The Unicode Standard
Version 8.0 – Core Specification

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

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Chapter 12

South and Central Asia-I

Official Scripts of India

The following South Asian scripts are described in this chapter:

Devanagari	*Gujarati*	*Telugu*
Bengali	*Oriya*	*Kannada*
Gurmukhi	*Tamil*	*Malayalam*

The scripts of South Asia share so many common features that a side-by-side comparison of a few will often reveal structural similarities even in the modern letterforms. With minor historical exceptions, they are written from left to right. They are all *abugidas* in which most symbols stand for a consonant plus an inherent vowel (usually the sound /a/). Word-initial vowels in many of these scripts have distinct symbols, and word-internal vowels are usually written by juxtaposing a vowel sign in the vicinity of the affected consonant. Absence of the inherent vowel, when that occurs, is frequently marked with a special sign. In the Unicode Standard, this sign is denoted by the Sanskrit word *virāma*. In some languages, another designation is preferred. In Hindi, for example, the word *hal* refers to the character itself, and *halant* refers to the consonant that has its inherent vowel suppressed; in Tamil, the word *puḷḷi* is used. The virama sign nominally serves to suppress the inherent vowel of the consonant to which it is applied; it is a combining character, with its shape varying from script to script.

Most of the scripts of South Asia, from north of the Himalayas to Sri Lanka in the south, from Pakistan in the west to the easternmost islands of Indonesia, are derived from the ancient Brahmi script. The oldest lengthy inscriptions of India, the edicts of Ashoka from the third century BCE, were written in two scripts, Kharoshthi and Brahmi. These are both ultimately of Semitic origin, probably deriving from Aramaic, which was an important administrative language of the Middle East at that time. Kharoshthi, written from right to left, was supplanted by Brahmi and its derivatives. The descendants of Brahmi spread with myriad changes throughout the subcontinent and outlying islands. There are said to be some 200 different scripts deriving from it. By the eleventh century, the modern script known as Devanagari was in ascendancy in India proper as the major script of Sanskrit literature.

The North Indian branch of scripts was, like Brahmi itself, chiefly used to write Indo-European languages such as Pali and Sanskrit, and eventually the Hindi, Bengali, and Gujarati languages, though it was also the source for scripts for non-Indo-European languages such as Tibetan, Mongolian, and Lepcha.

The South Indian scripts are also derived from Brahmi and, therefore, share many structural characteristics. These scripts were first used to write Pali and Sanskrit but were later adapted for use in writing non-Indo-European languages—namely, the languages of the Dravidian family of southern India and Sri Lanka. Because of their use for Dravidian languages, the South Indian scripts developed many characteristics that distinguish them from the North Indian scripts. South Indian scripts were also exported to southeast Asia and were the source of scripts such as Tai Tham (Lanna) and Myanmar, as well as the insular scripts of the Philippines and Indonesia.

The shapes of letters in the South Indian scripts took on a quite distinct look from the shapes of letters in the North Indian scripts. Some scholars suggest that this occurred because writing materials such as palm leaves encouraged changes in the way letters were written.

The major official scripts of India proper, including Devanagari, are documented in this chapter. They are all encoded according to a common plan, so that comparable characters are in the same order and relative location. This structural arrangement, which facilitates transliteration to some degree, is based on the Indian national standard (ISCII) encoding for these scripts.

The first six columns in each script are isomorphic with the ISCII-1988 encoding, except that the last 11 positions (U+0955..U+095F in Devanagari, for example), which are unassigned or undefined in ISCII-1988, are used in the Unicode encoding. The seventh column in each of these scripts, along with the last 11 positions in the sixth column, represent additional character assignments in the Unicode Standard that are matched across some or all of the scripts. For example, positions U+xx66..U+xx6F and U+xxE6..U+xxEF code the Indic script digits for each script. The eighth column for each script is reserved for script-specific additions that do not correspond from one Indic script to the next.

While the arrangement of the encoding for the scripts of India is based on ISCII, this does not imply that the rendering behavior of South Indian scripts in particular is the same as that of Devanagari or other North Indian scripts. Implementations should ensure that adequate attention is given to the actual behavior of those scripts; they should not assume that they work just as Devanagari does. Each block description in this chapter describes the most important aspects of rendering for a particular script as well as unique behaviors it may have.

Many of the character names in this group of scripts represent the same sounds, and common naming conventions are used for the scripts of India.

12.1 Devanagari

Devanagari: U+0900–U+097F

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

All other Indic scripts, as well as the Sinhala script of Sri Lanka, the Tibetan script, and the Southeast Asian scripts, are historically connected with the Devanagari script as descendants of the ancient Brahmi script. The entire family of scripts shares a large number of structural features.

The principles of the Indic scripts are covered in some detail in this introduction to the Devanagari script. The remaining introductions to the Indic scripts are abbreviated but highlight any differences from Devanagari where appropriate.

Standards. The Devanagari block of the Unicode Standard is based on ISCII-1988 (Indian Script Code for Information Interchange). The ISCII standard of 1988 differs from and is an update of earlier ISCII standards issued in 1983 and 1986.

The Unicode Standard encodes Devanagari characters in the same relative positions as those coded in positions A0–$F4_{16}$ in the ISCII-1988 standard. The same character code layout is followed for eight other Indic scripts in the Unicode Standard: Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam. This parallel code layout emphasizes the structural similarities of the Brahmi 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, Tibetan, Thai, Lao, Khmer, Myanmar, and other scripts depart to a greater extent from the Devanagari structural pattern, so the Unicode Standard does not attempt to provide any direct mappings for these scripts to the Devanagari order.

In November 1991, at the time *The Unicode Standard, Version 1.0,* was published, the Bureau of Indian Standards published a new version of ISCII in Indian Standard (IS) 13194:1991. This new version partially modified the layout and repertoire of the ISCII-1988 standard. Because of these events, the Unicode Standard does not precisely follow the layout of the current version of ISCII. Nevertheless, the Unicode Standard remains a superset of the ISCII-1991 repertoire. Modern, non-Vedic texts encoded with ISCII-1991 may be automatically converted to Unicode code points and back to their original encoding without loss of information. The Vedic extension characters defined in IS 13194:1991 *Annex G—Extended Character Set for Vedic* are are now fully covered by the Unicode Standard, but the conversions between ISCII and Unicode code points in some cases are more complex than for modern texts.

Encoding Principles. The writing systems that employ Devanagari and other Indic scripts constitute abugidas—a cross between syllabic writing systems and alphabetic writing systems. The effective unit of these writing systems is the orthographic syllable, consisting of a consonant and vowel (CV) core and, optionally, one or more preceding consonants, with a canonical structure of (((C)C)C)V. The orthographic 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 orthographic syllable is built up of alphabetic pieces, the actual letters of the Devanagari script. These pieces consist of three distinct character types: consonant letters, independent vowels, and dependent vowel signs. In a text sequence, these characters are stored in logical (phonetic) order. Consonant letters by themselves constitute a CV unit, where the V is an *inherent* vowel, whose exact phonetic value may vary by writing system. Independent vowels also constitute a CV unit, where the C is considered to be null.

A dependent vowel sign is used to represent a V in CV units where C is not null and V is not the inherent vowel. CV units are not represented by sequences of a consonant followed by virama followed by independent vowel. In some cases, a phonological diphthong (such as Hindi जाओ /jāo/) is actually written as two orthographic CV units, where the second of these units is an independent vowel letter, whose C is considered to be null.

Principles of the Devanagari Script

Rendering Devanagari Characters. Devanagari characters, like characters from many other scripts, can combine or change shape depending on their context. A character's appearance is affected by its ordering with respect to other characters, the font used to render the character, and the application or system environment. These variables can cause the appearance of Devanagari characters to differ from their nominal glyphs (used in the code charts).

Additionally, a few Devanagari characters cause a change in the order of the displayed characters. This reordering is not commonly seen in non-Indic scripts and occurs independently of any bidirectional character reordering that might be required.

Consonant Letters. Each consonant letter represents a single consonantal sound but also has the peculiarity of having an *inherent vowel*, generally the short vowel /a/ in Devanagari and the other Indic scripts. Thus U+0915 DEVANAGARI LETTER KA represents not just /k/ but also /ka/. In the presence of a dependent vowel, however, the inherent vowel associated with a consonant letter is overridden by the dependent vowel.

Consonant letters may also be rendered as *half-forms*, which are presentation forms used to depict non-final consonants in consonant clusters. These half-forms 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 sign denoting the inherent /a/ vowel.)

Some Devanagari consonant letters have alternative presentation forms whose choice depends on neighboring consonants. This variability is especially notable for U+0930 DEVANAGARI LETTER RA, which has numerous different forms, both as the initial element and as the final element of a consonant cluster. Only the nominal forms, rather than the contextual alternatives, are depicted in the code charts.

The traditional Sanskrit/Devanagari alphabetic encoding 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 Vowel Letters. The independent vowels in Devanagari are letters that stand on their own. The writing system treats independent vowels as orthographic CV syllables in which the consonant is null. The independent vowel letters are used to write syllables that start with a vowel.

Dependent Vowel Signs (Matras). The dependent vowels serve as the common manner of writing noninherent vowels and are generally referred to as *vowel signs*, or as *matras* in Sanskrit. The dependent vowels do not stand alone; rather, they are visibly depicted in combination with a base letterform. A single consonant or a consonant cluster may have a dependent vowel applied to it to indicate the vowel quality of the syllable, when it is different from the inherent vowel. Explicit appearance of a dependent vowel in a syllable overrides the inherent vowel of a single consonant letter.

The greatest variation among different Indic scripts is found in the way that the dependent vowels are applied to base letterforms. Devanagari has a collection of nonspacing dependent vowel signs that may appear above or below a consonant letter, as well as spacing dependent vowel signs that may occur to the right or to the left of a consonant letter or consonant cluster. Other Indic scripts generally have one or more of these forms, but what is a nonspacing mark in one script may be a spacing mark in another. Also, some of the Indic scripts have single dependent vowels that are indicated by two or more glyph components—and those glyph components may *surround* a consonant letter both to the left and to the right or may occur both above and below it.

In modern usage the Devanagari script has only one character denoting a left-side dependent vowel sign: U+093F DEVANAGARI VOWEL SIGN I. In the historic Prishthamatra orthography, Devanagari also made use of one additional left-side dependent vowel sign: U+094E DEVANAGARI VOWEL SIGN PRISHTHAMATRA E. Other Indic scripts either have no such vowel signs (Telugu and Kannada) or include as many as three of these signs (Bengali, Tamil, and Malayalam).

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-1* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-1. Devanagari Vowel Letters

For	Use	Do Not Use
ऍ	0904	<0905, 0946>
आ	0906	<0905, 093E>
ई	0908	<0930, 094D, 0907>
ऊ	090A	<0909, 0941>
ऍ	090D	<090F, 0945>
ऍ	090E	<090F, 0946>
ऐ	0910	<090F, 0947>
ऑ	0911	<0905, 0949> or <0906, 0945>
ओ	0912	<0905, 094A> or <0906, 0946>
ओ	0913	<0905, 094B> or <0906, 0947>
औ	0914	<0905, 094C> or <0906, 0948>
ऍ	0972	<0905, 0945>
अं	0973	<0905, 093A>
आं	0974	<0905, 093B> or <0906, 093A>
औ	0975	<0905, 094F>
ु	0976	<0905, 0956>
ू	0977	<0905, 0957>

Virama (Halant). Devanagari employs a sign known in Sanskrit as the *virama* or vowel omission sign. In Hindi, it is called *hal* or *halant*, and that term is used in referring to the virama or to a consonant with its vowel suppressed by the virama. The terms are used interchangeably in this section.

A dead consonant should not be followed by an independent vowel letter in an attempt to create an alternative representation of a CV orthographic syllable. The ordinary representation of a CV orthographic syllable is with a (live) consonant followed by a dependent vowel.

The virama sign, U+094D DEVANAGARI SIGN VIRAMA, nominally serves to cancel (or kill) the inherent vowel of the consonant to which it is applied. When a consonant has lost its inherent vowel by the application of virama, it is known as a *dead consonant*; in contrast, a *live consonant* is one that retains its inherent vowel or is written with an explicit dependent

vowel sign. In the Unicode Standard, a dead consonant is defined as a sequence consisting of a consonant letter followed by a virama. The default rendering for a dead consonant is to position the virama as a combining mark bound to the consonant letterform.

For example, if C_n denotes the nominal form of consonant C, and C_d denotes the dead consonant form, then a dead consonant is encoded as shown in *Figure 12-1*.

Figure 12-1. Dead Consonants in Devanagari

$$TA_n + VIRAMA_n \rightarrow TA_d$$

त + ◌् → त्

Consonant Conjuncts. The Indic scripts are noted for a large number of consonant conjunct forms that serve as orthographic abbreviations (ligatures) of two or more adjacent letterforms. This abbreviation takes place only in the context of a *consonant cluster*. An orthographic consonant cluster is defined as a sequence of characters that represents one or more dead consonants (denoted C_d) followed by a normal, live consonant letter (denoted C_l).

Under normal circumstances, a consonant cluster is depicted with a conjunct glyph if such a glyph is available in the current font. In the absence of a conjunct glyph, the one or more dead consonants that form part of the cluster are depicted using half-form glyphs. In the absence of half-form glyphs, the dead consonants are depicted using the nominal consonant forms combined with visible virama signs (see *Figure 12-2*).

Figure 12-2. Conjunct Formations in Devanagari

(1) $GA_d + DHA_l \rightarrow GA_h + DHA_n$

ग् + ध → ग्ध

(2) $KA_d + KA_l \rightarrow K.KA_n$

क् + क → क्क

(3) $KA_d + SSA_l \rightarrow K.SSA_n$

क् + ष → क्ष

(4) $RA_d + KA_l \rightarrow KA_l + RA_{sup}$

र् + क → र्क

A number of types of conjunct formations appear in these examples: (1) a half-form of GA in its combination with the full form of DHA; (2) a vertical conjunct K.KA; and (3) a fully ligated conjunct K.SSA, in which the components are no longer distinct. In example (4) in

Figure 12-2, the dead consonant RA_d is depicted with the nonspacing combining mark RA_{sup} (*repha*).

A well-designed Indic script font may contain hundreds of conjunct glyphs, but they are not encoded as Unicode characters because they are the result of ligation of distinct letters. Indic script rendering software must be able to map appropriate combinations of characters in context to the appropriate conjunct glyphs in fonts.

Explicit Virama (Halant). Normally a virama character serves to create dead consonants that are, in turn, combined with subsequent consonants to form conjuncts. This behavior usually results in a virama sign not being depicted visually. Occasionally, this default behavior is not desired when a dead consonant should be excluded from conjunct formation, in which case the virama sign is visibly rendered. To accomplish this goal, the Unicode Standard adopts the convention of placing the character U+200C ZERO WIDTH NON-JOINER immediately after the encoded dead consonant that is to be excluded from conjunct formation. In this case, the virama sign is always depicted as appropriate for the consonant to which it is attached.

For example, in *Figure 12-3*, the use of ZERO WIDTH NON-JOINER prevents the default formation of the conjunct form क्ष (K.SSA_n).

Figure 12-3. Preventing Conjunct Forms in Devanagari

$$KA_d + ZWNJ + SSA_l \rightarrow KA_d + SSA_n$$

क् + ZWNJ + ष → क्ष

Explicit Half-Consonants. When a dead consonant participates in forming a conjunct, the dead consonant form is often absorbed into the conjunct form, such that it is no longer distinctly visible. In other contexts, the dead consonant may remain visible as a *half-consonant form*. In general, a half-consonant form is distinguished from the nominal consonant form by the loss of its inherent vowel stem, a vertical stem appearing to the right side of the consonant form. In other cases, the vertical stem remains but some part of its right-side geometry is missing.

In certain cases, it is desirable to prevent a dead consonant from assuming full conjunct formation yet still not appear with an explicit virama. In these cases, the half-form of the consonant is used. To explicitly encode a half-consonant form, the Unicode Standard adopts the convention of placing the character U+200D ZERO WIDTH JOINER immediately after the encoded dead consonant. The ZERO WIDTH JOINER denotes a nonvisible letter that presents linking or cursive joining behavior on either side (that is, to the previous or following letter). Therefore, in the present context, the ZERO WIDTH JOINER may be consid-

ered to present a context to which a preceding dead consonant may join so as to create the half-form of the consonant.

For example, if C_h denotes the half-form glyph of consonant C, then a half-consonant form is represented as shown in *Figure 12-4*.

Figure 12-4. Half-Consonants in Devanagari

$$KA_d + ZWJ + SSA_l \rightarrow KA_h + SSA_n$$

क् + [ZWJ] + ष → क्ष

In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 12-4* would normally produce the full conjunct form क्ष (K.SSA_n).

This encoding of half-consonant forms also applies in the absence of a base letterform. That is, this technique may be used to encode independent half-forms, as shown in *Figure 12-5*.

Figure 12-5. Independent Half-Forms in Devanagari

$$GA_d + ZWJ \rightarrow GA_h$$

ग् + [ZWJ] → ग

Other Indic scripts have similar half-forms for the initial consonants of a conjunct. Some, such as Oriya, also have similar half-forms for the final consonants; those are represented as shown in *Figure 12-6*.

Figure 12-6. Half-Consonants in Oriya

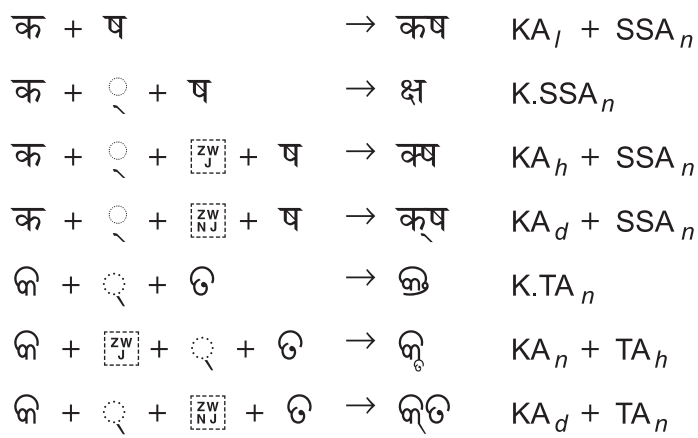
$$KA_n + ZWJ + VIRAMA + TA_l \rightarrow KA_l + TA_h$$

କ + [ZWJ] + ◌୍ + ତ → କ୍ତ

In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 12-6* would normally produce the full conjunct form କ୍ତ (K.TA_n).

Consonant Forms. In summary, each consonant may be encoded such that it denotes a live consonant, a dead consonant that may be absorbed into a conjunct, the half-form of a dead consonant, or a dead consonant with an overt halant that does not get absorbed into a conjunct (see *Figure 12-7*).

Figure 12-7. Consonant Forms in Devanagari and Oriya



As the rendering of conjuncts and half-forms depends on the availability of glyphs in the font, the following fallback strategy should be employed:

- If the coded character sequence would normally render with a full conjunct, but such a conjunct is not available, the fallback rendering is to use half-forms. If those are not available, the fallback rendering should use an explicit (visible) virama.
- If the coded character sequence would normally render with a half-form (it contains a ZWJ), but half-forms are not available, the fallback rendering should use an explicit (visible) virama.

Rendering Devanagari

Rules for Rendering. This section provides more formal and detailed rules for minimal rendering of Devanagari as part of a plain text sequence. It describes the mapping between Unicode characters and the glyphs in a Devanagari font. It also describes the combining and ordering of those glyphs.

These rules provide minimal requirements for legibly rendering interchanged Devanagari text. As with any script, a more complex procedure can add rendering characteristics, depending on the font and application.

> *In a font that is capable of rendering Devanagari, the number of glyphs is greater than the number of Devanagari characters.*

Notation. In the next set of rules, the following notation applies:

C_n	Nominal glyph form of consonant C as it appears in the code charts.
C_l	A live consonant, depicted identically to C_n.

C_d	Glyph depicting the dead consonant form of consonant C.
C_h	Glyph depicting the half-consonant form of consonant C.
L_n	Nominal glyph form of a conjunct ligature consisting of two or more component consonants. A conjunct ligature composed of two consonants X and Y is also denoted $X.Y_n$.
RA_{sup}	A nonspacing combining mark glyph form of U+0930 DEVANAGARI LETTER RA positioned above or attached to the upper part of a base glyph form. This form is also known as *repha*.
RA_{sub}	A nonspacing combining mark glyph form of U+0930 DEVANAGARI LETTER RA positioned below or attached to the lower part of a base glyph form.
V_{vs}	Glyph depicting the dependent vowel sign form of a vowel V.
$VIRAMA_n$	The nominal glyph form of the nonspacing combining mark depicting U+094D DEVANAGARI SIGN VIRAMA.

A virama character is not always depicted. When it is depicted, it adopts this nonspacing mark form.

Dead Consonant Rule. The following rule logically precedes the application of any other rule to form a dead consonant. Once formed, a dead consonant may be subject to other rules described next.

R1 ***When a consonant*** C_n ***precedes a*** $VIRAMA_n$***, it is considered to be a dead consonant*** C_d***. A consonant*** C_n ***that does not precede*** $VIRAMA_n$ ***is considered to be a live consonant*** C_l***.***

$$TA_n + VIRAMA_n \rightarrow TA_d$$

त + ◌् → त्

Consonant RA ***Rules.*** The character U+0930 DEVANAGARI LETTER RA takes one of a number of visual forms depending on its context in a consonant cluster. By default, this letter is depicted with its nominal glyph form (as shown in the code charts). In some contexts, it is depicted using one of two nonspacing glyph forms that combine with a base letterform.

R2 ***If the dead consonant*** RA_d ***precedes a consonant, then it is replaced by the superscript nonspacing mark*** RA_{sup} ***, which is positioned so that it applies to the logically subsequent element in the memory representation.***

$$RA_d + KA_l \rightarrow KA_l + RA_{sup}$$

Displayed Output

र् + क → क + ◌ं → र्क

$$RA^1_d + RA^2_d \rightarrow RA^2_d + RA^1_{sup}$$

र् + र् → र् + ◌ं → र्र

R3 ***If the superscript mark*** RA_{sup} ***is to be applied to a dead consonant and that dead consonant is combined with another consonant to form a conjunct ligature, then the mark is positioned so that it applies to the conjunct ligature form as a whole.***

$$RA_d + JA_d + NYA_l \rightarrow J.NYA_n + RA_{sup}$$

Displayed Output

र् + ज् + ञ → ज्ञ + ◌ं → ज्ञ्

R4 ***If the superscript mark*** RA_{sup} ***is to be applied to a dead consonant that is subsequently replaced by its half-consonant form, then the mark is positioned so that it applies to the form that serves as the base of the consonant cluster.***

$$RA_d + GA_d + GHA_l \rightarrow GA_h + GHA_l + RA_{sup}$$

Displayed Output

र् + ग् + घ → ग + घ + ◌ं → ग्घ्

R5 ***In conformance with the ISCII standard, the half-consonant form*** RRA_h ***is represented as eyelash-***RA***. This form of*** RA ***is commonly used in writing Marathi and Newari.***

$$RRA_n + VIRAMA_n \rightarrow RRA_h$$

र् + ◌् → ऱ

R5a ***For compatibility with*** **The Unicode Standard, Version 2.0,** ***if the dead consonant*** RA_d ***precedes*** ZERO WIDTH JOINER***, then the half-consonant form*** RA_h ***, depicted as eyelash-***RA***, is used instead of*** RA_{sup} ***.***

$$RA_d + ZWJ \rightarrow RA_h$$

र् + ZWJ → ऱ

R6 ***Except for the dead consonant*** RA_d ***, when a dead consonant*** C_d ***precedes the live consonant*** RA_l***, then*** C_d ***is replaced with its nominal form*** C_n***, and*** RA ***is replaced by the subscript nonspacing mark*** RA_{sub}***, which is positioned so that it applies to*** C_n***.***

$$TTHA_d + RA_l \rightarrow TTHA_n + RA_{sub}$$

Displayed Output

ठ् + र → ठ + ◌् → ठ्र

R7 ***For certain consonants, the mark*** RA_{sub} ***may graphically combine with the consonant to form a conjunct ligature form. These combinations, such as the one shown here, are further addressed by the ligature rules described shortly.***

$$PHA_d + RA_l \rightarrow PHA_n + RA_{sub}$$

Displayed Output

फ् + र → फ + ◌् → फ्र

R8 ***If a dead consonant (other than*** RA_d ***) precedes*** RA_d***, then the substitution of*** RA ***for*** RA_{sub} ***is performed as described above; however, the*** VIRAMA ***that formed*** RA_d ***remains so as to form a dead consonant conjunct form.***

$$TA_d + RA_d \rightarrow TA_n + RA_{sub} + VIRAMA_n \rightarrow T.RA_d$$

त् + र् → त + ◌् + ◌् → त्र

A dead consonant conjunct form that contains an absorbed RA_d ***may subsequently combine to form a multipart conjunct form.***

$$T.RA_d + YA_l \rightarrow T.R.YA_n$$

त्र् + य → त्र्य

Modifier Mark Rules. In addition to vowel signs, three other types of combining marks may be applied to a component of an orthographic syllable or to the syllable as a whole: *nukta*, *bindus*, and *svaras*.

R9 ***The nukta sign, which modifies a consonant form, is placed immediately after the consonant in the memory representation and is attached to that consonant in rendering. If the consonant represents a dead consonant, then*** NUKTA ***should precede*** VIRAMA ***in the memory representation.***

$$KA_n + NUKTA_n + VIRAMA_n \rightarrow QA_d$$

क + ◌़ + ◌् → क़्

R10 ***Other modifying marks, in particular bindus and svaras, apply to the orthographic syllable as a whole and should follow (in the memory representation) all other characters that constitute the syllable. The bindus should follow any vowel signs, and the svaras should come last. The relative placement of these marks is horizontal rather than vertical; the horizontal rendering order may vary according to typographic concerns.***

$$KA_n + AA_{vs} + CANDRABINDU_n$$

क + ा + ँ → काँ

Ligature Rules. Subsequent to the application of the rules just described, a set of rules governing ligature formation apply. The precise application of these rules depends on the availability of glyphs in the current font being used to display the text.

R11 ***If a dead consonant immediately precedes another dead consonant or a live consonant, then the first dead consonant may join the subsequent element to form a two-part conjunct ligature form.***

$JA_d + NYA_l \rightarrow J.NYA_n$ $\qquad$ $TTA_d + TTHA_l \rightarrow TT.TTHA_n$

ज् + ञ → ज्ञ $\qquad$ ट् + ठ → ट्ठ

R12 ***A conjunct ligature form can itself behave as a dead consonant and enter into further, more complex ligatures.***

$$SA_d + TA_d + RA_n \rightarrow SA_d + T.RA_n \rightarrow S.T.RA_n$$

स् + त् + र → स् + त्र → स्त्र

A conjunct ligature form can also produce a half-form.

$$K.SSA_d + YA_l \rightarrow K.SS_h + YA_n$$

क्ष् + य → क्ष्य

R13 ***If a nominal consonant or conjunct ligature form precedes*** RA_{sub} ***as a result of the application of rule R6, then the consonant or ligature form may join with*** RA_{sub} ***to form a multipart conjunct ligature (see rule R6 for more information).***

$KA_n + RA_{sub} \rightarrow K.RA_n$ $\qquad$ $PHA_n + RA_{sub} \rightarrow PH.RA_n$

क + ◌् र → क्र $\qquad$ फ + ◌् र → फ्र

R14 ***In some cases, other combining marks will combine with a base consonant, either attaching at a nonstandard location or changing shape. In minimal rendering, there are only two cases:*** RA_l ***with*** U_{vs} ***or*** UU_{vs} ***.***

$$RA_l + U_{vs} \rightarrow RU_n \qquad RA_l + UU_{vs} \rightarrow RUU_n$$

र + ◌ु → रु र + ◌ू → रू

Memory Representation and Rendering Order. The storage of plain text in Devanagari and all other Indic scripts generally follows phonetic order; that is, a CV syllable with a dependent vowel is always encoded as a consonant letter C followed by a vowel sign V in the memory representation. This order is employed by the ISCII standard and corresponds to both the phonetic order and the keying order of textual data (see *Figure 12-8*).

Figure 12-8. Rendering Order in Devanagari

Character Order Glyph Order

$$KA_n + I_{vs} \rightarrow I + KA_n$$

क + ◌ि → कि

Because Devanagari and other Indic scripts have some dependent vowels that must be depicted to the left side of their consonant letter, the software that renders the Indic scripts must be able to reorder elements in mapping from the logical (character) store to the presentational (glyph) rendering. For example, if C_n denotes the nominal form of consonant C, and V_{vs} denotes a left-side dependent vowel sign form of vowel V, then a reordering of glyphs with respect to encoded characters occurs as just shown.

R15 ***When the dependent vowel*** I_{vs} ***is used to override the inherent vowel of a syllable, it is always written to the extreme left of the orthographic syllable. If the orthographic syllable contains a consonant cluster, then this vowel is always depicted to the left of that cluster.***

$$TA_d + RA_l + I_{vs} \rightarrow T.RA_n + I_{vs} \rightarrow I_{vs} + T.RA_d$$

त् + र + ◌ि → त्र + ◌ि → त्रि

R16 ***The presence of an explicit virama (either caused by a ZWNJ or by the absence of a conjunct in the font) blocks this reordering, and the dependent vowel*** I_{vs} ***is rendered after the rightmost such explicit virama.***

$$TA_d + ZWNJ + RA_l + I_{vs} \rightarrow TA_d + I_{vs} + RA_l$$

त् + ZWNJ + र + ि → त्रि

Sample Half-Forms. *Table 12-2* shows examples of half-consonant forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. They may be encoded explicitly using ZERO WIDTH JOINER as shown. In normal conjunct formation, they may be used spontaneously to depict a dead consonant in combination with subsequent consonant forms.

Table 12-2. Sample Devanagari Half-Forms

क	+	◌्	+	ZWJ	→	क्	न	+	◌्	+	ZWJ	→	न्
ख	+	◌्	+	ZWJ	→	ख्	प	+	◌्	+	ZWJ	→	प्
ग	+	◌्	+	ZWJ	→	ग्	फ	+	◌्	+	ZWJ	→	फ्
घ	+	◌्	+	ZWJ	→	घ्	ब	+	◌्	+	ZWJ	→	ब्
च	+	◌्	+	ZWJ	→	च्	भ	+	◌्	+	ZWJ	→	भ्
ज	+	◌्	+	ZWJ	→	ज्	म	+	◌्	+	ZWJ	→	म्
झ	+	◌्	+	ZWJ	→	झ्	य	+	◌्	+	ZWJ	→	य्
ञ	+	◌्	+	ZWJ	→	ञ्	ल	+	◌्	+	ZWJ	→	ल्
ण	+	◌्	+	ZWJ	→	ण्	व	+	◌्	+	ZWJ	→	व्
त	+	◌्	+	ZWJ	→	त्	श	+	◌्	+	ZWJ	→	श्
थ	+	◌्	+	ZWJ	→	थ्	ष	+	◌्	+	ZWJ	→	ष्
ध	+	◌्	+	ZWJ	→	ध्	स	+	◌्	+	ZWJ	→	स्

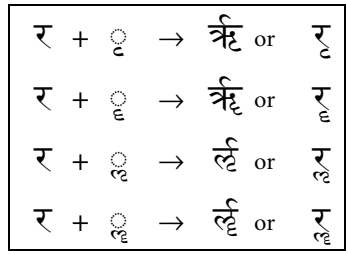
Sample Ligatures. *Table 12-3* shows examples of conjunct ligature forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. Not every writing system that employs this script uses all of these forms; in particular, many of these forms are used only in writing Sanskrit texts. Furthermore, individual fonts may provide fewer or more ligature forms than are depicted here.

Table 12-3. Sample Devanagari Ligatures

क	+	◌्	+	क	→	क्क	ट	+	◌्	+	ठ	→	ट्ठ
क	+	◌्	+	त	→	क्त	ठ	+	◌्	+	ठ	→	ठ्ठ
क	+	◌्	+	र	→	क्र	ड	+	◌्	+	ग	→	ड्ग
क	+	◌्	+	ष	→	क्ष	ड	+	◌्	+	ड	→	ड्ड
ङ	+	◌्	+	क	→	ङ्क	ड	+	◌्	+	ढ	→	ड्ढ
ङ	+	◌्	+	ख	→	ङ्ख	त	+	◌्	+	त	→	त्त
ङ	+	◌्	+	ग	→	ङ्ग	त	+	◌्	+	र	→	त्र
ङ	+	◌्	+	घ	→	ङ्घ	न	+	◌्	+	न	→	न्न
ञ	+	◌्	+	ज	→	ञ्ज	फ	+	◌्	+	र	→	फ्र
ज	+	◌्	+	ञ	→	ज्ञ	श	+	◌्	+	र	→	श्र
द	+	◌्	+	घ	→	द्ध	ह	+	◌्	+	म	→	ह्म
द	+	◌्	+	द	→	द्द	ह	+	◌्	+	य	→	ह्य
द	+	◌्	+	ध	→	द्ध	ह	+	◌्	+	ल	→	ह्ल
द	+	◌्	+	ब	→	द्ब	ह	+	◌्	+	व	→	ह्व
द	+	◌्	+	भ	→	द्भ	ह	+	◌ृ			→	ह्र
द	+	◌्	+	म	→	द्म	र	+	◌ु			→	रु
द	+	◌्	+	य	→	द्य	र	+	◌ू			→	रू
द	+	◌्	+	व	→	द्व	स	+	◌्	+	त्र	→	स्त्र
ट	+	◌्	+	ट	→	ट्ट							

Ligature Forms for Ra + Vocalic Liquids. The phonological sequence /r vocalic_r/ can graphically appear either as RA_{l} with a vowel sign for the vocalic_r, or as VOCALIC R with a superscript mark RA_{sup}. Similarly, the phonological sequences with the other vocalic sounds (*rr*, *l*, *ll*) have two written forms, as shown in *Table 12-4*.

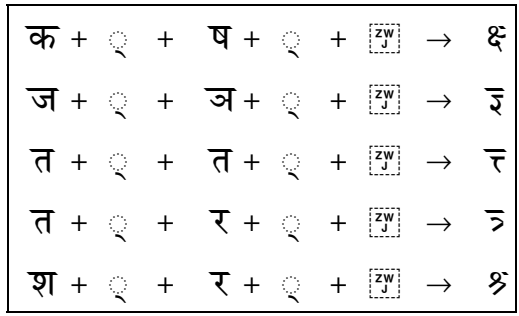
Table 12-4. RA + Vocalic Letter Ligature Forms



The graphical forms displayed above with the reph (RA_{sup}) should not be represented by sequences of RA + virama + independent vowel, as such sequences violate the general encoding principles of the script. CV orthographic syllables are not represented by consonant + virama + independent vowel.

Sample Half-Ligature Forms. In addition to half-form glyphs of individual consonants, half-forms are used to depict conjunct ligature forms. A sample of such forms is shown in *Table 12-5*. These forms are glyphs, not characters. They may be encoded explicitly using ZERO WIDTH JOINER as shown. In normal conjunct formation, they may be used spontaneously to depict a conjunct ligature in combination with subsequent consonant forms.

Table 12-5. Sample Devanagari Half-Ligature Forms



Language-Specific Allographs. In Marathi, Nepali, and some South Indian orthographies, variant glyphs are preferred for certain letters and digits. These include U+091D DEVANAGARI LETTER JHA, U+0932 DEVANAGARI LETTER LA, U+0936 DEVANAGARI LETTER SHA, and the digits five, eight, and nine, as shown in *Table 12-6*. Marathi also makes use of the "eyelash" form of the letter RA, as discussed in rule R5.

Table 12-6. Marathi and Nepali Allographs

Code Point	Hindi	Marathi	Nepali
U+091D JHA	झ	झ	झ
U+0932 LA	ल	ल	ल
U+0936 SHA	श	श	श
U+096B FIVE	५	५	५
U+096E EIGHT	८	८	८
U+096F NINE	९	९	९

In addition, various languages written in Devanagari (or sometimes their various orthographic traditions) tend to have different preferences for formation of certain ligatures (see the text on "Sample Ligatures," earlier in this section). For example, modern Nepali orthographies prefer a smaller number of ligatures than commonly used in Hindi or Marathi.

Combining Marks. Devanagari and other Indic scripts have a number of combining marks that could be considered diacritic. One class of these marks, known as bindus, is represented by U+0901 DEVANAGARI SIGN CANDRABINDU and U+0902 DEVANAGARI SIGN ANUSVARA. These marks 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 combining marks used in transcription of Sanskrit texts.

Devanagari Digits, Punctuation, and Symbols

Digits. Each Indic script has a distinct set of digits appropriate to that script. These digits may or may not be used in ordinary text in that script. European digits have displaced the Indic script forms in modern usage in many of the scripts. Some Indic scripts—notably Tamil—lacked a distinct digit for zero in their traditional numerical systems, but adopted a zero based on general Indian practice.

Punctuation. U+0964 । DEVANAGARI DANDA is similar to a full stop. U+0965 ॥ DEVANAGARI DOUBLE DANDA marks the end of a verse in traditional texts. The term *danda* is from Sanskrit, and the punctuation mark is generally referred to as a *viram* instead in Hindi. Although the *danda* and *double danda* are encoded in the Devanagari block, the intent is that they be used as common punctuation for all the major scripts of India covered by this chapter. *Danda* and *double danda* punctuation marks are not separately encoded for Bengali, Gujarati, and so on. However, analogous punctuation marks for other Brahmi-derived scripts *are* separately encoded, particularly for scripts used primarily outside of India.

Many modern languages written in the Devanagari script intersperse punctuation derived from the Latin script. Thus U+002C COMMA and U+002E FULL STOP are freely used in writing Hindi, and the *danda* is usually restricted to more traditional texts. However, the *danda* may be preserved when such traditional texts are transliterated into the Latin script.

Other Symbols. U+0970 ॰ DEVANAGARI ABBREVIATION SIGN appears after letters or combinations of letters and marks the sequence as an abbreviation. It is intended specifically for Devanagari script-based abbreviations, such as the Devanagari rupee sign. Other symbols and signs most commonly occurring in Vedic texts are encoded in the Devanagari Extended and Vedic Extensions blocks and are discussed in the text that follows.

The *svasti* (or well-being) signs often associated with the Hindu, Buddhist, and Jain traditions are encoded in the Tibetan block. See *Section 13.3, Tibetan* for further information.

Extensions in the Main Devanagari Block

Sindhi Letters. The characters U+097B DEVANAGARI LETTER GGA, U+097C DEVANAGARI LETTER JJA, U+097E DEVANAGARI LETTER DDDA, and U+097F DEVANAGARI LETTER BBA are used to write Sindhi implosive consonants. Previous versions of the Unicode Standard recommended representing those characters as a combination of the usual consonants with *nukta* and *anudātta*, but those combinations are no longer recommended.

Konkani. Konkani makes use of additional sounds that can be represented with combinations such as U+091A DEVANAGARI LETTER CA plus U+093C DEVANAGARI SIGN NUKTA and U+091F DEVANAGARI LETTER TTA plus U+0949 DEVANAGARI VOWEL SIGN CANDRA O.

Bodo, Dogri, and Maithili. The orthographies of the Bodo, Dogri, and Maithili languages of India make use of U+02BC " ʼ " MODIFIER LETTER APOSTROPHE, either as a tone mark or as a length mark. In Bodo and Dogri, this character functions as a tone mark, called *gojau kamaa* in Bodo and *sur chinha* in Dogri. In Dogri, the tone mark occurs after short vowels, including inherent vowels, and indicates a high-falling tone. After Dogri long vowels, a high-falling tone is written instead using U+0939 DEVANAGARI LETTER HA.

In Maithili, U+02BC " ʼ " MODIFIER LETTER APOSTROPHE is used to indicate the prolongation of a short *a* and to indicate the truncation of words. This sign is called *bikari kaamaa*.

Examples illustrating the use of U+02BC " ʼ " MODIFIER LETTER APOSTROPHE in Bodo, Dogri, and Maithili are shown in *Figure 12-9*. The Maithili examples show the same sentence, first in full form, and then using U+02BC to show truncation of words.

Figure 12-9. Use of Apostrophe in Bodo, Dogri and Maithili

Language	Examples	Meaning
Bodo	खर’ दख’ना	head type of Bodo dress
Dogri	ख’ल्ल ति’लकना	down to slip
Maithili	कतए पड़ाए गेलह? कत’ पड़ा’ गेल’?	} Where did you go away?

In both Dogri and Maithili, an *avagraha sign*, U+093D DEVANAGARI SIGN AVAGRAHA, is used to indicate extra-long vowels. An example of the contrastive use of this *avagraha sign* is shown for Dogri in *Figure 12-10*.

Figure 12-10. Use of Avagraha in Dogri

Example	Meaning
तला	sole
तलाऽ	pond

Kashmiri Letters. There are several letters for use with Kashmiri when written in Devanagari script. Long and short versions of the independent vowel letters are encoded in the range U+0973..U+0977. The corresponding dependent vowel signs are U+093A DEVANAGARI VOWEL SIGN OE, U+093B DEVANAGARI VOWEL SIGN OOE, and U+094F DEVANAGARI VOWEL SIGN AW. The forms of the independent vowels for Kashmiri are constructed by using the glyphs of the matras U+093B DEVANAGARI VOWEL SIGN OOE, U+094F DEVANAGARI VOWEL SIGN AW, U+0956 DEVANAGARI VOWEL SIGN UE, and U+0957 DEVANAGARI VOWEL SIGN UUE as diacritics on U+0905 DEVANAGARI LETTER A. However, for representation of independent vowels in Kashmiri, use the encoded, composite characters in the range U+0973..U+0977 and not the visually equivalent sequences of U+0905 DEVANAGARI LETTER A plus the matras. See *Table 12-1*. A few of the letters identified as being used for Kashmiri are also used to write the Bihari languages.

Letters for Bihari Languages. A number of the Devanagari vowel letters have been used to write the Bihari languages Bhojpuri, Magadhi, and Maithili, as listed in *Table 12-7*.

Table 12-7. Devanagari Vowels Used in Bihari Languages

U+090E	DEVANAGARI LETTER SHORT E
U+0912	DEVANAGARI LETTER SHORT O
U+0946	DEVANAGARI VOWEL SIGN SHORT E
U+094A	DEVANAGARI VOWEL SIGN SHORT O
U+0973	DEVANAGARI LETTER OE
U+0974	DEVANAGARI LETTER OOE
U+0975	DEVANAGARI LETTER AW
U+093A	DEVANAGARI VOWEL SIGN OE
U+093B	DEVANAGARI VOWEL SIGN OOE
U+094F	DEVANAGARI VOWEL SIGN AW

Prishthamatra Orthography. In the historic Prishthamatra orthography, the vowel signs for *e*, *ai*, *o*, and *au* are represented using U+094E DEVANAGARI VOWEL SIGN PRISHTHAMATRA E (which goes on the left side of the consonant) alone or in combination with one of U+0947 DEVANAGARI VOWEL SIGN E, U+093E DEVANAGARI VOWEL SIGN AA or U+094B DEVANAGARI VOWEL SIGN O. *Table 12-8* shows those combinations applied to *ka*. In the

underlying representation of text, U+094E should be first in the sequence of dependent vowel signs after the consonant, and may be followed by U+0947, U+093E or U+094B.

Table 12-8. Prishthamatra Orthography

	Prishthamatra Orthography	Modern Orthography
ke	कि <0915, 094E>	के <0915, 0947>
kai	कि <0915, 094E, 0947>	कै <0915, 0948>
ko	कि <0915, 094E, 093E>	को <0915, 094B>
kau	कि <0915, 094E, 094B>	कौ <0915, 094C>

Devanagari Extended: U+A8E0–U+A8FF

This block of characters is used chiefly for Vedic Sanskrit, although many of the characters are generic and can be used by other Indic scripts. The block includes a set of combining digits, letters, and *avagraha* which is used as a system of cantillation marks in the early Vedic Sanskrit texts. The Devanagari Extended block also includes nasalization marks (*candrabindu*), and a number of editorial marks.

The Devanagari Extended block, as well as the Vedic Extensions block and the Devanagari block, include characters that are used to indicate tone in Vedic Sanskrit. Indian linguists describe tone as a feature of vowels, shared by the consonants in the same syllable, or as a feature of syllables. In Vedic, vowels are marked for tone, as are certain non-vocalic characters that are syllabified in Vedic recitation (*visarga* and *anusvāra*); the tone marks directly follow the vowel or other character that they modify. Vowels are categorized according to tone as either *udātta* (high-toned or "acute"), *anudātta* (low-toned or "non-acute"), *svarita* ("modulated" or dropping from high to low tone) or *ekaśruti* (monotone). Some of the symbols used for marking tone indicate different tones in different traditions. *Visarga* may be marked for all three tones. The tone marks also can indicate other modifications of vocal text, such as vibration, lengthening a vowel, or skipping a tone in a descending scale.

Cantillation marks are used to indicate length, tone, and other features in the recited text of *Sāmaveda*, and in the Kauthuma and Rāṇāyanīya traditions of *Sāmagāna*. These marks are encoded as a series of combining digits, alphabetic characters, and *avagraha* in the range U+A8E0..U+A8F1.

Cantillation Marks for the Sāmaveda. One of the four major Vedic texts is *Sāmaveda*. The text is both recited (*Sāmaveda-Saṁhitā*) and sung (*Sāmagāna*), and is marked differently for the purposes of each. Cantillation marks are used to indicate length, tone, and other features in the recited text of *Sāmaveda*, and in the Kauthuma and Rāṇāyanīya traditions of *Sāmagāna*. These marks are encoded as a series of combining digits, alphabetic characters, and *avagraha* in the range U+A8E0..U+A8F1. The marks are rendered directly over the base letter. They are represented in text immediately after the syllable they modify.

In certain cases, two marks may occur over a letter: U+A8E3 COMBINING DEVANAGARI DIGIT THREE may be followed by U+A8EC COMBINING DEVANAGARI LETTER KA, for example. Although no use of U+A8E8 COMBINING DEVANAGARI DIGIT EIGHT has been found in the *Sāmagāna*, it is included to provide a complete set of 0-9 digits. The combining marks encoded for the *Sāmaveda* do not include characters that may appear as subscripts and superscripts in the Jaiminīya tradition of *Sāmagāna*, which used interlinear annotation. Interlinear annotation may be rendered using Ruby and may be represented by means of markup or other higher-level protocols.

Nasalization Marks. The set of spacing marks in the range U+A8F2..U+A8F7 include the term *candrabindu* in their names and indicate nasalization. These marks are all aligned with the headline. Note that U+A8F2 DEVANAGARI SIGN SPACING CANDRABINDU is lower than the U+0901 DEVANAGARI SIGN CANDRABINDU.

Editorial Marks. A set of editorial marks is encoded in the range U+A8F8..U+A8FB for use with Devanagari. U+A8F9 DEVANAGARI GAP FILLER signifies an intentional gap that would ordinarily be filled with text. In contrast, U+A8FB DEVANAGARI HEADSTROKE indicates illegible gaps in the original text. The glyph for DEVANAGARI HEADSTROKE should be designed so that it does not connect to the headstroke of the letters beside it, which will make it possible to indicate the number of illegible syllables in a given space. U+A8F8 DEVANAGARI SIGN PUSHPIKA acts as a filler in text, and is commonly flanked by double dandas. U+A8FA DEVANAGARI CARET, a zero-width spacing character, marks the insertion point of omitted text, and is placed at the insertion point between two orthographic syllables. It can also be used to indicate word division.

Vedic Extensions: U+1CD0–U+1CFF

The Vedic Extensions block includes characters that are used in Vedic texts; they may be used with Devanagari, as well as many other Indic scripts. This block includes a set of characters designating tone, grouped by the various Vedic traditions in which they occur. Characters indicating tone marks directly follow the character they modify. Most of these marks indicate the tone of vowels, but three of them specifically indicate the tone of *visarga*. Nasal characters are also included in the block. U+1CD3 VEDIC SIGN NIHSHVASA indicates where a breath may be taken. Finally, the block includes U+1CF2 VEDIC SIGN ARDHAVISARGA.

Tone Marks. The Vedic tone marks are all combining marks. The tone marks are grouped together in the code charts based upon the tradition in which they appear: they are used in the four core texts of the Vedas (*Sāmaveda*, *Yajurveda*, *Rigveda*, and *Atharvaveda*) and in the prose text on Vedic ritual (*Śatapathabrāhmaṇa*). The character U+1CD8 VEDIC TONE CANDRA BELOW is also used to identify the short vowels *e* and *o*. In this usage, the prescribed order is the Indic syllable (*aksara*), followed by U+1CD8 VEDIC TONE CANDRA BELOW and the tone mark (*svara*). When a tone mark is placed below, it appears below the VEDIC TONE CANDRA BELOW.

In addition to the marks encoded in this block, Vedic texts may use other nonspacing marks from the General Diacritics block and other blocks. For example, U+20F0 COMBINING ASTERISK ABOVE would be used to represent a mark of that shape above a Vedic letter.

Diacritics for the Visarga. A set of combining marks that serve as diacritics for the *visarga* is encoded in the range U+1CE2..U+1CE8. These marks indicate that the *visarga* has a particular tone. For example, the combination U+0903 DEVANAGARI SIGN VISARGA plus U+1CE2 VEDIC SIGN VISARGA SVARITA represents a *svarita visarga*. The upward-shaped diacritic is used for the *udātta* (high-toned), the downward-shaped diacritic for *anudātta* (low-toned), and the midline glyph indicates the *svarita* (modulated tone).

In Vedic manuscripts the tonal mark (that is, the horizontal bar, upward curve and downward curve) appears in colored ink, while the two dots of the *visarga* appear in black ink. The characters for accents can be represented using separate characters, to make it easier for color information to be maintained by means of markup or other higher-level protocols.

Nasalization Marks. A set of spacing marks and one combining mark, U+1CED VEDIC SIGN TIRYAK, are encoded in the range U+1CE9..U+1CF1. They describe phonetic distinctions in the articulation of nasals. The *gomukha* characters from U+1CE9..U+1CEC may be combined with U+0902 DEVANAGARI SIGN ANUSVARA or U+0901 DEVANAGARI SIGN CANDRABINDU. U+1CF1 VEDIC SIGN ANUSVARA UBHAYATO MUKHA may indicate a *visarga* with a tonal mark as well as a nasal. The three characters, U+1CEE VEDIC SIGN HEXIFORM LONG ANUSVARA, U+1CEF VEDIC SIGN LONG ANUSVARA, and U+1CF0 VEDIC SIGN RTHANG LONG ANUSVARA, are all synonymous and indicate a long *anusvāra* after a short vowel. U+1CED VEDIC SIGN TIRYAK is the only combining character in this set of nasalization marks. While it appears similar to the U+094D DEVANAGARI SIGN VIRAMA, it is used to render glyph variants of nasal marks that occur in manuscripts and printed texts.

Ardhavisarga. U+1CF2 VEDIC SIGN ARDHAVISARGA is a character that marks either the *jihvāmūlīya*, a velar fricative occurring only before the unvoiced velar stops *ka* and *kha*, or the *upadhmānīya*, a bilabial fricative occurring only before the unvoiced labial stops *pa* and *pha*. *Ardhavisarga* is a spacing character. It is represented in text in visual order before the consonant it modifies.

12.2 Bengali (Bangla)

Bengali: U+0980–U+09FF

Scripts encoded in the Unicode Standard often are used to write many languages. The script termed *Bengali* in Unicode is no exception. It is used for writing languages such as Bengali, Assamese, Bishnupriya Manipuri, Daphla, Garo, Hallam, Khasi, Mizo, Munda, Naga, Rian, and Santali. In the Indian state of West Bengal and the People's Republic of Bangladesh, the preferred name for the Bengali script and language is *Bangla*. In the Indian state of Assam, the preferred name for the script is *Asamiya* or *Assamese*. Although the Assamese language has been written historically using regional scripts, known generally as "Kamrupi," its modern writing system is similar to that presently used for Bengali, with the addition of extra characters. The Unicode Bengali block fully supports modern Assamese orthography.

The Bengali script is a North Indian script closely related to Devanagari.

Virama (Hasant). The Bengali script uses the Unicode virama model to form conjunct consonants. In Bengali, the virama is known as *hasant*.

Vowel Letters. Vowel letters of Indic scripts are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-9* shows the Bengali vowel letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-9. Bengali Vowel Letters

For	Use	Do Not Use
আ	0986	<0985, 09BE>
ৠ	09E0	<098B, 09C3>
ৡ	09E1	<098C, 09E2>

There is an exception to this general pattern for the representation of Bengali independent vowel letters, for the Bengali script orthography of Kokborok, a major language of Tripura state in Northeast India. Kokborok has diphthongs which can occur as initial letters. To reflect existing practice, these diphthongs are represented with two character sequences, rather than as atomic characters, as shown in *Table 12-10*. Rendering systems which support display of the Kokborok orthography need to be aware of these exceptional sequences. The sequence for *vowel letter aw* uses U+09D7 BENGALI AU LENGTH MARK, also noted in the following discussion of two-part vowel signs.

Two-Part Vowel Signs. The Bengali script, along with a number of other Indic scripts, makes use of two-part dependent vowel signs. In these dependent vowels (*matras*) one-half of the vowel is displayed on each side of a consonant letter or cluster—for example,

Table 12-10. Diphthong Vowel Letters in Kokborok

For	Use	Description
অৌ	<0985, 09D7>	*vowel letter aw*
উা	<0989, 09BE>	*vowel letter ua*

U+09CB BENGALI VOWEL SIGN O and U+09CC BENGALI VOWEL SIGN AU. To provide compatibility with existing implementations of the scripts that use two-part vowel signs, the Unicode Standard explicitly encodes the right half of these vowel signs. For example, U+09D7 BENGALI AU LENGTH MARK represents the right-half glyph component of U+09CC BENGALI VOWEL SIGN AU. In Bengali orthography, the *au length mark* is always used in conjunction with the left part and does not have a meaning on its own.

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

Historic Characters. The characters *vocalic rr*, *vocalic l* and *vocalic ll*, both in their independent and dependent forms (U+098C, U+09C4, U+09E0..U+09E3), are only used to write Sanskrit words in the Bengali script.

Characters for Assamese. Assamese employs two letters not used for the Bengali language. The Assamese letter *ra* is represented in Unicode by U+09F0 ৰ BENGALI LETTER RA WITH MIDDLE DIAGONAL, and the Assamese letter *wa* is represented by U+09F1 ৱ BENGALI LETTER RA WITH LOWER DIAGONAL.

Assamese uses a conjunct character called *kssa.* Although *kssa* is often considered a separate letter of the alphabet, it is not separately encoded. The conjunct is represented by the sequence <U+0995 ক BENGALI LETTER KA, U+09CD ◌্ BENGALI SIGN VIRAMA, U+09B7 ষ BENGALI LETTER SSA>. This same sequence is also used to represent the Bengali letter *khinya* (or *khiya*).

Assamese uses two additional consonant-vowel ligatures formed with U+09F0 BENGALI LETTER RA WITH MIDDLE DIAGONAL, which are not used for the Bengali language. These consonant-vowel ligatures are shown in the "ligated" column in *Table 12-11*.

Table 12-11. Assamese Consonant-Vowel Combinations

	Code Points	Ligated	Non-ligated
ru	<09F0, 09C1>	ক্ৰু	ক্ৰ◌ু
rū	<09F0, 09C2>	ক্ৰূ	ক্ৰ◌ূ

Rendering Behavior. Like other Brahmic scripts in the Unicode Standard, Bengali uses the *hasant* to form conjunct characters. For example, U+09B8 স BENGALI LETTER SA + U+09CD ◌্ BENGALI SIGN VIRAMA + U+0995 ক BENGALI LETTER KA yields the conjunct স্ক

SKA. For general principles regarding the rendering of the Bengali script, see the rules for rendering in *Section 12.1, Devanagari*.

Consonant-Vowel Ligatures. Some Bengali consonant plus vowel combinations have two distinct visual presentations. The first visual presentation is a traditional ligated form, in which the vowel combines with the consonant in a novel way. In the second presentation, the vowel is joined to the consonant but retains its nominal form, and the combination is not considered a ligature. These consonant-vowel combinations are illustrated in *Table 12-12*.

Table 12-12. Bengali Consonant-Vowel Combinations

	Code Points	Ligated	Non-ligated
gu	<0997, 09C1>	গু	গু
ru	<09B0, 09C1>	ক্রু	ক্রু
rū	<09B0, 09C2>	ক্রূ	ক্রূ
śu	<09B6, 09C1>	শু	শু
hu	<09B9, 09C1>	হু	হু
hṛ	<09B9, 09C3>	হৃ	হৃ

The ligature forms of these consonant-vowel combinations are traditional. They are used in handwriting and some printing. The "non-ligated" forms are more common; they are used in newspapers and are associated with modern typefaces. However, the traditional ligatures are preferred in some contexts.

No semantic distinctions are made in Bengali text on the basis of the two different presentations of these consonant-vowel combinations. However, some users consider it important that implementations support both forms and that the distinction be representable in plain text. This may be accomplished by using U+200D ZERO WIDTH JOINER and U+200C ZERO WIDTH NON-JOINER to influence ligature glyph selection. (See "Cursive Connection and Ligatures" in *Section 23.2, Layout Controls*.) Joiners are rarely needed in this situation. The rendered appearance will typically be the result of a font choice.

A given font implementation can choose whether to treat the ligature forms of the consonant-vowel combinations as the defaults for rendering. If the non-ligated form is the default, then ZWJ can be inserted to request a ligature, as shown in *Figure 12-11*.

If the ligated form is the default for a given font implementation, then ZWNJ can be inserted to block a ligature, as shown in *Figure 12-12*.

Khiya. The letter ক্ষ, known as *khiya* or *khinya*, is often considered as a distinct letter of the Bengla alphabet. However, it is not encoded separately. It is represented by the sequence <U+0995 ক BENGALI LETTER KA, U+09CD ◌্ BENGALI SIGN VIRAMA, U+09B7 ষ BENGALI LETTER SSA>.

Figure 12-11. Requesting Bengali Consonant-Vowel Ligature

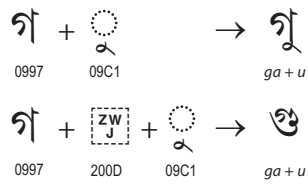
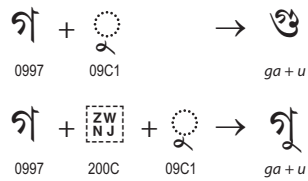


Figure 12-12. Blocking Bengali Consonant-Vowel Ligature



Khanda Ta. In Bengali, a dead consonant *ta* makes use of a special form, U+09CE BENGALI LETTER KHANDA TA. This form is used in all contexts except where it is immediately followed by one of the consonants: *ta*, *tha*, *na*, *ba*, *ma*, *ya*, or *ra*.

Khanda ta cannot bear a vowel matra or combine with a following consonant to form a conjunct *aksara*. It can form a conjunct *aksara* only with a preceding dead consonant *ra*, with the latter being displayed with a *repha* glyph placed on the *khanda ta*.

Versions of the Unicode Standard prior to Version 4.1 recommended that *khanda ta* be represented as the sequence <U+09A4 BENGALI LETTER TA, U+09CD BENGALI SIGN VIRAMA, U+200D ZERO WIDTH JOINER> in all circumstances. U+09CE BENGALI LETTER KHANDA TA should instead be used explicitly in newly generated text, but users are cautioned that instances of the older representation may exist.

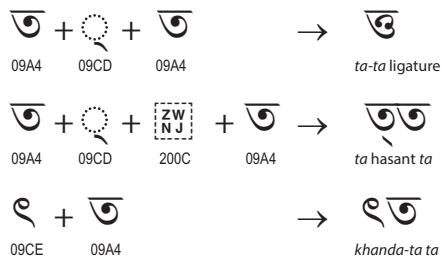
The Bengali syllable *tta* illustrates the usage of *khanda ta* when followed by *ta*. The syllable *tta* is normally represented with the sequence <U+09A4 *ta*, U+09CD *hasant*, U+09A4 *ta*>. That sequence will normally be displayed using a single glyph *tta* ligature, as shown in the first example in *Figure 12-13*.

It is also possible for the sequence <*ta*, *hasant*, *ta*> to be displayed with a full *ta* glyph combined with a *hasant* glyph, followed by another full *ta* glyph ত্ত. The choice of form actually displayed depends on the display engine, based on the availability of glyphs in the font.

The Unicode Standard also provides an explicit way to show the *hasant* glyph. To do so, a ZERO WIDTH NON-JOINER is inserted after the *hasant*. That sequence is always displayed with the explicit *hasant*, as shown in the second example in *Figure 12-13*.

When the syllable *tta* is written with a *khanda ta*, however, the character U+09CE BENGALI LETTER KHANDA TA is used and no *hasant* is required, as *khanda ta* is already a dead consonant. The rendering of *khanda ta* is illustrated in the third example in *Figure 12-13*.

Figure 12-13. Bengali Syllable *tta*



Ya-phalaa. *Ya-phalaa* is a presentation form of U+09AF য BENGALI LETTER YA. Represented by the sequence <U+09CD ◌্ BENGALI SIGN VIRAMA, U+09AF য BENGALI LETTER YA>, *ya-phalaa* has a special form ্য. When combined with U+09BE া BENGALI VOWEL SIGN AA, it is used for transcribing [æ] as in the "a" in the English word "bat." The *ya-phalaa* appears in র‍্যাশ [ræʃ] "rash," which provides a minimal pair with রাশ [raʃ] "a whole lot."

Ya-phalaa can be applied to initial vowels as well:

অ্যা = <0985, 09CD, 09AF, 09BE> (*a- hasant ya -aa*)

এ্যা = <098F, 09CD, 09AF, 09BE> (*e- hasant ya -aa*)

If a candrabindu or other combining mark needs to be added in the sequence, it comes at the end of the sequence. For example:

অ্যাঁ = <0985, 09CD, 09AF, 09BE, 0981> (*a- hasant ya -aa candrabindu*)

Further examples:

অ + ◌্ + য + া → অ্যা

এ + ◌্ + য + া → এ্যা

ত + ◌্ + য + া → ত্যা

Interaction of Repha and Ya-phalaa. The formation of the *repha* form is defined in *Section 12.1, Devanagari*, "Rules for Rendering," R2. Basically, the *repha* is formed when a *ra* that has the inherent vowel killed by the *hasant* begins a syllable. This scenario is shown in the following example:

র + ◌্ + ম → র্ম as in কর্ম (karma)

The *ya-phalaa* is a post-base form of *ya* and is formed when the *ya* is the final consonant of a syllable cluster. In this case, the previous consonant retains its base shape and the *hasant* is combined with the following *ya*. This scenario is shown in the following example:

ক + ◌্ + য → ক্য as in বাক্য (bakyô)

An ambiguous situation is encountered when the combination of *ra* + *hasant* + *ya* is encountered:

র + ◌্ + য → র্য or র‍্য

To resolve the ambiguity with this combination, the Unicode Standard adopts the convention of placing the character U+200D ZERO WIDTH JOINER immediately after the *ra* to obtain the *ya-phalaa*. The *repha* form is rendered when no ZWJ is present, as shown in the following example:

র + ◌্ + য → র্য

09B0 09CD 09AF

র + [ZWJ] + ◌্ + য → র‍্য

09B0 200D 09CD 09AF

When the first character of the cluster is not a *ra*, the *ya-phalaa* is the normal rendering of a *ya*, and a ZWJ is not necessary but can be present. Such a convention would make it possible, for example, for input methods to consistently associate *ya-phalaa* with the sequence <ZWJ, *hasant*, *ya*>.

Punctuation. Bengali uses punctuation marks shared across many Indic scripts, including the *danda* and *double danda* marks. In Bangla these are called the *dahri* and *double dahri*. For a description of these common punctuation marks, see *Section 12.1, Devanagari*.

Truncation. The orthography of the Bangla language makes use of U+02BC " ' " MODIFIER LETTER APOSTROPHE to indicate the truncation of words. This sign is called *urdha-comma*. Examples illustrating the use of U+02BC " ' " MODIFIER LETTER APOSTROPHE are shown in *Table 12-13*.

Table 12-13. Use of Apostrophe in Bangla

Example	Meaning
ক'রে	after, on doing (something)
'পরে ওপরে	} above

12.3 Gurmukhi

Gurmukhi: U+0A00–U+0A7F

The Gurmukhi script is a North Indian script used to write the Punjabi (or Panjabi) language of the Punjab state of India. Gurmukhi, which literally means "proceeding from the mouth of the Guru," is attributed to Angad, the second Sikh Guru (1504–1552 CE). It is derived from an older script called Landa and is closely related to Devanagari structurally. The script is closely associated with Sikhs and Sikhism, but it is used on an everyday basis in East Punjab. (West Punjab, now in Pakistan, uses the Arabic script.)

Encoding Principles. The Gurmukhi block is based on ISCII-1988, which makes it parallel to Devanagari. Gurmukhi, however, has a number of peculiarities described here.

The additional consonants (called *pairin bindi*; literally, "with a dot in the foot," in Punjabi) are primarily used to differentiate Urdu or Persian loan words. They include U+0A36 GURMUKHI LETTER SHA and U+0A33 GURMUKHI LETTER LLA, but do not include U+0A5C GURMUKHI LETTER RRA, which is genuinely Punjabi. For unification with the other scripts, ISCII-1991 considers *rra* to be equivalent to *dda+nukta*, but this decomposition is not considered in Unicode. At the same time, ISCII-1991 does not consider U+0A36 to be equivalent to <0A38, 0A3C>, or U+0A33 to be equivalent to <0A32, 0A3C>.

Two different marks can be associated with U+0902 DEVANAGARI SIGN ANUSVARA: U+0A02 GURMUKHI SIGN BINDI and U+0A70 GURMUKHI TIPPI. Present practice is to use *bindi* only with the dependent and independent forms of the vowels *aa*, *ii*, *ee*, *ai*, *oo*, and *au*, and with the independent vowels *u* and *uu*; *tippi* is used in the other contexts. Older texts may depart from this requirement. ISCII-1991 uses only one encoding point for both marks.

U+0A71 GURMUKHI ADDAK is a special sign to indicate that the following consonant is geminate. ISCII-1991 does not have a specific code point for addak and encodes it as a cluster. For example, the word ਪੱਗ *pagg*, "turban," can be represented with the sequence <0A2A, 0A71, 0A17> (or <pa, addak, ga>) in Unicode, while in ISCII-1991 it would be <pa, ga, virama, ga>.

U+0A75 ◌ੵ GURMUKHI SIGN YAKASH probably originated as a subjoined form of U+0A2F ਯ GURMUKHI LETTER YA. However, because its usage is relatively rare and not entirely predictable, it is encoded as a separate character. This character should occur after the consonant to which it attaches and before any vowel sign.

U+0A51 ◌ੑ GURMUKHI SIGN UDAAT occurs in older texts and indicates a high tone. This character should occur after the consonant to which is attaches and before any vowel sign.

Punjabi does not have complex combinations of consonant sounds. Furthermore, the orthography is not strictly phonetic, and sometimes the inherent /a/ sound is not pronounced. For example, the word ਗੁਰਮੁਖੀ *gurmukhī* is represented with the sequence <0A17, 0A41, 0A30, 0A2E, 0A41, 0A16, 0A40>, which could be transliterated as *guramukhī*; this lack of pronunciation is systematic at the end of a word. As a result, the virama sign is seldom used with the Gurmukhi script.

In older texts, such as the *Sri Guru Granth Sahib* (the Sikh holy book), one can find typographic clusters with a vowel sign attached to a vowel letter, or with two vowel signs attached to a consonant. The most common cases are ◌ੁ *u* attached to ਓ, as in ਓਮਾਹਾ and both the vowel signs ◌ੋ and ◌ੁ attached to a consonant, as in ਗੋਬਿੰਦ *goubinda*; this is used to indicate the metrical shortening of /o/ or the lengthening of /u/ depending on the context. Other combinations are attested as well, such as ਗ੍ਯਾਨ *ghiana*, represented by the sequence <U+0A17, U+0A4D, U+0A39, U+0A3F, U+0A3E, U+0A28>.

Because of the combining classes of the characters U+0A4B GURMUKHI VOWEL SIGN OO and U+0A41 GURMUKHI VOWEL SIGN U, the sequences <consonant, U+0A4B, U+0A41> and <consonant, U+0A41, U+0A4B> are not canonically equivalent. To avoid ambiguity in representation, the first sequence, with U+0A4B before U+0A41, should be used in such cases. More generally, when a consonant or independent vowel is modified by multiple vowel signs, the sequence of the vowel signs in the underlying representation of the text should be: left, top, bottom, right.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-14* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-14. Gurmukhi Vowel Letters

For	Use	Do Not Use
ਆ	0A06	<0A05, 0A3E>
ਇ	0A07	<0A72, 0A3F>
ਈ	0A08	<0A72, 0A40>
ਉ	0A09	<0A73, 0A41>
ਊ	0A0A	<0A73, 0A42>
ਏ	0A0F	<0A72, 0A47>
ਐ	0A10	<0A05, 0A48>
ਓ	0A13	<0A73, 0A4B>
ਔ	0A14	<0A05, 0A4C>

Tones. The Punjabi language is tonal, but the Gurmukhi script does not contain any specific signs to indicate tones. Instead, the voiced aspirates (*gha*, *jha*, *ddha*, *dha*) and the letter *ha* combine consonantal and tonal functions.

Ordering. U+0A73 GURMUKHI URA and U+0A72 GURMUKHI IRI are the first and third "letters" of the Gurmukhi syllabary, respectively. They are used as bases or bearers for some of

the independent vowels, while U+0A05 GURMUKHI LETTER A is both the second "letter" and the base for the remaining independent vowels. As a result, the collation order for Gurmukhi is based on a seven-by-five grid:

- The first row is U+0A73 *ura*, U+0A05 *a*, U+0A72 *iri*, U+0A38 *sa*, U+0A39 *ha*.
- This row is followed by five main rows of consonants, grouped according to the point of articulation, as is traditional in all South and Southeast Asian scripts.
- The semiconsonants follow in the seventh row: U+0A2F *ya*, U+0A30 *ra*, U+0A32 *la*, U+0A35 *va*, U+0A5C *rra*.
- The letters with *nukta*, added later, are presented in a subsequent eighth row if needed.

Rendering Behavior. For general principles regarding the rendering of the Gurmukhi script, see the rules for rendering in *Section 12.1, Devanagari*. In many aspects, Gurmukhi is simpler than Devanagari. In modern Punjabi, there are no half-consonants, no half-forms, no *repha* (upper form of U+0930 DEVANAGARI LETTER RA), and no real ligatures. Rules R2–R5, R11, and R14 do not apply. Conversely, the behavior for subscript RA (rules R6–R8 and R13) applies to U+0A39 GURMUKHI LETTER HA and U+0A35 GURMUKHI LETTER VA, which also have subjoined forms, called *pairin* in Punjabi. The subjoined form for RA is like a knot, while the subjoined HA and VA are written the same as the base form, without the top bar, but are reduced in size. As described in rule R13, they attach at the bottom of the base consonant, and will "push" down any attached vowel sign for U or UU. When U+0A2F GURMUKHI LETTER YA follows a dead consonant, it assumes a different form called *addha* in Punjabi, without the leftmost part, and the dead consonant returns to the nominal form, as shown in *Table 12-15*.

Table 12-15. Gurmukhi Conjuncts

ਮ	+	◌੍	+	ਹ	→	ਮ੍ਹ	(*mha*)	pairin ha
ਪ	+	◌੍	+	ਰ	→	ਪ੍ਰ	(*pra*)	pairin ra
ਦ	+	◌੍	+	ਵ	→	ਦ੍ਵ	(*dva*)	pairin va
ਦ	+	◌੍	+	ਯ	→	ਦ੍ਯ	(*dya*)	addha ya

Other letters behaved similarly in old inscriptions, as shown in *Table 12-16*.

Table 12-16. Additional Pairin and Addha Forms in Gurmukhi

ਸ + ◌੍ + ਗ	→ ਸ੍ਗ	(*sga*)	pairin ga
ਸ + ◌੍ + ਚ	→ ਸ੍ਚ	(*sca*)	pairin ca
ਸ + ◌੍ + ਟ	→ ਸ੍ਟ	(*stta*)	pairin tta
ਸ + ◌੍ + ਠ	→ ਸ੍ਠ	(*sttha*)	pairin ttha
ਸ + ◌੍ + ਤ	→ ਸ੍ਤ	(*sta*)	pairin ta
ਸ + ◌੍ + ਦ	→ ਸ੍ਦ	(*sda*)	pairin da
ਸ + ◌੍ + ਨ	→ ਸ੍ਨ	(*sna*)	pairin na
ਸ + ◌੍ + ਥ	→ ਸ੍ਥ	(*stha*)	pairin tha
ਸ + ◌੍ + ਯ	→ ਸ੍ਯ	(*sya*)	pairin ya
ਸ + ◌੍ + ਥ	→ ਸ੝ਥ	(*stha*)	addha tha
ਸ + ◌੍ + ਮ	→ ਸ੝ਮ	(*sma*)	addha ma

Older texts also exhibit another feature that is not found in modern Gurmukhi—namely, the use of a half- or reduced form for the first consonant of a cluster, whereas the modern practice is to represent the second consonant in a half- or reduced form. Joiners can be used to request this older rendering, as shown in *Table 12-17*. The reduced form of an initial U+0A30 GURMUKHI LETTER RA is similar to the Devanagari superscript RA (*repha*), but this usage is rare, even in older texts.

Table 12-17. Use of Joiners in Gurmukhi

ਸ + ◌੍ + ਵ	→ ਸ੍ਵ	(*sva*)
ਰ + ◌੍ + ਵ	→ ਰ੍ਵ	(*rva*)
ਸ + ◌੍ + ZWJ + ਵ	→ ਸ੍ਰ	(*sva*)
ਰ + ◌੍ + ZWJ + ਵ	→ ੱਵ	(*rva*)
ਸ + ◌੍ + ZWNJ + ਵ	→ ਸ੍ਵ	(*sva*)
ਰ + ◌੍ + ZWNJ + ਵ	→ ਰ੍ਵ	(*rva*)

A rendering engine for Gurmukhi should make accommodations for the correct positioning of the combining marks (see *Section 5.13, Rendering Nonspacing Marks*, and particularly *Figure 5-11*). This is important, for example, in the correct centering of the marks above and below U+0A28 GURMUKHI LETTER NA and U+0A20 GURMUKHI LETTER TTHA,

which are laterally symmetrical. It is also important to avoid collisions between the various upper marks, vowel signs, *bindi*, and/or *addak*.

Other Symbols. The religious symbol *khanda* sometimes used in Gurmukhi texts is encoded at U+262C ADI SHAKTI in the Miscellaneous Symbols block. U+0A74 GURMUKHI EK ONKAR, which is also a religious symbol, can have different presentation forms, which do not change its meaning. The font used in the code charts shows a highly stylized form; simpler forms look like the digit one, followed by a sign based on *ura*, along with a long upper tail.

Punctuation. Danda and double danda marks as well as some other unified punctuation used with Gurmukhi are found in the Devanagari block. See *Section 12.1, Devanagari*, for more information. Punjabi also uses Latin punctuation.

12.4 Gujarati

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 letterforms, a characteristic of the older Kaithi script to which Gujarati is related. The Gujarati script is used to write the Gujarati language of the Gujarat state in India.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-18* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-18. Gujarati Vowel Letters

For	Use	Do Not Use
આ	0A86	<0A85, 0ABE>
ઍ	0A8D	<0A85, 0AC5>
એ	0A8F	<0A85, 0AC7>
ઐ	0A90	<0A85, 0AC8>
ઑ	0A91	<0A85, 0AC9>
ઓ	0A93	<0A85, 0ACB> or <0A85, 0ABE, 0AC5>
ઔ	0A94	<0A85, 0ACC> or <0A85, 0ABE, 0AC8>
◌ૉ	0AC9	<0AC5, 0ABE>

Rendering Behavior. For rendering of the Gujarati script, see the rules for rendering in *Section 12.1, Devanagari*. Like other Brahmic scripts in the Unicode Standard, Gujarati uses the virama to form conjunct characters. The virama is informally called *khoḍo*, which means "lame" in Gujarati. Many conjunct characters, as in Devanagari, lose the vertical stroke; there are also vertical conjuncts. U+0AB0 GUJARATI LETTER RA takes special forms when it combines with other consonants, as shown in *Table 12-19*.

Punctuation. Words in Gujarati are separated by spaces. Danda and double danda marks as well as some other unified punctuation used with Gujarati are found in the Devanagari block; see *Section 12.1, Devanagari*.

Table 12-19. Gujarati Conjuncts

ક	+	◌્	+	ષ	→	ક્ષ	(*kṣa*)
જ	+	◌્	+	ઞ	→	જ્ઞ	(*jña*)
ત	+	◌્	+	ય	→	ત્ય	(*tya*)
ટ	+	◌્	+	ટ	→	ટ્ટ	(*ṭṭa*)
ર	+	◌્	+	ક	→	ર્ક	(*rka*)
ક	+	◌્	+	ર	→	ક્ર	(*kra*)

12.5 Oriya (Odia)

Oriya: U+0B00–U+0B7F

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

Languages and scripts can be referred to in many different ways, and these terms may evolve over time. The Oriya script is an example of this: The preferred Latin transcription used in India for this script has shifted to the spelling Odia (as shown, for example, by changes to the Indian constitution). The Unicode Standard retains the traditional English spelling Oriya in discussion, to minimize the potential for confusion when referring to immutable, standardized character names in the standard, which were assigned long ago.

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

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-20* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-20. Oriya Vowel Letters

For	Use	Do Not Use
ଆ	0B06	<0B05, 0B3E>
ଐ	0B10	<0B0F, 0B57>
ଔ	0B14	<0B13, 0B57>

Rendering Behavior. For rendering of the Oriya script, see the rules for rendering in *Section 12.1, Devanagari*. Like other Brahmic scripts in the Unicode Standard, Oriya uses the virama to suppress the inherent vowel. Oriya has a visible virama, often being a lengthening of a part of the base consonant:

କ + ◌୍ → କ୍ (*k*)

The virama is also used to form conjunct consonants, as shown in *Table 12-21*.

Table 12-21. Oriya Conjuncts

କ	+	◌୍	+	ଷ	→	କ୍ଷ	(*kṣa*)
କ	+	◌୍	+	ତ	→	କ୍ତ	(*kta*)
ତ	+	◌୍	+	କ	→	ତ୍କ	(*tka*)
ତ	+	◌୍	+	ୟ	→	ତ୍ୟ	(*tya*)

Consonant Forms. In the initial position in a cluster, RA is reduced and placed above the following consonant, while it is also reduced in the second position:

ର + ◌୍ + ପ → ର୍ପ (*rpa*)

ପ + ◌୍ + ର → ପ୍ର (*pra*)

Nasal and stop clusters may be written with conjuncts, or the anusvara may be used:

ଅ + ଙ + ◌୍ + କ → ଅଙ୍କ (*aṅka*)

ଅ + ◌ଂ + କ → ଅଂକ (*aṅka*)

Vowels. As with other scripts, some dependent vowels are rendered in front of their consonant, some appear after it, and some are placed above or below it. Some are rendered with parts both in front of and after their consonant. A few of the dependent vowels fuse with their consonants. U+0B01 ORIYA SIGN CANDRABINDU is used for nasal vowels. See *Table 12-22*.

Table 12-22. Oriya Vowel Placement

କ	+	◌ା	→	କା	(*kā*)
କ	+	◌ି	→	କି	(*ki*)
କ	+	◌ୀ	→	କୀ	(*kī*)
କ	+	◌ୁ	→	କୁ	(*ku*)
କ	+	◌ୂ	→	କୂ	(*kū*)
କ	+	◌ୃ	→	କୃ	(*kṛ*)
କ	+	◌େ	→	କେ	(*ke*)
କ	+	◌ୈ	→	କୈ	(*kai*)
କ	+	◌ୋ	→	କୋ	(*ko*)
କ	+	◌ୌ	→	କୌ	(*kau*)
କ	+	◌ଁ	→	କଁ	(*kaṁ*)

Oriya VA and WA. These two letters are extensions to the basic Oriya alphabet. Because Sanskrit वन *vana* becomes Oriya ବନ *bana* in orthography and pronunciation, an extended letter U+0B35 ଵ ORIYA LETTER VA was devised by dotting U+0B2C ବ ORIYA LETTER BA for use in academic and technical text. For example, basic Oriya script cannot distinguish Sanskrit ବବ *bava* from ବବ *baba* or ବବ *vava*, but this distinction can be made with the modified version of *ba*. In some older sources, the glyph ୰ is sometimes found for *va*; in others, ୱ and ୬ have been shown, which in a more modern type style would be ଵ. The letter *va* is not in common use today.

In a consonant conjunct, subjoined U+0B2C ବ ORIYA LETTER BA is usually—but not always—pronounced [wa]:

U+0B15 କ *ka* + U+0B4D ◌୍ *virama* + U+0B2C ବ *ba* → କ୍ୱ [kwa]

U+0B2E ମ *ma* + U+0B4D ◌୍ *virama* + U+0B2C ବ *ba* → ମ୍ୱ [mba]

The extended Oriya letter U+0B71 ୱ ORIYA LETTER WA is sometimes used in Perso-Arabic or English loan words for [w]. It appears to have originally been devised as a ligature of ଓ *o* and ବ *ba*, but because ligatures of independent vowels and consonants are not normally used in Oriya, this letter has been encoded as a single character that does not have a decomposition. It is used initially in words or orthographic syllables to represent the foreign consonant; as a native semivowel, *virama* + *ba* is used because that is historically accurate. Glyph variants of *wa* are ୲, ୳, and ୴.

Punctuation and Symbols. Danda and double danda marks as well as some other unified punctuation used with Oriya are found in the Devanagari block; see *Section 12.1, Devanagari*. The mark U+0B70 ORIYA ISSHAR is placed before names of persons who are deceased.

The sacred syllable *om* is formed by U+0B13 ORIYA LETTER O and U+0B01 ORIYA SIGN CANDRABINDU. Ligation of the two glyphs can be encouraged or discouraged by the use of U+200D ZERO WIDTH JOINER or U+200C ZERO WIDTH NON-JOINER between the two characters, as seen in *Table 12-23*. In the absence of a joiner, both the non-ligated and the ligated forms are acceptable renderings.

Table 12-23. Ligation for the Syllable *om*

ଓ	+	ZWJ	+	◌ଁ	→	ଓଁ or ଔଁ
ଓ	+	ZWNJ	+	◌ଁ	→	ଓଁ

Fraction Characters. As for many other scripts of India, Oriya has characters used to denote factional values. These were more commonly used before the advent of decimal weights, measures, and currencies. Oriya uses six signs: three for quarter values (1/4, 1/2, 3/4) and three for sixteenth values (1/16, 1/8, and 3/16). These are used additively, with quarter values appearing before sixteenths. Thus U+0B73 ORIYA FRACTION ONE HALF followed by U+0B75 ORIYA FRACTION ONE SIXTEENTH represents the value 5/16.

12.6 Tamil

Tamil: U+0B80–U+0BFF

The Tamil script is descended from the South Indian branch of Brahmi. It is used to write the Tamil language of the Tamil Nadu state in India as well as minority languages such as the Dravidian language Badaga and the Indo-European language Saurashtra. Tamil is also used in Sri Lanka, Singapore, and parts of Malaysia.

The Tamil script has fewer consonants than the other Indic scripts. When representing the "missing" consonants in transcriptions of languages such as Sanskrit or Saurashtra, superscript European digits are often used, so ப^{2} = *pha*, ப^{3} = *ba*, and ப^{4} = *bha*. The characters U+00B2, U+00B3, and U+2074 can be used to preserve this distinction in plain text. The Grantha script is often also used by Tamil speakers to write Sanskrit because Grantha contains these missing consonants.

The Tamil script also avoids the use conjunct consonant forms, although a few conventional conjuncts are used.

Virama (Puḷḷi). Because the Tamil encoding in the Unicode Standard is based on ISCII-1988 (Indian Script Code for Information Interchange), it makes use of the *abugida* model. An abugida treats the basic consonants as containing an inherent vowel, which can be canceled by the use of a visible mark, called a *virama* in Sanskrit. In most Brahmi-derived scripts, the placement of a virama between two consonants implies the deletion of the inherent vowel of the first consonant and causes a conjoined or subjoined consonant cluster. In those scripts, ZERO WIDTH NON-JOINER is used to display a visible virama, as shown previously in the Devangari example in *Figure 12-3*.

The situation is quite different for Tamil because the script uses very few consonant conjuncts. An orthographic cluster consisting of multiple consonants (represented by <C1, U+0BCD TAMIL SIGN VIRAMA, C2, ...>) is normally displayed with explicit viramas, which are called *puḷḷi* in Tamil. The *puḷḷi* is typically rendered as a dot centered above the character. It occasionally appears as small circle instead of a dot, but this glyph variant should be handled by the font, and not be represented by the similar-appearing U+0B82 TAMIL SIGN ANUSVARA.

The conjuncts *kssa* and *shrii* are traditionally displayed by conjunct ligatures, as illustrated for *kssa* in *Figure 12-14*, but nowadays tend to be displayed using an explicit *puḷḷi* as well.

Figure 12-14. Kssa Ligature in Tamil

க + ◌் + ஷ → க்ஷ *kṣa*

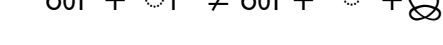
To explicitly display a *puḷḷi* for such sequences, ZERO WIDTH NON-JOINER can be inserted after the *puḷḷi* in the sequence of characters.

Rendering of the Tamil Script. The Tamil script is complex and requires special rules for rendering. The following discussion describes the most important features of Tamil rendering behavior. As with any script, a more complex procedure can add rendering characteristics, depending on the font and application.

> *In a font that is capable of rendering Tamil, the number of glyphs is greater than the number of Tamil characters.*

Tamil Vowels

Independent Versus Dependent Vowels. In the Tamil script, the dependent vowel signs are not equivalent to a sequence of *virama + independent vowel*. For example:



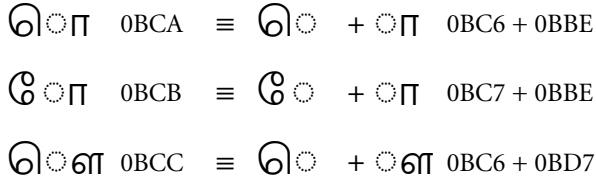
Left-Side Vowels. The Tamil vowels U+0BC6 ◌ெ, U+0BC7 ◌ே, and U+0BC8 ◌ை are reordered in front of the consonant to which they are applied. When occurring in a syllable, these vowels are rendered to the left side of their consonant, as shown in *Table 12-24*.

Table 12-24. Tamil Vowel Reordering

Memory Representation		Display
க	◌ெ	கெ
க	◌ே	கே
க	◌ை	கை

Two-Part Vowels. Tamil also has several vowels that consist of elements which flank the consonant to which they are applied. A sequence of two Unicode code points can be used to express equivalent spellings for these vowels, as shown in *Figure 12-15*.

Figure 12-15. Tamil Two-Part Vowels



In these examples, the representation on the left, which is a single code point, is the preferred form and the form in common use for Tamil. Note that the ◌ௗ in the third example is *not* U+0BB3 TAMIL LETTER LLA; it is U+0BD7 TAMIL AU LENGTH MARK.

In the process of rendering, these two-part vowels are transformed into the two separate glyphs equivalent to those on the right, which are then subject to vowel reordering, as shown in *Table 12-25*.

Table 12-25. Tamil Vowel Splitting and Reordering

Memory Representation			Display
க	◌ெ◌ா		கோ
க	◌ெ◌	◌ா	கோ
க	◌ே◌ா		கோ
க	◌ே◌	◌ா	கோ
க	◌ெ◌ள		கௌ
க	◌ெ◌	◌ள	கௌ

Even in the case where a two-part vowel occurs with a conjunct consonant or consonant cluster, the left part of the vowel is reordered around the conjunct or cluster, as shown in *Figure 12-16*.

Figure 12-16. Vowel Reordering Around a Tamil Conjunct

க + ◌் + ஷ + ◌ெ◌ + ◌ா → க்ஷோ *kṣo*

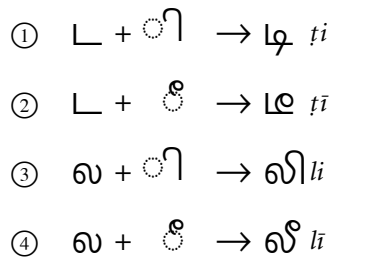
For either left-side vowels or two-part vowels, the ordering of the elements is unambiguous: the consonant or consonant cluster occurs first in the memory representation, followed by the vowel.

Tamil Ligatures

A number of ligatures are conventionally used in Tamil. Most ligatures involve the shape taken by a consonant plus vowel sequence. A wide variety of modern Tamil words are written without a conjunct form, with a fully visible *puḷḷi*.

***Ligatures with Vowel* i.** The vowel signs *i* ◌ி and *ii* ◌ீ form ligatures with the consonant *tta* ட as shown in examples 1 and 2 of *Figure 12-17*. These vowels often change shape or position slightly so as to join cursively with other consonants, as shown in examples 3 and 4 of *Figure 12-17*.

Figure 12-17. Tamil Ligatures with *i*



***Ligatures with Vowel* u.** The vowel signs *u* ◌ு and *uu* ◌ூ normally ligate with their consonant, as shown in *Table 12-26*. In the first column, the basic consonant is shown; the second column illustrates the ligation of that consonant with the *u* vowel sign; and the third column illustrates the ligation with the *uu* vowel sign.

Table 12-26. Tamil Ligatures with *u*

x	*x* + ◌ு	*x* + ◌ூ
க	கு	கூ
ங	ங	ஙூ
ச	சு	சூ
ஞ	ஞு	ஞூ
ட	டு	டூ
ண	ணு	ணூ
த	து	தூ
ந	நு	நூ
ன	னு	னூ

x	*x* + ◌ு	*x* + ◌ூ
ப	பு	பூ
ம	மு	மூ
ய	யு	யூ
ர	ரு	ரூ
ற	று	றூ
ல	லு	லூ
ள	ளு	ளூ
ழ	ழு	ழூ
வ	வு	வூ

With certain consonants, ஜ, ஷ, ஸ, ஹ, and the conjunct க்ஷ, the vowel signs *u* ◌ு and *uu* ◌ூ take a distinct spacing form, as shown in *Figure 12-18*.

***Ligatures with* ra.** Based on typographical preferences, the consonant *ra* ர may change shape to ர, when it ligates. Such change, if it occurs, will happen only when the ர form of U+0BB0 ர TAMIL LETTER RA would not be confused with the nominal form ர of U+0BBE TAMIL VOWEL SIGN AA (namely, when ர is combined with ◌் , ◌ி , or ◌ீ). This change in shape is illustrated in *Figure 12-19*.

Figure 12-18. Spacing Forms of Tamil *u*

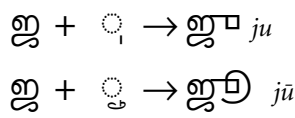
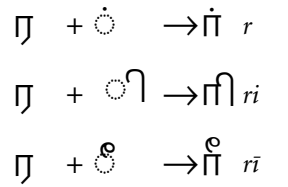


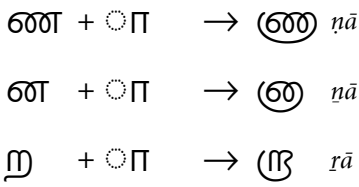
Figure 12-19. Tamil Ligatures with *ra*



However, various governmental bodies mandate that the basic shape of the consonant *ra* ர should be used for these ligatures as well, especially in school textbooks. Media and literary publications in Malaysia and Singapore mostly use the unchanged form of *ra* ர. Sri Lanka, on the other hand, specifies the use of the changed forms shown in *Figure 12-19*.

Ligatures with* aa *in Traditional Tamil Orthography. In traditional Tamil orthography, the vowel sign *aa* ◌ா optionally ligates with ண, ன, or ற, as illustrated in *Figure 12-20*.

Figure 12-20. Traditional Tamil Ligatures with *aa*



These ligations also affect the right-hand part of two-part vowels, as shown in *Figure 12-21*.

Figure 12-21. Traditional Tamil Ligatures with *o*

ண + ொ → ெண *ṇo*

ண + ோ → ேண *ṇō*

ன + ொ → ென *n̲o*

ன + ோ → ேன *n̲ō*

ற + ொ → ெற *r̲o*

ற + ோ → ேற *r̲ō*

Ligatures with* ai *in Traditional Tamil Orthography. In traditional Tamil orthography, the left-side vowel sign *ai* ை is also subject to a change in form. It is rendered as ை when it occurs on the left side of ண, ன, ல, or ள, as illustrated in *Figure 12-22*.

Figure 12-22. Traditional Tamil Ligatures with *ai*

ண + ை → ணை *ṇai*

ன + ை → னை *n̲ai*

ல + ை → லை *lai*

ள + ை → ளை *l̤ai*

By contrast, in modern Tamil orthography, this vowel does not change its shape, as shown in *Figure 12-23*.

Figure 12-23. Vowel *ai* in Modern Tamil

ண + ை → ணை *ṇai*

***Tamil* aytham.** The character U+0B83 TAMIL SIGN VISARGA is normally called *aytham* in Tamil. It is historically related to the *visarga* in other Indic scripts, but has become an ordinary spacing letter in Tamil. The *aytham* occurs in native Tamil words, but is frequently used as a modifying prefix before consonants used to represent foreign sounds. In particular, it is used in the spelling of words borrowed into Tamil from English or other languages.

Punctuation. Danda and double danda marks as well as some other unified punctuation used with Tamil are found in the Devanagari block; see *Section 12.1, Devanagari.*

Numbers. Modern Tamil decimal digits are encoded at U+0BE6..U+0BEF. Tamil also has distinct numerals for ten, one hundred, and one thousand at U+0BF0..U+0BF2 used for historical numbers.

Tamil Named Character Sequences

Tamil is less complex than some of the other Indic scripts, and both conceptually and in processing can be treated as an atomic set of elements: consonants, stand-alone vowels, and syllables. *Table 12-27* shows these atomic elements, with the corresponding Unicode characters or sequences. In cases where the atomic elements for Tamil correspond to sequences of Unicode characters, those sequences have been added to the approved list of Unicode named character sequences. See NamedSequences.txt in the Unicode Character Database for details.

In implementations such as natural language processing, where it may be useful to treat such Tamil text elements as single code points for ease of processing. Tamil named character sequences could be mapped to code points in a contiguous segment of the Private Use Area.

In *Table 12-27*, the first row shows the transliterated representation of the Tamil vowels in abbreviated form, while the first column shows the transliterated representation of the Tamil consonants. Those row and column labels, together with identifying strings such as "TAMIL SYLLABLE" or "TAMIL CONSONANT" are concatenated to form formal names for these sequences. For example, the sequence shown in the table in the K row and the AA column, with the sequence <0B95, 0BBE>, gets the associated name TAMIL SYLLABLE KAA. The sequence shown in the table in the K row in the first column, with the sequence <0B95, 0BCD>, gets the associated name TAMIL CONSONANT K.

Details on the complete names for each element can be found in NamedSequences.txt.

Table 12-27. Tamil Vowels, Consonants, and Syllables

		A	AA	I	II	U	UU	E	EE	AI	O	OO	AU
	◌ஂ ◌ஃ 0B83	அ 0B85	ஆ 0B86	இ 0B87	ஈ 0B88	உ 0B89	ஊ 0B8A	எ 0B8E	ஏ 0B8F	ஐ 0B90	ஒ 0B92	ஓ 0B93	ஔ 0B94
K	க் 0B95 0BCD	க 0B95	கா 0B95 0BBE	கி 0B95 0BBF	கீ 0B95 0BC0	கு 0B95 0BC1	கூ 0B95 0BC2	கெ 0B95 0BC6	கே 0B95 0BC7	கை 0B95 0BC8	கொ 0B95 0BCA	கோ 0B95 0BCB	கௌ 0B95 0BCC
NG	ங் 0B99 0BCD	ங 0B99	ஙா 0B99 0BBE	ஙி 0B99 0BBF	ஙீ 0B99 0BC0	ஙு 0B99 0BC1	ஙூ 0B99 0BC2	ஙெ 0B99 0BC6	ஙே 0B99 0BC7	ஙை 0B99 0BC8	ஙொ 0B99 0BCA	ஙோ 0B99 0BCB	ஙௌ 0B99 0BCC
C	ச் 0B9A 0BCD	ச 0B9A	சா 0B9A 0BBE	சி 0B9A 0BBF	சீ 0B9A 0BC0	சு 0B9A 0BC1	சூ 0B9A 0BC2	செ 0B9A 0BC6	சே 0B9A 0BC7	சை 0B9A 0BC8	சொ 0B9A 0BCA	சோ 0B9A 0BCB	சௌ 0B9A 0BCC
NY	ஞ் 0B9E 0BCD	ஞ 0B9E	ஞா 0B9E 0BBE	ஞி 0B9E 0BBF	ஞீ 0B9E 0BC0	ஞு 0B9E 0BC1	ஞூ 0B9E 0BC2	ஞெ 0B9E 0BC6	ஞே 0B9E 0BC7	ஞை 0B9E 0BC8	ஞொ 0B9E 0BCA	ஞோ 0B9E 0BCB	ஞௌ 0B9E 0BCC
TT	ட் 0B9F 0BCD	ட 0B9F	டா 0B9F 0BBE	டி 0B9F 0BBF	டீ 0B9F 0BC0	டு 0B9F 0BC1	டூ 0B9F 0BC2	டெ 0B9F 0BC6	டே 0B9F 0BC7	டை 0B9F 0BC8	டொ 0B9F 0BCA	டோ 0B9F 0BCB	டௌ 0B9F 0BCC
NN	ண் 0BA3 0BCD	ண 0BA3	ணா 0BA3 0BBE	ணி 0BA3 0BBF	ணீ 0BA3 0BC0	ணு 0BA3 0BC1	ணூ 0BA3 0BC2	ணெ 0BA3 0BC6	ணே 0BA3 0BC7	ணை 0BA3 0BC8	ணொ 0BA3 0BCA	ணோ 0BA3 0BCB	ணௌ 0BA3 0BCC
T	த் 0BA4 0BCD	த 0BA4	தா 0BA4 0BBE	தி 0BA4 0BBF	தீ 0BA4 0BC0	து 0BA4 0BC1	தூ 0BA4 0BC2	தெ 0BA4 0BC6	தே 0BA4 0BC7	தை 0BA4 0BC8	தொ 0BA4 0BCA	தோ 0BA4 0BCB	தௌ 0BA4 0BCC
N	ந் 0BA8 0BCD	ந 0BA8	நா 0BA8 0BBE	நி 0BA8 0BBF	நீ 0BA8 0BC0	நு 0BA8 0BC1	நூ 0BA8 0BC2	நெ 0BA8 0BC6	நே 0BA8 0BC7	நை 0BA8 0BC8	நொ 0BA8 0BCA	நோ 0BA8 0BCB	நௌ 0BA8 0BCC
P	ப் 0BAA 0BCD	ப 0BAA	பா 0BAA 0BBE	பி 0BAA 0BBF	பீ 0BAA 0BC0	பு 0BAA 0BC1	பூ 0BAA 0BC2	பெ 0BAA 0BC6	பே 0BAA 0BC7	பை 0BAA 0BC8	பொ 0BAA 0BCA	போ 0BAA 0BCB	பௌ 0BAA 0BCC
M	ம் 0BAE 0BCD	ம 0BAE	மா 0BAE 0BBE	மி 0BAE 0BBF	மீ 0BAE 0BC0	மு 0BAE 0BC1	மூ 0BAE 0BC2	மெ 0BAE 0BC6	மே 0BAE 0BC7	மை 0BAE 0BC8	மொ 0BAE 0BCA	மோ 0BAE 0BCB	மௌ 0BAE 0BCC
Y	ய் 0BAF 0BCD	ய 0BAF	யா 0BAF 0BBE	யி 0BAF 0BBF	யீ 0BAF 0BC0	யு 0BAF 0BC1	யூ 0BAF 0BC2	யெ 0BAF 0BC6	யே 0BAF 0BC7	யை 0BAF 0BC8	யொ 0BAF 0BCA	யோ 0BAF 0BCB	யௌ 0BAF 0BCC
R	ர் 0BB0 0BCD	ர 0BB0	ரா 0BB0 0BBE	ரி 0BB0 0BBF	ரீ 0BB0 0BC0	ரு 0BB0 0BC1	ரூ 0BB0 0BC2	ரெ 0BB0 0BC6	ரே 0BB0 0BC7	ரை 0BB0 0BC8	ரொ 0BB0 0BCA	ரோ 0BB0 0BCB	ரௌ 0BB0 0BCC
L	ல் 0BB2 0BCD	ல 0BB2	லா 0BB2 0BBE	லி 0BB2 0BBF	லீ 0BB2 0BC0	லு 0BB2 0BC1	லூ 0BB2 0BC2	லெ 0BB2 0BC6	லே 0BB2 0BC7	லை 0BB2 0BC8	லொ 0BB2 0BCA	லோ 0BB2 0BCB	லௌ 0BB2 0BCC

Table 12-27. Tamil Vowels, Consonants, and Syllables (Continued)

		A	AA	I	II	U	UU	E	EE	AI	O	OO	AU
V	வ் 0BB5 0BCD	வ 0BB5	வா 0BB5 0BBE	வி 0BB5 0BBF	வீ 0BB5 0BC0	வு 0BB5 0BC1	வூ 0BB5 0BC2	வெ 0BB5 0BC6	வே 0BB5 0BC7	வை 0BB5 0BC8	வொ 0BB5 0BCA	வோ 0BB5 0BCB	வௌ 0BB5 0BCC
LLL	ழ் 0BB4 0BCD	ழ 0BB4	ழா 0BB4 0BBE	ழி 0BB4 0BBF	ழீ 0BB4 0BC0	ழு 0BB4 0BC1	ழூ 0BB4 0BC2	ழெ 0BB4 0BC6	ழே 0BB4 0BC7	ழை 0BB4 0BC8	ழொ 0BB4 0BCA	ழோ 0BB4 0BCB	ழௌ 0BB4 0BCC
LL	ள் 0BB3 0BCD	ள 0BB3	ளா 0BB3 0BBE	ளி 0BB3 0BBF	ளீ 0BB3 0BC0	ளு 0BB3 0BC1	ளூ 0BB3 0BC2	ளெ 0BB3 0BC6	ளே 0BB3 0BC7	ளை 0BB3 0BC8	ளொ 0BB3 0BCA	ளோ 0BB3 0BCB	ளௌ 0BB3 0BCC
RR	ற் 0BB1 0BCD	ற 0BB1	றா 0BB1 0BBE	றி 0BB1 0BBF	றீ 0BB1 0BC0	று 0BB1 0BC1	றூ 0BB1 0BC2	றெ 0BB1 0BC6	றே 0BB1 0BC7	றை 0BB1 0BC8	றொ 0BB1 0BCA	றோ 0BB1 0BCB	றௌ 0BB1 0BCC
NNN	ன் 0BA9 0BCD	ன 0BA9	னா 0BA9 0BBE	னி 0BA9 0BBF	னீ 0BA9 0BC0	னு 0BA9 0BC1	னூ 0BA9 0BC2	னெ 0BA9 0BC6	னே 0BA9 0BC7	னை 0BA9 0BC8	னொ 0BA9 0BCA	னோ 0BA9 0BCB	னௌ 0BA9 0BCC
J	ஜ் 0B9C 0BCD	ஜ 0B9C	ஜா 0B9C 0BBE	ஜி 0B9C 0BBF	ஜீ 0B9C 0BC0	ஜு 0B9C 0BC1	ஜூ 0B9C 0BC2	ஜெ 0B9C 0BC6	ஜே 0B9C 0BC7	ஜை 0B9C 0BC8	ஜொ 0B9C 0BCA	ஜோ 0B9C 0BCB	ஜௌ 0B9C 0BCC
SH	ஶ் 0BB6 0BCD	ஶ 0BB6	ஶா 0BB6 0BBE	ஶி 0BB6 0BBF	ஶீ 0BB6 0BC0	ஶு 0BB6 0BC1	ஶூ 0BB6 0BC2	ஶெ 0BB6 0BC6	ஶே 0BB6 0BC7	ஶை 0BB6 0BC8	ஶொ 0BB6 0BCA	ஶோ 0BB6 0BCB	ஶௌ 0BB6 0BCC
SS	ஷ் 0BB7 0BCD	ஷ 0BB7	ஷா 0BB7 0BBE	ஷி 0BB7 0BBF	ஷீ 0BB7 0BC0	ஷு 0BB7 0BC1	ஷூ 0BB7 0BC2	ஷெ 0BB7 0BC6	ஷே 0BB7 0BC7	ஷை 0BB7 0BC8	ஷொ 0BB7 0BCA	ஷோ 0BB7 0BCB	ஷௌ 0BB7 0BCC
S	ஸ் 0BB8 0BCD	ஸ 0BB8	ஸா 0BB8 0BBE	ஸி 0BB8 0BBF	ஸீ 0BB8 0BC0	ஸு 0BB8 0BC1	ஸூ 0BB8 0BC2	ஸெ 0BB8 0BC6	ஸே 0BB8 0BC7	ஸை 0BB8 0BC8	ஸொ 0BB8 0BCA	ஸோ 0BB8 0BCB	ஸௌ 0BB8 0BCC
H	ஹ் 0BB9 0BCD	ஹ 0BB9	ஹா 0BB9 0BBE	ஹி 0BB9 0BBF	ஹீ 0BB9 0BC0	ஹு 0BB9 0BC1	ஹூ 0BB9 0BC2	ஹெ 0BB9 0BC6	ஹே 0BB9 0BC7	ஹை 0BB9 0BC8	ஹொ 0BB9 0BCA	ஹோ 0BB9 0BCB	ஹௌ 0BB9 0BCC
KSS	க்ஷ் 0B95 0BCD 0BB7 0BCD	க்ஷ 0B95 0BCD 0BB7	க்ஷா 0B95 0BCD 0BB7 0BBE	க்ஷி 0B95 0BCD 0BB7 0BBF	க்ஷீ 0B95 0BCD 0BB7 0BC0	க்ஷு 0B95 0BCD 0BB7 0BC1	க்ஷூ 0B95 0BCD 0BB7 0BC2	க்ஷெ 0B95 0BCD 0BB7 0BC6	க்ஷே 0B95 0BCD 0BB7 0BC7	க்ஷை 0B95 0BCD 0BB7 0BC8	க்ஷொ 0B95 0BCD 0BB7 0BCA	க்ஷோ 0B95 0BCD 0BB7 0BCB	க்ஷௌ 0B95 0BCD 0BB7 0BCC

SHRII	ஸ்ரீ 0BB8 0BCD 0BB0 0BC0

12.7 Telugu

Telugu: U+0C00–U+0C7F

The Telugu script is a South Indian script used to write the Telugu language of the Andhra Pradesh state in India as well as minority languages such as Gondi (Adilabad and Koi dialects) and Lambadi. The script is also used in Maharashtra, Odisha (Orissa), Madhya Pradesh, and West Bengal. The Telugu script became distinct by the thirteenth century CE and shares ancestors with the Kannada script.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-28* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-28. Telugu Vowel Letters

For	Use	Do Not Use
ఓ	0C13	<0C12, 0C55>
ఔ	0C14	<0C12, 0C4C>
◌ి	0C40	<0C3F, 0C55>
◌ీ	0C47	<0C46, 0C55>
◌ో	0C4B	<0C4A, 0C55>

Rendering Behavior. Telugu script rendering is similar to that of other Brahmic scripts in the Unicode Standard—in particular, the Tamil script. Unlike Tamil, however, the Telugu script writes conjunct characters with subscript letters. Many Telugu letters have a v-shaped headstroke, which is a structural mark corresponding to the horizontal bar in Devanagari and the arch in Oriya script. When a virama (called *virāmamu* in Telugu) or certain vowel signs are added to a letter with this headstroke, it is replaced:

U+0C15 క *ka* + U+0C4D ◌్ *virama* + U+200C ZWNJ ZERO WIDTH NON-JOINER → క్ (*k*)

U+0C15 క *ka* + U+0C3F ◌ి *vowel sign i* → కి (*ki*)

Telugu consonant clusters are most commonly represented by a subscripted, and often transformed, consonant glyph for the second element of the cluster:

U+0C17 గ *ga* + U+0C4D ◌్ *virama* + U+0C17 గ *ga* → గ్గ (*gga*)

U+0C15 క *ka* + U+0C4D ◌్ *virama* + U+0C15 క *ka* → క్క (*kka*)

U+0C15 క *ka* + U+0C4D ◌్ *virama* + U+0C2F య *ya* → క్య (*kya*)

U+0C15 క *ka* + U+0C4D ◌్ *virama* + U+0C37 ష *ssa* → క్ష (*kṣa*)

Nakāra-Pollu. The sequence <U+0C28 TELUGU LETTER NA, U+0C4D TELUGU SIGN VIRAMA> can have two representations in Telugu text. The first is the "regular" or "new style" form న్, which takes its shape from the glyphs in the sequence <U+0C28 న TELUGU LETTER NA , U+0C4D ◌్ TELUGU SIGN VIRAMA>. Older texts display the other vowel-less form ౯, called *nakāra-pollu*. The two forms are semantically identical. Fonts should render the sequence <U+0C28 TELUGU LETTER NA, U+0C4D TELUGU SIGN VIRAMA> with either the old-style glyph ౯ or the new style glyph న్. The character U+200C ZERO WIDTH NON-JOINER can be used to prevent interaction of this sequence with following consonants, as shown in *Table 12-29*.

Table 12-29. Rendering of Telugu *na* + *virama*

Font	Sequence	Rendering
Old Style	*na* + *virama*	౯
	na + *virama* + ZWNJ + *da*	౯ద
New Style	*na* + *virama*	న్
	na + *virama* + ZWNJ + *da*	న్ద
All Fonts	*na* + *virama* + *da*	న్ద

Reph. In modern Telugu, U+0C30 TELUGU LETTER RA behaves in the same manner as most other initial consonants in a consonant cluster. That is, the *ra* appears in its nominal form, and the second consonant takes the C2-conjoining or subscripted form:

U+0C30 ర *ra* + U+0C4D ◌్ *virama* + U+0C2E మ *ma* → ర్మ (*rma*)

However, in older texts, U+0C30 TELUGU LETTER RA takes the reduced (or *reph*) form ౚ when it appears first in a consonant cluster, and the following consonant maintains its nominal form:

U+0C30 ర *ra* + U+0C4D ◌్ *virama* + U+0C2E మ ma → మ్ర (*rma*)

U+200D ZERO WIDTH JOINER is placed immediately after the *virama* to render the *reph* explicitly in modern texts:

U+0C30 ర *ra* + U+0C4D ◌్ *virama* + U+200D ZWJ + U+0C2E మ *ma* → మ్ర

To prevent display of a *reph*, U+200D ZERO WIDTH JOINER is placed after the *ra*, but preceding the *virama*:

U+0C30 ర *ra* + U+200D ZWJ + U+0C4D ◌్ *virama* + U+0C2E మ *ma* → ర్మ

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 surroundrant vowel U+0C48 TELUGU VOWEL SIGN AI. The length marks are both nonspacing characters. For a detailed discussion of the use of two-part vowels, see "Two-Part Vowels" in *Section 12.6, Tamil*.

Fractions. Prior to the adoption of the metric system, Telugu fractions were used as part of the system of measurement. Telugu fractions are quaternary (base-4), and use eight marks, which are conceptually divided into two sets. The first set represents odd-numbered negative powers of four in fractions. The second set represents even-numbered negative powers of four in fractions. Different zeros are used with each set. The zero from the first set is known as *haḷḷi*, U+0C78 TELUGU FRACTION DIGIT ZERO FOR ODD POWERS OF FOUR. The zero for the second set is U+0C66 TELUGU DIGIT ZERO.

Punctuation. Danda and double danda are used primarily in the domain of religious texts to indicate the equivalent of a comma and full stop, respectively. The danda and double danda marks as well as some other unified punctuation used with Telugu are found in the Devanagari block; see *Section 12.1, Devanagari*.

12.8 Kannada

Kannada: U+0C80–U+0CFF

The Kannada script is a South Indian script. It is used to write the Kannada (or Kanarese) language of the Karnataka state in India and to write minority languages such as Tulu. The Kannada language is also used in many parts of Tamil Nadu, Kerala, Andhra Pradesh, and Maharashtra. This script is very closely related to the Telugu script both in the shapes of the letters and in the behavior of conjunct consonants. The Kannada script also shares many features common to other Indic scripts. See *Section 12.1, Devanagari*, for further information.

The Unicode Standard follows the ISCII layout for encoding, which also reflects the traditional Kannada alphabetic order.

Principles of the Kannada Script

Like Devanagari and related scripts, the Kannada script employs a halant, which is also known as a virama or vowel omission sign, U+0CCD ◌್ KANNADA SIGN VIRAMA. The halant nominally serves to suppress the inherent vowel of the consonant to which it is applied. The halant functions as a combining character. When a consonant loses its inherent vowel by the application of halant, it is known as a dead consonant. The dead consonants are the presentation forms used to depict the consonants without an inherent vowel. Their rendered forms in Kannada resemble the full consonant with the vertical stem replaced by the halant sign, which marks a character core. The stem glyph is graphically and historically related to the sign denoting the inherent /a/ vowel, U+0C85 ಅ KANNADA LETTER A. In contrast, a live consonant is a consonant that retains its inherent vowel or is written with an explicit dependent vowel sign. The dead consonant is defined as a sequence consisting of a consonant letter followed by a halant. The default rendering for a dead consonant is to position the halant as a combining mark bound to the consonant letterform.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-30* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-30. Kannada Vowel Letters

For	Use	Do Not Use
ಊ	0C8A	<0C89, 0CBE>
ೌ	0C94	<0C92, 0CCC>
ೠ	0CE0	<0C8B, 0CBE>

Consonant Conjuncts. Kannada is also noted for a large number of consonant conjunct forms that serve as ligatures of two or more adjacent forms. This use of ligatures takes place

in the context of a consonant cluster. A written consonant cluster is defined as a sequence of characters that represent one or more dead consonants followed by a normal live consonant. A separate and unique glyph corresponds to each part of a Kannada consonant conjunct. Most of these glyphs resemble their original consonant forms—many without the implicit vowel sign, wherever applicable.

In Kannada, conjunct formation tends to be graphically regular, using the following pattern:

- The first consonant of the cluster is rendered with the implicit vowel or a different dependent vowel appearing as the terminal element of the cluster.
- The remaining consonants (consonants between the first consonant and the terminal vowel element) appear in conjunct consonant glyph forms in phonetic order. They are generally depicted directly below or to the lower right of the first consonant.

A Kannada script font contains the conjunct glyph components, but they are not encoded as separate Unicode characters because they are simply ligatures. Kannada script rendering software must be able to map appropriate combinations of characters in context to the appropriate conjunct glyphs in fonts.

> *In a font that is capable of rendering Kannada, the number of glyphs is greater than the number of encoded Kannada characters.*

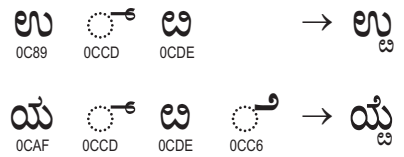
Special Characters. U+0CD5 ◌ೕ KANNADA LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0CC7 ◌ೇ KANNADA VOWEL SIGN EE should it be necessary for processing. Likewise, U+0CD6 ◌ೖ KANNADA AI LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0CC8 ◌ೈ KANNADA VOWEL SIGN AI. The Kannada two-part vowels actually consist of a nonspacing element above the consonant letter and one or more spacing elements to the right of the consonant letter. These two length marks have no independent existence in the Kannada writing system and do not play any part as independent codes in the traditional collation order.

Kannada Letter LLLA. U+0CDE ೞ KANNADA LETTER FA is actually an archaic Kannada letter that is transliterated in Dravidian scholarship as *z*, *l̤*, or *r̤*. This form should have been named "LLLA", rather than "FA", so the name in this standard is simply a mistake. A formal name alias KANNADA LETTER LLLA has been added to the Unicode Character Database for this character, to clarify its identity. Collations should treat U+0CDE as following U+0CB3 KANNADA LETTER LLA.

The *letter llla* has not been actively used in writing the Kannada language since the end of the tenth century. However, the letter does have modern use in writing the closely related Badaga language. Badaga is noteworthy for having distinct retroflexion in its vowel system, and a subjoined form of U+0CDE is often seen in Badaga written documents, to indicate retroflexed pronunciation of the vowel in a syllable. This subjoined form of U+0CDE may occur below consonants, but it also may be subjoined to an independent vowel, to indicate retroflexion of that vowel. In either case, the subjoined form of U+0CDE should be represented by a sequence including U+0CCD KANNADA SIGN VIRAMA. Implementations of the Kannada script need to be aware that these sequences involving independent vowels fol-

lowed by virama and U+0CDE are valid and required in orthographies for Badaga. Examples of the use of subjoined U+0CDE to indicate retroflexion, both for independent vowel letters and for dependent vowels, are shown in *Figure 12-24*.

Figure 12-24. Indicating Retroflexion in Badaga Vowels



Rendering Kannada

Plain text in Kannada is generally stored in phonetic order; that is, a CV syllable with a dependent vowel is always encoded as a consonant letter C followed by a vowel sign V in the memory representation. This order is employed by the ISCII standard and corresponds to the phonetic and keying order of textual data. Unlike in Devanagari and some other Indian scripts, all of the dependent vowels in Kannada are depicted to the right of their consonant letters. Hence there is no need to reorder the elements in mapping from the logical (character) store to the presentation (glyph) rendering, and vice versa.

Explicit Virama (Halant). Normally, a halant character creates dead consonants, which in turn combine with subsequent consonants to form conjuncts. This behavior usually results in a halant sign not being depicted visually. Occasionally, this default behavior is not desired when a dead consonant should be excluded from conjunct formation, in which case the halant sign is visibly rendered. To accomplish this, U+200C ZERO WIDTH NON-JOINER is introduced immediately after the encoded dead consonant that is to be excluded from conjunct formation. See *Section 12.1, Devanagari*, for examples.

Vowelless NA. The sequence <U+0CA8 KANNADA LETTER NA, U+0CCD KANNADA SIGN VIRAMA> can have two representations in Kannada text. The first is the "regular" or "new style" form ನ್, which takes its shape from the glyphs in the sequence <U+0CA8 KANNADA LETTER NA, U+0CCD KANNADA SIGN VIRAMA>. Older texts display the other vowel-less form ನ. The two forms are semantically identical. Fonts should render the sequence <U+0CA8 KANNADA LETTER NA, U+0CCD KANNADA SIGN VIRAMA> with either the old-style glyph ನ or the new style glyph ನ್. The character U+200C ZERO WIDTH NON-JOINER can be used to prevent interaction of this sequence with the following consonants, as shown in *Table 12-31*.

See the discussion of the analogous rendering of *na* in Telugu, called nakāra-pollu, in *Section 12.7, Telugu*.

Consonant Clusters Involving RA. Whenever a consonant cluster is formed with the U+0CB0 ರ KANNADA LETTER RA as the first component of the consonant cluster, the letter

Table 12-31. Rendering of Kannada *na* + *virama*

Font	Sequence	Rendering
Old Style	*na* + *virama* *na* + *virama* + ZWNJ + *da*	ನ್ ನ್ದ
New Style	*na* + *virama* *na* + *virama* + ZWNJ + *da*	ನ್ ನ್ದ
All Fonts	*na* + *virama* + *da*	ನ್ದ

ra is depicted with two different presentation forms: one as the initial element and the other as the final display element of the consonant cluster.

U+0CB0 ರ *ra* + U+0CCD ◌್ *halant* + U+0C95 ಕ *ka* → ರ್ಕ *rka*

U+0CB0 ರ *ra* + ZWJ + U+0CCD ◌್ *halant* + U+0C95 ಕ *ka* → ರ್ಕ *rka*

U+0C95 ಕ *ka* + U+0CCD ◌್ *halant* + U+0CB0 ರ *ra* → ಕ್ರ *kra*

Jihvamuliya and Upadhmaniya. Voiceless velar and bilabial fricatives in Kannada are represented by U+0CF1 KANNADA SIGN JIHVAMULIYA and U+0CF2 KANNADA SIGN UPADHMANIYA, respectively. When the signs appear with a following homorganic voiceless stop consonant, the combination should be rendered in the font as a stacked ligature, without a virama:

U+0CF1 ೱ *jihvamuliya* + U+0C95 ಕ *ka* → ೱ_ಕ

U+0CF2 ೲ *upadhmaniya* + U+0CAB ಫ *pha* → ೲ_ಫ

Dependent vowels signs can also be added to the stack:

U+0CF1 ೱ *jihvamuliya* + U+0C95 ಕ *ka* + U+0CBF ◌ಿ *vowel sign i* → ೱ_{ಕಿ}

Modifier Mark Rules. In addition to the vowel signs, one or more types of combining marks may be applied to a component of a written syllable or the syllable as a whole. If the consonant represents a dead consonant, then the nukta should precede the halant in the memory representation. The nukta is represented by a double-dot mark, U+0CBC ◌಼ KANNADA SIGN NUKTA. Two such modified consonants are used in the Kannada language: one representing the syllable *za* and one representing the syllable *fa*.

Avagraha Sign. A spacing mark called U+0CBD ಽ KANNADA SIGN AVAGRAHA is used when rendering Sanskrit texts.

Punctuation. Danda and double danda marks as well as some other unified punctuation used with this script are found in the Devanagari block; see *Section 12.1, Devanagari*.

12.9 Malayalam

Malayalam: U+0D00–U+0D7F

The Malayalam script is a South Indian script used to write the Malayalam language of the Kerala state. Malayalam is a Dravidian language like Kannada, Tamil, and Telugu. Throughout its history, it has absorbed words from Tamil, Sanskrit, Arabic, and English.

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

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-32* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 12-32. Malayalam Vowel Letters

For	Use	Do Not Use
ഈ	0D08	<0D07, 0D57>
ഊ	0D0A	<0D09, 0D57>
ഐ	0D10	<0D0E, 0D46>
ഓ	0D13	<0D12, 0D3E>
ഔ	0D14	<0D12, 0D57>

Two-Part Vowels. The Malayalam script uses several two-part vowel characters. In modern times, the dominant practice is to write the dependent form of the *au* vowel using only "ൗ", which is placed on the right side of the consonant it modifies; such texts are represented in Unicode using U+0D57 MALAYALAM AU LENGTH MARK. In the past, this dependent form was written using both "ൌ" on the left side and "ൗ" on the right side; U+0D4C MALAYALAM VOWEL SIGN AU can be used for documents following this earlier tradition. This historical simplification started much earlier than the orthographic reforms mentioned in the text that follows.

For a detailed discussion of the use of two-part vowels, see "Two-Part Vowels" in *Section 12.6, Tamil*.

Historic Characters. The four characters, *avagraha*, *vocalic rr sign*, *vocalic l sign*, and *vocalic ll sign*, are only used to write Sanskrit words in the Malayalam script. The *avagraha* is the most common of the four. The *vocalic l sign* is also commonly used in Sanskrit words.

Malayalam Orthographic Reform

In the 1970s and 1980s, Malayalam underwent orthographic reform due to printing difficulties. The treatment of the combining vowel signs *u* and *uu* was simplified at this time. These vowel signs had previously been represented using special cluster graphemes where

the vowel signs were fused beneath their consonants, but in the reformed orthography they are represented by spacing characters following their consonants. *Table 12-33* lists a variety of consonants plus the *u* or *uu* vowel sign, yielding a syllable. Each syllable is shown as it would be displayed in the older orthography, contrasted with its display in the reformed orthography.

Table 12-33. Malayalam Orthographic Reform

Syllable		Older Orthography	Reformed Orthography
ku	ക + ു	കൃ	കു
gu	ഗ + ു	ഗൃ	ഗു
chu	ച + ു	ച്ചൃ	ചു
ju	ജ + ു	ജൃ	ജു
ṇu	ണ + ു	ണൃ	ണു
tu	ത + ു	ത്തൃ	തു
nu	ന + ു	ന്നൃ	നു
bhu	ഭ + ു	ഭൃ	ഭു
ru	ര + ു	ര്രൃ	രു
śu	ശ + ു	ശ്രൃ	ശു
hu	ഹ + ു	ഹ്നൃ	ഹു
kū	ക + ൂ	കൄ	കൂ
gū	ഗ + ൂ	ഗൄ	ഗൂ
chū	ച + ൂ	ച്ചൄ	ചൂ
jū	ജ + ൂ	ജൄ	ജൂ
ṇū	ണ + ൂ	ണൄ	ണൂ
tū	ത + ൂ	ത്തൄ	തൂ
nū	ന + ൂ	ന്നൄ	നൂ
bhū	ഭ + ൂ	ഭൄ	ഭൂ
rū	ര + ൂ	ര്രൄ	രൂ
śū	ശ + ൂ	ശ്രൄ	ശൂ
hū	ഹ + ൂ	ഹ്നൄ	ഹൂ

Rendering Malayalam

Candrakkala. As is the case for many other Brahmi-derived scripts in the Unicode Standard, Malayalam uses a virama character to form consonant conjuncts. The virama sign itself is known as *candrakkala* in Malayalam. *Table 12-34* provides a variety of examples of

consonant conjuncts. There are both horizontal and vertical conjuncts, some of which ligate, and some of which are merely juxtaposed.

Table 12-34. Malayalam Conjuncts

ക	+	◌്	+	ഷ	→	ക്ഷ	*(kṣa)*
ക	+	◌്	+	ക	→	ക്ക	*(kka)*
ജ	+	◌്	+	ഞ	→	ജ്ഞ	*(jña)*
ട	+	◌്	+	ട	→	ട്ട	*(ṭṭa)*
പ	+	◌്	+	പ	→	പ്പ	*(ppa)*
ച	+	◌്	+	ഛ	→	ച്ഛ	*(ccha)*
ബ	+	◌്	+	ബ	→	ബ്ബ	*(bba)*
ന	+	◌്	+	യ	→	ന്യ	*(nya)*
പ	+	◌്	+	ര	→	പ്ര	*(pra)*
ശ	+	◌്	+	വ	→	ശ്വ	*(śva)*

When the *candrakkala* sign is visibly shown in Malayalam, it indicates either the suppression of the preceding vowel or its replacement with a neutral vowel sound. This sound is often called "half-u" or *samvruthokaram*. In traditional orthography it is displayed with a vowel sign -u followed by *candrakkala*, and in modern orthography it is displayed with a *candrakkala* alone. In all cases, the *candrakkala* sign is represented by the character U+0D4D MALAYALAM SIGN VIRAMA, which follows any vowel sign that may be present and precedes any *anusvara* that may be present. Examples are shown in *Table 12-35*.

Table 12-35. Candrakkala Examples

പാലു്	/pālə/ milk	0D2A, 0D3E, 0D32, 0D41, 0D4D
ഏ്ന്നാ	/ənnā/ on which day?	0D0E, 0D4D, 0D28, 0D4D, 0D28, 0D3E
ഐശീല്ം	/aishīləm/ than ice	0D10, 0D36, 0D40, 0D32, 0D4D, 0D02

Explicit Candrakkala. The sequence <C1, virama, ZWNJ, C2>, where C1 and C2 are consonants, may be used to request display with an explicit visible *candrakkala*, instead of the default conjunct form. See *Table 12-36* for an example. This convention is consistent with the use of this sequence in other Indic scripts.

Requesting Traditional Ligatures. The sequence <C1, ZWJ, virama, C2> may be used to request traditional ligatures, even if the current font defaults to the conjuncts appropriate for the reformed orthography. When such sequences occur, a closed or cursively connected ligature should be displayed, if available. See *Table 12-36* for examples. This convention is consistent with the use of this sequence in some other Indic scripts, such as Kannada, Oriya, and Telugu.

Requesting Open Forms of Conjuncts. The sequence <C1, ZWNJ, virama, C2> may be used to request open ligatures or those used in the reformed orthography, even if the current font defaults to the conjuncts appropriate for the traditional orthography. When such sequences occur, an open or disconnected conjunct form should be displayed, if available. See *Table 12-36* for examples. Note that such sequences are defined for Malayalam only, and are left undefined for other Indic scripts.

Table 12-36. Use of Joiners in Malayalam

ക + ് + ര →	ക്ര or ക്ര *(kra)*
സ + ് + ക →	സ്ക or സ്ക *(ska)*
ത + ് + സ →	ത്സ or ത്സ *(tsa)*
ഴ + ് + വ →	ഴ്വ or ഴ്വ or ഴ്വ *(ḻva)*
യ + ് + യ →	യ്യ *(yya)*
ക + ് + ZWNJ + ര →	ക്ര *(kra)*
ക + ZWJ + ് + ര →	ക്ര *(kra)*
സ + ZWJ + ് + ക →	സ്ക *(ska)*
ത + ZWJ + ് + സ →	ത്സ *(tsa)*
ഴ + ZWJ + ് + വ →	ഴ്വ *(ḻva)*
ക + ZWNJ + ് + ര →	ക്ര *(kra)*
ഴ + ZWNJ + ് + വ →	ഴ്വ *(ḻva)*
യ + ZWNJ + ് + യ →	യ്യ *(yya)*

Anusvara. The *anusvara* can be seen multiple times after vowels, whether independent letters or dependent vowel signs, as in ഇംംംം <0D08, 0D02, 0D02, 0D02, 0D02>. Vowel signs can also be seen after digits, as in ൩൫൦ം <0033, 0035, 0035, 0D3E, 0D02>. More generally, rendering engines should be prepared to handle Malayalam letters (including vowel letters), digits (both European and Malayalam), dashes, U+00A0 NO-BREAK SPACE and U+25CC DOTTED CIRCLE as base characters for the Malayalam vowel signs, U+0D4D MALAYALAM SIGN VIRAMA, U+0D02 MALAYALAM SIGN ANUSVARA, and U+0D03 MALAYALAM SIGN VISARGA. They should also be prepared to handle multiple combining marks on those bases.

Chillu Forms. The six characters, U+0D7A..U+0D7F, encode dead consonants (those without an inherent vowel) known as *chillu* or *cillakṣaram*. In Malayalam language text, *chillu* forms never start a word. Occasionally, chillu forms may take vowels or be elements

of conjuncts. The *chillu* forms *nn*, *-n*, *-rr*, *-l*, and *-ll* are quite common; *chillu-k* is relatively rare in contemporary usage.

For backward-compatibility issues regarding the representation of *chillu* forms, see the discussion of legacy *chillu* sequences later in this section.

***Special Cases Involving* rra.** There are a number of textual representation and reading issues involving the letter *rra*. These issues are discussed here and tables of explicit examples are presented.

The letter റ *rra* is normally read /ṟa/. Repetition of that sound is naturally written by repeating the letter: ററ. Each occurrence can bear a vowel sign.

The same repetition of the letter *rra* as ററ is also used for /ṯṯa/, which can be unambiguously represented by റ്റ. The sequence of two റ letters fundamentally behaves as a digraph in this instance. The digraph can bear a vowel sign in which case the digraph as a whole acts graphically as an atom: a left vowel part goes to the left of the digraph and a right vowel part goes to the right of the digraph. Historically, the side-by-side form was used until around 1960 when the stacked form began appearing and supplanted the side-by-side form.

As a consequence the graphical sequence ററ in text is ambiguous in reading. The reader must generally use the context to understand if ററ is read /ṟaṟa/ or /ṯṯa/. It is only when a vowel part appears between the two റ that the reading cannot be /ṯṯa/. Note that similar situations are common in many other orthographies. For example, *th* in English can be a digraph (*cathode*) or two separate letters (*cathouse*); *gn* in French can be a digraph (*oignon*) or two separate letters (*gnome*).

The sequence <0D31, 0D31> is rendered as ററ, regardless of the reading of that text. The sequence <0D31, 0D4D, 0D31> is rendered as റ്റ. In both cases, vowels signs can be used as appropriate, as shown in *Table 12-37*.

Table 12-37. Malayalam /ṟaṟa/ and /ṯṯa/

പാററ	0D2A 0D3E 0D31 0D31	/pāṯṯa/	cockroach
പാറ്റ	0D2A 0D3E 0D31 0D4D 0D31		
മാററൊലി	0D2E 0D3E 0D31 0D46 0D31 0D3E 0D32 0D3F	/māṯṯoli/	echo
മാറ്റൊലി	0D2E 0D3E 0D31 0D4D 0D31 0D46 0D3E 0D32 0D3F		
ബാററരി	0D2C 0D3E 0D31 0D31 0D31 0D3F	/bāṯṯaṟi/	battery
ബാറ്ററി	0D2C 0D3E 0D31 0D4D 0D31 0D31 0D3F		

Table 12-37. Malayalam /r̲ar̲a/ and /t̲t̲a/ (Continued)

സൂററ്റ്	0D38 0D42 0D31 0D31 0D31 0D4D	/sūr̲at̲t̲/	Surat, a town in Gujarat
സൂറ്റ്	0D38 0D42 0D31 0D31 0D4D 0D31 0D4D		
ടെംപററി	0D1F 0D46 0D02 0D2A 0D31 0D31 0D3F	/ṭempar̲ar̲i/	temporary
ലെക്ചറോട്	0D32 0D46 0D15 0D4D 0D1A 0D31 0D31 0D4B 0D1F 0D4D	/lekcar̲ar̲ōṭ/	to the lecturer

A very similar situation exists for the combination of ൻ *chillu-n* and ഩ *rra*. When used side by side, ൻഩ can be read either /n̲r̲a/ or /n̲t̲a/, while stacked ന്റ is always read /n̲t̲a/.

The sequence <0D7B, 0D31> is rendered as ൻഩ, regardless of the reading of that text. The sequence <0D7B, 0D4D, 0D31> is rendered as ന്റ. In both cases, vowels signs can be used as appropriate, as shown in *Table 12-38*.

Table 12-38. Malayalam /nr/ and /nt/

ആൻറോ	0D06 0D7B 0D47 0D31 0D3E	/āntō/	a proper name
ആന്റോ	0D06 0D7B 0D4D 0D31 0D47 0D3E		
എൻറോൾ	0D0E 0D7B 0D31 0D47 0D3E 0D7A	/enr̲ōl/	enroll

Dot Reph. U+0D4E MALAYALAM LETTER DOT REPH is used to represent the dead consonant form of U+0D30 MALAYALAM LETTER RA, when it is displayed as a dot or small vertical stroke above the consonant that follows it in logical order. It has the character properties of a letter rather than those of a combining mark, but special behavior is required in implementations. Conceptually, *dot reph* is analogous to the sequence <*ra*, *virama*> which, in many Indic scripts, is rendered as a *reph* mark over the following consonant. This same behavior is expected for *dot reph*: it should be rendered as a mark over the following consonant. In standard Malayalam, the sequence <*ra*, *virama*> would normally occur only within the sequence <*ra*, *virama, ya*>, which should be rendered as the nominal form of *ra* with a conjoining form of *ya*.

The sequence <*ra*, *virama*, ZWJ> is not used to represent the *dot reph*, because that sequence has considerable preexisting usage to represent the *chillu* form of *ra*, prior to the encoding of the *chillu* form as a distinct character, U+0D7C MALAYALAM LETTER CHILLU RR.

The Malayalam *dot reph* was in common print usage until 1970, but has fallen into disuse. Words that formerly used *dot reph* on a consonant are now spelled instead with a *chillu-rr* form preceding the consonant. (See the discussion of *chillu* characters earlier in this section.) The *dot reph* form is predominantly used by those who completed elementary education in Malayalam prior to 1970.

Legacy Chillu Sequences. Prior to Unicode Version 5.1, the representation of text with *chillu* forms was problematic, and not clearly described in the text of the standard. Because older data will use different representation for *chillu* forms, implementations must be pre-

pared to handle both kinds of data. For *chillu* forms considered in isolation, the following table shows the relationship between their representation in Version 5.0 and earlier, and the recommended representation starting with Version 5.1. Note that only the first five *chillu* forms listed in *Table 12-39* were represented in legacy text by <virama, ZWJ> sequences. The other *chillu* forms are only represented as atomically encoded *chillu* characters.

Table 12-39. Atomic Encoding of Malayalam *Chillus*

Visual	Legacy Representation (5.0)	Preferred Representation
ൺ	NNA, VIRAMA, ZWJ 0D23, 0D4D, 200D	0D7A MALAYALAM LETTER CHILLU NN
ൻ	NA, VIRAMA, ZWJ 0D28, 0D4D, 200D	0D7B MALAYALAM LETTER CHILLU N
ർ	RA, VIRAMA, ZWJ 0D30, 0D4D, 200D	0D7C MALAYALAM LETTER CHILLU RR
ൽ	LA, VIRAMA, ZWJ 0D32, 0D4D, 200D	0D7D MALAYALAM LETTER CHILLU L
ൾ	LLA, VIRAMA, ZWJ 0D33, 0D4D, 200D	0D7E MALAYALAM LETTER CHILLU LL
ൿ	*undefined*	0D7F MALAYALAM LETTER CHILLU K

Malayalam Numbers and Punctuation

Archaic Numbers. There are six characters used for the archaic number system, including characters for numbers 10, 100, 1000 and fractions.

Date Mark. The *date mark* is used only for the day of the month in dates; it is roughly the equivalent of "th" in "June 5th." While it has been used in modern times it is not seen as much in contemporary use.

Punctuation. *Danda* and *double danda* marks as well as some other unified punctuation used with Malayalam are found in the Devanagari block; see *Section 12.1, Devanagari*.