This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (http://www.unicode.org/errata/). For information on more recent versions of the standard, see http://www.unicode.org/standard/versions/enumeratedversions.html.

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

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

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

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

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

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

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

QA268.U545 2004
005.7'2—dc21

2003052158

For information on obtaining permission for use of material from this work, please submit a written request to the Unicode Consortium, Post Office Box 39146, Mountain View, CA 94039-1476, USA, Fax +1 650 693 3010 or to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300 Boston, MA 02116, USA, Fax: +1 617 848 7047.

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

Chapter 7

European Alphabetic Scripts

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

The European alphabetic scripts described in this chapter are:

- Latin
- Greek
- Cyrillic
- Armenian
- Georgian

They 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 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.

The Latin alphabet is derived from the alphabet used by the Etruscans, who had adopted a Western variant of the classical Greek alphabet. Originally it contained only 24 capital letters. The modern Latin alphabet as it is found in the Basic Latin block owes its appearance to innovations of scribes during the Middle Ages and practices of the early Renaissance printers.

The Cyrillic script was developed in the ninth century and is also based on Greek. 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 upper- and lowercase forms.

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

The block descriptions for archaic European alphabetic scripts, such as Gothic, Ogham, Old Italic, and Runic can be found in *Chapter 13, Archaic Scripts*.

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, "Line Breaking Properties.") Latin letters come in upper- and lowercase pairs.

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

In the Unicode Standard, all diacritical marks are encoded in sequence *after the base characters to which they apply*. For more details, see subsection on "Combining Diacritical Marks" in *Section 7.7, Combining Marks*, and also *Section 2.10, Combining Characters*.

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. The use of the additional characters is not restricted to the context of Latin script usage. The description of these characters is found in *Chapter 6, Writing Systems and Punctuation*.

Related Characters. For other Latin or Latin-derived characters, see Letterlike Symbols (U+2100..U+214F), Currency Symbols (U+20A0..U+20CF), Miscellaneous Symbols (U+2600..U+26FF), Enclosed Alphanumerics (U+2460..U+24FF), 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, such as Italian, which does not use "j" or "w".

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

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

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

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

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

Spacing Clones of Diacritics. ISO/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 character block for Combining Diacritical Marks, which can be used to represent accented Latin letters by means of composed character sequences. U+00B0 ° DEGREE SIGN is also occasionally used ambiguously by implementations of ISO/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.

Latin Extended-A: U+0100–U+017F

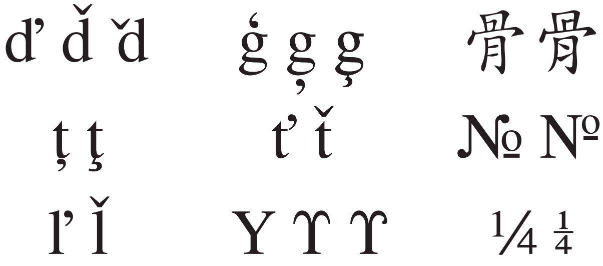
The Latin Extended-A block contains a collection of letters that, when added to the letters contained in the Basic Latin and Latin-1 Supplement blocks, allow for the representation of most European languages that employ the Latin script. Many other languages can also be written with the characters in 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 2.10, Combining Characters.*

Standards. This 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.

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

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

Figure 7-1. Alternative Glyphs



When Czech is typeset, U+010F LATIN SMALL LETTER D WITH CARON and U+0165 LATIN SMALL LETTER T WITH CARON are often rendered by glyphs with apostrophe instead of with caron (háček). 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, on the other hand, glyphs with háčeks are preferred. Languages other than Czech and Slovak that make use of these characters may simply choose to always use the forms with háčeks.

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

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

In Turkish and Romanian, a cedilla and a comma below sometimes replace one another depending on the font style. However, the form with cedilla is preferred in Turkish, and the form with comma below is preferred in Romanian. The characters with explicit commas below are provided to permit the distinction from characters with cedilla. However, legacy encodings for these characters contain only a single form of each of these characters. ISO/IEC 8859-2 maps these to the form with cedilla, while ISO/IEC 8859-16 maps them to the form with comma below. Migrating Romanian 8-bit data to Unicode should be done with care.

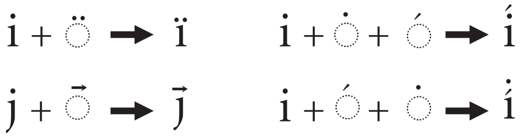
In general, characters with cedillas or ogoneks below are subject to variable typographical usage, depending on the availability and quality of fonts used, the technology, and the geographic area. Various hooks, commas, and squiggles may be substituted for the nominal forms of these diacritics below,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

Exceptional Case Pairs. The characters U+0130 LATIN CAPITAL LETTER I WITH DOT ABOVE and U+0131 LATIN SMALL LETTER DOTLESS I (used primarily in Turkish) are assumed to take ASCII "i" and "I", 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 top nonspacing mark loses the dot in rendering. Thus, in the word *naïve,* the *ï* could be spelled with *i + diaeresis*. Just as Cyrillic A is not equivalent to Latin A, a *dotted-i* is not equivalent to a Turkish *dotless-i + overdot*, nor are other cases of accented *dotted-i* equivalent to accented *dotless-i* (for example, i + ¨≠ ı + ¨). The same pattern is used for *j.*

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

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



Latin Extended-B: U+0180–U+024F

The Latin Extended-B block contains 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).

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

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

Croatian Digraphs Matching Serbian Cyrillic Letters. Serbo-Croatian is a single language with paired alphabets: a Latin script (Croatian) and a Cyrillic script (Serbian). A set of compatibility digraph codes is provided for one-to-one transliteration. There are two potential uppercase forms for each digraph, depending on whether only the initial letter is to be capitalized (titlecase), or both (all uppercase). The Unicode Standard offers both forms so that software can convert one form to the other without changing font sets. The appropriate cross references are given for the lowercase letters.

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

Case Pairs. A number of characters in this block are uppercase forms of characters whose lowercase form is part of some other grouping. Many of these characters came from the International Phonetic Alphabet; they acquired novel uppercase forms when they were adopted into Latin script-based writing systems. Occasionally, however, *alternative* uppercase forms arose in this process. In some instances, research has shown that alternative uppercase forms are merely variants of the same character. If so, such variants are assigned a single Unicode 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 if desired: U+01DD LATIN SMALL LETTER TURNED E is a copy of U+0259 LATIN SMALL LETTER SCHWA.

For historical reasons, the names of some case pairs differ. For example, U+018E LATIN CAPITAL LETTER REVERSED E is the uppercase of U+01DD LATIN SMALL LETTER TURNED E—not of U+0258 LATIN SMALL LETTER REVERSED E. (For default case mappings of Unicode characters, see *Section 4.2, Case—Normative.*)

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

IPA Extensions: U+0250–U+02AF

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

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

Standards. The characters in this block are taken from the 1989 revision of the International Phonetic Alphabet, published by the International Phonetic Association. The International Phonetic Association standard considers IPA to be a separate alphabet, so it includes the entire Latin lowercase alphabet *a–z*, a number of extended Latin letters such as

U+0153 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. Note that unlike other character standards referenced by the Unicode Standard, IPA constitutes an extended alphabet and phonetic transcriptional standard, rather than a character encoding standard.

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

IPA Alternates. In a few cases IPA practice has, over time, produced alternate forms, such as U+0269 LATIN SMALL LETTER IOTA "ɩ" versus U+026A LATIN LETTER SMALL CAPITAL I "ɪ." The Unicode Standard provides separate encodings for the two 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* ɸ, as well as the borrowed letter Greek *iota* ɩ, which has a unique Latin uppercase form. These forms are encoded as separate characters in the Unicode Standard because they have distinct semantics in plain text.

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

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

Phonetic Extensions: U+1D00–U+1D6A

Most of the characters encoded in this block 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°) rotation is used where turning (180°) might result in an ambiguous representation.

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

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

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

Vietnamese Vowel Plus Tone Mark Combinations. A portion of this block (U+1EA0.. U+1EF9) comprises vowel letters of the modern Vietnamese alphabet (*quốc ngữ*) combined with a diacritic mark that denotes the phonemic tone that applies to the syllable. In the modern Vietnamese alphabet, there are 12 vowel letters and 5 tone marks (see *Figure 7-3*).

Figure 7-3. Vietnamese Letters and Tone Marks

a ă â e ê i o ô ơ u ư y

◌́ ◌̀ ◌̉ ◌̃ ◌̣

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

Latin Ligatures: FB00–FB06

This section of the Alphabetic Presentation forms block (U+FB00..U+FB4F) contains several common Latin ligatures, which occur in legacy encodings. 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 15.2, Layout Controls*.

7.2 Greek

Greek: U+0370–U+03FF

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

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. The Unicode Standard encodes Greek characters in the same relative positions as in ISO/IEC 8859-7. A number of variant and archaic characters are taken from the bibliographic standard ISO 5428.

Polytonic Greek. Polytonic Greek, used for ancient Greek (classical and Byzantine) 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 7-1*).

Table 7-1. Nonspacing Marks Used with Greek

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

Because the characters in the Combining Diacritical Marks block are encoded by shape, not by meaning, they are appropriate for use in Greek where applicable. However, the character U+0344 COMBINING GREEK DIALYTIKA TONOS should not be used. When normalized, it is replaced by a U+301 COMBINING ACUTE. For example, the combination of *dialytika* plus *tonos* is instead represented by the sequence U+0308 COMBINING DIAERESIS plus U+0301 COMBINING ACUTE.

Multiple nonspacing marks applied to the same baseform character are encoded in inside-out sequence. See the general rules for applying nonspacing marks in *Section 2.10, 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 ◌̂, an inverted breve ◌̑, a tilde ◌̃, or occasionally a macron ◌̄. 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 *prosgegrammeni* (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 7-1*). 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.

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 final position or 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.

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. With *The Unicode Standard, Version 3.0*, and the concurrent second edition of ISO/IEC 10646-1, the representative glyphs for U+03C6 GREEK LETTER SMALL PHI and U+03D5 GREEK PHI SYMBOL were swapped. 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 (the "loopy" form) and less often with a glyph more like the representative glyph shown for U+03D5 (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 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 µ MICRO SIGN in the Latin-1 Supplement character block and U+2126 Ω 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 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.

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

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

Coptic-Unique Letters. In the Unicode Standard, Version 4.0, the Coptic script is regarded primarily as a stylistic variant of the Greek alphabet. The letters unique to Coptic are encoded in a separate range at the end of the Greek character block. Those characters may be used together with the basic Greek characters to represent the complete Coptic alphabet. Coptic text may be rendered using a font that contains the Coptic style of depicting the characters it shares with the Greek alphabet. Texts that mix Greek and Coptic languages together must employ appropriate font style associations.

The Unicode Technical Committee and ISO/IEC JTC1/SC2 have determined that Coptic would be better handled as a separate script. A future version of the Unicode Standard is therefore likely to contain separate encoded characters for those Coptic letters currently represented by Greek letters.

Related Characters. For math symbols, see *Section 14.4, 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; however, they do not cover all possible combinations in use, so some combining 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 diacritic marks are provided in this character block for use in the representation of polytonic Greek texts. Each has an alternative representation for use with systems that support nonspacing marks. The nonspacing alternatives appear in *Table 7-2*. The spacing forms are 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 7-2*.

Table 7-2. Greek Spacing and Nonspacing Pairs

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

7.3 Cyrillic

Cyrillic: U+0400–U+04FF

The Cyrillic script is one of several scripts that were derived from the Greek script. Cyrillic has traditionally been used for writing various Slavic languages, among which Russian is predominant. In the nineteenth and early twentieth centuries, Cyrillic was extended to write the non-Slavic minority languages of Russia and neighboring countries. The Cyrillic script is written in linear sequence from left to right with the occasional use of nonspacing marks. Cyrillic letters come in upper- and lowercase pairs, with the exception of U+04C0 CYRILLIC LETTER PALOCHKA, which has no lowercase form.

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

Historic Letters. The historical form of the Cyrillic alphabet is treated as a font style variation of modern Cyrillic because the historical forms are relatively close to the modern appearance, and because some of them are still in modern use in languages other than Russian (for example, U+0406 "I"CYRILLIC CAPITAL LETTER I is used in modern Ukrainian and Byelorussian). Some of the letters in this range 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 the *big yus*.

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.

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

Cyrillic Supplement: U+0500–U+052F

Komi. The characters in the range U+0500..U+050F are found in ISO 10754, and 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 *љ* and *њ*, which are ligatures of base letters *л* and *н* with a palatalizing soft sign *ь*. The Molodtsov orthography made use of a different kind of palatalization hook for Komi ԉ, ԋ, ԏ, ԁ, and so on.

7.4 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.

The Armenian script was devised about 406 CE by Mesrop Maštocʿ 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 mutually intelligible literary dialects of Modern Armenian—namely, 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 *Kʿē*

U+0561..U+0584 ա..ք *ayb* to *kʿē*

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 *verǰakēt* 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 *miǰakēt*; 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 *bowtʿ*.

In Armenian it is possible to differentiate between word-joining and word-splitting hyphens. To join words, the *miowtʿjan gic* - is used; it can be represented by U+002D HYPHEN-MINUS or by U+2010 ‐ HYPHEN. At the end of the line, to split words across lines, the *entʿamna* 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. By design, the Unicode Standard does not provide a general mechanism to indicate where ligatures should be displayed.

7.5 Georgian

Georgian: U+10A0–U+10FF

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.

Script Forms. The Georgian script originates from an inscriptional form called *Asomtavruli*, from which was derived a manuscript form called *Nuskhuri*. Together these forms are categorized as *Khutsuri* (ecclesiastical), but *Khutsuri* is not itself the name of a script form. Although no longer seen in most modern texts, the *Nuskhuri* style is still used for liturgical purposes. It 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.

Case Forms. The Georgian alphabet is fundamentally caseless and is used as such in most texts. The scholar Akaki Shanidze attempted to introduce a casing practice for Georgian in the 1950s, but it failed to gain popularity. In this typographic departure, the *Asomtavruli* forms serve to represent uppercase letters, while the lowercase is *Mkhedruli* or *Nuskhuri*. This usage parallels the evolution of the Latin alphabet, in which the original linear monumental style came to be considered uppercase, while manuscript styles of the same alphabet came to be represented as lowercase. The Unicode encoding of Georgian follows the Latin analogy: The range U+10A0..U+10CF is used to encode the uppercase capital forms (*Asomtavruli*), and the basic alphabetic range U+10D0..U+10FF may be regarded as lowercase (*Mkhedruli* or *Nuskhuri*). In lowercase (that is, normal caseless) Georgian text, *Mkhedruli* or *Nuskhuri* are distinguished via font, as are regular and italic forms in Latin lowercase. *Table 7-3* summarizes the relationship between the Georgian forms.

Table 7-3. Font Styles and Georgian Forms

Font style	“uppercase” U+10A0..U+10CF	basic/“lower” U+10D0..U+10FF
Secular	Asomtavruli	Mkhedruli
Ecclesiastical	Asomtavruli	Nuskhuri

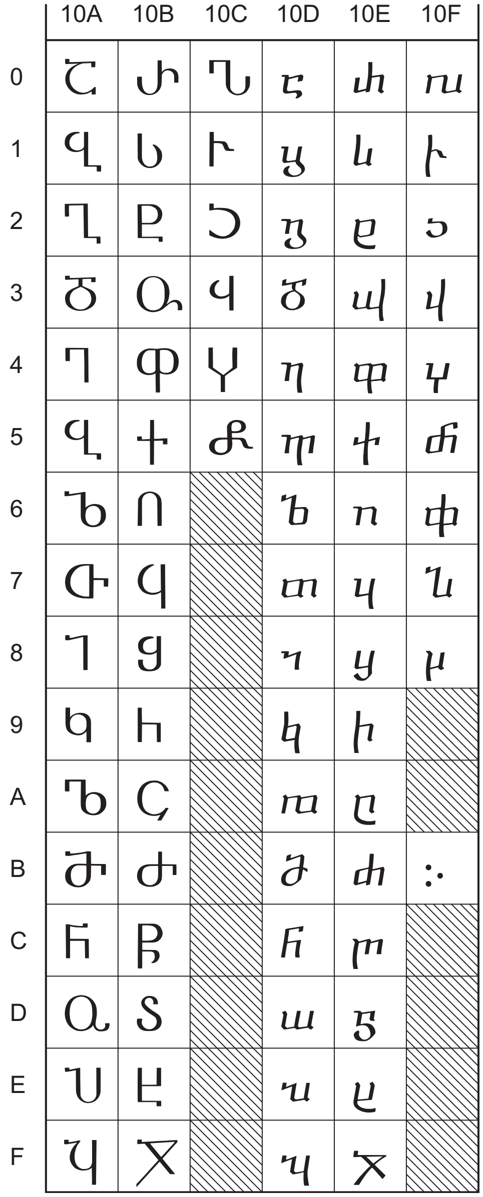
Figure 7-4 shows how the Georgian code chart would appear if presented in an ecclesiastical font.

Because Georgian is predominantly used as a caseless alphabet, no default case mappings are provided for Georgian in the Unicode Character Database. It is inadvisable for generic Unicode text processing to convert Georgian *Mkhedruli* text to *Asomtavruli* via a casing operation. In instances where software dealing with Georgian text treats *Asomtavruli* forms as uppercase letters and requires case folding, this should be done via extended casing rules that constitute a higher-level protocol.

Georgian Paragraph Separator. The Georgian paragraph separator has a distinct representation, so it has been separately encoded as U+10FB. It visually marks a paragraph end, but it must be followed by a newline character to cause a paragraph termination, as described in *Section 5.8, Newline Guidelines*.

Other Punctuation. For the Georgian full stop, use U+0589 ARMENIAN FULL STOP or U+002E FULL STOP.

Figure 7-4. Georgian Displayed with Ecclesiastical Font



	10A	10B	10C	10D	10E	10F
0	Ⴀ	Ⴐ	Ⴠ	ა	რ	ჰ
1	Ⴁ	Ⴑ	Ⴡ	ბ	ს	ჱ
2	Ⴂ	Ⴒ	Ⴢ	გ	ტ	ჲ
3	Ⴃ	Ⴓ	Ⴣ	დ	უ	ჳ
4	Ⴄ	Ⴔ	Ⴤ	ე	ფ	ჴ
5	Ⴅ	Ⴕ	Ⴥ	ვ	ქ	ჵ
6	Ⴆ	Ⴖ		ზ	ღ	ჶ
7	Ⴇ	Ⴗ		თ	ყ	ჷ
8	Ⴈ	Ⴘ		ი	შ	ჸ
9	Ⴉ	Ⴙ		კ	ჩ	
A	Ⴊ	Ⴚ		ლ	ც	
B	Ⴋ	Ⴛ		მ	ძ	჻
C	Ⴌ	Ⴜ		ნ	წ	
D	Ⴍ	Ⴝ		ო	ჭ	
E	Ⴎ	Ⴞ		პ	ხ	
F	Ⴏ	Ⴟ		ჟ	ჯ	

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).

7.6 Modifier Letters

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

Modifier letters are an assorted collection of small signs that are generally used to indicate modifications of a preceding letter. A few may modify the following letter, and some may serve as independent letters. These signs are distinguished from diacritical marks in that modifier letters are treated as free-standing, spacing characters. They are distinguished from similar- or identical-appearing punctuation or symbols by the fact that the members of this block are considered to be letter characters that do not break up a word. They mostly have the "letter" character property (see *Section 4.9, Letters, Alphabetic, and Ideographic*). The majority of these signs are phonetic modifiers, including the characters required for coverage of the International Phonetic Alphabet (IPA).

Phonetic Usage. Modifier letters have relatively well-defined phonetic interpretations. Their usage generally indicates a specific articulatory modification of a sound represented by another letter or intended to convey a particular level of stress or tone. In phonetic usage, the modifier letters are sometimes called "diacritics," which is correct in the logical sense that they are modifiers of the preceding letter. However, in the Unicode Standard, the term "diacritical marks" refers specifically to nonspacing marks, whereas the codes in this block specify *spacing characters*. For this reason, many of the modifier letters in this block correspond to separate diacritical mark codes, which are cross-referenced in *Chapter 16, 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.

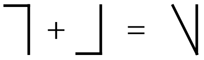
Latin Superscripts. Graphically, some of the phonetic modifier signs are raised or superscripted, some are lowered or subscripted, and some are vertically centered. Only those few forms that have specific usage in IPA, UPA, or other major phonetic systems are encoded.

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

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

Tone Letters. U+02E5..U+02E9 comprise a set of basic tone letters, defined in IPA and commonly used in detailed tone transcriptions of African and other languages. Each tone letter refers to one of five distinguishable tone levels. To represent contour tones, the tone letters are used in combinations. The rendering of contour tones follows a regular set of ligation rules that results in a graphic image of the contour (see *Figure 7-5*).

Figure 7-5. Tone Letters



7.7 Combining Marks

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

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

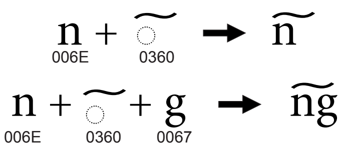
Standards. The combining diacritical marks are derived from a variety of sources, including IPA, ISO 5426, and ISO 6937.

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

The Unicode Standard convention is consistent with the logical order of other nonspacing marks in Semitic and Indic scripts, the great majority of which follow the base characters with respect to which they are positioned. This convention is also in line with the way modern font technology handles the rendering of nonspacing glyphic forms, so that mapping from character memory representation to rendered glyphs is simplified. (For more information on the use of diacritical marks, see *Section 2.10, Combining Characters*, and *Section 3.11, Canonical Ordering Behavior*.)

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. 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 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 7-6*.

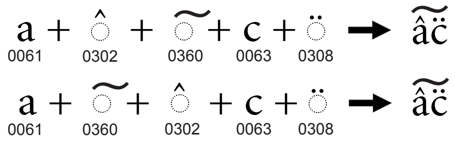
Figure 7-6. Double Diacritics



These double diacritic marks have a very high combining class, higher than all other nonspacing 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 7-7*.

In *Figure 7-7*, 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, 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.

Figure 7-7. Positioning of Double Diacritics



As a consequence, there is currently no mechanism for representing the occasionally seen orthographic convention of using a dot *above* a *ligature tie*—that is, U+0361 combining double inverted breve.

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, 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.

Marks as Spacing Characters. By convention, combining marks may be exhibited in (apparent) isolation by applying them to U+0020 space or to U+00A0 no-break space. This approach might be taken, for example, when referring to the diacritical mark itself as a mark, rather than using it in its normal way in text. The use of U+0020 space versus U+00A0 no-break space affects line breaking behavior.

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

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

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

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

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

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

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

Diacritical marks for symbols are generally applied to mathematical or technical symbols. They can be used to extend the range of the symbol set. For example, U+20D2 COMBINING LONG VERTICAL LINE OVERLAY can be used to express negation. Its presentation may change in those circumstances, changing length or slant. That is, U+2261 IDENTICAL TO followed by U+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.

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.

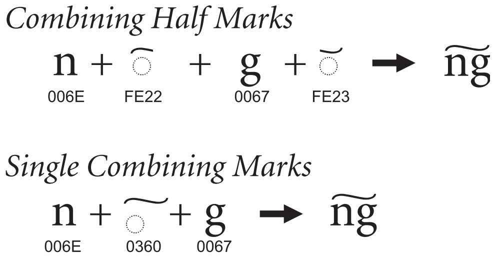
The combining enclosing marks apply to a preceding default grapheme cluster. See UAX #29, "Text Boundaries." These marks are intended for application to free-standing symbols. See "Application of Combining Marks" in *Section 3.11, Canonical Ordering Behavior*.

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 the other compatibility characters, these characters do not correspond to a single nominal character or a sequence of nominal characters; rather, a discontiguous sequence of these combining half marks corresponds to a single combining mark, as depicted in *Figure 7-8*. The preferred forms are the double diacritics, U+0360 and U+0361.

Figure 7-8. Combining Half Marks


Combining Half Marks
n + ◌︢ + g + ◌︣ → n︢g︣
006E
FE22
0067
FE23
Single Combining Marks
n + ◌͠ + g → n͠g
006E
0360
0067