Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

http://www.unicode.org/book/aboutbook.html

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see Joining the Unicode Consortium at

http://www.unicode.org/consortium/join.html

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.
Includes bibliographical references and index.
ISBN 0-321-48091-0 (hardcover : alk. paper)
1. Unicode (Computer character set) I. Allen, Julie D.
II. Unicode Consortium.
QA268.U545 2007
005.7'22—dc22

2006023526

ISBN 0-321-48091-0
Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.
First printing, October 2006

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 *puḷḷi* is used. The virama sign nominally serves to suppress the inherent vowel of the consonant to which it is applied; it is a combining character, with its shape varying from script to script.

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

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

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

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, Kanauji, Konkani, Kului, Kumaoni, Kurku, Kurukh, Marwari, Mundari, Newari, Palpa, and Santali.

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

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

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

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

In November 1991, at the time *The Unicode Standard, Version 1.0,* was published, the Bureau of Indian Standards published a new version of ISCII in Indian Standard (IS) 13194:1991. This new version partially modified the layout and repertoire of the ISCII-1988 standard. Because of these events, the Unicode Standard does not precisely follow the layout of the current version of ISCII. Nevertheless, the Unicode Standard remains a superset of the ISCII-1991 repertoire except for a number of new Vedic extension characters defined in IS 13194:1991 *Annex G—Extended Character Set for Vedic*. 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.

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 chart.

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.

The Devanagari script has only one character denoting a left-side dependent vowel sign: U+093F DEVANAGARI VOWEL SIGN I. 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.

Table 9-1. Devanagari Vowel Letters

To Represent	Use	Do Not Use
ऑ	0904	<0905, 0946>
आ	0906	<0905, 093E>
ऊ	090A	<0909, 0941>
ऍ	090D	<090F, 0945>
ऎ	090E	<090F, 0946>
ऐ	0910	<090F, 0947>
ऑ	0911	<0905, 0949>
ओ	0912	<0905, 094A>
ओ	0913	<0905, 094B>
औ	0914	<0905, 094C>

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

$$TA_n + VIRAMA_n \rightarrow TA_d$$

त + ◌् → त्

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

Under normal circumstances, a consonant cluster is depicted with a conjunct glyph if such a glyph is available in the current font. In the absence of a conjunct glyph, the one or more dead consonants that form part of the cluster are depicted using half-form glyphs. In the absence of half-form glyphs, the dead consonants are depicted using the nominal consonant forms combined with visible virama signs (see *Figure 9-2*).

Figure 9-2. Conjunct Formations in Devanagari

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

ग् + ध → ग्ध

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

क् + क → क्क

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

क् + ष → क्ष

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

र् + क → र्क

A number of types of conjunct formations appear in these examples: (1) a half-form of GA in its combination with the full form of DHA; (2) a vertical conjunct K.KA; and (3) a fully ligated conjunct K.SSA, in which the components are no longer distinct. In example (4) in *Figure 9-2*, the dead consonant RA_d is depicted with the nonspacing combining mark RA_{sup} (*repha*).

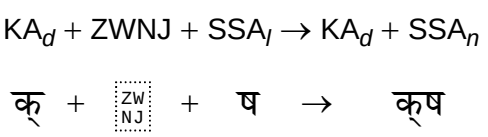
A well-designed Indic script font may contain hundreds of conjunct glyphs, but they are not encoded as Unicode characters because they are the result of ligation of distinct letters.

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 क्ष (K.SSA$_n$).

Figure 9-3. Preventing Conjunct Forms in Devanagari



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

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

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

In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 9-4* would normally produce the full conjunct form क्ष (K.SSA$_n$).

Figure 9-4. Half-Consonants in Devanagari

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

क् + ZWJ + ष → क्ष

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

$GA_d + ZWJ \rightarrow GA_h$

ग् + ZWJ → ग

Other Indic scripts have similar half-forms for the initial consonants of a conjunct. Some, such as Oriya, also have similar half-forms for the final consonants; those are represented as shown in *Figure 9-6*.

Figure 9-6. Half-Consonants in Oriya

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

କ୍ + ZWJ + ◌୍ + ତ → କ୍ତ

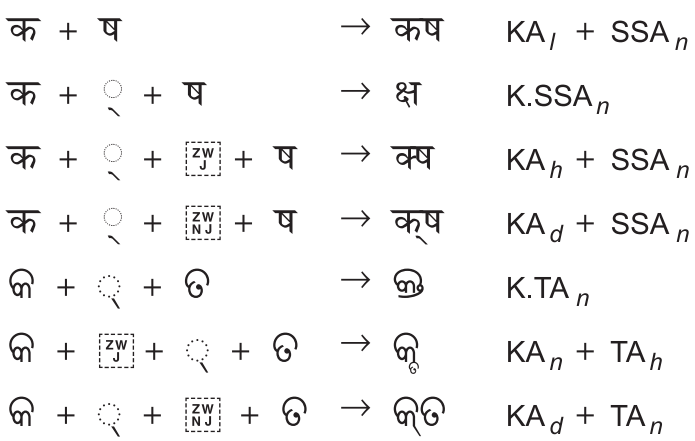
In the absence of the ZERO WIDTH JOINER, the sequence in *Figure 9-6* would normally produce the full conjunct form କ୍ତ ($K.TA_n$).

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

Figure 9-7. Consonant Forms in Devanagari and Oriya



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

Rendering Devanagari

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

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

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

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

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

RA_{sup}	A nonspacing combining mark glyph form of U+0930 DEVANAGARI LETTER RA positioned above or attached to the upper part of a base glyph form. This form is also known as *repha*.
RA_{sub}	A nonspacing combining mark glyph form of U+0930 DEVANAGARI LETTER RA positioned below or attached to the lower part of a base glyph form.
V_{vs}	Glyph depicting the dependent vowel sign form of a vowel V.
$VIRAMA_n$	The nominal glyph form of the nonspacing combining mark depicting U+094D DEVANAGARI SIGN VIRAMA.

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

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

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

$$TA_n + VIRAMA_n \rightarrow TA_d$$

त + ◌् → त्

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

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

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

Displayed Output

र् + क → क + ◌ं → र्क

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

र् + र् → र् + ◌ं → र्र्

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

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

Displayed Output

र् + ज् + ञ → ज्ञ + ◌ॅ → ज्ञॅ

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

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

Displayed Output

र् + ग् + घ → ग + घ + ◌ॅ → ग्घॅ

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

$$RRA_n + VIRAMA_n \rightarrow RRA_h$$

ऱ् + ◌् → ऱ

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

$$RA_d + ZWJ \rightarrow RA_h$$

र् + ZWJ → ऱ

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

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

Displayed Output

ठ् + र → ठ + ◌़ → ठ्र

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

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

Displayed Output

फ् + र → फ + ◌् → फ्र

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

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

त् + र् → त + ◌् + ◌् → त्र्

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

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

त्र् + य → त्र्य

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

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

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

क + ◌़ + ◌् → क़्

R10 ***Other modifying marks, in particular bindus and svaras, apply to the orthographic syllable as a whole and should follow (in the memory representation) all other characters that constitute the syllable. The bindus should follow any vowel signs, and the svaras should come last. The relative placement of these marks is***

horizontal rather than vertical; the horizontal rendering order may vary according to typographic concerns.

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

क + ा + ँ → काँ

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

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

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

ज् + ञ → ज्ञ $\qquad$ ट् + ठ → ट्ठ

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

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

स् + त् + र → स् + त्र → स्त्र

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

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

क्ष् + य → क्ष्य

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

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

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

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

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

र + ◌ु → रु र + ◌ू → रू

Memory Representation and Rendering Order. The storage of plain text in Devanagari and all other Indic scripts generally follows phonetic order; that is, a CV syllable with a dependent vowel is always encoded as a consonant letter C followed by a vowel sign V in the memory representation. This order is employed by the ISCII standard and corresponds to both the phonetic order and the keying order of textual data (see *Figure 9-8*).

Figure 9-8. Rendering Order in Devanagari

Character Order Glyph Order

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

क + ◌ि → कि

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

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

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

त् + र + ◌ि → त्र + ◌ि → त्रि

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

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

त् + ZWNJ + र + ि → त्रि

Sample Half-Forms. *Table 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 9-2. Sample Devanagari Half-Forms

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

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

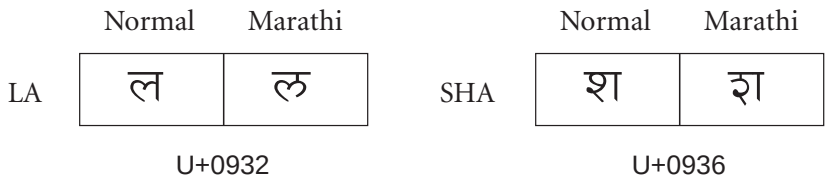
Sample Half-Ligature Forms. In addition to half-form glyphs of individual consonants, half-forms are used to depict conjunct ligature forms. A sample of such forms is shown in *Table 9-4*. These forms are glyphs, not characters. They may be encoded explicitly using ZERO WIDTH JOINER as shown. In normal conjunct formation, they may be used spontaneously to depict a conjunct ligature in combination with subsequent consonant forms.

Table 9-4. Sample Devanagari Half-Ligature Forms

क	+	◌्	+	ष	+	◌्	+	ZWJ	→	क्ष
ज	+	◌्	+	ञ	+	◌्	+	ZWJ	→	ज्ञ
त	+	◌्	+	त	+	◌्	+	ZWJ	→	त्त
त	+	◌्	+	र	+	◌्	+	ZWJ	→	त्र
श	+	◌्	+	र	+	◌्	+	ZWJ	→	श्र

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.

Figure 9-9. Marathi Allographs



Combining Marks. Devanagari and other Indic scripts have a number of combining marks that could be considered diacritic. One class of these marks, known as bindus, is represented by U+0901 DEVANAGARI SIGN CANDRABINDU and U+0902 DEVANAGARI SIGN ANUSVARA. These marks indicate nasalization or final nasal closure of a syllable. U+093C DEVANAGARI SIGN NUKTA is a true diacritic. It is used to extend the basic set of consonant letters by modifying them (with a subscript dot in Devanagari) to create new letters. U+0951..U+0954 are a set of combining marks used in transcription of Sanskrit texts.

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—lack a distinct digit for zero.

Punctuation and Symbols. 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.

U+0970 ° DEVANAGARI ABBREVIATION SIGN appears after letters or combinations of letters and marks the sequence as an abbreviation.

Encoding Structure. The Unicode Standard organizes the nine principal Indic scripts in blocks of 128 encoding points each. 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.

Other Languages. 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 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.

9.2 Bengali

Bengali: U+0980–U+09FF

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

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 9-5* 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 9-5. Bengali Vowel Letters

To Represent	Use	Do Not Use
আ	0986	<0985, 09BE>

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 placed on each side of a consonant letter or cluster—for example, U+09CB BENGALI VOWEL SIGN O and U+09CC BENGALI VOWEL SIGN AU. The vowel signs are coded in each case in the position in the charts isomorphic with the corresponding vowel in Devanagari. Hence U+09CC BENGALI VOWEL SIGN AU is isomorphic with U+094C DEVANAGARI VOWEL SIGN AU. 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.

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

Rendering Behavior. Like other Brahmic scripts in the Unicode Standard, Bengali uses the *hasant* to form conjunct characters. For example, U+0995 ক BENGALI LETTER KA + U+09CD ◌্ BENGALI SIGN VIRAMA + U+09B7 ষ BENGALI LETTER SSA yields the conjunct ক্ষ KSSA, which is pronounced *khya* in Assamese. 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 9-6*.

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.

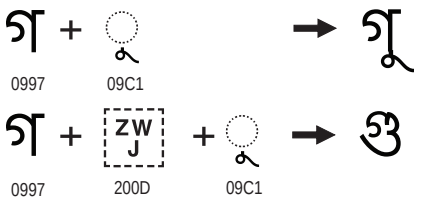
No semantic distinctions are made in Bengali text on the basis of the two different presentations of these consonant-vowel combinations. However, some users consider it important that implementations support both forms and that the distinction be representable in plain text. This may be accomplished 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.*)

Table 9-6. Bengali Consonant-Vowel Combinations

	Code Points	Ligated	Non-ligated
gu	<0997, 09C1>	গু	গু
ru	<09B0, 09C1>	রু	র়ু
rū	<09B0, 09C2>	রূ	র়ূ
śu	<09B6, 09C1>	শু	শ়ু
hu	<09B9, 09C1>	হু	হ়ু
hṛ	<09B9, 09C3>	হৃ	হ়ৃ

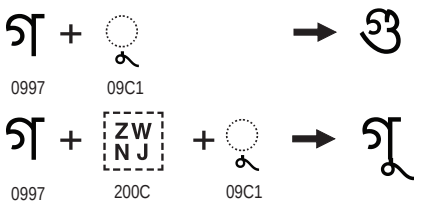
A given font implementation can choose whether to treat the ligature forms of the consonant-vowel combinations as the defaults for rendering. If the non-ligated form is the default, then ZWJ can be inserted to request a ligature, as shown in *Figure 9-10*.

Figure 9-10. Requesting Bengali Consonant-Vowel Ligature



If the ligated form is the default for a given font implementation, then ZWNJ can be inserted to block a ligature, as shown in *Figure 9-11*.

Figure 9-11. Blocking Bengali Consonant-Vowel Ligature



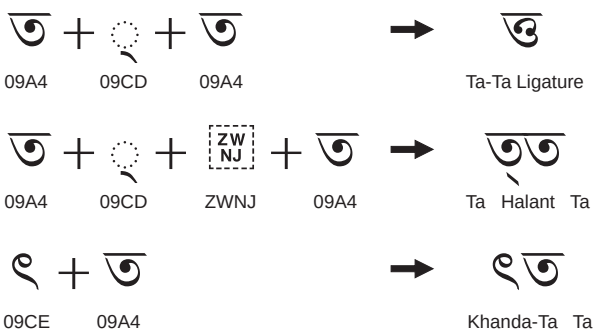
Khanda Ta. In Bengali, a dead consonant *ta* makes use of a special form, U+09CE BENGALI LETTER KHANDA TA. This form is used in all contexts except where it is immediately followed by one of the consonants: *ta*, *tha*, *na*, *ba*, *ma*, *ya*, or *ra*.

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

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

The Bengali syllable *tta* illustrates the usage of *khanda ta* when followed by *ta*. The syllable *tta* is normally represented with the sequence <U+09A4 *ta*, U+09CD *hasant*, U+09A4 *ta*>. That sequence will normally be displayed using a single glyph *tta* ligature, as shown in the first example in *Figure 9-12*.

Figure 9-12. Bengali Syllable *tta*



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

The Unicode Standard also provides an explicit way to show the *hasant* glyph. To do so, a ZERO WIDTH NON-JOINER is inserted after the *hasant*. That sequence is always displayed with the explicit *hasant*, as shown in the second example in *Figure 9-12*.

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 9-12*.

Ya-phalaa. *Ya-phalaa* (pronounced *jo-phola* in Bengali) is a presentation form of U+09AF য BENGALI LETTER YA. Represented by the sequence <U+09CD ◌্ BENGALI SIGN VIRAMA, U+09AF য BENGALI LETTER YA>, *ya-phalaa* has a special form ্য. When combined with U+09BE ◌া BENGALI VOWEL SIGN AA, it is used for transcribing [æ] as in the "a" in the English word "bat." *Ya-phalaa* can be applied to initial vowels as well:

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

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

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

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

Further examples:

অ + ◌্ + য + ◌া = অ্যা

এ + ◌্ + য + ◌া = এ্যা

ত + ◌্ + য + ◌া = ত্যা

Interaction of Repha aand 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:

র + ◌্ + ম ➡ র্ম as in কর্ম (karma)

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

ক + ◌্ + য ➡ ক্য as in বাক্য (bakyô)

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

র + ◌্ + য ➡ র্য or র‍্য

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

র + ◌্ + য ➡ র্য

09B0 09CD 09AF

র + ZWJ + ◌্ + য ➡ র‍্য

09B0 200D 09CD 09AF

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

Punctuation. 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*.

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 ਪੱਗ *pagg*, "turban," can be represented with the sequence <0A2A, 0A71, 0A17> (or <pa, addak, ga>) in Unicode, while in ISCII-1991 it would be <pa, ga, virama, ga>.

Punjabi does not have complex combinations of consonant sounds. Furthermore, the orthography is not strictly phonetic, and sometimes the inherent /a/ sound is not pronounced. For example, the word ਗੁਰਮੁਖੀ *gurmukhī* is represented with the sequence <0A17, 0A41, 0A30, 0A2E, 0A41, 0A16, 0A40>, which could be transliterated as *gura-*

mukhii; 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 consonants modified by both U+0A4B GURMUKHI VOWEL SIGN OO, rendered above the consonant, and U+0A41 GURMUKHI VOWEL SIGN U, rendered below the consonant. Because of the combining classes of those characters, the sequences <C, 0A41, 0A4B> and <C, 0A4B, 0A41> are not canonically equivalent. To avoid ambiguity in representation, the second sequence, with U+0A4B before U+0A41, should be used in such cases. When a consonant is not present, the same sequence of vowels may be represented by attaching the dependent vowel sign *u* to the independent vowel letter U+0A13 GURMUKHI LETTER OO: <0A13, 0A41>. More generally, when a consonant or independent vowel is modified by two vowel signs with one above and one below, the vowel sign above should occur first.

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. Gurmukhi Vowel Letters

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

Ordering. U+0A73 GURMUKHI URA and U+0A72 GURMUKHI IRI are the first and third "letters" of the Gurmukhi syllabary, respectively. They are used as bases or bearers for some of the independent vowels, while U+0A05 GURMUKHI LETTER A is both the second "letter" and the base for the remaining independent vowels. As a result, the collation order for Gurmukhi is based on a seven-by-five grid:

- The first row is U+0A73 *ura*, U+0A05 *a*, U+0A72 *iri*, U+0A38 *sa*, U+0A39 *ha*.
- This row is followed by five main rows of consonants, grouped according to the point of articulation, as is traditional in all South and Southeast Asian scripts.

- The semiconsonants follow in the seventh row: U+0A2F *ya*, U+0A30 *ra*, U+0A32 *la*, U+0A35 *va*, U+0A5C *rra*.
- The letters with *nukta*, added later, are presented in a subsequent eighth row if needed.

Rendering Behavior. For general principles regarding the rendering of the Gurmukhi script, see the rules for rendering in *Section 9.1, Devanagari*. In many aspects, Gurmukhi is simpler than Devanagari. There are no half-consonants, no half-forms, no *repha* (upper form of U+0930 DEVANAGARI LETTER RA), and no real ligatures. Rules R2–R5, R11, and R14 do not apply. Conversely, the behavior for subscript RA (rules R6–R8 and R13) applies to U+0A39 GURMUKHI LETTER HA and U+0A35 GURMUKHI LETTER VA, which also have subjoined forms, called *pairin* in Punjabi. The subjoined form for RA is like a knot, while the subjoined HA and VA are written the same as the base form, without the top bar, but are reduced in size. As described in rule R13, they attach at the bottom of the base consonant, and will "push" down any attached vowel sign for U or UU. When U+0A2F GURMUKHI LETTER YA follows a dead consonant, it assumes a different form called *addha* in Punjabi, without the leftmost part, and the dead consonant returns to the nominal form, as shown in *Table 9-8*.

Table 9-8. Gurmukhi Conjuncts

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

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

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

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

Table 9-9. Additional Pairin and Addha Forms in Gurmukhi (Continued)

ਸ	+	◌੍	+	ਯ	→	ਸ੍ਯ	(*sya*)	pairin ya
ਸ	+	◌੍	+	ਥ	→	ਸ੍ਥ	(*stha*)	addha tha
ਸ	+	◌੍	+	ਮ	→	ਸ੍ਮ	(*sma*)	addha ma

Older texts also exhibit another feature that is not found in modern Gurmukhi—namely, the use of a half- or reduced form for the first consonant of a cluster, whereas the modern practice is to represent the second consonant in a half- or reduced form. Joiners can be used to request this older rendering, as shown in *Table 9-10*. The reduced form of an initial U+0A30 GURMUKHI LETTER RA is similar to the Devanagari superscript RA (*repha*), but this usage is rare, even in older texts.

Table 9-10. Use of Joiners in Gurmukhi

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

A rendering engine for Gurmukhi should make accommodations for the correct positioning of the combining marks (see *Section 5.13, Rendering Nonspacing Marks*, and particularly *Figure 5-12*). 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 9-11* 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 9-11. Gujarati Vowel Letters

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

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 9-12*.

Table 9-12. Gujarati Conjuncts

ક	+	◌્	+	ષ	→	ક્ષ	(*kṣa*)
જ	+	◌્	+	ઞ	→	જ્ઞ	(*jña*)
ત	+	◌્	+	ય	→	ત્ય	(*tya*)

Table 9-12. Gujarati Conjuncts (Continued)

ટ + ◌્ + ટ → ટ્ટ	(*ṭṭa*)
ર + ◌્ + ક → ર્ક	(*rka*)
ક + ◌્ + ર → ક્ર	(*kra*)

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 surroundrant vowel U+0B4C ORIYA VOWEL SIGN AU.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 9-13* 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 9-13. Oriya Vowel Letters

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

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

କ + ◌୍ → କ୍ (*k*)

The virama is also used to form conjunct consonants, as shown in *Table 9-14*.

Table 9-14. Oriya Conjuncts

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

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

ର + ◌୍ + ପ → ର୍ପ (*rpa*)

ପ + ◌୍ + ର → ପ୍ର (*pra*)

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

ଅ + ଙ + ◌୍ + କ → ଅଙ୍କ (*aṅka*)

ଅ + ◌ଂ + କ → ଅଂକ (*aṁka*)

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 9-15*.

Table 9-15. Oriya Vowel Placement

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

U+0B01 ORIYA SIGN CANDRABINDU is used for nasal vowels:

କ + ◌ଁ → କଁ (*kaṁ*)

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

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

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

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

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

Punctuation and Symbols. Danda and double danda marks as well as some other unified punctuation used with Oriya are found in the Devanagari block; see *Section 9.1, Devanagari*. The mark U+0B70 ORIYA ISSHAR is placed before names of persons who are deceased.

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 க^2 = *pha*, க^3 = *ba*, and க^4 = *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 (Puḷḷi). Because the Tamil encoding in the Unicode Standard is based on ISCII-1988 (Indian Script Code for Information Interchange), it makes use of the *abugida* model. An abugida treats the basic consonants as containing an inherent vowel, which can be canceled by the use of a visible mark, called a *virama* in Sanskrit. In most Brahmi-derived scripts, the placement of a virama between two consonants implies the deletion of the inherent vowel of the first consonant and causes a conjoined or subjoined consonant cluster. In those scripts, ZERO WIDTH NON-JOINER is used to display a visible virama, as shown previously in the Devangari example in *Figure 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 <C1, U+0BCD TAMIL SIGN VIRAMA, C2, ...>) is normally displayed with explicit viramas (which are called *puḷḷi* in Tamil). The conjuncts *kssa* and *shra* are traditionally displayed by conjunct ligatures, as illustrated for *kssa* in *Figure 9-13*, but nowadays tend to be displayed using an explicit *puḷḷi* as well.

Figure 9-13. Kssa Ligature in Tamil

க + ◌் + ஷ → க்ஷ *kṣa*

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

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

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

Tamil Vowels

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

ன + ி ≠ ன + ◌் + இ

Left-Side Vowels. The Tamil vowels U+0BC6 ெ◌, U+0BC7 ே◌, and U+0BC8 ை◌ are reordered in front of the consonant to which they are applied. When occurring in a syllable, these vowels are rendered to the left side of their consonant, as shown in *Table 9-16*.

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 9-14*.

Table 9-16. Tamil Vowel Reordering

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

Figure 9-14. Tamil Two-Part Vowels

◌ொ 0BCA ≡ ◌ெ + ◌ா 0BC6 + 0BBE

◌ோ 0BCB ≡ ◌ே + ◌ா 0BC7 + 0BBE

◌ௌ 0BCC ≡ ◌ெ + ◌ௗ 0BC6 + 0BD7

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

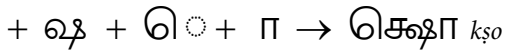
In the process of rendering, these two-part vowels are transformed into the two separate glyphs equivalent to those on the right, which are then subject to vowel reordering, as shown in *Table 9-17*.

Table 9-17. Tamil Vowel Splitting and Reordering

Memory Representation			Display
க	◌ொ		கொ
க	◌ெ	◌ா	கொ
க	◌ோ		கோ
க	◌ே	◌ா	கோ
க	◌ெ◌ௗ		கௌ
க	◌ெ	◌ௗ	கௌ

Even in the case where a two-part vowel occurs with a conjunct consonant or consonant cluster, the left part of the vowel is reordered around the conjunct or cluster, as shown in *Figure 9-15*.

Figure 9-15. Vowel Reordering Around a Tamil Conjunct



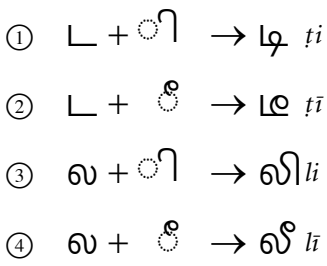
For either left-side vowels or two-part vowels, the ordering of the elements is unambiguous: the consonant or consonant cluster occurs first in the memory representation, followed by the vowel.

Tamil Ligatures

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

***Ligatures with Vowel* i.** The vowel signs *i* ி and *ii* ீ form ligatures with the consonant *tta* ட as shown in examples 1 and 2 of *Figure 9-16*. 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 9-16*.

Figure 9-16. Tamil Ligatures with *i*



***Ligatures with Vowel* u.** The vowel signs *u* ு and *uu* ூ normally ligate with their consonant, as shown in *Table 9-18*. In the first column, the basic consonant is shown; the second column illustrates the ligation of that consonant with the *u* vowel sign; and the third column illustrates the ligation with the *uu* vowel sign.

Table 9-18. Tamil Ligatures with *u*

x	*x* + ு	*x* + ூ
க	கு	கூ
ங	ங	ஙூ
ச	சு	சூ
ஞ	ஞு	ஞூ

x	*x* + ு	*x* + ூ
ப	பு	பூ
ம	மு	மூ
ய	யு	யூ
ர	ரு	ரூ

Table 9-18. Tamil Ligatures with *u* (Continued)

x	*x* + ◌ு◌	*x* + ◌ூ
ட	டு	டூ
ண	ணு	ணூ
த	து	தூ
ந	நு	நூ
ன	னு	னூ

x	*x* + ◌ு◌	*x* + ◌ூ
ற	று	றூ
ல	லு	லூ
ள	ளு	ளூ
ழ	ழு	ழூ
வ	வு	வூ

With certain consonants, ஜ, ஷ, ஸ, ஹ, and the conjunct க்ஷ, the vowel signs *u* ◌ுand *uu* ◌ூ take a distinct spacing form, as shown in *Figure 9-17*.

Figure 9-17. Spacing Forms of Tamil *u*

ஜ + ◌ு → ஜு *ju*

ஜ + ◌ூ → ஜூ *jū*

***Ligatures with* ra.** Based on typographical preferences, the consonant *ra* ர may change shape to ர், when it ligates. Such change, if it occurs, will happen only when the ர் form of U+0BB0 ர TAMIL LETTER RA would not be confused with the nominal form ர of U+0BBE TAMIL VOWEL SIGN AA (namely, when ர is combined with ◌், ◌ி , or ◌ீ). This change in shape is illustrated in *Figure 9-18*.

Figure 9-18. Tamil Ligatures with *ra*

ர + ◌் → ர் *r*

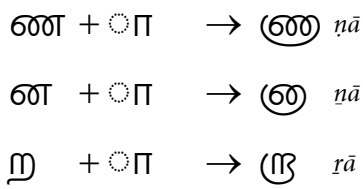
ர + ◌ி → ரி *ri*

ர + ◌ீ → ரீ *rī*

However, various governmental bodies mandate that the basic shape of the consonant *ra* ர should be used for these ligatures as well, especially in school textbooks. Media and literary publications in Malaysia and Singapore mostly use the unchanged form of *ra* ர.

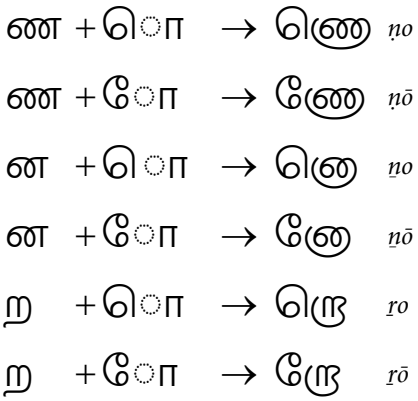
Ligatures with* aa *in Traditional Tamil Orthography. In traditional Tamil orthography, the vowel sign *aa* ◌ா optionally ligates with ணா, னா, or றா, as illustrated in *Figure 9-19*.

Figure 9-19. Tamil Ligatures with *aa*



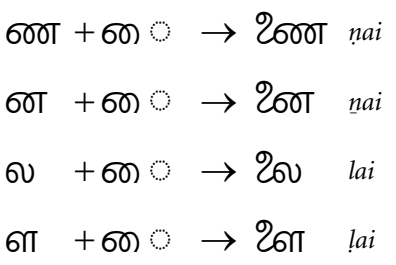
These ligations also affect the right-hand part of two-part vowels, as shown in *Figure 9-20*.

Figure 9-20. Tamil Ligatures with *o*



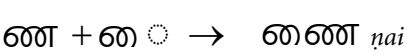
Ligatures with* ai *in Traditional Tamil Orthography. In traditional Tamil orthography, the left-side vowel sign *ai* ை◌ is also subject to a change in form. It is rendered as ◌ை when it occurs on the left side of ணை, னை, ல, or ளை, as illustrated in *Figure 9-21*.

Figure 9-21. Tamil Ligatures with *ai*



By contrast, in modern Tamil orthography, this vowel does not change its shape, as shown in *Figure 9-22*.

Figure 9-22. Vowel *ai* in Modern Tamil



***Tamil* aytham.** The character U+0B83 TAMIL SIGN VISARGA is normally called *aytham* in Tamil. It is historically related to the *visarga* in other Indic scripts, but has become an ordinary spacing letter in Tamil. It is used to modify the sound of other consonants and, in particular, to represent 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*.

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 9-19* 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. 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

Table 9-19. Telugu Vowel Letters

To Represent	Use	Do Not Use
ఓ	0C13	<0C12, 0C55>
ఔ	0C14	<0C12, 0C4C>

script writes conjunct characters with subscript letters. Many Telugu letters have a v-shaped headstroke, which is a structural mark corresponding to the horizontal bar in Devanagari and the arch in Oriya script. When a virama (called *virāmamu* in Telugu) or certain vowel signs are added to a letter with this headstroke, it is replaced:

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

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

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

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

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

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

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

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*.

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 ಅ KANNADA LETTER A. In contrast, a live consonant is a consonant that retains its inherent vowel or is written with an explicit dependent vowel sign. The dead consonant is defined as a sequence consisting of a consonant letter followed by a halant. The default rendering for a dead consonant is to position the halant as a combining mark bound to the consonant letterform.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 9-20* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 9-20. Kannada Vowel Letters

To Represent	Use	Do Not Use
ೆೌ	0C94	<0C92, 0CCC>

Consonant Conjuncts. Kannada is also noted for a large number of consonant conjunct forms that serve as ligatures of two or more adjacent forms. This use of ligatures takes place in the context of a consonant cluster. A written consonant cluster is defined as a sequence of characters that represent one or more dead consonants followed by a normal live consonant. A separate and unique glyph corresponds to each part of a Kannada consonant conjunct. Most of these glyphs resemble their original consonant forms—many without the implicit vowel sign, wherever applicable.

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

- The first consonant of the cluster is rendered with the implicit vowel or a different dependent vowel appearing as the terminal element of the cluster.

- The remaining consonants (consonants between the first consonant and the terminal vowel element) appear in conjunct consonant glyph forms in phonetic order. They are generally depicted directly below or to the lower right of the first consonant.

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

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

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

Kannada Letter LLLA. U+0CDE ೞ KANNADA LETTER FA is actually an obsolete Kannada letter that is transliterated in Dravidian scholarship as *ẓ*, *l̤*, or *ṟ*. This form should have been named "LLLA", rather than "FA", 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 KANNADA LETTER LLA.

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.

If any invisible base is required for the display of dependent vowels without any consonant base, U+200C ZERO WIDTH NON-JOINER can be used. It can also be used to provide proper collation of the words containing dead consonants.

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

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

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

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 ◌಼ 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 ऽ 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 9-21* 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. In *Table 9-22*, an initial con-

Table 9-21. Malayalam Vowel Letters

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

sonant plus the vowel sign yields a syllable. Both the older orthography and the newer orthography are shown on the right.

Table 9-22. Malayalam Orthographic Reform

Syllable		Older Orthography	Newer Orthography
ku	ക + ു	കു	കു
gu	ഗ + ു	ഗു	ഗു
chu	ച + ു	ചു	ചു
ju	ജ + ു	ജു	ജു
ṇu	ണ + ു	ണു	ണു
tu	ത + ു	തു	തു
nu	ന + ു	നു	നു
bhu	ഭ + ു	ഭു	ഭു
ru	ര + ു	രു	രു
śu	ശ + ു	ശു	ശു
hu	ഹ + ു	ഹു	ഹു
kū	ക + ൂ	കൂ	കൂ
gū	ഗ + ൂ	ഗൂ	ഗൂ
chū	ച + ൂ	ചൂ	ചൂ
jū	ജ + ൂ	ജൂ	ജൂ
ṇū	ണ + ൂ	ണൂ	ണൂ
tū	ത + ൂ	തൂ	തൂ
nū	ന + ൂ	നൂ	നൂ
bhū	ഭ + ൂ	ഭൂ	ഭൂ

Table 9-22. Malayalam Orthographic Reform (Continued)

rū	ര + ൂ	ൠ	രൂ
śū	ശ + ൂ	ൡ	ശൂ
hū	ഹ + ൂ	ൣ	ഹൂ

Like other Brahmic scripts in the Unicode Standard, Malayalam uses the virama to form conjunct characters (see *Table 9-23*); this is known as *candrakala* in Malayalam. There are both horizontal and vertical conjuncts. The visible virama usually shows the suppression of the inherent vowel, but sometimes indicates a reduced schwa sound [ə], often called "half-u".

Table 9-23. Malayalam Conjuncts

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

Five sonorant consonants merge with the virama when they appear in syllable-final position with no inherent vowel. A consonant when so merged is called *cillakṣaram* or *chillu*:

ൺ	*ṇ*
ൻ	*n*
ർ	*r*
ൽ	*l*
ൾ	*ḷ*

It is important to note the use of the ZERO WIDTH JOINER and ZERO WIDTH NON-JOINER in these environments:

ന + ് + മ → ന്മ *(nma)*

ന + ് + [ZWNJ] + മ → ന്‌മ *(nma)*

ന + ് + [ZWJ] + മ → ന്മ *(nma)*

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

