Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

Chapter 6

Writing Systems and Punctuation

This chapter begins the portion of the Unicode Standard devoted to the detailed description of each script or other related group of Unicode characters. Each of the subsequent chapters presents a historically or geographically related group of scripts. This chapter presents a general introduction to writing systems, explains how they can be used to classify scripts, and then presents a detailed discussion of punctuation characters that are shared across scripts.

Scripts and Blocks. The codespace of the Unicode Standard is divided into subparts called *blocks*. Character blocks generally contain characters from a single script, and in many cases, a script is fully represented in its character block; however, some scripts are encoded using several blocks, which are not always adjacent. Discussion of scripts and other groups of characters are structured by character blocks. Corresponding subsection headers identify each block and its associated range of Unicode code points. The code charts in *Chapter 17, Code Charts*, are also organized by character blocks.

Scripts and Writing Systems. 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.

Punctuation. 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. Some scripts use both common and script-specific punctuation characters, usually as the result of recent adoption of standard Western punctua-

tion marks. While punctuation characters vary in details of appearance and function between different languages and scripts, 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. Certain punctuation characters also occur in mathematical and scientific formulae.

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.

The term *writing system* has two mutually exclusive meanings in this standard. As used in this section, "writing system" refers to a way that families of scripts may be classified by how they represent the sounds or words of human language. For example, the writing system of the Latin script is alphabetic. In other places in the standard, "writing system" refers to the way a particular *language* is written. For example, the modern Japanese writing system uses four scripts: Han ideographs, Hiragana, Katakana and Latin (Romaji).

Alphabets. A writing system that consists of letters for the writing of both consonants and vowels is called an *alphabet*. The term "alphabet" is derived from the first two letters of the Greek script: *alpha, beta*. Consonants and vowels have equal status as letters in such a system. The Latin alphabet is the most widespread and well-known example of an alphabet, having been adapted for use in writing thousands of languages.

The correspondence between letters and sounds may be either more or less exact. Many alphabets do not exhibit a one-to-one correspondence between distinct sounds and letters or groups of letters used to represent them; often this is an indication of original spellings that were not changed as the language changed. Not only are many sounds represented by letter combinations, such as "th" in English, but the language may have evolved since the writing conventions were settled. Examples 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.

Phonetic alphabets, in contrast, 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. A writing system in which only consonants are indicated is an *abjad*. The main letters are all consonants (or long vowels), with other vowels either left out entirely or option-

ally indicated with the use of secondary marks on the consonants. The Phoenician script is a prototypical abjad; a better-known example is the Arabic writing system. The term "abjad" is derived from the first four letters of the traditional order of the Arabic script: *alef, beh, jeem, dal*. Abjads are often, although not exclusively, associated with Semitic languages, which have word structures particularly well suited to the use of consonantal writing. Some abjads allow consonant letters to mark long vowels, as the use of *waw* and *yeh* in Arabic for /u:/ or /i:/.

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*. This can lead to some confusion, however, 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 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. The best example of this 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.

The jamos of the Hangul writing system have another characteristic: their shapes are not completely arbitrary, but were devised with intentionally iconic shapes relating them to articulatory features of the sounds they represent in Korean. The Hangul writing system has thus also been classified as a *featural syllabary*.

Abugidas. *Abugidas* represent a kind of blend of syllabic and alphabetic characteristics in a writing system. The Ethiopic script is an abugida. The term "abugida" is derived from the first four letters of the letters of the Ethiopic script in the Semitic order: *alf, bet, gaml, dant*. The order of vowels (-ä -u -i -a) is that of the traditional vowel order in the first four columns of the Ethiopic syllable chart. Historically, abugidas spread across South Asia and were adapted by many languages, often of phonologically very different types.

This process has also resulted in many extensions, innovations, and/or simplifications of the original patterns. 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.

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.

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 is traditionally 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. Historically, 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"), a *logograph* ("word writing"), or a *sinogram*, 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, such as Tangut, 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.

Typology of Scripts in the Unicode Standard. *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 type of script 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*).

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

Alphabets	Latin, Greek, Cyrillic, Armenian, Thaana, Georgian, Ogham, Runic, Mongolian, Glagolitic, Coptic, Tifinagh, Old Italic, Gothic, Ugaritic, Old Persian, Deseret, Shavian, Osmanya, N'Ko
Abjads	Hebrew, Arabic, Syriac, Phoenician
Abugidas	Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala, Thai, Lao, Tibetan, Myanmar, Tagalog, Hanunóo, Buhid, Tagbanwa, Khmer, Limbu, Tai Le, New Tai Lue, Buginese, Syloti Nagri, Kharoshthi, Balinese, Phags-pa
Logosyllabaries	Han, Sumero-Akkadian
Simple Syllabaries	Cherokee, Hiragana, Katakana, Bopomofo, Yi, Linear B, Cypriot, Ethiopic, Canadian Aboriginal Syllabics
Featural Syllabaries	Hangul

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, and 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.

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 principally used with a specific script is found in the block corresponding to that script, such as U+058A ARMENIAN HYPHEN, U+061B "؛" ARABIC SEMICOLON, or the punctuation used with CJK ideographs in the CJK Symbols and Punctuation block. Script-specific punctuation characters may be unique in function, have different directionality, or be distinct in appearance or usage from their generic counterparts.

Punctuation intended for use with several related scripts is often encoded with the principal script for the group. For example, U+1735 PHILIPPINE SINGLE PUNCTUATION is encoded in a single location in the Hanunóo block, but it is intended for use with all four of the Philippine scripts.

Use and Interpretation. 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.

Many Unicode algorithms, such as the Bidirectional Algorithm and Line Breaking Algorithm, both of which treat numeric punctuation differently from text punctuation, resolve the status of any ambiguous punctuation mark depending on whether it is part of a number context.

Legacy character encoding standards commonly 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 marks, em dash, en dash, minus, hyphen, em space, en space, hair space, zero width space, and so on.

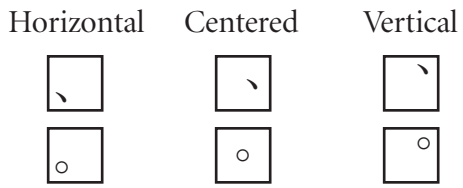
Rendering. Punctuation characters vary in appearance with the font style, just like the surrounding text characters. In some cases, where used in the context of a particular script, a specific glyph style is preferred. For example, U+002E FULL STOP should appear square when used with Armenian, but is 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 render the character faithfully.

Writing Direction. Punctuation characters shared across scripts have no inherent directionality. In a bidirectional context, their display direction is resolved according to the rules in Unicode Standard Annex #9, "The Bidirectional Algorithm." Certain script-specific punctuation marks have an inherent directionality that matches the writing direction of the script. For an example, see "Dandas" later in this section. The image of certain paired punctuation marks, specifically those that are brackets, is mirrored when the character is part of a right-to-left directional run (see *Section 4.7, Bidi Mirrored—Normative*). Mirroring ensures that the opening and closing semantics of the character remains independent of the writing direction. The same is generally not true for other punctuation marks even when their image is not bilaterally symmetric, such as *slash* or the *curly quotes*. See also "Paired Punctuation" later in this section.

In vertical writing, many punctuation characters have special vertical glyphs. Normally, fonts contain both the horizontal and vertical glyphs, and the selection of the appropriate glyph is based on the text orientation in effect at rendering time. However, see "CJK Compatibility Forms: Vertical Forms" later in this section.

Figure 6-2 shows a set of three common shapes used for *ideographic comma* and *ideographic full stop*. The first shape in each row is that used for horizontal text, the last shape is that for vertical text. The centered form may be used with both horizontal and vertical text. See also *Figure 6-4* for an example of vertical and horizontal forms for quotation marks.

Figure 6-2. Forms of CJK Punctuation



Layout Controls. A number of characters in the blocks described in this section are not graphic punctuation characters, but rather affect the operation of layout algorithms. For a description of those characters, see *Section 16.2, Layout Controls*.

Encoding Characters with Multiple Semantic Values. Some of the punctuation characters in the ASCII range (U+0020..U+007F) 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 more information, see "Apostrophes," "Space Characters," and "Dashes and Hyphens" later in this section.

Blocks Devoted to Punctuation

For compatibility with widely used legacy character sets, the Basic Latin (ASCII) block (U+0000..U+007F) and the Latin-1 Supplement block (U+0080..U+00FF) contain several of the most common punctuation signs. They are isolated from the larger body of Unicode punctuation, signs, and symbols only because their relative code locations within ASCII and Latin-1 are so widely used in standards and software. The Unicode Standard has a number of blocks devoted specifically to encoding collections of punctuation characters.

The General Punctuation block (U+2000..U+206F) contains the most common punctuation characters widely used in Latin typography, as well as a few specialized punctuation marks and a large number of format control characters. All of these punctuation characters are intended for generic use, and in principle they could be used with any script.

The Supplemental Punctuation block (U+2E00..U+2E7F) is devoted to less commonly encountered punctuation marks, including those used in specialized notational systems or occurring primarily in ancient manuscript traditions.

The CJK Symbols and Punctuation block (U+3000..U+303F) has the most commonly occurring punctuation specific to East Asian typography—that is, typography involving the rendering of text with CJK ideographs.

The Vertical Forms block (U+FE10..U+FE1F), the CJK Compatibility Forms block (U+FE30..U+FE4F), the Small Form Variants block (U+FE50..U+FE6F), and the Halfwidth and Fullwidth Forms block (U+FF00..U+FFEF) contain many compatibility characters for punctuation marks, encoded for compatibility with a number of East Asian character encoding standards. Their primary use is for round-trip mapping with those legacy standards. For vertical text, the regular punctuation characters are used instead, with alternate glyphs for vertical layout supplied by the font.

The punctuation characters in these various blocks are discussed below in terms of their general types.

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. All format control characters have General_Category=Cf. 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 16.2, Layout Controls*. Bidirectional ordering controls are also discussed in *Section 16.2, Layout Controls*, but their detailed use is specified in Unicode Standard Annex #9, "The Bidirectional Algorithm."

Invisible operators are explained in *Section 15.5, Invisible Mathematical Operators*. Deprecated format characters related to obsolete models of Arabic text processing are described in *Section 16.3, 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 Standard 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 (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.

In addition to the various fixed-width space characters, there are a few script-specific space characters in the Unicode Standard. U+1680 OGHAM SPACE MARK is unusual in that it is generally rendered with a visible horizontal line, rather than being blank.

Space characters with special behavior in word or line breaking are described in "Line and Word Breaking" in *Section 16.2, Layout Controls*, and Unicode Standard Annex #14, "Line Breaking."

U+00A0 NO-BREAK SPACE has an additional, important function in the Unicode Standard. It may serve as the base character for displaying a nonspacing combining mark in apparent isolation. Versions of the standard prior to Version 4.1 indicated that U+0020 SPACE could also be used for this function, but SPACE is no longer recommended, because of potential interactions with the handling of SPACE in XML and other markup languages. See *Section 2.11, Combining Characters*, for further discussion.

Space characters are found in several 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+1680	OGHAM SPACE MARK
U+180E	MONGOLIAN VOWEL SEPARATOR
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+202F	NARROW NO-BREAK SPACE
U+205F	MEDIUM MATHEMATICAL SPACE
U+3000	IDEOGRAPHIC SPACE

The space characters in the Unicode Standard can be identified by their General Category, [gc=Zs], in the Unicode Character Database. One exceptional "space" character is U+200B ZERO WIDTH SPACE. This character, although called a "space" in its name, does not actually have any width or visible glyph in display. It functions primarily to indicate word boundaries in writing systems that do not actually use orthographic spaces to separate words in text. It is given the General Category [gc=Cf] and is treated as a format control character, rather than as a space character, in implementations. Further discussion of U+200B ZERO WIDTH SPACE, as well as other zero-width characters with special properties, can be found in *Section 16.2, 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 has the same semantic value as U+2010 HYPHEN, but should not be broken across lines.

U+2012 FIGURE DASH 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, although in some languages *hyphen* is used for that purpose. The *en dash* should be distinguished from the U+2212 MINUS SIGN, which is

an arithmetic operator. Although it is not preferred in mathematical typesetting, typographers sometimes use U+2013 EN DASH to represent the *minus sign*, particularly a *unary minus*. When 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", unless a higher-level protocol precisely defines which of these characters serves that function.

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.) Like many other conventions for punctuation characters, such usage may depend on language. This kind of dash is commonly represented with a typewriter as a double hyphen. In older mathematical typography, U+2014 EM DASH may 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+05BE	HEBREW PUNCTUATION MAQAF
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+2E17	DOUBLE OBLIQUE HYPHEN
U+301C	WAVE DASH
U+3030	WAVY DASH
U+30A0	KATAKANA-HIRAGANA DOUBLE HYPHEN
U+FE31	PRESENTATION FORM FOR VERTICAL EM DASH
U+FE32	PRESENTATION FORM FOR VERTICAL EN DASH
U+FE58	SMALL EM DASH
U+FE63	SMALL HYPHEN-MINUS
U+FF0D	FULLWIDTH HYPHEN-MINUS

Soft Hyphen. Despite its name, U+00AD SOFT HYPHEN is not a hyphen, but rather an invisible format character used to indicate optional intraword breaks. As described in

Section 16.2, Layout Controls, its effect on the appearance of the text depends on the language and script used.

Tilde. Although several shapes are commonly 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.

Paired Punctuation

Mirroring 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, the orientation of these characters toward the enclosed text is maintained by the software, independent of the writing direction. In a bidirectional context, the glyphs are adjusted as described in Unicode Standard Annex #9, "The Bidirectional Algorithm." (See also *Section 4.7, Bidi Mirrored—Normative*.) During display, the software must ensure that the rendered glyph is the correct one in the context of bidirectional or vertical texts.

Paired punctuation marks containing the qualifier "LEFT" in their name are taken to denote *opening*; characters whose name contains 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 a right-to-left directional run, U+0028 is rendered as ")". In a left-to-right run, the same character is rendered as "(". In some mathematical usage, brackets may not be paired, or may be deliberately used in the reversed sense, such as]a,b[. Mirroring assures that in a right-to-left environment, such specialized mathematical text continues to read]b,a[and not [b, a]. See also "Language-Based Usage of Quotation Marks" later in this section.

Quotation Marks and Brackets. Like brackets, quotation marks occur in pairs, with some overlap in usage and semantics between these two types of punctuation marks. For example, some of the CJK quotation marks resemble brackets in appearance, and they are often used when brackets would be used in non-CJK text. Similarly, both single and double *guillemets* may be treated more like brackets than quotation marks.

Some of the editing marks used in annotated editions of scholarly texts exhibit features of both quotation marks and brackets. The particular convention employed by the editors determines whether editing marks are used in pairs, which editing marks form a pair, and which is the opening character. Unlike brackets, quotation marks are not mirrored in a bidirectional context.

Horizontal brackets—for example, those used in annotating mathematical expressions—are not paired punctuation, even though the set includes both top and bottom brackets. See "Horizontal Brackets" in *Section 15.6, Technical Symbols*, for more information.

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. Most keyboard layouts support only U+0022 QUOTATION MARK, therefore word processors commonly offer a facility for automatically converting the U+0022 QUOTATION MARK to a contextually selected curly quote glyph.

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 may be used 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 6-3*. 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 6-3. 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 6-4*. 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.

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

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 to wide quotes from an Asian font used with other wide characters. This situation can be handled with some suc-

Figure 6-4. Asian Quotation Marks

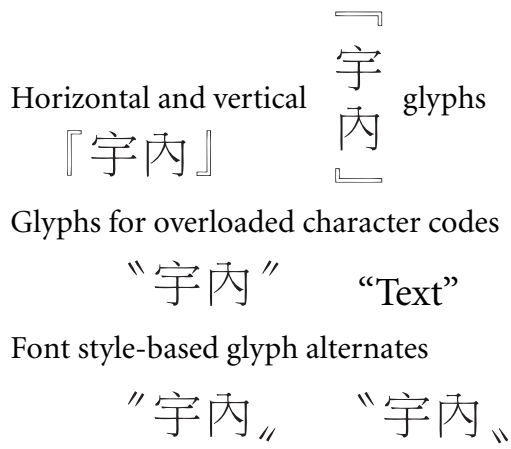


Table 6-5. Opening and Closing Forms

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

cess where the text is marked up with language tags. For more information on narrow and wide characters, see Unicode Standard Annex #11, “East Asian Width.”

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. All other quotation marks may represent opening or closing quotation marks depending on the usage.

Apostrophes

U+0027 APOSTROPHE is the most commonly used character for apostrophe. 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.

When text is set, U+2019 RIGHT SINGLE QUOTATION MARK is preferred as apostrophe, but only U+0027 is present on keyboards. Word processors commonly offer a facility for auto-

matically converting the U+0027 APOSTROPHE to a contextually selected curly quotation glyph. 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.

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 and 3/9. 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 $\frac{3}{4}$. The precise choice of display can depend on 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 Overscores and Underscores. 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 overscores and 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 char-

acters: 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, when rendered vertically.

Period or Full Stop. The *period*, or U+002E FULL STOP, can be circular or square in appearance, depending on the font or script. The hollow circle period used in East Asian texts is separately encoded as U+3002 IDEOGRAPHIC FULL STOP. Likewise, Armenian, Arabic, Ethiopic, and several other script-specific periods are coded separately because of their significantly different appearance.

In contrast, the various functions of the period, such as its use as sentence-ending punctuation, an abbreviation mark, or a decimal point, are not separately encoded. The specific semantic therefore depends on context.

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.

Ellipsis. The omission of text is often indicated by a sequence of three dots "...", a punctuation convention called *ellipsis*. Typographic traditions vary in how they lay out these dots. In some cases the dots are closely spaced; in other cases the dots are spaced farther apart. U+2026 HORIZONTAL ELLIPSIS is the ordinary Unicode character intended for the representation of an ellipsis in text and typically shows the dots separated with a moderate degree of spacing. A sequence of three U+002E FULL STOP characters can also be used to indicate an ellipsis, in which case the space between the dots will depend on the font used for rendering. For example, in a monowidth font, a sequence of three *full stops* will be wider than the *horizontal ellipsis*, but in a typical proportional font, a *full stop* is very narrow and a sequence of three of them will be more tightly spaced than the the dots in *horizontal ellipsis*.

Conventions that use four dots for an ellipsis in certain grammatical contexts should represent them either as a sequence of <full stop, horizontal ellipsis> or <horizontal ellipsis, full stop> or simply as a sequence of four *full stop* characters, depending on the requirements of those conventions.

In East Asian typographic traditions, particularly in Japan, an ellipsis is raised to the center line of text. This effect requires the use of a Japanese-specific font, or at least a specific glyph for the *horizontal ellipsis* character.

Vertical Ellipsis. When text is laid out vertically, the ellipsis is normally oriented so that the dots run from top to bottom. Most commonly, an East Asian font will contain a vertically oriented glyph variant of U+2026 for use in vertical text layout. U+FE19 PRESENTATION FORM FOR VERTICAL HORIZONTAL ELLIPSIS is a compatibility character for use in mapping

to the GB 18030 standard; it would not usually be used for an ellipsis except in systems that cannot handle the contextual choice of glyph variants for vertical rendering. U+22EE VERTICAL ELLIPSIS and U+22EF MIDLINE HORIZONTAL ELLIPSIS are part of a set of special ellipsis characters used for row or column elision in matrix notation. Their use is restricted to mathematical contexts; they should not be used as glyph variants of the ordinary punctuation ellipsis for East Asian typography.

U+205D TRICOLON has a superficial resemblance to a *vertical ellipsis*, but is part of a set of dot delimiter punctuation marks for various manuscript traditions. As for the *colon*, the dots in the *tricolon* are always oriented vertically.

Leader Dots. Leader dots are typically seen in contexts such as a table of contents or in indices, where they represent a kind of style line, guiding the eye from an entry in the table to its associated page number. Usually leader dots are generated automatically by page formatting software and do not require the use of encoded characters. However, there are occasional plain text contexts in which a string of leader dots is represented as a sequence of characters. U+2024 ONE DOT LEADER and U+2025 TWO DOT LEADER are intended for such usage. U+2026 HORIZONTAL ELLIPSIS can also serve as a three-dot version of leader dots. These leader dot characters can be used to control, to a certain extent, the spacing of leader dots based on font design, in contexts where a simple sequence of *full stops* will not suffice.

U+2024 ONE DOT LEADER also serves as a "semicolon" punctuation in Armenian, where it is distinguished from U+002E FULL STOP. See *Section 7.6, Armenian*.

Other Basic Latin Punctuation Marks. The interword punctuation marks encoded in the Basic Latin block are used for a variety of other purposes. This can complicate the tasks of parsers trying to determine sentence boundaries. As noted later in this section, some can be used as numeric separators. Both *period* and U+003A ":" COLON can be used to mark abbreviations as in "etc." or as in the Swedish abbreviation "S:ta" for "Sankta". U+0021 "!" EXCLAMATION MARK is used as a mathematical operator (*factorial*). U+003F "?" QUESTION MARK is often used as a substitution character when mapping Unicode characters to other character sets where they do not have a representation. This practice can lead to unexpected results when the converted data are file names from a file system that supports "?" as a wildcard character. U+003B ";" SEMICOLON is the preferred representation for the Greek question mark. (U+037E ";" GREEK QUESTION MARK is canonically equivalent to U+003B, so all normalized data will use the *semicolon*.)

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.

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

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.

At Sign. U+0040 COMMERCIAL AT has acquired a prominent modern use as part of the syntax for e-mail addresses. As a result, users in practically every language community suddenly needed to use and refer to this character. Consequently, many colorful names have been invented for this character. Some of these contain references to animals or even pastries. *Table 6-6* gives a sample.

Table 6-6. Names for the @

Language	Name and Comments
Chinese	= xiao laoshu (means "little mouse" in Mandarin Chinese), laoshu hao (means "mouse mark" in Mandarin Chinese)
Danish	= grishale, snabel-a (common, humorous slang)
Dutch	= apenstaartje (common, humorous slang)
Finnish	= ät, ät-merkki (Finnish standard) = kissanhäntä, miukumauku (common, humorous slang)
French	= arobase, arrobe, escargot, a crolle (common, humorous slang)
German	= Klammeraffe
Hebrew	= shtrudl ("Strudel", modern Hebrew) = krukhit (more formal Hebrew)
Hungarian	= kukac (common, humorous slang)
Italian	= chiocciola
Polish	= atka, małpa, małpka (common, humorous slang)
Portuguese	= arroba
Russian	= sobachka (common, humorous slang)
Slovenian	= afna (common, humorous slang)
Spanish	= arroba
Swedish	= snabel-a, kanelbulle (common, humorous slang)

Archaic Punctuation and Editorial Marks

Archaic Punctuation. Many archaic scripts use punctuation marks consisting of a set of multiple dots, such as U+2056 THREE DOT PUNCTUATION. The semantics of these marks can vary by script, and some of them are also used for special conventions, such as the use of U+205E VERTICAL FOUR DOTS in modern dictionaries. U+205B FOUR DOT MARK and U+205C DOTTED CROSS were used by scribes in the margin to highlight a piece of text.

These kinds of punctuation marks commonly occur in ancient scripts. Their specific function may be different in each script. However, encoding only a single set in the Unicode Standard simplifies the task of deciding which character to use for a given mark.

There are some exceptions to this general rule. Archaic scripts with script-specific punctuation include Runic, Aegean Numbers, and Cuneiform. In particular, the appearance of punctuation written in the Cuneiform style is sufficiently different that no unification was attempted.

Editorial Marks. The Greek text of the New Testament exists in a large number of manuscripts with many textual variants. The most widely used critical edition of the New Testament, the Nestle-Aland edition published by the United Bible Societies (UBS), introduced a set of editorial characters that are regularly used in a number of journals and other publications. As a result, these editorial marks have become the recognized method of annotating the New Testament.

U+2E00 RIGHT ANGLE SUBSTITUTION MARKER is placed at the start of a single word when that word is replaced by one or more different words in some manuscripts. These alternative readings are given in the *apparatus criticus*. If there is a second alternative reading in one verse, U+2E01 RIGHT ANGLE DOTTED SUBSTITUTION MARKER is used instead.

U+2E02 LEFT SUBSTITUTION BRACKET is placed at the start of a sequence of words where an alternative reading is given in the *apparatus criticus*. This bracket is used together with the U+2E03 RIGHT SUBSTITUTION BRACKET. If there is a second alternative reading in one verse, the dotted forms at U+2E04 and U+2E05 are used instead.

U+2E06 RAISED INTERPOLATION MARKER is placed at a point in the text where another version has additional text. This additional text is given in the *apparatus criticus*. If there is a second piece of interpolated text in one verse, the dotted form U+2E07 RAISED DOTTED INTERPOLATION MARKER is used instead.

U+2E08 DOTTED TRANSPOSITION MARKER is placed at the start of a word or verse that has been transposed. The transposition is explained in the *apparatus criticus*. When the words are preserved in different order in some manuscripts, U+2E09 LEFT TRANSPOSITION BRACKET is used. The end of such a sequence of words is marked by U+2E0A RIGHT TRANSPOSITION BRACKET.

The characters U+2E0B RAISED SQUARE and U+2E0C LEFT RAISED OMISSION BRACKET are conventionally used in pairs to bracket text, with RAISED SQUARE marking the start of a passage of omitted text and LEFT RAISED OMISSION BRACKET marking its end. In other editorial traditions, U+2E0C LEFT RAISED OMISSION BRACKET may be paired with U+2E0D RIGHT

RAISED OMISSION BRACKET. Depending on the conventions used, either may act as the starting or ending bracket.

Two other bracket characters, U+2E1C LEFT LOW PARAPHRASE BRACKET and U+2E1D RIGHT LOW PARAPHRASE BRACKET, have particular usage in the N'Ko script, but also may be used for general editorial punctuation.

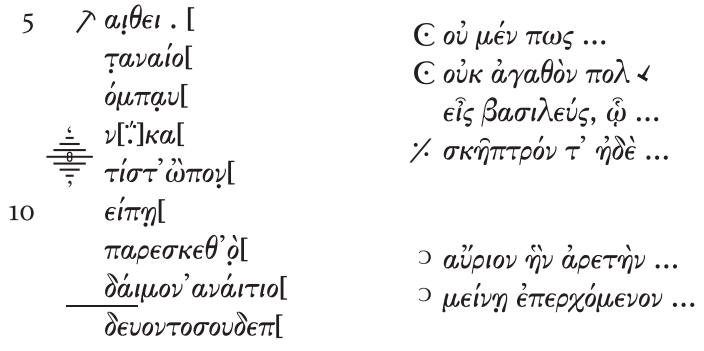
Ancient Greek Editorial Marks. Ancient Greek scribes generally wrote in continuous uppercase letters without separating letters into words. On occasion, the scribe added punctuation to indicate the end of a sentence or a change of speaker or to separate words. Editorial and punctuation characters appear abundantly in surviving papyri and have been rendered in modern typography when possible, often exhibiting considerable glyphic variation. A number of these editorial marks are encoded in the range U+2E0E..U+2E16.

The punctuation used in Greek manuscripts can be divided into two categories: marginal or semi-marginal characters that mark the end of a section of text (for example, *coronis*, *paragraphos*), and characters that are mixed in with the text to mark pauses, end of sense, or separation between words (for example, *stigme*, *hypodiastole*). The *hypodiastole* is used in contrast with *comma* and is not a glyph variant of it.

A number of editorial characters are attributed to and named after Aristarchos of Samothrace (circa 216–144 BCE), fifth head of the Library at Alexandria. Aristarchos provided a major edition of the works of Homer, which forms the basis for modern editions.

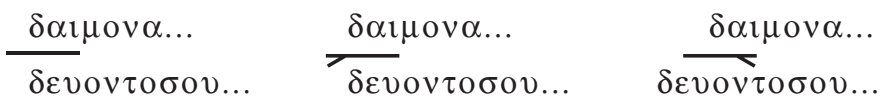
A variety of Ancient Greek editorial marks are shown in the text of *Figure 6-5*, including the *editorial coronis* and *upwards ancora* on the left. On the right are illustrated the *dotted obelos*, *capital dotted lunate sigma symbol*, *capital reversed lunate sigma symbol*, and a glyph variant of the *downards ancora*. The numbers on the left indicate text lines. A *paragraphos* appears below the start of line 12. The opening brackets "[" indicate fragments, where text is illegible or missing in the original. These examples are slightly adapted and embellished from editions of the *Oxyrhynchus Papyri* and Homer's *Iliad*.

Figure 6-5. Examples of Ancient Greek Editorial Marks



U+2E0F PARAGRAPHOS is placed at the beginning of the line but may refer to a break in the text at any point in the line. The *paragraphos* should be a horizontal line, generally stretching under the first few letters of the line it refers to, and possibly extending into the margin. It should be given a no-space line of its own and does not itself constitute a line or paragraph break point for the rest of the text. Examples of the *paragraphos*, *forked paragraphos*, and *reversed forked paragraphos* are illustrated in *Figure 6-6*.

Figure 6-6. Use of Greek Paragraphos



Double Oblique Hyphen. U+2E17 "⸗" DOUBLE OBLIQUE HYPHEN is used in ancient Near Eastern linguistics to indicate certain morphological boundaries while continuing to use the ordinary hyphen to indicate other boundaries. This symbol is also semantically distinct from U+003D "=" EQUALS SIGN. Fraktur fonts use an oblique glyph of similar appearance for the hyphen, but that is merely a font variation of U+002D HYPHEN-MINUS or U+2010 HYPHEN, not the distinctly encoded DOUBLE OBLIQUE HYPHEN.

Indic Punctuation

Dandas. Dandas are phrase-ending punctuation common to the scripts of South and South East Asia. The Devanagari *danda* and *double danda* characters are intended for generic use across the scripts of India. They are also occasionally used in Latin transliteration of traditional texts from Indic scripts.

There are minor visual differences in the appearance of the dandas, which may require script-specific fonts or a font that can provide glyph alternates based on script environment. See *Chapter 9, South Asian Scripts-I*, for a list of scripts in question. For the four Philippine scripts, the analogues to the dandas are encoded once in Hanunóo and shared across all four scripts. The other Brahmi-derived scripts have separately encoded equivalents for the danda and double danda. See *Chapter 10, South Asian Scripts-II*, and *Chapter 11, Southeast Asian Scripts*.

The Bidirectional Class of the dandas matches that for the scripts they are intended for. Kharoshthi, which is written from right to left, has Bidirectional Class R for U+10A56 KHAROSHTHI PUNCTUATION DANDA. For more on bidirectional classes, see Unicode Standard Annex #9, "The Bidirectional Algorithm."

Note that the name of the danda in Hindi is *viram*, while the different Unicode character named *virama* is called *halant* in Hindi. If this distinction is not kept in mind, it can lead to confusion as to which character is meant.

CJK Punctuation

CJK Punctuation comprises punctuation marks and symbols used by writing systems that employ Han ideographs. Most of these characters are found in East Asian standards. Typical for many of these wide punctuation characters is that the actual image occupies only the left or the right half of the normal square character cell. The extra whitespace is frequently removed in a kerning step during layout, as shown in *Figure 6-7*. Unlike ordinary kerning, which uses tables supplied by the font, the character space adjustment of wide punctuation characters is based on their character code.

Figure 6-7. CJK Parentheses

U+3000 IDEOGRAPHIC SPACE is provided for compatibility with legacy character sets. It is a fixed-width wide space appropriate for use with an ideographic font. For more information about wide characters, see Unicode Standard Annex #11, "East Asian Width."

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*.

The U+3008 and U+3009 angle brackets are unambiguously wide, as are other bracket characters in this block, such as double angle brackets, tortoise shell brackets, and white square brackets. Where mathematical and other non-CJK contexts use brackets of similar shape, the Unicode Standard encodes them separately.

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 characters (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.

Sesame Dots. U+FE45 SESAME DOT and U+FE46 WHITE SESAME DOT are used in vertical text, where a series of sesame dots may appear beside the main text, as a sidelining to provide visual emphasis. In this respect, their usage is similar to such characters as U+FE34 PRESENTATION FORM FOR VERTICAL WAVY LOW LINE, which are also used for sidelining vertical text for emphasis. Despite being encoded in the block for CJK compatibility forms, the sesame dots are not compatibility characters. They are in general typographic use and are found in the Japanese standard, JIS X 0213.

U+FE45 SESAME DOT is historically related to U+3001 IDEOGRAPHIC COMMA, but is not simply a vertical form variant of it. The function of an *ideographic comma* in connected text is distinct from that of a *sesame dot*.

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 12.1, Han*.

CJK Compatibility Forms

Vertical Forms. CJK vertical forms are compatibility characters encoded for compatibility with legacy implementations that encode these characters explicitly when Chinese text is

being set in vertical rather than horizontal lines. The preferred Unicode approach to representation of such text is to simply use the nominal characters that correspond to these vertical variants. Then, at display time, the appropriate glyph is selected according to the line orientation.

The Unicode Standard contains two blocks devoted primarily to these CJK vertical forms. The CJK Vertical Forms block, U+FE10..U+FE1F, contains compatibility characters needed for round-trip mapping to the Chinese standard, GB 18030. The CJK Compatibility Forms block, U+FE30..U+FE4F, contains forms found in the Chinese standard, CNS 11643.

Styled Overscores and Underscores. The CJK Compatibility Forms block also contains a number of compatibility characters from CNS 11643, which consist of different styles of overscores or underscores. They were intended, in the Chinese standard, for the representation of various types of overlining or underlining, for emphasis of text when laid out *horizontally.* Except for round-trip mapping with legacy character encodings, the use of these characters is discouraged; use of styles is the preferred way to handle such effects in modern text rendering.

Small Form Variants. CNS 11643 also contains a number of small variants of ASCII punctuation characters. The Unicode Standard encodes those variants as compatibility characters in the Small Form Variants block, U+FE50..U+FE6F. Those characters, while construed as fullwidth characters, are nevertheless depicted using small forms that are set in a fullwidth display cell. (See the discussion in *Section 12.4, Hiragana and Katakana*.) These characters are provided for compatibility with legacy implementations.

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.

Fullwidth and Halfwidth Variants. For compatibility with East Asian legacy character sets, the Unicode Standard encodes fullwidth variants of ASCII punctuation and halfwidth variants of CJK punctuation. See *Section 12.5, Halfwidth and Fullwidth Forms*, for more information.

