3.1 General Scripts

The General Script area of the Unicode standard includes the encoding of all Latin and non-ideographic script characters.

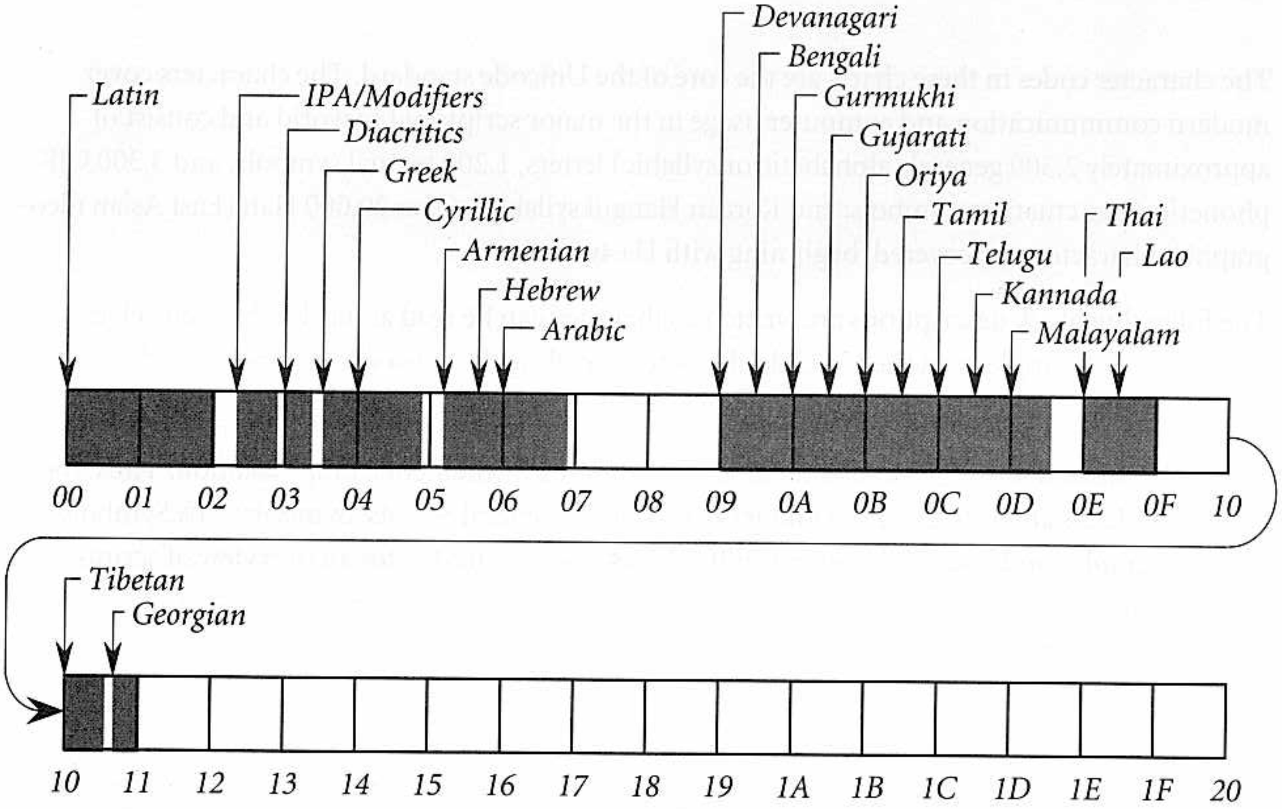

Figure 3-1. General Scripts

ASCII U+0000 → U+007F

Standards. The Unicode standard adapts the ISO 7-bit standard for character encoding by retaining the semantics and numeric values of these codes, merely supplying enough leading zeros to convert them into 16-bit values. The content and arrangement of these standards is far from optimal in the context of a 16-bit space, but the Unicode standard retains them without change because of their prevalence in existing usage.

ISO 646 The ISO character encoding standards are founded on ISO 646, *ISO 7-bit Coded Character Set for Information Interchange.* This standard provides an international set of interpretations for code values 00 to 7F, which are intended to be localized into national standard codes.

ANSI X3.4 This is ASCII: *American National Standard Code for Information Interchange.* ASCII is the version of ISO 646 localized into the national standard code for the USA. In the few places where ISO 646 and ASCII differ, the Unicode standard gives priority to the specific interpretations of ASCII rather than to the generic interpretations of ISO 646. (For example, at code point 24, ISO 646 has the generic *international currency symbol,* whereas in ASCII and for the Unicode value U+0024 this is localized to the *dollar sign.*) The Unicode principle is to designate unambiguous character codes. U+0024 is interpreted as DOLLAR SIGN because code point 24 is the *dollar sign* in ASCII. Other currency symbols are likewise given their own code points within the appropriate blocks, principally U+20A0 → U+20CF, Currency Symbols. The graphic used at code point 24 in ISO 646 is assigned to U+00A4, CURRENCY SIGN.

ASCII C0 Control Codes. The Unicode standard makes no particular use of these control codes, but provides for the passage of the numeric code values intact, neither adding to nor subtracting from their semantics. For more information on control codes, see Section 2.4.

There is a simple one-to-one mapping between 8-bit control codes and Unicode control codes: every 8-bit control code is simply zero-extended to a 16-bit code. For example, if line feed (0A) is to be used for terminal control, then "AB<LF>CD" would be transmitted in plain Unicode text as the following 16-bit values: "0041 0042 000A 0043 0044." Any interpretation of these control codes is outside the scope of the Unicode standard; programmers should refer to the relevant standard (for example, ISO 6429) which specifies control code interpretations.

ASCII Graphic Characters. (U+007F DELETE is a control code, and the remaining 95 codes in this range are graphic characters.) Some of the non-letter characters in this range suffer from overburdened usage as a result of the limited number of codes in a 7-bit space. Some coding consequences of this are discussed below under "Simple Semantic Encoding versus Encoding Characters with Multiple Semantic Value" and "Loose vs. Precise Semantics." The rather haphazard ASCII collection of punctuation and mathematical signs are isolated from the larger body of Unicode punctuation, signs, and symbols (which are encoded in ranges starting at U+2000) only because the relative locations within ASCII are so widely used in standards and software.

Simple Semantic Encoding vs. Encoding Characters with Multiple Semantic Values. Code values in the ASCII range are well-established and used in widely various implementations. The Unicode standard therefore provides only minimal specifications on the typographic appearance of corresponding glyphs. For example, the value U+0024 ($) (derived from ASCII 24) has the semantic *dollar sign,* leaving open the question of whether the dollar sign is to be rendered with one vertical stroke or two. The Unicode value U+0024 refers to the *dollar sign semantic,* not to its precise appearance. Likewise, for other characters in this range that have alternative glyphs, the Unicode character is displayed with the basic or most common glyph, and rendering software may present any other graphical form of that character.

Loose vs. Precise Semantics. Some ASCII characters have multiple uses, either through ambiguity in the original standards or through accumulated reinterpretations of a limited codeset. For example, 27 hex is defined in ANSI X3.4 as *apostrophe (closing single quotation mark; acute accent),* and 2D hex as *hyphen minus.* In general, the Unicode standard provides the same interpretation for the equivalent code values, without adding to or subtracting from their semantics. The Unicode standard supplies *unambiguous* codes elsewhere for the most useful particular interpretations of these ASCII values; the corresponding unambiguous characters are cross-referenced in the character names list for this block. In a few cases, the Unicode standard indicates the generic interpretation of an ASCII code in the name of the corresponding Unicode character, for example U+0027 is APOSTROPHE-QUOTE '.

Diacritics. ASCII contains four codes which are ambiguous: they may be treated either as spacing or as non-spacing marks. The equivalents in the Unicode encoding are: U+005E CIRCUMFLEX ^; U+005F SPACING UNDERSCORE _ ; U+0060 GRAVE `; U+007E TILDE ~. In the Unicode encoding, these code points are restricted to use as spacing characters. The Unicode standard provides unambiguous non-spacing marks in other blocks which can be used to represent accented Latin letters as composed character sequences.

Semantics of Paired Punctuation. Paired punctuation marks such as parentheses (U+0028, U+0029), square brackets (U+005B, U+005D), and braces (U+007B, U+007D) are defined in the Unicode standard as "Opening" and "Closing" rather than "Left" and "Right" so that the semantic will not change in such languages as Arabic, Hebrew or Chinese, where text may flow from right to left or top to bottom. *These characters therefore have consistant semantics but alternative glyphs depending upon the directional flow rendered by a given software program.* The software must ensure that the rendered glyph is the correct one.

Encoding Structure. The Unicode character block for the ASCII character set is divided into the following ranges:

U+0000 → U+001F	ASCII C0 control codes
U+0020 → U+007F	ASCII graphic characters

Latin1 Characters U+0080 → U+00FF

Standards. ISO 8859-1, *Latin1,* is intended to extend ISO 646 by providing additional letters for major languages of Europe (listed below). Like ASCII, the Latin1 set also includes a miscellaneous set of punctuation and mathematical signs. Punctuation, signs, and symbols not included in ASCII and Latin1 are encoded in the range starting at U+2000.

Languages. The languages targeted for coverage by Latin1, ISO 8859-1 are Danish, Dutch, Faroese, Finnish, French, German, Icelandic, Irish, Italian, Norwegian, Portuguese, Spanish, and Swedish. Many other languages can be written with this set of letters, including Hawaiian, Indonesian/ Malay, and Swahili.

C1 Control Codes. Control codes in the C1 range are assigned interpretations in various ISO standards, but do not have the force of long established usage as do those in the ASCII C0 range. Whatever the eventual assignments in the C1 range may be, the Unicode standard makes no particular use of them; their definition is left to other standards.

Diacritics. Latin1 contains five codes which are ambiguous; they may be treated either as spacing or as non-spacing marks. The equivalents in the Unicode encoding are U+00A8, U+00AF, U+00B0, U+00B4, U+00B8. In the Unicode standard, these code points are restricted to use as spacing characters. The Unicode standard provides unambiguous non-spacing marks in other blocks which can be used to represent accented Latin letters as composed character sequences.

U+00AD SOFT HYPHEN indicates a hyphenation point, where a line-break is preferred when a word is to be hyphenated. Depending on the script, the visible rendering of this character when a line break occurs may differ (for example, in some scripts it is rendered as a *hyphen* -, while in others it may be invisible). See also U+2027 HYPHENATION POINT.

U+00A0 NON-BREAKING SPACE is included for compatibility with existing standards. The appearance (width) is the same as the standard space (U+0020), but the NON-BREAKING SPACE generally disallows line breaks on either side.

Quotation Marks. The *guillemets,* U+00AB « and U+00BB », have heterogeneous semantics. They may represent open or close quotation marks, depending on which language they are used with: «in quotes» or »in quotes«.

Encoding Structure. The Unicode Latin1 block is divided into the following ranges:

U+0080 → U+009F	C1 control codes
U+00A0 → U+00FF	Latin1 graphic characters

European Latin U+0100 → U+017F

The European Latin block contains a collection of letters which, when added to the letters contained in the ASCII and Latin1 blocks, allow the representation of most European languages using the Latin script. Many other languages can also be written with this set of letters. Most of these letters are precomposed characters, which can also be represented as composed character sequences. See Section 2.5, Non-spacing Marks.

Standards. This block includes characters contained in the International Standard *Information Processing—8-bit Single-byte Coded Graphic Character Sets*—ISO 8859 Part 2, Part 3, Part 4, and Part 9, also known as Latin2, Latin3, Latin4 and Latin5. This block also includes the repertoire specified for ISO 6937. Many of the other graphic characters contained in these standards such as punctuation, signs, symbols, and diacritical marks are already encoded in the Latin1 block.

Languages. Most languages covered by this block also require characters contained in the ASCII and Latin1 blocks. When combined with these two blocks, the European Latin block covers Afrikaans, Breton, Basque, Catalan, Croatian, Czech, Danish, Dutch, Esperanto, Estonian, Faroese, Finnish, Flemish, Frisian, Greenlandic, Hungarian, Icelandic, Italian, Latin, Latvian, Lithuanian, Malay, Maltese, Norwegian, Polish, Portuguese, Provençal, Rhaeto-Romanic, Romanian, Romany, Slovak, Slovenian, Serbian, Spanish, Swedish, Turkish, Welsh.

Alternative Graphics. Some characters have alternative representations, although they have a common semantic. When Czech is printed in books, letter/apostrophe forms are frequently used. In typewritten or handwritten documents, letter/hacek forms are preferred.

Exceptional Case Pairs. The letters U+0130 LATIN CAPITAL LETTER I DOT and U+0131 LATIN SMALL LETTER DOTLESS I (used primarily in Turkish) are assumed to take ASCII "i" and "I" as their case alternates. This mapping makes the corresponding back mapping ambiguous, but the Unicode standard follows industry practice in this regard.

Encoding Structure. The letters are laid out in alphabetic order, the small letters following the capital letters.

Extended Latin U+0180 → U+01FF

The Extended Latin block contains letterforms used to extend Latin scripts to represent non-European languages. It also contains phonetic symbols not in the International Phonetic Alphabet (the Standard Phonetic block, U+0250 → U+02AF).

Standards. This block covers, among other things, ISO 6438, graphic characters of African languages, *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 Lapp characters from ECMA Registry #144 under ISO 2375.

Arrangement. The characters are arranged in approximate Latin 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 rather 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, Greek-derived.

Croatian Digraphs Matching Serbian Cyrillic Letters. The Unicode standard generally avoids encoding digraphs and other multiple letterforms. An exception is made for Serbo-Croatian, which is a single language with paired alphabets: a Latin script (Croatian) and a Cyrillic script (Serbian). A set of digraph codes is provided to enable direct one-to-one character transliteration between Serbian and Croatian. There are two potential uppercase forms, depending on whether only the initial letter is to be capitalized, or both (for the case of 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 PRC standard GB 2312, as well as JIS X 0212, includes a set of codes for *Pinyin*, the Latin transcription of Mandarin Chinese. Most of the letters used in Pinyin romanization (even those with non-spacing marks) are already covered in the preceding Latin blocks. The group of sixteen codes here complete the Pinyin character set specified in GB 2312 and in 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 came from the International Phonetic Alphabet; they acquired novel uppercase forms when they were adopted into the Latin-based scripts of real languages. Occasionally *alternative* uppercase forms arose by this process. If research has found that alternative uppercase forms are merely variants of the same character they are assigned a single Unicode value, as is the case of U+01B7 LATIN CAPITAL LETTER YOGH. When research has found 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 TURNED-E and U+018F LATIN CAPITAL LETTER SCHWA . In this case, the shared lowercase form is cloned: U+01D5 LATIN SMALL LETTER TURNED E is a clone of U+0259 LATIN SMALL LETTER SCHWA to enable unique case-pair mappings if desired.

Languages. Some indication of language or other usage is given for most characters.

Encoding Structure. The Unicode block for Extended Latin is divided into the following ranges:

U+0180 → U+01C3	Extended Latin
U+01C4 → U+01CC	Croatian digraphs matching Serbian Cyrillic letters
U+01CD → U+01DC	Pinyin diacritic-vowel combinations
U+01DD → U+01F0	Additional characters
U+01F1 → U+01FF	Currently unassigned

Standard Phonetic U+0250 → U+02AF

The Standard Phonetic block contains primarily the unique symbols of the International Phonetic Alphabet (IPA), which is a standard system for indicating specific 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 which are no longer currently sanctioned. The use of non-spacing diacritical marks for close phonetic transcription is an integral part of IPA, as is the use of modifier letters. The IPA diacritics and modifiers are encoded in the two blocks following this one. 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 extended Latin letters rather than IPA forms. Note also that a few non-standard or obsolete phonetic symbols are encoded in the block preceding this one (Extended Latin).

Standards. The characters in this block are taken from the 1989 revision of the International Phonetic Alphabet, published by the International Phonetic Association. This standard considers IPA forms to be a separate alphabet, so it includes the Latin alphabet a-z and other symbols as separate and distinct characters. In contrast, the Unicode standard does not duplicate the Latin alphabet a-z in encoding IPA.

The alphabetic run of characters from U+0299 → U+02A8 represents additions to the Unicode standard reflecting IPA (1989). Some obsolete characters from earlier versions of IPA have also been included.

Unifications. IPA includes the entire lowercase Latin alphabet a-z, a number of extended Latin letters such as U+0153 LATIN SMALL LETTER OE œ, and a few Greek letters. The question of whether these characters are identical when used in an IPA context, or whether all IPA forms should be considered unique characters of a separate alphabet, has many reasonable arguments on both sides. Unicode principles prescribe unification whenever practical and consistent with other principles. Therefore, the IPA symbols are unified as much as possible with other letters (though not with non-letter symbols such as U+222B INTEGRAL SIGN ∫.) A primary reason, aside from reduced duplication, is that the IPA symbols have been adopted into Latin scripts for many languages (such as in Africa). It is futile to attempt to distinguish a transcription from an actual script in such cases. Therefore, many IPA symbols are found outside the IPA block. The Latin alphabet is, of course, coded in the first block. Other IPA characters that are not found in the IPA 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 evolved 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 alternates.

Case Pairs. IPA does not sanction case distinctions, so in effect its phonetic symbols are all lowercase. When IPA symbols are adopted into the actual script of a language, as for example has occurred in Africa, they acquire uppercase forms. Since these uppercase forms are not themselves IPA symbols, they are encoded in the block preceding this one (Extended Latin), and are cross-referenced with the IPA names list.

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

Non-spacing Marks. The Unicode character encoding scheme presumes the necessity of dynamically-applied marks. This principle is an essential element of IPA orthography. Therefore IPA diacritical mark characters are coded in the Generic Diacritical Marks block, U+0300 → U+036F. In IPA, diacritical marks can be freely applied to baseform letters to indicate fine degrees of phonetic differentiation required for precise recording of different languages. In the Unicode standard, all diacritical marks are encoded in sequence *after the base characters* to which they apply. For more details, see the block description for Generic Diacritical Marks, and Section 2.5, Non-spacing Marks.

Affricate Ligatures. IPA officially sanctions six ligatures used in transcription of coronal affricates. These are encoded at U+02A3 → U+02A8. The Unicode standard does not normally encode typographical ligatures, but the IPA ligatures differ in being explicitly defined in IPA and also in having semantic values which make them not simply rendering forms. Thus, for example, while U+02A6 LATIN SMALL LETTER T S is a transcription for the sounds which could also be transcribed in IPA as U+0074 U+0073, "ts," the choice of the ligature may be the result of a deliberate distinction made by the transcriber regarding the systematic phonetic status of the affricate. It certainly should not be a choice left up to rendering software whether to ligate or not based on the font available.

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

Modifier Letters U+02B0 → U+02FF

Modifier Letters are an assorted collection of small signs that are generally used to indicate modifications of the 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 should be interpreted as having the "alphabetic" character property. The majority of these signs are phonetic modifiers, including the characters required for coverage of the International Phonetic Alphabet (IPA).

Phonetic Usage. 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 non-spacing marks, whereas the codes in the current 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 the character names list. Modifier letters have relatively well-defined phonetic interpretations. Their usage is generally to indicate a specific articulatory modification of a sound represented by another letter, or to convey a particular level of stress or tone.

Standards. The modifier letters in the Unicode standard are collated from a variety of sources, the most important of which is the IPA.

Encoding Principles. In this set, there are different characters for the same semantic values, and there are also different semantic values attributed to the same character in different contexts. For example, the letters U+02BC, U+02BE, and U+02C0 have all been used in various publications as a Latin transliteration of the glottal stop (Arabic *hamza*), while U+02BC, at least, has numerous other usages as well. There are also instances where an IPA modifier letter is explicitly equated in semantic value to an IPA non-spacing diacritic form. 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 to choose from. The list of usages given for each modifier letter should not be considered exhaustive.

Latin Superscripts. Graphically, some of the phonetic modifier signs are raised or superscripted, some are lowered or subscripted, and some are vertically centered. The raised signs that derive from Latin letters might suggest the superscripting of the entire Latin alphabet, but the intention here is to encode only those few forms that have specific usage in IPA or other major phonetic systems. The Unicode standard does not in general provide separate codes for superscripted or subscripted characters (although an exception is also made for a limited set of numeric forms to preserve one-to-one mapping with existing standards).

Spacing Clones of Diacritics. Some corporate standards distinguish spacing and non-spacing forms of diacritical marks, and the Unicode standard provides matching codes for these interpretations when practical. The majority of the spacing forms are covered in the Unicode Latin1 block (derived from ISO 8859-1). The six common European diacritics which do not have encodings in ISO 8859-1 are added as spacing characters in the current block. Since the characters frame multiple semantics, these forms may be used with any suitable semantic interpretation (such as U+02D9 SPACING DOT ABOVE ˙ as an indicator of Mandarin Chinese fifth tone).

Rhotic Hook. U+02DE 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 SCHWA + U+02DE RHOTIC HOOK can be treated as equivalent to U+025A SCHWA HOOK.

Tone Letters. U+02E5 → U+02E9 comprise a set of basic tone letters, defined in IPA and commonly used in detailed tone transcription of African and other languages. Each tone letter refers to one of five distinguishable tone levels. In order to represent contour tones, the tone letters may be used in combinations, and their rendering is handled by a regular set of ligation rules which result in a graphic image of the contour:

˥ + ˩ = ˥˩

Encoding Structure. The block for Modifier Letters is divided into the following ranges:

U+02B0 → U+02B8	Phonetic modifiers derived from Latin letters
U+02B9 → U+02D7	Miscellaneous phonetic modifiers
U+02D8 → U+02DD	Spacing clones of non-spacing diacritic marks
U+02DE	Miscellaneous phonetic modifier
U+02DF	Currently unassigned
U+02E0 → U+02E4	Phonetic modifiers derived from Latin letters
U+02E5 → U+02E9	IPA tone letters
U+02EA → U+02FF	Currently unassigned

Generic Diacritical Marks U+0300 → U+036F

The diacritical marks in this block are intended for generic use with any scripts. Diacritical marks which are specific to some particular script are encoded along with the alphabet for that script. Diacritical marks which are primarily used with symbols are defined in code range U+20D0 → U+20FF, Diacritical Marks for Symbols.

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

Sequence of Base Letters and Diacritics. In the Unicode character encoding, all non-spacing marks, including diacritics, are encoded *after* the base character. For example, the Unicode sequence U+0061 "a", U+0308 "¨", U+0075 "u" unambiguously encodes äu, not aü.

The Unicode convention is consistent with the logical order of other non-spacing 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 non-spacing glyphic forms, so that mapping from character store to font rendering is simplified.

Details concerning the use of diacritics are included in the general introduction in Section 2.5.

Diacritics Positioned Over Two or More Base Characters. IPA and a few languages such as Tagalog use diacritics which are applied to two Latin baseform characters. These diacritics will be included in a future version of the Unicode standard.

Marks as Spacing Characters. By convention, Unicode non-spacing marks may be exhibited in (apparent) isolation by applying them to the SPACE character U+0020 or to the NON-BREAKING SPACE U+00A0. This might be done, for example, when talking about the diacritical mark itself as a mark, rather than using it in its normal way in text. Also, the Unicode standard separately encodes clones of many common European diacritical marks that are spacing characters, largely to provide compatibility with existing character sets. These related characters are cross-referenced in the character names list.

Encoding Principles. Because non-spacing 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, U+0326, and U+0327.

Encoding Structure. The Unicode block for generic diacritical marks is divided into the following ranges:

U+0300 → U+0333	Ordinary diacritics
U+0334 → U+0338	Overstruck diacritics
U+0339 → U+033F	Ordinary diacritics
U+0340 → U+0341	Vietnamese tone mark diacritics
U+0342 → U+036F	Currently unassigned

Greek U+0370 → U+03FF

The Greek script is used for writing the Greek language, and (in an extended variant) for the Coptic language. Greek is ancestral to the family of scripts including Latin and Cyrillic.

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

Standards. The ECMA registry under ISO 2375 for use with ISO 2022 contains many Greek subsets. Unicode is based on the latest and most prominent of these: ISO 8859-7, which equals the Greek national standard ELOT 928, and also ECMA-118.

ISO 8859-7 The Unicode standard encodes Greek characters in the same relative positions as in 8859-7. Generic punctuation characters (17 of them) are unified with characters in other Unicode ranges; cross-references to such codes are given in the character names list.

ISO 5428 A number of variant and archaic characters are taken into the Unicode standard from this bibliographic standard.

Polytonic Greek. Polytonic Greek, used for ancient Greek (classical and Byzantine) is coded in the Unicode encoding scheme with composed character sequences, rather than as single precomposed characters.

Non-spacing Marks. Several non-spacing marks may be used with Greek. These are found in the Generic Diacritical Marks range:

U+0300	NON-SPACING GRAVE
U+0301	NON-SPACING ACUTE
U+0302	NON-SPACING CIRCUMFLEX
U+0303	NON-SPACING TILDE
U+0304	NON-SPACING MACRON
U+0306	NON-SPACING BREVE
U+0308	NON-SPACING DIAERESIS
U+0313	NON-SPACING COMMA ABOVE
U+0314	NON-SPACING REVERSED COMMA ABOVE

Since the marks in this range are encoded by shape, not by meaning, they are appropriate for use in Greek where applicable. Multiple non-spacing marks applied onto the same baseform character are to be spelled as the baseform character followed by the several non-spacing mark characters in sequence. The order of non-spacing marks is from the base form outward. See the general rules for applying non-spacing marks in Section 2.5.

Variant Letterforms. Variant forms of Greek letters (sigma and beta) are encoded as separate characters in ISO 8859-7 and ISO 5428, therefore this approach is taken in the Unicode character set.

Greek Letters as Symbols. A few of the Greek variants that are used primarily as technical symbols are placed in this range since they are clearly forms of Greek letters. In some cases, however, Greek letters borrowed into symbol usage may be said to have acquired separate identities, such as U+2126 OHM SIGN Ω vs. U+03A9 GREEK CAPITAL LETTER OMEGA Ω, or U+00B5 MICRO SIGN µ vs. U+03BC GREEK SMALL LETTER MU μ. Despite identical glyphs, the semantic distinctions are so great that these characters are assigned separate codes which are cross-referenced to distinguish them.

Punctuation-like Characters. The question of which punctuation-like characters are uniquely Greek and which ones can be unified with generic Western punctuation has no definitive answer. The Greek question mark U+03D7 ";" was retained for use by systems which treat it as a sentence-final punctuation distinct from the semicolon.

Historic Letters. Historic letters have been retained from ISO 5428. The Unicode standard also includes their lower-case forms.

Coptic-Unique Letters. The Coptic script is regarded as a font/style variant of the Greek alphabet. The letters unique to Coptic have been added, including their lower-case forms. A complete Coptic set would be obtained by rendering the whole Greek alphabet in that same style.

Encoding Structure. The Unicode block for the Greek script is divided into the following ranges:

U+0370 → U+03CF	Letters, punctuation, and diacritical marks from ISO 8859-7
U+03D0 → U+03D6	Variant letterforms
U+03D7 → U+03D9	Punctuation-like characters
U+03DA → U+03E1	Historic letters
U+03E2 → U+03EF	Coptic-unique letters
U+03F0 → U+03F2	Variant letter forms
U+03F3 → U+03F5	Spacing clones of diacritical marks
U+03F6 → U+03FF	Currently unassigned

Cyrillic U+0400 → U+ 04FF

The Cyrillic script is a member of the family of scripts derived from ancient Greek. Cyrillic has traditionally been used for writing various Slavic languages, among which Russian is now predominant. In the 19th and early 20th Centuries Cyrillic was extended to write the non-Slavic minority languages of the Soviet Union.

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.

Standards. The ECMA registry under ISO 2375 for use with ISO 2022 contains several Cyrillic subsets. The Cyrillic block of the Unicode standard is based on the latest and most prominent of these: ISO 8859-5.The Unicode standard encodes Cyrillic characters in the same relative positions as in 8859-5. Four generic punctuation characters are unified with characters in other Unicode character ranges; cross-references to these characters appear in the character names list for this block.

Accented Characters. Some letter forms that might be considered decomposable are not so considered by the Unicode standard when the mark appears integral to the body of the letter. Such letters are encoded as independent characters in order to avoid dispute over whether they have marks or protrusions. For example, many extended Cyrillic characters contain a protrusion at the lower right corner which is subject to wide typographic variability. Also, a few combinations used in archaic Cyrillic are encoded whole because their attachments are not productive.

Unifications. The recently-created alphabets including Extended Cyrillic characters for Soviet minority languages are not well-established. Latin characters included in those alphabets (such as q and w for Kurdish, or U+0292 LATIN SMALL LETTER YOGH for Abkhasian) are not given unique Cyrillic encodings.

Historic Letters. The early form of the Cyrillic alphabet is regarded as a *font change* from modern Cyrillic, because the early Cyrillic forms are relatively close to the modern appearance and because some of them are still in modern use in languages other than Russian (such as U+0406 CYRILLIC CAPITAL LETTER I "I" used in modern Ukrainian and Belorussian). Since the early Cyrillic letters outside of ISO 8859-5, that is, those in the range U+0460 → U+0486, rarely occur in modern form, those letters are shown in the charts in an archaic font.

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

If, at some future date, the old letterforms are adequately documented and the need for them demonstrated, then they can be added to this block.

Glagolitic. The history of the creation of the Cyrillic and Glagolitic 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 is primarily because Glagolitic appears unrecognizably different from Cyrillic, and secondarily because Glagolitic has not grown to match the expansion of Cyrillic. Since Glagolitic is essentially extinct, it is not encoded in the Unicode standard version 1.0.

Encoding Structure. The Unicode block for the Cyrillic script is divided into the following ranges:

U+0400	Currently unassigned
U+0401 → U+040C	Cyrillic characters from ISO 8859-5
U+040D	Currently unassigned
U+040E → U+044F	Cyrillic characters from ISO 8859-5
U+0450	Currently unassigned
U+0451 → U+045C	Cyrillic characters from ISO 8859-5
U+045D	Currently unassigned
U+045E → U+045F	Cyrillic characters from ISO 8859-5
U+0460 → U+0481	Historic letters
U+0482 → U+0486	Historic miscellaneous
U+0487 → U+048F	Currently unassigned
U+0490 → U+04CC	Extended Cyrillic
U+04CD → U+04FF	Currently unassigned

Armenian U+0530 → U+058F

The Armenian script is used primarily for writing the Armenian language. The script is simple, without diacritics. It does have upper- and lower-case pairs.

Modifier Letters. The small marks in the group called Armenian modifier letters are sometimes said to be placed above the alphabetic letters of the words to which they apply, but in modern Armenian typography they are quite uniformly placed *above and to the right*, so that they actually occupy a letter position of their own. Therefore, in the Unicode standard these objects are treated as spacing letters rather than as non-spacing marks.

Encoding Structure. The block for the Armenian script is divided into the following ranges:

U+0530	Currently unassigned
U+0531 → U+0556	Uppercase letters
U+0557 → U+0558	Currently unassigned
U+0559 → U+055F	Modifier letters
U+0560	Currently unassigned
U+0561 → U+0586	Lowercase letters
U+0587 → U+0588	Currently unassigned
U+0589	Punctuation
U+058A → U+058F	Currently unassigned

Hebrew U+0590 → U+05FF

The Hebrew script is used for writing the Hebrew language as well as other languages, primarily Yiddish and Judezmo (Ladino). Vowels and various other marks are written as *points* applied to consonantal base letters; these points are usually omitted in Hebrew. Five Hebrew letters assume a different graphic form when last in a word.

The Hebrew script is written from right to left (the only other right-to-left script currently encoded in Unicode is Arabic). For a general discussion of character ordering including right-to-left scripts, see Appendix A, Directionality.

Standards. The Unicode standard encodes the Hebrew alphabetic characters in the same relative positions as in ISO 8859-8; however, there are no points or Hebrew punctuation characters in ISO 8859-8.

Vowels and Other Marks of Pronunciation. These non-spacing marks, generically called *points,* indicate vowels or other modifications of consonantal letters. The occurrence of a character in the points and punctuation range, depicted with relation to a dashed circle, constitutes an assertion that this character is intended to be applied via some process *to the consonantal character that precedes it* in the text stream. General rules for applying non-spacing marks are given in Section 2.5.

Shin *and* Sin. Separate characters for the dotted letters *shin* and *sin* are not included in the Unicode standard. When it is necessary to distinguish between the two forms, they should be encoded as U+05E9 HEBREW LETTER SHIN followed by the appropriate dot (U+05C1 or U+05C2). This is consistent with Israeli standard encoding.

Cantillation Accents. Cantillation accents have not been included in this edition of the Unicode standard because the Standards Institution of Israel is currently working on a definitive standard for these.

Punctuation. Most punctuation marks used with the Hebrew script are not given independent codes (that is, they are unified with Latin punctuation), except for the few cases where the mark has a unique form in Hebrew: namely *maqaf* (U+05BE), *paseq* (U+05C0), *sof pasuq* (U+05C3), *geresh* (U+05F3), and *gershayim* (U+05F4).

Final (Contextual Variant) Letterforms. Variant forms of five Hebrew letters are encoded as separate characters in all Hebrew standards; therefore this practice is followed in the Unicode standard.

Yiddish Digraphs. These are considered to be independent characters in Yiddish. The Unicode standard has included them as separate characters in order to distinguish certain letter combinations in Yiddish text; for example, to distinguish the digraph *double vav* from an occurrence of a consonantal *vav* followed by a vocalic *vav*. The use of digraphs is consistent with standard Yiddish orthography. Other letters of the Yiddish alphabet, such as *pasekh alef* can be composed from other characters and therefore do not have independent Unicode values.

Encoding Structure. The Unicode character block for the Hebrew script is divided into the following ranges:

U+0590 → U+05AF	Currently unassigned
U+05B0 → U+05C3	Points and punctuation
U+05C4 → U+05CF	Unassigned
U+05D0 → U+05EA	Hebrew letters
U+05EB → U+05EF	Currently unassigned
U+05F0 → U+05F2	Yiddish digraphs
U+05F3 → U+05F4	Additional punctuation
U+05F5	Additional point
U+05F6 → U+05FF	Currently unassigned

Arabic U+0600 → U+ 06FF

The Arabic script is used for writing the Arabic language, and has been extended for representing a number of other languages both major and minor: Persian, Urdu, Pashto, Sindhi and Kurdish among others. Some languages which formerly used the Arabic script now employ the Latin or Cyrillic scripts: Indonesian/Malay, Turkish, Ingush and so on.

The Arabic script is cursive, even in its printed form. As a result, in the handwritten tradition the same letter may be written in different forms depending on how it joins with its neighbors. Vowels and various other marks may be written as *harakat* applied to consonantal base letters; in normal writing, however, these *harakat* are omitted. The script is written from right to left. Arabic and Hebrew are the only scripts currently encoded in the Unicode standard that are written right to left. (See Appendix A, Directionality.)

Standards. The Unicode standard encodes the basic Arabic characters in the same relative positions as in ISO 8859-6 (= ECMA-114 = ASMO 449).

Encoding Principles. The alphabet of the Arabic language is well-defined. Each letter receives only one Unicode character value, no matter how many different contextual appearances it may exhibit in text. Each Unicode character value may be said to represent the inherent semantic identity of the letter. A word is spelled as a sequence of these letters. The graphic form (glyph) shown in the Unicode character chart for an Arabic letter (usually the form of the letter when standing by itself) is *not* the identity of that character. The process of converting electronic or abstract code to visual text form and the graphic elements used to compose visual forms are beyond the scope of the Unicode standard. (See also Compatibility Zone.)

Punctuation. Most punctuation marks used with the Arabic script are *not* given independent codes (that is, are unified with Latin punctuation), *except* for the few cases where the mark has a significantly different appearance in Arabic, namely: U+060C COMMA, U+061B SEMICOLON, U+061F QUESTION MARK, and U+066A PERCENT SIGN. Note that for paired punctuation such as parentheses, the Unicode standard follows a semantic encoding scheme, so that the glyph chosen to represent U+0028 OPENING PARENTHESIS, will depend upon the direction of the rendered text.

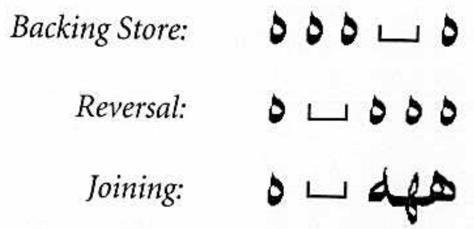

Figure 3-2. Reversal and Cursive Connection

The Non-joiner and the Joiner. The Unicode standard provides two user-selectable zero-width formatting codes: U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER. The use of a non-joiner between two letters prevents them from attaching to each other when rendered. Each letter assumes the correct form, while the context analysis algorithm stays perfectly regular. Examples include the Persian plural suffix, some Persian proper names, and Ottoman Turkish vowels. For further discussion of joiners and non-joiners, see the General Punctuation block description.

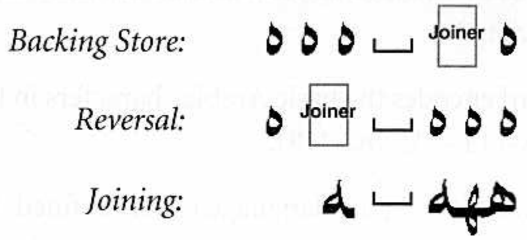

Figure 3-3. Using Joiner

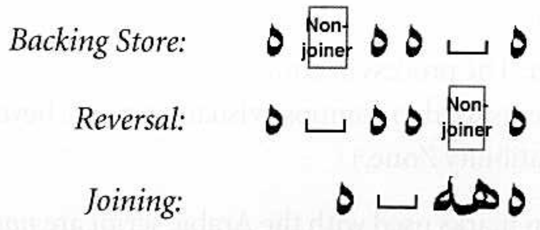

Figure 3-4. Using Non-Joiner

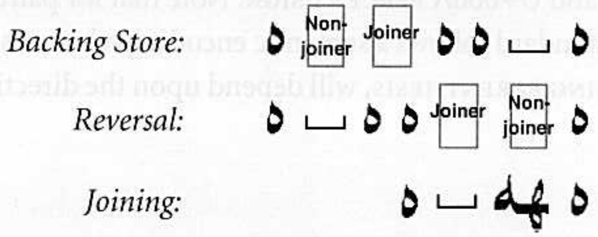

Figure 3-5. Using Joiner & Non-Joiner

Harakat (Vowel) Non-spacing Marks. *Harakat* are marks that indicate vowels or other modifications of consonant letters. The occurrence of a character in the *harakat* range, and its depiction in relation to a dashed circle, constitute an assertion that this character is intended to be applied via some process *to the consonantal character that precedes it* in the text stream, the base character. General rules for applying non-spacing marks are given in the Generic Diacritical Mark block description section. The Unicode standard does not specify a sequence order in case of multiple marks applied to the same Arabic base character, since there is no possible ambiguity of interpretation.

Arabic-Indic Digits. The names for the forms of decimal digits vary widely across different languages. The decimal numbering system originated in India (for example, Devanagari ०१२३ . . .) and was subsequently adopted in the Arabic world with a different appearance (for example, Arabic ٠١٢٣ . . .). The Europeans adopted decimal numbers from the Arabic world, although once again the forms of the digits changed greatly (European 0 1 2 3 . . .). The European forms were later adopted widely around the world, and are used even in many Arabic-speaking countries in North Africa. In each case, the interpretation of decimal numbers remained the same. However, the forms of the digits changed to a degree that they are no longer recognizably the same characters. Because of the origin of these characters, the European decimal numbers are widely known as "Arabic numerals" or "Hindi-Arabic numerals," while the decimal numbers in use in the Arabic world are widely known there as "Hindi numbers."

The Unicode standard includes both Indic digits (including forms used with different Indic scripts), Arabic digits (with forms used in most of the Arabic world), and European digits (now used internationally). Because of this, the traditional names could not be retained without confusion. In addition, there are two main variants of the Arabic-Indic digits, those used in Eastern countries for languages such as Persian and Urdu, and those used in other parts of the Arabic world. The Eastern variant digits are given separate codes in the Unicode standard to account for the differences in appearance and directional treatment when rendering them. (For a complete discussion of directional formatting in the Unicode standard, see Appendix A.)

In summary, the following names will be used in the running text of the Unicode standard:

Name	*Code Points*	*Forms*
European	U+0030 → U+0039	0 1 2 3 . . .
Arabic-Indic	U+0660 → U+0669	٠١٢٣ . . .
Eastern Arabic-Indic	U+06F0 → U+06F9	۰۱۲۳ . . .
Indic	U+0966 → U+096F	Devanagari ०१२३ . . .
	U+09E6 → U+09EF	Bengali

and so on.

These names have been chosen so as to reduce the confusion involved in the use of the decimal number forms. They do not have any normative content; as with the choice of any other names, they are meant to be unique distinguishing labels, and should not be viewed as favoring one culture over another.

Extended Arabic Letters. Arabic script is used to write major languages, such as Persian and Urdu, but it has also been used to transcribe some relatively obscure languages, such as Baluchi and Lahnda, which have little tradition in printed typography. As a result, the set of characters encoded in this section unavoidably contains spurious forms. The Unicode standard encodes multiple forms of the Extended Arabic letters, because for a number of languages, the character forms and usages are not well documented. This approach was felt to be the most practical in the interest of minimizing the risk of omitting valid characters.

Languages. The languages using a given character are indicated, even though this information is incomplete. When such an annotation ends with an ellipsis (...) then the languages cited are merely the principal ones among many.

Reserved Range. The characters in the range U+0700 → U+08FF are unassigned and reserved for use in future right-to-left scripts. See Appendix B, Implementation Guidelines for a discussion of the implications of this.

Encoding Structure. The Arabic block is divided into the following ranges:

U+0600 → U+064A	Arabic letters and punctuation from ISO 8859-6
U+064B → U+0652	Non-spacing characters from 8859-6
U+0653 → U+065F	Currently unassigned
U+0660 → U+0669	Arabic-Indic digits
U+066A → U+066C	Arabic punctuation
U+066D → U+066F	Currently unassigned
U+0670	Additional non-spacing character
U+0671 → U+06D5	Extended Arabic letters
U+06D6 → U+06EF	Currently unassigned
U+06F0 → U+06F9	Eastern Arabic-Indic digits
U+06FA → U+06FF	Currently unassigned

Devanagari U+0900 → U+097F

The Devanagari script is used for writing classical Sanskrit and its modern derivative, Hindi. Extensions to Devanagari are used to write other related languages of northern India (such as Marathi) and of Nepal (Nepali). In addition to the languages mentioned above, the Devanagari script is also used to write the following languages: Awadhi, Bagheli, Bhatneri, Bhili, Bihari, Braj Bhasha, Chhattisgarhi, Garhwali, Gondi (Betul, Chhindwara, Mandla dialects), Harauti, Ho, Jaipuri, Kachchhi, Kanauji, Konkani, Kului, Kumaoni, Kurku, Kurukh, Marwari, Mundari, Newari, Palpa, and Santali.

All other Indian scripts, as well as the Sinhala script of Sri Lanka and the Southeast Asian scripts (Thai, Lao, Khmer, and Burmese) are historically connected with the Devanagari script as descendants of the ancient Brahmi script, and the entire family of scripts shares a large number of structural features.

The Unicode standard follows the ISCII (Indian Standard Code for Information Interchange) code standard in treating all nine of the official Indian scripts (Devanagari, Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam) in a parallel way. This emphasizes the structural similarities of the scripts and follows the stated intention of the Indian coding standards to enable one-to-one mappings between analogous coding positions in different scripts in the family. Sinhala, Thai, Lao, Khmer, and Burmese depart to a greater extent from the Devanagari structural pattern. The Unicode standard does not attempt to provide any direct mappings for Thai or Lao to the Devanagari order, and future versions of the Unicode standard will not provide direct mappings for Sinhala, Khmer, or Burmese to the Devanagari order.

The principles of the Indian scripts are covered in some detail in this introduction to the Devanagari scripts. The remaining introductions to the Indian scripts are abbreviated, so as to highlight the differences from Devanagari, where appropriate.

General Principles of Indian Scripts. Devanagari and other Indian scripts constitute a cross between syllabic writing systems and alphabets. The effective unit for understanding how the system works is a graphemic syllable, consisting of a CV core and any number of preceding consonants. The canonical structure is ((C)C)CV. The graphemic syllable need not correspond exactly with a phonological syllable, especially when a consonant cluster is involved, but the writing system is built on phonological principles and tends to correspond quite closely to pronunciation.

The graphemic syllable is built up of alphabetic pieces, the actual letters of the Devanagari script. These consist of three major types: consonants, dependent vowels, and independent vowels.

Consonants. The consonant letters each represent a single consonantal sound, but also have the peculiarity of having an inherent vowel, generally *a.* Thus U+0915 DEVANAGARI LETTER KA represents not just *k* but *ka.* In the presence of a dependent vowel, the inherent vowel in a consonant letter is overridden by the dependent vowel.

Consonant letters also occur in half-forms, which are presentational forms used as the initial consonant in CCV consonant clusters. These half-consonants do not have an inherent vowel. Their rendered forms in Devanagari often resemble the full consonant, but are missing the vertical stem which marks a syllabic core. (The stem glyph is graphically and historically related to the *a* vowel.)

Some Devanagari consonant letters have multiple glyphic forms which are contextually dependent on neighboring consonants. This is especially true of U+0930 DEVANAGARI LETTER RA, which has numerous different forms, both as the initial and as the final element of a consonant cluster.

The traditional Sanskrit/Devanagari alphabetic order for consonants follows articulatory phonetic principles, starting with velar consonants and moving forward to bilabial consonants, followed by liquids and then fricatives. ISCII and the Unicode standard both observe this traditional order.

Independent Vowels. The independent vowels in Devanagari are consonant letters which can stand on their own. The writing system treats independent vowels as CV syllables in which the consonant is a *zero.* (This means there is no actual consonant articulation involved, but there is a phonological placeholder treated as if a consonant were present.) The independent vowel letters are used to write syllables which start with a vowel.

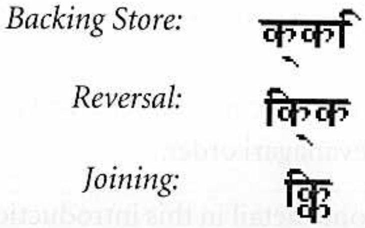

Figure 3-6. Reordering

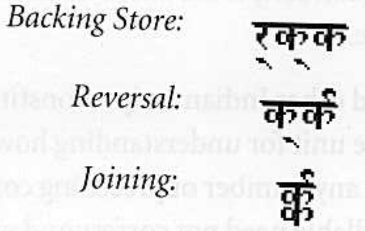

Figure 3-7. Conversion to Non-Spacing Mark and Reordering

Consonant Conjuncts. The Indian scripts are also noted for a large number of consonant conjunct letters. The default behavior for some classes of consonants when combined in clusters is to stack the two consonants vertically. Such forms often involve modification of the glyph forms, or complete substitution of distinct forms for the combination. Glyph modification for combinations of half-consonants with consonant letters is also noted. Such combinations are called conjuncts in

Indian scripts. Well-designed Indian script fonts may contain hundreds of conjuncts, but since they are the result of ligation of distinct consonant letters, those conjuncts are not encoded as Unicode characters. Indian script rendering software must be able to map appropriate combinations of characters in context to the available conjuncts in fonts.

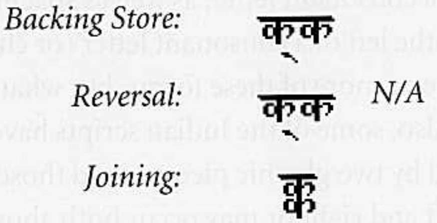

Figure 3-8. Conjunct Characters

Virama. Devanagari and other Indian scripts contain a special character known as the *virama*, or *halant*. The virama combines two successive consonant letters into a cluster, while cancelling the inherent vowel of the first consonant. In Devanagari, the *virama* (U+094D) is a non-spacing subscript character, but its shape may vary from script to script. It is written (in Devanagari) underneath the first of the sequence of consonant letters to be conjoined.

In the Unicode standard, explicit coding of a virama is taken as an indication that the text should be rendered with a conjunct if possible in the font. If the conjunct cannot be formed, then the virama is rendered in relation to the previous character. The ZERO WIDTH NON-JOINER can be used to prevent the formation of conjuncts, if desired. For this purpose, the non-joiner must be placed after the virama in the code stream:

<ka><ka>	/-kaka-/ rendered with sequence of <ka>s
<ka><virama><ka>	/-kka-/ rendered with conjunct if available, else with virama glyph and <ka>s
<ka><virama><non-joiner><ka>	/-kka-/ rendered with two adjacent <ka> characters and a virama under the first <ka>

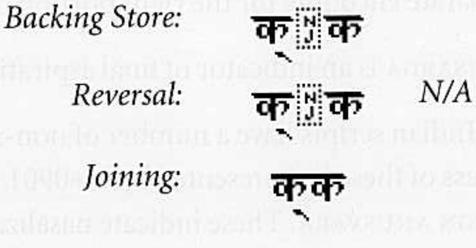

Figure 3-9. Using Non-joiner to Show Virama

Dependent Vowels (Matras). The dependent vowels are the usual vowel letters, generally referred to as *vowel signs* or as *matras* in Sanskrit. The dependent vowels do not stand alone; they depend on a consonant letter for their presentation. A single consonant, or a consonant cluster, has the dependent vowel applied to it to indicate the vowel quality of the syllable. Explicit appearance of a dependent vowel in a syllable effectively cancels the inherent vowel of a single consonant letter.

The greatest variation between different Indian scripts is found in the way that the dependent vowels are applied to the consonant letters. Devanagari has a collection of non-spacing dependent vowel signs which may appear above or below a consonant letter, as well as spacing dependent vowel signs which may occur to the right or to the left of a consonant letter (or cluster of consonant letters). Other Indian scripts generally have one or more of these forms, but what is non-spacing in one script may be a spacing letter in another. Also, some of the Indian scripts have one or more instances of a single dependent vowel indicated by two glyphic pieces—and those glyphic pieces may *surround* a consonant letter both to the left and right or may occur both above and below it.

The logical order for storage of text in Devanagari and all other Indian scripts follows the phonetic order: that is, a CV syllable with a dependent vowel is always encoded as C plus the dependent vowel V in the backing store. Since Devanagari and other Indian scripts have some dependent vowels that are rendered to the left of their consonant letter, the software that renders the Indian scripts must be able to reorder in mapping from the logical (character) store to the presentational (glyph) rendering. In Devanagari, the single vowel sign which is reordered in rendering (but not in the backing store) is U+093F DEVANAGARI VOWEL SIGN I. (Some other Indian scripts have more than one such vowel sign.) Therefore, in the Indic script charts, the dotted circle stands for the base character that precedes it in the backing store.

Devanagari does not have two-part vowel signs—vowels in which half of the vowel is placed on one side of a consonant letter, and the other on the opposite side—but some of the other Indian scripts do. Bengali, for example, has two: U+09CB and U+09CC. The vowel signs are coded in each case in the position in the charts isomorphic with the corresponding vowel in Devanagari. Hence U+09CC BENGALI VOWEL SIGN AU is isomorphic with U+094C DEVANAGARI VOWEL SIGN AU. In order to simplify computer implementations of scripts that use two-part vowel signs, the Unicode starndard uses three positions in the charts (for each Indian script) to encode one or more of the parts of the vowels, when they are not otherwise encodable as two separate characters. Thus U+09D7 BENGALI AU LENGTH MARK is a separate encoding for the right portion of U+09CC.

Other Letters. U+0903 DEVANAGARI SIGN VISARGA is an indicator of final aspiration on a syllable.

Non-spacing Marks. Devanagari and other Indian scripts have a number of non-spacing marks which could be considered diacritic. One class of these is represented by U+0901 DEVANAGARI SIGN CANDRABINDU and U+0902 DEVANAGARI SIGN ANUSVARA. These indicate nasalization or final nasal closure of a syllable.

U+093C DEVANAGARI SIGN NUKTA is a true diacritic. It is used to extend the basic set of consonant letters by modifying them (with a subscript dot in Devanagari) to create new letters.

U+0951 → U+0954 are a set of non-spacing marks used in transcription of Sanskrit texts.

Digits. Each Indian script has a distinct set of digits appropriate to that script. These may or may not be used in ordinary text in that script. European digits have displaced the Indian script forms in modern usage in many of the scripts. Some Indian scripts—notably Tamil—lack a distinct digit for zero.

Punctuation and Symbols. U+0964 DEVANAGARI DANDA is a functional period. Corresponding forms occur in many other Indian scripts. U+0965 DEVANAGARI DOUBLE DANDA marks the end of a verse in traditional texts.

Many modern languages written in the Devanagari script can intersperse punctuation derived from the Latin script. Thus U+002C COMMA and U+002E PERIOD are freely used in writing Hindi, and the *danda* is usually restricted to more traditional texts.

Standards. The Devanagari block of the Unicode standard is based on the ISCII Code as issued by the Department of Electronics of the Government of India, dated 5 July 1988, and as adopted by the Department of Official Language of the Ministry of Home Affairs, Government of India in their publication "Mechanical Facilities in Devanagari." The ISCII standard of 1988 (ISCII-1988) differs from and is an update of earlier ISCII standards issued in 1983 and in 1986. The Unicode standard encodes Devanagari characters in the same relative position as those coded in positions A0 - F4 in the ISCII-1988 standard. The same layout of character encoding points is followed for the other eight Indian scripts in the Unicode standard.

Unifications. ISCII-SPACE (A4) is unified with U+0020 SPACE.

ISCII-INV (DB) is an invisible consonant, used for presentation of dependent vowels in isolation. This character is unified with U+200C ZERO WIDTH NON-JOINER.

Encoding Structure. The Unicode standard lays out the nine Indian scripts in blocks of 128 encoding points, organized in eight columns of sixteen characters each.

The first six columns in each script (0 to 5 or 8 to D) are isomorphic with the ISCII-1988 encoding, except that the last eleven positions (U+0955 → U+095F in Devanagari, for example), which are unassigned or undefined in ISCII-1988, are in fact used in the Unicode encoding.

The seventh column in each script (6 or E), along with the last eleven positions in the sixth column, represent additional character assignments in the Unicode standard which are isomorphic across all nine scripts. For example, positions U+xx66 → U+xx6F or U+xxE6 → U+xxEF code the Indic script digits for each script.

The eighth column in each script (7 or F) is reserved for script-specific additions which do not correspond from one Indian script to the next.

The Unicode block for the Devanagari script is divided into the following specific ranges:

U+0900		Unassigned
U+0901	→ U+0903	Various signs
U+0904		Unassigned
U+0905	→ U+0914	Independent vowels
U+0915	→ U+0939	Consonants
U+093A	→ U+093B	Unassigned
U+093C	→ U+093D	Various signs
U+093E	→ U+094C	Dependent vowel signs
U+094D		Devanagari *Virama*
U+094E	→ U+094F	Unassigned
U+0950	→ U+0954	Various signs
U+0955	→ U+0957	Unassigned
U+0958	→ U+095F	Additional consonants composed with Nukta
U+0960	→ U+0961	Additional independent vowels
U+0962	→ U+0963	Additional dependent vowel signs
U+0964	→ U+0965	Additional punctuation
U+0966	→ U+096F	Devanagari digits
U+0970		Devanagari-specific addition: abbreviation sign
U+0971	→ U+097F	Unassigned

The numerous gaps in the Unicode encoding of the Devanagari script result from the following considerations:

U+0900 Reflects undefined position in ISCII-1988.

U+0904 Results from unification of ISCII-SPACE with Unicode SPACE.

U+0934 ISCII-1988 leaves an unassigned position at D4, with the inferred intention of leaving a collation position for Tamil l. The Unicode standard assigns the corresponding Devanagari character (not used in Hindi) to the U+0934 position.

U+093A ISCII-1988 leaves an unassigned position at DA, with the inferred intention of leaving a position for a marker to aid in the collation of Malayalam. The Unicode standard leaves this position unassigned.

U+093B Results from unification of ISCII-INV with Unicode ZERO WIDTH NON-JOINER.

U+094E Reflects unassigned position in ISCII-1988.

U+094F Reflects undefined position in ISCII-1988.

U+0955 → U+0957 Reflect unassigned positions in ISCII-1988 which are used in the Unicode standard to encode various length marks in Indian scripts other than Devanagari.

U+0958 → U+095E Reflect unassigned positions in ISCII-1988 which are used in the Unicode standard to encode composite consonant letters formed with U+093C DEVANAGARI SIGN NUKTA (or its correspondent form in other Indian scripts). In Devanagari these represent extended-Hindi letters used, for example, in the transcription of Arabic or Persian loanwords in Urdu.

U+095F Reflects undefined position in ISCII-1988 which is used in the Unicode standard analogously to the range U+0958 → U+095E.

U+0971 → U+097F Do not correspond to any position in ISCII-1988, so are simply unassigned positions in the Unicode standard.

Bengali U+0980 → U+09FF

The Bengali script is a North Indian script closely related to Devanagari. It is used to write the Bengali language in West Bengal state (India) and in the nation of Bangladesh. It is also used to write Assamese in Assam (India) and a number of other minority languages (Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Naga, Munda, Rian, and Santali) in northeastern India.

Special Characters. U+09D7 BENGALI AU LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+09CC BENGALI VOWEL SIGN AU.

U+09F2 → U+09F9 are a series of Bengali additions for writing currency and fractions.

Encoding Structure. The Unicode character block for the Bengali script is divided into the following ranges:

U+0980 → U+09EF	Follow the Devanagari prototype
U+09F0 → U+09F9	Bengali-specific additions: currency symbols and so on
U+09FA → U+09FF	Unassigned

Gurmukhi U+0A00 → U+0A7F

The Gurmukhi script is a North Indian script historically derived from an older script called Lahnda. It is quite closely related to Devanagari structurally. Gurmukhi is used to write the Punjabi language in the Punjab in India.

Encoding Structure. The Unicode character block for the Gurmukhi script is divided into the following ranges:

U+0A00 → U+0A6F	Follow the Devanagari prototype
U+0A70 → U+0A75	Gurmukhi-specific additions: letters, diacritics
U+0A76 → U+0A7F	Unassigned

Gujarati U+0A80 → U+0AFF

The Gujarati script is a North Indian script closely related to Devanagari. It is most obviously distinguished from Devanagari by not having a horizontal bar for its letter forms, a characteristic of the older Kaithi script which Gujarati is related to. The Gujarati script is used to write the Gujarati language, of Gujarat state, India.

Encoding Structure. The Unicode block for the Gujarati script is divided into the following ranges:

U+0A80 → U+0AEF	Follow the Devanagari prototype
U+0AF0 → U+0AFF	Unassigned

Oriya U+0B00 → U+0B7F

The Oriya script is a North Indian script structurally similar to Devanagari, but with semicircular lines at the top of most letters, instead of the straight horizontal bars of Devanagari. The actual shapes of the letters, particularly for vowel signs, show similarities to Tamil. The Oriya script is used to write the Oriya language, of Orissa state, India, as well as minority languages such as Khondi and Santali.

Special Characters. U+0B57 ORIYA AU LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0B4C ORIYA VOWEL SIGN AU.

Encoding Structure. The Unicode block for the Oriya script is divided into the following ranges:

U+0B00 → U+0B6F	Follow the Devanagari prototype
U+0B70	Oriya-specific addition
U+0B71 → U+0B7F	Unassigned

Tamil U+0B80 → U+0BFF

The Tamil script is a South Indian script. South Indian scripts are structurally related to the North Indian scripts, but they are used to write Dravidian languages of southern India and of Sri Lanka, which are genetically unrelated to the North Indian languages such as Hindi, Bengali, and Gujarati. The shapes of letters in the South Indian scripts are generally quite distinct from the shapes of letters in Devanagari and its related scripts. This is partly a result of the fact that the South Indian scripts were originally carved with needles on palm leaves, a technology which apparently favored rounded letter shapes rather than square, blocky shapes.

The Tamil script is used to write the Tamil language, of Tamil Nadu state in India as well as minority languages such as Badaga. Tamil is also used in Sri Lanka, Singapore, and parts of Malaysia.

The Tamil script has fewer consonants than the other Indian scripts. It also lacks conjunct consonant forms. Instead of conjunct consonant forms, the *virama* (U+0BCD) is widely used in Tamil text. A form of Tamil known as Tamil Granta adds the consonant letters needed to transcribe Sanskrit and Pali consonants.

Naming Conventions for Mid Vowels. The Unicode character encoding for Tamil uses a distinct set of naming conventions for mid vowels in the South Indian (Dravidian) scripts. These conventions are illustrated by U+0B8E TAMIL LETTER E and U+0B8F TAMIL LETTER EE, to be contrasted with the isomorphic positions in Devanagari: U+090E DEVANAGARI LETTER SHORT E and U+090F DEVANAGARI LETTER E. The Dravidian languages have a regular length distinction in the mid vowels which is not reflected in normal Devanagari. U+090E DEVANAGARI LETTER SHORT E is an addition to Devanagari to enable transcription of the Dravidian short vowel forms. The naming conventions are chosen to best reflect the actual character of the vowels in question in the Dravidian scripts, as well as in Devanagari and the other North Indian scripts.

Special Characters. U+0BD7 TAMIL AU LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0BCC TAMIL VOWEL SIGN AU.

Encoding Structure. The Unicode block for the Tamil script is divided into the following ranges:

U+0B80 → U+0BEF	Follow the Devanagari prototype
U+0BF0 → U+0BF2	Tamil-specific additions
U+0BF3 → U+0BFF	Unassigned

Telugu U+0C00 → U+0C7F

The Telugu script is a South Indian script used to write the Telugu language of Andhra Pradesh state in India, as well as minority languages such as Gondi (Adilabad and Koi dialects) and Lambadi.

Unlike Tamil, the Telugu script writes conjunct consonants with subscripted letters. There are also numerous consonant letters with contextual shape changes when used in conjuncts. Some vowel signs also change their shape in specified combinations.

Special Characters. U+0C55 TELUGU LENGTH MARK is provided as an encoding for the second element of the vowel U+0C47 TELUGU VOWEL SIGN EE. U+0C56 TELUGU AI LENGTH MARK is provided as an encoding for the second element of the two-part vowel U+0C48 TELUGU VOWEL SIGN AI. The length marks are both non-spacing characters.

Encoding Structure. The Unicode block for the Telugu script is divided into the following ranges:

U+0C00 → U+0C6F	Follow the Devanagari prototype
U+0C70 → U+0C7F	Unassigned

Kannada U+0C80 → U+0CFF

The Kannada script is a South Indian script used to write the Kannada (or Kanarese) language of Karnataka state, as well as minority languages such as Tulu.

The Kannada script is very closely related to the Telugu script both with regard to the shapes of the letters and in the way conjunct consonants behave.

Special Characters. U+0CD5 KANNADA LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0CC7 KANNADA VOWEL SIGN EE. U+0CD6 KANNADA AI LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0CC8 KANNADA VOWEL SIGN AI. The Kannada two-part vowels actually consist of a non-spacing element above the consonant letter and one or more spacing letters to the right of the consonant letter.

Encoding Structure. The Unicode block for the Kannada script is divided into the following ranges:

U+0C80 → U+0CEF	Follow the Devanagari prototype
U+0CF0 → U+0CFF	Unassigned

Malayalam U+0D00 → U+0D7F

The Malayalam script is a South Indian script used to write the Malayalam language of Kerala state.

The shapes of Malayalam letters closely resemble those of Tamil. However, Malayalam has a very full and complex set of conjunct consonant forms.

Special Characters. U+0D57 MALAYALAM AU LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0D4C MALAYALAM VOWEL SIGN AU.

Encoding Structure. The Unicode block for the Malayalam script is divided into the following ranges:

U+0D00 → U+0D6F	Follow the Devanagari prototype
U+0D70 → U+0D7F	Unassigned

Thai U+0E00 → U+0E7F

The Thai script is used to write Thai and other Southeast Asian languages, such as Kuy, Lavna, and Pali. It is a member of the Indic family of scripts descended from Brahmi. Thai modifies the original Brahmi letter shapes and extends the number of letters to accommodate features unique to the Thai language, including tone marks derived from superscript digits.

Standards. Thai layout in the Unicode standard is based on the Thai Industrial Standard 620-2529.

General Principles of the Thai Script. In common with the Indic scripts, each Thai letter is a consonant possessing an inherent vowel sound. Thai letters further feature inherent tones. The inherent vowel and tone can be modified by means of letters attached to the base consonant letter. These modifier letters consist of floating vowel signs, floating tone marks, and independent vowel letters. The floating and independent modifier letters are treated somewhat differently in implementations. The floating letters follow the modified consonant in the text stream. The independent vowels are treated as other independent letters and precede or follow the consonant depending on their visual position. The encoding for Thai differs from other Indic alphabets because TIS and ISCII standards made different choices.

Thai Punctuation. Common Thai punctuation includes U+0E46 THAI MAY YAMOK to mark repetition of preceding letters and U+0E5A THAI ANGKHANKHU for ellipsis. U+0E5B THAI KOMUT marks the beginning of religious texts. Thai also uses punctuation marks such as U+002E PERIOD and U+002C COMMA, encoded in the ASCII and Latin1 block.

Thai Transcription of Pali and Sanskrit. Thai is frequently used to write Pali and Sanskrit. When so used, consonant clusters are represented by the explicit use of U+0E3A THAI VOWEL SIGN PHINTHU (*virama*), to mark the removal of the inherent vowel. There is no conjoining behavior, unlike other Indic scripts. U+0E4D THAI VOWEL SIGH NIKKIHIT is the Pali *nigghahita* and Sanskrit *anusvara.* U+0E30 THAI VOWEL SIGN SARA A is the Sanskrit *visarga.* U+0E24 THAI LETTER RY and U+0E26 THAI LETTER LY are vocalic *r* and *l*, with U+0E45 THAI LAAK KANG used to indicate their lengthening.

Alternate Ordering. Alternate Thai vowel signs have been added to permit systems to handle Indic and Thai reordering in a consistent fashion. The phonetic order vowel signs behave as the DEVANAGARI VOWEL SIGN I: they reorder before the preceding consonant. In addition, the *virama* can be used to link consonants into a cluster, so that the phonetic order vowel signs appear before the whole cluster. As with Devanagari, the non-joiner can be used to make the *virama* visible for use in Pali; the non-joiner has no effect on the reordering of the phonetic order vowel signs.

Encoding Structure. The Unicode block for the Thai script is divided into the following ranges:

U+0E01	Currently unassigned
U+0E01 → U+0E2E	Consonant letters
U+0E2F	Punctuation
U+0E30 → U+0E3A	Vowel signs
U+0E3B → U+0E3E	Currently unassigned
U+0E3F	Currency symbol (*baht*)
U+0E40 → U+0E44	Vowel signs (written to the left of a consonant)
U+0E45 → U+0E46	Punctuation
U+0E47	Vowel sign
U+0E48 → U+0E4B	Tone marks (diacritics)
U+0E4C → U+0E4D	Vowel signs
U+0E4E → U+0E4F	Misc. signs
U+0E50 → U+0E59	Thai digits
U+0E5A → U+0E5B	Misc. signs
U+0E5C → U+0E7F	Currently unassigned

Lao U+0E80 → U+0EFF

The Lao language and script are closely related to Thai. The Unicode standard encodes the Lao script in the same relative order as Thai.

There are a few additional letters in Lao that have no match in Thai. These are U+0EBB LAO VOWEL SIGH MAI KON, U+0EBC LAO SEMIVOWEL SIGN LO, and U+0EBD LAO SEMIVOWEL SIGN NYO.

The preceding two semivowel signs are the last remnants "yl" of the system of subscript medials which in Burmese also includes original "rw." (In Burmese and Khmer, and in Indian scripts, there is in addition a full set of subscript consonant forms used for conjuncts. Thai no longer uses any of these. Lao has just the two.)

There are also two ligatures in the Unicode character encoding for Lao: U+0EDC LAO HO NO and U+0EDD LAO HO MO. These correspond to Thai sequences of [h] plus [n] or [h] plus [m] without ligaturing. Their function in Lao is to provide versions of the [n] and [m] consonants with a different inherent tonal implication. They are regarded as distinct letters.

Tibetan U+1000 → U+105F

The Tibetan script is used for writing Tibetan in Tibet proper and for Tibetan and related languages, such as Ladakhi and Lahuli, spoken elsewhere in the Himalayan region, including Bhutan, India, and Nepal. The Tibetan script is a member of the Indic family of scripts descended from Brahmi. The original Brahmi letter shapes can still be clearly discerned in Tibetan, but Tibetan removes the Brahmi voiced aspirates and adds letters for Tibetan sounds not found in Brahmi.

General Principles of the Tibetan Script. As in all Indic scripts, each Tibetan letter is a consonant containing an inherent vowel sound. Tibetan letters each also contain an inherent tone related to the voicing or non-voicing of the original Brahmi letters. This is not marked in the script. The inherent vowels are modified by means of floating non-spacing characters attached to the base letter. Removal of the inherent vowel is not always marked in native Tibetan words and must be determined from context. Consonant clusters are rendered as conjuncts formed by stacking letters along a vertical axis. Conjuncts are represented in the text stream by placing a Tibetan conjunct marker (virama) between letters to be conjoined.

Tibetan Punctuation. Common Tibetan punctuation includes U+1034 SHEY to mark phrases. *Shey* is doubled to mark full stops. U+1039 TIBETAN RINCHENPUNGSHEY is a decorative variant. U+1035 TSEK is a syllable delimiter. Automatic line wrapping processes can generally wrap after occurrences of *tsek* (there are no interword spaces in Tibetan). U+103A TIBETAN DRUISHEY is sometimes used at the beginning (and less frequently the end) of a text. The character U+1033 TIBETAN GOYIK is an honorific flourish, double (*swasti*) and triple forms of which are used at the beginnings of texts. It normally joins with one or two more occurrences of the same character to form ligatures, and is almost never used alone; it is often followed by *shey* in a decorative form. The *Wheel of Dharma*, which occurs sometimes in Tibetan texts, is encoded in the Miscellaneous Dingbats block at U+2638.

Tibetan Transcription of Sanskrit. Tibetan is also used to write Sanskrit. The Sanskrit retroflex letters are retained; the voiced aspirates are represented by conjuncts formed of consonants placed above the letter U+1021 HA. U+102C NAMCHEY is the *visarga* (see Devanagari), and U+102E NGARO is the *anusvara.*

When the Tibetan script is used to write Sanskrit, consonants are frequently stacked vertically in ways that do not occur in native Tibetan words; this usually indicates deletion of one or more vowel sounds. This stacking behavior is indicated by insertion of a U+104B VIRAMA between the consonants to be stacked. In native Tibetan, the letter U+101A AA sometimes appears as a subscript below another consonant to mark vowel lengthening, with or without another vowel sign. This stacking is also indicated by inserting a *virama* between the consonant and letter AA.

The Unicode standard does not encode superscript and subscript forms for the letters WA (U+1017), RA, (U+101C) and YA, (U+101B), as these shape changes can be determined from context and by the typographical rules for Tibetan. Examples of these modified forms are *ra-ta* (*ra* subjoined) and *ra-go* (*ra* head). In some rare cases, conjoining occurs in written Tibetan without the normal shape changes (non-morphological conjunction); such cases may be encoded in plain text by using a ZERO WIDTH JOINER (U+200D) instead of the *virama* between the letters to be conjoined. The normal rules for form changes in written Tibetan are contained in various grammatical treatises on the Tibetan language.

Encoding Structure. The Unicode block for the Tibetan script is divided into the following ranges:

U+1000 → U+1022	Consonant letters
U+1023 → U+1025	Currently unassigned
U+1026 → U+102A	Non-spacing Vowel Signs
U+102B → U+103C	Symbols
U+103D → U+103E	Double vowel signs
U+103F	Currently unassigned
U+1040 → U+1049	Digits
U+104A → U+104C	Symbols
U+104D → U+105F	Currently unassigned

Georgian U+10A0 → U+10FF

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

Archaic Script Form. The modern Georgian script is a style called *mkhedruli* (soldier's), which originated as the secular derivative of a form called *khutsuri* (ecclesiastical) that had uppercase/lowercase pairs. While *khutsuri* is essentially extinct, it is occasionally seen; the Unicode standard encodes the uppercase form of *khutsuri* as well as the lowercase letters of modern Georgian. The archaic Georgian paragraph separator has a distinct representation, so has been separately encoded at U+10FB instead of being unified with the PARAGRAPH SEPARATOR of the General Punctuation block.

For the Georgian full stop, use U+0589.

Encoding Structure. The Unicode block for the Georgian script is divided into the following ranges:

U+10A0 → U+10C5	Historic (uppercase) letters
U+10C6 → U+10CF	Currently unassigned
U+10D0 → U+10F0	Modern letters
U+10F1 → U+10F6	Archaic (lowercase) letters
U+10F7 → U+10FA	Currently unassigned
U+10FB	Punctuation
U+10FC → U+10FF	Currently unassigned