Chapter 7

European Alphabetic Scripts

Modern European alphabetic scripts are derived from or influenced by the Greek script. The word *alphabet* is derived from the Greek word *alphabetos*, itself derived from the names of the first two Greek letters, alpha and beta. The Greek script itself is an adaptation of the Phoenician alphabet. A Greek innovation was writing the letters from left to right, which is the writing direction for all the scripts derived from or inspired by Greek.

The European alphabetic scripts in the Unicode Standard,

- Latin
- Greek
- Cyrillic
- Armenian
- Georgian
- Ogham
- Runic

are written from left to right. Many have separate lowercase and uppercase forms of the alphabet. Spaces are used to separate words. Accents and diacritical marks are used to indicate phonetic features and to extend the use of base scripts to additional languages. The process of applying diacritical marks is potentially open-ended—one of the reasons combining marks are encoded in the Unicode Standard.

Latin and Cyrillic are used to write or transliterate texts in a wide variety of languages. The Latin alphabet is derived from the alphabet used by the Etruscans, who had adopted a western variant of the classical Greek alphabet. Originally it contained only 24 capital letters. The modern Latin alphabet as it is found in the Basic Latin block owes its appearance to innovations of scribes during the Middle Ages and practices of the early Renaissance printers. The Cyrillic script was developed in the ninth century and is also based on Greek.

The Georgian and Armenian scripts were invented in the fifth century and are influenced by Greek. Modern Georgian does not have separate upper- and lowercase forms.

The International Phonetic Alphabet is an extension of the Latin alphabet, enabling it to represent the phonetics of all languages.

The two historic scripts of northwestern Europe, Runic and Ogham, have a distinct appearance owing to their primary use in carving inscriptions in stone and wood. They are conventionally rendered left to right in scholarly literature, but on the original stone carvings often proceeded in an arch tracing the outline of the stone.

7.1 Latin

The Latin script was derived from the Greek script. Today it is used to write a wide variety of languages all over the world. In the process of adapting it to other languages, numerous extensions have been devised. The most common is the addition of diacritical marks. Furthermore, the creation of digraphs, inverse or reverse forms, and outright new characters have all been used to extend the Latin script.

The Latin script is written in linear sequence from left to right. Spaces are used to separate words and provide the primary line-breaking opportunities. Hyphens are used where lines are broken in the middle of a word. (For more information, see Unicode Technical Report #14, "Line Breaking Properties," on the CD-ROM or the up-to-date version on the Unicode Web site.) Latin letters come in upper- and lowercase pairs.

Diacritical Marks. Speakers of different languages treat the addition of a diacritical mark to a base letter differently. In some languages, the combination is treated as a letter in the alphabet for the language. In others, such as English, the same words can often be spelled with and without the diacritical mark without implying any difference. Most languages that use the Latin script treat letters with diacritical marks as variations of the base letter, but do not accord the combination the full status of an independent letter in the alphabet. The encoding for the Latin script in the Unicode standard is sufficiently flexible to allow implementations to support these letters according to the users' expectation, as long as the language is known. Widely used accented character combinations are provided as single characters to accommodate interoperation with pervasive practice in legacy encodings. Combining diacritical marks can express these and all other accented letters as combining character sequences.

In the Unicode Standard, all diacritical marks are encoded in sequence *after the base characters to which they apply.* For more details, see subsection on "Combining Diacritical Marks" in *Section 7.9, Combining Marks*, and also *Section 2.6, Combining Characters*.

Standards. Unicode follows ISO 8859-1 in the layout of Latin letters up to U+00FF. ISO 8859-1, in turn, is based on older standards, among others ASCII (ANSI X3.4), which is identical to ISO/IEC 646:1991-IRV. Like ASCII, ISO 8859-1 contains Latin letters, punctuation signs, and mathematical symbols. The use of the additional characters is not restricted to the context of Latin script usage. The description of these characters is found in *Chapter 6, Punctuation*.

Related Characters. For other Latin or Latin-derived characters, see letterlike symbols (U+2100..U+214F), currency symbols (U+20A0..U+20CF), miscellaneous symbols (U+2600..U+26FF), enclosed alphanumberics (U+2460..U+24FF), and fullwidth forms (U+FF21..U+FF5A).

Letters of Basic Latin: U+0041–U+007A

Only a small fraction of the languages written with the Latin script can be written entirely with the basic set of 26 uppercase and 26 lowercase Latin letters contained in this block. The 26 basic letter pairs form the core of the alphabets used by all the other languages that use the Latin script. A stream of text using one of these alphabets would therefore intermix characters from the Basic Latin block and other Latin blocks.

Occasionally a few of the basic letter pairs are omitted, such as in Italian, which does not use "j" or "w".

Alternative Graphics. Common typographical variations include the open- and closed-loop form of the lowercase letters "a" and "g". Phonetic transcription systems, such as IPA and Pinyin, make a distinction between such forms.

Letters of the Latin-1 Supplement: U+00C0–U+00FF

The Latin-1 supplement extends the basic 26 letter pairs of ASCII by providing additional letters for major languages of Europe (listed in the next paragraph). Like ASCII, the Latin-1 set also includes a miscellaneous set of punctuation and mathematical signs. Punctuation, signs, and symbols not included in the Basic Latin and Latin-1 Supplement blocks are encoded in character blocks starting with the General Punctuation block.

Languages. The languages supported by the Latin-1 supplement include Danish, Dutch, Faroese, Finnish, Flemish, German, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish, and Swedish.

Ordinals. U+00AA FEMININE ORDINAL INDICATOR and U+00BA MASCULINE ORDINAL INDICATOR can be depicted with an underscore, but many modern fonts show them as superscripted Latin letters with no underscore. In sorting and searching, these characters should be treated as weakly equivalent to their Latin character equivalents.

Spacing Clones of Diacritics. ISO 8859-1 contains eight characters that are ambiguous regarding whether they denote combining characters or separate spacing characters. In the Unicode Standard, the corresponding codepoints (U+005E ^ CIRCUMFLEX ACCENT, U+005F _ LOW LINE, U+0060 ` GRAVE ACCENT, U+007E ~ TILDE, U+00A8 ¨ DIAERESIS, U+00AF ¯ MACRON, U+00B4 ´ ACUTE ACCENT, and U+00B8 ¸ CEDILLA) are restricted to use as spacing characters. The Unicode Standard provides unambiguous combining characters in the character block for Combining Diacritical Marks, which can be used to represent accented Latin letters by means of composed character sequences. U+00B0 ° DEGREE SIGN is also occasionally used ambiguously by implementations of ISO 8859-1 to denote a spacing form of a diacritic ring above a letter; in the Unicode Standard, that spacing diacritical mark is denoted unambiguously by U+02DA ˚ RING ABOVE. U+007E "~" TILDE is ambiguous between usage as a spacing form of a diacritic and as an operator or other punctuation; it is generally rendered with a center line glyph, rather than as a diacritic raised tilde. The spacing form of the diacritic tilde is denoted unambiguously by U+02DC "˜" SMALL TILDE.

Latin Extended-A: U+0100–U+017F

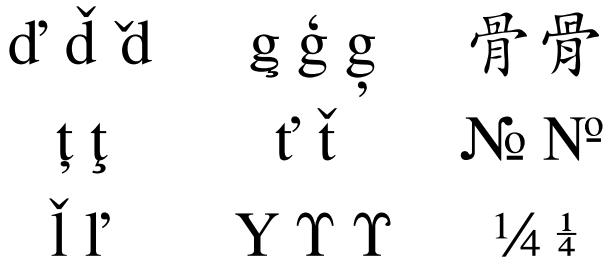
The Latin Extended-A block contains a collection of letters that, when added to the letters contained in the Basic Latin and Latin-1 Supplement blocks, allow for the representation of most European languages that employ the Latin script. Many other languages can also be written with the characters in this block. Most of these characters are equivalent to precomposed combinations of base character forms and combining diacritical marks. These combinations may also be represented by means of composed character sequences. See *Section 2.6, Combining Characters.*

Standards. This block includes characters contained in International Standard ISO 8859—Part 2. Latin alphabet No. 2, Part 3. Latin alphabet No. 3, Part 4. Latin alphabet No. 4, and Part 9. Latin alphabet No. 5. Many of the other graphic characters contained in these standards, such as punctuation, signs, symbols, and diacritical marks, are already encoded in the Latin-1 Supplement block. Other characters from these parts of ISO 8859 are encoded in other blocks, primarily in the Spacing Modifier Letters block (U+02B0..U+02FF) and in the character blocks starting at and following the General Punctuation block.

Languages. Most languages supported by this block also require the concurrent use of characters contained in the Basic Latin and Latin-1 Supplement blocks. When combined with these two blocks, the Latin Extended-A block supports Afrikaans, Basque, Breton, Catalan, Croatian, Czech, Esperanto, Estonian, French, Frisian, Greenlandic, Hungarian, Latin, Latvian, Lithuanian, Maltese, Polish, Provençal, Rhaeto-Romanic, Romanian, Romany, Sami, Slovak, Slovenian, Sorbian, Turkish, Welsh, and many others.

Alternative Glyphs. Some characters have alternative representations, although they have a common semantic. In such cases, a preferred glyph is chosen to represent the character in the code charts, even though it may not be the form used under all circumstances. Some examples to illustrate this point are provided in *Figure 7-1* and discussed in the text that follows.

Figure 7-1. Alternative Glyphs



When Czech is printed in books, U+010F LATIN SMALL LETTER D WITH CARON and U+0165 LATIN SMALL LETTER T WITH CARON letter forms with apostrophe are often used instead of letter forms with caron (hacek) over the base forms. In Slovak, this use also applies to U+013E LATIN SMALL LETTER L WITH CARON. The use of an apostrophe can avoid some line crashes over the ascenders of those letters and so result in better typography. In typewritten or handwritten documents, or in didactic and pedagogical material, on the other hand, letter forms with haceks are preferred. Languages other than Czech or Slovak that make use of these characters may simply choose to always use the forms with haceks.

A similar situation can be seen in the Latvian letter U+0123 LATIN SMALL LETTER G WITH CEDILLA. In good Latvian typography, this character is always shown with a rotated comma *over* the g, rather than a cedilla below the g, because of the typographical design and layout issues resulting from trying to place a cedilla below the descender loop of the g. Poor Latvian fonts may substitute an acute accent for the rotated comma, and handwritten or other printed forms may actually show the cedilla below the g. The uppercase form of the letter is always shown with a cedilla, as the rounded bottom of the G poses no problems for attachment of the cedilla.

Other Latvian letters with cedilla below (U+0137 LATIN SMALL LETTER K WITH CEDILLA, U+0146 LATIN SMALL LETTER N WITH CEDILLA, and U+0157 LATIN SMALL LETTER R WITH CEDILLA) always prefer a glyph with a floating comma below as there is no proper attachment point for a cedilla proper at the bottom of the base form.

In Turkish and Romanian, a cedilla and a comma below can replace one another depending on the font style. The letters U+015F LATIN SMALL LETTER S WITH CEDILLA and U+0163 LATIN SMALL LETTER T WITH CEDILLA (and their uppercase counterparts) have been duplicated at U+0219 LATIN SMALL LETTER S WITH COMMA BELOW and U+021B LATIN SMALL LETTER T WITH COMMA BELOW. The duplicated characters with explicit commas below are provided *solely* for compatibility with sociopolitical practices. Legacy encodings for these

characters, including ISO/IEC 8859-2, contain only a single form of each of these characters, which is mapped to the form with cedilla.

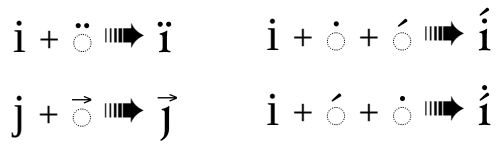
In general, characters with cedillas or ogoneks below are subject to variable typographical usage, depending on the availability and quality of fonts used, the technology, and the geographic area. Various hooks, commas, and squiggles may be substituted for the nominal forms of these diacritics below, and even the direction of the hooks may be reversed. Implementers should take care to become familiar with particular typographical traditions before assuming that characters are missing or are wrongly represented in the code charts in the Unicode Standard.

Exceptional Case Pairs. The characters U+0130 LATIN CAPITAL LETTER I WITH DOT ABOVE and U+0131 LATIN SMALL LETTER DOTLESS I (used primarily in Turkish) are assumed to take ASCII "i" and "I" as their case alternates, respectively. This mapping makes the corresponding reverse mapping language-specific; mapping in both directions requires special attention from the implementor (see *Section 5.17, Sorting and Searching*). See SpecialCasing.txt on the CD-ROM for more information.

***Diacritics on* i *and* j.** A dotted (normal) *i* or *j* followed by a top nonspacing mark loses the dot in rendering. Thus, in the word *naïve,* the *ï* could be spelled with *i + diaeresis.* Just as Cyrillic A is not equivalent to Latin A, a *dotted-i* is not equivalent to a Turkish *dotless-i + overdot,* nor are other cases of accented *dotted-i* equivalent to accented *dotless-i* (for example, i + ¨ ≠ ı + ¨). The same pattern is used for *j.*

To express the forms sometimes used in the Baltic (where the dot is retained under a top accent), use *i + overdot + accent* (see *Figure 7-2*).

Figure 7-2. Diacritics on *i* and *j*



Latin Extended-B: U+0180–U+024F

The Latin Extended-B block contains letter forms used to extend Latin scripts to represent additional languages. It also contains phonetic symbols not included in the International Phonetic Alphabet (see the IPA Extensions block, U+0250..U+02AF).

Standards. This block covers, among others, characters in ISO 6438 Documentation—African coded character set for bibliographic information interchange, *Pinyin* Latin transcription characters from the People's Republic of China national standard GB 2312 and from the Japanese national standard JIS X 0212, and Sami characters from ISO 8859 Part 10. Latin alphabet No. 6.

Arrangement. The characters are arranged in a nominal alphabetical order, followed by a small collection of Latinate forms. Upper- and lowercase pairs are placed together where possible, but in many instances the other case form is encoded at some distant location and so is cross-referenced. Variations on the same base letter are arranged in the following order: turned, inverted, hook attachment, stroke extension or modification, different style (script), small cap, modified basic form, ligature, and Greek-derived.

Croatian Digraphs Matching Serbian Cyrillic Letters. Serbo-Croatian is a single language with paired alphabets: a Latin script (Croatian) and a Cyrillic script (Serbian). A set of

compatibility digraph codes is provided for one-to-one transliteration. There are two potential uppercase forms for each digraph, depending on whether only the initial letter is to be capitalized (titlecase), or both (all uppercase). The Unicode Standard offers both forms so that software can convert one form to the other without changing font sets. The appropriate cross-references are given for the lowercase letters. For more information about canonical equivalence, see *Chapter 3, Conformance*.

Pinyin Diacritic-Vowel Combinations. The Chinese standard GB 2312, as well as the Japanese standard JIS X 0212, includes a set of codes for Pinyin, used for Latin transcription of Mandarin Chinese. Most of the letters used in Pinyin romanization (even those with combining diacritical marks) are already covered in the preceding Latin blocks. The group of 16 characters provided here completes the Pinyin character set specified in GB 2312 and JIS X 0212.

Case Pairs. A number of characters in this block are uppercase forms of characters whose lowercase form is part of some other grouping. Many of these characters came from the International Phonetic Alphabet; they acquired novel uppercase forms when they were adopted into Latin script-based writing systems. Occasionally, however, *alternative* uppercase forms arose in this process. In some instances, research has shown that alternative uppercase forms are merely variants of the same character. If so, such variants are assigned a single Unicode value, as is the case of U+01B7 LATIN CAPITAL LETTER EZH. But when research has shown that two uppercase forms are actually used in different ways, then they are given different codes; such is the case for U+018E LATIN CAPITAL LETTER REVERSED E and U+018F LATIN CAPITAL LETTER SCHWA. In this instance, the shared lowercase form is copied to enable unique case-pair mappings if desired: U+01DD LATIN SMALL LETTER TURNED E is a copy of U+0259 LATIN SMALL LETTER SCHWA.

For historical reasons, the names of some case pairs differ. For example, U+018E LATIN CAPITAL LETTER REVERSED E is the uppercase of U+01DD LATIN SMALL LETTER TURNED E—not of U+0258 LATIN SMALL LETTER REVERSED E. (For default case mappings of Unicode characters, see *Chapter 4, Character Properties.*)

Languages. Some indication of language or other usage is given for most characters within the names lists accompanying the character charts.

IPA Extensions: U+0250–U+02AF

The IPA Extensions block contains primarily the unique symbols of the International Phonetic Alphabet (IPA), which is a standard system for indicating specific speech sounds. The IPA was first introduced in 1886 and has undergone occasional revisions of content and usage since that time. The Unicode Standard covers all single symbols and all diacritics in the last published IPA revision (1989), as well as a few symbols in former IPA usage that are no longer currently sanctioned. A few symbols have been added to this block that are part of the transcriptional practices of Sinologists, Americanists, and other linguists. Some of these practices have usages independent of the IPA and may use characters from other Latin blocks rather than IPA forms. Note also that a few nonstandard or obsolete phonetic symbols are encoded in the Latin Extended-B block.

An essential feature of IPA is the use of combining diacritical marks. IPA diacritical mark characters are coded in the Combining Diacritical Marks block, U+0300..U+036F. In IPA, diacritical marks can be freely applied to base form letters to indicate fine degrees of phonetic differentiation required for precise recording of different languages.

Standards. The characters in this block are taken from the 1989 revision of the International Phonetic Alphabet, published by the International Phonetic Association. The International Phonetic Association standard considers IPA to be a separate alphabet, so it

includes the entire Latin lowercase alphabet *a–z*, a number of extended Latin letters such as U+0153 LATIN SMALL LIGATURE OE œ, and a few Greek letters and other symbols as separate and distinct characters. In contrast, the Unicode Standard does not duplicate either the Latin lowercase letters *a–z* or other Latin or Greek letters in encoding IPA. Note that unlike other character standards referenced by the Unicode Standard, IPA constitutes an extended alphabet and phonetic transcriptional standard, rather than a character encoding standard.

Unifications. The IPA symbols are unified as much as possible with other letters, albeit not with nonletter symbols like U+222B ∫ INTEGRAL. The IPA symbols have also been adopted into the Latin-based alphabets of many written languages, such as some used in Africa. It is futile to attempt to distinguish a transcription from an actual alphabet in such cases. Therefore, many IPA symbols are found outside the IPA Extensions block. IPA symbols that are not found in the IPA Extensions block are listed as cross-references at the beginning of the character names list for this block.

IPA Alternates. In a few cases IPA practice has, over time, produced alternate forms, such as U+0269 LATIN SMALL LETTER IOTA "ɩ" versus U+026A LATIN LETTER SMALL CAPITAL I "ɪ." The Unicode Standard provides separate encodings for the two alternate forms because they are used in a meaningfully distinct fashion.

Case Pairs. IPA does not sanction case distinctions; in effect, its phonetic symbols are all lowercase. When IPA symbols are adopted into a particular alphabet and used by a given written language (as has occurred, for example, in Africa) they acquire uppercase forms. Because these uppercase forms are not themselves IPA symbols, they are generally encoded in the Latin Extended-B block (or other Latin extension blocks) and are cross-referenced with the IPA names list.

Typographic Variants. IPA includes typographic variants of certain Latin and Greek letters that would ordinarily be considered variations of font style rather than of character identity, such as SMALL CAPITAL letter forms. Examples include a typographic variant of the Greek letter *phi* ɸ, as well as the borrowed letter Greek *iota* ɩ, which has a unique Latin uppercase form. These forms are encoded as separate characters in the Unicode Standard because they have distinct semantics in plain text.

Affricate Digraph Ligatures. IPA officially sanctions six digraph ligatures used in transcription of coronal affricates. These are encoded at U+02A3..U+02A8. The IPA digraph ligatures are explicitly defined in IPA and also have possible semantic values that make them not simply rendering forms. Thus, for example, while U+02A6 LATIN SMALL LETTER TS DIGRAPH is a transcription for the sounds that could also be transcribed in IPA as "ts" U+0074 U+0073, the choice of the digraph ligature may be the result of a deliberate distinction made by the transcriber regarding the systematic phonetic status of the affricate. The choice of whether to ligate cannot be left to rendering software based on the font available. This ligature also differs in typographical design from the ts ligature found in some old-style fonts.

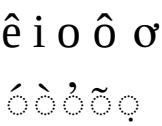
Encoding Structure. The IPA Extensions block is arranged in approximate alphabetical order according to the Latin letter that is graphically most similar to each symbol. This order has nothing to do with a phonetic arrangement of the IPA letters.

Latin Extended Additional: U+1E00–U+1EFF

The characters in this block constitute a number of precomposed combinations of Latin letters with one or more general diacritical marks. Each of the characters contained in this block may be alternatively represented with a base letter followed by one or more general diacritical mark characters found in the Combining Diacritical Marks block. A canonical form for such alternative representations is specified in *Chapter 3, Conformance*.

Vietnamese Vowel Plus Tone Mark Combinations. A portion of this block (U+1EA0.. U+1EF9) comprises vowel letters of the modern Vietnamese alphabet (quốc ngữ) combined with a diacritic mark that denotes the phonemic tone that applies to the syllable. In the modern Vietnamese alphabet, there are 12 vowel letters and five tone marks (see *Figure 7-3*).

Figure 7-3. Vietnamese Letters and Tone Marks



Some implementations of Vietnamese systems prefer storing the combination of vowel letter and tone mark as a singly encoded element; other implementations prefer storing the vowel letter and tone mark separately. The former implementations will use characters defined in this block along with combination forms defined in the Latin-1 Supplement and Latin Extended-A character blocks; the latter implementations will use the basic vowel letters in the Basic Latin, Latin-1 Supplement, and Latin Extended-A blocks along with characters from the Combining Diacritical Marks block. For these latter implementations, the characters U+0300 COMBINING GRAVE, U+0309 COMBINING HOOK ABOVE, U+0303 COMBINING TILDE, U+0301 COMBINING ACUTE, and U+0323 COMBINING DOT BELOW should be used in representing the Vietnamese tone marks. The characters U+0340 COMBINING GRAVE TONE MARK and U+0341 COMBINING ACUTE TONE MARK are deprecated and should not be used.

Latin Ligatures: FB00–FB06

This section of the Alphabetic Presentation forms block (U+FB00..U+FB4F) contains several common Latin ligatures, which occur in legacy encodings. By design, the Unicode standard does not provide a general mechanism to indicate where ligatures should be displayed. Where to place a Latin ligature is a matter of typographical style as well as the orthographical rules of the language. Some languages prohibit ligatures across word boundaries. In these cases, it is preferable for the implementations to use unligated characters in the backing store and provide out-of-band information to the display layer where ligatures may be placed.

7.2 Greek

Greek: U+0370–U+03FF

The Greek script is used for writing the Greek language and (in an extended variant) the Coptic language. The Greek script had a strong influence in the development of the Latin and Cyrillic scripts.

The Greek script is written in linear sequence from left to right with the occasional use of nonspacing marks. Greek letters come in upper- and lowercase pairs.

Standards. The Unicode encoding of Greek is based on ISO 8859-7, which is equivalent to the Greek national standard ELOT 928. The Unicode Standard encodes Greek characters in the same relative positions as in ISO 8859-7. A number of variant and archaic characters are taken from the bibliographic standard ISO 5428.

Polytonic Greek. Polytonic Greek, used for ancient Greek (classical and Byzantine), may be encoded using either combining character sequences or precomposed base plus diacritic combinations. For the latter, see the following subsection, "Greek Extended: U+1F00–U+1FFF."

Nonspacing Marks. Several nonspacing marks commonly used with the Greek script are found in the Combining Diacritical Marks range (see *Table 7-1*).

Table 7-1. Nonspacing Marks Used with Greek

Code	Name	Alternative Names
U+0300	COMBINING GRAVE ACCENT	*varia*
U+0301	COMBINING ACUTE ACCENT	*tonos, oxia*
U+0302	COMBINING CIRCUMFLEX ACCENT	
U+0303	COMBINING TILDE	
U+0304	COMBINING MACRON	
U+0306	COMBINING BREVE	
U+0308	COMBINING DIAERESIS	*dialytika*
U+0313	COMBINING COMMA ABOVE	*psili*
U+0314	COMBINING REVERSED COMMA ABOVE	*dasia*
U+0342	COMBINING GREEK PERISPOMENI	*circumflex, tilde*
U+0343	COMBINING GREEK KORONIS	*comma above*
U+0345	COMBINING GREEK YPOGEGRAMMENI	*iota subscript*

Because the characters in the Combining Diacritical Marks block are encoded by shape, not by meaning, they are appropriate for use in Greek where applicable. However, the character U+0344 COMBINING GREEK DIALYTIKA TONOS is deprecated and should not be used. The combination of *dialytika* plus *tonos* is instead represented by the sequence U+0308 COMBINING DIAERESIS plus U+0301 COMBINING ACUTE.

Multiple nonspacing marks applied to the same baseform character are encoded in inside-out sequence. See the general rules for applying nonspacing marks in *Section 2.6, Combining Characters*.

The basic Greek accent written in modern Greek is called *tonos*. It is represented by an acute accent (U+0301). The shape that the acute accent takes over Greek letters is generally steeper than that shown over Latin letters in Western European typographic traditions, and

in earlier editions of this standard was mistakenly shown as a vertical line over the vowel. Polytonic Greek has several contrastive accents, and the accent, or *tonos*, written with an acute accent is referred to as *oxia*, in contrast to the *varia*, which is written with a grave accent.

U+0342 COMBINING GREEK PERISPOMENI may appear as either a circumflex or a tilde: ◌͂ versus ◌̃. Because of this variation in form, the *perispomeni* was encoded distinctly from U+0303 COMBINING TILDE.

U+0313 COMBINING COMMA ABOVE and U+0343 COMBINING GREEK KORONIS both take the form of a raised comma over a baseform letter. U+0343 COMBINING GREEK KORONIS was included for compatibility reasons; U+0313 COMBINING COMMA ABOVE is the preferred form for general use.

The nonspacing mark *ypogegrammeni* (also known as *iota subscript* in English) can be applied to the vowels *alpha*, *eta*, and *omega* to represent historic diphthongs. This mark appears as a small *iota* below the vowel. When applied to a single uppercase vowel, the iota does not appear as a subscript, but is instead normally rendered as a regular lowercase iota to the right of the uppercase vowel. This form of the iota is called *prosgegrammeni* (also known as *iota adscript* in English). In completely uppercased words, the iota subscript should be replaced by a capital iota. See SpecialCasing.txt on the CD-ROM. Archaic representations of Greek words (which did not have lowercase or accents) use the Greek capital letter iota following the vowel for these diphthongs. Such archaic representations require special case mapping.

Variant Letterforms. U+03A5 GREEK CAPITAL LETTER UPSILON has two common forms—one looks essentially like the Latin capital Y, and the other has two symmetric upper branches that curl like rams' horns, "ϒ". The Y-form glyph has been chosen consistently for use in the code charts, both for monotonic and polytonic Greek. For mathematical usage, the rams' horn form of the glyph is required. Variant forms of certain other Greek letters are encoded as separate characters in ISO/IEC 8859-7 and ISO 5428; therefore, those forms are also included as separate characters in the Unicode Standard. They include U+03C2 GREEK SMALL LETTER FINAL SIGMA and U+03D0 GREEK BETA SYMBOL, as well as additional forms of the capital letter upsilon that have an asymmetric hook—for example, U+03D2 GREEK UPSILON WITH HOOK SYMBOL.

Greek Letters as Symbols. For compatibility purposes, a few Greek letters are separately encoded as symbols in other character blocks. Examples include U+00B5 MICRO SIGN µ in the Latin-1 Supplement character block and U+2126 OHM SIGN Ω in the Letterlike Symbols character block. The use of Greek letters for mathematical variables and operators is well established. Characters from the Greek block may be used for these symbols.

Punctuation-like Characters. The question of which punctuation-like characters are uniquely Greek and which ones can be unified with generic Western punctuation has no definitive answer. The Greek question mark U+037E GREEK QUESTION MARK *erotimatiko* ";" is encoded for compatibility. The preferred character is U+003B SEMICOLON.

Historic Letters. Historic Greek letters have been retained from ISO 5428.

Coptic-Unique Letters. The Coptic script is regarded primarily as a stylistic variant of the Greek alphabet. The letters unique to Coptic are encoded in a separate range at the end of the Greek character block. Those characters may be used together with the basic Greek characters to represent the complete Coptic alphabet. Coptic text may be rendered using a font that contains the Coptic style of depicting the characters it shares with the Greek alphabet. Texts that mix Greek and Coptic languages together must employ appropriate font style associations.

Related Characters. For math symbols, see mathematical operators (U+2200..U+22FF). For additional punctuation to be used with this script, see C0 Controls and ASCII Punctuation (U+0000..U+007F).

Greek Extended: U+1F00–U+1FFF

The characters in this block constitute a number of precomposed combinations of Greek letters with one or more general diacritical marks; in addition, a number of spacing forms of Greek diacritical marks are provided here. In particular, these characters facilitate the representation of Polytonic Greek texts without the use of combining marks.

Each of the characters contained in this block may be alternatively represented with a base letter from the Greek block followed by one or more general diacritical mark characters found in the Combining Diacritical Marks block. A canonical form for such alternative representations is specified in *Chapter 3, Conformance*.

Spacing Diacritics. Sixteen additional spacing diacritic marks are provided in this character block for use in the representation of Polytonic Greek texts. Each has an alternative representation for use with systems that support nonspacing marks. The Unicode Standard considers the nonspacing alternative forms to be the canonical Unicode representation of the information represented by the spacing forms. The nonspacing alternatives appear in *Table 7-2*.

Table 7-2. Greek Spacing and Nonspacing Pairs

Spacing Form	Nonspacing Form
1FBD GREEK KORONIS	0313 COMBINING COMMA ABOVE
037A GREEK YPOGEGRAMMENI	0345 COMBINING YPOGEGRAMMENI
1FBF GREEK PSILI	0313 COMBINING COMMA ABOVE
1FC0 GREEK PERISPOMENI	0342 COMBINING GREEK PERISPOMENI
1FC1 GREEK DIALYTIKA AND PERISPOMENI	0308 COMBINING DIAERESIS + 0342 COMBINING GREEK PERISPOMENI
1FCD GREEK PSILI AND VARIA	0313 COMBINING COMMA ABOVE + 0300 COMBINING GRAVE ACCENT
1FCE GREEK PSILI AND OXIA	0313 COMBINING COMMA ABOVE + 0301 COMBINING ACUTE ACCENT
1FCF GREEK PSILI AND PERISPOMENI	0313 COMBINING COMMA ABOVE + 0342 COMBINING GREEK PERISPOMENI
1FDD GREEK DASIA AND VARIA	0314 COMBINING REVERSED COMMA ABOVE + 0300 COMBINING GRAVE ACCENT
1FDE GREEK DASIA AND OXIA	0314 COMBINING REVERSED COMMA ABOVE + 0301 COMBINING ACUTE ACCENT
1FDF GREEK DASIA AND PERISPOMENI	0314 COMBINING REVERSED COMMA ABOVE + 0342 COMBINING GREEK PERISPOMENI
1FED GREEK DIALYTIKA AND VARIA	0308 COMBINING DIAERESIS + 0300 COMBINING GRAVE ACCENT
1FEE GREEK DIALYTIKA AND OXIA	0308 COMBINING DIAERESIS + 0301 COMBINING ACUTE ACCENT
1FEF GREEK VARIA	0300 COMBINING GRAVE ACCENT
1FFD GREEK OXIA	0301 COMBINING ACUTE ACCENT
1FFE GREEK DASIA	0314 COMBINING REVERSED COMMA ABOVE

Decomposition of Spacing Forms. When decomposing the spacing forms, the spacing status of the implied usage must be taken into account. Unless information is present to the

contrary, these spacing forms would be decomposed to U+0020 SPACE followed by the nonspacing form equivalents shown in *Table 7-2*.

In archaic forms of Greek, U+0345 COMBINING GREEK YPOGEGRAMMENI and the precomposed forms that contain it have special case mappings.

7.3 Cyrillic

Cyrillic: U+0400–U+04FF

The Cyrillic script is a member of the family of scripts strongly influenced by the Greek script. Cyrillic has traditionally been used for writing various Slavic languages, among which Russian is predominant. In the nineteenth and early twentieth centuries, Cyrillic was extended to write the non-Slavic minority languages of the former Soviet Union. The Cyrillic script is written in linear sequence from left to right with the occasional use of non-spacing marks. Cyrillic letters come in upper- and lowercase pairs.

Standards. The Cyrillic block of the Unicode Standard is based on ISO 8859-5. The Unicode Standard encodes Cyrillic characters in the same relative positions as in ISO 8859-5.

Unifications. Latin characters included in those alphabets that use both Latin and Cyrillic letters are not given duplicate Cyrillic encodings. Examples include *q* and *w* for Kurdish.

Historic Letters. The historic form of the Cyrillic alphabet is treated as a font style variation of modern Cyrillic because the historic forms are relatively close to the modern appearance and because some of them are still in modern use in languages other than Russian (for example, U+0406 CYRILLIC CAPITAL LETTER I "I"is used in modern Ukrainian and Byelorussian). Since the historic Cyrillic characters encoded in Unicode (U+0460.. U+0486) rarely occur in modern form, these letters are shown in the charts in an archaic font. A complete Old Cyrillic set would be obtained by rendering the whole Cyrillic section (that is, U+0400..U+0486) in that same style.

Extended Cyrillic. These letters are used in alphabets for minority languages of the former Soviet Union. The scripts of some of these languages have often been revised in the past; the Unicode Standard includes only the alphabets in current use, not the rejected old letterforms.

Glagolitic. The history of the creation of the Slavic scripts and their relationship has been lost. The Unicode Standard regards Glagolitic as a *separate* script from Cyrillic, not as a font change from Cyrillic. This position is taken primarily because Glagolitic appears unrecognizably different from Cyrillic, and secondarily because Glagolitic has not grown to match the expansion of Cyrillic. The Glagolitic script is not currently supported by the Unicode Standard.

7.4 Armenian

Armenian: U+0530–U+058F

The Armenian script is used primarily for writing the Armenian language. The script is written from left to right and generally does not use diacritics (except for the modifier letters specified below). It does have upper- and lowercase pairs.

Modifier Letters. In modern Armenian typography, the small marks in the group called Armenian modifier letters are placed above and to the right of other letters so that they occupy a letter position of their own. For example, the emphasis mark, the exclamation mark, and the interrogation mark are positioned to the right of the vowel in the syllable that is stressed. As the rendering of these modifiers may involve horizontal and vertical adjustments, it should ideally be done with enhanced kerning as described in *Section 5.15, Locating Text Element Boundaries*. Because these modifiers generally have their own width, they are treated as spacing letters in the Unicode Standard, rather than as nonspacing marks. There appears to be no recorded usage of U+0559 ARMENIAN MODIFIER LETTER LEFT HALF RING in Armenian text; its presence in this block is thus probably spurious.

Punctuation Letters. The Armenian writing system uses many punctuation letters from other blocks, such as U+002C COMMA and U+00B7 MIDDLE DOT. However, when used in Armenian, such punctuation letters should be rendered in a style compatible with the Armenian font style. In addition to U+055D ARMENIAN COMMA, which is already included in the modifier letters, two punctuation letters are specific to Armenian: U+0589 ARMENIAN FULL STOP and U+058A ARMENIAN HYPHEN. The behavior of the latter character is similar to U+00AD SOFT HYPHEN. It is used to indicate a line-breaking opportunity within a polysyllabic Armenian word. Its shape distinguishes it from the soft hyphen.

Ligatures. Five Armenian ligatures are encoded in the Alphabetic Presentation Forms block in the range U+FB13..U+FB17. By design, the Unicode Standard does not provide a general mechanism to indicate where ligatures should be displayed.

For additional punctuation to be used with this script, see C0 Controls and ASCII Punctuation (U+0000..U+007F).

7.5 Georgian

Georgian: U+10A0–U+10FF

The Georgian script is used primarily for writing the Georgian language. Upper- and lowercase pairs exist primarily in archaic forms of the script.

Script Forms. The modern Georgian script is a lowercase style called *mkhedruli* (soldier's). It originated as the secular derivative of a form called *khutsuri* (ecclesiastical) that had both uppercase and lowercase pairs. Although no longer used in most modern texts, the *khutsuri* style is still used for liturgical purposes; the Unicode Standard encodes the uppercase form of *khutsuri* as well as the lowercase letters of modern Georgian.

Georgian Paragraph Separator. The Georgian paragraph separator has a distinct representation, so it has been separately encoded at U+10FB. It visually marks a paragraph end, but it must be followed by a formatting character such as U+2029 PARAGRAPH SEPARATOR to cause a paragraph break. See the discussion of the *paragraph separator* in *Section 13.2, Layout Controls.*

Other Punctuation. For the Georgian full stop, use U+0589 ARMENIAN FULL STOP.

For additional punctuation to be used with this script, see C0 Controls and ASCII Punctuation (U+0000..U+007F).

7.6 Runic

Runic: U+16A0–U+16F0

The Runic script was historically used to write the languages of the early and medieval societies in the German, Scandinavian, and Anglo-Saxon areas. Use of the Runic script in various forms covers a period from the first century to the nineteenth century. Some 6,000 Runic inscriptions are known. They form an indispensable source of information about the development of the Germanic languages.

Historical Script. The Runic script is one of the first "historical" or "extinct" scripts to be incorporated into the Unicode Standard. The only important use of runes today is in scholarly and popular works about the old Runic inscriptions and their interpretation. The Runic script illustrates many technical problems that are typical for this kind of script. Unlike other scripts in the Unicode Standard, which predominantly serve the needs of the modern user community—with occasional extensions for historic forms—the encoding of the Runic script attempts to suit the needs of texts from different periods of time and from distinct societies that had little contact with each other.

Direction. Like other early writing systems, runes could be written either from left to right or from right to left, or moving first in one direction, then the other (*boustrophedon*), or following the outlines of the inscribed object. At times, characters appear in mirror image, or upside down, or both. In modern scholarly literature, Runic is written left to right. Therefore, the letters of the Runic script have a default directionality of strong left-to-right in this standard.

The Runic Alphabet. Present-day knowledge about runes is incomplete. The set of graphemically distinct units shows greater variation in its graphical shapes than most modern scripts. The Runic alphabet changed several times during its history, both in the number and shapes of the letters contained in it. The shape of most runes can be related to some Latin capital letter, but not necessarily with a letter representing the same sound. The most conspicuous difference between the Latin and the Runic alphabets is the order of the letters.

The Runic alphabet is known as the *futhark* from the name of its first six letters. The original *old futhark* contained 24 runes:

ᚠ ᚢ ᚦ ᚨ ᚱ ᚲ ᚷ ᚹ ᚺ ᚾ ᛁ ᛃ ᛇ ᛈ ᛉ ᛊ ᛏ ᛒ ᛖ ᛗ ᛚ ᛜ ᛞ ᛟ

They are usually transliterated in this way:

f u þ a r k g w h n i j ï p z s t b e m l ŋ d o

In England and Friesland, seven additional runes were added from the fifth century to the ninth century.

In the Scandinavian countries, the *futhark* changed in a different way; in the eighth century, the simplified younger *futhark* appeared. It consists of only 16 runes, some of which are used in two different forms. The long-branch form is shown here:

ᚠ ᚢ ᚦ ᚬ ᚱ ᚴ ᚼ ᚾ ᛁ ᛅ ᛋ ᛏ ᛒ ᛘ ᛚ ᛦ

f u þ o r k h n i a s t b m l ʀ

The use of runes continued in Scandinavia during the Middle Ages. During that time, the *futhark* was influenced by the Latin alphabet and new runes were invented so that there was full correspondence with the Latin letters.

Representative Glyphs. The known inscriptions can include considerable variations of shape for a given rune, sometimes to the point where the nonspecialist will mistake the shape for a different rune. There is no dominant main form for some runes, particularly for many runes added in the Anglo-Friesian and medieval Nordic systems. When transcribing a Runic inscription into its Unicode-encoded form, one cannot rely on the idealized *reference glyph* shape in the character charts alone. One must take into account to which of the four Runic systems an inscription belongs, and be knowledgeable about the permitted form variations within each system. The reference glyphs were chosen to provide an image that distinguishes each rune visually from all other runes in the same system. For actual use, it might be advisable to use a separate font for each Runic system. Of particular note is the fact that the glyph for U+16C4 ᛄ RUNIC LETTER GER is actually a rare form, as the more common form is already used for U+16E1 ᛡ RUNIC LETTER IOR.

Unifications. When a rune in an earlier writing system evolved into several different runes in a later system, the unification of the earlier rune with one of the later runes was based on similarity in graphic form rather than similarity in sound value. In cases where a substantial change in the typical graphical form has occurred, though the historical continuity is undisputed, unification has not been attempted. When runes from different writing systems have the same graphic form but a different origin and denote different sounds, they have been coded as separate characters.

Long-Branch and Short-Twig. Two sharply different graphic forms, the *long-branch* and the *short-twig* form, were used for nine of the 16 Viking Age Nordic runes. Although only one form is used in a given inscription, there are runologically important exceptions. In some cases, the two forms were used to convey different meanings in later use in the medieval system. Therefore the two forms have been separated in the Unicode Standard.

Staveless Runes. Staveless runes are a third form of the Viking Age Nordic runes, a kind of runic shorthand. The number of known inscriptions is small and the graphic forms of many of the runes show great variability between inscriptions. For this reason, staveless runes have been unified with the corresponding Viking Age Nordic runes. The corresponding Viking Age Nordic runes must be used to encode these characters, specifically the short-twig characters, where both short-twig and long-branch characters exist.

Punctuation Marks. The wide variety of Runic punctuation marks has been reduced to three distinct characters based on simple aspects of their graphical form, as very little is known about any difference in intended meaning between marks that look different. Any other punctuation marks have been unified with shared punctuation marks elsewhere in the Unicode Standard.

Golden Numbers. Runes were used as symbols for Sunday letters and golden numbers on calendar staves used in Scandinavia during the Middle Ages. To complete the number series 1–19, three additional calendar runes were added. They are included after the punctuation marks.

Encoding. A total of 81 characters of the Runic script are included in the Unicode Standard. Of these, 75 are Runic letters, 3 are punctuation marks, and 3 are Runic symbols. The order of the Runic characters follows the traditional *futhark* order, with variants and derived runes inserted directly after the corresponding ancestor.

Runic character names are based as much as possible on the sometimes several traditional names for each rune, often with the Latin transliteration at the end of the name.

7.7 Ogham

Ogham: U+1680–U+169F

Ogham is an alphabetic script devised to write a a very early form of Irish. Monumental Ogham inscriptions are found in Ireland, Wales, Scotland, England, and on the Isle of Man. Many of the Scottish inscriptions are undeciphered and may be in Pictish. It is probable that Ogham (Old Irish "Ogam") was widely written in wood in early times. The main flowering of "classical" Ogham, rendered in monumental stone, was in the fifth and sixth centuries. Such inscriptions were mainly employed as territorial markers and memorials; the more ancient examples are standing stones.

The script was originally written along the edges of stone where two faces meet; when written on paper, the central "stemlines" of the script can be said to represent the edge of the stone. Inscriptions written on stemlines cut into the face of the stone, instead of along its edge, are known as "scholastic" and are of a later date (post-seventh century). Notes were also commonly written in Ogham in manuscripts as recently as the sixteenth century.

Structure. The Ogham alphabet consists of 26 distinct characters (*feda*), the first 20 of which are considered to be primary, the last 6 (*forfeda*) supplementary. The four primary series are called *aicmí* (plural of *aicme*, meaning "family"). Each *aicme* was named after its first character, (*Aicme Beithe*, *Aicme Uatha*, meaning "the B Family," "the H Family," and so forth). The character names used in this standard reflect the spelling of the names in modern Irish Gaelic, except that the acute accent is stripped from *Úr*, *Éabhadh*, *Ór*, and *Ifín*, and the mutation of *nGéadal* is not reflected.

Rendering. Ogham text is read beginning from the bottom left side of a stone, continuing upward, across the top and down the right side (in the case of long inscriptions). Monumental Ogham was incised chiefly in a bottom-to-top direction, though there are examples of left-to-right bilingual inscriptions in Irish and Latin. Manuscript Ogham accommodated the horizontal left-to-right direction of the Latin script, and the vowels were written as vertical strokes as opposed to the incised notches of the inscriptions. Ogham should therefore be rendered on computers from left-to-right or from bottom-to-top (never starting from top-to-bottom).

Forfeda (Supplementary Characters). In printed and in manuscript Ogham, the fonts are conventionally designed with a central stemline, but this convention is not necessary. In implementations without the stemline, the character U+1680 OGHAM SPACE MARK should be given its conventional width and simply left blank like U+0020 SPACE. U+169B OGHAM FEATHER MARK and U+169C OGHAM REVERSED FEATHER MARK are used at the beginning and the end of Ogham text, particularly in manuscript Ogham. In some cases, only the *Ogham feather mark* is used, which can indicate the direction of the text.

7.8 Modifier Letters

Spacing Modifier Letters: U+02B0–U+02FF

Modifier letters are an assorted collection of small signs that are generally used to indicate modifications of a preceding letter. A few may modify the following letter, and some may serve as independent letters. These signs are distinguished from diacritical marks in that modifier letters are treated as free-standing, spacing characters. They are distinguished from similar or identical appearing punctuation or symbols by the fact that the members of this block are considered to be letter characters that do not break up a word. They have the "letter" character property (see *Chapter 4, Character Properties*). The majority of these signs are phonetic modifiers, including the characters required for coverage of the International Phonetic Alphabet (IPA).

Phonetic Usage. Modifier letters have relatively well-defined phonetic interpretations. Their usage generally indicates a specific articulatory modification of a sound represented by another letter or intended to convey a particular level of stress or tone. In phonetic usage, the modifier letters are sometimes called "diacritics," which is correct in the logical sense that they are modifiers of the preceding letter. However, in the Unicode Standard, the term "diacritical marks" refers specifically to nonspacing marks, whereas the codes in this block specify *spacing characters*. For this reason, many of the modifier letters in this block correspond to separate diacritical mark codes, which are cross-referenced in *Section 14.1, Character Names List*.

Encoding Principles. This block includes characters that may have different semantic values attributed to them in different contexts. It also includes multiple characters that may represent the same semantic values—there is no necessary one-to-one relationship. The intention of the Unicode encoding is not to resolve the variations in usage, but merely to supply implementers with a set of useful forms from which to choose. The list of usages given for each modifier letter should not be considered exhaustive. For example, the glottal stop (Arabic *hamza*) in Latin transliteration has been variously represented by the characters U+02BC MODIFIER LETTER APOSTROPHE, U+02BE MODIFIER LETTER RIGHT HALF RING, and U+02C0 MODIFIER LETTER GLOTTAL STOP. Conversely, an apostrophe can have several uses; for a list, see the entry for U+02BC MODIFIER LETTER APOSTROPHE in the character names list. There are also instances where an IPA modifier letter is explicitly equated in semantic value to an IPA nonspacing diacritic form.

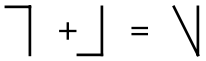
Latin Superscripts. Graphically, some of the phonetic modifier signs are raised or superscripted, some are lowered or subscripted, and some are vertically centered. Only those few forms that have specific usage in IPA or other major phonetic systems are encoded.

Spacing Clones of Diacritics. Some corporate standards explicitly specify spacing and nonspacing forms of combining diacritical marks, and the Unicode Standard provides matching codes for these interpretations when practical. A number of the spacing forms are covered in the Basic Latin and Latin-1 Supplement blocks. The six common European diacritics that do not have encodings there are added as spacing characters. These forms can have multiple semantics, such as U+02D9 DOT ABOVE, which is used as an indicator of the Mandarin Chinese fifth tone.

Rhotic Hook. U+02DE MODIFIER LETTER RHOTIC HOOK is defined in IPA as a free-standing modifier letter. However, in common usage, it is treated as a ligated hook on a baseform letter. Hence, U+0259 LATIN SMALL LETTER SCHWA + U+02DE MODIFIER LETTER RHOTIC HOOK may be treated as equivalent to U+025A LATIN SMALL LETTER SCHWA WITH HOOK.

Tone Letters. U+02E5..U+02E9 comprise a set of basic tone letters, defined in IPA and commonly used in detailed tone transcriptions of African and other languages. Each tone letter refers to one of five distinguishable tone levels. To represent contour tones, the tone letters are used in combinations. The rendering of contour tones follows a regular set of ligation rules that results in a graphic image of the contour (see *Figure 7-4*).

Figure 7-4. Tone Letters



7.9 Combining Marks

Combining Diacritical Marks: U+0300–U+036F

The combining diacritical marks in this block are intended for general use with any script. Diacritical marks specific to a particular script are encoded with the alphabet for that script. Diacritical marks that are primarily used with symbols are defined in the Combining Diacritical Marks for Symbols character block (U+20D0..U+20FF).

Standards. The combining diacritical marks are derived from a variety of sources, including IPA, ISO 5426, and ISO 6937.

Sequence of Base Letters and Diacritics. In the Unicode character encoding, all nonspacing marks, including diacritics, are encoded *after* the base character. For example, the Unicode character sequence U+0061 "a" LATIN SMALL LETTER A, U+0308 "◌̈" COMBINING DIAERESIS, U+0075 "u" LATIN SMALL LETTER U unambiguously encodes "äu", *not* "aü".

The Unicode Standard convention is consistent with the logical order of other nonspacing marks in Semitic and Indic scripts, the great majority of which follow the base characters with respect to which they are positioned. This convention is also in line with the way modern font technology handles the rendering of nonspacing glyphic forms, so that mapping from character memory representation to rendered glyphs is simplified. (For more information on the use of diacritical marks, see *Chapter 2, General Structure*, and *Chapter 3, Conformance*.)

Diacritics Positioned Over Two Base Characters. IPA and a few languages such as Tagalog use diacritics that are applied to two baseform characters. These marks apply to the previous base character—just like all other combining nonspacing marks—but hang over the following letter as well. The two characters U+0360 COMBINING DOUBLE TILDE and U+0361 COMBINING DOUBLE INVERTED BREVE are intended to be displayed as depicted in *Figure 7-5*.

Figure 7-5. Double Diacritics

o + ◌͠ ➠ o͠

o + ◌͠ + o ➠ o͠o

These double diacritics always bind more loosely than all other nonspacing marks except U+0345 *iota subscript*, and thus sort near the end in the canonical representation. In rendering, the double diacritic will float above other diacritics (excluding surrounding diacritics), as in *Figure 7-6*.

Figure 7-6. Ordering of Double Diacritics

o + ◌̂ + ◌͠ + o + ◌̈ ➠ ô̲̈

o + ◌͠ + ◌̂ + o + ◌̈ ➠ ô̲̈

Marks as Spacing Characters. By convention, combining marks may be exhibited in (apparent) isolation by applying them to U+0020 SPACE or to U+00A0 NO-BREAK SPACE.

This approach might be taken, for example, when referring to the diacritical mark itself as a mark, rather than using it in its normal way in text. The use of U+0020 SPACE versus U+00A0 NO-BREAK SPACE affects line-breaking behavior.

In charts and illustrations in this standard, the combining nature of these marks is illustrated by applying them to U+25CC DOTTED CIRCLE, as shown in the examples throughout this standard.

The Unicode Standard separately encodes clones of many common European diacritical marks as spacing characters. These related characters are cross-referenced in the character names list.

Encoding Principles. Because nonspacing marks have such a wide variety of applications, the characters in this block may have multiple semantic values. For example, U+0308 = *diaeresis* = *umlaut* = *double derivative*. There are also cases of several different Unicode characters for equivalent semantic values; variants of CEDILLA include at least U+0312 COMBINING TURNED COMMA ABOVE, U+0326 COMBINING COMMA BELOW, and U+0327 COMBINING CEDILLA. (For more information about the difference between nonspacing marks and combining characters, see *Chapter 2, General Structure*.)

Glyphic Variation. When rendered in the context of a language or script, like ordinary letters, combining marks may be subjected to systematic stylistic variation. For example, when used in Polish, U+0301 COMBINING ACUTE ACCENT appears at a steeper angle than when it is used in French. When it is used for Greek (as *oxia*), it can appear nearly upright. U+030C COMBINING CARON is commonly rendered as an apostrophe when used with certain letterforms. U+0326 COMBINING COMMA BELOW is sometimes rendered as U+0312 COMBINING TURNED COMMA ABOVE on a lowercase “g” to avoid conflict with the decender. In many fonts, there is no clear distinction made between COMBINING COMMA BELOW and U+0327 COMBINING CEDILLA.

Combining accents above the base glyph are usually adjusted in height for use with uppercase versus lowercase forms. In the absence of specific font protocols, combining marks are often designed as if they were applied to typical base characters in the same font.

For more information, see *Section 5.14, Rendering Nonspacing Marks.*

Combining Marks for Symbols: U+20D0–U+20FF

Diacritical marks for symbols are generally applied to mathematical or technical symbols. They can be used to extend the range of the symbol set. For example, U+20D3 COMBINING SHORT VERTICAL LINE OVERLAY can be used to express negation. Its presentation may change in those circumstances, changing length or slant. That is, U+2261 IDENTICAL TO followed by U+20D3 is equivalent to U+2262 NOT IDENTICAL TO. In this case, there is a precomposed form for the negated symbol. However, this statement does not always hold true, and U+20D3 can be used with other symbols to form the negation. For example, U+2258 CORRESPONDS TO followed by U+20D3 can be used to express *does not correspond to*, without requiring that a precomposed form be part of the Unicode Standard.

Other nonspacing characters can also be used in mathematical expressions, of course. For example, a U+0304 COMBINING MACRON is commonly used in propositional logic to indicate logical negation.

Enclosing Diacritics. These nonspacing characters are supplied for compatibility with existing standards, allowing individual base characters to be enclosed in several ways. For example, U+2460 CIRCLED DIGIT ONE ① can be expressed as U+0030 DIGIT ONE “1” + U+20DD COMBINING ENCLOSING CIRCLE ⃝. As with other combining characters, this one can also be applied productively; *circled letter alef* can be produced by the sequence:

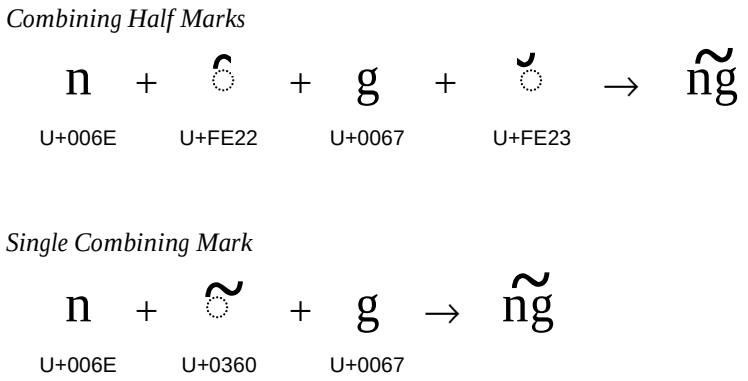
U+05D0 HEBREW LETTER ALEF א + U+20DD COMBINING ENCLOSING CIRCLE ◌⃝. The combining enclosing diacritics cannot be used to enclose a sequence of base characters in plain text. For example, there is no way to represent U+246A CIRCLED NUMBER ELEVEN with the ENCLOSING CIRCLE, as there is no single character NUMBER ELEVEN.

Combining Half Marks: U+FE20–U+FE2F

This block consists of a number of presentation form (glyph) encodings that may be used to visually encode certain combining marks that apply to multiple base letterforms. These characters are intended to facilitate the support of such combining marks in legacy implementations.

Unlike the other compatibility characters, these characters do not correspond to a single nominal character or a sequence of nominal characters; rather, a discontiguous sequence of these combining half marks corresponds to a single combining mark, as depicted in *Figure 7-7*. The preferred forms are the double diacritics, U+0360 and U+0361.

Figure 7-7. Combining Half Marks



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

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

ISBN 0-201-61633-5

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

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

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

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

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

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

First printing, January 2000.