

# The Unicode Standard

## Version 8.0 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://www.unicode.org/versions/latest/>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

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

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

Copyright © 1991–2015 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0

Includes bibliographical references and index.

ISBN 978-1-936213-10-8 (<http://www.unicode.org/versions/Unicode8.0.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2015

ISBN 978-1-936213-10-8

Published in Mountain View, CA

August 2015

## Chapter 7

# *Europe-I*

## *Modern and Liturgical Scripts*

Modern European alphabetic scripts are derived from or influenced by the Greek script, which itself was 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 alphabetic scripts and additional characters described in this chapter are:

<i>Latin</i>	<i>Cyrillic</i>	<i>Georgian</i>
<i>Greek</i>	<i>Glagolitic</i>	<i>Modifier letters</i>
<i>Coptic</i>	<i>Armenian</i>	<i>Combining marks</i>

Some scripts whose geographic area of primary usage is outside Europe are included in this chapter because of their relationship with Greek script. Coptic is used primarily by the Coptic church in Egypt and elsewhere; Armenian and Georgian are primarily associated with countries in the Caucasus (which is often not included as part of Europe), although Armenian in particular is used by a large diaspora.

These scripts are all 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. Some of these modification marks have evolved into small free-standing signs that can be treated as characters in their own right.

The Latin script is used to write or transliterate texts in a wide variety of languages. The International Phonetic Alphabet (IPA) is an extension of the Latin alphabet, enabling it to represent the phonetics of all languages. Other Latin phonetic extensions are used for the Uralic Phonetic Alphabet and the Teuthonista transcription system.

The Latin alphabet is derived from the alphabet used by the Etruscans, who had adopted a Western variant of the classical Greek alphabet (*Section 8.5, Old Italic*). 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. Like Latin, Cyrillic is used to write or transliterate texts in many languages. The Georgian and Armenian scripts were devised in the fifth century and are influenced by Greek. Modern Georgian does not have separate uppercase and lowercase forms.

The Coptic script was the last stage in the development of Egyptian writing. It represented the adaptation of the Greek alphabet to writing Egyptian, with the retention of forms from Demotic for sounds not adequately represented by Greek letters. Although primarily used in Egypt from the fourth to the tenth century, it is described in this chapter because of its close relationship to the Greek script.

Glagolitic is an early Slavic script related in some ways to both the Greek and the Cyrillic scripts. It was widely used in the Balkans but gradually died out, surviving the longest in Croatia. Like Coptic, however, it still has some modern use in liturgical contexts.

This chapter also describes modifier letters and combining marks used with the Latin script and other scripts.

The block descriptions for other archaic European alphabetic scripts, such as Gothic, Ogham, Old Italic, and Runic, can be found in *Chapter 8, Europe-II*.

## 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 Standard Annex #14, “Unicode Line Breaking Algorithm.”) Latin letters come in uppercase and lowercase pairs.

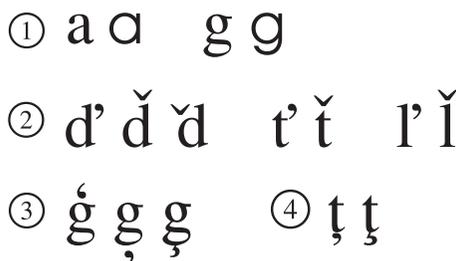
**Languages.** Some indication of language or other usage is given for many characters within the names lists accompanying the character charts.

**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. 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 the subsection “Combining Diacritical Marks” in Section 7.9, *Combining Marks*, and also Section 2.11, *Combining Characters*.

**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 Latin examples to illustrate this point are provided in Figure 7-1 and discussed in the text that follows.

Figure 7-1. Alternative Glyphs in Latin



Common typographical variations of basic Latin letters include the open- and closed-loop forms of the lowercase letters “a” and “g”, as shown in the first example in *Figure 7-1*. In ordinary Latin text, such distinctions are merely glyphic alternates for the same characters; however, phonetic transcription systems, such as IPA and Pinyin, often make systematic distinctions between these forms.

**Variations in Diacritical Marks.** The shape and placement of diacritical marks can be subject to considerable variation that might surprise a reader unfamiliar with such distinctions. For example, when Czech is typeset, U+010F LATIN SMALL LETTER D WITH CARON and U+0165 LATIN SMALL LETTER T WITH CARON are often rendered by glyphs with an apostrophe instead of with a caron, commonly known as a háček. See the second example in *Figure 7-1*. In Slovak, this use also applies to U+013E LATIN SMALL LETTER L WITH CARON and U+013D LATIN CAPITAL 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, glyphs with háčeks are preferred.

Characters with cedillas, commas or ogoneks below often are subject to variable typographical usage, depending on the availability and quality of fonts used, the technology, the era and the geographic area. Various hooks, cedillas, commas, and squiggles may be substituted for the nominal forms of these diacritics below, and even the directions of the hooks may be reversed.

The character U+0327 COMBINING CEDILLA can be displayed by a wide variety of forms, including cedillas and commas below. This variability also occurs for the precomposed characters whose decomposition includes U+0327. For text in some languages, a specific form is typically preferred. In particular, Latvian and Romanian prefer a comma below, while a cedilla is preferred in Turkish and Marshallese. These language-specific preferences are discussed in more detail in the text that follows.

Also, as a result of legacy encodings and practices, and the mapping of those legacy encodings to Unicode, some particular shapes for U+0327 COMBINING CEDILLA are preferred in the absence of language or locale context. A rendering as cedilla is preferred for the letters listed in the first column, while rendering as comma below is preferred for those listed in the second column of *Table 7-1*.

**Table 7-1.** Preferred Rendering of Cedilla versus Comma Below

Cedilla	Comma Below
c, e, h, s	d, g, k, l, n, r, t

**Latvian Cedilla.** There is specific variation involved in the placement and shapes of cedillas on Latvian characters. This is illustrated by the Latvian letter U+0123 LATIN SMALL LETTER G WITH CEDILLA, as shown in example 3 in *Figure 7-1*. 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 a 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 at the bottom of the base form.

**Cedilla and Comma Below in Turkish and Romanian.** The Latin letters *s* and *t* with comma below or with cedilla diacritics pose particular interpretation issues for Turkish and Romanian data, both in legacy character sets and in the Unicode Standard. Legacy character sets generally include a single form for these characters. While the formal interpretation of legacy character sets is that they contain only one of the forms, in practice this single character has been used to represent any of the forms. For example, 0xBA in ISO 8859-2 is formally defined as a *lowercase s with cedilla*, but has been used to represent a *lowercase s with comma below* for Romanian.

The Unicode Standard provides unambiguous representations for all of the forms, for example, U+0219 Ș LATIN SMALL LETTER S WITH COMMA BELOW versus U+015F ș LATIN SMALL LETTER S WITH CEDILLA. In modern usage, the preferred representation of Romanian text is with U+0219 Ș LATIN SMALL LETTER S WITH COMMA BELOW, while Turkish data is represented with U+015F ș LATIN SMALL LETTER S WITH CEDILLA.

However, due to the prevalence of legacy implementations, a large amount of Romanian data will contain U+015F ș LATIN SMALL LETTER S WITH CEDILLA or the corresponding code point 0xBA in ISO 8859-2. When converting data represented using ISO 8859-2, 0xBA should be mapped to the appropriate form. When processing Romanian Unicode data, implementations should treat U+0219 Ș LATIN SMALL LETTER S WITH COMMA BELOW and U+015F ș LATIN SMALL LETTER S WITH CEDILLA as equivalent.

**Exceptional Case Pairs.** The characters U+0130 İ LATIN CAPITAL LETTER I WITH DOT ABOVE and U+0131 İ LATIN SMALL LETTER DOTLESS I (used primarily in Turkish) are assumed to take ASCII “i” and “I”, respectively, as their case alternates. This mapping makes the corresponding reverse mapping language-specific; mapping in both directions requires special attention from the implementer (see Section 5.18, *Case Mappings*).

**Diacritics on *i* and *j*.** A dotted (normal) *i* or *j* followed by a nonspacing mark above loses the dot in rendering. Thus, in the word *naïve*, the *i* could be spelled with *i* + *diaeresis*. 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,  $i + \grave{\ } \neq i + \grave{\grave{\}}$ ). The same pattern is used for *j*. *Dotless-j* is used in the *Landsmålsalfabet*, where it does not have a case pair.

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

All characters that use their dot in this manner have the `Soft_Dotted` property in Unicode.

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

$$\begin{array}{l} \dot{\mathbf{i}} + \ddot{\circ} \rightarrow \ddot{\mathbf{i}} \quad \dot{\mathbf{i}} + \dot{\circ} + \acute{\circ} \rightarrow \acute{\mathbf{i}} \\ \mathbf{j} + \bar{\circ} \rightarrow \bar{\mathbf{j}} \quad \dot{\mathbf{i}} + \acute{\circ} + \dot{\circ} \rightarrow \acute{\mathbf{i}} \end{array}$$

**Vietnamese.** In the modern Vietnamese alphabet, there are 12 vowel letters and 5 tone marks (see Figure 7-3). Normalization Form C represents the combination of vowel letter and tone mark as a single unit—for example, U+1EA8  $\grave{A}$  LATIN CAPITAL LETTER A WITH CIRCUMFLEX AND HOOK ABOVE. Normalization Form D decomposes this combination into the combining character sequence, such as <U+0041, U+0302, U+0309>. Some widely used implementations prefer storing the vowel letter and the tone mark separately.

Figure 7-3. Vietnamese Letters and Tone Marks

a  $\grave{a}$   $\hat{a}$  e  $\hat{e}$  i o  $\hat{o}$   $\sigma$  u  $\acute{u}$  y  
 $\grave{\circ}$   $\acute{\circ}$   $\tilde{\circ}$   $\acute{\circ}$   $\grave{\circ}$

The Vietnamese vowels and other letters are found in the Basic Latin, Latin-1 Supplement, and Latin Extended-A blocks. Additional precomposed vowels and tone marks are found in the Latin Extended Additional block.

The characters U+0300 COMBINING GRAVE ACCENT, U+0309 COMBINING HOOK ABOVE, U+0303 COMBINING TILDE, U+0301 COMBINING ACUTE ACCENT, 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 have canonical equivalences to U+0300 COMBINING GRAVE ACCENT and U+0301 COMBINING ACUTE ACCENT, respectively; they are not recommended for use in representing Vietnamese tones, despite the presence of *tone mark* in their character names.

**Standards.** Unicode follows ISO/IEC 8859-1 in the layout of Latin letters up to U+00FF. ISO/IEC 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/IEC 8859-1 contains Latin letters, punctuation signs, and mathematical symbols. These additional characters are widely used with scripts other than Latin. The descriptions of these characters are found in *Chapter 6, Writing Systems and Punctuation*, and *Chapter 22, Symbols*.

The Latin Extended-A block includes characters contained in ISO/IEC 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/IEC 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. The Latin Extended-A block also covers additional characters from ISO/IEC 6937.

The Latin Extended-B 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/IEC 8859 Part 10. *Latin alphabet No. 6*.

The characters in the IPA block are taken from the 1989 revision of the International Phonetic Alphabet, published by the International Phonetic Association. Extensions from later IPA sources have also been added.

**Related Characters.** For other Latin-derived characters, see Letterlike Symbols (U+2100..U+214F), Currency Symbols (U+20A0..U+20CF), Number Forms (U+2150..U+218F), Enclosed Alphanumerics (U+2460..U+24FF), CJK Compatibility (U+3300..U+33FF), Fullwidth Forms (U+FF21..U+FF5A), and Mathematical Alphanumeric Symbols (U+1D400..U+1D7FF).

### ***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 not used to write a language. For example, Italian does not use “j” or “w”.

### ***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 the major languages of Europe listed in the next paragraph.

**Languages.** The languages supported by the Latin-1 supplement include Catalan, 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.

### ***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 2.11, Combining Characters*, and *Section 7.9, Combining Marks*.

**Compatibility Digraphs.** The Latin Extended-A block contains five compatibility digraphs, encoded for compatibility with ISO/IEC 6937:1984. Two of these characters, U+0140 LATIN SMALL LETTER L WITH MIDDLE DOT and its uppercase version, were originally encoded in ISO/IEC 6937 for support of Catalan. In current conventions, the representation of this digraphic sequence in Catalan simply uses a sequence of an ordinary “l” and U+00B7 MIDDLE DOT.

Another pair of characters, U+0133 LATIN SMALL LIGATURE IJ and its uppercase version, was provided to support the digraph “ij” in Dutch, often termed a “ligature” in discussions of Dutch orthography. When adding intercharacter spacing for line justification, the “ij” is kept as a unit, and the space between the *i* and *j* does not increase. In titlecasing, both the *i* and the *j* are uppercased, as in the word “IJsselmeer.” Using a single code point might simplify software support for such features; however, because a vast amount of Dutch data is encoded without this digraph character, under most circumstances one will encounter an <i, j> sequence.

Finally, U+0149 LATIN SMALL LETTER N PRECEDED BY APOSTROPHE was encoded for use in Afrikaans. The character is deprecated, and its use is strongly discouraged. In nearly all cases it is better represented by a sequence of an apostrophe followed by “n”.

**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, Croatian, Czech, Esperanto, Estonian, French, Frisian, Greenlandic, Hungarian, Latin, Latvian, Lithuanian, Maltese, Polish, Provençal, Rhaeto-Romanic, Romanian, Romany, Sámi, Slovak, Slovenian, Sorbian, Turkish, Welsh, and many others.

### **Latin Extended-B: U+0180–U+024F**

The Latin Extended-B block contains letterforms 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).

**Arrangement.** The characters are arranged in a nominal alphabetical order, followed by a small collection of Latinate forms. Uppercase 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, 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.

**Pinyin Diacritic–Vowel Combinations.** The Chinese standard GB 2312, the Japanese standard JIS X 0212, and some other standards include codes for Pinyin, which is used for Latin transcription of Mandarin Chinese. Most of the letters used in Pinyin romanization 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 forms are part of some other grouping. Many of these characters came from the International Phonetic Alphabet; they acquired 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 code point, 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: 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 *Section 4.2, Case*.

**Caseless Letters.** A number of letters used with the Latin script are caseless—for example, the caseless *glottal stop* at U+0294 and U+01BB LATIN LETTER TWO WITH STROKE, and the various letters denoting click sounds. Caseless letters retain their shape when uppercased. When titlecasing words, they may also act transparently; that is, if they occur in the leading position, the next following cased letter may be uppercased instead.

Over the last several centuries, the trend in typographical development for the Latin script has tended to favor the eventual introduction of case pairs. See the following discussion of the glottal stop. The Unicode Standard may encode additional uppercase characters in such instances. However, for reasons of stability, the standard will never add a new lowercase form for an existing uppercase character. See also “Caseless Matching” in *Section 5.18, Case Mappings*.

**Glottal Stop.** There are two patterns of usage for the *glottal stop* in the Unicode Standard. U+0294 ʔ LATIN LETTER GLOTTAL STOP is a caseless letter used in IPA. It is also widely seen in language orthographies based on IPA or Americanist phonetic usage, in those instances where no casing is apparent for *glottal stop*. Such orthographies may avoid casing for *glottal*

*stop* to the extent that when titlecasing strings, a word with an initial *glottal stop* may have its second letter uppercased instead of the first letter.

In a small number of orthographies for languages of northwestern Canada, and in particular, for Chipewyan, Dogrib, and Slavey, case pairs have been introduced for *glottal stop*. For these orthographies, the cased *glottal stop* characters should be used: U+0241 ʔ LATIN CAPITAL LETTER GLOTTAL STOP and U+0242 ʔ LATIN SMALL LETTER GLOTTAL STOP.

The glyphs for the *glottal stop* are somewhat variable and overlap to a certain extent. The glyph shown in the code charts for U+0294 ʔ LATIN LETTER GLOTTAL STOP is a cap-height form as specified in IPA, but the same character is often shown with a glyph that resembles the top half of a question mark and that may or may not be cap height. U+0241 ʔ LATIN CAPITAL LETTER GLOTTAL STOP, while shown with a larger glyph in the code charts, often appears identical to U+0294. U+0242 ʔ LATIN SMALL LETTER GLOTTAL STOP is a small form of U+0241.

Various small, raised hook- or comma-shaped characters are often substituted for a *glottal stop*—for instance, U+02BC ’ MODIFIER LETTER APOSTROPHE, U+02BB ‘ MODIFIER LETTER TURNED COMMA, U+02C0 ʔ MODIFIER LETTER GLOTTAL STOP, or U+02BE ’ MODIFIER LETTER RIGHT HALF RING. U+02BB, in particular, is used in Hawaiian orthography as the *’okina*.

### **IPA Extensions: U+0250–U+02AF**

The IPA Extensions block contains primarily the unique symbols of the International Phonetic Alphabet, 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 (1999) 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 the fine degrees of phonetic differentiation required for precise recording of different languages.

**Standards.** 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 œ LATIN SMALL LIGATURE OE, 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. 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 characters are unified as much as possible with other letters, albeit not with nonletter symbols such as U+222B ∫ INTEGRAL. The IPA characters 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 characters are found outside the IPA Extensions block. IPA characters 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 “ı” versus U+026A LATIN LETTER SMALL CAPITAL I “ı̇.” The Unicode Standard provides separate encodings for the two 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 letterforms. Examples include a typographic variant of the Greek letter *phi* ϕ and the borrowed letter Greek *iota* ι, 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 have possible semantic values that make them not simply rendering forms. 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.

**Arrangement.** 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.

### ***Phonetic Extensions: U+1D00–U+1DBF***

Most of the characters in the first of the two adjacent blocks comprising the phonetic extensions are used in the Uralic Phonetic Alphabet (UPA; also called Finno-Ugric Transcription, FUT), a highly specialized system that has been used by Uralicists globally for more than 100 years. Originally, it was chiefly used in Finland, Hungary, Estonia, Germany, Norway, Sweden, and Russia, but it is now known and used worldwide, including in North America and Japan. Uralic linguistic description, which treats the phonetics, phonology, and etymology of Uralic languages, is also used by other branches of linguistics, such as Indo-European, Turkic, and Altaic studies, as well as by other sciences, such as archaeology.

A very large body of descriptive texts, grammars, dictionaries, and chrestomathies exists, and continues to be produced, using this system.

The UPA makes use of approximately 258 characters, some of which are encoded in the Phonetic Extensions block; others are encoded in the other Latin blocks and in the Greek and Cyrillic blocks. The UPA takes full advantage of combining characters. It is not uncommon to find a base letter with three diacritics above and two below.

***Typographic Features of the UPA.*** Small capitalization in the UPA means voicelessness of a normally voiced sound. Small capitalization is also used to indicate certain either voiceless or half-voiced consonants. Superscripting indicates very short schwa vowels or transition vowels, or in general very short sounds. Subscripting indicates co-articulation caused by the preceding or following sound. Rotation (turned letters) indicates reduction; sideways (that is, 90 degrees counterclockwise) rotation is used where turning (180 degrees) might result in an ambiguous representation.

UPA phonetic material is generally represented with italic glyphs, so as to separate it from the surrounding text.

***Other Phonetic Extensions.*** The remaining characters in the phonetics extension range U+1D6C..U+1DBF are derived from a wide variety of sources, including many technical orthographies developed by SIL linguists, as well as older historic sources.

All attested phonetic characters showing struckthrough tildes, struckthrough bars, and retroflex or palatal hooks attached to the basic letter have been separately encoded here. Although separate combining marks exist in the Unicode Standard for overstruck diacritics and attached retroflex or palatal hooks, earlier encoded IPA letters such as U+0268 LATIN SMALL LETTER I WITH STROKE and U+026D LATIN SMALL LETTER L WITH RETROFLEX HOOK have never been given decomposition mappings in the standard. For consistency, all newly encoded characters are handled analogously to the existing, more common characters of this type and are not given decomposition mappings. Because these characters do not have decompositions, they require special handling in some circumstances. See the discussion of single-script confusables in Unicode Technical Standard #39, “Unicode Security Mechanisms.”

The Phonetic Extensions Supplement block also contains 37 superscript modifier letters. These complement the much more commonly used superscript modifier letters found in the Spacing Modifier Letters block.

U+1D77 LATIN SMALL LETTER TURNED G and U+1D78 MODIFIER LETTER CYRILLIC EN are used in Caucasian linguistics. U+1D79 LATIN SMALL LETTER INSULAR G is used in older Irish phonetic notation. It is to be distinguished from a Gaelic style glyph for U+0067 LATIN SMALL LETTER G.

**Digraph for th.** U+1D7A LATIN SMALL LETTER TH WITH STRIKETHROUGH is a digraphic notation commonly found in some English-language dictionaries, representing the voiceless (inter)dental fricative, as in *thin*. While this character is clearly a digraph, the obligatory strikethrough across two letters distinguishes it from a “th” digraph per se, and there is no mechanism involving combining marks that can easily be used to represent it. A common alternative glyphic form for U+1D7A uses a horizontal bar to strike through the two letters, instead of a diagonal stroke.

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

The characters in this block are mostly precomposed combinations of Latin letters with one or more general diacritical marks. With the exception of U+1E9A LATIN SMALL LETTER A WITH RIGHT HALF RING, each of the precomposed characters contained in this block is a canonical decomposable character and may alternatively be represented with a base letter followed by one or more general diacritical mark characters found in the Combining Diacritical Marks block.

The non-decomposable characters in this block, particularly in the range U+1EFA..U+1EFF, are mostly specialized letters used in Latin medieval manuscript traditions. These characters complement the larger set of medieval manuscript characters encoded in the Latin Extended-D block.

**Capital Sharp S.** U+1E9E LATIN CAPITAL LETTER SHARP S is for use in German. It is limited to specialized circumstances, such as uppcased strings in shop signage and book titles. The casing behavior of this character is unusual, as the recommended uppcase form for most casing operations on U+00DF ß LATIN SMALL LETTER SHARP S continues to be “SS”. See the discussion of tailored casing in *Section 3.13, Default Case Algorithms*, for more about the casing of this character.

**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 diacritical mark that denotes the phonemic tone that applies to the syllable.

### **Latin Extended-C: U+2C60–U+2C7F**

This small block of additional Latin characters contains orthographic Latin additions for minority languages, a few historic Latin letters, and further extensions for phonetic notations, particularly UPA.

**Uighur.** The Latin orthography for the Uighur language was influenced by widespread conventions for extension of the Cyrillic script for representing Central Asian languages. In particular, a number of Latin characters were extended with a Cyrillic-style descender diacritic to create new letters for use with Uighur.

**Claudian Letters.** The Roman emperor Claudius invented three additional letters for use with the Latin script. Those letters saw limited usage during his reign, but were abandoned soon afterward. The *half h* letter is encoded in this block. The other two letters are encoded in other blocks: U+2132 TURNED CAPITAL F and U+2183 ROMAN NUMERAL REVERSED ONE HUNDRED (unified with the Claudian letter *reversed c*). Claudian letters in inscriptions are uppercase only, but may be transcribed by scholars in lowercase.

### **Latin Extended-D: U+A720–U+A7FF**

This block contains a variety of historic letters for the Latin script and other uncommon phonetic and orthographic extensions to the script.

**Egyptological Transliteration.** The letters in the range U+A722..U+A725 are specialized letters used for the Latin transliteration of *alef* and *ain* in ancient Egyptian texts. Their forms are related to the modifier letter half rings (U+02BE..U+02BF) which are sometimes used in Latin transliteration of Arabic.

**Historic Mayan Letters.** The letters in the range U+A726..U+A72F are obsolete historic letters seen only in a few early Spanish manuscripts of Mayan languages. They are not used in modern Mayan orthographies.

**European Medievalist Letters.** The letters in the range U+A730..U+A778 occur in a variety of European medievalist manuscript traditions. None of these have any modern orthographic usage. A number of these letterforms constitute abbreviations, often for common Latin particles or suffixes.

**Insular and Celticist Letters.** The Insular manuscript tradition was current in Anglo-Saxon England and Gaelic Ireland throughout the early Middle Ages. The letters *d*, *f*, *g*, *r*, *s*, and *t* had unique shapes in that tradition, different from the Carolingian letters used in the modern Latin script. Although these letters can be considered variant forms of ordinary Latin letters, they are separately encoded because of their use by antiquarian Edward Lhuyd in his 1707 work *Archæologia Britannica*, which described the Late Cornish language in a phonetic alphabet using these Insular characters. Other specialists may make use of these letters contrastively in Old English or Irish manuscript contexts or in secondary material discussing such manuscripts.

**Orthographic Letter Additions.** The letters and modifier letters in the range U+A788..U+A78C occur in modern orthographies of a few small languages of Africa, Mexico, and New Guinea. Several of these characters were based on punctuation characters originally, so their shapes are confusingly similar to ordinary ASCII punctuation. Because of this potential confusion, their use is not generally recommended outside the specific context of the few orthographies already incorporating them.

**Sinological Dot.** U+A78F LATIN LETTER SINOLOGICAL DOT is a middle dot used in the sinological tradition to represent a glottal stop. This convention of representing a glottal stop with a middle dot was introduced by Bernhard Karlgren in the early 20th century for Middle Chinese reconstructions, and was adopted by other influential sinologists and Tangutologists. This dot is also used in Latin transliterations of Phags-pa text.

The representative glyph for U+A78F is larger than a typical middle dot used as punctuation, to avoid visual confusion with U+00B7 MIDDLE DOT. Use of the sinological dot should be limited to the appropriate scholarly contexts; it is not intended as a letter substitution for other functions of U+00B7 MIDDLE DOT.

**Latvian Letters.** The letters with strokes in the range U+A7A0..U+A7A9 are for use in the pre-1921 orthography of Latvian. During the 19th century and early 20th century, Latvian was usually typeset in a Fraktur typeface. Because Fraktur typefaces do not work well with detached diacritical marks, the extra letters required for Latvian were formed instead with overstruck bars. The new orthography introduced in 1921 replaced these letters with the current Latvian letters with cedilla diacritics. The *barred s* letters were also used in Fraktur representation of Lower Sorbian until about 1950.

**Ancient Roman Epigraphic Letters.** There are a small number of additional Latin epigraphic letters known from Ancient Roman inscriptions. These letters only occurred as monumental capitals in the inscriptions, and were not part of the regular Latin alphabet which later developed case distinctions.

### **Latin Extended-E: U+AB30–U+AB6F**

This block contains a number of Latin letters and modifier letters for phonetic transcription systems. The majority of these are letters specifically associated with the Böhmer-Ascoli transcription system, more generally known as “Teuthonista.” The Teuthonista system was extensively used in the 20th century to transcribe Germanic dialects. Teuthonista or closely related systems were also used in Switzerland and Italy to transcribe Romance dialects. For related characters, see the Combining Diacritical Marks Extended block, which contains a number of specialized combining diacritics for use in Teuthonista.

The Latin Extended-E block also contains a few rarely used letters from other transcription systems.

### **Latin Ligatures: U+FB00–U+FB06**

This range in the Alphabetic Presentation Forms block (U+FB00..U+FB4F) contains several common Latin ligatures, which occur in legacy encodings. Whether to use a Latin ligature is a matter of typographical style as well as a result of 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.

Some format controls in the Unicode Standard can affect the formation of ligatures. See “Controlling Ligatures” in *Section 23.2, Layout Controls*.

## 7.2 Greek

### **Greek: U+0370–U+03FF**

The Greek script is used for writing the Greek language. The Greek script had a strong influence on the development of the Latin, Cyrillic, and Coptic scripts.

The Greek script is written in linear sequence from left to right with the frequent use of nonspacing marks. There are two styles of such use: monotonic, which uses a single mark called *tonos*, and polytonic, which uses multiple marks. Greek letters come in uppercase and lowercase pairs. Spaces are used to separate words and provide the primary line breaking opportunities. Archaic Greek texts do not use spaces.

**Standards.** The Unicode encoding of Greek is based on ISO/IEC 8859-7, which is equivalent to the Greek national standard ELOT 928, designed for monotonic Greek. 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) and occasionally for modern Greek, 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 7-2*).

**Table 7-2. Nonspacing Marks Used with Greek**

Code	Name	Alternative Names
U+0300	COMBINING GRAVE ACCENT	<i>varia</i>
U+0301	COMBINING ACUTE ACCENT	<i>tonos, oxia</i>
U+0304	COMBINING MACRON	
U+0306	COMBINING BREVE	
U+0308	COMBINING DIAERESIS	<i>dialytika</i>
U+0313	COMBINING COMMA ABOVE	<i>psili, smooth breathing mark</i>
U+0314	COMBINING REVERSED COMMA ABOVE	<i>dasia, rough breathing mark</i>
U+0342	COMBINING GREEK PERISPOMENI	<i>circumflex, tilde, inverted breve</i>
U+0343	COMBINING GREEK KORONIS	<i>comma above</i>
U+0345	COMBINING GREEK YPOGRAMMENI	<i>iota subscript</i>

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. The character U+0344 COMBINING GREEK DIALYTIKA TONOS should not be used. The combination of *dialytika* plus *tonos* is instead represented by the sequence <U+0308 COMBINING DIAERESIS, U+0301 COMBINING ACUTE ACCENT>.

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 2.11, 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 a circumflex  $\hat{\text{̂}}$ , an inverted breve  $\text{̆}$ , a tilde  $\text{̃}$ , or occasionally a macron  $\text{̄}$ . 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. Greek uses guillemets for quotation marks; for Ancient Greek, the quotations tend to follow local publishing practice. Because of the possibility of confusion between smooth breathing marks and curly single quotation marks, the latter are best avoided where possible. When either breathing mark is followed by an acute or grave accent, the pair is rendered side-by-side rather than vertically stacked.

Accents are typically written above their base letter in an all-lowercase or all-uppercase word; they may also be omitted from an all-uppercase word. However, in a titlecase word, accents applied to the first letter are commonly written to the left of that letter. This is a matter of presentation only—the internal representation is still the base letter followed by the combining marks. It is *not* the stand-alone version of the accents, which occur before the base letter in the text stream.

**Iota.** 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 *proseggrammeni* (also known as *iota adscript* in English). In completely uppercased words, the *iota subscript* should be replaced by a capital *iota* following the vowel. Precomposed characters that contain *iota subscript* or *iota adscript* also have special mappings. (See *Section 5.18, Case Mappings*.) 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, which may not be automatically derivable.

**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 to distinguish it from the *Latin Y*. A third form is also encoded as U+03D2 GREEK UPSILON WITH HOOK SYMBOL (see *Figure 7-4*). The precomposed characters U+03D3 GREEK UPSILON WITH ACUTE AND HOOK SYMBOL and U+03D4 GREEK UPSILON WITH DIAERESIS AND HOOK SYMBOL should not normally be needed, except where necessary for backward compatibility for legacy character sets.

**Figure 7-4.** Variations in Greek Capital Letter Upsilon

Variant forms of several other Greek letters are encoded as separate characters in this block. Often (but not always), they represent different forms taken on by the character when it appears in the final position of a word. Examples include U+03C2 GREEK SMALL LETTER FINAL SIGMA used in a final position and U+03D0 GREEK BETA SYMBOL, which is the form that U+03B2 GREEK SMALL LETTER BETA would take on in a medial or final position.

Of these variant letterforms, only *final sigma* should be used in encoding standard Greek text to indicate a final sigma. It is also encoded in ISO/IEC 8859-7 and ISO 5428 for this purpose. Because use of the final sigma is a matter of spelling convention, software should not automatically substitute a final form for a nominal form at the end of a word. However, when performing lowercasing, the final form needs to be generated based on the context. See *Section 3.13, Default Case Algorithms*.

In contrast, U+03D0 GREEK BETA SYMBOL, U+03D1 GREEK THETA SYMBOL, U+03D2 GREEK UPSILON WITH HOOK SYMBOL, U+03D5 GREEK PHI SYMBOL, U+03F0 GREEK KAPPA SYMBOL, U+03F1 GREEK RHO SYMBOL, U+03F4 GREEK CAPITAL THETA SYMBOL, U+03F5 GREEK LUNATE EPSILON SYMBOL, and U+03F6 GREEK REVERSED LUNATE EPSILON SYMBOL should be used only in mathematical formulas—never in Greek text. If positional or other shape differences are desired for these characters, they should be implemented by a font or rendering engine.

**Representative Glyphs for Greek Phi.** Starting with *The Unicode Standard, Version 3.0*, and the concurrent second edition of ISO/IEC 10646-1, the representative glyphs for U+03C6 ϕ GREEK SMALL LETTER PHI and U+03D5 φ GREEK PHI SYMBOL were swapped compared to earlier versions. In ordinary Greek text, the character U+03C6 is used exclusively, although this character has considerable glyphic variation, sometimes represented with a glyph more like the representative glyph shown for U+03C6 ϕ (the “loopy” form) and less often with a glyph more like the representative glyph shown for U+03D5 φ (the “straight” form).

For mathematical and technical use, the straight form of the small phi is an important symbol and needs to be consistently distinguishable from the loopy form. The straight-form phi glyph is used as the representative glyph for the symbol phi at U+03D5 to satisfy this distinction.

The representative glyphs were reversed in versions of the Unicode Standard prior to Unicode 3.0. This resulted in the problem that the character explicitly identified as the mathe-

mathematical symbol did not have the straight form of the character that is the preferred glyph for that use. Furthermore, it made it unnecessarily difficult for general-purpose fonts supporting ordinary Greek text to add support for Greek letters used as mathematical symbols. This resulted from the fact that many of those fonts already used the loopy-form glyph for U+03C6, as preferred for Greek body text; to support the phi symbol as well, they would have had to disrupt glyph choices already optimized for Greek text.

When mapping symbol sets or SGML entities to the Unicode Standard, it is important to make sure that codes or entities that require the straight form of the phi symbol be mapped to U+03D5 and not to U+03C6. Mapping to the latter should be reserved for codes or entities that represent the small phi as used in ordinary Greek text.

Fonts used primarily for Greek text may use either glyph form for U+03C6, but fonts that also intend to support technical use of the Greek letters should use the loopy form to ensure appropriate contrast with the straight form used for U+03D5.

**Greek Letters as Symbols.** The use of Greek letters for mathematical variables and operators is well established. Characters from the Greek block may be used for these symbols.

For compatibility purposes, a few Greek letters are separately encoded as symbols in other character blocks. Examples include U+00B5  $\mu$  MICRO SIGN in the Latin-1 Supplement character block and U+2126  $\Omega$  OHM SIGN in the Letterlike Symbols character block. The *ohm sign* is canonically equivalent to the *capital omega*, and normalization would remove any distinction. Its use is therefore discouraged in favor of *capital omega*. The same equivalence does not exist between *micro sign* and *mu*, and use of either character as a micro sign is common. For Greek text, only the *mu* should be used.

**Symbols Versus Numbers.** The characters *stigma*, *koppa*, and *sampi* are used only as numerals, whereas *archaic koppa* and *digamma* are used only as letters.

**Compatibility Punctuation.** Two specific modern Greek punctuation marks are encoded in the Greek and Coptic block: U+037E “;” GREEK QUESTION MARK and U+0387 “.” GREEK ANO TELEIA. The *Greek question mark* (or *erotimatiko*) has the shape of a semicolon, but functions as a question mark in the Greek script. The *ano teleia* has the shape of a middle dot, but functions as a semicolon in the Greek script.

These two compatibility punctuation characters have canonical equivalences to U+003B SEMICOLON and U+00B7 MIDDLE DOT, respectively; as a result, normalized Greek text will lose any distinctions between the Greek compatibility punctuation characters and the common punctuation marks. Furthermore, ISO/IEC 8859-7 and most vendor code pages for Greek simply make use of semicolon and middle dot for the punctuation in question. Therefore, use of U+037E and U+0387 is not necessary for interoperating with legacy Greek data, and their use is not generally encouraged for representation of Greek punctuation.

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

**Coptic-Unique Letters.** In the Unicode Standard prior to Version 4.1, the Coptic script was regarded primarily as a stylistic variant of the Greek alphabet. The letters unique to Coptic

were encoded in a separate range at the end of the Greek character block. Those characters were to be used together with the basic Greek characters to represent the complete Coptic alphabet. Coptic text was supposed to be rendered with a font using the Coptic style of depicting the characters it shared with the Greek alphabet. Texts that mixed Greek and Coptic languages using that encoding model could be rendered only by associating an appropriate font by language.

The Unicode Technical Committee and ISO/IEC JTC1/SC2 determined that Coptic is better handled as a separate script. Starting with Unicode 4.1, a new Coptic block added all the letters formerly unified with Greek characters as separate Coptic characters. (See *Section 7.3, Coptic*.) Implementations that supported Coptic under the previous encoding model may, therefore, need to be modified. Coptic fonts may need to continue to support the display of both the Coptic and corresponding Greek character with the same shape to facilitate their use with older documents.

**Related Characters.** For math symbols, see *Section 22.5, Mathematical Symbols*. 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 can be used for the representation of polytonic Greek texts without the use of combining marks. Because they do not cover all possible combinations in use, some combining character sequences may be required for a given text.

Each of the letters 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.

**Spacing Diacritics.** Sixteen additional spacing diacritical 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 nonspacing alternatives appear in *Table 7-3*. The spacing forms are meant for keyboards and pedagogical use and are not to be used in the representation of titlecase words. The compatibility decompositions of these spacing forms consist of the sequence U+0020 SPACE followed by the nonspacing form equivalents shown in *Table 7-3*.

Table 7-3. Greek Spacing and Nonspacing Pairs

Spacing Form	Nonspacing Form
1FBD GREEK KORONIS	0313 COMBINING COMMA ABOVE
037A GREEK YPOGEGRAMMENI	0345 COMBINING GREEK 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

### ***Ancient Greek Numbers: U+10140–U+1018F***

Ancient Greeks primarily used letters of the Greek alphabet to represent numbers. However, some extensions to this usage required quite a few nonalphabetic symbols or symbols derived from letters. Those symbols are encoded in the Ancient Greek Numbers block.

***Acrophonic Numerals.*** Greek acrophonic numerals are found primarily in ancient inscriptions from Attica and other Greek regions. *Acrophonic* means that the character used to represent each number is the initial letter of the word by which the number is called—for instance, H for “HECATON” = 100.

The Attic acrophonic system, named for the greater geographic area that includes the city of Athens, is the most common and well documented. The characters in the Ancient Greek Numbers block cover the Attic acrophonic numeral system as well as non-Attic characters that cannot be considered glyph variants of the Attic acrophonic repertoire. They are the standard symbols used to represent weight or cost, and they appear consistently in modern

editions and scholarly studies of Greek inscriptions. Uppercase Greek letters from the Greek block are also used for acrophonic numerals.

The Greek acrophonic number system is similar to the Roman one in that it does not use decimal position, does not require a placeholder for zero, and has special symbols for 5, 50, 500, and so on. The system is language specific because of the acrophonic principle. In some cases the same symbol represents different values in different geographic regions. The symbols are also differentiated by the unit of measurement—for example, talents versus staters.

**Other Numerical Symbols.** Other numerical symbols encoded in the range U+10175..U+1018A appear in a large number of ancient papyri. The standard symbols used for the representation of numbers, fractions, weights, and measures, they have consistently been used in modern editions of Greek papyri as well as various publications related to the study and interpretation of ancient documents. Several of these characters have considerable glyphic variation. Some of these glyph variants are similar in appearance to other characters.

**Symbol for Zero.** U+1018A GREEK ZERO SIGN occurs whenever a sexagesimal notation is used in historical astronomical texts to record degrees, minutes and seconds, or hours, minutes and seconds. The most common form of zero in the papyri is a small circle with a horizontal stroke above it, but many variations exist. These are taken to be scribal variations and are considered glyph variants.

## 7.3 Coptic

### **Coptic:** U+2C80–U+2CFF

The Coptic script is the final stage in the development of the Egyptian writing system. Coptic was subject to strong Greek influences because Greek was more identified with the Christian tradition, and the written demotic Egyptian no longer matched the spoken language. The Coptic script was based on the Greek uncial alphabets with several Coptic additional letters unique to Coptic. The Coptic language died out in the fourteenth century, but it is maintained as a liturgical language by Coptic Christians. Coptic is written from left to right in linear sequence; in modern use, spaces are used to separate words and provide the primary line breaking opportunities.

Prior to Version 4.1, the Unicode Standard treated Coptic as a stylistic variant of Greek. Seven letters unique to Coptic (14 characters with the case pairs) were encoded in the Greek and Coptic block. In addition to these 14 characters, Version 4.1 added a Coptic block containing the remaining characters needed for basic Coptic text processing. This block also includes standard logotypes used in Coptic text as well as characters for Old Coptic and Nubian.

**Development of the Coptic Script.** The best-known Coptic dialects are Sahidic and Bohairic. Coptic scholarship recognizes a number of other dialects that use additional characters. The repertoires of Sahidic and Bohairic reflect efforts to standardize the writing of Coptic, but attempts to write the Egyptian language with the Greek script preceded that standardization by several centuries. During the initial period of writing, a number of different solutions to the problem of representing non-Greek sounds were made, mostly by borrowing letters from Demotic writing. These early efforts are grouped by Copticians under the general heading of Old Coptic.

**Casing.** Coptic is considered a bicameral script. Historically, it was caseless, but it has acquired case through the typographic developments of the last centuries. Already in Old Coptic manuscripts, letters could be written larger, particularly at the beginning of paragraphs, although the capital letters tend to have the most distinctive shapes in the Bohairic tradition. To facilitate scholarly and other modern casing operations, Coptic has been encoded as a bicameral script, including uniquely Old Coptic characters.

**Font Styles.** Bohairic Coptic uses only a subset of the letters in the Coptic repertoire. It also uses a font style distinct from that for Sahidic. Prior to Version 5.0, the Coptic letters derived from Demotic, encoded in the range U+03E2..U+03EF in the Greek and Coptic block, were shown in the code charts in a Bohairic font style. Starting from Version 5.0, all Coptic letters in the standard, including those in the range U+03E2..U+03EF, are shown in the code charts in a Sahidic font style, instead.

**Characters for Cryptogrammic Use.** U+2CB7 COPTIC SMALL LETTER CRYPTOGRAMMIC EIE and U+2CBD COPTIC SMALL LETTER CRYPTOGRAMMIC NI are characters for cryptogrammic use. A common Coptic substitution alphabet that was used to encrypt texts had the disadvantageous feature whereby three of the letters (*eie*, *ni*, and *fi*) were substituted by

themselves. However, because *ie* and *ni* are two of the highest-frequency characters in Coptic, Copts felt that the encryption was not strong enough, so they replaced those letters with these cryptogrammic ones. Two additional cryptogrammic letters in less frequent use are also encoded: U+2CEC COPTIC SMALL LETTER CRYPTOGRAMMIC SHEI and U+2CEE COPTIC SMALL LETTER CRYPTOGRAMMIC GANGIA. Coptacists preserve these letter substitutions in modern editions of these encrypted texts and do not consider them to be glyph variants of the original letters.

U+2CC0 COPTIC CAPITAL LETTER SAMPI has a numeric value of 900 and corresponds to U+03E0 GREEK LETTER SAMPI. It is not found in abecedaria, but is used in cryptogrammic contexts as a letter.

**Crossed Shei.** U+2CC3 𐪓 COPTIC SMALL LETTER CROSSED SHEI is found in Dialect I of Old Coptic, where it represents a sound /ç/. It is found alongside U+03E3 𐪓 COPTIC SMALL LETTER SHEI, which represents /ʃ/. The diacritic is not productive.

**Supralineation.** In Coptic texts, a line is often drawn across the top of two or more characters in a row. There are two distinct conventions for this supralineation, each of which is represented by different sequences of combining marks.

The first of these is a convention for abbreviation, in which words are shortened by removal of certain letters. A line is then drawn across the tops of all of the remaining letters, extending from the beginning of the first to the end of the last letter of the abbreviated form. This convention is represented by following each character of the abbreviated form with U+0305 COMBINING OVERLINE. When rendered together, these combining overlines should connect into a continuous line.

The other convention is to distinguish the spelling of certain common words or to highlight proper names of divinities and heroes—a convention related to the use of cartouches in hieroglyphic Egyptian. In this case the supralineation extends from the *middle* of the first character in the sequence to the *middle* of the last character in the sequence. Instead of using U+0305 COMBINING OVERLINE for the entire sequence, one uses U+FE24 COMBINING MACRON LEFT HALF after the first character, U+FE25 COMBINING MACRON RIGHT HALF after the last character, and U+FE26 COMBINING CONJOINING MACRON after any intervening characters. This gives the effect of a line starting and ending in the middle of letters, rather than at their edges.

**Combining Diacritical Marks.** Bohairic text uses a mark called *jinkim* to represent syllabic consonants, which is indicated by either U+0307 COMBINING DOT ABOVE or U+0300 COMBINING GRAVE ACCENT. Other dialects, including Sahidic, use U+0304 COMBINING MACRON for the same purpose. A number of other generic diacritical marks are used with Coptic.

U+2CEF COPTIC COMBINING NI above is a script-specific combining mark, typically used at the end of a line to indicate a final *ni* after a vowel. In rendering, this mark typically hangs over the space to the right of its base character.

The characters U+2CF0 COPTIC COMBINING SPIRITUS ASPER and U+2CF1 COPTIC COMBINING SPIRITUS LENIS are analogues of the Greek breathing marks. They are used rarely in Coptic. When used, they typically occur over the letter U+2C8F COPTIC SMALL LETTER

HATE, sometimes to indicate that it is the borrowed Greek conjunction “or”, written with the cognate Greek letter *eta*.

**Punctuation.** Coptic texts use common punctuation, including *colon*, *full stop*, *semicolon* (functioning, as in Greek, as a question mark), and *middle dot*. Quotation marks are found in edited texts. In addition, Coptic-specific punctuation occurs: U+2CFE COPTIC FULL STOP and U+2CFF COPTIC MORPHOLOGICAL DIVIDER. Several other historic forms of punctuation are known only from Old Nubian texts.

**Numerical Use of Letters.** Numerals are indicated with letters of the alphabet, as in Greek. Sometimes the numerical use is indicated specifically by marking a line above, represented with U+0305 COMBINING OVERLINE. U+0375 GREEK LOWER NUMERAL SIGN or U+033F COMBINING DOUBLE OVERLINE can be used to indicate multiples of 1,000, as shown in *Figure 7-5*.

Figure 7-5. Coptic Numerals

Coptic	Value
ⲁ	1
ⲁ, or ⲁ̅	1,000
ⲁ, ⲰⲎⲐ	1,888

U+0374 GREEK NUMERAL SIGN is used to indicate fractions. For example, ρ indicates the fractional value 1/3. There is, however, a special symbol for 1/2: U+2CFD COPTIC FRACTION ONE HALF.

## 7.4 Cyrillic

The Cyrillic script is one of several scripts that were ultimately derived from the Greek script. The details of the history of that development and of the relationship between early forms of writing systems for Slavic languages has been lost. Cyrillic has traditionally been used for writing various Slavic languages, among which Russian is predominant. The earliest attestations of Cyrillic are for Old Church Slavonic manuscripts, dating to the 10th century CE. Old Church Slavonic is also commonly referred to as Old Church Slavic, and is abbreviated as OCS.

In the nineteenth and early twentieth centuries, Cyrillic was extended to write the non-Slavic minority languages of Russia and neighboring countries.

**Structure.** The Cyrillic script is written in linear sequence from left to right with the occasional use of nonspacing marks. Cyrillic letters have uppercase and lowercase pairs. Spaces are used to separate words and provide the primary line breaking opportunities.

**Historic Letterforms.** The historic form of the Cyrillic alphabet—most notably that seen in Old Church Slavonic manuscripts—is treated as a font style variation of modern Cyrillic. The historic forms of the letters are relatively close to their modern appearance, and some of the historic letters are still in modern use in languages other than Russian. For example, U+0406 “І” CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I is used in modern Ukrainian and Byelorussian, and is encoded amidst other modern Cyrillic extensions. Some of the historic letterforms were used in modern typefaces in Russian and Bulgarian. Prior to 1917, Russian made use of *yat*, *fita*, and *izhitsa*; prior to 1945, Bulgaria made use of these three as well as *big yus*.

**Glagolitic.** The particular early Slavic writing known as Glagolitic is treated as a distinct script from Cyrillic, rather than as a font style variation. The letterforms for Glagolitic, even though historically related, appear unrecognizably different from most modern Cyrillic letters. Glagolitic was also limited to a certain historic period; it did not grow to match the repertoire expansion of the Cyrillic script. See *Section 7.5, Glagolitic*.

### **Cyrillic: U+0400–U+04FF**

**Standards.** The Cyrillic block of the Unicode Standard is based on ISO/IEC 8859-5.

**Extended Cyrillic.** These letters are used in alphabets for Turkic languages such as Azerbaijani, Bashkir, Kazakh, and Tatar; for Caucasian languages such as Abkhasian, Avar, and Chechen; and for Uralic languages such as Mari, Khanty, and Kildin Sami. The orthographies of some of these languages have often been revised in the past; some of them have switched from Arabic to Latin to Cyrillic, and back again. Azerbaijani, for instance, is now officially using a Turkish-based Latin script.

**Abkhasian.** The Cyrillic orthography for Abkhasian has been updated fairly frequently over the course of the 20th and early 21st centuries. Some of these revisions involved changes in letterforms, often for the diacritic descenders used under extended Cyrillic letters for Abkhasian. The most recent such reform has been reflected in glyph changes for

Abkhaz-specific Cyrillic letters in the code charts. In particular, U+04BF CYRILLIC SMALL LETTER ABKHASIAN CHE WITH DESCENDER, is now shown with a straight descender diacritic. In code charts for Version 5.1 and earlier, that character was displayed with a representative glyph using an ogonek-type hook descender, more typical of historic orthographies for Abkhasian. The glyph for U+04A9 CYRILLIC SMALL LETTER ABKHASIAN HA was also updated.

Other changes for Abkhasian orthography represent actual respellings of text. Of particular note, the character added in Version 5.2, U+0525 CYRILLIC SMALL LETTER PE WITH DESCENDER, is intended as a replacement for U+04A7 CYRILLIC SMALL LETTER PE WITH MIDDLE HOOK, which was used in older orthographies.

***Palochka.*** U+04C0 “I” CYRILLIC LETTER PALOCHKA is used in Cyrillic orthographies for a number of Caucasian languages, such as Adyghe, Avar, Chechen, and Kabardian. The name *palochka* itself is based on the Russian word for “stick,” referring to the shape of the letter. The glyph for *palochka* is usually indistinguishable from an uppercase Latin “I” or U+0406 “I” CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I; however, in some serified fonts it may be displayed without serifs to make it more visually distinct.

In use, *palochka* typically modifies the reading of a preceding letter, indicating that it is an ejective. The *palochka* is generally caseless and should retain its form even in lowercased Cyrillic text. However, there is some evidence of distinctive lowercase forms; for those instances, U+04CF CYRILLIC SMALL LETTER PALOCHKA may be used.

### ***Cyrillic Supplement: U+0500–U+052F***

***Komi.*** The characters in the range U+0500..U+050F are found in ISO 10754; they were used in Komi Cyrillic orthography from 1919 to about 1940. These letters use glyphs that differ structurally from other characters in the Unicode Standard that represent similar sounds—namely, Serbian *ѣ* and *ѣ*, which are ligatures of the base letters *а* and *и* with a palatalizing soft sign *ь*. The Molodtsov orthography made use of a different kind of palatalization hook for Komi *ѣ*, *ѣ*, *ѣ*, *ѣ*, and so on.

***Kurdish Letters.*** Although the Kurdish language is almost always written in either the Arabic script or the Latin script, there also exists a Cyrillic orthography which saw some usage for Kurdish in the former Soviet Union. The Cyrillic letters *qa* and *we* in this block are encoded to enable the representation of Cyrillic Kurdish entirely in the Cyrillic script, without use of the similar Latin letters q and w, from which these Kurdish letters were ultimately derived.

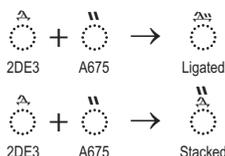
### ***Cyrillic Extended-A: U+2DE0–U+2DFF***

***Titlo Letters.*** This block contains a set of superscripted (written above), or *titlo*, letters, used in manuscript Old Church Slavonic texts and in modern Church Slavonic, usually to indicate abbreviations of words in the text. They can be found alone or in pairs that typically form digraphs or ligatures above one base character. These may occur with or without

the generic titlo character, U+0483 COMBINING CYRILLIC TITLO, or with U+0487 COMBINING CYRILLIC POKRYTIE, or with U+A66F COMBINING CYRILLIC VZMET.

When used in combination, two titlo letters normally form a composite combining letter, in which the components appear side-by-side or ligated, a behavior which deviates from the default vertical stacking of multiple combining characters. Occasionally, titlo letters can also be found vertically stacked in Old Church Slavonic texts, in this case exhibiting default stacking behavior. As there is no semantic distinction associated with the two presentations, both are handled at the font level, without requiring the use of format characters. The usual ligated form and the less common vertical stacking of titlo letters are contrasted in *Figure 7-6* for the sequence <U+2DE3 COMBINING CYRILLIC LETTER DE, U+A675 COMBINING CYRILLIC LETTER I>.

**Figure 7-6. Combination of Titlo Letters**



A wide variety of composite titlo letters can be encountered in Old Church Slavonic manuscripts, including such combinations as *ghe-o*, *de-ie*, *de-i*, *de-o*, *de-uk*, *el-i*, *em-i*, *es-te*, and many others. One of these combinations has been encoded atomically in Unicode as U+2DF5 COMBINING CYRILLIC LETTER ES-TE. However, the preferred representation of a composite titlo *es-te* is the sequence <U+2DED COMBINING CYRILLIC LETTER ES, U+2DEE COMBINING CYRILLIC LETTER TE>.

The glyphs in the code chart for the Cyrillic Extended-A block are based on the modern Cyrillic letters to which these titlo letters correspond, but in Old Church Slavonic manuscripts, the actual glyphs used are related to the older forms of Cyrillic letters.

### ***Cyrillic Extended-B: U+A640–U+A69F***

This block contains an extended set of historic Cyrillic characters used in Old Cyrillic manuscript materials, particularly Old Church Slavonic.

**Numeric Enclosing Signs.** The combining numeric signs in the range U+A670..U+A672 extend the series of such combining signs from the main Cyrillic block. These enclosing signs were used around letters to indicate high decimal multiples of the basic numeric values of the letters.

**Titlo Letters.** Several additional titlo letters based on manuscript sources are encoded in the ranges U+A674..U+A67B and U+A69E..U+A69F. For a description of titlo letters, see the subsection “Cyrillic Extended-A: U+2DE0–U+2DFF” earlier in this section.

**Old Abkhasian Letters.** The letters in the range U+A680..U+A697 are obsolete letters for an old orthography of the Abkhaz language. These characters are no longer in use, and the Abkhaz language is currently represented using various Cyrillic extensions in the main Cyrillic block.

## 7.5 Glagolitic

### **Glagolitic:** U+2C00–U+2C5F

Glagolitic, from the Slavic root *glagol*, meaning “word,” is an alphabet considered to have been devised by Saint Cyril in or around 862 CE for his translation of the Scriptures and liturgical books into Slavonic. The relatively few Glagolitic inscriptions and manuscripts that survive from this early period are of great philological importance. Glagolitic was eventually supplanted by the alphabet now known as Cyrillic.

Like Cyrillic, the Glagolitic script is written in linear sequence from left to right with no contextual modification of the letterforms. Spaces are used to separate words and provide the primary line breaking opportunities.

In parts of Croatia where a vernacular liturgy was used, Glagolitic continued in use until modern times: the last Glagolitic missal was printed in Rome in 1893 with a second edition in 1905. In these areas Glagolitic is still occasionally used as a decorative alphabet.

**Glyph Forms.** Glagolitic exists in two styles, known as round and square. Round Glagolitic is the original style and more geographically widespread, although surviving examples are less numerous. Square Glagolitic (and the cursive style derived from it) was used in Croatia from the thirteenth century. There are a few documents written in a style intermediate between the two. The letterforms used in the charts are round Glagolitic. Several of the letters have variant glyph forms, which are not encoded separately.

**Ordering.** The ordering of the Glagolitic alphabet is largely derived from that of the Greek alphabet, although nearly half the Glagolitic characters have no equivalent in Greek and not every Greek letter has its equivalent in Glagolitic.

**Punctuation and Diacritics.** Glagolitic texts use common punctuation, including *comma*, *full stop*, *semicolon* (functioning, as in Greek, as a question mark), and *middle dot*. In addition, several forms of multiple-dot, archaic punctuation occur, including U+2056 THREE DOT PUNCTUATION, U+2058 FOUR DOT PUNCTUATION, and U+2059 FIVE DOT PUNCTUATION. Quotation marks are found in edited texts. Glagolitic also used numerous diacritical marks, many of them shared in common with Cyrillic.

**Numerical Use of Letters.** Glagolitic letters have inherent numerical values. A letter may be rendered with a line above or a tilde above to indicate the numeric usage explicitly. Alternatively, U+00B7 MIDDLE DOT may be used, flanking a letter on both sides, to indicate numeric usage of the letter.

## 7.6 Armenian

### *Armenian: U+0530–U+058F*

The Armenian script is used primarily for writing the Armenian language. It is written from left to right. Armenian letters have uppercase and lowercase pairs. Spaces are used to separate words and provide the primary line breaking opportunities.

The Armenian script was devised about 406 CE by Mesrop Maštoc' to give Armenians access to Christian scriptural and liturgical texts, which were otherwise available only in Greek and Syriac. The script has been used to write Classical or *Grabar* Armenian, Middle Armenian, and both of the literary dialects of Modern Armenian: East and West Armenian.

**Orthography.** Mesrop's original alphabet contained 30 consonants and 6 vowels in the following ranges:

U+0531..U+0554 *Ա..Ք* *Ayb* to *K'ē*

U+0561..U+0584 *ա..բ* *ayb* to *k'ē*

Armenian spelling was consistent during the *Grabar* period, from the fifth to the tenth centuries CE; pronunciation began to change in the eleventh century. In the twelfth century, the letters *ō* and *fē* were added to the alphabet to represent the diphthong [aw] (previously written *աւ* *aw*) and the foreign sound [f], respectively. The Soviet Armenian government implemented orthographic reform in 1922 and again in 1940, creating a difference between the traditional Mesropian orthography and what is known as Reformed orthography. The 1922 reform limited the use of *w* to the digraph *ow* (or *u*) and treated this digraph as a single letter of the alphabet.

**User Community.** The Mesropian orthography is presently used by West Armenian speakers who live in the diaspora and, rarely, by East Armenian speakers whose origins are in Armenia but who live in the diaspora. The Reformed orthography is used by East Armenian speakers living in the Republic of Armenia and, occasionally, by West Armenian speakers who live in countries formerly under the influence of the former Soviet Union. Spell-checkers and other linguistic tools need to take the differences between these orthographies into account, just as they do for British and American English.

**Punctuation.** Armenian makes use of a number of punctuation marks also used in other European scripts. Armenian words are delimited with spaces and may terminate on either a space or a punctuation mark. U+0589 : ARMENIAN FULL STOP, called *verjakēt* in Armenian, is used to end sentences. A shorter stop functioning like the semicolon (like the *ano teleia* in Greek, but normally placed on the baseline like U+002E FULL STOP) is called *mijakēt*; it is represented by U+2024 . ONE DOT LEADER. U+055D ` ARMENIAN COMMA is actually used more as a kind of colon than as a comma; it combines the functionality of both elision and pause. Its Armenian name is *bowt'*.

In Armenian it is possible to differentiate between word-joining and word-splitting hyphens. To join words, the *miowt'jan gic* - is used; it can be represented by either U+002D

HYPHEN-MINUS OF U+2010 - HYPHEN. At the end of the line, to split words across lines, the *ent'amna* U+058A ֊ ARMENIAN HYPHEN may also be used. This character has a curved shape in some fonts, but a hyphen-like shape in others. Both the word-joiner and the word-splitter can also break at word boundaries, but the two characters have different semantics.

Several other punctuation marks are unique to Armenian, and these function differently from other kinds of marks. The tonal punctuation marks (U+055B ARMENIAN EMPHASIS MARK, U+055C ARMENIAN EXCLAMATION MARK, and U+055E ARMENIAN QUESTION MARK) are placed directly above and slightly to the right of the vowel whose sound is modified, instead of at the end of the sentence, as European punctuation marks are. Because of the mechanical limitations of some printing technologies, these punctuation marks have often been typographically rendered as spacing glyphs above and to the right of the modified vowel, but this practice is not recommended. Depending on the font, the kerning sometimes presents them as half-spacing glyphs, which is somewhat more acceptable. The placement of the Armenian tonal mark can be used to distinguish between different questions.

U+055F ARMENIAN ABBREVIATION MARK, or *patiw*, is one of four abbreviation marks found in manuscripts to abbreviate common words such as God, Jesus, Christos, Lord, Saint, and so on. It is placed above the abbreviated word and spans all of its letters.

**Preferred Characters.** The apostrophe at U+055A has the same shape and function as the Latin apostrophe at U+2019, which is preferred. There is no left half ring in Armenian. Unicode character U+0559 is not used. It appears that this character is a duplicate character, which was encoded to represent U+02BB MODIFIER LETTER TURNED COMMA, used in Armenian transliteration. U+02BB is preferred for this purpose.

**Ligatures.** Five Armenian ligatures are encoded in the Alphabetic Presentation Forms block in the range U+FB13..U+FB17. These shapes (along with others) are typically found in handwriting and in traditional fonts that mimic the manuscript ligatures. Of these, the *men-now* ligature is the one most useful for both traditional and modern fonts.

## 7.7 Georgian

**Georgian:** U+10A0–U+10FF, U+2D00–U+2D2F

The Georgian script is used primarily for writing the Georgian language and its dialects. It is also used for the Svan and Mingrelian languages and in the past was used for Abkhaz and other languages of the Caucasus. It is written from left to right. Spaces are used to separate words and provide the primary line breaking opportunities.

**Script Forms.** The script name “Georgian” in the Unicode Standard is used for what are really two closely related scripts. The original Georgian writing system was an inscriptional form called *Asomtavruli*, from which a manuscript form called *Nuskhuri* was derived. Together these forms are categorized as *Khutsuri* (ecclesiastical), in which *Asomtavruli* is used as the uppercase and *Nuskhuri* as the lowercase. This development of a bicameral script parallels the evolution of the Latin alphabet, in which the original linear monumental style became the uppercase and manuscript styles of the same alphabet became the lowercase. The *Khutsuri* script is still used for liturgical purposes, but was replaced, through a history now uncertain, by an alphabet called *Mkhedruli* (military), which is now the form used for nearly all modern Georgian writing.

Both the *Mkhedruli* alphabet and the *Asomtavruli* inscriptional form are encoded in the Georgian block. The *Nuskhuri* script form is encoded in the Georgian Supplement block.

**Case Forms.** The Georgian *Mkhedruli* alphabet is fundamentally caseless. The scholar Akaki Shanidze attempted to introduce a casing practice for Georgian in the 1950s, but this system failed to gain popularity. In his typographic departure, he used the *Asomtavruli* forms to represent uppercase letters, alongside “lowercase” *Mkhedruli*. This practice is anomalous—the Unicode Standard instead provides case mappings between the two *Khutsuri* forms: *Asomtavruli* and *Nuskhuri*. Figure 7-7 uses Akaki Shanidze’s name to illustrate the various forms of Georgian and its case usage.

Figure 7-7. Georgian Scripts and Casing

Asomtavruli majuscule	ՇԳՇԳՂԳՇԸԻՂԺԻ
Nuskhuri minuscule	շգշգրյցհրძილ
Casing Khutsuri	ՇԳշգրյցհრძილ
Mkhedruli	აკაკი შანიძე
Mtavruli style	აკაკი შანიძე
Shanidze’s orthography	ՇԳշգրյցհრძილ

**Mtavruli Style.** *Mtavruli* is a particular style of *Mkhedruli* in which the distinction between letters with ascenders and descenders is not maintained. All letters appear with an equal height standing on the baseline; *Mtavruli*-style letters are never used as capitals. A word is always entirely presented in *Mtavruli* or not. *Mtavruli* is a font style, similar to CAPS in the Latin script.

**Punctuation.** Modern Georgian text uses generic European conventions for punctuation. See the common punctuation marks in the Basic Latin and General Punctuation blocks.

**Historic Punctuation.** Historic Georgian manuscripts, particularly text in the older, ecclesiastical styles, use manuscript punctuation marks common to the Byzantine tradition. These include single, double, and multiple dot punctuation. For a single dot punctuation mark, U+00B7 MIDDLE DOT or U+2E31 WORD SEPARATOR MIDDLE DOT may be used. Historic double and multiple dot punctuation marks can be found in the U+2056..U+205E range in the General Punctuation block and in the U+2E2A..U+2E2D range in the Supplemental Punctuation block.

U+10FB GEORGIAN PARAGRAPH SEPARATOR is a historic punctuation mark commonly used in Georgian manuscripts to delimit text elements comparable to a paragraph level. Although this punctuation mark may demarcate a paragraph in exposition, it does not force an actual paragraph termination in the text flow. To cause a paragraph termination, U+10FB must be followed by a newline character, as described in *Section 5.8, Newline Guidelines*.

Prior to Version 6.0 the Unicode Standard recommended the use of U+0589 ARMENIAN FULL STOP as the two dot version of the full stop for historic Georgian documents. This is no longer recommended because designs for Armenian fonts may be inconsistent with the display of Georgian text, and because other, generic two dot punctuation characters are available in the standard, such as U+205A TWO DOT PUNCTUATION or U+003A COLON.

For additional punctuation to be used with this script, see C0 Controls and ASCII Punctuation (U+0000..U+007F) and General Punctuation (U+2000..U+206F).

## 7.8 Modifier Letters

Modifier letters, in the sense used in the Unicode Standard, are letters or symbols that are typically written adjacent to other letters and which modify their usage in some way. They are not formally combining marks (gc=Mn or gc=Mc) and do not *graphically* combine with the base letter that they modify. They are base characters in their own right. The sense in which they modify other letters is more a matter of their semantics in usage; they often tend to function as if they were diacritics, indicating a change in pronunciation of a letter, or otherwise distinguishing a letter's use. Typically this diacritic modification applies to the character preceding the modifier letter, but modifier letters may sometimes modify a following character. Occasionally a modifier letter may simply stand alone representing its own sound.

Modifier letters are commonly used in technical phonetic transcriptional systems, where they augment the use of combining marks to make phonetic distinctions. Some of them have been adapted into regular language orthographies as well. For example, U+02BB MODIFIER LETTER TURNED COMMA is used to represent the *ʻokina* (glottal stop) in the orthography for Hawaiian.

Many modifier letters take the form of superscript or subscript letters. Thus the IPA modifier letter that indicates labialization (U+02B7) is a superscript form of the letter *w*. As for all such superscript or subscript form characters in the Unicode Standard, these modifier letters have compatibility decompositions.

**Case and Modifier Letters.** Most modifiers letters are derived from letters in the Latin script, although some modifier letters occur in other scripts. Latin-derived modifier letters may be based on either minuscule (lowercase) or majuscule (uppercase) forms of the letters, but never have case mappings. Modifier letters which have the shape of capital or small capital Latin letters, in particular, are used exclusively in technical phonetic transcriptional systems. Strings of phonetic transcription are notionally lowercase—all letters in them are considered to be lowercase, whatever their shapes. In terms of formal properties in the Unicode Standard, modifier letters based on letter shapes are Lowercase=True; modifier letters not based on letter shapes are simply caseless. All modifier letters, regardless of their shapes, are operationally caseless; they need to be unaffected by casing operations, because changing them by a casing operation would destroy their meaning for the phonetic transcription. Only those superscript or subscript forms that have specific usage in IPA, the Uralic Phonetic Alphabet (UPA), or other major phonetic transcription systems are encoded.

**General Category.** Modifier letters in the Unicode Standard are indicated by either one of two General\_Category values: gc=Lm or gc=Sk. The General\_Category Lm is given to modifier letters derived from regular letters. It is also given to some other characters with more punctuation-like shapes, such as raised commas, which nevertheless have letterlike behavior and which occur on occasion as part of the orthography for regular words in one language or another. The General\_Category Sk is given to modifier letters that typically have more symbol-like origins and which seldom, if ever, are adapted to regular orthogra-

phies outside the context of technical phonetic transcriptional systems. This subset of modifier letters is also known as “modifier symbols.”

This distinction between  $gc=Lm$  and  $gc=Sk$  is reflected in other Unicode specifications relevant to identifiers and word boundary determination. Modifier letters with  $gc=Lm$  are included in the set definitions that result in the derived properties `ID_Start` and `ID_Continue` (and `XID_Start` and `XID_Continue`). As such, they are considered part of the default definition of Unicode identifiers. Modifier *symbols* ( $gc=Sk$ ), on the other hand, are *not* included in those set definitions, and so are excluded by default from Unicode identifiers.

Modifier letters ( $gc=Lm$ ) have the derived property `Alphabetic`, while modifier symbols ( $gc=Sk$ ) do not. Modifier letters ( $gc=Lm$ ) also have the word break property value (`wb=ALetter`), while modifier symbols ( $gc=Sk$ ) do not. This means that for default determination of word break boundaries, modifier symbols will cause a word break, while modifier letters proper will not.

**Blocks.** Most general use modifier letters (and modifier symbols) were collected together in the Spacing Modifier Letters block (U+02B0..U+02FF), the UPA-related Phonetic Extensions block (U+1D00..U+1D7F), the Phonetic Extensions Supplement block (U+1D80..U+1DBF), and the Modifier Tone Letters block (U+A700..U+A71F). However, some script-specific modifier letters are encoded in the blocks appropriate to those scripts. They can be identified by checking for their `General_Category` values.

**Character Names.** There is no requirement that the Unicode character names for modifier letters contain the label “MODIFIER LETTER”, although most of them do.

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

**Phonetic Usage.** The majority of the modifier letters in this block are phonetic modifiers, including the characters required for coverage of the International Phonetic Alphabet. In many cases, modifier letters are used to indicate that the pronunciation of an adjacent letter is different in some way—hence the name “modifier.” They are also used to mark stress or tone, or may simply represent their own sound. Many of these modifiers letters correspond to separate, nonspacing diacritical marks; the specific cross-references can be found in the code charts.

**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.

**Superscript Letters.** Some of the modifier letters are superscript forms of other letters. The most commonly occurring of these superscript letters are encoded in this block, but many others, particularly for use in UPA, can be found in the Phonetic Extensions block (U+1D00..U+1D7F) and in the Phonetic Extensions Supplement block (U+1D80..U+1DBF). The superscript forms of the *i* and *n* letters can be found in the Superscripts and Subscripts block (U+2070..U+209F). The fact that the latter two letters contain the word “superscript” in their names instead of “modifier letter” is an historical artifact of original sources for the characters, and is not intended to convey a functional distinction in the use of these characters in the Unicode Standard.

Superscript modifier letters are intended for cases where the letters carry a specific meaning, as in phonetic transcription systems, and are not a substitute for generic styling mechanisms for superscripting of text, as for footnotes, mathematical and chemical expressions, and the like.

The superscript modifier letters are *spacing* letters, and should be distinguished from superscripted *combining* Latin letters. The superscripted combining Latin letters, as for example those encoded in the Combining Diacritical Marks block in the range U+0363..U+036F, are associated with the Latin historic manuscript tradition, often representing various abbreviatory conventions in text.

**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 encoded 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 (neutral) tone.

**Rhotic Hook.** U+02DE MODIFIER LETTER RHOTIC HOOK is defined in IPA as a free-standing modifier letter. 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 comprises 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 7-8*).

For example, the sequence “1 + 5” in the first row of *Figure 7-8* indicates the sequence of the lowest tone letter, U+02E9 MODIFIER LETTER EXTRA-LOW TONE BAR, followed by the highest tone letter, U+02E5 MODIFIER LETTER EXTRA-HIGH TONE BAR. In that sequence, the tone letter is drawn with a ligation from the iconic position of the low tone to that of the

**Figure 7-8. Tone Letters**

1 + 5	→	∧ (rising contour)
5 + 1	→	∨ (falling contour)
3 + 5	→	↗ (high rising contour)
1 + 3	→	↘ (low rising contour)
1 + 3 + 1	→	↗↘ (rising-falling contour)

high tone to indicate the sharp rising contour. A sequence of three tone letters may also be ligated, as shown in the last row of *Figure 7-8*, to indicate a low rising-falling contour tone.

### **Modifier Tone Letters: U+A700–U+A71F**

The Modifier Tone Letters block contains modifier letters used in various schemes for marking tones. These supplement the more commonly used tone marks and tone letters found in the Spacing Modifier Letters block (U+02B0..U+02FF).

The characters in the range U+A700..U+A707 are corner tone marks used in the transcription of Chinese. They were invented by Bridgman and Wells Williams in the 1830s. They have little current use, but are seen in a number of old Chinese sources.

The tone letters in the range U+A708..U+A716 complement the basic set of IPA tone letters (U+02E5..U+02E9) and are used in the representation of Chinese tones for the most part. The dotted tone letters are used to represent short (“stopped”) tones. The left-stem tone letters are mirror images of the IPA tone letters; like those tone letters, they can be ligated in sequences of two or three tone letters to represent contour tones. Left-stem versus right-stem tone letters are sometimes used contrastively to distinguish between tonemic and tonetic transcription or to show the effects of tonal sandhi.

The modifier letters in the range U+A717..U+A71A indicate tones in a particular orthography for Chinantec, an Oto-Manguean language of Mexico. These tone marks are also spacing modifier letters and are not meant to be placed over other letters.

## 7.9 Combining Marks

Combining marks are a special class of characters in the Unicode Standard that are intended to combine with a preceding character, called their *base*. They have a formal syntactic relationship—or *dependence*—on their base, as defined by the standard. This relationship is relevant to the definition of combining character sequences, canonical reordering, and the Unicode Normalization Algorithm. For formal definitions, see *Section 3.6, Combination*.

Combining marks usually have a visible glyphic form, but some of them are invisible. When visible, a combining mark may interact graphically with neighboring characters in various ways. Visible combining marks are divided roughly into two types: nonspacing marks and spacing marks. In rendering, the nonspacing marks generally have no baseline advance of their own, but instead are said to *apply* to their *grapheme base*. Spacing marks behave more like separate letters, but in some scripts they may have complex graphical interactions with other characters. For an extended discussion of the principles for the application of combining marks, see *Section 3.6, Combination*.

Nonspacing marks come in two types: diacritic and other. The diacritics are exemplified by such familiar marks as the *acute accent* or the *macron*, which are applied to letters of the Latin script (or similar scripts). They tend to indicate a change in pronunciation or a particular tone or stress. They may also be used to derive new letters. However, in some scripts, such as Arabic and Hebrew, other kinds of nonspacing marks, such as *vowel points*, represent separate sounds in their own right and are not considered diacritics.

**Sequence of Base Letters and Combining Marks.** In the Unicode character encoding, all combining marks are encoded *after* their 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 “äü”, *not* “aü”, as shown in *Figure 2-18*.

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 formal behavior of combining marks, see *Section 2.11, Combining Characters*, and *Section 3.6, Combination*.)

**Multiple Semantics.** Because nonspacing combining marks have such a wide variety of applications, they may have multiple semantic values. For example, U+0308 = *diaeresis* = *trema* = *umlaut* = *double derivative*. Such multiple functions for a single combining mark are not separately encoded in the standard.

**Glyphic Variation.** When rendered in the context of a language or script, like ordinary letters, combining marks may be subjected to systematic stylistic variation, as discussed in *Section 7.1, Latin*. 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 a *turned comma above* on a lowercase “g” to avoid conflict with the descender. In many fonts, there is no clear distinction made between U+0326 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. However, this will result in suboptimal appearance in rendering and may cause security problems. See Unicode Technical Report #36, “Unicode Security Considerations.”

For more information, see *Section 5.13, Rendering Nonspacing Marks*.

**Overlaid Diacritics.** A few combining marks are encoded to represent overlaid diacritics such as U+0335 COMBINING SHORT STROKE OVERLAY (= “bar”) or hooks modifying the shape of base characters, such as U+0322 COMBINING RETROFLEX HOOK BELOW. Such overlaid diacritics are not used in decompositions of characters in the Unicode Standard. Overlaid combining marks for the indication of negation of mathematical symbols are an exception to this rule and are discussed later in this section.

One should use the combining marks for overlaid diacritics sparingly and with care, as rendering them on letters may create opportunities for spoofing and other confusion. Sequences of a letter followed by an overlaid diacritic or hook character are *not* canonically equivalent to any preformed encoded character with diacritic even though they may appear the same. See “Non-decomposition of Overlaid Diacritics” in *Section 2.12, Equivalent Sequences* for more discussion of the implications of overlaid diacritics for normalization and for text matching operations.

**Marks as Spacing Characters.** By convention, combining marks may be exhibited in (apparent) isolation by applying them 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. Prior to Version 4.1 of the Unicode Standard, the standard also recommended the use of U+0020 SPACE for display of isolated combining marks. This is no longer recommended, however, because of potential conflicts with the handling of sequences of U+0020 SPACE characters in such contexts as XML.

In charts and illustrations in this standard, the combining nature of these marks is illustrated by applying them to a dotted circle, as shown in the examples throughout this standard.

In a bidirectional context, using any character with neutral directionality (that is, with a Bidirectional Class of ON, CS, and so on) as a base character, including U+00A0 NO-BREAK SPACE, a dotted circle, or any other symbol, can lead to unintended separation of the base character from certain types of combining marks during bidirectional ordering. The result is that the combining mark will be graphically applied to something other than the correct base. This affects spacing combining marks (that is, with a General Category of Mc) but not nonspacing combining marks. The unintended separation can be prevented by bracketing the combining character sequence with RLM or LRM characters as appropriate. For

more details on bidirectional reordering, see Unicode Standard Annex #9, “Unicode Bidirectional Algorithm.”

**Spacing Clones of Diacritical Marks.** The Unicode Standard separately encodes clones of many common European diacritical marks, primarily for compatibility with existing character set standards. These cloned accents and diacritics are *spacing* characters and can be used to display the mark in isolation, without application to a NO-BREAK SPACE. They are cross-referenced to the corresponding combining mark in the names list in the Unicode code charts. For example, U+02D8 BREVE is cross-referenced to U+0306 COMBINING BREVE. Most of these spacing clones also have compatibility decomposition mappings involving U+0020 SPACE, but implementers should be cautious in making use of those decomposition mappings because of the complications that can arise from replacing a spacing character with a SPACE + combining mark sequence.

**Relationship to ISO/IEC 8859-1.** ISO/IEC 8859-1 contains eight characters that are ambiguous regarding whether they denote combining characters or separate spacing characters. In the Unicode Standard, the corresponding code points (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 used only as spacing characters. The Unicode Standard provides unambiguous combining characters in the Combining Diacritical Marks block, 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/IEC 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.

**Diacritics Positioned Over Two Base Characters.** IPA, pronunciation systems, some transliteration systems, and a few languages such as Tagalog use diacritics that are applied to a sequence of two letters. This display of diacritics over two letters, also known as the use of *double diacritics*, is most often noted for the Latin script, which is widely used for transcription and transliteration. However, the use of double diacritics is not limited to the Latin script.

In rendering, these marks of unusual size appear as wide diacritics spanning across the top (or bottom) of the two base characters. The Unicode Standard contains a set of double-diacritic combining marks to represent such forms. Like all other combining nonspacing marks, these marks apply to the previous base character, but they are intended to hang over the following letter as well. For example, the character U+0360 COMBINING DOUBLE TILDE is intended to be displayed as depicted in *Figure 7-9*.

The Unicode Standard also contains a set of combining half diacritical marks, which can be used as an alternative, but not generally recommended, way of representing diacritics over

Figure 7-9. Double Diacritics

$$\begin{array}{c}
 \mathbf{n} + \overset{\circ}{\sim} \rightarrow \overset{\sim}{\mathbf{n}} \\
 \text{006E} \quad \text{0360} \\
 \\
 \mathbf{n} + \overset{\circ}{\sim} + \mathbf{g} \rightarrow \overset{\sim}{\mathbf{ng}} \\
 \text{006E} \quad \text{0360} \quad \text{0067}
 \end{array}$$

a sequence of two (or more) letters. See “Combining Half Marks” later in this section and Figure 7-15.

The double-diacritical marks have a very high combining class—higher than all other non-spacing marks except U+0345 *iota subscript*—and so always are at or near the end of a combining character sequence when canonically reordered. In rendering, the double diacritic will float above other diacritics above (or below other diacritics below)—excluding surrounding diacritics—as shown in Figure 7-10.

Figure 7-10. Positioning of Double Diacritics

$$\begin{array}{c}
 \mathbf{a} + \overset{\circ}{\hat{}} + \overset{\circ}{\sim} + \mathbf{c} + \overset{\circ}{\ddot{}} \rightarrow \overset{\sim}{\hat{\mathbf{a}}\overset{\circ}{\mathbf{c}}} \\
 \text{0061} \quad \text{0302} \quad \text{0360} \quad \text{0063} \quad \text{0308} \\
 \\
 \mathbf{a} + \overset{\circ}{\sim} + \overset{\circ}{\hat{}} + \mathbf{c} + \overset{\circ}{\ddot{}} \rightarrow \overset{\sim}{\hat{\mathbf{a}}\overset{\circ}{\mathbf{c}}} \\
 \text{0061} \quad \text{0360} \quad \text{0302} \quad \text{0063} \quad \text{0308}
 \end{array}$$

In Figure 7-10, the first line shows a combining character sequence in canonical order, with the double-diacritic tilde following a circumflex accent. The second line shows an alternative order of the two combining marks that is canonically equivalent to the first line. Because of this canonical equivalence, the two sequences should display identically, with the double diacritic floating above the other diacritics applied to single base characters.

Occasionally one runs across orthographic conventions that use a dot, an acute accent, or other simple diacritic *above* a *ligature tie*—that is, U+0361 COMBINING DOUBLE INVERTED BREVE. Because of the considerations of canonical order just discussed, one cannot represent such text simply by putting a *combining dot above* or *combining acute* directly after U+0361 in the text. Instead, the recommended way of representing such text is to place U+034F COMBINING GRAPHEME JOINER (CGJ) between the *ligature tie* and the combining mark that follows it, as shown in Figure 7-11.

Figure 7-11. Use of CGJ with Double Diacritics

$$\mathbf{u} + \overset{\circ}{\hat{}} + \overset{\square}{\text{CGJ}} + \overset{\circ}{\acute{}} + \mathbf{i} \rightarrow \overset{\hat{}}{\mathbf{u}\overset{\acute{}}{\mathbf{i}}}$$

0075
0361
034F
0301
0069

Because CGJ has a combining class of zero, it blocks reordering of the double diacritic to follow the second combining mark in canonical order. The sequence of <CGJ, acute> is then rendered with default stacking, placing it centered above the *ligature tie*. This convention can be used to create similar effects with combining marks above other double diacritics (or below double diacritics that render below base characters).

For more information on the combining grapheme joiner, see “Combining Grapheme Joiner” in *Section 23.2, Layout Controls*.

***Diacritics Positioned Over Three or More Base Characters.*** Some transcriptional systems extend the convention of double-diacritic display and show diacritics above (or below) three or more base letters. There are no characters encoded in the Unicode Standard which are specifically designated for plain text representation of triple diacritics. Instead, the recommendation of the Unicode Standard is to use text markup for such representation. The application of modifying text marks to arbitrary spans of text exceeds the normal scope of plain text and is usually better dealt with by conventions designed for rich text. In some limited circumstances, the combining half mark diacritics can be used in combinations to represent triple diacritics, but the display of half mark diacritics used in this way often is unsatisfactory in plain text rendering.

***Subtending Marks.*** An additional class of marks called subtending marks is positioned under (or occasionally over or around) a sequence of several other characters. These marks are typically used in the Arabic script, which has several marks that occur before a sequence of digits and are displayed with an extended swash underneath the digits. These marks often indicate whether the sequence of digits is to be interpreted as a number or a date, for example.

Formally, the subtending marks are not treated as combining marks (gc=M) in the Unicode Standard. They are format characters (gc=Cf) and precede the sequence of characters they subtend, rather than following a single base character, as combining marks do. Proper display of subtending marks requires specialized rendering support. Similar subtending marks are encoded for several other scripts, including Syriac and Kaithi. (See *Section 9.2, Arabic*, *Section 9.3, Syriac*, and *Section 15.2, Kaithi* for a number of examples and further discussion.)

***Combining Marks with Ligatures.*** According to *Section 3.6, Combination*, for a simple combining character sequence such as <*i*, ◌̂>, the nonspacing mark ◌̂ both *applies* to and *depends* on the base character *i*. If the *i* is preceded by a character that can ligate with it, additional considerations apply.

*Figure 7-12* shows typical examples of the interaction of combining marks with ligatures. The sequence <*f*, *i*, ◌̂> is canonically equivalent to <*f*, *î*>. This implies that both sequences should be rendered identically, if possible. The precise way in which the sequence is rendered depends on whether the *f* and *i* of the first sequence ligate. If so, the result of applying ◌̂ should be the same as ligating an *f* with an *î*. The appearance depends on whatever typographical rules are established for this case, as illustrated in the first example of *Figure 7-12*. Note that the two characters *f* and *î* may not ligate, even if the sequence <*f*, *î*> does.

Figure 7-12. Interaction of Combining Marks with Ligatures

$$\begin{aligned}
 \textcircled{1} \quad \mathbf{f} + \mathbf{i} + \hat{\circ} &\equiv \mathbf{f} + \hat{\mathbf{i}} \rightarrow \mathbf{f}\hat{\mathbf{i}}, \hat{\mathbf{f}}\mathbf{i}, \hat{\mathbf{f}}\hat{\mathbf{i}} \\
 \textcircled{2} \quad \mathbf{f} + \tilde{\circ} + \mathbf{i} + \hat{\circ} &\rightarrow \tilde{\mathbf{f}}\hat{\mathbf{i}}, \tilde{\mathbf{f}}\hat{\mathbf{i}} \\
 \textcircled{3} \quad \mathbf{f} + \hat{\circ} + \mathbf{i} + \tilde{\circ} &\rightarrow \hat{\mathbf{f}}\tilde{\mathbf{i}}, \hat{\mathbf{f}}\tilde{\mathbf{i}} \\
 \textcircled{4} \quad \mathbf{f} + \tilde{\circ} + \mathbf{i} + \hat{\circ} &\not\equiv \mathbf{f} + \hat{\circ} + \mathbf{i} + \tilde{\circ}
 \end{aligned}$$

The second and third examples show that by default the sequence  $\langle \mathbf{f}, \tilde{\circ}, \mathbf{i}, \hat{\circ} \rangle$  is visually distinguished from the sequence  $\langle \mathbf{f}, \hat{\circ}, \mathbf{i}, \tilde{\circ} \rangle$  by the relative placement of the accents. This is true whether or not the  $\langle \mathbf{f}, \tilde{\circ} \rangle$  and the  $\langle \mathbf{i}, \hat{\circ} \rangle$  ligate. Example 4 shows that the two sequences are not canonically equivalent.

In some writing systems, established typographical rules further define the placement of combining marks with respect to ligatures. As long as the rendering correctly reflects the identity of the character sequence containing the marks, the Unicode Standard does not prescribe such fine typographical details.

Compatibility characters such as the *fi*-ligature are not canonically equivalent to the sequence of characters in their compatibility decompositions. Therefore, sequences like  $\langle \mathbf{fi}\text{-ligature}, \textcircled{\circ} \rangle$  may legitimately differ in visual representation from  $\langle \mathbf{f}, \mathbf{i}, \textcircled{\circ} \rangle$ , just as the visual appearance of other compatibility characters may be different from that of the sequence of characters in their compatibility decompositions. By default, a compatibility character such as *fi*-ligature is treated as a single base glyph.

### 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 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.

**Underlining and Overlining.** The characters U+0332 COMBINING LOW LINE, U+0333 COMBINING DOUBLE LOW LINE, U+0305 COMBINING OVERLINE, and U+033F COMBINING DOUBLE OVERLINE are intended to connect on the left and right. Thus, when used in combination, they could have the effect of continuous lines above or below a sequence of characters. However, because of their interaction with other combining marks and other layout considerations such as intercharacter spacing, their use for underlining or overlining of text is discouraged in favor of using styled text.

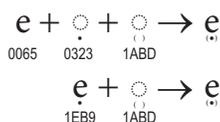
### Combining Diacritical Marks Extended: U+1AB0–U+1AFF

This block contains a set of combining diacritical marks used primarily in phonetic transcription for German dialectology.

**Combining Parentheses.** The combining diacritical marks U+1ABB COMBINING PARENTHESSES ABOVE, U+1ABC COMBINING DOUBLE PARENTHESSES ABOVE, and U+1ABD COMBINING PARENTHESSES BELOW are used in German dialectology to indicate that the effect of a modifier on pronunciation is weakened.

The positioning of these three combining parentheses diacritics deviates from the default stacking behavior of nonspacing marks. Instead of stacking vertically, they are placed side-by-side, surrounding the preceding diacritic above or below the base character. U+1ABB COMBINING PARENTHESSES ABOVE and U+1ABC COMBINING DOUBLE PARENTHESSES ABOVE are intended to be used with diacritics placed above, and U+1ABD COMBINING PARENTHESSES BELOW is intended to be used with diacritics placed below. Correct positioning is illustrated in *Figure 7-13*.

Figure 7-13. Positioning of Combining Parentheses



In contrast with the three combining parentheses diacritical marks above or below, which combine with other diacritics, U+1ABE COMBINING PARENTHESSES OVERLAY is a regular enclosing mark, intended to surround a base character. The exact placement of the overlay U+1ABE with respect to a base character is not specified by the Unicode Standard, but may be adjusted for a particular base character as needed in fonts. For example, in the context of phonetic transcription for German dialectology, the combining character sequence <U+014B LATIN SMALL LETTER ENG, U+1ABE COMBINING PARENTHESSES OVERLAY> could be rendered with the parentheses placed lower to surround the descender of the letter *eng*.

### Combining Diacritical Marks Supplement: U+1DC0–U+1DFF

This block is the supplement to the Combining Diacritical Marks block in the range U+0300..U+036F. It contains lesser-used combining diacritical marks.

U+1DC0 COMBINING DOTTED GRAVE ACCENT and U+1DC1 COMBINING DOTTED ACUTE ACCENT are marks occasionally seen in some Greek texts. They are variant representations of the accent combinations *dialytika varia* and *dialytika oxia*, respectively. They are, however, encoded separately because they cannot be reliably formed by regular stacking rules involving U+0308 COMBINING DIAERESIS and U+0300 COMBINING GRAVE ACCENT OR U+0301 COMBINING ACUTE ACCENT.

U+1DC3 COMBINING SUSPENSION MARK is a combining mark specifically used in Glagolitic. It is not to be confused with a combining breve.

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

The combining marks in this block are generally applied to mathematical or technical symbols. They can be used to extend the range of the symbol set. For example, U+20D2 ◌ COMBINING LONG VERTICAL LINE OVERLAY can be used to express negation, as shown in *Figure 7-14*. Its presentation may change in those circumstances—changing its length or slant, for example. That is, U+2261 ≡ IDENTICAL TO followed by U+20D2 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+20D2 can be used with other symbols to form the negation. For example, U+2258 CORRESPONDS TO followed by U+20D2 can be used to express *does not correspond to*, without requiring that a precomposed form be part of the Unicode Standard.

Figure 7-14. Use of Vertical Line Overlay for Negation



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

**Enclosing Marks.** These nonspacing characters are supplied for compatibility with existing standards, allowing individual base characters to be enclosed in several ways. For example, U+2460 ① CIRCLED DIGIT ONE can be expressed as U+0031 DIGIT ONE “1” + U+20DD ○ COMBINING ENCLOSING CIRCLE. For additional examples, see *Figure 2-17*.

The combining enclosing marks surround their grapheme base and any intervening nonspacing marks. These marks are intended for application to free-standing symbols. See “Application of Combining Marks” in *Section 3.11, Normalization Forms*.

Users should be cautious when applying combining enclosing marks to other than free-standing symbols—for example, when using a combining enclosing circle to apply to a letter or a digit. Most implementations assume that application of any nonspacing mark will not change the character properties of a base character. This means that even though the intent might be to create a circled symbol (General\_Category=So), most software will continue to treat the base character as an alphabetic letter or a numeric digit. Note that there is no *canonical* equivalence between a symbolic character such as U+24B6 CIRCLED LATIN CAPITAL LETTER A and the sequence <U+0041 LATIN CAPITAL LETTER A, U+20DD COMBINING ENCLOSING CIRCLE>, partly because of this difference in treatment of properties.

### 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 marks in legacy implementations.

Unlike other compatibility characters, these half marks do not correspond directly to a single character or a sequence of characters; rather, a discontinuous sequence of the combining half marks corresponds to a single combining mark, as depicted in *Figure 7-15*. The preferred forms are the double diacritics, such as U+0360 COMBINING DOUBLE TILDE. See the earlier discussion of “Diacritics Positioned Above Two Base Characters.”

**Figure 7-15.** Double Diacritics and Half Marks

Using Combining Half Marks

$$\underset{006E}{\mathbf{n}} + \underset{FE22}{\circ} + \underset{0067}{\mathbf{g}} + \underset{FE23}{\circ} \rightarrow \widetilde{\mathbf{ng}}$$

Using Double Diacritics

$$\underset{006E}{\mathbf{n}} + \underset{0360}{\circ} + \underset{0067}{\mathbf{g}} \rightarrow \widetilde{\mathbf{ng}}$$

This block also contains half marks for macrons and conjoining macrons, both above and below. These marks can be used in combinations on successive letters to support particular styles of supralineation or sublineation in some historic scripts. See, for example, *Section 7.3, Coptic*. However, lines which extend across more than two letters may be better rendered if expressed in terms of explicit text styles, rather than by a series of combining half marks, applied one letter at a time in the plain text sequence.

### Combining Marks in Other Blocks

In addition to the blocks of characters in the standard specifically set aside for combining marks, many combining marks are associated with particular scripts or occasionally with groups of scripts. Thus the Arabic block contains a large collection of combining marks used to indicate vowing of Arabic text as well as another collection of combining marks used in annotation of Koranic text. Such marks are mostly intended for use with the Arabic script, but in some instances other scripts, such as Syriac, may use them as well.

Nearly every Indic script has its own collection of combining marks, notably including sets of combining marks to represent dependent vowels, or *matras*.

In some instances a combining mark encoded specifically for a given script, and located in the code chart for that script, may look very similar to a diacritical mark from one of the blocks dedicated to generic combining marks. In such cases, a variety of reasons, including rendering behavior in context or patterning considerations, may have led to separate

encoding. The general principle is that if a correctly identified script-specific combining mark of the appropriate shape is available, that character is intended for use with that script, in lieu of a generic combining mark that might look similar. If a combining mark of the appropriate shape is not available in the relevant script block or blocks, then one should make use of whichever generic combining mark best suits the intended purpose.

For example, in representing Syriac text, to indicate a dot above a letter that was identified as a *qushshaya*, one would use U+0741 SYRIAC QUSHSHAYA rather than the generic U+0307 COMBINING DOT ABOVE . When attempting to represent a *hamza* above a Syriac letter, one would use U+0654 ARABIC HAMZA ABOVE, which is intended for both Arabic and Syriac, because there is no specifically Syriac *hamza* combining mark. However, if marking up Syriac text with diacritics such as a macron to indicate length or some other feature, one would then make use of U+0304 COMBINING MACRON from the generic block of combining diacritical marks.