The Unicode® Standard Version 12.0 – Core Specification

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

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

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

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

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

© 2019 Unicode, Inc.

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

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 12.0.

Includes index.

ISBN 978-1-936213-22-1 (http://www.unicode.org/versions/Unicode12.0.0/)

 $1.\ Unicode\ (Computer\ character\ set) \quad \ I.\ Unicode\ Consortium.$

QA268.U545 2019

ISBN 978-1-936213-22-1 Published in Mountain View, CA March 2019

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/). Wordinitial 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–F4₁₆ 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 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 within an orthographic syllable to depict initial consonants in a consonant cluster. 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, as discussed later in this section.

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 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.

For Use Do Not Use ऄ 0904 <0905, 0946> 0906 <0905, 093E> 0908 <0930, 094D, 0907> 090A <0909, 0941> 090D <090F, 0945> 090E <090F, 0946> 0910 <090F, 0947> 0911 <0905, 0949> or <0906, 0945> 0912 <0905, 094A> or <0906, 0946> 0913 <0905, 094B> or <0906, 0947> 0914 <0905, 094C> or <0906, 0948> 0972 <0905, 0945> 0973 <0905, 093A> <0905, 093B> or <0906, 093A> 0975 <0905, 094F> 0976 <0905, 0956> 0977 <0905, 0957>

Table 12-1. Devanagari Vowel Letters

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.

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

$$\mathsf{TA}_n + \mathsf{VIRAMA}_n \to \mathsf{TA}_d$$

$$\pi + \quad \bigcirc \quad \to \quad \overline{\pi}$$

It could be assumed that a dead consonant may be combined with a vowel letter or sign to represent a CV orthographic syllable. Some non-Unicode implementations have used this approach; however, this is not done in implementations of the Unicode Standard. Instead, a CV orthographic syllable is represented with a (live) consonant followed by a dependent vowel. A dead consonant should not be followed either by an independent vowel letter or by a dependent vowel sign in an attempt to create an alternative representation of a CV orthographic syllable.

Atomic Representation of Consonant Letters. Consonant letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. In particular, consonant half forms are dead-consonant forms that often resemble a full consonant form minus a vertical stem. This vertical stem is visually similar to the vowel sign denoting /ā/, U+093E DEVANAGARI VOWEL SIGN AA. Table 12-2 shows atomic consonant letters in Devanagari that could be graphically analyzed this way, 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 12-2. Devanagari Atomic Consonants

For	Use	Do Not Use
ख	0916	<0916, 094D, 093E>, <0916, 094D, 200D, 093E>
ग	0917	<0917, 094D, 093E>, <0917, 094D, 200D, 093E>
घ	0918	<0918, 094D, 093E>, <0918, 094D, 200D, 093E>
च	091A	<091A, 094D, 093E>, <091A, 094D, 200D, 093E>
ज	091C	<091C, 094D, 093E>, <091C, 094D, 200D, 093E>
झ	091D	<091D, 094D, 093E>, <091D, 094D, 200D, 093E>
ञ	091E	<091E, 094D, 093E>, <091E, 094D, 200D, 093E>
ण	0923	<0923, 094D, 093E>, <0923, 094D, 200D, 093E>
त	0924	<0924, 094D, 093E>, <0924, 094D, 200D, 093E>
थ	0925	<0925, 094D, 093E>, <0925, 094D, 200D, 093E>
ध	0927	<0927, 094D, 093E>, <0927, 094D, 200D, 093E>

 Table 12-2.
 Devanagari Atomic Consonants (Continued)

For	Use	Do Not Use
न	0928	<0928, 094D, 093E>, <0928, 094D, 200D, 093E>
न	0929	<0929, 094D, 093E>, <0929, 094D, 200D, 093E>, <0928, 093C, 094D, 093E>, <0928, 093C, 094D, 093E>,
प	092A	<092A, 094D, 093E>, <092A, 094D, 200D, 093E>
ৰ	092C	<092C, 094D, 093E>, <092C, 094D, 200D, 093E>
भ	092D	<092D, 094D, 093E>, <092D, 094D, 200D, 093E>
म	092E	<092E, 094D, 093E>, <092E, 094D, 200D, 093E>
य	092F	<092F, 094D, 093E>, <092F, 094D, 200D, 093E>
ল	0932	<0932, 094D, 093E>, <0932, 094D, 200D, 093E>
व	0935	<0935, 094D, 093E>, <0935, 094D, 200D, 093E>
থ	0936	<0936, 094D, 093E>, <0936, 094D, 200D, 093E>
ष	0937	<0937, 094D, 093E>, <0937, 094D, 200D, 093E>
स	0938	<0938, 094D, 093E>, <0938, 094D, 200D, 093E>
ख़	0959	<0959, 094D, 093E>, <0959, 094D, 200D, 093E>, <0916, 093C, 094D, 093E>, <0916, 093C, 094D, 093E>,
ग़	095A	<095A, 094D, 093E>, <095A, 094D, 200D, 093E>, <0917, 093C, 094D, 093E>, <0917, 093C, 094D, 093E>
ज़	095B	<095B, 094D, 093E>, <095B, 094D, 200D, 093E>, <091C, 093C, 094D, 093E>, <091C, 093C, 094D, 093E>, <091C, 093C, 094D, 200D, 093E>
य़	095F	<095F, 094D, 093E>, <095F, 094D, 200D, 093E>, <092F, 093C, 094D, 093E>, <092F, 093C, 094D, 093E>, <092F, 093C, 094D, 200D, 093E>
ॹ	0979	<0979, 094D, 093E>, <0979, 094D, 200D, 093E>
ষ	097A	<097A, 094D, 093E>, <097A, 094D, 200D, 093E>
ग	097B	<097B, 094D, 093E>, <097B, 094D, 200D, 093E>
ত্র	097C	<097C, 094D, 093E>, <097C, 094D, 200D, 093E>
<u>ਫ</u>	097E	<097E, 094D, 093E>, <097E, 094D, 200D, 093E>
ॿ	097F	<097F, 094D, 093E>, <097F, 094D, 200D, 093E>

The principle of using atomic consonant representations, rather than representations analyzing the consonant into a half form plus stem, also applies to other Indic scripts, such as Gujarati and Bengali.

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)
$$GA_d + DHA_I \rightarrow GA_h + DHA_n$$
 (3) $KA_d + SSA_I \rightarrow K.SSA_n$

$$\underbrace{\eta} + \underbrace{u} \rightarrow \underbrace{u} \qquad \underbrace{\pi} + \underbrace{u} \rightarrow \underbrace{u} \rightarrow \underbrace{\pi} + \underbrace{u} \rightarrow \underbrace{u} \rightarrow \underbrace{\pi} + \underbrace{u} \rightarrow \underbrace{u} \rightarrow$$

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_{SUD} (repha).

A consonant conjunct form can take a virama and so become a dead consonant conjunct form. A dead consonant conjunct form can be followed by another consonant letter, and so form a multi-consonant conjunct. For example, *Figure 12-3* illustrates a three-consonant conjunct form, P.S.YA.

Figure 12-3. Multi-Consonant Conjuncts in Devanagari

$$PA_d + SA_d + YA_I \rightarrow P.S.YA_n$$

 $\Psi + \Psi + \Psi \rightarrow \Psi \Psi$

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.

A dead consonant conjunct may have an appearance like a half form, because the vertical stem of the last consonant is removed. As a result, a live consonant conjunct could be analyzed visually as consisting of the dead, consonant-conjunct half form plus the vowel sign $/\bar{a}$ /. As in the case of consonant letters, the live form should not be represented using a half form followed by U+093E DEVANAGARI VOWEL SIGN AA. *Table 12-3* shows some examples of live consonant conjuncts that exhibit this visual pattern, but that should not be represented with fully analyzed sequences. *Table 12-3* also shows the sequence of code points that should be used to represent these conjuncts in text, and the sequence of code points resulting from analysis that should not be used.

For	Use	Do Not Use
क्च	<0915, 094D, 091A>	<0915, 094D, 091A, 094D, 093E>, <0915, 094D, 091A, 094D, 200D, 093E>
क्ष	<0915, 094D, 0937>	<0915, 094D, 0937, 094D, 093E>, <0915, 094D, 0937, 094D, 200D, 093E>
त्त	<0924, 094D, 0924>	<0924, 094D, 0924, 094D, 093E>, <0924, 094D, 0924, 094D, 200D, 093E>
न्त	<0928, 094D, 0924>	<0928, 094D, 0924, 094D, 093E>, <0928, 094D, 0924, 094D, 200D, 093E>

Table 12-3. Devanagari Consonant Conjuncts

Note that these are illustrative examples only. There are many consonant conjuncts that could be visually analyzed in the same way, and the same principle applies to all such cases: these should not be represented as dead conjunct plus vowel sign sequences. The principle of using atomic consonant representations, rather than representations analyzing the consonant into a half form plus stem, also applies to other Indic scripts, such as Gujarati and Bengali.

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 NONJOINER 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 12-4*, the use of ZERO WIDTH NON-JOINER prevents the default formation of the conjunct form $\mathcal{E}(K.SSA_n)$.

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 dis-

Figure 12-4. Preventing Conjunct Forms in Devanagari

$$KA_d + ZWNJ + SSA_I \rightarrow KA_d + SSA_n$$
 $\phi + W \rightarrow \phi$

tinctly 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 considered 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 12-5*.

Figure 12-5. Half-Consonants in Devanagari

$$KA_d + ZWJ + SSA_I \rightarrow KA_h + SSA_n$$
 $\overline{A}_h + \overline{A}_h + \overline{A}_h + \overline{A}_h + \overline{A}_h$
 $\overline{A}_h + \overline{A}_h + \overline{A}_h + \overline{A}_h + \overline{A}_h$

In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 12-5* would normally produce the full conjunct form \mathcal{H} (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 12-6*.

Figure 12-6. Independent Half-Forms in Devanagari

$$GA_d + ZWJ \rightarrow GA_h$$

$$\boxed{\eta} + \boxed{2} \rightarrow \boxed{1}$$

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 12-7*.

Figure 12-7. Half-Consonants in Oriya

In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 12-7* would normally produce the full conjunct form \mathfrak{D} (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 12-8*).

Figure 12-8. 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_{I}	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 <i>repha</i> .
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 .

$$\mathsf{TA}_n + \mathsf{VIRAMA}_n \to \mathsf{TA}_d$$

$$\mathsf{TA}_n + \quad \circlearrowleft \quad \to \quad \mathsf{TA}_d$$

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.

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.

$${
m RA}_d + {
m JA}_d + {
m NYA}_l
ightarrow {
m J.NYA}_n + {
m RA}_{sup}$$
 Displayed Output ${
m c}$ ${
m t}$ + ${
m v}$ + ${
m v}$ + ${
m v}$ ${
m t}$

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.

$$RA_d + GA_d + GHA_I \rightarrow GA_h + GHA_I + RA_{sup}$$
 Displayed Output
 $\chi + \eta + \Xi \rightarrow \tau + \Xi + \dot{\gamma} \rightarrow \tau \dot{\Xi}$

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.

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_{SUD} .

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_{SUD} , which is positioned so that it applies to C_n .

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.

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.

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

$$T.RA_d + YA_I \rightarrow T.R.YA_n$$
 $\overline{A} + \overline{A} \rightarrow \overline{A}$

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.

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.

$$\mathsf{KA}_n + \mathsf{AA}_{vs} + \mathsf{CANDRABINDU}_n$$
 क $+$ ा $+$ ँ \to 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.

$$JA_d + NYA_l \rightarrow J.NYA_n$$
 $TTA_d + TTHA_l \rightarrow TT.TTHA_n$ $\overline{q} + \overline{a} \rightarrow \overline{q}$ $\overline{c} + \overline{c} \rightarrow \overline{c}$

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

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

$$\text{K.SSA}_d + \text{YA}_l \rightarrow \text{K.SS}_h + \text{YA}_n$$

क्ष $+$ य \rightarrow 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).

$$\mathsf{KA}_n + \mathsf{RA}_{sub} \to \mathsf{K.RA}_n$$
 $\mathsf{PHA}_n + \mathsf{RA}_{sub} \to \mathsf{PH.RA}_n$ $\mathsf{PHA}_n + \mathsf{PHA}_n \to \mathsf{PH.RA}_n$ $\mathsf{PHA}_n + \mathsf{PH.RA}_n \to \mathsf{PH.RA}_n$

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_I with U_{VS} or UU_{VS} .

$$\begin{aligned} \mathsf{RA}_I + \mathsf{U}_{vs} \to \mathsf{RU}_n & \mathsf{RA}_I + \mathsf{UU}_{vs} \to \mathsf{RUU}_n \\ & \mathsf{T} + \mathsf{U} \to \mathsf{T} & \mathsf{T} + \mathsf{U} \to \mathsf{T} \end{aligned}$$

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 12-9*).

Figure 12-9. Rendering Order in Devanagari

Character Order Glyph Order
$$KA_n + I_{vs} \rightarrow I + KA_n$$
 क $+ \circ \rightarrow$ 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 $|_{VS}$ 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.

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 l_{vs} is rendered after the rightmost such explicit virama.

$$\mathsf{TA}_d + \mathsf{ZWNJ} + \mathsf{RA}_I + \mathsf{I}_{vs} \to \mathsf{TA}_d + \mathsf{I}_{vs} + \mathsf{RA}_I$$

$$\bar{\mathsf{q}} + \boxed{\mathsf{RV}} + \bar{\mathsf{t}} + \bar{\mathsf{t}} \to \bar{\mathsf{q}}\bar{\mathsf{t}}\bar{\mathsf{t}}$$

Alternative Forms of Cluster-Initial RA. In addition to reph (rule R2) and eyelash (rule R5a), a cluster-initial RA may also take its nominal form while the following consonant takes a reduced form. This behavior is required by languages that make a morphological distinction between "reph on YA" and "RA with reduced YA", such as Braj Bhasha. To trigger this behavior, a ZWJ is placed immediately before the virama to request a reduced form of the following consonant, while preventing the formation of reph, as shown in the third example below.

$$\begin{array}{ccc}
\mathbf{7} + \mathbf{7} + \mathbf{7} & \rightarrow & \mathbf{2} \\
\mathbf{7} + \mathbf{7} + \mathbf{7} + \mathbf{2} & \rightarrow & \mathbf{7} \\
\mathbf{7} + \mathbf{7} + \mathbf{7} + \mathbf{7} & \rightarrow & \mathbf{7} \\
\mathbf{7} + \mathbf{7} + \mathbf{7} + \mathbf{7} & \rightarrow & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} \\
\mathbf{7} + \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7} & \mathbf{7}$$

Similar, special rendering behavior of cluster-initial RA is noted in other scripts of India. See, for example, "Interaction of Repha and Ya-phalaa" in *Section 12.2, Bengali (Bangla)*, "Reph" in *Section 12.7, Telugu*, and "Consonant Clusters Involving RA" in *Section 12.8, Kannada*.

Sample Half-Forms. Table 12-4 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 12-4. Sample Devanagari Half-Forms

क + ्	+	ZW	\rightarrow	क
ख + ्	+	ZW J	\rightarrow	ख
ग + ्	+	ZW J	\rightarrow	1
घ + ्	+	ZW J	\rightarrow	૩
च + ्	+	ZW J	\rightarrow	٦
ज + ्	+	ZW J	\rightarrow	7
झ + ्	+	ZW J	\rightarrow	इ
স + ্	+	ZW J	\rightarrow	6
ण + ्	+	ZW J	\rightarrow	σ
त + ्	+	ZW J	\rightarrow	7
थ + ्	+	ZW J	\rightarrow	દ
ध + ्	+	ZW J	\rightarrow	જ

न + ्	+	zw	\rightarrow	-
प + ्	+	zw	\rightarrow	τ
फ + ्	+	zw	\rightarrow	फ्
ब + ्	+	zw	\rightarrow	10
भ + ्	+	zw	\rightarrow	£
म + ्	+	zw	\rightarrow	Ŧ
य + ्	+	zw	\rightarrow	ट
ल + ्	+	zw	\rightarrow	₹
व + ्	+	zw	\rightarrow	c
श + ्	+	zw	\rightarrow	ठ
ष + ्	+	zw	\rightarrow	$\boldsymbol{\sigma}$
स + ्	+	zw	\rightarrow	£

Sample Ligatures. Table 12-5 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 12-5.
 Sample Devanagari Ligatures

क +	् +	क →	露
क +	् +	$\overline{n} \to$	क्त
क +	् +		क्र
	् +	ॺ →	क्ष
ङ +	् +	\overline{a} $ o$	ङ्क
ङ +	् +	ख →	ड्खं
ङ +	् +	ग →	ङ्ग
ङ +	् +	घ →	ङ्घ
স +	् +	$\overline{\mathbf{J}}$ \rightarrow	ञ्ज
স +	् +	স →	ज्ञ
द +	् +	घ →	द्ध
द +	् +	$\overline{\zeta} \rightarrow$	ह
द +	् +	ध →	ब्द्र
द +	् +	ৰ $ o$	द्ध
द +	् +	$\Psi \rightarrow$	ब्द
द +	् +	म →	द्म
	् +	य →	द्य
	् +	ਕ →	द्व
र +	् +	ੋ →	ट्ट

ट	+	্	+	ਰ →	ट्ठ
ਠ	+	্	+	$\mathcal{Q} \rightarrow$	ट्ठ
ड	+	্	+	ग →	ङ्ग
ड	+	্	+	ड →	डु
ड	+	্	+	$\mathcal{G} \rightarrow$	ङु
त	+	্	+	$\overline{d} \to$	त्त
त	+	্	+		त्र
न	+	্	+		न्न
फ	+	্	+		फ्र
श	+	্	+		श्र
ह	+	্	+	म →	ह्म
ह	+	্	+	य →	ह्य
ह	+	্	+	$\overline{\sigma}$ →	ह्न
ह	+	্	+	ਕ →	ह्न
ह	+	ૃ		\rightarrow	ह
र	+	ુ		\rightarrow	रु
र	+	ૄ		\rightarrow	रू
स	+	্	+	त्र →	स्त्र

Ligature Forms for Ra + Vocalic Liquids. The phonological sequence /r vocalic_r/, expressed with the character sequence <U+0930 ra, U+0943 vocalic_r>, can graphically appear as either of two forms, as shown in the first row of *Table 12-6*. It may appear as the full independent vowel form of the vocalic_r, with a superscript *repha* form of the ra (V + RA_{sup}): $\frac{1}{R}$. Alternatively, it may appear as the full letter form of the ra with the subscript, dependent form of the vocalic_r ($RA_n + V_{VS}$): $\frac{1}{R}$. Similarly, the phonological sequences with the other vocalic sounds (rr, l, ll) have two written forms, as shown in *Table 12-6*.

Table 12-6. 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.

The practice of writing these phonological sequences as a *reph* on an independent vocalic liquid letter is also observed in other Indic scripts, such as Bengali, Gujarati, Oriya, Telugu, Kannada, and Bhaiksuki.

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 12-7*. 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 12-7. 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 12-8*. Marathi also makes use of the "eyelash" form of the letter RA, as discussed in rule R5.

Code Point	Hindi	Marathi	Nepali
U+091D јна	झ	झ	भ
U+0932 la	ल	ਲ	ल
U+0936 sна	श	হা	श
U+096B FIVE	ų	ų	Ą
U+096E eight	۷	۷	ζ
U+096F nine	९	٩	ረ

Table 12-8. Marathi and Nepali Allographs

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 Ben-

gali, 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 13.4*, *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 0.

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 ooe, 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.

Bodo, Dogri, and Maithili. The orthographies of the Bodo, Dogri, and Maithili languages of India make use of U+02BC "'" 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 "' "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 " '" MODIFIER LETTER APOSTROPHE in Bodo, Dogri, and Maithili are shown in *Figure 12-10*. The Maithili examples show the same sentence, first in full form, and then using U+02BC to show truncation of words.

Language	Examples	Meaning
Bodo	खर'	head
Dodo	दख'ना	type of Bodo dress
Dogri	ख'ल्ल	down
2 0811	ति'लकना	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 12-11*.

Figure 12-11. Use of Avagraha in Dogri

Example	Meaning
तला	sole
तलाऽ	pond

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 12-9*.

Table 12-9. 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

Letter Short a. The character U+0904 DEVANAGARI LETTER SHORT A is used to denote a short *e* in the Awadi language, an Indo-Aryan language spoken in the north Indian state of Uttar Pradesh and southern Nepal. A publisher in Lucknow, Uttar Pradesh also uses it in Hindi translations and Devanagari transliterations of the Kannada, Telugu, Tamil, Malayalam and Kashmiri languages.

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 12-10* 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.

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

Table 12-10. Prishthamatra Orthography

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-Saṃhitā) 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*.

A number of marks for nasalization are also included in the block. U+1CD3 VEDIC SIGN NIHSHVASA is a breaking mark which separates sections of Samavedic singing between

which a pause is disallowed. The block also contains several Vedic signs for *ardhavisarga*, *jihvamuliya*, *upadhmaniya* and *atikrama*.

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 12-11* 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.

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

Table 12-11. Bengali Vowel Letters

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 12-12*. 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, U+09CB BENGALI VOWEL SIGN O and U+09CC BENGALI VOWEL SIGN AU. To provide com-

For Use Description আ <0985, 09D7> vowel letter aw ভা <0989, 09BE> vowel letter ua

Table 12-12. Diphthong Vowel Letters in Kokborok

patibility 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 ◀BENGALI LETTER RA WITH MIDDLE DIAGONAL, and the Assamese letter *wa* is represented by U+09F1 록 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 \$\tilde{\pi}\$ BENGALI LETTER KA, U+09CD \$\Q \tilde{\Q}\$ BENGALI SIGN VIRAMA, U+09B7 \$\tilde{\q}\$ 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 12-13*.

Table 12-13. Assamese Consonant-Vowel Combinations

	Code Points	Ligated	Non-ligated
ru	<09F0, 09C1>	ৰু	ৰু
rū	<09F0, 09C2>	ৰ	ৰু

Rendering Behavior. Like other Brahmic scripts in the Unicode Standard, Bengali uses the hasant to form conjunct characters. For example, U+09B8 ञ BENGALI LETTER SA + U+09CD © BENGALI SIGN VIRAMA + U+0995 क BENGALI LETTER KA yields the conjunct अ 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 12-14*.

	Code Points	Ligated	Non-ligated
gu	<0997, 09C1>	3	গু
ru	<09B0, 09C1>	রু	রু
rū	<09B0, 09C2>	র	রূ
śи	<09B6, 09C1>	শু	শু
hu	<09B9, 09C1>	ছ	र्
hṛ	<09B9, 09C3>	হ্ন	ङ्

Table 12-14. Bengali Consonant-Vowel Combinations

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 12-12*.

Figure 12-12. Requesting Bengali Consonant-Vowel Ligature

$$\begin{array}{cccc}
\uparrow & + & & \rightarrow & \uparrow \\
0997 & 09C1 & & & ga+u
\end{array}$$

$$\uparrow & + & \boxed{zw} & + & \rightarrow & \mathfrak{G}$$

$$0997 & 200D & 09C1 & & aa+u \text{ ligature}$$

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 12-13*.

Figure 12-13. Blocking Bengali Consonant-Vowel Ligature

Khiya. The letter ቕ, 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 ቕ BENGALI LETTER KA, U+09CD ் BENGALI SIGN VIRAMA, U+09B7 ቕ BENGALI LETTER SSA>.

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 12-14*.

Figure 12-14. Bengali Syllable tta

It is also possible for the sequence < ta, ta > to be displayed with a full ta glyph combined with a ta glyph, followed by another full ta glyph \overline{y} . 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 12-14*.

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 12-14*.

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

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

Further examples:

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
$$+$$
 \circ $+$ ম \rightarrow ম as in কম (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:

$$\overline{\Phi} + \overline{\psi} + \overline{V} \rightarrow \overline{\Phi}$$
 as in বাক্য (bakyô)

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

$$\overline{A} + 0 + \overline{A} \rightarrow \overline{A}$$
 or \overline{A}

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:

$$\overline{A} + \overline{\downarrow} + \overline{4} \rightarrow \overline{4}$$
09B0 09CD 09AF
$$\overline{A} + \overline{\underline{2}} \overline{\underline{W}} + \overline{\downarrow} + \overline{4} \rightarrow \overline{A} \overline{\underline{J}}$$
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>*.

Jihvamuliya and Upadhmaniya. In Bengali, the voiceless velar and bilabial fricatives are represented by U+1CF5 X VEDIC SIGN JIHVAMULIYA and U+1CF6 W VEDIC SIGN UPADHMANIYA, respectively. When the signs appear with a following homorganic voiceless stop consonant, they can be rendered in a font as a stacked ligature without a virama:

$$\begin{array}{ccc} \boxed{\mathbb{X}} + \overline{\Phi} \to \overset{\mathsf{X}}{\Phi} \\ \boxed{\mathbb{W}} + \overline{\Psi} \to \overset{\mathsf{W}}{\Psi} \end{array}$$

The sequences can also be represented linearly by inserting a U+200C ZERO WIDTH NON-JOINER after the *jihvamuliya* or *upadhmaniya*, but before the following consonant:

$$\begin{array}{cccc} \left[X \right] + & \left[\begin{matrix} zw \\ NJ \end{matrix} \right] + & \overline{\Phi} & \rightarrow & X\overline{\Phi} \\ \\ \hline \omega & + & \left[\begin{matrix} zw \\ NJ \end{matrix} \right] + & \overrightarrow{P} & \rightarrow & \omega \overrightarrow{P} \end{array}$$

Dependent vowel signs can also be added to the stack or linear sequence. Consonant clusters containing U+1CF5 VEDIC SIGN JIHVAMULIYA and U+1CF6 VEDIC SIGN UPADHMANIYA can occur with more than two consonants, such as *hkra* and *hpra*.

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 12-15*.

Table 12-15. Use of Apostrophe in Bangla

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

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 Чॅdl 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 ч GURMUKHI LETTER YA. However, because its usage is relatively rare and not entirely predictable, it is encoded as a separate character. Some modern fonts render YAKASH with the glyph ♀, which varies from the traditional shape found in the code charts. 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 it 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 데ਰ빗네 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 ਉਮਾਹਾ and both the vowel signs ੋ and ੁ attached to a consonant, as in ਗੁੱਥਿੰਦ *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 ਗ੍ਰਾਿਨ *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 00 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 12-16* 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.

For	Use	Do Not Use
ਆ	0A06	<0A05, 0A3E>
ਇ	0A07	<0A72, 0A3F>
ਈ	0A08	<0A72, 0A40>
₿	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 va 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 12-17.

Table 12-17. Gurmukhi Conjuncts

ਮ +	্	+ ਹ	→ ਮ੍ਹ	(mha)	pairin ha
ਪ +	্	+ ਰ	\rightarrow ψ	(pra)	pairin ra
ਦ +	্	+ ਵ	→ ਦੂ	(dva)	pairin va
ਦ +	্	+ ਯ	→ ਦੁਯ	(dya)	addha ya

Other letters behaved similarly in old inscriptions, as shown in *Table 12-18*.

Table 12-18. Additional Pairin and Addha Forms in Gurmukhi

ਸ	+	্	+	ਗ	→ ਸ	(sga)	pairin ga
ਸ	+	্	+	ਚ	→ ਸ਼੍ਰ	(sca)	pairin ca
ਸ	+	্	+	ट	→ ਸ਼੍ਰ	(stta)	pairin tta
ਸ	+	্	+	ਠ	ightarrow ਸ਼੍ਰ	(sttha)	pairin ttha
ਸ	+	্	+	ਤ	\rightarrow ਸੂ	(sta)	pairin ta
ਸ	+	্	+	ਦ	ightarrow ਸ਼ੂ	(sda)	pairin da
ਸ	+	্	+	ਨ	ightarrow ਸ਼੍ਰ	(sna)	pairin na
ਸ	+	্	+	ਥ	\rightarrow Ξ	(stha)	pairin tha
ਸ	+	্	+	ਯ	\rightarrow $\stackrel{r_{i}}{H}$	(sya)	pairin ya
ਸ	+	্	+	ਥ	→ ਸਭ	(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 12-19*. 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 12-19. Use of Joiners in Gurmukhi

ਸ	+	্	+	ਵ			\rightarrow	ਸੂ	(sva)
ਰ	+	্	+	ਵ			\rightarrow	ਰ੍ਵ	(rva)
						ਵ			(sva)
ਰ	+	্	+	zw	+	ਵ	\rightarrow	ਵੰ	(rva)
ਸ	+	্	+	ZW N J	+	ਵ	\rightarrow	ਸ਼ਵ	(sva)
					+	₹	\rightarrow	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 12-20* 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.

For	Use	Do Not Use
આ	0A86	<0A85, 0ABE>
ઍ	0A8D	<0A85, 0AC5>
એ	0A8F	<0A85, 0AC7>
એ	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>

Table 12-20. Gujarati Vowel Letters

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+0ABO GUJARATI LETTER RA takes special forms when it combines with other consonants, as shown in *Table 12-21*.

Marks for Transliteration of Arabic. The combining marks encoded in the range U+0AFA..U+0AFF are used for the transliteration of the Arabic script into Gujarati. This system of transliteration was devised in the late 19th century, and is used by Ismaili Khoja communities. These marks occur both in manuscripts and in printed materials.

The three forms of *nukta* encoded in the range U+0AFD..U+0AFF are diacritics, placed above regular Gujarati letters to create new letters corresponding to Arabic letters for non-Gujarati sounds. U+0AFF GUJARATI SIGN TWO-CIRCLE NUKTA ABOVE is used only with U+0A9D GUJARATI LETTER JHA, to transliterate the Arabic *zah*. U+0AFE GUJARATI SIGN

Table 12-21. Gujarati Conjuncts

CIRCLE NUKTA ABOVE is used with U+0A9D GUJARATI LETTER JHA to transliterate the Arabic *thal* and with U+0AB8 GUJARATI LETTER SA to transliterate the Arabic *theh*. U+0AFD GUJARATI SIGN THREE-DOT NUKTA ABOVE occurs with a number of different Gujarati letters, to transliterate a variety of Arabic letters.

U+0AFA GUJARATI SIGN SUKUN, U+0AFB GUJARATI SIGN SHADDA, and U+0AFC GUJARATI SIGN MADDAH are used to transliterate the Arabic *sukun*, *shadda*, and *maddah above*, respectively. These marks may be applied to a Gujarati letter which also uses one of the three above-base *nukta* diacritic marks. In such cases, the *nukta* occurs first in the combining sequence, followed by the *sukun*, *shadda*, or *maddah* mark. However, instead of being rendered above the *nukta* mark on the letter, the *sukun*, *shadda*, or *maddah* mark is rendered to the left of the *nukta* mark.

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*.

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 12-22* 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.

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

Table 12-22. Oriya Vowel Letters

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:

The virama is also used to form conjunct consonants, as shown in *Table 12-23*.

Table 12-23. Oriya Conjuncts

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:

$$0 + 0 + 0 \rightarrow 0 \text{ (rpa)}$$

$$0 + 0 \rightarrow 0 \text{ (pra)}$$

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

$$\mathbb{C}^2 + \mathbb{C}^2 +$$

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 12-24*.

Table 12-24. Oriya Vowel Placement

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

In a consonant conjunct, subjoined U+0B2C \circ окіча Letter ва is usually—but not always—pronounced [wa]:

U+0B15 କି
$$ka$$
 + U+0B4D ୍ $virama$ + U+0B2C ବି $ba \rightarrow \mathfrak{P}_{0}$ [kwa] U+0B2E ମି ma + U+0B4D ୍ $virama$ + U+0B2C ବି $ba \rightarrow \mathfrak{P}_{0}$ [mba]

The extended Oriya letter U+0B71 $\mbox{\ensuremath{\ensuremath{\varnothing}}}$ Oriya letter W+0B71 $\mbox{\ensuremath{\ensuremath{\otimes}}}$ Oriya letter U+0B71 $\mbox{\ensuremath{\ensuremath{\otimes}}}$ Oriya to have originally been devised as a ligature of $\mbox{\ensuremath{\ensuremath{\otimes}}}$ or and $\mbox{\ensuremath{\ensuremath{\otimes}}}$ 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 $\mbox{\ensuremath{\ensuremath{\otimes}}}$, $\mbox{\ensuremath{\ensuremath{\otimes}}}$, and $\mbox{\ensuremath{\ensuremath{\otimes}}}$.

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 12-25*. In the absence of a joiner, both the non-ligated and the ligated forms are acceptable renderings.

Table 12-25. Ligation for the Syllable *om*

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 Irula, 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 $\Box^2 = pha$, $\Box^3 = ba$, and $\Box^4 = bha$. 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 of conjunct consonant forms, although a few conventional conjuncts are used.

Virama (Pulli). 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 12-4.

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 *pulli* in Tamil. The *pulli* 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 12-15*, but nowadays tend to be displayed using an explicit *puḷḷi* as well.

Figure 12-15. Kssa Ligature in Tamil

க
$$+$$
 ் $+$ ஷ \rightarrow க்ஷ k șa

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

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 12-26* 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 12-26. Tamil Vowel Letters

For	Use	Do Not Use
ஆ	0B86	<0B85, 0BC2>

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 60 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 *Figure 12-16*.

Figure 12-16. Tamil Vowel Reordering

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 12-17*.

Figure 12-17. Tamil Two-Part Vowels

$$\bigcirc$$
 ∩ 0BCA \equiv \bigcirc 0BC6+0BBE
 \bigcirc ∩ 0BCB \equiv \bigcirc 0BC7+0BBE
 \bigcirc ∩ 0BCC \equiv \bigcirc 0BC6+0BD7

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.

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 *Figure 12-18*.

Figure 12-18. Tamil Vowel Splitting and Reordering

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 12-19*.

Figure 12-19. Vowel Reordering Around a Tamil Conjunct

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.

Confusable Vowels. U+0B94 TAMIL LETTER AU and U+0BCC TAMIL VOWEL SIGN AU are visually indistinguishable from two semantically unrelated sequences, as shown in *Figure 12-20*. In the decompositions of these two vowel characters, the rightmost part is represented as the character U+0BD7 TAMIL AU LENGTH MARK, which looks exactly like the separate character, U+0BB3 TAMIL LETTER LLA.

Figure 12-20. Confusable Vowels in Tamil

ଡ଼େଶୀ
$$0B94 \equiv 9 + 1 = 0B92 + 0BD7 \neq 9 + 1 = 0B92 + 0BB3$$

େଶୀ $0BCC \equiv 1 + 1 = 0BC6 + 0BD7 \neq 1 = 0BC6 + 0BB3$

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 *pulli*.

Ligatures with Vowel i. The vowel signs i \cap and ii form ligatures with the consonant $tta \perp$ as shown in examples 1 and 2 of *Figure 12-21*. 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 12-21*.

Figure 12-21. Tamil Ligatures with *i*

①
$$\Box + \cap \rightarrow \downarrow \downarrow ti$$
② $\Box + \circ \rightarrow \Box ti$
③ $\odot + \cap \rightarrow \odot \Box i$
④ $\odot + \circ \rightarrow \odot \Box i$

Ligatures with Vowel u. The vowel signs u and uu normally ligate with their consonant, as shown in *Table 12-27*. 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 12-27. Tamil Ligatures with *u*

x	$x + \bigcirc$	x + ಼
க	ஞ	Б
囮	回	நு
ச	ᅲ	ம
ஞ	ஞ	ஞூ
L	டு	Ĝ

x	x +	x + ु.
П	Т	±)
Ф	மு	மூ
Ш	Щ	J)
Л	ரு	ரு
ഉ	று	றூ

x	x + ○	x + ु.
ண	ത്ത്വ	ணூ
த	து	தூ
ந	நு	நூ
ன	ത്വ	னூ

Table 12-27. Tamil Ligatures with u (Continued)

x	x + ○	x + ु
ಉ	லு	லூ
តា	ளு	ளு
மி	9	ழ
ഖ	ഖ	வூ

With certain consonants, \mathfrak{P}_{0} , \mathfrak{A}_{0} , \mathfrak{P}_{0} , \mathfrak{P}_{0} , and the conjunct \mathfrak{P}_{0} , the vowel signs u and uu take a distinct spacing form, as shown in *Figure 12-22*.

Figure 12-22. Spacing Forms of Tamil *u*

$$\mathfrak{g} + \mathfrak{g} \to \mathfrak{g}^{\mathbb{D}} ju$$
 $\mathfrak{g} + \mathfrak{g} \to \mathfrak{g} \mathfrak{D} j\bar{u}$

Ligatures with ra. Based on typographical preferences, the consonant ra Π may change shape to Π , when it ligates. Such change, if it occurs, will happen only when the Π form of U+0BB0 Π Tamil Letter ra would not be confused with the nominal form Π of U+0BBE Tamil vowel sign as (namely, when Π is combined with $\dot{}$, $\dot{}$, or $\ddot{}$). This change in shape is illustrated in *Figure 12-23*.

Figure 12-23. Tamil Ligatures with *ra*

However, various governmental bodies mandate that the basic shape of the consonant $ra \, \Pi$ 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 $ra \, \Pi$. Sri Lanka, on the other hand, specifies the use of the changed forms shown in *Figure 12-24*.

Tamil Ligature shri. Prior to Unicode 4.1, the best mapping to represent the ligature *shri* was to the sequence <U+0BB8, U+0BCD, U+0BB0, U+0BC0>. Unicode 4.1 in 2005 added the character U+0BB6 TAMIL LETTER SHA and as a consequence, the best mapping became

<U+0BB6, U+0BCD, U+0BB0, U+0BC0>. Due to slow updates to implementations, both representations are widespread in existing text. Therefore, treating both representations as equivalent sequences is recommended. *Figure 12-24* shows the two sequences.

Figure 12-24. Tamil Ligatures for shri

Ligatures with aa in Traditional Tamil Orthography. In traditional Tamil orthography, the vowel sign aa \Box optionally ligates with 600, 60, or \bigcirc , as illustrated in *Figure 12-25*.

Figure 12-25. Traditional Tamil Ligatures with *aa*

These ligations also affect the right-hand part of two-part vowels, as shown in Figure 12-26.

Figure 12-26. Traditional Tamil Ligatures with *o*

Ligatures with ai in Traditional Tamil Orthography. In traditional Tamil orthography, the left-side vowel sign *ai* 600 is also subject to a change in form. It is rendered as 20 when it occurs on the left side of 6001, 601, or 611, as illustrated in *Figure 12-27*.

Figure 12-27. Traditional Tamil Ligatures with *ai*

By contrast, in modern Tamil orthography, this vowel does not change its shape, as shown in *Figure 12-28*.

Figure 12-28. Vowel *ai* in Modern Tamil

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. Note that some digits are confusable with letters, as shown in *Table 12-28*. In some Tamil fonts, the digits for two and eight look exactly like the letters *u* and *a*, respectively. In other fonts, as shown here, the shapes for the digits two and eight are adjusted to minimize confusability.

Table 12-28. Confusable Tamil Digits

U+0BE7 க tamil digit one	U+0B95 க tamil letter ka
U+0BE8 2_ TAMIL DIGIT TWO	U+0B89 2_ tamil letter u
U+0BED 6T TAMIL DIGIT SEVEN	U+0B8E 6T tamil letter e
U+0BEE 🔑 TAMIL DIGIT EIGHT	U+0B85 🎒 tamil letter a

Tamil also has distinct numerals for ten, one hundred, and one thousand at U+0BF0..U+0BF2 used for historical numbers.

Use of Nukta. In addition to Tamil, several other languages of southern India are written using the Tamil script. For example, Irula is written with the Tamil script. Some of these languages contain sounds distinct from those normally written for the Tamil language. In

such cases, the writing systems of these languages apply diacritic nukta marks to Tamil letters to represent their distinct sounds. For example, Irula uses a double dot nukta below for some sounds. That nukta can be represented with U+1133C GRANTHA SIGN NUKTA.

Tamil Supplement: U+11FC0-U+11FFF

The Tamil Supplement block contains a set of fractions in the range U+11FC0..U+11FD4 used for generic measurement and calculations and for money. The block also includes symbols indicating various forms of measurement, old units of currency, agricultural and clerical signs, and other miscellaneous abbreviations. Most characters in this block are no longer in use, but a few appear in traditional contexts, such as on marriage invitations printed in a traditional format.

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 12-29* 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 12-29*, 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 12-29.
 Tamil Vowels, Consonants, and Syllables

╙		Α	AA	I	II	U	UU	E	EE	Al	0	00	AU
	0B83	அ 0B85	ஆ 0886	இ 0B87	FT: 0B88	ഉ 0B89	<u>ഉണ്</u> _{0B8A}	6T 0B8E	ஏ 0B8F	ഇ 0B90	ஒ 0B92	⊚ 0B93	ஒள 0B94
	க்	க	கா	கி	තී	கு	ቻዉ	கெ	கே	கை	கொ	கோ	கௌ
K	0B95 0BCD	0B95	0B95 0BBE	0B95 0BBF	0B95 0BC0	0B95 0BC1	0B95 0BC2	0B95 0BC6	0B95 0BC7	0B95 0BC8	0B95 0BCA	0B95 0BCB	0B95 0BCC
NG	ାଲ	Ш	ΙШΠ	ஙி	ឃ្នេ	Щ	阻	ஙெ	ஙே	ஙை	ஙொ	ஙோ	ஙெள
ING	0B99 0BCD	0B99	0B99 0BBE	0B99 0BBF	0B99 0BC0	0B99 0BC1	0B99 0BC2	0B99 0BC6	0B99 0BC7	0B99 0BC8	0B99 0BCA	0B99 0BCB	0B99 0BCC
С	ச்	æ	σπ	Ð	ቆ	சு	சூ	செ	சே	சை	சொ	சோ	சௌ
Ľ	0B9A 0BCD	0B9A	0B9A 0BBE	0B9A 0BBF	0B9A 0BC0	0B9A 0BC1	0B9A 0BC2	0B9A 0BC6	0B9A 0BC7	0B9A 0BC8	0B9A 0BCA	0B9A 0BCB	0B9A 0BCC
NY	ஞ்	ஞ	ஞா	ஞி	ණී	ஞ	ஞூ	ெஞ	ஞே	ത്രെ	ஞொ	ஞோ	ஞௌ
	0B9E 0BCD	0B9E	0B9E 0BBE	0B9E 0BBF	0B9E 0BC0	0B9E 0BC1	0B9E 0BC2	0B9E 0BC6	0B9E 0BC7	0B9E 0BC8	0B9E 0BCA	0B9E 0BCB	0B9E 0BCC
TT	Ŀ	L	டா	IД	<u>lo</u>	$^{\circ}$	G	டெ	டே	டை	டொ	டோ	டௌ
L	0B9F 0BCD	0B9F	0B9F 0BBE	0B9F 0BBF	0B9F 0BC0	0B9F 0BC1	0B9F 0BC2	0B9F 0BC6	0B9F 0BC7	0B9F 0BC8	0B9F 0BCA	0B9F 0BCB	0B9F 0BCC
NN	ண்	ண	ணா	ணி	ഞ്ഞ്	ഞ്ഞ	ணூ	ணெ	ணே	ത്തെ	ணொ	ணோ	ணௌ
	0BA3 0BCD	0BA3	0BA3 0BBE	0BA3 0BBF	0BA3 0BC0	0BA3 0BC1	0BA3 0BC2	0BA3 0BC6	0BA3 0BC7	0BA3 0BC8	0BA3 0BCA	0BA3 0BCB	0BA3 0BCC
 -	த்	த	தா	தி	ජි	து	தூ	தெ	தே	தை	தொ	தோ	தௌ
т	. 0ВА4 0ВСD	Б 0ВА4	தா 0BA4 0BBE	தி 0BA4 0BBF	ජි 0BA4 0BC0	Б Ј 0ВА4 0ВС1	தூ 0BA4 0BC2	தெ 0BA4 0BC6	தே 0BA4 0BC7	தை 0BA4 0BC8	தொ 0BA4 0BCA	தோ 0BA4 0BCB	தௌ ^{0BA4} 0BCC
-	0BA4	_	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4	0BA4
T	0BA4 0BCD	0BA4	0BA4 0BBE	0BA4 0BBF	0BA4 0BC0	0BA4 0BC1	0BA4 0BC2	0BA4 0BC6	0BA4 0BC7	0BA4 0BC8	0BA4 0BCA	0BA4 0BCB	0BA4 0BCC
N	OBA4 OBCD	ова4	0BA4 0BBE நா 0BA8	OBA4 OBBF	0BA4 0BC0 நீ 0BA8	OBA4 OBC1	OBA4 OBC2 DIT OBA8 OBC2	0ВА4 0ВС6 О Б 0ВА8	0ВА4 0ВС7 С Б 0ВА8	0BA4 0BC8 のந 0BA8	0ВА4 0ВСА ОБП 0ВА8	ова4 овсв СБ П ова8	0BA4 0BCC Бр ள 0BA8
-	OBA4 OBCD	ОВА4	OBA4 OBBE DIT OBA8 OBBE	OBA4 OBBF	OBA4 OBCO OBA8 OBCO	OBA4 OBC1 OBA8 OBC1	0BA4 0BC2 IJIT 0BA8	0ВА4 0ВС6 О <u>Б</u> 0ВА8 0ВС6	0ВА4 0ВС7 СБ 0ВА8 0ВС7	ОВА4 ОВС8 ОВС8 ОВА8 ОВС8	OBA4 OBCA	OBA4 OBCB	OBA4 OBCC
N	OBA4 OBCD DBA8 OBCD U OBAA	ова4 Б ова8	OBA4 OBBE OBA8 OBBE LIT OBAA	OBA4 OBBF OBA8 OBBF	OBA4 OBCO OBA8 OBCO OBAA	OBA4 OBC1 OBA8 OBC1 U OBAA	OBA4 OBC2 OBA8 OBC2	0BA4 0BC6 OJD 0BA8 0BC6 OLL 0BAA	0BA4 0BC7 G D 0BA8 0BC7 G L 0BAA	のBA4 のBC8 のDD のBA8 のBC8 のBU のBAA	OBA4 OBCA OBCA OBCA OBCA	ОВА4 ОВСВ СБП ОВА8 ОВСВ СШП ОВАА	OBA4 OBCC OBG OBA8 OBCC OLIM OBAA
N P	OBA4 OBCD OBA8 OBCD U OBAA OBCD	0BA4 D	OBA4 OBBE DIT OBA8 OBBE LIT OBAA OBBE	OBA4 OBBF OBA8 OBBF OBAA OBBF	OBA4 OBCO OBA8 OBCO OBAA OBCO	0BA4 0BC1 0BA8 0BC1 U 0BAA 0BC1	OBA4 OBC2 OBA8 OBC2 UBA8 OBC2 OBAA OBC2	OBA4 OBC6 OBBA8 OBC6 OBA8 OBC6 OBAA OBC6	0BA4 0BC7 GID 0BA8 0BC7 GLI 0BAA 0BC7	0BA4 0BC8 60) ID 0BA8 0BC8 60) LI 0BAA 0BC8	OBA4 OBCA OBCA OBA8 OBCA OBAA OBAA OBAA OBAA	OBA4 OBCB CIDIT OBA8 OBCB CLITT OBAA OBCB	OBA4 OBCC OBA8 OBCC OLIM OBAA OBCC
N P	OBA4 OBCD DBA8 OBCD U OBAA OBCD OBAA OBCD OBAA	OBA4 DOBA8 UOBAA OBAE UU	OBA4 OBBE DIT OBA8 OBBE LIT OBAA OBBE LOT OBAA	OBAA OBBF OBAS OBBF OBAA OBBF OBAE OBBF	OBAA OBCO OBAB OBCO OBAA OBCO OBAE OBCO	OBA4 OBC1 OBA8 OBC1 OBAA OBC1 OBAA OBC1 OBAE OBC1 OBAE OBC1	OBA4 OBC2 OBA8 OBC2 OBAA OBC2 OBAA OBC2 OBAE OBC2 OBAE OBC2	OBAA OBC6 OBA8 OBC6 OBAA OBC6 OBAA OBC6	OBA4 OBC7 OBA8 OBC7 GU OBAA OBC7 GLI OBAA OBC7	のBA4 のBC8 のDD のBA8 のBC8 のBU のBAA のBC8 のBC8	OBA4 OBCA	OBA4 OBCB CIDIT OBA8 OBCB CUIT OBAA OBCB CUIT OBAA	OBA4 OBCC ODG OBA8 OBCC OLIM OBAA OBCC OLIM OBAA OBCC
N P	OBA4 OBCD OBA8 OBCD OBAA OBCD OBAA OBCD OBAE OBCD	OBA4 DOBA8 U OBAA OBAE	OBA4 OBBE DIT OBA8 OBBE LIT OBAA OBBE LOTI OBAE OBBE	OBA4 OBBF OBA8 OBBF OBAA OBBF OBAA OBBF	OBA4 OBCO OBA8 OBCO OBAA OBCO OBAA OBCO OBAA OBCO	OBA4 OBC1 OBA8 OBC1 OBAA OBC1 OBAA OBC1 OBAE OBC1	OBA4 OBC2 OBA8 OBC2 OBAA OBC2 OBAA OBC2 OBAE OBC2	OBAA OBC6 OBAB OBC6 OBAA OBC6 OBAA OBC6 OBAE OBAE	OBA4 OBC7 OBA8 OBC7 CLL OBAA OBC7 CLD OBAE OBC7	ОВА4 ОВС8 ОВС8 ОВС8 ОВС8 ОВС8 ОВС8 ОВС8	OBA4 OBCA OBCA	OBA4 OBCB CIDIT OBA8 OBCB CLIT OBAA OBCB CLIT OBAA OBCB	OBA4 OBCC OIDM OBA8 OBCC OLIM OBAA OBCC OLIM OBAA OBCC OLIM OBAA OBCC
N P	OBAA OBCD DBAA OBCD OBAA OBCD OBAE OBCD OBAE OBCD OBAE OBCD OBAE OBCD	OBA4 DOBAA OBAA OBAA OBAF	OBA4 OBBE DIT OBA8 OBBE LIT OBAA OBBE LOT OBAE OBBE UIT OBAF OBBE	OBAA OBBF OBAA OBBF OBAA OBBF OBAE OBBF OBAF OBBF	OBA4 OBCO OBA8 OBCO OBAA OBCO OBAE OBCO OBAE OBCO OBAF OBCO OBAF OBCO	OBA4 OBC1 OBA8 OBC1 OBAA OBC1 OBAC OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1	OBA4 OBC2 OBA8 OBC2 OBAA OBC2 OBAA OBC2 OBAE OBC2 OBAF OBC2	OBA4 OBC6 ODD OBA8 OBC6 ODD OBAA OBC6 ODD OBAE OBC6 OBAE OBC6 OBAF OBC6	OBA4 OBC7 OBA8 OBC7 OBAA OBC7 OBAA OBC7 OBAE OBC7 OBAF OBC7 OBAF OBC7	OBA4 OBC8 OBA8 OBC8 OBAA OBC3 OBAA OBC3 OBAE OBC8 OBAE OBC8 OBAF OBC8	OBA4 OBCA OBCA OBCA OBCA OBCA OBCA OBCA OBCA	OBA4 OBCB CIJI OBAA OBCB CUIT OBAA OBCB	OBA4 OBCC Digin OBA8 OBCC GUII OBAA OBCC GUII OBAE OBCC GUII OBAF OBCC GUII OBAF OBCC GUII OBAF OBCC
N P M Y	OBA4 OBCD OBA8 OBCD OBAA OBCD OBAA OBCD OBAE OBCD OBAE OBCD OBAE OBCD	OBA4 DOBAA OBAA OBAE UL OBAF	OBA4 OBBE DIT OBAA OBBE LIT OBAA OBBE LOT OBAE OBBE UIT OBAE OBBE	OBA4 OBBF OBA8 OBBF OBAA OBBF OBAE OBBF OBAF OBAF	OBA4 OBCO OBCO OBAB OBCO OBCO OBAB OBCO OBCO	OBA4 OBC1 OBA8 OBC1 OBAA OBC1 OBAE OBC1 OBAF OBC1	OBA4 OBC2 OBA8 OBC2 OBAA OBC2 OBAA OBC2 OBAE OBC2 OBAF OBC2	OBA4 OBC6 ODD OBA8 OBC6 ODD OBAA OBC6 ODD OBAE OBC6 OBAE OBC6 OBAF OBC6	0BA4 0BC7 0BA8 0BC7 0BAA 0BC7 0BAE 0BC7 0BAF 0BC7 0BAF 0BC7	OBA4 OBC8 OBA8 OBC8 OBAA OBC8 OBAA OBC8 OBAE OBAE OBAE OBAF OBC8	OBA4 OBCA OBCA OBCA OBCA OBCA OBCA OBCA OBCA	OBA4 OBCB CIDT OBAA OBCB CLIT OBAA OBCB CLIT OBAA OBCB CLIT OBAE OBCB CUIT OBAE OBCB CUIT OBAE OBCB	OBA4 OBCC OIDM OBA8 OBCC OLIM OBAA OBCC OLIM OBAE OBCC OBAE OBCC
N P M Y	OBA4 OBCD OBA8 OBCD OBAA OBCD OBAE OBCD OBAE OBCD OBAE OBCD OBAE OBCD	OBA4 DOBAA OBAA OBAA OBAF	OBA4 OBBE DIT OBAA OBBE LIT OBAA OBBE LOT OBAE OBBE UIT OBAE OBBE UIT OBAE OBBE UIT OBAF OBBE	OBA4 OBBF OBAA8 OBBF OBAA6 OBBF OBAA6 OBBF OBAA6 OBBF OBAA6 OBBF	OBA4 OBCO OBA8 OBCO OBA6 OBCO OBAF OBCO OBAF OBCO OBAF OBCO OBCO OBAF OBCO OBO OBBO	OBA4 OBC1 OBA8 OBC1 OBAA OBC1 OBAA OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1 OBAE OBC1	OBA4 OBC2 OBA8 OBC2 OBAA OBC2 OBAE OBC2 OBAE OBC2 OBAF OBC2 OBAF OBC2	OBA4 OBC6 OJD OBAA OBC6 OLD OBAA OBC6 OLD OBAE OBC6 OBC6 OLD OBAF OBC6 OGC OGC OBC7 OBC7 OBC7 OBC7 OBC7 OBC7 OBC7 OBC	OBA4 OBC7 OBA8 OBC7 OBAA OBC7 OBAE OBC7 OBAE OBC7 OBAF OBC7 OBAF OBC7 OBAF OBC7	OBAA OBCS OBAA OBCS OBAA OBCS OBAA OBCS OBAE OBCS OBAE OBCS OBAE OBCS OBAE OBCS OBAE OBCS	OBA4 OBCA OBCA OBCA OBCA OBCA OBCA OBCA OBCA	OBA4 OBCB CIDT OBAA OBCB CUIT OBAA OBCB CUIT OBAA OBCB CUIT OBAE OBCB CUIT OBAE OBCB CUIT OBAE OBCB	OBA4 OBCC OIDM OBA8 OBCC OLIM OBAA OBCC OLIM OBAE OBCC OLIM OBAF OBCC OLIM OBAF OBCC OLIM OBAF OBCC

Table 12-29. Tamil Vowels, Consonants, and Syllables (Continued)

		Α	AA	I	II	U	UU	E	EE	Al	0	00	AU
V	வ் _{0BB5}	വ 0BB5	வா 0BB5	ബി ₀₈₈₅	ഖ് _{0BB5}	ഖ 0BB5	வூ 0BB5	0BB5	வே _{0BB5}	ബെ 0BB5	வொ	வோ 0BB5	வெள _{0BB5}
	0BCD	0000	0BBE	0BBF	0BC0	0BC1	0BC2	0BC6	0BC7	0BC8	0BCA	0BCB	0BCC
LLL	<u>і</u> р 0ВВ4	Б	Ю П 0ВВ4	ழ 0BB4	ழீ 0BB4	9 0BB4	<u>С</u> 0ВВ4	<u>நெ</u> 0BB4	ழே _{0BB4}	ழை _{0BB4}	ழொ _{0BB4}	ழோ _{0BB4}	ழௌ _{0BB4}
	0BCD	000-	OBBE	0BBF	0BC0	0BC1	0BC2	0BC6	0BC7	0BC8	0BCA	0BCB	0BCC
LL	ள்	តា	ளா	ଶୀ	ଶଂ	ளு	ளு	ளெ	ளே	ளை	ளொ	ளோ	ளெள
	0BB3 0BCD	0BB3	0BB3 0BBE	0BB3 0BBF	0BB3 0BC0	0BB3 0BC1	0BB3 0BC2	0BB3 0BC6	0BB3 0BC7	0BB3 0BC8	0BB3 0BCA	0BB3 0BCB	0BB3 0BCC
RR	ற்	ற	றா	றி	றீ	று	றூ	றெ	றே	றை	றொ	றோ	றௌ
	0BB1 0BCD	0BB1	0BB1 0BBE	0BB1 0BBF	0BB1 0BC0	0BB1 0BC1	0BB1 0BC2	0BB1 0BC6	0BB1 0BC7	0BB1 0BC8	0BB1 0BCA	0BB1 0BCB	0BB1 0BCC
NNN	ळा	ன	னா	னி	ത്	னு	னூ	னெ	னே	னை	னொ	னோ	னௌ
ININ	0BA9 0BCD	0BA9	0BA9 0BBE	0BA9 0BBF	0BA9 0BC0	0BA9 0BC1	0BA9 0BC2	0BA9 0BC6	0BA9 0BC7	0BA9 0BC8	0BA9 0BCA	0BA9 0BCB	0BA9 0BCC
J	ஐ	ജ	ஜா	ஜி	කී	ஜ	ഈ	ജെ	ලෙ	ജെ	ஜொ	ஜோ	ஜௌ
Ľ	0B9C 0BCD	0B9C	0B9C 0BBE	0B9C 0BBF	0B9C 0BC0	0B9C 0BC1	0B9C 0BC2	0B9C 0BC6	0B9C 0BC7	0B9C 0BC8	0B9C 0BCA	0B9C 0BCB	0B9C 0BCC
SH	ஶ்	w	חסט	மி	ஶீ	υυ°	ത	ஶெ	ஶே	ஶை	ஶொ	ஶோ	ஶௌ
	0BB6 0BCD	0BB6	0BB6 0BBE	0BB6 0BBF	0BB6 0BC0	0BB6 0BC1	0BB6 0BC2	0BB6 0BC6	0BB6 0BC7	0BB6 0BC8	0BB6 0BCA	0BB6 0BCB	0BB6 0BCC
SS	ல்	ல	லூ	ഖ്	െ	ஹு	മൌ	லெ	ஷே	ക്ക	ஹொ	ஹோ	தெள
	0BB7 0BCD	0BB7	0BB7 0BBE	0BB7 0BBF	0BB7 0BC0	0BB7 0BC1	0BB7 0BC2	0BB7 0BC6	0BB7 0BC7	0BB7 0BC8	0BB7 0BCA	0BB7 0BCB	0BB7 0BCC
s	ஸ்	സ	ஸா	സി	സ്	സൗ		സെ	സേ	സൈ	ஸொ	ஸோ	ஸௌ
5	0BB8 0BCD	0BB8	0BB8 0BBE	0BB8 0BBF	0BB8 0BC0	0BB8 0BC1	0BB8 0BC2	0BB8 0BC6	0BB8 0BC7	0BB8 0BC8	0BB8 0BCA	0BB8 0BCB	0BB8 0BCC
Н	ஹ்	ஹ	ஹா	ஹி	ஹீ	ஹு	ஹൗ	ஹெ	ஹേ	ഞ്ചെ	ஹொ	ஹோ	ஹௌ
Ľ	0BB9 0BCD	0BB9	0BB9 0BBE	0BB9 0BBF	0BB9 0BC0	0BB9 0BC1	0BB9 0BC2	0BB9 0BC6	0BB9 0BC7	0BB9 0BC8	0BB9 0BCA	0BB9 0BCB	0BB9 0BCC
	கூற	கூற	கூலா	கூறி	ക്കൂ	ው ያ"	ক্ষক্ত	க்ஷெ	கேழ	கைழ	க்ஷொ	க்ஷோ	க்ஷௌ
KSS	0B95 0BCD 0BB7 0BCD	0B95 0BCD 0BB7	0B95 0BCD 0BB7 0BBE	0B95 0BCD 0BB7 0BBF	0B95 0BCD 0BB7 0BC0	0B95 0BCD 0BB7 0BC1	0B95 0BCD 0BB7 0BC2	0B95 0BCD 0BB7 0BC6	0B95 0BCD 0BB7 0BC7	0B95 0BCD 0BB7 0BC8	0B95 0BCD 0BB7 0BCA	0B95 0BCD 0BB7 0BCB	0B95 0BCD 0BB7 0BCC



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 12-30* 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.

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

Table 12-30. Telugu Vowel Letters

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 §
$$ka$$
 + U+0C4D δ $virama$ + U+200C δ $virama$ zero width nonjoiner \rightarrow δ δ

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

U+0C17
$$\mbox{\ensuremath{\mbox{\'e}}} ga + \text{U+0C4D} \mbox{\ensuremath{\mbox{\oe}}} virama + \text{U+0C17} \mbox{\ensuremath{\mbox{\oe}}} ga \rightarrow \mbox{\ensuremath{\mbox{\oe}}} (gga)$$
U+0C15 $\mbox{\ensuremath{\mbox{\oe}}} ka + \text{U+0C4D} \mbox{\ensuremath{\mbox{\oe}}} virama + \text{U+0C2F} \mbox{\ensuremath{\mbox{\oe}}} ya \rightarrow \mbox{\ensuremath{\mbox{\oe}}} (kya)$
U+0C15 $\mbox{\ensuremath{\mbox{\oe}}} ka + \text{U+0C4D} \mbox{\ensuremath{\mbox{\oe}}} virama + \text{U+0C37} \mbox{\ensuremath{\mbox{\oe}}} ssa \rightarrow \mbox{\ensuremath{\mbox{\oe}}} (kşa)$

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 5, which takes its shape from the glyphs in the sequence <U+0C28 5 telugu letter na , U+0C4D 5 telugu sign virama>. Older texts display the other vowel-less form 6, called $nak\bar{a}ra$ -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 6 or the new style glyph 6. The character U+200C zero width non-joiner can be used to prevent interaction of this sequence with following consonants, as shown in Table 12-31.

Font Sequence Rendering

Old Style $na + virama = \sum_{\substack{na + virama \\ na + virama + |N| \\ |N|} + da \in \Delta$

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

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:

న్ద

U+0C30 ర
$$ra$$
 + U+0C4D δ $virama$ + U+0C2E మ $ma \rightarrow \delta_{\lambda}$ (rma)

 $na + virama + \begin{vmatrix} zw \\ NJ \end{vmatrix} + da$

na + virama + da

na + virama

New Style

All Fonts

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

U+0C30 ర
$$ra$$
 + U+0C4D δ $virama$ + U+0C2E మ $ma \rightarrow \Delta \mathcal{E}$ (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 5 $virama$ + U+200D $\boxed{^{2V}}$ ZWJ + U+0C2E మ ma \rightarrow మ ϵ

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 \fbox{ZW} ZWJ + U+0C4D \raiset{virama} + U+0C2E మ ma \rightarrow 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 F 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 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 12-32* 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. Kannada Vowel Letters

For	Use	Do Not Use
ಊ	0C8A	<0C89, 0CBE>
컌	0C94	<0C92, 0CCC>
<u> </u> ಮಾ	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 & KANNADA LETTER FA is actually an archaic Kannada letter that is transliterated in Dravidian scholarship as *z*, *l*, or *r*. 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 followed 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 12-29*.

Figure 12-29. 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 NONJOINER 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 \vec{v}^6 , 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 $\vec{\epsilon}$. The two forms are semantically identical. Fonts should render the sequence <U+0CA8 kannada letter na, U+0CCD kannada sign virama> with either the oldstyle glyph $\vec{\epsilon}$ or the new style glyph \vec{v}^6 . 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 12-33*.

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+0CBO T KANNADA LETTER RA as the first component of the consonant cluster, the letter

Font Sequence Rendering na + virama E Old Style $na + virama + \begin{vmatrix} zw \\ nJ \end{vmatrix} + da$ ೯ದ ನ್ na + virama New Style $na + virama + \begin{vmatrix} zw \\ nJ \end{vmatrix} + da$ ನ್ದ All Fonts na + virama + da ನ್ದ

Table 12-33. Rendering of Kannada *na* + *virama*

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.

$$U+0CB0$$
 ਹ $ra+U+0CCD$ ਾ $halant+U+0C95$ ਵੱ $ka \rightarrow$ ਵੱ ϵrka
 $U+0CB0$ ਹ $ra+\frac{2W}{U}+U+0CCD$ ਾ $halant+U+0C95$ ਵੱ $ka \rightarrow$ ਹੈ, rka
 $U+0C95$ ਵੱ $ka+U+0CCD$ ਾ $halant+U+0CB0$ ਹ $ra \rightarrow$ ਵੱ, 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
$$X$$
 jihvamuliya + U+0C95 $\forall ka \rightarrow X$
U+0CF2 oo upadhmaniya + U+0CAB \vec{a} pha $\rightarrow \mathcal{A}$

Dependent vowels signs can also be added to the stack:

$$U+0CF1$$
 \boxed{X} jihvamuliya + $U+0C95$ \overrightarrow{b} ka + $U+0CBF$ $\overset{9}{\circ}$ vowel sign $i \rightarrow \overset{9}{\lambda}$

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 \subseteq 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, U+0CBD 5 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 12-34* 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.

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>

Table 12-34. Malayalam Vowel Letters

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 "0", 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 "6" on the left side and "0" 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.

Two specific forms of viramas are found in historical materials. The U+0D3B MALAYALAM SIGN VERTICAL BAR VIRAMA was used to indicate a pure consonant when transliterating foreign words, while the U+0D3C MALAYALAM SIGN CIRCULAR VIRAMA was employed to indicate a pure consonant in native Malayalam texts.

Suriyani Malayalam. The Suriyani dialect of Malayalam is written using the Syriac script. It is also called Garshuni (Karshoni) or Syriac Malayalam. This usage requires eleven additional letters encoded in the Syriac Supplement block (U+0860..U+086F) to represent the sounds of Malayalam. The dialect was widely used by the St. Thomas Christians living in Kerala, India, in the 19th century.

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 12-35* 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 12-35. Malayalam Orthographic Reform

Syllable		Older Orthography	Reformed Orthography	
ku	ക + ു	∂₃	കു	
gu	ഗ + ു	S	ഗു	
chu	ഛ + ു	20	ചരു	
ju	ജ + ೃ	3	ജു	
ņи	ണ + ു	ണു	ണു	
tu	ത + ു	<u>ത</u>	തു	
nu	ന + ു	നു	നു	
bhu	ഭ + ു	ල	ഭു	
ru	ര + ു	ത്ദ	രു	
śu	ળ + ૂ	ශු	ശു	
hu	ഹ + ു	ച	ഹു	
kū	ക + ൂ	<u>ச</u> ூ	കൂ	
gū	ഗ + ၘ	2	ഗൂ	
chū	ഛ + ൂ	<u>ഛ</u>	ചരു	
jū	ಜ + ೃ	E	ಜ್ಮ	
ņū	ണ + ൂ	ഞ	ണൂ	
tū	ത + ൂ	<u></u>	തൂ	
пū	ന + ൂ	ന്ത	നൂ	
bhū	ഭ + ൂ	C	ල ු	

ഹ + ൂ

Table 12-35. Malayalam Orthographic Reform (Continued)

മ

ഹു

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 12-36* 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 12-36. Malayalam Conjuncts

ക	+	്	+	ഷ	\rightarrow	ക്ഷ	(kṣa)
ക	+	័	+	ക	\rightarrow	ക്ക	(kka)
ജ	+	័	+	ഞ	\rightarrow	ඐා	(jña)
S	+	័	+	S	\rightarrow	S	(ṭṭa)
പ	+	័	+	വ	\rightarrow	심	(ppa)
ച	+	័	+	ഛ	\rightarrow	ച്ച	(ccha)
ബ	+	័	+	ബ	\rightarrow	ബ്ബ	(bba)
ന	+	័	+	∞	\rightarrow	ന്യ	(nya)
പ	+	័	+	0	\rightarrow	വ	(pra)
ശ	+	័	+	വ	\rightarrow	പ്ര	(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 12-37*.

Table 12-37. Candrakkala Examples

പാലു്	/pālə/ milk	0D2A, 0D3E, 0D32, 0D41, 0D4D
എ്ന്നാ	/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 12-38* 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 12-38* 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 12-38* for examples. Note that such sequences are defined for Malayalam only, and are left undefined for other Indic scripts.

Table 12-38. Use of Joiners in Malayalam

ക + ്	+ (0 → (£) or (£)	(kra)
സ+്	$+$ ക \rightarrow സ്ക \circ r സ്ക	(ska)
ത + ്	$+ m \rightarrow m or m$	(tsa)
y + ँ	+ $\Omega \rightarrow \mathcal{G} \text{ or } \mathcal{G} \text{ or } \mathcal{G} $	(<u>l</u> va)
ထ + ိ	$+ \omega \rightarrow \omega$	(yya)
ക+് + 📆	+ (0 → ക്	(kra)
ക + 📆 + ്	+ 0 → 🖨	(kra)
က + 💯 + ိ	$+ \Leftrightarrow \rightarrow m$	(ska)
ത + [zw] + ്	+	(tsa)
9 + [zw] + °	$+ \Omega \rightarrow \mathcal{Y}$	(<u>l</u> va)
က + [xw] + ိ	+ (0 → (db	(kra)
9 + [xw] + °	+ OJ → YJ	(<u>l</u> va)
w + [xw] + °	+ W → WJ	(yya)

Anusvara. The anusvara can be seen multiple times after vowels, whether independent letters or dependent vowel signs, as in ഈoooo <0D08, 0D02, 0D02, 0D02, 0D02>. Vowel

signs can also be seen after digits, as in 35500 <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.

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 following discussion of *chillu* characters.) The *dot reph* form is predominantly used by those who completed elementary education in Malayalam prior to 1970.

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 Ω *rra* is normally read /<u>r</u>a/. Repetition of that sound is naturally written by repeating the letter: $\Omega\Omega$. Each occurrence can bear a vowel sign.

The same repetition of the letter rra as OO is also used for $/\underline{tt}a$, which can be unambiguously represented by Θ . The sequence of two O 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 OO in text is ambiguous in reading. The reader must generally use the context to understand if OO is read $/\underline{rara}/$ or $/\underline{tta}/$. It is only when a vowel part appears between the two O that the reading cannot be $/\underline{tta}/$. 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 OO, regardless of the reading of that text. The sequence <0D31, 0D4D, 0D31> is rendered as α . In both cases, vowels signs can be used as appropriate, as shown in *Table 12-39*.

പാററ	0D2A 0D3E 0D31 0D31	/pā <u>tt</u> a/	cockroach	
പാറ്റ	0D2A 0D3E 0D31 0D4D 0D31			
മാറെറാലി	0D2E 0D3E 0D31 0D46 0D31 0D3E 0D32 0D3F	/mā <u>tt</u> oli/	echo	
മാറ്റെലി	0D2E 0D3E 0D31 0D4D 0D31 0D46 0D3E 0D32 0D3F			
ബാറററി	0D2C 0D3E 0D31 0D31 0D31 0D3F	/bā <u>tt</u> ari/	battery	
ബാറ്ററി	0D2C 0D3E 0D31 0D4D 0D31 0D31 0D3F			
സൂറററ്	0D38 0D42 0D31 0D31 0D31 0D4D	/sū <u>r</u> a <u>tt</u> /	Surat, a town in	
സൂററ്റ്	0D38 0D42 0D31 0D31 0D4D 0D31 0D4D		Gujarat	
ടെംപററി	0D1F 0D46 0D02 0D2A 0D31 0D31 0D3F	/ṭemparari/	temporary	
ലെക്ചററോട്	0D32 0D46 0D15 0D4D 0D1A 0D31 0D31 0D4B 0D1F 0D4D	/lekcararōt/	to the lecturer	

Table 12-39. Malayalam /rara/ and /tta/

A very similar situation exists for the combination of \mathfrak{M} *chillu-n* and \mathfrak{N} *rra.* When used side by side, \mathfrak{M} can be read either $/\underline{n}\underline{r}a/$ or $/\underline{n}\underline{t}a/$, while stacked \mathfrak{M} is always read $/\underline{n}\underline{t}a/$.

The sequence <0D7B, 0D31> is rendered as not, regardless of the reading of that text. The sequence <0D7B, 0D4D, 0D31> is rendered as not. In both cases, vowels signs can be used as appropriate, as shown in *Table 12-40*.

Table 12-40. Malayalam /nr/ and /nt/

ആൻേറാ	0D06 0D7B 0D47 0D31 0D3E	/āntō/	a proper name
ആന്റോ	0D06 0D7B 0D4D 0D31 0D47 0D3E		
എൻറോൾ	0D0E 0D7B 0D31 0D47 0D3E 0D7E	/en <u>r</u> ōl/	enroll

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 prepared 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 12-41 were represented in legacy text by <virama, ZWJ> sequences. The other chillu forms are only represented as atomically encoded chillu characters.

Visual **Legacy Representation (5.0)** Preferred Representation NNA, VIRAMA, ZWJ ൺ 0D7A malayalam letter chillu nn 0D23, 0D4D, 200D NA, VIRAMA, ZWJ ൻ 0D7B malayalam letter chillu n 0D28, 0D4D, 200D RA, VIRAMA, ZWJ ര 0D7C MALAYALAM LETTER CHILLU RR 0D30, 0D4D, 200D LA, VIRAMA, ZWJ ൽ 0D7D MALAYALAM LETTER CHILLU L 0D32, 0D4D, 200D LLA, VIRAMA, ZWJ ൾ 0D7E MALAYALAM LETTER CHILLU LL 0D33, 0D4D, 200D ക്ഷ undefined 0D7F malayalam letter chillu k

Table 12-41. Atomic Encoding of Malayalam Chillus

Malayalam Numbers and Punctuation

Archaic Numbers. The archaic numbering system for Malayalam included numbers for 10, 100, and 1000, as well as signs for fractions. Many Malayalam-specific fraction signs are encoded in the Malayalam block. Malayalam also made use of the fraction signs for one quarter, one half, and three quarters encoded in the Common Indic Number Forms block.

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.