The Unicode Standard
Version 6.2 – Core Specification

To learn about the latest version of the Unicode Standard, see http://www.unicode.org/versions/latest/.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.2.
Includes bibliographical references and index.
ISBN 978-1-936213-07-8) (http://www.unicode.org/versions/Unicode6.2.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2012

ISBN 978-1-936213-07-8
Published in Mountain View, CA
September 2012

Chapter 17

About the Code Charts

> ***Disclaimer***
>
> Character images shown in the code charts are not prescriptive. In actual fonts, considerable variations are to be expected.

The online Unicode code charts present the characters of the Unicode Standard. This chapter explains the conventions used in the code charts and provides other useful information about the accompanying names lists.

Characters are organized into related groups called *blocks*. Many scripts are fully contained within a single character block, but other scripts, including some of the most widely used scripts, have characters divided across several blocks. Separate blocks contain common punctuation characters and different types of symbols.

A character names list follows the code chart for each block. The character names list itemizes every character in that block and provides supplementary information in many cases. A full set of character names, in machine-readable form, appears in the Unicode Character Database.

An index to distinctive character names can also be found on the Unicode Web site.

For information about access to the code charts and the character name index, see *Section B.6, Other Unicode Online Resources*.

17.1 Character Names List

The following illustration identifies the components of typical entries in the character names list.

code	*image*	*entry*	
00AE	®	REGISTERED SIGN	
		= registered trade mark sign (1.0)	*(Version 1.0 name)*
00AF	¯	MACRON	*(Unicode name)*
		= overline, APL overbar	*(alternative names)*
		• this is a spacing character	*(informative note)*
		→ 02C9 ¯ modifier letter macron	*(cross reference)*
		→ 0304 ◌̄ combining macron	
		→ 0305 ◌̅ combining overline	
		≈ 0020 SP 0304 ◌̄	*(compatibility decomposition)*
00E5	å	LATIN SMALL LETTER A WITH RING ABOVE	
		• Danish, Norwegian, Swedish, Walloon	*(sample of language use)*
		≡ 0061 a 030A ◌̊	*(canonical decomposition)*

2272 ≤ LESS-THAN OR EQUIVALENT TO
~ 2272 FE00 following the slant of the lower leg *(standardized variation sequence)*

Images in the Code Charts and Character Lists

Each character in these code charts is shown with a representative glyph. A representative glyph is not a prescriptive form of the character, but rather one that enables recognition of the intended character to a knowledgeable user and facilitates lookup of the character in the code charts. In many cases, there are more or less well-established alternative glyphic representations for the same character.

Designers of high-quality fonts will do their own research into the preferred glyphic appearance of Unicode characters. In addition, many scripts require context-dependent glyph shaping, glyph positioning, or ligatures, none of which is shown in the code charts. The Unicode Standard contains many characters that are used in writing minority languages or that are historical characters, often used primarily in manuscripts or inscriptions. Where there is no strong tradition of printed materials, the typography of a character may not be settled. Because of these factors, the glyph image chosen as the representative glyph in these code charts should not be considered a definitive guide to best practice for typographical design.

Fonts. The representative glyphs for the Latin, Greek, and Cyrillic scripts in the code charts are based on a serifed, Times-like font. For non-European scripts, typical typefaces were selected that allow as much distinction as possible among the different characters.

The fonts used for other scripts are similar to Times in that each represents a common, widely used design, with variable stroke width and serifs or similar devices, where applicable, to show each character as distinctly as possible. Sans-serif fonts with uniform stroke width tend to have less visibly distinct characters. In the code charts, sans-serif fonts are used for archaic scripts that predate the invention of serifs, for example.

Alternative Forms. Some characters have alternative forms. For example, even the ASCII character U+0061 LATIN SMALL LETTER A has two common alternative forms: the "a" used in Times and the "ɑ" that occurs in many other font styles. In a Times-like font, the character U+03A5 GREEK CAPITAL LETTER UPSILON looks like "Y"; the form ϒ is common in other font styles.

A different case is U+010F LATIN SMALL LETTER D WITH CARON, which is commonly typeset as ď instead of ǎ. In such cases, the code charts show the more common variant in preference to a more didactic archetypical shape.

Many characters have been unified and have different appearances in different language contexts. The shape shown for U+2116 № NUMERO SIGN is a fullwidth shape as it would be used in East Asian fonts. In Cyrillic usage, № is the universally recognized glyph. See *Figure 15-2*.

In certain cases, characters need to be represented by more or less condensed, shifted, or distorted glyphs to make them fit the format of the code charts. For example, U+0D10 ഐ MALAYALAM LETTER AI is shown in a reduced size to fit the character cell.

When characters are used in context, the surrounding text gives important clues as to identity, size, and positioning. In the code charts, these clues are absent. For example, U+2075 ^{5} SUPERSCRIPT FIVE is shown much smaller than it would be in a Times-like text font.

Whenever a more obvious choice for representative glyph may be insufficient to aid in the proper identification of the encoded character, a more distinct variant has been selected as representative glyph instead.

Orientation. Representative glyphs for characters in the code charts are oriented as they would appear in normal text. This is true regardless of whether the script in question is predominantly laid out in horizontal lines, as for most scripts, or is predominantly laid out in vertical lines, as for Mongolian and Phags-pa. In cases such as Mongolian, this is accomplished using specialized chart fonts which show the glyphs correctly oriented, even though the chart production software lays out all glyphs horizontally in their boxes. Note that commercial production fonts for Mongolian do not behave this way; if used with common charting tools, including those for the Unicode code charts, such fonts will show Mongolian glyphs with their images turned 90 degrees counterclockwise.

Special Characters and Code Points

The code charts and character lists use a number of notational conventions for the representation of special characters and code points. Some of these conventions indicate those code points which are *not* assigned to encoded characters, or are permanently reserved. Other conventions convey information about the type of character encoded, or provide a possible fallback rendering for non-printing characters.

Combining Characters. Combining characters are shown with a dotted circle. This dotted circle is not part of the representative glyph and it would not ordinarily be included as part of any actual glyph for that character in a font. Instead, the relative position of the dotted circle indicates an an approximate location of the base character in relation to the combining mark.

093F ◌ि DEVANAGARI VOWEL SIGN I
• stands to the left of the consonant

0940 ◌ी DEVANAGARI VOWEL SIGN II

0941 ◌ु DEVANAGARI VOWEL SIGN U

The detailed rules for placement of combining characters with respect to various base characters are implemented by the selected font in conjunction with the rendering system.

During rendering, additional adjustments are necessary. Accents such as U+0302 COMBINING CIRCUMFLEX ACCENT are adjusted vertically and horizontally based on the height and width of the base character, as in " î " versus "Ŵ".

If the display of a combining mark with a dotted circle is desired, U+25CC ◌ DOTTED CIRCLE is often chosen as the base character for the mark.

Dashed Box Convention. There are a number of characters in the Unicode Standard which in normal text rendering have no visible display, or whose only effect is to modify the display of *other* characters in proximity to them. Examples include space characters, control characters, and format characters.

To make such characters easily recognizable and distinguishable in the code charts and in any discussion about the characters, they are represented by a square dashed box. This box surrounds a short mnemonic abbreviation of the character's name.

0020 SP SPACE
• sometimes considered a control code
• other space characters: 2000 NQSP - 200A HSP

Where such characters have a typical visual appearance in some contexts, an additional representative image may be used, either alone or with a mnemonic abbreviation.

00AD SHY SOFT HYPHEN
= discretionary hyphen
• commonly abbreviated as SHY

This convention is also used for some graphic characters which are only distinguished by special behavior from another character of the same appearance.

2011 NB‑ NON-BREAKING HYPHEN
→ 002D - hyphen-minus
→ 00AD SHY soft hyphen
≈ <noBreak> 2010 -

The dashed box convention also applies to the glyphs of combining characters which have no visible display of their own, such as variation selectors (see *Section 16.4, Variation Selectors*).

FE00 VS 1 VARIATION SELECTOR-1
• these are abbreviated VS1, and so on

Sometimes, the combining status of the character is indicated by including a dotted circle inside the dashed box, for example for the consonant-stacking viramas.

17D2 ◌្ KHMER SIGN COENG
• functions to indicate that the following Khmer letter is to be rendered subscripted
• shape shown is arbitrary and is not visibly rendered

Even though the presence of the dashed box in the code charts indicates that a character is likely to be a space character, a control character, a format character, or a combining character, it cannot be used to infer the actual General_Category value of that character.

Reserved Characters. Character codes that are marked "<reserved>" are unassigned and reserved for future encoding. Reserved codes are indicated by a ▨ glyph. To ensure readability, many instances of reserved characters have been suppressed from the names list. Reserved codes may also have cross references to assigned characters located elsewhere.

2073 ▨ <reserved>
→ 00B3 ³ superscript three

Noncharacters. Character codes that are marked "<not a character>" refer to noncharacters. They are designated code points that will never be assigned to a character. These codes are indicated by a ■ glyph. Noncharacters are shown in the code charts only where they occur together with other characters in the same block. For a complete list of noncharacters, see *Section 16.7, Noncharacters*.

FFFF ■ <not a character>
• the value FFFF ■ is guaranteed not to be a Unicode character at all

Deprecated Characters. Deprecated characters are characters whose use is strongly discouraged, but which are retained in the standard indefinitely so that existing data remain well defined and can be correctly interpreted. (See D13 in *Section 3.4, Characters and Encoding*.) Deprecated characters are explicitly indicated in the Unicode Code Charts using annotations or subheads.

Character Names

The character names in the code charts precisely match the normative character names in the Unicode Character Database. Character names are unique and stable. By convention, they are in uppercase. For more information on character names, see *Section 4.8, Name*.

Informative Aliases

An informative alias (preceded by =) is an alternate name for a character. Characters may have several aliases, and aliases for different characters are not guaranteed to be unique. Aliases are informative and may be updated. By convention, aliases are in lowercase, except where they contain proper names. Where an alias matches the name of a character in *The*

Unicode Standard, Version 1.0, it is listed first, followed by "1.0" in parentheses. Because the formal character names may differ in unexpected ways from commonly used names (for example, PILCROW SIGN = paragraph sign), some aliases may be useful alternate choices for indicating characters in user interfaces. In the Hangul Jamo block, U+1100..U+11FF, the normative short jamo names are given as aliases.

Normative Aliases

A normative character name alias (one preceded by ※) is a formal, unique, and stable alternate name for a character. Characters are given normative character name aliases in certain cases where there is a defect in the character name. These aliases do not replace the character name, but rather allow users to formally refer to the character without requiring the use of a defective name. Normative character name aliases which provide information about corrections to defective character names are always printed in the character names list. Normative aliases serving other purposes, such as defining abbreviations for characters, may be omitted or may be presented with an alternative symbol to distinguish them. For a definite list, suitable for machine parsing, see, NameAliases.txt in the UCD. For more information, see *Section 4.8, Name*. By convention, normative character aliases are written in uppercase letters.

FE18 ﹈ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRAKCET
※ PRESENTATION FORM FOR VERTICAL RIGHT WHITE LENTICULAR BRACKET
• misspelling of "BRACKET" in character name is a known defect
≈ <vertical> 3017

Cross References

Cross references (preceded by →) are used to indicate a related character of interest, but without indicating the nature of the relation. Possibilities are a different character of similar appearance or name, the other member of a case pair, or some other linguistic relationship.

Explicit Inequality. The two characters are not identical, although the glyphs that depict them are identical or very close.

003A : COLON
→ 0589 : armenian full stop
→ 2236 : ratio

Other Linguistic Relationships. These relationships include transliterations (such as between Serbian and Croatian), typographically unrelated characters used to represent the same sound, and so on.

01C9 lj LATIN SMALL LETTER LJ
→ 0459 љ cyrillic small letter lje
≈ 006C l 006A j

Cross references are neither exhaustive nor symmetric. Typically a general character would have cross references to more specialized characters, but not the other way around.

Information About Languages

An informative note may include a list of one or more of the languages using that character where this information is considered useful. For case pairs, the annotation is given only for the lowercase form to avoid needless repetition. An ellipsis "..." indicates that the listed languages cited are merely the principal ones among many.

Case Mappings

When a case mapping corresponds *solely* to a difference based on SMALL versus CAPITAL in the names of the characters, the case mapping is not given in the names list but only in the Unicode Character Database.

0041 A LATIN CAPITAL LETTER A

01F2 Dz LATIN CAPITAL LETTER D WITH SMALL LETTER Z
≈ 0044 D 007A z

When the case mapping cannot be predicted from the name, the casing information is sometimes given in a note.

00DF ß LATIN SMALL LETTER SHARP S
= Eszett
• German
• uppercase is "SS"
• in origin a ligature of 017F ſ and 0073 s
→ 03B2 β greek small letter beta

For more information about case and case mappings, see *Section 4.2, Case*.

Decompositions

The decomposition sequence (one or more letters) given for a character is either its canonical mapping or its compatibility mapping. The canonical mapping is marked with an *identical to* symbol ≡.

00E5 å LATIN SMALL LETTER A WITH RING ABOVE
• Danish, Norwegian, Swedish, Walloon
≡ 0061 a 030A ◌̊

212B Å ANGSTROM SIGN
≡ 00C5 Å latin capital letter a with ring above

Compatibility mappings are marked with an *almost equal to* symbol ≈. Formatting information may be indicated with a formatting tag, shown inside angle brackets.

01F2 Dz LATIN CAPITAL LETTER D WITH SMALL LETTER Z
≈ 0044 D 007A z

FF21 A FULLWIDTH LATIN CAPITAL LETTER A
≈ <wide> 0041 A

The following compatibility formatting tags are used in the Unicode Character Database:

<font>	A font variant (for example, a blackletter form)
<noBreak>	A no-break version of a space, hyphen, or other punctuation
<initial>	An initial presentation form (Arabic)
<medial>	A medial presentation form (Arabic)
<final>	A final presentation form (Arabic)
<isolated>	An isolated presentation form (Arabic)
<circle>	An encircled form
<super>	A superscript form
<sub>	A subscript form
<vertical>	A vertical layout presentation form
<wide>	A fullwidth (or zenkaku) compatibility character
<narrow>	A halfwidth (or hankaku) compatibility character
<small>	A small variant form (CNS compatibility)

<square>	A CJK squared font variant
<fraction>	A vulgar fraction form
<compat>	Otherwise unspecified compatibility character

In the character names list accompanying the code charts, the "<compat>" label is suppressed, but all other compatibility formatting tags are explicitly listed in the compatibility mapping.

Decomposition mappings are not necessarily full decompositions. For example, the decomposition for U+212B Å ANGSTROM SIGN can be further decomposed using the canonical mapping for U+00C5 Å LATIN CAPITAL LETTER A WITH RING ABOVE. (For more information on decomposition, see *Section 3.7, Decomposition*.)

Compatibility decompositions do not attempt to retain or emulate the formatting of the original character. For example, compatibility decompositions with the <noBreak> formatting tag do not use U+2060 WORD JOINER to emulate nonbreaking behavior; compatibility decompositions with the <circle> formatting tag do not use U+20DD COMBINING ENCLOSING CIRCLE; and compatibility decompositions with formatting tags <initial>, <medial>, <final>, or <isolate> for explicit positional forms do not use ZWJ or ZWNJ. The one exception is the use of U+2044 FRACTION SLASH to express the <fraction> semantics of compatibility decompositions for vulgar fractions.

Standardized Variation Sequences

Characters for which one or more standardized variants have been defined are displayed in the code charts with a special convention: the code chart cell for such characters has a small black triangle in its upper-right corner.

In the character names list, each variation sequence for standardized variants is listed in the entry for the base character for that sequence. In some cases a character may be associated with multiple variation sequences. A standardized variation sequence is identified in the character names list with an initial tilde symbol "~".

The list of standardized variation sequences in the character names list exactly matches the list defined in the data file StandardizedVariants.txt in the Unicode Character Database. Ideographic variation sequences defined in the Ideographic Variation Database are not included. See *Section 16.4, Variation Selectors* for more information.

Subheads

The character names list contains a number of informative subheads that help divide up the list into smaller sublists of similar characters. For example, in the Miscellaneous Symbols block, U+2600..U+26FF, there are subheads for "Astrological symbols," "Chess symbols," and so on. Such subheads are editorial and informative; they should not be taken as providing any definitive, normative status information about characters in the sublists they mark or about any constraints on what characters could be encoded in the future at reserved code points within their ranges. The subheads are subject to change.

17.2 CJK Unified and Compatibility Ideographs

The code charts for CJK ideographs differ significantly from those for other characters in the standard.

CJK Unified Ideographs

Character names are not provided for any of the code charts of CJK Unified Ideograph character blocks, because the name of a unified ideograph simply consists of its Unicode code point preceded by CJK UNIFIED IDEOGRAPH-.

In other code charts each character is represented with a single representative glyph, but in the code charts for CJK Unified and Compatibility Ideographs, each character may have multiple representative glyphs. Each character is shown with as many representative glyphs as there are Ideographic Rapporteur Group (IRG) sources defined for that character. Each representative glyph is accompanied with its detailed source information provided in alphanumeric form. Altogether, there are nine IRG sources, as shown in *Table 17-1*. Data for these IRG sources are also documented in Unicode Standard Annex #38, "Unicode Han Database (Unihan)".

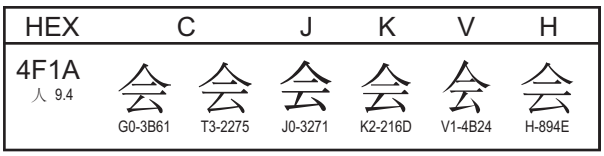
Table 17-1. IRG Sources

Name	Source Identity
G source	China PRC and Singapore
H source	Hong Kong SAR
J source	Japan
KP source	North Korea
K source	South Korea
M source	Macau SAR
T source	Taiwan
U source	Unicode/USA
V source	Vietnam

To assist in reference and lookup, each CJK Unified Ideograph is accompanied by a representative glyph of its Unicode radical and by its Unicode radical-stroke counts. These are printed directly underneath the Unicode code point for the character. A radical-stroke index to all of the CJK ideographs is also provided separately on the Unicode Web site.

Chart for the Main CJK Block. For the CJK Unified Ideographs block (U+4E00..U+9FFF) the glyphs are arranged in the following order: G source and T sources are grouped under the header "C," and J, K, V and H sources are listed under their respective headers. Each row contains positions for all six sources, and if a particular source is undefined for CJK Unified Ideograph, that position is left blank in the row. This format is illustrated by *Figure 17-1*. If a character also has a U source, an additional line is used for that character. Note that this block does not contain any characters with M sources. The KP sources are not shown due to lack of reliable glyph information.

Figure 17-1. CJK Chart Format for the Main CJK Block

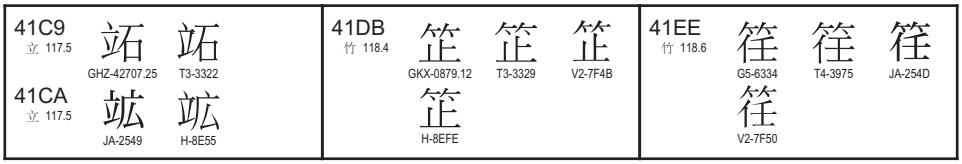


Charts for CJK Extensions. The code charts for all of the extension blocks for CJK Unified Ideographs use a more condensed format. That format dispenses with the "C, J, K, V, and H" headers and leaves no holes for undefined sources. For those blocks, sources are always shown in the following order: G, T, J, K, KP, V, H, M, and U. The first letters of the source information provide the source type for all sources except G. (For Unicode 6.0, KP sources

are omitted from the code charts because of the lack of an appropriately vetted font for display.)

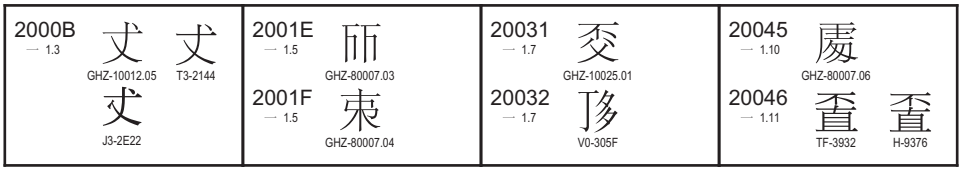
The multicolumn code chart for the CJK Unified Ideographs Extension A block (U+3400..U+4DBF) uses the condensed format with three source columns per entry, and with entries arranged in three columns per page. An entry may have additional rows, if it is associated with more than three sources, as illustrated in *Figure 17-2*.

Figure 17-2. CJK Chart Format for CJK Extension A



The multicolumn code charts for the other extension blocks for CJK Unicode Ideographs use the condensed format with two source columns per entry, and with entries arranged in four columns per page. An entry may have additional rows, if it is associated with more than two sources, as illustrated in *Figure 17-3*.

Figure 17-3. CJK Chart Format for CJK Extension B



Compatibility Ideographs

The format of the code charts for the CJK Compatibility Ideograph blocks is largely similar to the CJK chart format for Extension A. However, several additional notational elements described in *Section 17.1, Character Names List* are used. In particular, for each CJK compatibility ideograph other than the small list of unified ideographs included in these charts, a canonical decomposition is shown. Each CJK unified ideograph in these charts has an annotation identifying it as such. Character names are not provided for any CJK Compatibility Ideograph blocks because the name of a compatibility ideograph simply consists of its Unicode code point preceded by CJK COMPATIBILITY IDEOGRAPH-.

Figure 17-4. CJK Chart Format for Compatibility Ideographs

17.3 Hangul Syllables

As in the case of CJK Unified Ideographs, a character names list is not provided for the online chart of characters in the Hangul Syllables block, U+AC00..U+D7AF, because the name of a Hangul syllable can be determined by algorithm as described in *Section 3.12,*

Conjoining Jamo Behavior. The short names used in that algorithm are listed in the code charts as aliases in the Hangul Jamo block, U+1100..U+11FF, as well as in Jamo.txt in the Unicode Character Database.