This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, 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 (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/standard/versions/enumeratedversions.html.

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

Cover and CD-ROM label design: Steve Mehallo, http://www.mehallo.com

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: http://www.awprofessional.com

Library of Congress Cataloging-in-Publication Data
The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].
p. cm.
Includes bibliographical references and index.
ISBN 0-321-18578-1 (alk. paper)
1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004
005.7'2—dc21

2003052158

ISBN 0-321-18578-1
Text printed on recycled paper
1 2 3 4 5 6 7 8 9 10—CRW—0706050403
First printing, August 2003

Chapter 6

Writing Systems and Punctuation

This chapter begins the portion of the Unicode Standard devoted to the detailed description of various related groups of Unicode characters. This chapter itself presents a general introduction to writing systems and then discusses punctuation characters in more detail. Subsequent chapters deal with historically or geographically related groups of scripts.

In this standard, characters are organized into subparts of the codespace called *blocks*. Character blocks generally contain characters from a single script, and in many cases, a script is fully represented in its character block. Because of this, character blocks are used to help structure the discussion of scripts and other groups of characters in this and subsequent chapters. The blocks can be identified by the section headers and associated ranges of Unicode code points listed for them. The code charts in *Chapter 16, Code Charts*, are also organized by character blocks.

There are many different kinds of writing systems in the world. Their variety poses some significant issues for character encoding in the Unicode Standard as well as for implementers of the standard. Those who first approach the Unicode Standard without a background in writing systems may find the huge list of scripts bewilderingly complex. Therefore, before considering the script descriptions in detail, this chapter first presents a brief introduction to the types of writing systems. That introduction explains basic terminology about scripts and character types that will be used again and again when discussing particular scripts.

The rest of this chapter deals with a special case: punctuation marks, which tend to be scattered about in different blocks and which may be used in common by many scripts. 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. There are also occasional punctuation characters in character blocks for specific scripts.

Most punctuation characters are intended for common usage with any script, although some of them are script-specific. Their appearance and detailed functions may vary between languages and scripts, but their overall purpose is shared: They serve to separate or otherwise organize units of text, such as sentences and phrases, thereby helping to clarify the meaning of the text. Their use is not limited to linguistic text, however, as some groups of punctuation characters also occur with sets of symbols in mathematical and scientific formulae, for example.

6.1 Writing Systems

This section presents a brief introduction to writing systems. It describes the different kinds of writing systems and relates them to the encoded scripts found in the Unicode Standard. This framework may help to make the variety of scripts, modern and historic, a little less daunting. The terminology used here follows that developed by Peter T. Daniels, a leading expert on writing systems of the world.

Alphabets. Writing systems that consist of letters for the writing of both consonants and vowels are called *alphabets*. The word *alphabet* is derived from the Greek word *alphabetos*, itself derived from the names of the first two Greek letters, alpha and beta. Consonants and vowels have equal status as letters in these systems. The Latin alphabet is the most widespread and well-known example of an alphabet, having been adapted for use in writing thousands of languages.

While in principle alphabets are used to write the consonantal and vocalic sounds of a language, in practice the relationship between sound and letter may be rather rough. The history of an alphabet's use for writing a particular language may be long enough that the sound system of the language changes out from under the writing conventions, resulting in arbitrary and sometimes incomprehensible spelling rules for a language. The situations range from cases such as Italian or Finnish, where the match between letter and sound is rather close, to English, which has notoriously complex and arbitrary spelling. The most important thing to keep in mind is that one should never assume that the encoded character for a particular letter is necessarily associated with any particular sound, even for a given, single language.

Phonetic alphabets are exceptions. They are used specifically for the precise transcription of the sounds of languages. The best known of these alphabets is the *International Phonetic Alphabet*, an adaptation and extension of the Latin alphabet by the addition of new letters and marks for specific sounds and modifications of sounds. Unlike normal alphabets, the intent of phonetic alphabets is that their letters exactly represent sounds. Phonetic alphabets are not used as general-purpose writing systems per se, but it is not uncommon for a formerly unwritten language to have an alphabet developed for it based on a phonetic alphabet.

Abjads. An *abjad* is a consonant writing system. The main letters are all consonants (or long vowels), with other vowels either left out entirely or indicated with secondary marking of the consonants. The best-known example of an abjad is the Arabic writing system. Indeed, the term "abjad" is derived from the first four letters of the traditional order of the Arabic script. Abjads are often, although not exclusively, associated with Semitic languages, which have word structures particularly well suited to the use of consonantal writing.

Hebrew and Arabic are typically written without any vowel marking at all. The vowels, when they do occur in writing, are referred to as *points* or *harakat*, and are indicated by the use of diacritic dots and other marks placed above and below the consonantal letters.

Syllabaries. In a *syllabary* each symbol of the system typically represents both a consonant and a vowel, or in some instances more than one consonant and a vowel. One of the best-known examples of a syllabary is Hiragana, used for Japanese, in which the units of the system represent the syllables *ka*, *ki*, *ku*, *ke*, *ko*, *sa*, *si*, *su*, *se*, *so*, and so on. In general parlance, the elements of a syllabary are not called *letters*, but rather *syllables*, but this can lead to some confusion, because letters of alphabets and units of other writing systems are also used, singly or in combinations, to write syllables of languages. So in a broad sense, the term "letter" can also be used to refer to the syllables of a syllabary.

In syllabaries such as Cherokee, Hiragana, Katakana, and Yi, each symbol has a unique shape, with no particular shape relation to any of the consonant(s) or vowels of the syllables. In other cases, however, the syllabic symbols of a syllabary are not atomic; they can be built up out of parts that have a consistent relationship to the phonological parts of the syllable. Such systems are called *featural syllabaries*. The best example of a featural syllabary is the Hangul writing system for Korean. Each Hangul syllable is made up of a part for the initial consonant (or consonant cluster), a part for the vowel (or diphthong), and an optional part for the final consonant (or consonant cluster). The relationship between the sounds and the graphic parts to represent them is systematic enough for Korean that the graphic parts collectively are known as *jamos* and constitute a kind of alphabet on their own. In other featural syllabaries, such as the Canadian Aboriginal Syllabics, the relationship of sound and graphic parts is less systematic.

Abugidas. The many scripts of South and Southeast Asia that are historically derived from the ancient Brahmi script share a special type of writing system known as an *abugida*. The term "abugida" is actually derived from the North Semitic alphabetic order: *alef*, *bet*, *gimel*, *dalet*, and, more particularly, from the first four vowels and first four consonants of the traditional order of the Ethiopic script.

In an abugida, each consonant letter carries an inherent vowel, usually /a/. There are also vowel letters, often distinguished between a set of independent vowel letters, which occur on their own, and dependent vowel letters, or *matras*, which are subordinate to consonant letters. When a dependent vowel letter follows a consonant letter, the vowel overrides the inherent vowel of the consonant. This is shown schematically in *Figure 6-1*.

Figure 6-1. Overriding Inherent Vowels

ka + i → ki ka + e → ke

ka + u → ku ka + o → ko

Abugidas also typically contain a special element usually referred to as a *halant*, *virama*, or *killer*, which, when applied to a consonant letter with its inherent vowel, has the effect of *removing* the inherent vowel, resulting in a bare consonant sound.

The best-known example of an abugida is the Devanagari script, used in modern times to write Hindi and many other Indian languages, and used classically to write Sanskrit. See *Section 9.1, Devanagari*, for a detailed description of how Devanagari works and is rendered.

Abugidas represent a kind of blend of syllabic and alphabetic characteristics in a writing system. Historically, abugidas spread across South Asia and were adapted by many languages, often of phonologically very different types. This has also resulted in many extensions, innovations, and/or simplifications of the original patterns.

Because of legacy practice, three distinct approaches have been taken in the Unicode Standard for the encoding of abugidas: the Devanagari model, the Tibetan model, and the Thai model. The Devanagari model, used for most abugidas, encodes an explicit virama character and represents text in its logical order. The Thai model departs from the Devanagari model in that it represents text in its visual display order, based on the typewriter legacy, rather than in logical order. The Tibetan model avoids an explicit virama, instead encoding a sequence of *subjoined consonants* to represent consonants occurring in clusters in a syllable.

The Ethiopic script may also be analyzed as an abugida, because the base character for each consonantal series is understood as having an inherent vowel. However, Ethiopic lacks some of the typical features of Brahmi-derived scripts, such as halants and matras. Histor-

ically, it was derived from early Semitic scripts and in its earliest form was an abjad. In its traditional presentation and its encoding in the Unicode Standard, it is now treated more like a syllabary.

Logosyllabaries. The final major category of writing system is known as the *logosyllabary*. In a logosyllabary, the units of the writing system are used primarily to write words and/or morphemes of words, with some subsidiary usage to represent syllabic sounds per se.

The best example of a logosyllabary is the Han script, used for writing Chinese and borrowed by a number of other East Asian languages for use as part of their writing systems. The term for a unit of the Han script is *hànzì* 漢字 in Chinese, *kanji* 漢字 in Japanese, and *hanja* 漢字 in Korean. In many instances this unit also constitutes a word, but more typically, two or more units together are used to write a word.

This unit has variously been referred to as an *ideograph* ("idea writing") or a *logograph* ("word writing"), as well as other terms. No single English term is completely satisfactory or uncontroversial. In this standard, *CJK ideograph* is used because it is a widely understood term.

There are a number of other historical examples of logosyllabaries, many of which may eventually be encoded in the Unicode Standard. They vary in the degree to which they combine logographic writing principles, where the symbols stand for morphemes or entire words, and syllabic writing principles, where the symbols come to represent syllables per se, divorced from their meaning as morphemes or words. In some notable instances, as for Sumero-Akkadian cuneiform, a logosyllabary may evolve through time into a syllabary or alphabet by shedding its use of logographs. In other instances, as for the Han script, the use of logographic characters is very well entrenched and persistent. However, even for the Han script a small number of characters are used purely to represent syllabic sounds, so as to be able to represent such things as foreign personal names and place names.

The classification of a writing system is often somewhat blurred by complications in the exact ways in which it matches up written elements to the phonemes or syllables of a language. For example, although Hiragana is classified as a syllabary, it does not always have an exact match between syllables and written elements. Syllables with long vowels are not written with a single element, but rather with a sequence of elements. Thus the syllable with a long vowel *kū* is written with two separate Hiragana symbols, {ku}+{u}. Because of these kinds of complications, one must always be careful not to assume too much about the structure of a writing system from its nominal classification.

Table 6-1 lists all of the scripts currently encoded in the Unicode Standard, showing the writing system type for each. The list is an approximate guide, rather than a definitive classification, because of the mix of features seen in many scripts. The writing systems for some languages may be quite complex, mixing more than one writing system together in a composite system. Japanese is the best example; it mixes a logosyllabary (Han), two syllabaries (Hiragana and Katakana), and one alphabet (Latin, for *romaji*).

Notational Systems. In addition to scripts for written natural languages, there are notational systems for other kinds of information. Some of these more closely resemble text than others. The Unicode Standard encodes symbols for use with mathematical notation, Western and Byzantine musical notation, Braille, as well as symbols for use in divination, such as the Yijing hexagrams. Notational systems can be classified by how closely they resemble text. Even notational systems that do not fully resemble text may have symbols used in text. In the case of musical notation, for example, while the full notation is two-dimensional, many of the encoded symbols are frequently referenced in texts about music and musical notation.

Table 6-1. Typology of Scripts in the Unicode Standard

Alphabets	Latin, Greek, Cyrillic, Armenian, Thaana, Georgian, Ogham, Runic, Mongolian, Old Italic, Gothic, Ugaritic, Deseret, Shavian, Osmanya
Abjads	Hebrew, Arabic, Syriac
Abugidas	Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao, Tibetan, Myanmar, Tagalog, Hanunóo, Buhid, Tagbanwa, Khmer, Limbu, Tai Le
Logosyllabaries	Han
Simple Syllabaries	Cherokee, Hiragana, Katakana, Bopomofo, Yi, Linear B, Cypriot
Featural Syllabaries	Ethiopic, Canadian Aboriginal Syllabics, Hangul

6.2 General Punctuation

Punctuation characters—for example, U+002C COMMA and 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 and 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.

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.

Numeric Separators. Any of the characters U+002C, U+002E, U+060C, U+066B, or U+066C (and possibly others) can be used as numeric separator characters, depending on the locale and user customizations.

Rendering. 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 Unicode Standard Annex #9, "The Bidirectional Algorithm"), 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, Bidi 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 rather affect the operation of layout algorithms. For a description of these characters, see *Section 15.2, Layout Controls*.

Punctuation: U+0020–U+00BF

Standards. The Unicode Standard adapts the ASCII (ISO 646) 7-bit standard by retaining the semantics and numeric code points. The content and arrangement of the ASCII standard are far from optimal in the context of a large codespace, 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 Usage of Quotation Marks." The rather haphazard ASCII collection of punctuation and mathematical signs is 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 points 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 code point 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 character 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_{16} is defined in ANSI X3.4 as *apostrophe* (*closing single quotation mark; acute accent*), and $2D_{16}$ is defined as *hyphen-minus*. In general, the Unicode Standard provides the same interpretation for the equivalent code points, 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 Usage of Quotation Marks" later in this section.

Tilde. Although there are several common shapes used to render U+007E "~" TILDE, modern fonts generally render it with a center line glyph, as shown here and in the code charts. However, it may also appear as a raised, spacing tilde, serving as a spacing clone of U+0303 "◌̃" COMBINING TILDE (see "Spacing Clones of Diacritics" in *Section 7.1, Latin*). This is a form common in older implementations, particularly for terminal emulation and typewriter style fonts.

Some of the common uses of a tilde include indication of alternation, an approximate value or, in some notational systems, indication of a logical negation. In the latter context, it is really being used as a shape-based substitute character for the more precise U+00AC "¬" NOT SIGN. A tilde is also used in dictionaries to repeat the defined term in examples. In that usage, as well as when used as punctuation to indicate alternation, it is more appropriately represented by a wider form, encoded as U+2053 "⁓" SWUNG DASH. U+02DC "˜" SMALL TILDE is a modifier letter encoded explicitly as the spacing form of the combining tilde as a diacritic. For mathematical usage, U+223C "∼" TILDE OPERATOR should be used to unambiguously encode the operator.

General Punctuation: U+2000–U+206F

The General Punctuation block contains punctuation characters and characterlike elements used to achieve certain text layout effects.

Format Control Characters

Format control characters are special characters that have no visible glyph of their own, but that affect the display of characters to which they are adjacent, or that have other specialized functions such as serving as invisible anchor points in text. A significant number of format control characters are encoded in the General Punctuation block, but their descriptions are found in other sections.

Cursive joining controls, as well as U+200B ZERO WIDTH SPACE, U+2028 LINE SEPARATOR, U+2029 PARAGRAPH SEPARATOR, and U+2060 WORD JOINER, are described in *Section 15.2, Layout Controls*. Bidirectional ordering controls are also discussed in *Section 15.2, Layout Controls*, but their detailed use is specified in Unicode Standard Annex #9, "The Bidirectional Algorithm."

Invisible operators are explained in *Section 15.3, Invisible Operators*. Deprecated format characters related to obsolete models of Arabic text processing are described in *Section 15.4, Deprecated Format Characters*.

The reserved code points U+2064..U+2069 and U+FFF0..U+FFF8, as well as any reserved code points in the range U+E0000..U+E0FFF, are reserved for the possible future encoding of other format control characters. Because of this, they are treated as default ignorable code points. For more information, see *Section 5.20, Default Ignorable Code Points*.

Space Characters

The most commonly used space character is U+0020 SPACE. Also often used is its nonbreaking counterpart, U+00A0 NO-BREAK SPACE. These two characters have the same width, but behave differently for line breaking. For more information, see Unicode Stan-

dard Annex #14, "Line Breaking." U+00A0 NO-BREAK SPACE behaves like a numeric separator for the purposes of bidirectional layout. (See Unicode Standard Annex #9, "The Bidirectional Algorithm," 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 15.2, Layout Controls*, and Unicode Standard Annex #14, "Line Breaking."

Space characters may also be found in other character blocks in the Unicode Standard. The list of space characters appears in *Table 6-2*.

Table 6-2. 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+205F	MEDIUM MATHEMATICAL SPACE
U+3000	IDEOGRAPHIC SPACE

U+200B ZERO WIDTH SPACE and several spacelike, zero-width characters with special properties are described in *Section 15.2, 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+2212 MINUS SIGN, 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. (Some typographers prefer to use U+2013 EN DASH set off with spaces – like this – to make the same kind of break.) This kind of dash 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 6-3*. For a description of the line breaking behavior of dashes and hyphens, see Unicode Standard Annex #14, "Line Breaking Properties."

Table 6-3. Unicode Dash Characters

Code	Name
U+002D	HYPHEN-MINUS
U+007E	TILDE (when used as *swung dash*)
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+2053	SWUNG DASH
U+207B	SUPERSCRIPT MINUS
U+208B	SUBSCRIPT MINUS
U+2212	MINUS SIGN
U+301C	WAVE DASH
U+3030	WAVY DASH

For information about the function of U+00AD SOFT HYPHEN, see *Section 15.2, Layout Controls*.

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 6-2*. 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 6-2. 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.

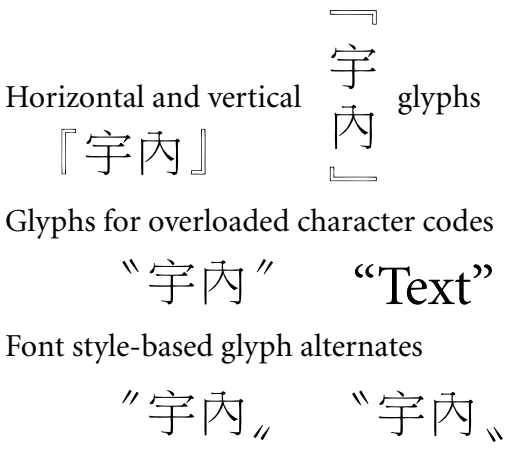
The pairs of quotation characters are listed in *Table 6-4*.

Table 6-4. 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 6-3*. 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 6-3. Asian Quotation Marks



Three pairs of quotation marks are used with Western-style, horizontal text, as shown in *Table 6-5*.

Table 6-5. 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 MARK 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+301F 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 (General Category = Nd), 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 + THIN 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 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.

Commercial Minus. U+2052 ⁒ COMMERCIAL MINUS SIGN is used in commercial or tax-related forms or publications in several European countries, including Germany and Scandinavia. The string "./." is used as a fallback representation for this character.

The symbol may also appear as a marginal note in letters, denoting enclosures. One variation replaces the top dot with a digit indicating the number of enclosures.

An additional usage of the sign appears in the Uralic Phonetic Alphabet (UPA), where it marks a structurally related borrowed element of different pronunciation. In Finland and a number of other European countries, the dingbats ⁒ and ✓ are always used for "correct" and "incorrect," respectively, in marking a student's paper. This contrasts with American practice for example, where ✓ and ✗ might be used for "correct" and "incorrect," respectively, in the same context.

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

This block encodes punctuation marks and symbols used 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 6-2* and *Table 6-3*.)

U+3037 IDEOGRAPHIC TELEGRAPH 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+3008, U+3009 angle brackets are unambiguously wide. The Unicode Standard encodes different characters for use in other contexts, such as mathematics. There are other characters in this block that have the same characteristics, including double angle brackets, tortoise shell brackets, and white square brackets.

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+3036 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 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 as <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 11.1, 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 11.3, Hiragana and 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.