- Draft -- for UTC/L2 Consideration only --

L2/98-054

February 23, 1998

A new Unicode Character Property "Width"

Submitted by: *Asmus Freytag*,
Version 0.02

Background

In mixed-width legacy encodings there is a concept of an inherent width of a character. For a fixed pitch font, this width translates to a display width of either one half or a whole standard unit (hereafter referred to as "Em"). By convention, 1/2 Em wide characters are called "half-width" or *hankaku* characters, the others are called correspondingly "full-width" or *zenkaku* characters. Legacy encodings often use a single byte for the half-width characters and two bytes for the full-width characters.

What ISO/IEC 10646 says today

>> As far as I can tell, ISO 10646 is silent on the terms "half-width" and "full-width" except to say that the characters so named are provided for compatibility.

What the Unicode Standard says today

The Unicode Standard states (p. 6-130):

> In the context of conversion to and from such mixed-width encodings, all characters in the General Scripts area [i.e. 0000-1FFF] should be construed as half-width (*hankaku*) characters.

This sentence, as it stands, is misleading in that it implies that everything in the range U+0000..U+1FFF is half-width.

> All characters in the CJK Phonetics and Symbols area [i.e. 3000-33FF] and the Unified CJK Ideograph area [i.e. 4E00-9FFF], along with the characters in the CJK Compatibility Ideographs [i.e. F900-FAFF], CJK Compatibility Forms [i.e. FE30-FE4F], and Small Form Variants blocks [i.e. FE50-FE6F], should be construed as full-width (*zenkaku*) characters. Other Compatibility Area [i.e. F900-FFFF] characters outside of the current block should be construed as half-width characters. The characters of the Symbols Area are neutral regarding their width semantics.

It should clearly be noted that statements made in the Unicode Standard in Chapter 6 (Character Block Descriptions) do not have normative status. Chapters 3, 4, and 7 (Charts) have normative status. The rest of the book, including Chapter 6 is provided basically to give as much information as possible to help people understand and implement the characters correctly. But it is dangerous to make legalistic arguments based on the text of Chapter 6, since there is rather large leeway for the editors of the Unicode Standard to modify and augment such explanatory text as new issues arise or old ones require more clarification.

The intent of the statement on page 6-130

The intent of the existing paragraph is not to create a property but to account for the fact that there are full-width forms encoded in the ranges U+FF01..U+FF5E and U+FFE0..U+FFE6. When converting a DBCS mixed-width encoding to and from Unicode, the full-width characters in such a mixed-width encoding are mapped to the full-width compatibility characters in the FFxx block, whereas the corresponding half-width characters are mapped to ordinary Unicode characters (e.g. ASCII in U+0021..U+007E, plus a few other scattered characters).

In the context of interoperability with DBCS character encodings, that restricted set of Unicode characters in the General Scripts area can be construed as half-width, rather than full-width. (This applies only to the restricted set of characters which can be paired with the full-width compatibility characters.)

In the context of interoperability with DBCS character encodings, all other Unicode characters which are not explicitly marked as half-width can be construed as full-width.

In any other context, Unicode characters not explicitly marked as being either full-width or half-width compatibility forms should be construed as unmarked as to half-width versus full-width status.

Seen in this light, the "half-width" and "full-width" properties are not unitary character properties in the same sense as "space" or "combining" or "alphabetic". They are, instead, relational properties of a pair of characters, one of which is explicitly encoded as a half-width or full-width form for compatibility in mapping to DBCS mixed-width character encodings.

What is "full-width" by default today could in theory become "half-width" tomorrow by the introduction of another character on the SBCS part of a mixed-width code page somewhere, requiring the introduction of another full-width compatibility character to complete the mapping. Since the single byte part of mixed-width character sets is limited, there are not going to be many candidates and UTC and WG2 both will resist adding compatibility characters unless they are truly critical.

Width as an inherent property

In East Asian typography there is an inherent distinction between 'wide' (or ideographic like) and 'narrow' (or Western like) characters. These two categories follow different rules in everything from line breaking to their appearance in vertical writing. In legacy encodings this distinction has conveniently been equivalent to the half-width / full-width dichotomy. For a globally useful set of character properties, a universal encoding like Unicode must go beyond the simple half-width and full-width model

The distinction between full-width and half-width was made "in the context of interoperability" between Unicode and legacy encodings. However, the rules of East Asian typography apply even if characters were never stored in legacy character sets. A generalized concept of width is therefore more useful..

A generalized concept of 'width'

There are 4 properties

Wide

From the above we recognize that the are wide characters that are *defined* as full-width and also wide characters that are *implicitly* wide (such as the Unified Han Ideographs or Squared Katakana Symbols) because they occur *only* in the context of East Asian typography where they are wide characters.

Narrow

There are narrow characters that are *defined* as half-width and also characters that are half-width by implication because they have full-width clones (all of ASCII is an example).

Half-width

Because half-width punctuation behaves in some important ways like ideographic punctuation, it is useful to distinguish characters defined as half-width from characters that are narrow by implication. Alternatively, it is useful to distinguish characters defined as half-width from general purpose characters that are narrow by implication where there are duplicate pairs (this is a smaller number). Since the latter cannot be trivially derived from the block names, it is what is proposed explicitly below.

Ambiguous (Wide or Narrow depending on context)

For the rest there are lots of characters that are either unspecified or ambiguous."Ambiguous" characters are those that occur in both East Asian legacy character sets as wide characters, and in their own local usage (The standard examples are the Greek and Cyrillic Alphabet, but also some of the mathematical symbols). Ambiguous characters require context to resolve their width.

Neutral (same as Narrow)

"Neutral" characters are all characters that do not occur in legacy mixed width sets at all, and, by extension, also do not occur in East Asian typography. (There is no traditional Japanese way of typesetting Devanagari, for example). In order to keep the scheme simple, neutral and narrow are considered as the same property.

Usage

When interchanging data

- Wide characters map to full-width characters in the mixed-width set
- Wide characters never map to non East Asian legacy character encodings
- Narrow (and neutral) characters map to half-width characters in the mixed-width set
- Half-width characters map to half-width characters in the mixed-width set
- Ambiguous characters map to full-width characters in East Asian legacy character encodings

When processing or displaying data

- Wide characters behave like ideographs in important ways. In fixed pitched fonts, they take up one Em of space.
- Half-width characters behave like ideographs in some ways, In fixed pitched fonts, they take up 1/2 Em of space.
- Narrow characters behave like Western characters in important ways, In fixed pitched East Asian fonts, they take up 1/2 Em of space.
- Ambiguous characters behave like wide or narrow characters depending on context (language tag, associated font, source of data, or explicit markup all can provide the context)

Acknowledgments

Michel Suignard provided extensive input into the analysis and source material for the detail assigments of these properties.

Part of this document draws on e-mail discussion contribution by Ken Whistler, heavily edited, so don't blame him.

Proposed width classification of Unicode 2.0 characters

A - Ambiguous	H - Halfwidth	N - Narrow	W - Wide	X - Unassigned
001F	20A9	0020..00A0	1100..11F9	02A9..02AF
00A1	FF61..FF64	00A2..00A3	3000..303F	02DF
00A4		00A5..00A6	3041..3094	02EA..02FF
00A7..00A8		00A9	3099..309E	0362..0373
00AA		00AB..00AC	30A1..30FE	03F0..03FF
00AD		00AE	3131..318E	0487..048F
00AF..00B4		00B5	3190..319F	04FA..0530
00B6..00BA		00BB	3200..321C	0557..0558
00BC..00BF		00C0..00C5	3220..3243	0560
00C6		00C7..00CF	3260..32B0	0588
00D0		00D1..00D6	32C0..3376	058A..0590
00D7..00D8		00D9..00DD	337B..33DD	05F5..060B
00DE..00E1		00E2..00E5	33E0..33FE	06FA..0900
00E6		00E7	4E00..9FA5	0971..0980
00E8..00EA		00EB	AC00..D7A3	09FB..0A01
00EC..00ED		00EE..00EF	E000..E757	0A75..0A80
00F0		00F1	F900..FA2D	0AF0..0B00
00F2..00F3		00F4..00F6	FE30..FE44	0B71..0B81
00F7..00FA		00F8	FE49..FE52	0BF3..0C00
00FC		00FD	FE54..FE6B	0C70..0C81
00FE		00FF..0100	FF01..FF5E	0CF0..0D01
0101		0102..0110	FFE0..FFE6	0D70..0E00
0111		0112		0E5C..0E80
0113		0114..011A		0EDE..0EFF
011B		011C..0125		0FBA..109F

0126..0127
012B
0131..0133
0138
013F..0142
0144
0148..014B
014D
0152..0153
0166..0167
016B
01CE
01D0
01D2
01D4
01D6
01D8
01DA
01DC
0251
0261
02C7
02C9..02CB
02CD
02D0
02D8..02DB
02DD
0300..0361
0391..03A9
03B1..03C1
03C3..03C9
0401
0410..044F
0451
2010
2013..2016
2018..2019
201C..201D
2020..2021
2025..2027
2030
2032..2033
2035
203B
2074
207F
2081..2084
2103
2105
2109
2113
2116
2121..2122
2126
212B
2153..2154
215B..215E
2160..216B
2170..2179
2190..2199
21D2
21D4
2200
2202..2203
2207..2208
220B
220F
2211
2215
221A
221D..2220

0128..012A
012C..0130
0134..0137
0139..013E
0143
0145..0147
014C
014E..0151
0154..0165
0168..016A
016C..01CD
01CF
01D1
01D3
01D5
01D7
01D9
01DB
01DD..0250
0252..0260
0262..02A8
02B0..02C6
02C8
02CC
02CE..02CF
02D1..02D7
02DC
02DE
02E0..02E9
0374..0390
03AA..03B0
03C2
03CA..03EF
0400
0402..040F
0450
0452..0486
0490..04F9
0531..0556
0559..055F
0561..0587
0589
0591..05F4
060C..06F9
0901..0970
0981..09FA
0A02..0A74
0A81..0AEF
0B01..0B70
0B82..0BF2
0C01..0C6F
0C82..0CEF
0D02..0D6F
0E01..0E5B
0E81..0EDD
0F00..0FB9
10A0..10F6
10FB
1E00..1EF9
1F00..1FFE
2000..200F
2011..2012
2017
201A..201B
201E..201F
2022..2024
2028..202E
2031
2034
2036..203A
203C..2046

10F7..10FA
10FC..10FF
11FA..1DFF
1EFA..1EFF
1FFF
202F
2047..2069
2071..2073
208F..209F
20AC..20CF
2139..2152
2183..218F
21EB..21FF
244B..245F
24EB..24FF
25F0..25FF
2670..2700
27BF..2FFF
3040
3095..3098
309F..30A0
30FF..3104
312D..3130
318F
31A0..31FF
321D..321F
3244..325F
32B1..32BF
3377..337A
33DE..33DF
33FF..4DFF
9FA6..ABFF
D7A4..DFFF
E758..F8FF
FA2E..FAFF
FB07..FB12
FB18..FB1D
FDFC..FE1F
FE24..FE2F
FE45..FE48
FE53
FE6C..FE6F
FEFD..FEFE
FF00
FF5F..FF60
FFDD..FFDF
FFE7
FFEF..FFFB

2223
2225
2227..222C
222E
2234..2237
223C..223D
2248
224C
2252
2260..2261
2264..2267
226A..226B
226E..226F
2282..2283
2286..2287
2295
2299
22A5
22BF
2312
2460..24B5
24D0..24E9
2500..254B
2550..2574
2581..258F
2592..25A1
25A3..25A9
25B2..25B3
25B6..25B7
25BC..25BD
25C0..25C1
25C6..25C8
25CB
25CE..25D1
25E2..25E5
25EF
2605..2606
2609
260E..260F
261C
261E
2640
2642
2660..2661
2663..2665
2667..266A
266C..266D
266F
206A..2070
2075..207E
2080
2085..208E
20A0..20A8
20AA..20AB
20D0..2102
2104
2106..2108
210A..2112
2114..2115
2117..2120
2123..2125
2127..212A
212C..2138
2155..215A
215F
216C..216F
217A..2182
219A..21D1
21D3
21D5..21EA
2201
2204..2206
2209..220A
220C..220E
2210
2212..2214
2216..2219
221B..221C
2221..2222
2224
2226
222D
222F..2233
2238..223B
223E..2247
2249..224B
224D..2251
2253..225F
2262..2263
2268..2269
226C..226D
2270..2281
2284..2285
2288..2294
2296..2298
229A..22A4
22A6..22BE
22C0..2311
2313..244A
24B6..24CF
24EA
254C..254F
2575..2580
2590..2591
25A2
25AA..25B1
25B4..25B5
25B8..25BB
25BE..25BF
25C2..25C5
25C9..25CA
25CC..25CD
25D2..25E1
25E6..25EE
2600..2604
2607..2608
260A..260D
2610..261B
261D

		261F..263F 2641 2643..265F 2662 2666 266B 266E 2701..27BE 3105..312C FB00..FB06 FB13..FB17 FB1E..FDFB FE20..FE23 FE70..FEFC FEFF FF65..FFDC FFE8..FFEE FFFC..FFFD		