Title: **Summary of proposed changes to EAW classification and documentation**
From: Asmus Freytag
Date: 2002-02-13

L2/02-078

1) Based on a detailed review I carried out, the following are currently supported:

- ShiftJIS (the MS version, a superset of IBM and NEC) (MS code page 932)
- Korean KSC C 5061 (as implemented in code page 1361)
- Big 5/TCA/Eten sets (as implemented in Win2K cp 950)
- GBK (code page 936)

as are any and all other legacy sets that are subsets of these.

Assignments in the data file not covered by these sets:

(a) Vulgar Fraction
2155;A # VULGAR FRACTION ONE FIFTH

I could not find this in any of the sets we used to generate the EA width assignment. This looks like a mistake in generating our tables. It's at the end of a range, so it might be a 'one too much' error.
Suggested action: A → N

(b) Circled digits and letters
24B6..24BF;A # CIRCLED LATIN CAPITAL LETTER A..LETTER J (sic!)
24EB..24F4;A # NEGATIVE CIRCLED NUMBER ELEVEN..TWENTY
24F5..24FE;A # DOUBLE CIRCLED DIGIT ONE..NUMBER TEN
These look like mistakes in generating our initial data file, esp. the first range is totally goofy!! See item (3) for a proposed resolution

(c) Box Drawing
2574;A # BOX DRAWINGS LIGHT LEFT
The rest of the same sub-set of box-drawing symbols is not A. This looks like a mapping mistake (it's in the Win2K mapping for code page 950).

Details: If viewed with the PMingLIU font, the glyph is not at all compatible with the definition of the character. Instead of a 'single-pixel' half-line extending from the center of the cell to the left, it's a long heavier underline. It looks like other Taiwanese sets map to FF3F for this, judging by the surrounding characters.

Suggested action: Change A → N

(d) Misc. Special characters
0300..034F;A # (treating combining marks like the preceding character)
0360..036F;A # (ditto)

1160..11A2;A # (treating the trailing conjoing jamos like combining marks)
11A8..11F9;A # (ditto)
FE00..FE0F;A # VARIATION SELECTOR-1..-16
FFFD;A # REPLACEMENT CHARACTER
E000..F8FF, etc.;A # <private use>

These are deliberate structural choices we've made - I think they are defensible as they are - may need better documentation in the TR. 'A' means something slightly different for these sets than for the individual characters and symbols elsewhere.

Suggested action: No action
[The Korean set includes some, but not all of the conjoining Jamo.]

2) Decide that JIS X 0213 should not be added to this set of supported legacy code sets.
Legacy practice is evolving, and many of the recently added characters can be expected to be supported as neutral characters by modern implementations. Removing JIS X 0213 requires several changes to the data file (version 6d3)

Suggested Action: Change from A to N:
23BE..23CC Dental symbols
23CE return symbol
2394..2395 curved corner arrows
2616..2617 Shogi
29BF cicled bullet
26FA..26FB double and triple plus
2985 and 2986

3) Special case the circled digits and letters.

By nature, these are used as wide characters (i.e. stay upright in vertical context). Therefore it makes sense to continue to assign 'A' to these characters, even if they are not found in any of the specific sets in (1).

2776..277F;N # DINGBAT NEGATIVE CIRCLED DIGIT ONE..NUMBER TEN
24B6..24BF;A # CIRCLED LATIN CAPITAL LETTER A..LETTER J (sic!)
24C0..24CF;N # CIRCLED LATIN CAPITAL LETTER K..LETTER Z (sic!)
24EB..24F4;A # NEGATIVE CIRCLED NUMBER ELEVEN..TWENTY
24F5..24FE;A # DOUBLE CIRCLED DIGIT ONE..NUMBER TEN

Suggested action: Change N → A, or keep as A, resp.

4) Document important changes in resolving ambiguous width characters

See the proposed changes in the text of the proposed update to the UAX (appended).

Technical Reports

Proposed Update Unicode Standard Annex #11

EAST ASIAN WIDTH

Version	3.2.0
Authors	Asmus Freytag (asmus@unicode.org)
Date	2001-01-16
This Version	http://www.unicode.org/unicode/reports/tr11/tr11-10
Previous Version	http://www.unicode.org/unicode/reports/tr11/tr11-9
Latest Version	http://www.unicode.org/unicode/reports/tr11
Tracking Number	10 [Appendix to Doc L2/02-078]

Summary

This report presents the specifications of a informative property for Unicode characters that is useful when interoperating with East Asian Legacy character sets.

Status

This document is a proposed update of a previously approved Unicode Standard Annex. Publication does not imply endorsement by the Unicode Consortium. This is a draft document which may be updated, replaced, or superseded by other documents at any time. This is not a stable document; it is inappropriate to cite this document as other than a work in progress. The links in this document to the data files do not work. Preliminary datafiles for the proposed update are available at http://www.unicode.org/Public/BETA.

> *A Unicode Standard Annex (UAX) forms an integral part of the Unicode Standard, but is published as a separate document. Note that conformance to a version of the Unicode Standard includes conformance to its Unicode Standard Annexes. The version number of a UAX document corresponds to the version number of the Unicode Standard at the last point that the UAX document was updated.*
>
> *A list of current Unicode Technical Reports is found on http://www.unicode.org/unicode/reports/. For more information about*

versions of the Unicode Standard, see http://www.unicode.org/unicode/standard/versions/.

The References provide related information that is useful in understanding this document. Please mail corrigenda and other comments to the author(s).

Contents

1 Overview

In mixed-width, East Asian, 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 unit width. A common name for this unit width is "Em". It is customarily the *height* of the letter 'M', but since in East Asian fonts the standard character cell is square, it is the same as the unit *width*.

> Note: the character width for a fixed pitch Latin font like Courier is generally 3/5 of an em.

Layout and line breaking (to cite only two examples) in an East Asian context show systematic variations depending on the value of the East-Asian Width property (even for non-fixed pitch fonts). Further, the same information is useful in creating correct transcoding tables for East Asian character sets.

2 Scope

The East Asian Width property provides a useful concept for implementations that

- have to interwork with East Asian legacy character encodings
- support both East Asian and Western typography and line layout

- need to associate fonts with unmarked text runs containing East Asian characters

This Unicode Technical Report does not provide rules or specifications of how this property might be used in font design or line layout, since, while a useful property for this purpose, it is only one of several character properties that would need to be considered.

3 Description

By convention, 1/2 Em wide characters of East Asian legacy encodings are called "half-width" (or *hankaku* characters in Japanese), 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. In the Unicode Standard, no such distinction is made, but understanding the distinction is often necessary when interchanging data with legacy systems, especially when fixed size buffers are involved. Some character blocks in the compatibility zone contain characters that are explicitly marked "half-width" and "full-width" in their character name but for all other characters the width property must be implicitly derived. Some characters behave differently in East Asian context than in non-East Asian content. Their default width property is considered ambiguous and needs to be resolved into an actual width property based on context. This technical report assigns to each Unicode character one of the six values *Ambiguous, Full Width, Half Width, Narrow, Wide*, or *Not East Asian Neutral* (defined below) as its default width property. For any given operation, these six default property values resolve into only two property values *narrow* and *wide,* depending on context.

4 Definitions

All terms not defined here shall be as defined in the Unicode Standard.

ED1. East Asian Width – in the context of interoperating with East Asian legacy character encodings and implementing East Asian typography, the East Asian Width is a categorization of character. It can take on two abstract values, narrow and wide. In legacy implementations, there is often a corresponding difference in encoding length (one or two bytes) as well as a difference in displayed width. However, the *actual* display width of a glyph is given by the font and may be further adjusted by layout. An important class of fixed width legacy fonts contains glyphs of just two widths with the wider glyphs twice as wide as the narrower glyphs.

> Note: For convenience, the classification further distinguishes among explicitly or implicitly wide and narrow characters.

ED2. East Asian Full-width (*F*) – all characters that are defined as FULL WIDTH in the Unicode Standard ~~and therefore are~~by having a compatibility ~~equivalents of~~decomposition of type <wide> to characters elsewhere in the Unicode Standard that are implicitly narrow but unmarked.

ED3. East Asian Half-width (*H*) – all characters that are explicitly defined as HALF WIDTH in the Unicode Standard ~~and therefore are~~by having a compatibility ~~characters of~~decomposition of type <narrow> to characters elsewhere in the Unicode Standard that are implicitly wide but unmarked plus the WON SIGN.

ED4. East Asian Wide (W) – all other characters that are always wide. These characters occur only in the context of East Asian typography where they are wide characters (such as the Unified Han Ideographs or Squared Katakana Symbols). This category includes characters that have explicit half-width counterparts.

ED5. East Asian Narrow (Na) – all other characters that are always narrow and have explicit full-width or wide counterparts. These characters are implicitly narrow in East Asian typography and legacy character sets since they have explicit full-width or wide counterparts. All of ASCII is an example of East Asian Narrow characters.

> It is useful to distinguish characters explicitly defined as half-width from other characters that have a full-width equivalent. In particular, half-width punctuation behaves in some important ways like ideographic ~~punctuation.~~
> punctuation, and knowing a character is a halfwidth character can aid in font selection when binding a font to unstyled text.

ED6. East Asian Ambiguous (A) – all characters that can be sometimes wide and sometimes narrow. Ambiguous characters require additional information not contained in the character code to further resolve their width.

> Ambiguous characters occur in East Asian legacy character sets as *wide* characters, but as *narrow* (i.e. normal-width) characters in non-East Asian usage (Examples are the Greek and Cyrillic alphabet found in East Asian character sets, but also some of the mathematical symbols). Private Use characters are considered ambiguous, since additional information is required to know whether they should be treated as wide or narrow.

Error!

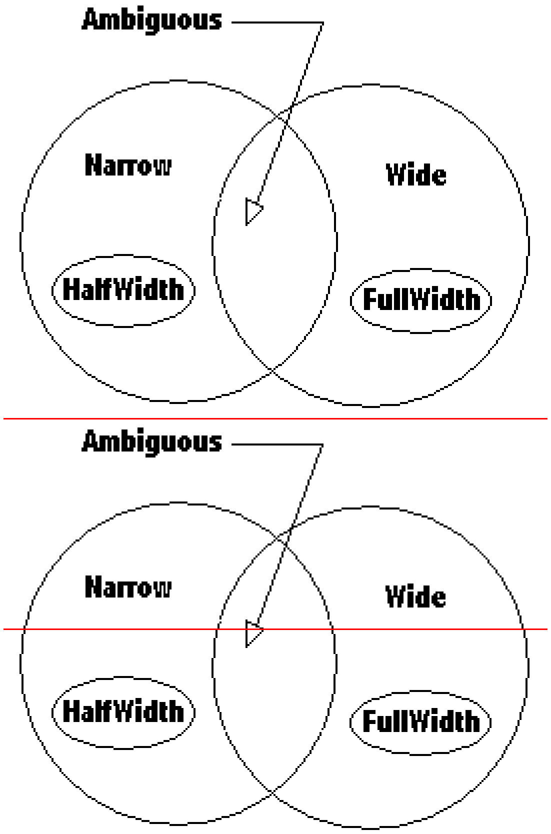

Figure 1: Venn diagram showing the set relations for five of the six categories.

Because East Asian legacy character sets do not always include complete case pairs of Latin characters, two members of a case pair may have different East Asian Width properties:

Ambiguous:	01D4	LATIN SMALL LETTER U WITH CARON
NEA Neutral:	01D3	LATIN CAPITAL LETTER U WITH CARON

ED7. Not East Asian (Neutral) – all other characters. Neutral characters do not occur in legacy East Asian character sets. By extension, they also do not occur in East Asian typography. For example, there is no traditional Japanese way of typesetting Devanagari.

Strictly speaking, it makes no sense to talk of narrow and wide for neutral characters, but since for all practical purposes they behave like Na, they are treated as narrow characters (the same as Na) under the recommendations below.

In a broad sense, *wide characters* include W, F, and A (when in EA context), while *narrow characters* include N, Na, H, and A (when not in EA context).

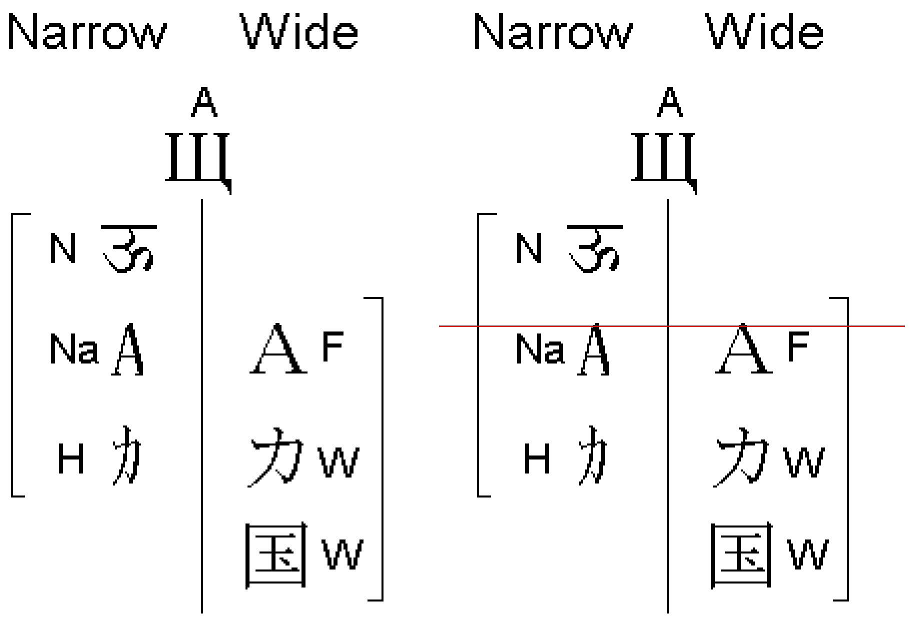

Figure 2: Examples for each character class and their resolved widths

4.1 Relation to the terms "full-width" and "half-width"

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, this 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 that 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 are neither half-width nor full-width.

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. However, since the single byte part of mixed-width character sets is limited, there are not going to be many candidates and neither UTC nor WG2 have any intention to add additional compatibility characters for this purpose.

4.2 Ambiguous width characters

Ambiguous width characters are all those characters that can occur as full-width characters in any of a number of East Asian legacy character encodings. They have a 'resolved' width of either narrow or wide depending on the context of their use. If they are not used in context of the specific legacy encoding they belong to, their width resolves to narrow. Otherwise it resolves to full-width or half-width. The term context as used here includes extra information such as explicit markup, knowledge of the source codepage, font information, or language identification. For example:

- Greek characters resolve to narrow when used with a standard Greek font, since there is no East Asian legacy context.
- Private use character codes and the replacement character have ambiguous width, since they may stand in for characters of any width.
- Ambiguous quotation marks are generally resolved to wide, when they enclose and are adjacent to a wide character, and to narrow otherwise.

Note: Modern practice is evolving towards rendering ever more of the ambiguous characters with proportionally spaced, narrow forms that rotate with the direction of writing, making a distinction *within* the character set. In other words, context information beyond the choice of font or source character set is employed to resolve the width of the character. This document does not attempt to track such changes in practice, therefore the set of characters that have been assigned ambiguous width form a superset of the set of characters that can be rendered as wide characters depending on context.

5 Conformance

East Asian Width is an informative character property and implies no conformance requirements.

6 Recommendation (informative)

When mapping Unicode to legacy character encodings

- Wide Unicode characters always map to full-width characters
- Narrow (and neutral) Unicode characters always map to half-width characters
- Half-width Unicode characters always map to half-width characters
- Ambiguous Unicode characters always map to full-width characters

- Wide Unicode characters never map to non-East Asian legacy character encodings
- Ambiguous Unicode characters always map to regular (narrow) characters in non-East Asian legacy character encodings

When processing or displaying data

- Wide characters behave like ideographs in important ways, such as layout. Except for certain punctuation, they are not rotated, when appearing in vertical text runs. 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~~.~~, f~~F~~or example, in line breaking. They are rotated sideways, when appearing in vertical text. In fixed pitched East Asian fonts, they take up 1/2 Em of space, but in rendering, a non-East Asian, proportional font is often substituted.~~.~~
- Ambiguous characters behave like wide or narrow characters depending on context (language tag, script identification, associated font, source of data, or explicit markup; all can provide the context).

7 Classifications (informative)

The classifications presented here are based on the most widely used mixed-width legacy character sets in use in East Asia as of this writing. In particular, the assignments of the neutral or ambiguous categories depend on the contents of these character sets. For example, an implementation that knows a-priori that it *only* needs to interchange data with the Japanese Shift-JIS character set, but not other East Asian character sets, could reduce the number of characters in the ambiguous classification to those actually encoded in Shift-JIS. Or such a reduction could be done implicitly at runtime in the context of interoperating with Shift-JIS fonts or data sources. Conversely, if additional character sets are created and widely adopted for legacy purposes, more characters would need to be classified as ambiguous.

7.1 Unassigned and Private Use characters

All unassigned characters are by default classified as non-East Asian neutral, except for the range ~~U-00020000 to U-0002FFFD,~~U+20000 to U+2FFFD, since all code positions from ~~U-00020000 to U-0002FFFD~~U+20000 to U+2FFFD are intended for CJK ideographs (W). All Private use characters are by default classified as ambiguous, since their definition depends on context.

7.2 Combining Marks

Combining marks have been classified and are given a property assignment based on their typical applicability. For example combining marks typically applied to characters of class N, Na or W are classified as A. Combining marks for purely non-East Asian scripts are marked as N, and non-spacing marks used only with wide characters are given a W. Even more so than for other characters, the *East Asian width* property for combining marks is not the same as their display width.

In particular, non-spacing marks do not possess actual advance width. Therefore, even when displaying combining marks, the East Asian Width property cannot be related to the advance width of these characters. However, it can be useful in determining the encoding length in a legacy encoding, or the choice of font for the range of characters including that non-spacing mark. The width of the glyph image of a non-spacing mark should always be chosen as the appropriate one for the width of the base character.

7.3 Data File

The East Asian Width classification of all Unicode characters at the time of this writing is available in the current version of the file EastAsianWidht.txt [Data] in the Unicode Character Database [UCD]. This is a tab-delimited, two column plain text file, with code position, East Asian Width designator. A comment at the end of each line indicates the character name. Ideographic, Hangul, Surrogate, and Private Use ranges are collapsed by giving a range in the first column.

As more character are added to the Unicode Standard, or if additional character sets are created and widely adopted for legacy purposes, the assignment of East Asian Width may be changed for some characters. Implementations should not make any assumptions to the contrary. Any future updates will be reflected in the latest version of the data file. (See the Unicode Character Database [UCD] for any specific version of the datafile).

7.4 Adding Characters

The sets of East Asian Narrow, East Asian Full-width and East Asian Half-width are fixed for all practical purposes. New characters for most scripts will be East Asian Neutral characters, unless the script is an East Asian script using wide characters, and then the new characters will be classified as East Asian Wide. One important exception consists of those new symbol characters that are, or are expected to be used both as wide characters in East Asian usage and narrow characters in non-East Asian usage. These would need to be classified as East Asian Ambiguous.

References

~~[Data]~~

~~http://www.unicode.org/Public/Unicode3.1.1-Update/EastAsianWidth-5.txt~~
~~The latest version of the data file is~~
~~http://www.unicode.org/Public/UNIDATA/EastAsianWidth.txt~~

[Data] The current version of the East Asian width property data file is
http://www.unicode.org/Public/3.2-Update/EastAsianWidth-6.txt
The latest version of the data file is
http://www.unicode.org/Public/UNIDATA/EastAsianWidth.txt

[FAQ] Unicode Frequently Asked Questions
http://www.unicode.org/unicode/faq/
For answers to common questions on technical issues.

[Glossary] Unicode Glossary
http://www.unicode.org/glossary/
For explanations of terminology used in this and other documents.

[Reports] Unicode Technical Reports
http://www.unicode.org/unicode/reports/
For information on the status and development process for technical reports, and for a list of technical reports.

[U3.0] *The Unicode Standard, Version 3.0,* (Reading, Massachusetts: Addison-Wesley Developers Press 2000) or online as http://www.unicode.org/unicode/uni2book/u2.html

[U3.1] Unicode Standard Annex #27: *Unicode 3.1*
http://www.unicode.org/unicode/reports/tr27/

~~[UCD] Unicode Character Database.~~
~~http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html~~
~~*For and overview of the Unicode Character Database and a list of its associated files*~~

[UCD] Unicode Character Database
http://www.unicode.org/ucd/
For an overview of the Unicode Character Database and a list of its associated files see
http://www.unicode.org/Public/UNIDATA/UnicodeCharacterDatabase.html

[Versions] Versions of the Unicode Standard
http://www.unicode.org/unicode/standard/versions/
For details on the precise contents of each version of the Unicode Standard, and how to cite them.

Acknowledgments

Michel Suignard provided extensive input into the analysis and source material for the detail assignments of these properties. Mark Davis and Ken Whistler performed consistency checks on the data files. Tomohiro Kubota reviewed the East Asian Width assignments against some common legacy encodings.

Modifications

Revision 10: [TBD: Subject to change as result of the changes to the document in discussion.] Reworded the definitions of F and H to explicitly refer to the compatibility decomposition. Changed 3000 from W to F, to align with the revised definitions. Extended the definition for Na to include all characters that have a wide equivalent, whether F or W. Changed 2329, 232A, 3008, 3009, 3018, 301B from A to W to reflect the addition of 27E6..27EB which are their Na equivalents. Also changed 3018..3019 from A to W to reflect their intended use as CJK-only punctuation.

Revision 9: Updated links to new version of data file. Changed 00AE, 014B, 02C4, 02DF, 2022, 2024, 203E, 2116, 2153, 215C..215D, 21B8..21B9, 21E7and 273D from N to A . This is a result of a recent review of existing mapping tables showing their use as wide characters in widely implemented East Asian legacy encodings. [Revision 9 was never published].

Revision 8: Change in header for Unicode 3.1. New status section and new format for references. Properties assigned to the new characters added to Unicode 3.1. Changed 2329..232A and 3008-3009 from N to A and W to A respectively. This is a result of their canonical equivalence.

Revision 7: Change in header for Unicode 3.0.1, change in file versioning format.

Version 6.0: Restated the definitions so that the wording more clearly reflects the intent. No changes to the assignments of properties to any character were made. Added a section on classifying characters that are to be added to the standard in the future. Also added figure 2.

Version 5.0: Changed the spelling of the title and made minor clarifying changes to the definitions and the description of ambiguous characters and combining marks. As result of the Unicode 3.0 beta process, changed some CJK punctuation characters from W to A since they are also used in Western mathematical notation. Removed some historic information and made other edits to prepare TR for publication as part of Unicode 3.0.