Chapter 6

Punctuation

The appearance and usage of punctuation marks varies between languages and scripts, but punctuation marks have a common function. They separate units of text, such as sentences and phrases, thus clarifying the meaning of the text. The use of punctuation marks is not limited to prose; they are also used in mathematical and scientific formulae, for example. In the Unicode Standard, punctuation marks are called punctuation characters.

In the standard, characters are organized into related groups called blocks. 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. Punctuation characters occur in several widely separated places in the character blocks, including Basic Latin, Latin-1 Supplement, General Punctuation, and CJK Symbols and Punctuation, as well as occasional characters in character blocks for specific scripts.

Punctuation characters—for example, U+002C comma or U+2022 bullet—are encoded only once, rather than being encoded again and again for particular scripts; such general-purpose punctuation may be used for any script or mixture of scripts. In contrast, punctuation characters that are encoded in a given script block—for example, U+058A armenian hyphen or U+060C arabic comma—are intended primarily for use in the context of that script. They are unique in function, have different directionality, or are distinct in appearance or usage from their generic counterparts.

The use and interpretation of punctuation characters can be heavily context-dependent. For example, U+002E full stop can be used as sentence-ending punctuation, an abbreviation indicator, a decimal point, and so on.

Punctuation characters vary in appearance with the font style, just like the surrounding text characters. In some cases, where used in the context of a particular script, a specific glyph style is preferred. For example, U+002E full stop should appear square when used with Armenian, but is typically circular when used with Latin. For mixed Latin/Armenian text, two fonts (or one font allowing for context-dependent glyph variation) may need to be used to faithfully render the character.

In a bidirectional context (see *Section 3.12, Bidirectional Behavior*), shared punctuation characters have no inherent directionality, but resolve according to the Unicode bidirectional algorithm. Where the image of a punctuation character is not bilaterally symmetric, the mirror image is used when the character is part of the right-to-left text stream (see *Section 4.7, Mirrored—Normative*). In vertical writing, many punctuation characters have special vertical glyphs.

A number of characters in the blocks described in this chapter are not graphic punctuation characters, but nevertheless affect the operation of layout algorithms. For a description of these characters, see *Section 13.2, Layout Controls*.

6.1 General Punctuation

Punctuation: U+0020–U+00BF

Standards. The Unicode Standard adapts the ASCII (ISO 646) 7-bit standard by retaining the semantics and numeric code values, merely supplying enough leading zeros to convert them into 16-bit values. The content and arrangement of the ASCII standard are far from optimal in the context of a 16-bit space, but the Unicode Standard retains it without change because of its prevalence in existing usage. The ASCII (ANSI X3.4) standard is identical to ISO/IEC 646:1991-IRV.

ASCII Graphic Characters. Some of the nonletter characters in this range suffer from overburdened usage as a result of the limited number of codes in a 7-bit space. Some coding consequences of this problem are discussed in this section under “Encoding Characters with Multiple Semantic Values” and “General Punctuation,” but also see “Language-Based Use of Quotation Marks.” The rather haphazard ASCII collection of punctuation and mathematical signs are isolated from the larger body of Unicode punctuation, signs, and symbols (which are encoded in ranges starting at U+2000) only because the relative locations within ASCII are so widely used in standards and software.

Typographic Variations. Code values in the ASCII range are well established and used in widely varying implementations. The Unicode Standard therefore provides only minimal specifications on the typographic appearance of corresponding glyphs. For example, the value U+0024 ($) (derived from ASCII 24) has the semantic *dollar sign*, leaving open the question of whether the dollar sign is to be rendered with one vertical stroke or two. The Unicode value U+0024 refers to the *dollar sign semantic*, not to its precise appearance.

Likewise, in old-style numerals, where numbers vary in placement above and below the baseline, a decimal or thousands separator may be displayed with a dot that is raised above the baseline. Because it would be inadvisable to have a stylistic variation between old-style and new-style numerals that actually changes the underlying representation of text, the Unicode Standard considers this raised dot to be merely a glyphic variant of U+002E “.” FULL STOP. For other characters in this range that have alternative glyphs, the Unicode character is displayed with the basic or most common glyph; rendering software may present any other graphical form of that character.

Encoding Characters with Multiple Semantic Values. Some ASCII characters have multiple uses, either through ambiguity in the original standards or through accumulated reinterpretations of a limited code set. For example, 27 hex is defined in ANSI X3.4 as *apostrophe* (*closing single quotation mark; acute accent*), and 2D hex is defined as *hyphen-minus*. In general, the Unicode Standard provides the same interpretation for the equivalent code values, without adding to or subtracting from their semantics. The Unicode Standard supplies *unambiguous* codes elsewhere for the most useful particular interpretations of these ASCII values; the corresponding unambiguous characters are cross-referenced in the character names list for this block. For a complete list of space characters and dash characters in the Unicode Standard, see “General Punctuation” later in this section.

For historical reasons, U+0027 is a particularly overloaded character. In ASCII, it is used to represent a punctuation mark (such as right single quotation mark, left single quotation mark, apostrophe punctuation, vertical line, or prime) or a modifier letter (such as apostrophe modifier or acute accent). Punctuation marks generally break words; modifier letters generally are considered part of a word.

The preferred character for apostrophe is U+2019, but U+0027 is commonly present on keyboards. In modern software, it is therefore common to substitute U+0027 by the appropriate character in input. In these systems, a U+0027 in the data stream is always represented as a straight vertical line and can never represent a curly apostrophe or a right quotation mark. For more information, see “Apostrophes” later in this section.

Semantics of Paired Punctuation. Paired punctuation marks such as parentheses (U+0028, U+0029), square brackets (U+005B, U+005D), and braces (U+007B, U+007D) are interpreted semantically rather than graphically in the context of bidirectional or vertical texts; that is, *these characters have consistent semantics but alternative glyphs depending upon the directional flow rendered by a given software program.* The software must ensure that the rendered glyph is the correct one. When interpreted semantically rather than graphically, characters containing the qualifier “LEFT” are taken to denote *opening*; characters containing the qualifier “RIGHT” are taken to denote *closing*. For example, U+0028 LEFT PARENTHESIS and U+0029 RIGHT PARENTHESIS are interpreted as opening and closing parentheses, respectively, in the context of bidirectional or vertical texts. In a right-to-left directional flow, U+0028 is rendered as “)”. In a left-to-right flow, the same character is rendered as “(”. See also “Language-Based Use of Quotation Marks” later in this section.

Tilde. U+007E TILDE can be used either as a Spacing Clone of Combining Tilde (see “Spacing Clones of Diacritics” in *Section 7.1, Latin*) or more often as a center line tilde, similar in appearance to U+223C “∼” TILDE OPERATOR. Two common uses are to indicate an approximate value or in dictionaries to repeat the defined term in the definition of the ~. Although U+007E “~” TILDE is ambiguous in its rendering, modern fonts generally render it with a center line glyph, as shown in the code charts.

General Punctuation: U+2000–U+206F

General punctuation combines punctuation characters and characterlike elements used to achieve certain text layout effects. Some punctuation characters can be used with many different scripts. Many general punctuation characters can also be found in the Basic Latin (ASCII) and Latin-1 Supplement blocks.

In many cases, current standards include generic characters for punctuation instead of the more precisely specified characters used in printing. Examples include the single and double quotes, period, dash, and space. The Unicode Standard includes these generic characters, but also encodes the unambiguous characters independently: various forms of quotation mark, decimal period, em dash, en dash, minus, hyphen, em space, en space, hair space, zero-width space, and so on.

Punctuation principally used with a specific script is found in the block corresponding to that script, such as U+061B “؛” ARABIC SEMICOLON or the punctuation used with ideographs in the CJK Symbols and Punctuation Block.

Space Characters

The most commonly used space character is U+0020 SPACE. Also often used is its non-breaking counterpart, U+00A0 NO-BREAK SPACE. These two characters have the same width, but behave differently for line breaking. U+00A0 NO-BREAK SPACE behaves like a numeric separator for the purposes of bidirectional layout. (See *Section 3.12, Bidirectional Behavior*, for a detailed discussion of the Unicode bidirectional algorithm.) In ideographic text, U+3000 IDEOGRAPHIC SPACE is commonly used because its width matches that of the ideographs.

The main difference among other space characters is their width. U+2000..U+2006 are standard quad widths used in typography. U+2007 FIGURE SPACE has a fixed width, known as *tabular width*, which is the same width as digits used in tables. U+2008 PUNCTUATION SPACE is a space defined to be the same width as a period. U+2009 THIN SPACE and U+200A HAIR SPACE are successively smaller-width spaces used for narrow word gaps and for justification of type. The fixed-width space characters (U+2000..U+200A) are derived from conventional (hot lead) typography. Algorithmic kerning and justification in computerized typography do not use these characters. However, where they are used, as, for example, in typesetting mathematical formulae, their width is generally font-specified, and they typically do not expand during justification. The exception is U+2009 THIN SPACE, which sometimes gets adjusted.

Space characters with special behavior in word or line breaking are described in "Line and Word Breaking" in *Section 13.2, Layout Controls*.

Space characters may also be found in other character blocks in the Unicode Standard. The list of space characters appears in *Table 6-1*.

Table 6-1. Unicode Space Characters

Code	Name
U+0020	SPACE
U+00A0	NO-BREAK SPACE
U+2000	EN QUAD
U+2001	EM QUAD
U+2002	EN SPACE
U+2003	EM SPACE
U+2004	THREE-PER-EM SPACE
U+2005	FOUR-PER-EM SPACE
U+2006	SIX-PER-EM SPACE
U+2007	FIGURE SPACE
U+2008	PUNCTUATION SPACE
U+2009	THIN SPACE
U+200A	HAIR SPACE
U+200B	ZERO WIDTH SPACE
U+202F	NARROW NO-BREAK SPACE
U+3000	IDEOGRAPHIC SPACE

U+200B ZERO WIDTH SPACE, as well as several spacelike, zero-width characters with special properties, are described in *Section 13.2, Layout Controls.*

Dashes and Hyphens

Because of its prevalence in legacy encodings, U+002D HYPHEN-MINUS is the most common of the dash characters used to represent a hyphen. It has ambiguous semantic value and is rendered with an average width. U+2010 HYPHEN represents the hyphen as found in words such as "left-to-right." It is rendered with a narrow width. When typesetting text, U+2010 HYPHEN is preferred over U+002D HYPHEN-MINUS. U+2011 NON-BREAKING HYPHEN is present for compatibility with existing standards. It has the same semantic value as U+2010 HYPHEN, but should not be broken across lines.

U+2012 FIGURE DASH is present for compatibility with existing standards; it has the same (ambiguous) semantic as the U+002D HYPHEN-MINUS, but has the same width as digits (if they are monospaced). U+2013 EN DASH is used to indicate a range of values, such as 1973–1984. It should be distinguished from the U+2122 MINUS, which is an arithmetic operator; however, typographers have typically used U+2013 EN DASH in typesetting to represent the *minus sign*. For general compatibility in interpreting formulas, U+002D HYPHEN-MINUS, U+2012 FIGURE DASH, and U+2212 MINUS SIGN should each be taken as indicating a *minus sign*, as in "x = a - b."

U+2014 EM DASH is used to make a break—like this—in the flow of a sentence. It is commonly represented with a typewriter as a double-hyphen. In older mathematical typography, U+2014 EM DASH is also used to indicate a *binary minus sign*. U+2015 HORIZONTAL BAR is used to introduce quoted text in some typographic styles.

Dashes and hyphen characters may also be found in other character blocks in the Unicode Standard. A list of dash and hyphen characters appears in *Table 6-2*. For a description of the line-breaking behavior of dashes and hyphens, see Unicode Technical Report #14, "Line Breaking Properties," on the CD-ROM or the up-to-date version on the Unicode Web site.

Table 6-2. Unicode Dash Characters

Code	Name
U+002D	HYPHEN-MINUS
U+007E	TILDE (= *swung dash*)
U+00AD	SOFT HYPHEN
U+058A	ARMENIAN HYPHEN
U+1806	MONGOLIAN TODO SOFT HYPHEN
U+2010	HYPHEN
U+2011	NON-BREAKING HYPHEN
U+2012	FIGURE DASH
U+2013	EN DASH
U+2014	EM DASH
U+2015	HORIZONTAL BAR (= *quotation dash*)
U+207B	SUPERSCRIPT MINUS
U+208B	SUBSCRIPT MINUS
U+2212	MINUS SIGN
U+301C	WAVE DASH
U+3030	WAVY DASH

Language-Based Usage of Quotation Marks

U+0022 QUOTATION MARK is the most commonly used character for quotation mark. However, it has ambiguous semantics and direction. Word processors commonly offer a facility for automatically converting the U+0022 QUOTATION MARK to a contextually-selected curly quote glyph.

Low Quotation Marks. U+201A SINGLE LOW-9 QUOTATION MARK and U+201E DOUBLE LOW-9 QUOTATION MARK are unambiguously opening quotation marks. All other quotation marks have heterogeneous semantics. They may represent opening or closing quotation marks depending on the usage.

European Usage. The use of quotation marks differs systematically by language and by medium. In European typography, it is common to use *guillemets* (single or double angle quotation marks) for books and, except for some languages, curly quotation marks in office automation. Single guillemets can be found for quotes inside quotes. The following description does not attempt to be complete, but intends to document a range of known usages of quotation mark characters. Some of these usages are also illustrated in *Figure 6-1*. In this section, the words *single* and *double* are omitted from character names where there is no conflict or both are meant.

Dutch, English, Italian, Portugese, Spanish, and Turkish use a *left quotation mark* and a *right quotation mark* for opening and closing quotations, respectively. It is typical to alternate single and double quotes for quotes within quotes. Whether single or double quotes are used for the outer quotes depends on local and stylistic conventions.

Czech, German, and Slovak use the low-9 style of quotation mark for opening instead of the standard open quotes. They employ the *left quotation mark* style of quotation mark for closing instead of the more common *right quotation mark* forms. When guillemets are used in German books, they point to the quoted text. This style is the inverse of French usage.

Danish, Finnish, Norwegian, and Swedish use the same *right quotation mark* character for both the opening and closing quotation character. This usage is employed both for office automation purposes and for books. Books sometimes use the guillemet, U+00BB right-pointing double angle quotation mark, for both opening and closing.

Hungarian and Polish usage of quotation marks is similar to the Scandinavian usage, except that they use low double quotes for opening quotations. Presumably, these languages avoid the low single quote so as to prevent confusion with the comma.

French, Greek, Russian and Slovenian, among others, use the guillemets, but Slovenian usage is the same as German usage in their direction. Of these languages, at least French inserts space between text and quotation marks. In the French case, U+00A0 no-break space can be used for the space that is enclosed between quotation mark and text; this choice helps line-breaking algorithms.

Figure 6-1. European Quotation Marks

Single right quote = apostrophe

‘quote’ don’t

Usage depends on language

“English” «French»
„German“ »Slovenian«
”Swedish” »Swedish books»

East Asian Usage. The glyph for each quotation mark character for an Asian character set occupies predominantly a single quadrant of the character cell. The quadrant used depends on whether the character is opening or closing and whether the glyph is for use with horizontal or vertical text.

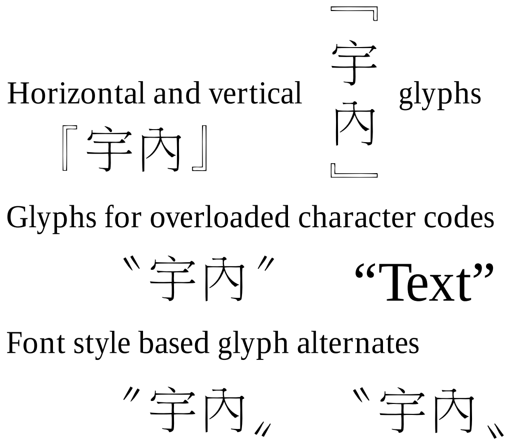
The pairs of quotation characters are listed in *Table 6-3*.

Table 6-3. East Asian Quotation Marks

Style	Opening	Closing
Corner bracket	300C	300D
White corner bracket	300E	300F
Double prime	301D	301F

Glyph Variation. The glyphs for “double-prime” quotation marks consist of a pair of wedges, slanted either forward or backward, with the tips of the wedges pointing either up or down. In a pair of double-prime quotes, the closing and the opening character of the pair slant in opposite directions. Two common variations exist as shown in *Figure 6-2*. To confuse matters more, another form of double-prime quotation marks is used with Western style, horizontal text, in addition to the curly single or double quotes.

Figure 6-2. Asian Quotation Marks


Horizontal and vertical glyphs
『宇内』
Glyphs for overloaded character codes
〝宇内〞 “Text”
Font style based glyph alternates
〞宇内〟 〝宇内〟

Three pairs of quotation marks are used with Western-style, horizontal text, as shown in *Table 6-4*.

Table 6-4. Opening and Closing Forms

Style	Opening	Closing	Comment
Single	2018	2019	Rendered as “wide” character
Double	201C	201D	Rendered as “wide” character
Double prime	301D	301E	

Overloaded Character Codes. The character codes for standard quotes can refer to regular narrow quotes from a Latin font used with Latin text, as well as wide quotes from an Asian font used with other wide characters. This situation can be handled with some success where the text is marked up with language tags.

Consequences for Semantics. The semantics of U+00AB, U+00BB (double guillemets), and U+201D (right double quotation) are context-dependent. The semantics of U+201A and U+201B (low-9 quotation marks) are always opening; this usage is distinct from the usage of U+301E LOW DOUBLE PRIME QUOTATION MARK, which is unambiguously closing.

Apostrophes

U+0027 APOSTROPHE is the most commonly used character for apostrophe. However, it has ambiguous semantics and direction. When text is set, U+2019 RIGHT SINGLE QUOTATION MARK is preferred as apostrophe. Word processors commonly offer a facility for automatically converting the U+0027 APOSTROPHE to a contextually-selected curly quotation glyph.

Letter Apostrophe. U+02BC MODIFIER LETTER APOSTROPHE is preferred where the apostrophe is to represent a modifier letter (for example, in transliterations to indicate a glottal stop). In the latter case, it is also referred to as a *letter apostrophe*.

Punctuation Apostrophe. U+2019 RIGHT SINGLE QUOTATION MARK is preferred where the character is to represent a punctuation mark, as for contractions: "*We've been here before.*" In this latter case, U+2019 is also referred to as a *punctuation apostrophe*.

An implementation cannot assume that users' text always adheres to the distinction between these characters. The text may come from different sources, including mapping from other character sets that do not make this distinction between the letter apostrophe and the punctuation apostrophe/right single quotation mark. In that case, *all* of them will generally be represented by U+2019.

The semantics of U+2019 are therefore context-dependent. For example, if surrounded by letters or digits on both sides, it behaves as an in-text punctuation character and does not separate words or lines.

Other Punctuation

Hyphenation Point. U+2027 HYPHENATION POINT is a raised dot used to indicate correct word breaking, as in dic·tion·ar·ies. It is a punctuation mark, to be distinguished from U+00B7 MIDDLE DOT, which has multiple semantics.

Fraction Slash. U+2044 FRACTION SLASH is used between digits to form numeric fractions, such as 2/3, 3/9, and so on. The standard form of a fraction built using the fraction slash is defined as follows: Any sequence of one or more decimal digits, followed by the fraction slash, followed by any sequence of one or more decimal digits. Such a fraction should be displayed as a unit, such as ¾ or as $\frac{3}{4}$. The precise choice of display can depend upon additional formatting information.

If the displaying software is incapable of mapping the fraction to a unit, then it can also be displayed as a simple linear sequence as a fallback (for example, 3/4). If the fraction is to be separated from a previous number, then a space can be used, choosing the appropriate width (normal, thin, zero width, and so on). For example, 1 + ZERO WIDTH SPACE + 3 + FRACTION SLASH + 4 is displayed as 1¾.

Spacing Overscore. U+203E OVERLINE is the above-the-line counterpart to U+005F LOW LINE. It is a spacing character, not to be confused with U+0305 COMBINING OVERLINE. As with all over- or underscores, a sequence of these characters should connect in an unbroken line. The overscoring characters also must be distinguished from U+0304 COMBINING MACRON, which does not connect horizontally in this way.

Doubled Punctuation. Several doubled punctuation characters that have compatibility decompositions into a sequence of two punctuation marks are also encoded as single characters: U+203C DOUBLE EXCLAMATION MARK, U+2048 QUESTION EXCLAMATION MARK, and U+2049 EXCLAMATION QUESTION MARK. These doubled punctuation marks are included as an implementation convenience for East Asian and Mongolian text, which is rendered vertically.

Bullets. U+2022 BULLET is the typical character for a bullet. Within the general punctuation, several alternative forms for bullets are separately encoded: U+2023 TRIANGULAR BULLET, U+204C BLACK LEFTWARDS BULLET, and so on. U+00B7 MIDDLE DOT also often functions as a small bullet. Bullets mark the head of specially formatted paragraphs, often occurring in lists, and may in fact use arbitrary graphics or dingbat forms as well as more conventional bullet forms. U+261E WHITE RIGHT POINTING INDEX, for example, is often used to highlight a note in text, as a kind of gaudy bullet.

Paragraph Marks. U+00A7 SECTION SIGN and U+00B6 PILCROW SIGN are often used as visible indications of sections or paragraphs of text, in editorial markup, to show format modes, and so on. Which character indicates sections and which character indicates paragraphs may vary by convention. U+204B REVERSED PILCROW SIGN is a fairly common alternate representation of the paragraph mark.

CJK Symbols and Punctuation: U+3000–U+303F

This block encodes punctuation marks and symbols used primarily by writing systems that employ Han ideographs. Most of these characters are found in East Asian standards.

U+3000 IDEOGRAPHIC SPACE is provided for compatibility with legacy character sets. It is a fixed-width space appropriate for use with an ideographic font. U+301C WAVE DASH and U+3030 WAVY DASH are special forms of dashes found in East Asian character standards. (For a list of other space and dash characters in the Unicode Standard, see *Table 6-1* and *Table 6-2*.)

U+3037 IDEOGRAPHIC TELEGRAPHIC LINE FEED SEPARATOR SYMBOL is a visible indicator of the line feed separator symbol used in the Chinese telegraphic code; it is comparable to the pictures of control codes found in the Control Pictures block.

U+3005 IDEOGRAPHIC ITERATION MARK is used to stand for the second of a pair of identical ideographs occurring in adjacent positions within a document.

U+3006 IDEOGRAPHIC CLOSING MARK is used frequently on signs to indicate that a store or booth is closed for business. The Japanese pronunciation is *shime*, most often encountered in the compound *shime-kiri*.

U+3012 POSTAL MARK is used in Japanese addresses immediately preceding the numerical postal code. It is also used on forms and applications to indicate the blank space in which a postal code is to be entered. U+3020 POSTAL MARK FACE and U+3016 CIRCLED POSTAL MARK are properly glyphic variants of U+3012, and are included for compatibility.

U+3031 VERTICAL KANA REPEAT MARK and U+3032 VERTICAL KANA REPEAT MARK WITH VOICED SOUND MARK are used only in *vertically written* Japanese to repeat pairs of kana characters occurring immediately prior in a document. The voiced variety U+3032 is used in cases where the repeated kana are to be voiced. For instance, a repetitive phrase like *toki-doki* could be expressed <U+3068 U+304D U+3032> in vertical writing. Both of these characters are intended to be represented by "double height" glyphs requiring two ideographic "cells" to print; this intention also explains the existence in source standards of the characters representing the top and bottom halves of these (that is, the characters U+3033, U+3034, and U+3035). In horizontal writing, similar characters are used, and they are separately encoded. In Hiragana, the equivalent repeat marks are encoded at U+309D and U+309E; in Katakana, they are U+30FD and U+30FE.

Unknown or Unavailable Ideographs

U+3013 GETA MARK is used to indicate the presence of, or to hold a place for, an ideograph that is not available when a document is printed. It has no other use. Its name comes from its resemblance to the mark left by traditional Japanese sandals (*geta*); a variety of light and heavy glyphic variants occur.

U+303E IDEOGRAPHIC VARIATION INDICATOR is a graphic character that is to be rendered visibly. It alerts the user that the intended character is similar to, but not equal to, the character that follows. Its use is similar to the existing character U+3013 GETA MARK. A GETA MARK substitutes for the unknown or unavailable character, but does not identify it. The IDEOGRAPHIC VARIATION INDICATOR is the head of a two-character sequence that gives some indication about the intended glyph or intended character. Ultimately, the IDEOGRAPHIC VARIATION INDICATOR and the character following it are intended to be replaced by the correct character, once it has been identified or a font resource or input resource has been provided for it.

U+303F IDEOGRAPHIC HALF-FILL SPACE is a visible indicator of a display cell filler used when ideographic characters have been split during display on systems using a double-byte character encoding. It is included in the Unicode Standard for compatibility.

See also “Ideographic Description Sequences” in *Section 10.1, Han*.

CJK Compatibility Forms: U+FE30–U+FE4F

A number of presentation forms encoded in this block are found in the Republic of China (Taiwan) national standard CNS 11643. These vertical forms of punctuation characters are provided for compatibility with those legacy implementations that encode these characters explicitly when Chinese text is being set in vertical rather than horizontal lines. The preferred Unicode encoding is to encode the nominal characters that correspond to these vertical variants. Then, at display time, the appropriate glyph is selected according to the line orientation.

Small Form Variants: U+FE50–U+FE6F

The Republic of China (Taiwan) national standard CNS 11643 also encodes a number of small variants of ASCII punctuation.

The characters of this block, while construed as fullwidth characters, are nevertheless depicted using small forms that are set in a fullwidth display cell. (See the discussion in *Section 10.3, Katakana*.) These characters are provided for compatibility with legacy implementations.

Unifications. Two small form variants from CNS 11643/plane 1 were unified with other characters outside the ASCII block: 2131_{16} was unified with U+00B7 MIDDLE DOT, and 2261_{16} was unified with U+2215 DIVISION SLASH.

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (see http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html). More recent versions of the Unicode standard exist (see http://www.unicode.org/unicode/standard/versions/).

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.