Chapter 9

South Asian Scripts-I

The following South Asian scripts are described in this chapter:

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

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

### 9.1 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, Kanaui, 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.

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.

**Principles of the Devanagari Script**

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

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

**Consonant Letters.** Each consonant letter represents a single consonantal sound but also has the peculiarity of having an inherent vowel, generally the short vowel /a/ in Devanagari and the other Indic scripts. Thus U+0915 DEVANAGARI LETTER KA represents not just /k/ but also /ka/. In the presence of a dependent vowel, however, the inherent vowel associated with a consonant letter is overridden by the dependent vowel.
Consonant letters may also be rendered as *half-forms*, which are presentation forms used to depict the initial consonant in consonant clusters. These half-forms do not have an inherent vowel. Their rendered forms in Devanagari often resemble the full consonant but are missing the vertical stem, which marks a syllabic core. (The stem glyph is graphically and historically related to the sign denoting the inherent /a/ vowel.)

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

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

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

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

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

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

**Vowel Letters.** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. Table 9-1 shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

**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 9-1.

**Figure 9-1. Dead Consonants in Devanagari**

$$TA_n + \text{VIRAMA}_n \rightarrow TA_d$$

$$\text{त} + \text{०} \rightarrow \text{ट}$$

**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 9-2).

**Figure 9-2. Conjunct Formations in Devanagari**

1. $GA_d + DHA_l \rightarrow GA_h + DHA_n$
2. $KA_d + KA_l \rightarrow K.KA_n$
3. $KA_d + SSA_l \rightarrow K.SSA_n$
4. $RA_d + KA_l \rightarrow KA_l + RA_{sup}$

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 9-2, the dead consonant $RA_d$ is depicted with the nonspacing combining mark $RA_{sup}$ (repha).

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

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

For example, in Figure 9-3, the use of zero width non-joiner prevents the default formation of the conjunct form $\text{क्र्न} (K.SSA_n)$.

**Figure 9-3. Preventing Conjunct Forms in Devanagari**

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

$\text{क्र्} + \text{्न} + ष \rightarrow \text{कष}$

**Explicit Half-Consonants.** When a dead consonant participates in forming a conjunct, the dead consonant form is often absorbed into the conjunct form, such that it is no longer distinctly visible. In other contexts, the dead consonant may remain visible as a half-consonant form. In general, a half-consonant form is distinguished from the nominal consonant form
by the loss of its inherent vowel stem, a vertical stem appearing to the right side of the con-
sonant form. In other cases, the vertical stem remains but some part of its right-side geom-
etry 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 fol-
lowing letter). Therefore, in the present context, the zero width joiner may be consid-
ered to present a context to which a preceding dead consonant may join so as to create the
half-form of the consonant.

For example, if Cₜ denotes the half-form glyph of consonant C, then a half-consonant form
is represented as shown in Figure 9-4.

**Figure 9-4. Half-Consonants in Devanagari**

\[ KAₜ + ZWJ + SSA \rightarrow KAₜ + SSA \]

\[ क + ळ + ष \rightarrow कष \]

In the absence of the zero width joiner, the sequence in Figure 9-4 would normally pro-
duce the full conjunct form कण (K.SSAₙ).

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 9-5.

**Figure 9-5. Independent Half-Forms in Devanagari**

\[ GAₜ + ZWJ \rightarrow GAₜ \]

\[ ग + ळ \rightarrow र \]

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 9-6.

**Figure 9-6. Half-Consonants in Oriya**

\[ KAₜ + ZWJ + VIRAMA + TA \rightarrow KAₜ + TA \]

\[ क + ळ + ṭ + ે \rightarrow ṭ \]

In the absence of the zero width joiner, the sequence in Figure 9-6 would normally pro-
duce the full conjunct form ṭ (K.TAₙ).

**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 con-
junct (see Figure 9-7).
As the rendering of conjuncts and half-forms depends on the availability of glyphs in the font, the following fallback strategy should be employed:

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

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

**Rendering Devanagari**

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

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

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

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

- $C_n$: Nominal glyph form of consonant $C$ as it appears in the code charts.
- $C_l$: A live consonant, depicted identically to $C_n$.
- $C_d$: Glyph depicting the dead consonant form of consonant $C$.
- $C_h$: Glyph depicting the half-consonant form of consonant $C$.
- $L_n$: Nominal glyph form of a conjunct ligature consisting of two or more component consonants. A conjunct ligature composed of two consonants $X$ and $Y$ is also denoted $X,Y_n$.  

![Figure 9-7. Consonant Forms in Devanagari and Oriya](image)
9.1 Devanagari

RA_{sup} A nonspacing combining mark glyph form of U+0930 devanagari letter ra positioned above or attached to the upper part of a base glyph form. This form is also known as repha.

RA_{sub} A nonspacing combining mark glyph form of U+0930 devanagari letter ra positioned below or attached to the lower part of a base glyph form.

V_v A glyph depicting the dependent vowel sign form of a vowel V.

VIRAMA 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, it is considered to be a dead consonant \( C_d \). A consonant \( C_n \) that does not precede VIRAMA is considered to be a live consonant \( C_l \).

\[ TA_n + \text{VIRAMA}_n \rightarrow TA_d \]

\[ \text{त} + \text{ँ} \rightarrow \text{त} \]

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 precedes a consonant, then it is replaced by the superscript nonspacing mark RA\text{sup}, which is positioned so that it applies to the logically subsequent element in the memory representation.

\[ \text{RA}_d + \text{KA}_l \rightarrow \text{KA}_l + \text{RA}_{sup} \]

\[ \text{र} + \text{क} \rightarrow \text{क} + \text{ँ} \rightarrow \text{क} \]

\[ \text{RA}_d^1 + \text{RA}_d^2 \rightarrow \text{RA}_d^2 + \text{RA}_d^{sup} \]

\[ \text{र} + \text{र} \rightarrow \text{र} + \text{ँ} \rightarrow \text{र} \]

R3 If the superscript mark RA\text{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.

\[ \text{RA}_d + \text{JA}_d + \text{NYA}_l \rightarrow \text{J.NYA}_n + \text{RA}_{sup} \]

\[ \text{र} + \text{ज} + \text{ज} \rightarrow \text{ज} + \text{ँ} \rightarrow \text{ज} \]
R4  If the superscript mark \textsuperscript{RA} 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.

\[ \text{RA}_d + \text{GA}_d + \text{GHA}_l \rightarrow \text{GA}_n + \text{GHA}_l + \text{RA}_{sup} \]

\[ \underline{र} + \underline{ग} + \underline{घ} \rightarrow \underline{र} + \underline{ग} + \underline{ः} \rightarrow \underline{रघ} \]

R5  In conformance with the ISCII standard, the half-consonant form \text{RRA}_h is represented as eyelash-RA. This form of RA is commonly used in writing Marathi and Newari.

\[ \text{RRA}_n + \text{VIRAMA}_n \rightarrow \text{RRA}_h \]

\[ \underline{र} + \underline{ऽ} \rightarrow \underline{रऽ} \]

R5a  For compatibility with The Unicode Standard, Version 2.0, if the dead consonant \text{RA}_d precedes zero width joiner, then the half-consonant form \text{RA}_h, depicted as eyelash-RA, is used instead of \text{RA}_{sup}.

\[ \text{RA}_d + \text{ZWJ} \rightarrow \text{RA}_h \]

\[ \underline{र} + \underline{ऽ} \rightarrow \underline{रऽ} \]

R6  Except for the dead consonant RA\textsubscript{d}, when a dead consonant C\textsubscript{d} precedes the live consonant RA\textsubscript{l}, then C\textsubscript{d} is replaced with its nominal form C\textsubscript{n}, and RA is replaced by the subscript nonspacing mark \textsubscript{RA}_{sub}, which is positioned so that it applies to C\textsubscript{n}.

\[ \text{TTHA}_d + \text{RA}_l \rightarrow \text{TTHA}_n + \text{RA}_{sub} \]

\[ \underline{ट} + \underline{र} \rightarrow \underline{ट} + \underline{ः} \rightarrow \underline{टः} \]

R7  For certain consonants, the mark RA\textsubscript{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.

\[ \text{PHA}_d + \text{RA}_l \rightarrow \text{PHA}_n + \text{RA}_{sub} \]

\[ \underline{फ} + \underline{र} \rightarrow \underline{फ} + \underline{ः} \rightarrow \underline{फः} \]
R8 If a dead consonant (other than RA<sub>d</sub>) precedes RA<sub>d</sub>, then the substitution of RA for RA<sub>sub</sub> is performed as described above; however, the VIRAMA that formed RA<sub>d</sub> remains so as to form a dead consonant conjunct form.

\[
\text{T}_d + \text{RA}_d \rightarrow \text{T}_n + \text{RA}_{sub} + \text{VIRAMA}_n \rightarrow \text{T.RA}_d
\]

A dead consonant conjunct form that contains an absorbed RA<sub>d</sub> may subsequently combine to form a multipart conjunct form.

\[
\text{T.RA}_d + \text{YA}_l \rightarrow \text{T.R.YA}_n
\]

\[
\text{t} + \text{r} \rightarrow \text{t} + \text{v} + \text{v} \rightarrow \text{v}
\]

**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.

\[
\text{KA}_n + \text{NUKTA}_n + \text{VIRAMA}_n \rightarrow \text{QA}_d
\]

\[
\text{k} + \text{v} + \text{v} \rightarrow \text{v}
\]

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.

\[
\text{KA}_n + \text{AA}_v + \text{CANDRABINDU}_n
\]

\[
\text{k} + \text{a} + \text{v} \rightarrow \text{v}
\]

**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.

\[
\text{JA}_d + \text{NYA}_l \rightarrow \text{J.NYA}_n
\]

\[
\text{TTA}_d + \text{TTHA}_l \rightarrow \text{TT.TTHA}_n
\]

\[
\text{z} + \text{z} \rightarrow \text{z}
\]

\[
\text{t} + \text{t} \rightarrow \text{t}
\]
R12 A conjunct ligature form can itself behave as a dead consonant and enter into further, more complex ligatures.

\[ SA_d + TA_d + RA_n \rightarrow SA_d + T.RA_n \rightarrow S.T.RA_n \]

\[ स + त + र \rightarrow स + च \rightarrow सच \]

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

\[ K.SSA_d + YA_l \rightarrow K.SS_n + YA_n \]

\[ क्ष + य \rightarrow क्ष्य \]

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

\[ KA_n + RA\textsubscript{sub} \rightarrow K.RA_n \quad PHA_n + RA\textsubscript{sub} \rightarrow PH.RA_n \]

\[ क \ + \ ृ \ \rightarrow \ क्र \quad \ फ \ + \ ृ \ \rightarrow \ फ्र \]

R14 In some cases, other combining marks will combine with a base consonant, either attaching at a nonstandard location or changing shape. In minimal rendering, there are only two cases: RA\textsubscript{l} with U\textsubscript{vs} or UU\textsubscript{vs}.

\[ RA_l + U\textsubscript{vs} \rightarrow RU_n \quad RA_l + UU\textsubscript{vs} \rightarrow RUU_n \]

\[ र + \ ृ \ \rightarrow \ रु \quad र + \ ृ \ \rightarrow \ रु \]

**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 9-8).

**Figure 9-8.** Rendering Order in Devanagari

<table>
<thead>
<tr>
<th>Character Order</th>
<th>Glyph Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>KA\textsubscript{n} + I\textsubscript{vs} \rightarrow I + KA\textsubscript{n}</td>
<td>क + \ ृ \ \rightarrow \ कि</td>
</tr>
</tbody>
</table>

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\textsubscript{n} denotes the nominal form of consonant C, and V\textsubscript{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.
9.1 Devanagari

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

\[ \text{TA}_d + \text{RA}_l + I_{vs} \rightarrow \text{T.RA}_n + I_{vs} \rightarrow I_{vs} + \text{T.RA}_d \]

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

\[ \text{TA}_d + \text{ZWNJ} + \text{RA}_l + I_{vs} \rightarrow \text{TA}_d + I_{vs} + \text{RA}_l \]

Sample Half-Forms. Table 9-2 shows examples of half-consonant forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. They may be encoded explicitly using zero width joiner as shown. In normal conjunct formation, they may be used spontaneously to depict a dead consonant in combination with subsequent consonant forms.

<table>
<thead>
<tr>
<th>Table 9-2. Sample Devanagari Half-Forms</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>क + क + ZWJ → क</td>
<td>न + क + ZWJ → ऋ</td>
</tr>
<tr>
<td>ख + क + ZWJ → ख</td>
<td>प + क + ZWJ → ॠ</td>
</tr>
<tr>
<td>ग + क + ZWJ → ग</td>
<td>फ + क + ZWJ → ऌ</td>
</tr>
<tr>
<td>घ + क + ZWJ → घ</td>
<td>ब + क + ZWJ → ॡ</td>
</tr>
<tr>
<td>च + क + ZWJ → ङ</td>
<td>भ + क + ZWJ → ॢ</td>
</tr>
<tr>
<td>छ + क + ZWJ → छ</td>
<td>म + क + ZWJ → ॣ</td>
</tr>
<tr>
<td>ज + क + ZWJ → ज</td>
<td>य + क + ZWJ → ५</td>
</tr>
<tr>
<td>झ + क + ZWJ → झ</td>
<td>ल + क + ZWJ → ६</td>
</tr>
<tr>
<td>ञ + क + ZWJ → ञ</td>
<td>त + क + ZWJ → ७</td>
</tr>
<tr>
<td>ण + क + ZWJ → ण</td>
<td>थ + क + ZWJ → ८</td>
</tr>
<tr>
<td>त + क + ZWJ → ट</td>
<td>ध + क + ZWJ → ९</td>
</tr>
<tr>
<td>थ + क + ZWJ → ठ</td>
<td>ध + क + ZWJ → ॥</td>
</tr>
<tr>
<td>ध + क + ZWJ → ध</td>
<td></td>
</tr>
</tbody>
</table>

Sample Ligatures. Table 9-3 shows examples of conjunct ligature forms that are commonly used with the Devanagari script. These forms are glyphs, not characters. Not every writing system that employs this script uses all of these forms; in particular, many of these forms
are used only in writing Sanskrit texts. Furthermore, individual fonts may provide fewer or more ligature forms than are depicted here.

Table 9-3. Sample Devanagari Ligatures

| क + ऋ + क → क | ट + ◌ + ठ →ठ |
| क + ऋ + त → तक | ठ + ◌ + ठ →ठ |
| क + ऋ + र → क्र | ड + ◌ + ढ →ढ |
| क + ऋ + ष → क्ष | ड + ◌ + ढ →ढ |
| ड + ◌ + क → क्र | ड + ◌ + ढ →ढ |
| ड + ◌ + ख → ख्र | त + ◌ + त → त |
| ड + ◌ + ग → ग्र | त + ◌ + त → त |
| ड + ◌ + घ → घ्र | न + ◌ + न → न |
| ज + ◌ + ज → ज | फ + ◌ + र → फ |
| ज + ◌ + ज → ज | श + ◌ + र → श |
| द + ◌ + घ → घ | ह + ◌ + म → ह |
| द + ◌ + द → द | ह + ◌ + म → ह |
| द + ◌ + ध → ध | ह + ◌ + द → द |
| द + ◌ + ध → ध | ह + ◌ + द → द |
| द + ◌ + ध → ध | ह + ◌ + द → द |
| द + ◌ + ध → ध | ह + ◌ + द → द |
| ट + ◌ + ठ →ठ | स + ◌ + स्त → र्ष |

**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 9-4. 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.

**Language-Specific Allographs.** In Marathi and some South Indian orthographies, variant glyphs are preferred for U+0932 devanagari letter la and U+0936 devanagari letter sha, as shown in Figure 9-9. Marathi also makes use of the “eyelash” form of the letter RA, as discussed in rule R5.

**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 repre-
9.1 Devanagari

The Unicode Standard, Version 6.1
Copyright © 1991–2012 Unicode, Inc.

sent by U+0901 devanagari sign candrabindu and U+0902 devanagari sign anusvara. These marks indicate nasalization or final nasal closure of a syllable. U+093C devanagari sign nukta is a true diacritic. It is used to extend the basic set of consonant letters by modifying them (with a subscript dot in Devanagari) to create new letters. U+0951..U+0954 are a set of combining marks used in transcription of Sanskrit texts.

Devanagari Digits, Punctuation, and Symbols

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

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

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

Other Symbols. U+0970 devanagari abbreviation sign appears after letters or combinations of letters and marks the sequence as an abbreviation. It is intended specifically for Devanagari script-based abbreviations, such as the Devanagari rupee sign. Other symbols and signs most commonly occurring in Vedic texts are encoded in the Devanagari Extended and Vedic Extensions blocks and are discussed in the text that follows.
The svasti (or well-being) signs often associated with the Hindu, Buddhist, and Jain traditions are encoded in the Tibetan block. See Section 10.2, 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 anudatta, but those combinations are no longer recommended.

**Konkani.** Konkani makes use of additional sounds that can be represented with combinations such as U+091A devanagari letter ca plus U+093C devanagari sign nukta and U+091F devanagari letter tta plus U+0949 devanagari vowel sign candra o.

**Bodo, Dogri, and Maithili.** The orthographies of the Bodo, Dogri, and Maithili languages of India make use of U+02BC "'" 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 9-10. The Maithili examples show the same sentence, first in full form, and then using U+02BC to show truncation of words.

**Figure 9-10. Use of Apostrophe in Bodo, Dogri and Maithili**

<table>
<thead>
<tr>
<th>Language</th>
<th>Examples</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bodo</td>
<td>खर' दख'ना</td>
<td>head type of Bodo dress</td>
</tr>
<tr>
<td>Dogri</td>
<td>ख'ल्ल ति'लकना</td>
<td>down to slip</td>
</tr>
<tr>
<td>Maithili</td>
<td>कतए पद्धए गेलह? कत' पडा' गेल'?</td>
<td>} Where did you go away?</td>
</tr>
</tbody>
</table>

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 9-11.

**Figure 9-11. Use of Avagraha in Dogri**

<table>
<thead>
<tr>
<th>Example</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>तला तलांस</td>
<td>sole pond</td>
</tr>
</tbody>
</table>
Kashmiri Letters. There are several letters for use with Kashmiri when written in Devanagari script. Long and short versions of the independent vowel letters are encoded in the range U+0973..U+0977. The corresponding dependent vowel signs are U+093A DEVANAGARI VOWEL SIGN OE, U+093B DEVANAGARI VOWEL SIGN OOE, and U+094F DEVANAGARI VOWEL SIGN AW. The forms of the independent vowels for Kashmiri are constructed by using the glyphs of the matras U+093B DEVANAGARI VOWEL SIGN OOE, U+094F DEVANAGARI VOWEL SIGN AW, U+0956 DEVANAGARI VOWEL SIGN UE, and U+0957 DEVANAGARI VOWEL SIGN UUE as diacritics on U+0905 DEVANAGARI LETTER A. However, for representation of independent vowels in Kashmiri, use the encoded, composite characters in the range U+0973..U+0977 and not the visually equivalent sequences of U+0905 DEVANAGARI LETTER A plus the matras. See Table 9-1. A few of the letters identified as being used for Kashmiri are also used to write the Bihari languages.

Letters for Bihari Languages. A number of the Devanagari vowel letters have been used to write the Bihari languages Bhojputi, Magadhi, and Maithili, as listed in Table 9-5.

Table 9-5. Devanagari Vowels Used in Bihari Languages

<table>
<thead>
<tr>
<th>Devanagari Letter</th>
<th>Unicode Code Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short E</td>
<td>U+090E, U+0912</td>
</tr>
<tr>
<td>Short O</td>
<td>U+0946, U+094A</td>
</tr>
<tr>
<td>OE</td>
<td>U+0973</td>
</tr>
<tr>
<td>OOE</td>
<td>U+0974</td>
</tr>
<tr>
<td>AW</td>
<td>U+0975</td>
</tr>
<tr>
<td>OE</td>
<td>U+0976</td>
</tr>
<tr>
<td>OOE</td>
<td>U+0977</td>
</tr>
<tr>
<td>AW</td>
<td>U+093A</td>
</tr>
<tr>
<td>AW</td>
<td>U+093B</td>
</tr>
<tr>
<td>AW</td>
<td>U+094F</td>
</tr>
</tbody>
</table>

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 9-6 shows those combinations applied to ka. In the underlying representation of text, U+094E should be first in the sequence of dependent vowel signs after the consonant, and may be followed by U+0947, U+093E or U+094B.

Table 9-6. Prishthamatra Orthography

<table>
<thead>
<tr>
<th>Prishthamatra Orthography</th>
<th>Modern Orthography</th>
</tr>
</thead>
<tbody>
<tr>
<td>ke</td>
<td>के &lt;0915, 0947&gt;</td>
</tr>
<tr>
<td>kai</td>
<td>कै &lt;0915, 0948&gt;</td>
</tr>
<tr>
<td>ko</td>
<td>कौ &lt;0915, 094B&gt;</td>
</tr>
<tr>
<td>kau</td>
<td>कौ &lt;0915, 094C&gt;</td>
</tr>
</tbody>
</table>

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 marks of nasalization (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”), svarīta (“modulated” or dropping from high to low tone) or ekāruti (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 Rāṣṭreyāṇya 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-Sañhitā) 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 Rāṣṭreyāṇya 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.

Marks of Nasalization. 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 dan- das. U+A8FA devanagari caret, a zero-width spacing character, marks the insertion point of omitted text, and is placed at the insertion point between two orthographic syllables. It can also be used to indicate word division.

Vedic Extensions: U+1CD0-U+1CFF

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

\textbf{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 (S\=a\=naveda, Yajurveda, Rigveda, and Atharvaveda) and in the prose text on Vedic ritual (\textit{\=S\=atapathabr\=\=h\=\=ma}na). The character U+1CD8 \textit{vedic tone candra} below is also used to identify the short vowels \textit{e} and \textit{o}. In this usage, the prescribed order is the Indic syllable (\textit{aksara}), followed by U+1CD8 \textit{vedic tone candra} below and the tone mark (\textit{svara}). When a tone mark is placed below, it appears below the \textit{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 \textit{combining asterisk above} would be used to represent a mark of that shape above a Vedic letter.

\textbf{Diacritics for the Visarga.} A set of combining marks that serve as diacritics for the \textit{visarga} is encoded in the range U+1CE2..U+1CE8. These marks indicate that the \textit{visarga} has a particular tone. For example, the combination U+0903 \textit{devanagari sign visarga} plus U+1CE2 \textit{vedic sign visarga svarita} represents a \textit{svarita visarga}. The upward-shaped diacritic is used for the \textit{ud\=\=atta} (high-toned), the downward-shaped diacritic for \textit{anud\=\=atta} (low-toned), and the midline glyph indicates the \textit{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 \textit{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.

\textbf{Nasalization Marks.} A set of spacing marks and one combining mark, U+1CED \textit{vedic sign tiryak}, are encoded in the range U+1CE9..U+1CF1. They describe phonetic distinctions in the articulation of nasals. The \textit{gomukha} characters from U+1CE9..U+1CEC may be combined with U+0902 \textit{devanagari sign anusvara} or U+0901 \textit{devanagari sign candrabindu}. U+1CF1 \textit{vedic sign anusvara ubhayato mukha} may indicate a \textit{visarga} with a tonal mark as well as a nasal. The three characters, U+1CEE \textit{vedic sign hexiform long anusvara}, U+1CEF \textit{vedic sign long anusvara}, and U+1CF0 \textit{vedic sign rthang long anusvara}, are all synonymous and indicate a long \textit{anusv\=\=ra} after a short vowel. U+1CED \textit{vedic sign tiryak} is the only combining character in this set of nasalization marks. While it appears similar to the U+094D \textit{devanagari sign virama}, it is used to render glyph variants of nasal marks that occur in manuscripts and printed texts.

\textbf{Ardhavisarga.} U+1CF2 \textit{vedic sign ardhavisarga} is a character that marks either the \textit{jihv\=\=amal\=\=ya}, a velar fricative occurring only before the unvoiced velar stops \textit{ka} and \textit{kha}, or the \textit{upadhm\=\=\=nia}, a bilabial fricative occurring only before the unvoiced labial stops \textit{pa} and \textit{pha}. \textit{Ardhavisarga} is a spacing character. It is represented in text in visual order before the consonant it modifies.

9.2 Bengali (Bangla)

Bengali: U+0980–U+09FF

Scripts encoded in the Unicode Standard often are used to write many languages. The script termed \textit{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 are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. Table 9-7 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 9-7. Bengali Vowel Letters](image)

**Two-Part Vowel Signs**. The Bengali script, along with a number of other Indic scripts, makes use of two-part vowel signs. In these vowels 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 compatibility with existing implementations of the scripts that use two-part vowel signs, the Unicode Standard explicitly encodes the right half of these vowel signs. For example, U+09D7 Bengali au length mark represents the right-half glyph component of U+09CC Bengali vowel sign au. In Bengali orthography, the au length mark is always used in conjunction with the left part and does not have a meaning on its own.

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

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

**Characters for Assamese**. U+09F0 Bengali letter ra with middle diagonal and U+09F1 Bengali letter ra with lower diagonal are characters in this block required to write Assamese.

**Rendering Behavior**. Like other Brahmic scripts in the Unicode Standard, Bengali uses the hasant to form conjunct characters. For example, U+09B8 Bengalí letter sa + U+09CD Bengalí sign virama + U+0995 Bengalí letter ka yields the conjunct % SKA. For general principles regarding the rendering of the Bengali script, see the rules for rendering in Section 9.1, 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 9-8.
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 16.2, 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 9-12.

**Figure 9-12. Requesting Bengali Consonant-Vowel Ligature**

\[
\begin{align*}
\text{gu} & \quad \text{<0997, 09C1>} & \text{গু} & \quad \text{গু} \\
\text{ru} & \quad \text{<09B0, 09C1>} & \text{রু} & \quad \text{রু} \\
\text{r¯u} & \quad \text{<09B0, 09C2>} & \text{রু} & \quad \text{রু} \\
\text{šu} & \quad \text{<09B6, 09C1>} & \text{শু} & \quad \text{শু} \\
\text{hu} & \quad \text{<09B9, 09C1>} & \text{হু} & \quad \text{হু} \\
\text{hr} & \quad \text{<09B9, 09C3>} & \text{হৃ} & \quad \text{হৃ}
\end{align*}
\]

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 9-13.

**Figure 9-13. Blocking Bengali Consonant-Vowel Ligature**

\[
\begin{align*}
\text{gu} & \quad \text{<0997, 09C1>} & \text{গু} & \quad \text{গু} \\
\text{ru} & \quad \text{<09B0, 09C1>} & \text{রু} & \quad \text{রু} \\
\text{r¯u} & \quad \text{<09B0, 09C2>} & \text{রু} & \quad \text{রু} \\
\text{šu} & \quad \text{<09B6, 09C1>} & \text{শু} & \quad \text{শু} \\
\text{hu} & \quad \text{<09B9, 09C1>} & \text{হু} & \quad \text{হু} \\
\text{hr} & \quad \text{<09B9, 09C3>} & \text{হৃ} & \quad \text{হৃ}
\end{align*}
\]

**Khiya.** The letter ক্ষ, known as khiya, is often considered as a distinct letter of the Bengali alphabet. However, it is not encoded separately. It is represented by the sequence <U+0995 ক্ষ Bengali letter ka, U+09CD ঢ Bengali sign virama, U+09B7 ঢ 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 \(<U+09A4\) ta, \(U+09CD\) hasant, \(U+09A4\) ta> 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 9-14.

![Figure 9-14. Bengali Syllable tta](image)

It is also possible for the sequence \(<ta, hasant, ta>\) to be displayed with a full ta glyph combined with a hasant glyph, followed by another full ta glyph \(\text{తు}\). 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 9-14.

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 9-14.

Ya-phalaa. Ya-phalaa is a presentation form of U+09AF য Bengali letter ya. Represented by the sequence \(<U+09CD\) hasant virama, \(U+09AF\) য Bengali letter ya>, ya-phalaa has a special form ͏. When combined with \(U+09BE\) অ Bengali vowel sign aa, it is used for transcribing [æ] as in the “a” in the English word “bat.” Ya-phalaa can be applied to initial vowels as well:

\[
\text{অ্‌} = <0985, 09CD, 09AF, 09BE> \quad (a- \text{ hasant ya -aa})
\]

\[
\text{এ্঍‌} = <098F, 09CD, 09AF, 09BE> \quad (e- \text{ hasant ya -aa})
\]

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

\[
\text{অ্‌} = <0985, 09CD, 09AF, 09BE, 0981> \quad (a- \text{ hasant ya -aa candrabindu})
\]
Further examples:

- \( \text{অ + ও + য + ও} \rightarrow \text{ওয়া} \)
- \( \text{এ + ও + য + ও} \rightarrow \text{ওয়া} \)
- \( \text{ত + ও + য + ও} \rightarrow \text{ওয়া} \)

**Interaction of Repha and Ya-phalaa.** The formation of the *repha* form is defined in *Section 9.1, 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:

\[
\text{র + ও + ম} \rightarrow \text{ম} \quad \text{as in \ カーマ (karma)}
\]

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:

\[
\text{ক + ও + য} \rightarrow \text{ক্য} \quad \text{as in \ バク (bakyô)}
\]

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

\[
\text{র + ও + য} \rightarrow \text{য} \quad \text{or} \quad \text{র}
\]

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:

\[
\text{র + ও + য} \rightarrow \text{য}
\]

\[
\text{র + [zwj] + ও + য} \rightarrow \text{র}
\]

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

**Punctuation.** Danda and double danda marks as well as some other unified punctuation used with Bengali are found in the Devanagari block; see *Section 9.1, Devanagari*.

**Truncation.** The orthography of the Bangla language makes use of U+02BC “’” **MODIFIER LETTER APOSTROPE** to indicate the truncation of words. This sign is called *urdha-comma*. Examples illustrating the use of U+02BC “’” **MODIFIER LETTER APOSTROPE** are shown in *Table 9-9*.

**Table 9-9.** Use of Apostrophe in Bangla

<table>
<thead>
<tr>
<th>Example</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>কুরে</td>
<td>after, on doing (something)</td>
</tr>
<tr>
<td>‘পরে</td>
<td>above</td>
</tr>
<tr>
<td>ঔপরে</td>
<td></td>
</tr>
</tbody>
</table>
9.3 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 पंजा *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 *gurmukhi letter ya*. However, because its usage is relatively rare and not entirely predictable, it is encoded as a separate character. This character should occur after the consonant to which it attaches and before any vowel sign.

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

Punjabi does not have complex combinations of consonant sounds. Furthermore, the orthography is not strictly phonetic, and sometimes the inherent /a/ sound is not pronounced. For example, the word ਗੁਰਮੁਖੀ *gurmukhi* is represented with the sequence <0A17, 0A41, 0A30, 0A2E, 0A41, 0A16, 0A40>, which could be transliterated as *gurmukhi*; 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 ੰ attached to ੀ, as in ਗੁਬਿਦਾ and both the vowel signs ੰ and ੫ attached to a consonant, as in ਗੁਬਿਦਾ; this is used to indicate the metrical shortening of /o/ or the lengthening of /a/ depending on the context. Other combinations are attested as well, such as ਗ਼ਿਤਾ *ghiana*.
Because of the combining classes of the characters U+0A4B 
Punjabi vowel sign oo and U+0A41 
Punjabi 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 9-10 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 9-10. Gurmukhi Vowel Letters

<table>
<thead>
<tr>
<th>For</th>
<th>Use</th>
<th>Do Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>ਅ</td>
<td>0A06</td>
<td>&lt;0A05, 0A3E&gt;</td>
</tr>
<tr>
<td>ਖ</td>
<td>0A07</td>
<td>&lt;0A72, 0A3F&gt;</td>
</tr>
<tr>
<td>ਗ</td>
<td>0A08</td>
<td>&lt;0A72, 0A40&gt;</td>
</tr>
<tr>
<td>ਘ</td>
<td>0A09</td>
<td>&lt;0A73, 0A41&gt;</td>
</tr>
<tr>
<td>ਜ</td>
<td>0A0A</td>
<td>&lt;0A73, 0A42&gt;</td>
</tr>
<tr>
<td>ਝ</td>
<td>0A0F</td>
<td>&lt;0A72, 0A47&gt;</td>
</tr>
<tr>
<td>ਞ</td>
<td>0A10</td>
<td>&lt;0A05, 0A48&gt;</td>
</tr>
<tr>
<td>ਟ</td>
<td>0A13</td>
<td>&lt;0A73, 0A4B&gt;</td>
</tr>
<tr>
<td>ਠ</td>
<td>0A14</td>
<td>&lt;0A05, 0A4C&gt;</td>
</tr>
</tbody>
</table>

**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 
Punjabi ura and U+0A72 
Punjabi 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 
Punjabi 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 sa, 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 9.1, Devanagari. In many aspects, Gurmukhi is simpler than Devanagari. In modern Punjabi, there are no half-consonants, no half-forms, no repa (upper form of U+0930 
Devanagari letter ra), and no real ligatures. Rules R2–
R5, R11, and R14 do not apply. Conversely, the behavior for subscript RA (rules R6–R8 and R13) applies to U+0A39 GURMUKHI LETTER HA and U+0A35 GURMUKHI LETTER VA, which also have subjoined forms, called pairin in Punjabi. The subjoined form for RA is like a knot, while the subjoined HA and VA are written the same as the base form, without the top bar, but are reduced in size. As described in rule R13, they attach at the bottom of the base consonant, and will “push” down any attached vowel sign for U or UU. When U+0A2F GURMUKHI LETTER YA follows a dead consonant, it assumes a different form called addha in Punjabi, without the leftmost part, and the dead consonant returns to the nominal form, as shown in Table 9-11.

Table 9-11. Gurmukhi Conjuncts

<table>
<thead>
<tr>
<th>Pair</th>
<th>Representation</th>
<th>Hindi Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>ह + र + ा</td>
<td>हा (mha)</td>
<td>pairin ha</td>
</tr>
<tr>
<td>प + र + ा</td>
<td>पा (pra)</td>
<td>pairin ra</td>
</tr>
<tr>
<td>च + र + ा</td>
<td>चा (dva)</td>
<td>pairin va</td>
</tr>
<tr>
<td>ठ + र + ा</td>
<td>ठा (dy)</td>
<td>addha ya</td>
</tr>
</tbody>
</table>

Other letters behaved similarly in old inscriptions, as shown in Table 9-12.

Table 9-12. Additional Pairin and Addha Forms in Gurmukhi

<table>
<thead>
<tr>
<th>Pair</th>
<th>Representation</th>
<th>Hindi Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>म + र + ा</td>
<td>मा (sga)</td>
<td>pairin ga</td>
</tr>
<tr>
<td>म + र + ा</td>
<td>मा (sca)</td>
<td>pairin ca</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (stta)</td>
<td>pairin tta</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sttha)</td>
<td>pairin ttha</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sta)</td>
<td>pairin ta</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sda)</td>
<td>pairin da</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
<tr>
<td>म + र + ठ</td>
<td>मा (sna)</td>
<td>pairin na</td>
</tr>
</tbody>
</table>

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 9-13. The reduced form of an initial U+0A30 GURMUKHI LETTER RA is similar to the Devanagari superscript RA (repaha), but this usage is rare, even in older texts.

A rendering engine for Gurmukhi should make accommodations for the correct positioning of the combining marks (see Section 5.13, Rendering Nonspaceing 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 9.1, *Devanagari*, for more information. Punjabi also uses Latin punctuation.

### 9.4 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 9-14 shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

#### Table 9-14. Gujarati Vowel Letters

<table>
<thead>
<tr>
<th>For</th>
<th>Use</th>
<th>Do Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>अ</td>
<td>0A86 &lt;0A85, 0ABE&gt;</td>
<td></td>
</tr>
<tr>
<td>ओ</td>
<td>0A8D &lt;0A85, 0AC5&gt;</td>
<td></td>
</tr>
<tr>
<td>े</td>
<td>0A8F &lt;0A85, 0AC7&gt;</td>
<td></td>
</tr>
<tr>
<td>ऋ</td>
<td>0A90 &lt;0A85, 0AC8&gt;</td>
<td></td>
</tr>
<tr>
<td>ो</td>
<td>0A91 &lt;0A85, 0AC9&gt;</td>
<td></td>
</tr>
<tr>
<td>ो</td>
<td>0A93 &lt;0A85, 0ACB&gt; or &lt;0A85, 0ABE, 0AC5&gt;</td>
<td></td>
</tr>
<tr>
<td>ौ</td>
<td>0A94 &lt;0A85, 0ACC&gt; or &lt;0A85, 0ABE, 0AC8&gt;</td>
<td></td>
</tr>
<tr>
<td>ु</td>
<td>0AC9 &lt;0AC5, 0ABE&gt;</td>
<td></td>
</tr>
</tbody>
</table>
**Rendering Behavior.** For rendering of the Gujarati script, see the rules for rendering in Section 9.1, Devanagari. Like other Brahmic scripts in the Unicode Standard, Gujarati uses the virama to form conjunct characters. The virama is informally called *khoḍo*, which means “lame” in Gujarati. Many conjunct characters, as in Devanagari, lose the vertical stroke; there are also vertical conjuncts. U+0AB0 GUJARATI LETTER RA takes special forms when it combines with other consonants, as shown in Table 9-15.

<table>
<thead>
<tr>
<th>Table 9-15. Gujarati Conjuncts</th>
</tr>
</thead>
<tbody>
<tr>
<td>क + क + ल → ल (kṣa)</td>
</tr>
<tr>
<td>Ḍ + क + ल → ल (jña)</td>
</tr>
<tr>
<td>त + क + ल → ल (tya)</td>
</tr>
<tr>
<td>ट + क + ल → ल (ṭa)</td>
</tr>
<tr>
<td>ण + क + ल → ल (rka)</td>
</tr>
<tr>
<td>९ + क + ल → ल (kra)</td>
</tr>
</tbody>
</table>

**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 9.1, Devanagari.

---

**9.5 Oriya**

**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 actual shapes of the letters, particularly for vowel signs, show similarities to Tamil. The Oriya script is used to write the Oriya language of the Orissa state in India as well as minority languages such as Khondi and Santali.

**Special Characters.** U+0B57 ORIYA AU LENGTH MARK is provided as an encoding for the right side of the surroundant 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 9-16 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.

<table>
<thead>
<tr>
<th>Table 9-16. Oriya Vowel Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>For</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>ऐ</td>
</tr>
<tr>
<td>ओ</td>
</tr>
<tr>
<td>औ</td>
</tr>
</tbody>
</table>

**Rendering Behavior.** For rendering of the Oriya script, see the rules for rendering in Section 9.1, Devanagari. Like other Brahmic scripts in the Unicode Standard, Oriya uses the
Oriya suppress the inherent vowel. Oriya has a visible virama, often being a lengthening of a part of the base consonant:

\[ \text{Virama to suppress inherent vowel} \]

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Virama} \rightarrow \text{视听节目} (k)\)}
\end{align*} \]

The virama is also used to form conjunct consonants, as shown in Table 9-17.

**Table 9-17. Oriya Conjuncts**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (k\text{sa}))</td>
<td>Conjunct consonants</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (k\text{ta}))</td>
<td>Conjunct consonants</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (t\text{ka}))</td>
<td>Conjunct consonants</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (t\text{ya}))</td>
<td>Conjunct consonants</td>
</tr>
</tbody>
</table>

**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:

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (r\text{pa})\)}
\end{align*} \]

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (p\text{ra})\)}
\end{align*} \]

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

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Virama} + \text{Consonant} + \text{Consonant} \rightarrow \text{视听节目} (a\text{nika})\)}
\end{align*} \]

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Virama} + \text{Consonant} \rightarrow \text{视听节目} (a\text{nica})\)}
\end{align*} \]

**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. See Table 9-18.

**Table 9-18. Oriya Vowel Placement**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{a}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{i}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{i}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{u}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{u}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{r}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{e}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{i}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{o}))</td>
<td>Dependent vowels</td>
</tr>
<tr>
<td>(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{au}))</td>
<td>Dependent vowels</td>
</tr>
</tbody>
</table>

U+0B01 oriya sign candrabindu is used for nasal vowels:

\[ \begin{align*}
\text{\(\text{视听节目} + \text{Vowel} \rightarrow \text{视听节目} (k\text{a}m)\)}
\end{align*} \]
Oriya VA and WA. These two letters are extensions to the basic Oriya alphabet. Because Sanskrit वन vana becomes Oriya বন bana in orthography and pronunciation, an extended letter U+0B35 Oriya letter va was devised by dotting U+0B2C Oriya letter ba for use in academic and technical text. For example, basic Oriya script cannot distinguish Sanskrit व्य vyana from व्य baba or व्व vava, but this distinction can be made with the modified version of ba. In some older sources, the glyph र is sometimes found for va; in others, र and ल have been shown, which in a more modern type style would be त। The letter va is not in common use today.

In a consonant conjunct, subjoined U+0B2C Oriya letter ba is usually—but not always—pronounced [wa]:

\[
\begin{align*}
U+0B15 \text{ ka } + U+0B4D \text{ virama } + U+0B2C \text{ ba } & \rightarrow \text{ [kwa]} \\
U+0B2E \text{ ma } + U+0B4D \text{ virama } + U+0B2C \text{ ba } & \rightarrow \text{ [mba]}
\end{align*}
\]

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

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 9.1, Devanagari. The mark U+0B70 Oriya isshar is placed before names of persons who are deceased.

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.

9.6 Tamil

Tamil: U+0B80–U+0BFF

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

The Tamil script has fewer consonants than the other Indic scripts. When representing the “missing” consonants in transcriptions of languages such as Sanskrit or Saurashtra, superscript European digits are often used, so \( \text{L}^2 = \text{pha} \), \( \text{L}^3 = \text{ba} \), and \( \text{L}^4 = \text{bha} \). The characters U+00B2, U+00B3, and U+2074 can be used to preserve this distinction in plain text. The Tamil script also avoids the use conjunct consonant forms, although a few conventional conjuncts are used.

Virama (Pulii). 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, \texttt{zero width non-joiner} is used to display a visible virama, as shown previously in the Devangari example in Figure 9-3.

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

The conjuncts \textit{kssa} and \textit{shrii} are traditionally displayed by conjunct ligatures, as illustrated for \textit{kssa} in Figure 9-15, but nowadays tend to be displayed using an explicit \textit{pu\kki} as well.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{Figure9-15.png}
\caption{Kssa Ligature in Tamil}
\end{figure}

\begin{equation}
\text{\texttt{\#} + \cdot + U+0B8B \rightarrow \text{\#}} \text{\texttt{\#}} \text{\texttt{k\texttt{a}}}
\end{equation}

To explicitly display a \textit{pu\kki} for such sequences, \texttt{zero width non-joiner} can be inserted after the \textit{pu\kki} in the sequence of characters.

\textit{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.

\begin{quote}
\textit{In a font that is capable of rendering Tamil, the number of glyphs is greater than the number of Tamil characters.}
\end{quote}

\textbf{Tamil Vowels}

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

\begin{equation}
\text{\texttt{\#} + \circ \neq \text{\texttt{\#}} + \cdot + \circ}
\end{equation}

\textit{Left-Side Vowels.} The Tamil vowels \texttt{U+0BC6 01}, \texttt{U+0BC7 02}, and \texttt{U+0BC8 03} are reordered in front of the consonant to which they are applied. When occurring in a syllable, these vowels are rendered to the left side of their consonant, as shown in Table 9-19.

\begin{table}[h]
\centering
\caption{Tamil Vowel Reordering}
\begin{tabular}{|c|c|}
\hline
\textbf{Memory Representation} & \textbf{Display} \\
\hline
\texttt{\#} & \texttt{\#} \texttt{01} \\
\hline
\texttt{\#} & \texttt{\#} \texttt{02} \\
\hline
\texttt{\#} & \texttt{\#} \texttt{03} \\
\hline
\end{tabular}
\end{table}

\textit{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 9-16.
Figure 9-16. Tamil Two-Part Vowels

\[
\begin{align*}
\text{\text{ Tamil letter lla}} & \equiv 0BCA + 0BC6 + 0BBE \\
\text{\text{ Tamil au length mark}} & \equiv 0BCB + 0BC7 + 0BBE \\
\text{\text{ Tamil au length mark}} & \equiv 0BCC + 0BD7
\end{align*}
\]

In these examples, the representation on the left, which is a single code point, is the preferred form and the form in common use for Tamil. Note that the \(\text{\text{ Tamil au length mark}}\) in the third example is not \text{U+0BB3} TAMIL LETTER LLA; it is \text{U+0BD7} TAMIL AU LENGTH MARK.

In the process of rendering, these two-part vowels are transformed into the two separate glyphs equivalent to those on the right, which are then subject to vowel reordering, as shown in Table 9-20.

Table 9-20. Tamil Vowel Splitting and Reordering

<table>
<thead>
<tr>
<th>Memory Representation</th>
<th>Display</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{\text{ Tamil letter lla}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
<tr>
<td>(\text{\text{ Tamil au length mark}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
<tr>
<td>(\text{\text{ Tamil au length mark}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
<tr>
<td>(\text{\text{ Tamil au length mark}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
<tr>
<td>(\text{\text{ Tamil au length mark}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
<tr>
<td>(\text{\text{ Tamil au length mark}})</td>
<td>(\text{\text{ Tamil au length mark}})</td>
</tr>
</tbody>
</table>

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 9-17.

Figure 9-17. Vowel Reordering Around a Tamil Conjunct

\[
\text{\text{ Tamil letter lla}} + \text{\text{ Tamil au length mark}} + \text{\text{ Tamil letter lla}} + \text{\text{ Tamil au length mark}} \rightarrow \text{\text{ Tamil au length mark}} _{k50}
\]

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

Tamil Ligatures

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

Ligatures with Vowel \text{i}. The vowel signs \text{i} \(\text{\text{ Tamil letter lla}}\) and \text{\text{ Tamil au length mark}} \text{\text{ Tamil au length mark}} form ligatures with the consonant \text{\text{ Tamil letter lla}} as shown in examples 1 and 2 of Figure 9-18. 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 9-18.
Figure 9-18. Tamil Ligatures with i

1. \( \text{இ} + \text{ீ} \rightarrow \text{தி} \)
2. \( \text{ே} + \text{ீ} \rightarrow \text{ேதி} \)
3. \( \text{ே} + \text{ீ} \rightarrow \text{ேதி} \)
4. \( \text{ே} + \text{ீ} \rightarrow \text{ேதி} \)

Ligatures with Vowel u. The vowel signs \( \text{ீ} \) and \( \text{ீீ} \) normally ligate with their consonant, as shown in Table 9-21. In the first column, the basic consonant is shown; the second column illustrates the ligation of that consonant with the \( \text{ீ} \) vowel sign; and the third column illustrates the ligation with the \( \text{ீீ} \) vowel sign.

Table 9-21. Tamil Ligatures with u

<table>
<thead>
<tr>
<th>( x )</th>
<th>( x + \text{ீ} )</th>
<th>( x + \text{ீீ} )</th>
<th>( x + \text{ீ} )</th>
<th>( x + \text{ீீ} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{ெ} )</td>
<td>( \text{ெீ} )</td>
<td>( \text{ெீீ} )</td>
<td>( \text{ெீ} )</td>
<td>( \text{ெீீ} )</td>
</tr>
<tr>
<td>( \text{ே} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
</tr>
<tr>
<td>( \text{ை} )</td>
<td>( \text{ைீ} )</td>
<td>( \text{ைீீ} )</td>
<td>( \text{ைீ} )</td>
<td>( \text{ைீீ} )</td>
</tr>
<tr>
<td>( \text{ே} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
</tr>
<tr>
<td>( \text{ை} )</td>
<td>( \text{ைீ} )</td>
<td>( \text{ைீீ} )</td>
<td>( \text{ைீ} )</td>
<td>( \text{ைீீ} )</td>
</tr>
<tr>
<td>( \text{ே} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
<td>( \text{ேீ} )</td>
<td>( \text{ேீீ} )</td>
</tr>
</tbody>
</table>

With certain consonants, \( \text{ெ}, \text{ே}, \text{ை}, \text{ே}, \text{ை}, \text{ே} \), and the conjunct \( \text{ை} \), the vowel signs \( \text{ீ} \) and \( \text{ீீ} \) take a distinct spacing form, as shown in Figure 9-19.

Figure 9-19. Spacing Forms of Tamil u

1. \( \text{ீ} + \text{ீ} \rightarrow \text{ீீ} \)
2. \( \text{ீ} + \text{ீ} \rightarrow \text{ீீ} \)

Ligatures with ra. Based on typographical preferences, the consonant \( \text{ை} \) may change shape to \( \text{ை} \), when it ligates. Such change, if it occurs, will happen only when the \( \text{ை} \) form of U+0BB0 \( \text{Tamil Letter RA} \) would not be confused with the nominal form \( \text{ை} \) of U+0BBD \( \text{Tamil Vowel Sign AA} \) (namely, when \( \text{ை} \) is combined with \( \text{ீ}, \text{ீீ}, \) or \( \text{ீ} \)). This change in shape is illustrated in Figure 9-20.
Figure 9-20. Tamil Ligatures with ra

\[
\begin{align*}
\text{rn} + \ddot{o} & \rightarrow \dddot{r} \\
\text{rn} + \dddot{o} & \rightarrow \dddot{ri} \\
\text{rn} + \ddot{e} & \rightarrow \dddot{ri}
\end{align*}
\]

However, various governmental bodies mandate that the basic shape of the consonant \( ra \) 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 \( ra \). Sri Lanka, on the other hand, specifies the use of the changed forms shown in Figure 9-20.

Ligatures with aa in Traditional Tamil Orthography. In traditional Tamil orthography, the vowel sign \( aa \) optionally ligates with \( \dddot{s} \), \( \ddot{s} \), or \( \dddot{m} \), as illustrated in Figure 9-21.

Figure 9-21. Traditional Tamil Ligatures with \( aa \)

\[
\begin{align*}
\text{ss} + \ddot{a} & \rightarrow \dddot{s}a \\
\text{st} + \ddot{a} & \rightarrow \dddot{st}a \\
\text{m} + \ddot{a} & \rightarrow \dddot{ma}
\end{align*}
\]

These ligations also affect the right-hand part of two-part vowels, as shown in Figure 9-22.

Figure 9-22. Traditional Tamil Ligatures with \( o \)

\[
\begin{align*}
\text{ss} + \ddot{o} & \rightarrow \dddot{s}o \\
\text{st} + \ddot{o} & \rightarrow \dddot{st}o \\
\text{st} + \ddot{a} & \rightarrow \dddot{st}o \\
\text{m} + \ddot{o} & \rightarrow \dddot{mo} \\
\text{m} + \ddot{a} & \rightarrow \dddot{mo}
\end{align*}
\]

Ligatures with ai in Traditional Tamil Orthography. In traditional Tamil orthography, the left-side vowel sign \( ai \) is also subject to a change in form. It is rendered as \( \ddot{a} \) when it occurs on the left side of \( \dddot{s} \), \( \ddot{s} \), \( \ddot{u} \), or \( \ddot{t} \), as illustrated in Figure 9-23.
Figure 9-23. Traditional Tamil Ligatures with ai

\[
\begin{align*}
& \text{தை + ஓ} \rightarrow \text{நை} \\
& \text{தை + ஔ} \rightarrow \text{நய} \\
& \text{ஞ + ஔ} \rightarrow \text{லய} \\
& \text{ஞ + ஓ} \rightarrow \text{லந} \\
\end{align*}
\]

By contrast, in modern Tamil orthography, this vowel does not change its shape, as shown in Figure 9-24.

Figure 9-24. Vowel ai in Modern Tamil

\[
\begin{align*}
& \text{தை + ஔ} \rightarrow \text{நை} \\
\end{align*}
\]

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 9.1, Devanagari.

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 9-22 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 9-22, 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 9-22. Tamil Vowels, Consonants, and Syllables

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>AA</th>
<th>I</th>
<th>II</th>
<th>U</th>
<th>UU</th>
<th>E</th>
<th>EE</th>
<th>AI</th>
<th>O</th>
<th>OO</th>
<th>AU</th>
</tr>
</thead>
<tbody>
<tr>
<td>0B83</td>
<td>0B95</td>
<td>0B96</td>
<td>0B97</td>
<td>0B84</td>
<td>0B98</td>
<td>0B9A</td>
<td>0B9C</td>
<td>0B99</td>
<td>0B9B</td>
<td>0B9D</td>
<td>0B9E</td>
<td></td>
</tr>
<tr>
<td>0B95</td>
<td>0B96</td>
<td>0B97</td>
<td>0B84</td>
<td>0B98</td>
<td>0B99</td>
<td>0B9A</td>
<td>0B9C</td>
<td>0B9B</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td></td>
</tr>
<tr>
<td>0B9A</td>
<td>0B9B</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td></td>
</tr>
<tr>
<td>0B9E</td>
<td>0B84</td>
<td>0B9A</td>
<td>0B9B</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td></td>
</tr>
<tr>
<td>0B8A</td>
<td>0B8B</td>
<td>0B8C</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td></td>
</tr>
<tr>
<td>0B8D</td>
<td>0B8F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td></td>
</tr>
<tr>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td></td>
</tr>
<tr>
<td>0B98</td>
<td>0B99</td>
<td>0B9A</td>
<td>0B9B</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td></td>
</tr>
<tr>
<td>0B9B</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td>0B9F</td>
<td>0B9C</td>
<td>0B9D</td>
<td>0B9E</td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 1991–2012 Unicode, Inc. The Unicode Standard, Version 6.1
9.7 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, 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 9-23 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>
<thead>
<tr>
<th>For</th>
<th>Use</th>
<th>Do Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>ऐ</td>
<td>0C13</td>
<td>&lt;0C12, 0C55&gt;</td>
</tr>
<tr>
<td>ऑ</td>
<td>0C14</td>
<td>&lt;0C12, 0C4C&gt;</td>
</tr>
<tr>
<td>ओ</td>
<td>0C40</td>
<td>&lt;0C3F, 0C55&gt;</td>
</tr>
<tr>
<td>ठ</td>
<td>0C47</td>
<td>&lt;0C46, 0C55&gt;</td>
</tr>
<tr>
<td>ड</td>
<td>0C48</td>
<td>&lt;0C4A, 0C55&gt;</td>
</tr>
</tbody>
</table>

**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
South Asian Scripts-I

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:

\[
U+0C15  ka + U+0C4D  virama + U+200C  \text{ZWJ}  \text{zero width non-joiner} \rightarrow \mathbf{k}
\]

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

\[
\begin{align*}
U+0C17  ga & + U+0C4D  virama & + U+0C17  ga & \rightarrow \mathbf{g}\mathbf{g} \\
U+0C15  ka & + U+0C4D  virama & + U+0C15  ka & \rightarrow \mathbf{k}\mathbf{k} \\
U+0C15  ka & + U+0C4D  virama & + U+0C2F  ya & \rightarrow \mathbf{y}\mathbf{a} \\
U+0C15  ka & + U+0C4D  virama & + U+0C37  ssa & \rightarrow \mathbf{s}\mathbf{s}\mathbf{a}
\end{align*}
\]

Nakārā-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 \(\mathbf{n}\), which takes its shape from the glyphs in the sequence <U+0C28 telugu letter na, U+0C4D telugu sign virama>. Older texts display the other vowel-less form \(\mathbf{F}\), called nakārā-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 \(\mathbf{F}\) or the new style glyph \(\mathbf{n}\). The character U+200C ZERO WIDTH NON-JOINER can be used to prevent interaction of this sequence with following consonants, as shown in Table 9-24.

Table 9-24. Rendering of Telugu \textit{na + virama}

<table>
<thead>
<tr>
<th>Font</th>
<th>Sequence</th>
<th>Rendering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Old Style</td>
<td>\textit{na + virama}</td>
<td>(\mathbf{F})</td>
</tr>
<tr>
<td></td>
<td>\textit{na + virama} + \text{ZWJ} + da</td>
<td>(\mathbf{F}\mathbf{\delta})</td>
</tr>
<tr>
<td>New Style</td>
<td>\textit{na + virama}</td>
<td>(\mathbf{n})</td>
</tr>
<tr>
<td></td>
<td>\textit{na + virama} + \text{ZWJ} + da</td>
<td>(\mathbf{n}\mathbf{\delta})</td>
</tr>
<tr>
<td>All Fonts</td>
<td>\textit{na + virama} + da</td>
<td>(\mathbf{n})</td>
</tr>
</tbody>
</table>

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

\[
U+0C30  ra + U+0C4D  virama + U+0C2E  ma \rightarrow \mathbf{r}\mathbf{ma}
\]

However, in older texts, U+0C30 telugu letter ra takes the reduced (or reph) form \(\mathbf{r}\) when it appears first in a consonant cluster, and the following consonant maintains its nominal form:

\[
U+0C30  ra + U+0C4D  virama + U+0C2E  ma \rightarrow \mathbf{r}\mathbf{\alpha}\mathbf{\epsilon} \ (\text{reph})
\]

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

\[
U+0C30  ra + U+0C4D  virama + U+200D  \text{ZWJ} + U+0C2E  ma \rightarrow \mathbf{r}\mathbf{\alpha}\mathbf{\epsilon}\]
To prevent display of a reph, U+200D zero width joiner is placed after the ra, but preceding the virama:

\[ U+0C30 \text{ ra } + U+200D \text{ ZWJ } + U+0CAD \text{ virama } + U+0C2E \text{ } \]

\[ \text{ma} \rightarrow \text{ } \]

**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 9.6, 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 halli, 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 9.1, Devanagari.**

---

### 9.8 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 9.1, Devanagari,** for further information.

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

**Principles of the Kannada Script**

Like Devanagari and related scripts, the Kannada script employs a halant, which is also known as a virama or vowel omission sign, U+0CCD Kannada sign virama. The halant nominally serves to suppress the inherent vowel of the consonant to which it is applied. The halant functions as a combining character. When a consonant loses its inherent vowel by the application of halant, it is known as a dead consonant. The dead consonants are the presentation forms used to depict the consonants without an inherent vowel. Their rendered forms in Kannada resemble the full consonant with the vertical stem replaced by the halant sign, which marks a character core. The stem glyph is graphically and historically related to the sign denoting the inherent /a/ vowel, U+0C85 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 9-25* 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>
<thead>
<tr>
<th>For</th>
<th>Use</th>
<th>Do Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>ಖ</td>
<td>0C8A</td>
<td>&lt;0C89, 0CBE&gt;</td>
</tr>
<tr>
<td>ಙ</td>
<td>0C94</td>
<td>&lt;0C92, 0CCC&gt;</td>
</tr>
<tr>
<td>ಚ</td>
<td>0CE0</td>
<td>&lt;0C8B, 0CBE&gt;</td>
</tr>
</tbody>
</table>

**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 Ø कङ्गानादा लेंथ्ट मार्क is provided as an encoding for the right side of the two-part vowel U+0CC7 ए कङ्गानादा वोल्व्यो ए should it be necessary for processing. Likewise, U+0CD6 ज कङ्गानादा एआइ लेंथ्ट मार्क is provided as an encoding for the right side of the two-part vowel U+0CC8 एआइ कङ्गानादा वोल्व्यो एआइ. 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 फ कङ्गानादा लेट्टर फँ is actually an obsolete Kannada letter that is transliterated in Dravidian scholarship as ẑ, ƚ, or ȓ. This form should have been named “LLLA”, rather than “f̄”, so the name in this standard is simply a mistake. This letter has not been actively used in Kannada since the end of the tenth century. Collations should treat U+0CDE as following U+0CB3 कङ्गानादा लेट्टर ल्ला.
Rendering Kannada

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

Explicit Virama (Halant). Normally, a halant character creates dead consonants, which in turn combine with subsequent consonants to form conjuncts. This behavior usually results in a halant sign not being depicted visually. Occasionally, this default behavior is not desired when a dead consonant should be excluded from conjunct formation, in which case the halant sign is visibly rendered. To accomplish this, U+200C zero width non-joiner is introduced immediately after the encoded dead consonant that is to be excluded from conjunct formation. See Section 9.1, Devanagari, for examples.

Consonant Clusters Involving RA. Whenever a consonant cluster is formed with the U+0CB0 D kannada letter ra as the first component of the consonant cluster, the letter 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 D ra + U+0CCD halant + U+0C95 ka → ꞎ rka
- U+0CB0 D ra + Ꞷ + U+0CCD halant + U+0C95 ka → Ꞷ rka
- U+0C95 ka + U+0CCD halant + U+0CB0 D ra → ꞎ kra

Modifier Mark Rules. In addition to the vowel signs, one 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 E kannada sign nukta. Two such modified consonants are used in the Kannada language: one representing the syllable za and one representing the syllable fa.

Avagraha Sign. A spacing mark called U+0CBD F 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 9.1, Devanagari.

9.9 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.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. Table 9-26 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.

Rendering Behavior. The shapes of Malayalam letters closely resemble those of Tamil. Malayalam, however, has a very full and complex set of conjunct consonant forms. 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 9-27 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 9-26. Malayalam Vowel Letters

<table>
<thead>
<tr>
<th>For</th>
<th>Use</th>
<th>Do Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>0D08:</td>
<td></td>
<td>&lt;0D07, 0D57&gt;</td>
</tr>
<tr>
<td>0D09:</td>
<td></td>
<td>&lt;0D09, 0D57&gt;</td>
</tr>
<tr>
<td>0D10:</td>
<td></td>
<td>&lt;0D0E, 0D46&gt;</td>
</tr>
<tr>
<td>0D12:</td>
<td></td>
<td>&lt;0D12, 0D3E&gt;</td>
</tr>
<tr>
<td>0D14:</td>
<td></td>
<td>&lt;0D12, 0D57&gt;</td>
</tr>
</tbody>
</table>

Table 9-27. Malayalam Orthographic Reform

<table>
<thead>
<tr>
<th>Syllable</th>
<th>Older Orthography</th>
<th>Reformed Orthography</th>
</tr>
</thead>
<tbody>
<tr>
<td>ku</td>
<td>కూ</td>
<td>కూ</td>
</tr>
<tr>
<td>gu</td>
<td>గు</td>
<td>గు</td>
</tr>
<tr>
<td>chu</td>
<td>చు</td>
<td>చు</td>
</tr>
<tr>
<td>ju</td>
<td>జు</td>
<td>జు</td>
</tr>
<tr>
<td>mu</td>
<td>ము</td>
<td>ము</td>
</tr>
<tr>
<td>tu</td>
<td>తు</td>
<td>తు</td>
</tr>
<tr>
<td>nu</td>
<td>ను</td>
<td>ను</td>
</tr>
<tr>
<td>bhu</td>
<td>బు</td>
<td>బు</td>
</tr>
<tr>
<td>ru</td>
<td>రు</td>
<td>రు</td>
</tr>
<tr>
<td>si</td>
<td>సి</td>
<td>సి</td>
</tr>
<tr>
<td>lu</td>
<td>లు</td>
<td>లు</td>
</tr>
<tr>
<td>kā</td>
<td>కా</td>
<td>కా</td>
</tr>
<tr>
<td>gā</td>
<td>గా</td>
<td>గా</td>
</tr>
<tr>
<td>chū</td>
<td>చు</td>
<td>చు</td>
</tr>
<tr>
<td>jī</td>
<td>జి</td>
<td>జి</td>
</tr>
<tr>
<td>nū</td>
<td>ను</td>
<td>ను</td>
</tr>
<tr>
<td>tū</td>
<td>తు</td>
<td>తు</td>
</tr>
<tr>
<td>nū</td>
<td>ను</td>
<td>ను</td>
</tr>
<tr>
<td>bhu</td>
<td>బు</td>
<td>బు</td>
</tr>
<tr>
<td>rū</td>
<td>రు</td>
<td>రు</td>
</tr>
<tr>
<td>sū</td>
<td>సు</td>
<td>సు</td>
</tr>
<tr>
<td>hū</td>
<td>హు</td>
<td>హు</td>
</tr>
</tbody>
</table>

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
candrakala in Malayalam. Table 9-28 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>
<thead>
<tr>
<th>Table 9-28. Malayalam Conjuncts</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{ae} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) (k(\text{\textsuperscript{}})a)</td>
</tr>
<tr>
<td>( \text{ae} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) (k(\text{\textsuperscript{}})a)</td>
</tr>
<tr>
<td>( \text{\textsuperscript{}} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) ((\text{\textsuperscript{}})a)</td>
</tr>
<tr>
<td>( \text{\textsuperscript{}} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) ((\text{\textsuperscript{}})a)</td>
</tr>
<tr>
<td>( \text{\textsuperscript{}} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) ((\text{\textsuperscript{}})a)</td>
</tr>
<tr>
<td>( \text{\textsuperscript{}} + \text{\textsuperscript{}} + \text{\textsuperscript{}} \rightarrow \text{\textsuperscript{}} ) ((\text{\textsuperscript{}})a)</td>
</tr>
</tbody>
</table>

When the candrakala sign is visibly shown in Malayalam, it usually indicates the suppression of the inherent vowel, but it sometimes indicates instead a reduced schwa sound [\(\text{\textsuperscript{}}\)], often called “half-u” or samvruthokaram. In the later case, there can also be a -u vowel sign, and the base character can be a vowel letter. In all cases, the candrakala 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 9-29.

<table>
<thead>
<tr>
<th>Table 9-29. Candrakala Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\text{\textsuperscript{}}) /paal/ milk 0D2A, 0D3E, 0D32, 0D41, 0D4D</td>
</tr>
<tr>
<td>(\text{\textsuperscript{}}) /milk on which day? 0D0E, 0D4D, 0D28, 0D4D, 0D28, 0D3E</td>
</tr>
<tr>
<td>(\text{\textsuperscript{}}) /aishiil/ than ice 0D10, 0D36, 0D40, 0D28, 0D4D, 0D4D, 0D02</td>
</tr>
</tbody>
</table>

The anusvara can be seen after after vowel letters, as in \(\text{\textsuperscript{}}\)O \(\text{\textsuperscript{}}\)O \(\text{\textsuperscript{}}\)O <0D08, 0D02, 0D02, 0D02, 0D02>. Vowel signs can also be seen after digits, as in 3\(\text{\textsuperscript{}}\)O \(\text{\textsuperscript{}}\)O <0033, 0035, 0035, 0035E, 0032>. More generally, rendering engines should be prepared to handle Malayalam letters (including vowel letters), digits (both European and Malayalam), dashes, U+00A0 no-break space and U+25CC dotted circle as base characters for the Malayalam vowel signs, U+0D4D Malayalam sign virama, U+0D02 Malayalam sign anusvara, and U+0D03 Malayalam sign visarga. They should also be prepared to handle multiple combining marks on those bases.

Chillu Characters. The six chillu or cillak\(\textsuperscript{\text{\textsuperscript{}}}\)aram characters, U+0D7A..U+0D7F, encode dead consonants (those without an inherent vowel). To simplify the discussion here, the formal names of the characters are shortened to use the terms that are typically used in spoken discussion of the chillu characters: chillu-n for Malayalam letter chillu n, and so forth.

In Malayalam-language text, chillu characters never start a word. The chillu letters -\(\text{\textsuperscript{}}\)n, -\(\text{\textsuperscript{}}\)n, -rr, -l, and -ll are quite common; chillu-k is not very common.
Prior to Unicode Version 5.1, the representation of text with chillus was problematic, and not clearly described in the text of the standard. Because older data will use different representation for chillus, implementations must be prepared to handle both kinds of data. For the chillu letters considered in isolation, Table 9-30 shows the relation between their representation in Unicode Version 5.0 and earlier, and the recommended representation starting with Unicode Version 5.1.

**Table 9-30. Atomic Encoding of Malayalam Chillus**

<table>
<thead>
<tr>
<th>Visual</th>
<th>Representation in 5.0 and Prior</th>
<th>Preferred 5.1 Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ೊ</td>
<td>NNA, VIRAMA, ZWJ (0D23, 0D4D, 200D)</td>
<td>0D7A MALAYALAM LETTER CHILLU NN</td>
</tr>
<tr>
<td>ೋ</td>
<td>NA, VIRAMA, ZWJ (0D28, 0D4D, 200D)</td>
<td>0D7B MALAYALAM LETTER CHILLU N</td>
</tr>
<tr>
<td>ೌ</td>
<td>RA, VIRAMA, ZWJ (0D30, 0D4D, 200D)</td>
<td>0D7C MALAYALAM LETTER CHILLU RR</td>
</tr>
<tr>
<td>್</td>
<td>LA, VIRAMA, ZWJ (0D32, 0D4D, 200D)</td>
<td>0D7D MALAYALAM LETTER CHILLU L</td>
</tr>
<tr>
<td>೎</td>
<td>LLA, VIRAMA, ZWJ (0D33, 0D4D, 200D)</td>
<td>0D7E MALAYALAM LETTER CHILLU LL</td>
</tr>
<tr>
<td>೏</td>
<td>undefined</td>
<td>0D7F MALAYALAM LETTER CHILLU K</td>
</tr>
</tbody>
</table>

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

The letter Ɂ ra is normally read /r/. Repetition of that sound is written by two occurrences of the letter: ɁɁ. Each occurrence can bear a vowel sign.

Repetition of the letter, written either Ɂ or ɁɁ, is also used for the sound /tt/. The sequence of two Ɂ 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 ɁɁ in text is ambiguous in reading. The reader must generally use the context to understand if this is read /rr/ or /tt/. It is only when a vowel part appears between the two Ɂ that the reading is unambiguously /rr/. 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 (ignon) or two separate letters (gnome).

The sequence <0D31, 0D31> is rendered as ɁɁ, regardless of the reading of that text. The sequence <0D31, 0D4D, 0D31> is rendered as Ɂ. In both cases, vowels signs can be used as appropriate, as shown in Table 9-31.

**Table 9-31. Malayalam /rr/ and /tt/**

<table>
<thead>
<tr>
<th>Malayalam</th>
<th>Representation</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ೊ</td>
<td>0D2A 0D3E 0D31 0D31</td>
<td>/paatta/ cockroach</td>
</tr>
<tr>
<td>ೋ</td>
<td>0D2A 0D3E 0D31 0D4D 0D31</td>
<td>/maattoli/ echo</td>
</tr>
<tr>
<td>ೌ</td>
<td>0D2E 0D3E 0D31 0D46 0D31 0D3E 0D32 0D3F</td>
<td>/maattoli/ echo</td>
</tr>
<tr>
<td>್</td>
<td>0D2E 0D3E 0D31 0D46 0D31 0D4D 0D31 0D3E 0D32 0D3F</td>
<td>/maattoli/ echo</td>
</tr>
</tbody>
</table>
A very similar situation exists for the combination of ചില്ല്-n and ര ആ. When used side by side, ചില്ല് can be read either /nr/ or /nt/, while ര ആ is always read /nt/.

The sequence <0D7B, 0D31> is rendered as ചില്ല്, regardless of the reading of that text. The sequence <0D7B, 0D4D, 0D31> is rendered as ര ആ. In both cases, vowels signs can be used as appropriate, as shown in Table 9-32.

### Table 9-32. Malayalam /nr/ and /nt/

<table>
<thead>
<tr>
<th>Malayalam</th>
<th>U+0D2C 0D3E 0D31 0D31</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>ചില്ല്</td>
<td>ചില്ല്</td>
<td>/aantoo/</td>
</tr>
<tr>
<td>ര ആ</td>
<td>ര ആ</td>
<td>/engool/</td>
</tr>
<tr>
<td>മാതാമാതാ</td>
<td>മാതാമാതാ</td>
<td>/baattar/</td>
</tr>
<tr>
<td>മാതാത്ത</td>
<td>മാതാത്ത</td>
<td>(name of a place)</td>
</tr>
<tr>
<td>മാതാത്ത</td>
<td>മാതാത്ത</td>
<td>/surgatt/</td>
</tr>
<tr>
<td>മാതാത്ത</td>
<td>മാതാത്ത</td>
<td>/temparari/</td>
</tr>
<tr>
<td>മാതാത്ത</td>
<td>മാതാത്ത</td>
<td>/lekcaragoot/</td>
</tr>
</tbody>
</table>

### 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 over the consonant following it. Conceptually, the dot reph is analogous to the sequence <ര ആ, വിരാമ, C2> normally assumes the C2 conjoining form. U+0D4E MALAYALAM LETTER DOT REPH occurs first, in logical order, even though it displays as a dot above the succeeding consonant. It has the character properties of a letter, and is not considered a combining mark.

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

The Malayalam dot reph was in common print usage until 1970, but has fallen into disuse. Words that formerly used dot reph are now spelled using U+0D7C MALAYALAM CHILLU RR or the respective C2-conjoining forms. The dot reph form is predominantly used by those who completed elementary education in Malayalam prior to 1970.

### Historic Characters.

The four characters, അവഗ്രഹ, വോക്ലിക് ര്ര്, വോക്ലിക് ല്, and വോക്ലിക് ല്ല്, are only used to write Sanskrit words in the Malayalam script. The അവഗ്രഹ is the most common of the four, followed by the വോക്ലിക് ല് sign. There are six characters used for the archaic number system, including characters for numbers 10, 100, 1000 and fractions. 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.

### Special Characters.

In modern times, the dominant practice is to write the dependent form of the അ യ 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 “Ω” on the left side and “Π” 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 above.

For a detailed discussion of the use of two-part vowels, see “Two-Part Vowels” in Section 9.6, Tamil.

**Punctuation.** Danda and double danda marks as well as some other unified punctuation used with Malayalam are found in the Devanagari block; see Section 9.1, Devanagari.