols, orientation U.
- Small forms variants are treated like their fullwidth counterparts.
- Superscripts and subscript characters are now cl-27, western, orientation S
- Small kana: orientation U; class split in cl-11.1, smallHiragana and cl-11.2, smallKatakana
- U+3030 〜 WAVY DASH has orientation T.
- The two alternatives for math, etc. are described.

Revision 1

- First working draft.
