Chapter 14

Code Charts

Disclaimer

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

The code charts that follow present the characters of the Unicode Standard. Characters are organized into related groups called *blocks*. In the Unicode Standard, character blocks generally contain characters from a single script. In many cases, a script is fully represented in its character block. There are, however, important exceptions, most notably in the area of punctuation characters.

A character names list follows each character chart, except for CJK ideographs and Hangul syllables, as discussed in *Section 14.2, CJK Unified Ideographs*, and *Section 14.3, Hangul Syllables*. The character names list itemizes every character in the block and provides supplementary information in many cases.

An index to distinctive character names is at the back of this book; a full set of character names (including earlier Version 1.0 names) are in the *Unicode Character Database* on the CD-ROM.

14.1 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	*(Version 1.0 name)*
00AF	¯	MACRON	*(Unicode name)*
		= overline, APL overbar	*(alternative names)*
		• this is a spacing character	*(informative note)*
		→ 02C9 ˉ 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
≡ 0061 a 030A ◌̊ *(canonical decomposition)*

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 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 representative glyphs in the code charts are based on a serifed, Times-like font. 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 d’ instead of ď. 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.

In many 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.

Sometimes characters need to be given artificial shapes to make them recognizable in the code charts. Examples are U+00AD SHY SOFT HYPHEN and U+2011 NB NON-BREAKING HYPHEN, where the special behavior of the hyphen is indicated by the dashed box and the letters.

When characters are used in context, the surrounding text will give 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.

Combining characters are shown with a dotted circle—for example, U+0940 ◌ी DEVANAGARI VOWEL SIGN II. The relative position of the dotted circle gives an approximate indication of the location of the base character in relation to the combining mark. 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”.

For non-European scripts, typical typefaces were selected that allow as much distinction as possible among the different characters.

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.

Cross References

Cross-referenced characters (preceded by →) have various characteristics: explicit inequality, 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

Case Form 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* on the CD-ROM.

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 information is 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

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 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 black letter 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 wide (or zenkaku) compatibility character
<narrow>	A narrow (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.

Decompositions 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 small letter a with ring above. (For more information on decomposition, see *Section 3.6, Decomposition*.)

Information About Languages

An informative note may include a list of the language(s) 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.

Reserved Characters

Character codes that are marked "<reserved>" are unassigned and reserved for future encoding. Reserved codes are indicated by a ▨ glyph.

060D ▨ <reserved>

Reserved codes may also have cross references to assigned characters located elsewhere.

2073 ▨ <reserved>
→ 00B3 ^{3} superscript three

Character codes that are marked "<not a character>" are permanently unassigned; they will never be assigned a character. These codes are indicated by a ■ glyph.

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

14.2 CJK Unified Ideographs

A character names list is not provided for the *CJK Unified Ideographs* and *CJK Unified Ideographs Vertical Extension A* character blocks because the name of a unified ideograph simply consists of its Unicode value preceded by CJK UNIFIED IDEOGRAPH-.

As with other character charts, each Unicode character in these blocks is shown with its Unicode value and a single representative glyph. Note that varying typographic practices throughout East Asia may require glyphs other than the representative one to be used so that the display is correct for a particular country or language.

A table providing mappings between the CJK ideographs included in the Unicode Standard and those in other character set standards is included on the CD-ROM.

A radical-stroke index to CJK ideographs is in *Section 15.1, Han Radical-Stroke Index*.

An index in Shift-JIS order of the ideographs in JIS X 0208 can be found in *Section 15.2, Shift-JIS Index*.

14.3 Hangul Syllables

A character names list is not provided for characters in the *Hangul Syllables Area* (U+AC00..U+D7A3) because the name of a Hangul syllable can be determined by algorithm as described in *Section 3.11, Conjoining Jamo Behavior*.

The code charts, pages 336-846 in the book, are omitted here. Please see the online code charts at http://www.unicode.org/charts/

Note: The online code charts are continuously updated and may contain characters added after the publication of *The Unicode Standard, Version 3.0*. To find out whether a particular character was part of the Unicode Standard, Version 3.0, please consult either the printed edition of the standard (ISBN 0-201-61633-5) or version 3.0.0 of the Unicode Character Database. Normative references to the Unicode Standard, Version 3.0 should use the printed edition.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The authors and publisher have taken care in preparation of this book, but make no expressed or implied warranty of any kind and assume no responsibility for errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of the use of the information or programs contained herein.

The *Unicode Character Database* and other files are provided as-is by Unicode®, Inc. No claims are made as to fitness for any particular purpose. No warranties of any kind are expressed or implied. The recipient agrees to determine applicability of information provided.

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

ISBN 0-201-61633-5

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: http://www.awl.com/cseng/

First printing, January 2000.