

This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (<http://www.unicode.org/errata/>). For information on more recent versions of the standard, see <http://www.unicode.org/standard/versions/enumeratedversions.html>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

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

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

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

Cover and CD-ROM label design: Steve Mehallo, <http://www.mehallo.com>

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: <http://www.awprofessional.com>

Library of Congress Cataloging-in-Publication Data

The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].

p. cm.

Includes bibliographical references and index.

ISBN 0-321-18578-1 (alk. paper)

1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004

005.7'2—dc21

2003052158

Copyright © 1991–2003 by Unicode, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher or Unicode, Inc. Printed in the United States of America. Published simultaneously in Canada.

For information on obtaining permission for use of material from this work, please submit a written request to the Unicode Consortium, Post Office Box 39146, Mountain View, CA 94039-1476, USA, Fax +1 650 693 3010 or to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300 Boston, MA 02116, USA, Fax: +1 617 848 7047.

ISBN 0-321-18578-1

Text printed on recycled paper

1 2 3 4 5 6 7 8 9 10—CRW—0706050403

First printing, August 2003

Chapter 9

South Asian Scripts

The following South Asian scripts are described in this chapter:

- Devanagari
- Bengali
- Gurmukhi
- Gujarati
- Oriya
- Tamil
- Telugu
- Kannada
- Malayalam
- Sinhala
- Tibetan
- Limbu

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. This northern branch includes such modern scripts as Bengali, Gurmukhi, and Tibetan; the southern branch includes scripts such as Malayalam and Tamil.

The major official scripts of India proper, including Devanagari, 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. Sinhala has a virama-based model, but is not structurally mapped to ISCII. Tibetan stands apart, using a subjoined consonant model for conjoined consonants, reflecting its somewhat different structure and usage. The Limbu script makes use of an explicit encoding of syllable-final consonants.

Many of the character names in this group of scripts represent the same sounds, and naming conventions are similar across the range.

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–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 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)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 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 upon 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 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).

A one-to-one correspondence exists between the independent vowels and the dependent vowel signs. Independent vowels are sometimes represented by a sequence consisting of the independent form of the vowel /a/ followed by a dependent vowel sign. *Figure 9-1* illustrates this relationship (see the notation formally described in the “Rules for Rendering” later in this section).

Figure 9-1. Dependent Versus Independent Vowels

<u>/a/ + Dependent Vowel</u>		<u>Independent Vowel</u>
$A_n + I_{VS} \rightarrow I_{VS} + A_n$	\approx	I_n
अ + ि → अि	\approx	इ
$A_n + U_{VS} \rightarrow A_n + U_{VS}$	\approx	U_n
अ + ु → अु	\approx	उ

The combination of the independent form of the default vowel /a/ (in the Devanagari script, U+0905 DEVANAGARI LETTER A) with a dependent vowel sign may be viewed as an alternative spelling of the phonetic information normally represented by an isolated independent vowel form. However, these two representations should not be considered equivalent for the purposes of rendering. Higher-level text processes may choose to consider these alternative spellings equivalent in terms of information content, but such an equivalence is not stipulated by this standard.

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 9-2*.

Figure 9-2. Dead Consonants

$$TA_n + VIRAMA_n \rightarrow TA_d$$

$$त + ः \rightarrow त्$$

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(s). 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 9-3*).

Figure 9-3. Conjunct Formations

$$(1) GA_d + DHA_l \rightarrow GA_n + DHA_n \quad (3) KA_d + SSA_l \rightarrow K.SSA_n$$

$$ग + ध \rightarrow गध$$

$$क + ष \rightarrow क्ष$$

$$(2) KA_d + KA_l \rightarrow K.KA_n \quad (4) RA_d + KA_l \rightarrow KA_l + RA_{sup}$$

$$क + क \rightarrow क्क$$

$$र् + क \rightarrow कर्$$

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. Note that in example (4) in *Figure 9-3*, 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 9-4*, the use of ZERO WIDTH NON-JOINER prevents the default formation of the conjunct form क्ष (K.SSA_n).

Figure 9-4. Preventing Conjunct Forms

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

$$\text{क्} + \text{ZW NJ} + \text{ष} \rightarrow \text{क्ष}$$

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

In certain cases, it is desirable to prevent a dead consonant from assuming full conjunct formation yet still not appear with an explicit virama. In these cases, the half-form of the consonant is used. To explicitly encode a half-consonant form, the Unicode Standard adopts the convention of placing the character U+200D ZERO WIDTH JOINER immediately after the encoded dead consonant. The ZERO WIDTH JOINER denotes a nonvisible letter that presents linking or cursive joining behavior on either side (that is, to the previous or following letter). Therefore, in the present context, the ZERO WIDTH JOINER may be 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 encoded as shown in *Figure 9-5*.

Figure 9-5. Half-Consonants

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

$$\text{क्} + \text{ZW J} + \text{ष} \rightarrow \text{क्ष}$$

- In the absence of the ZERO WIDTH JOINER, this sequence would normally produce the full conjunct form क्ष (K.SSA_n).

This encoding of half-consonant forms also applies in the absence of a base letterform. That is, this technique may also be used to encode independent half-forms, as shown in *Figure 9-6*.

Figure 9-6. Independent Half-Forms

$$GA_d + ZWJ \rightarrow GA_h$$

$$\text{ग} + \text{ZW J} \rightarrow \text{ग}$$

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, or the half-form of a dead consonant (see Figure 9-7).

Figure 9-7. Consonant Forms

क	→	क	KA _l				
क	+	◌̣	→	क̣	KA _d		
क	+	◌̣	+	◌̣ ^{ZW} _J	→	क	KA _h

Rendering

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.

It is important to emphasize that 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 <i>repha</i> .
RA _{sub}	A nonspacing combining mark glyph form of U+0930 DEVANAGARI LETTER RA positioned below or attached to the lower part of a base glyph form.
V _{vs}	Glyph depicting the dependent vowel sign form of a vowel V.
VIRAMA _n	The nominal glyph form of the nonspacing combining mark depicting U+094D DEVANAGARI SIGN VIRAMA.

- A virama character is not always depicted; when it is depicted, it adopts this nonspacing mark form.

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

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

$$TA_n + VIRAMA_n \rightarrow TA_d$$

$$त + ष् \rightarrow त्$$

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.

$$RA_d + KA_l \rightarrow KA_l + RA_{sup} \quad \text{Displayed Output}$$

$$र् + क \rightarrow क + ि \rightarrow क्$$

$$RA_d^1 + RA_d^2 \rightarrow RA_d^2 + RA_{sup}^1$$

$$र् + र् \rightarrow र् + ि \rightarrow र्$$

- 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.

$$RA_d + JA_d + NYA_l \rightarrow J.NYA_n + RA_{sup} \quad \text{Displayed Output}$$

$$र् + ज् + ञ \rightarrow ज्ञ + ि \rightarrow ज्ञ्$$

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

$$RA_d + GA_d + GHA_l \rightarrow GA_n + GHA_l + RA_{sup} \quad \text{Displayed Output}$$

$$र् + ग् + घ \rightarrow र + घ + ि \rightarrow र्घ$$

- 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.

$$RRA_n + VIRAMA_n \rightarrow RRA_h$$

$$\text{र} + \text{्} \rightarrow \text{ः}$$

- 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} .

$$RA_d + ZWJ \rightarrow RA_h$$

$$\text{र} + \text{ZWJ} \rightarrow \text{ः}$$

- 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 .

$$TTHA_d + RA_l \rightarrow TTHA_n + RA_{sub} \text{ Displayed Output}$$

$$\text{ठ} + \text{र} \rightarrow \text{ठ} + \text{्} \rightarrow \text{ठ}$$

- 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.

$$PHA_d + RA_l \rightarrow PHA_n + RA_{sub} \text{ Displayed Output}$$

$$\text{फ} + \text{र} \rightarrow \text{फ} + \text{्} \rightarrow \text{फ्र}$$

- 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.

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

$$\text{त्} + \text{र्} \rightarrow \text{त} + \text{्} + \text{्} \rightarrow \text{त्र्}$$

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

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

$$\text{त्र्} + \text{य} \rightarrow \text{त्र्य}$$

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.*

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

$$\text{क} + \text{्} + \text{्} \rightarrow \text{क्}$$

- R10 *The other modifying marks, 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. In particular, 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.*

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

$$\text{क} + \text{ा} + \text{ँ} \rightarrow \text{काँ}$$

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(s) 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.*

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

$$\text{ज्} + \text{ञ} \rightarrow \text{ज्ञ} \quad \text{ट्} + \text{ठ} \rightarrow \text{ट्ठ}$$

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

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

$$\text{स्} + \text{त्} + \text{र} \rightarrow \text{स्} + \text{त्र} \rightarrow \text{स्त्र}$$

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

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

$$\text{क्ष्} + \text{य} \rightarrow \text{क्ष्य}$$

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).

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

$$क + ँ \rightarrow क्र \quad फ + ँ \rightarrow फ्र$$

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_l with U_{vs} or UU_{vs} .

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

$$र + ु \rightarrow रु \quad र + ू \rightarrow रू$$

Memory Representation and Rendering Order. The order for 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 and the keying order of textual data (see Figure 9-8).

Figure 9-8. Rendering Order

<u>Character Order</u>	<u>Glyph Order</u>
$KA_n + I_{vs} \rightarrow$	$I_{vs} + KA_n$
क + ि \rightarrow	कि

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

R15 When the dependent vowel 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. For example:

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

$$त् + र + ि \rightarrow त्र + ि \rightarrow त्रि$$

Sample Half-Forms. Table 9-1 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 9-1. Sample Half-Forms

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

Sample Ligatures. Table 9-2 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 9-2. Sample Ligatures

क	◌्	क	क्क
क	◌्	त	क्त
क	◌्	र	क्र
क	◌्	ष	क्ष
ड	◌्	क	ड्क
ड	◌्	ख	ड्ख
ड	◌्	ग	ड्ग
ड	◌्	घ	ड्घ
ञ	◌्	ज	ञ्ज
ज	◌्	ञ	ज्ञ
द	◌्	घ	द्घ
द	◌्	द	द्द
द	◌्	ध	द्ध
ट	◌्	ठ	ट्ठ
ठ	◌्	ठ	ठ्ठ
ड	◌्	ग	ड्ग
ड	◌्	ड	ड्ड
ड	◌्	ढ	ड्ढ
त	◌्	त	त्त
त	◌्	र	त्र
न	◌्	न	न्न
फ	◌्	र	फ्र
श	◌्	र	श्र
ह	◌्	म	ह्र
ह	◌्	य	ह्र
ह	◌्	ल	ह्र

Table 9-2. Sample Ligatures (Continued)

द	्	ब	द्व
द	्	भ	द्व
द	्	म	द्व
द	्	य	द्व
द	्	व	द्व
ट	्	ट	द्व

ह	्	व	ह्व
ह		े	ह
र		ु	रु
र		्	रु
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 9-3*. 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 9-3. Sample Half-Ligature Forms

क	्	ष	्	ZW J	क्ष
ज	्	ञ	्	ZW J	ज्ञ
त	्	त	्	ZW J	त्त
त	्	र	्	ZW J	त्र
श	्	र	्	ZW J	श्च

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 previously in rule R5.

Figure 9-9. Marathi Allographs

	normal	Marathi		normal	Marathi
LA	ल	लु	SHA	श	श
	U+0932			U+0936	

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. Corresponding forms occur in many other Indic scripts. U+0965 || DEVANAGARI DOUBLE DANDA marks the end of a verse in traditional texts. U+0970 ° DEVANAGARI ABBREVIATION SIGN appears after letters or combinations.

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.

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. Sindhi makes use of U+0974 DEVANAGARI LETTER SHORT YA. Several implosive consonants in Sindhi are realized as combinations with nukta and U+0952 DEVANAGARI STRESS SIGN ANUDATTA. Konkani makes use of additional sounds that can be made 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 Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Munda, Naga, Rian, and Santali, in north-eastern India.




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

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.



Khanda Ta. The Bengali syllable “tta” is notable. It is encoded with the following sequence:

```
U+09A4 BENGALI LETTER TA
U+09CD BENGALI SIGN VIRAMA (= hasant)
U+09A4 BENGALI LETTER TA
```

The sequence will normally be displayed using the single glyph *tta* ligature . It is also possible for the sequence to be displayed using a *khanda ta* glyph followed by a full *ta* glyph , or with a full *ta* glyph combined with a virama glyph and followed by a full *ta* glyph . The choice of form actually displayed depends on the display engine, based on availability of glyphs in the font.

The Unicode Standard provides an explicit way to encode a half-letter form. To do this, a ZERO WIDTH JOINER is inserted after the virama:

```
U+09A4 BENGALI LETTER TA
U+09CD BENGALI SIGN VIRAMA (= hasant)
U+200D ZERO WIDTH JOINER
U+09A4 BENGALI LETTER TA
```

This sequence is always displayed as a *khanda ta* glyph followed by a full *ta* glyph . Even if the consonant “ta” is not present, the sequence U+09A4, U+09CD, U+200D is displayed as a *khanda ta* glyph .

The Unicode Standard provides an explicit way to show the virama glyph. To do this, a ZERO WIDTH NON-JOINER is inserted after the virama:

U+09A4 BENGALI LETTER TA

U+09CD BENGALI SIGN VIRAMA (= *hasant*)

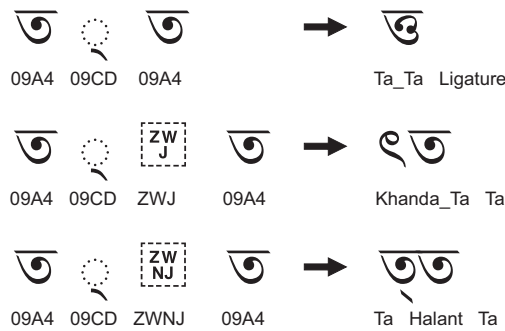
U+200C ZERO WIDTH NON-JOINER

U+09A4 BENGALI LETTER TA

This sequence is always displayed as a full *ta* glyph combined with a virama glyph and followed by a full *ta* glyph ত্ত.

A summary image of various sequences is shown in *Figure 9-10*.

Figure 9-10. Bengali Khanda Ta



Ya-phalaa. *Ya-phalaa* (pronounced *jo-phola* in Bengali) is a presentation form of U+09AF য় BENGALI LETTER YA. Represented by the sequence <U+09CD ্ BENGALI VIRAMA, U+09AF য় BENGALI LETTER YA>, *ya-phalaa* has a special form জ়. 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:

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

এ়া = <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:

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

Further examples:

অ + ্ + য + া = অ়া

এ + ্ + য + া = এ়া

ত + ্ + য + া = ত়া

Rendering Behavior. Like other Brahmic scripts in the Unicode Standard, Bengali uses the virama to form conjunct characters. For example, U+0995 ক BENGALI LETTER KA + U+09CD ্ BENGALI 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*.

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 Lahnda, and is closely related to Devanagari structurally. The script is closely associated with Sikh and Sikhism, but it is used everyday in East Punjab. (West Punjab, now in Pakistan, uses the Arabic script.)

Encoding Principles. The Gurmukhi block is based on ISCII-88, 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. Note that for unification with the other scripts, ISCII-91 considers *rri* to be equivalent to *dda+nukta*, but this decomposition is not considered in Unicode. On the other hand, ISCII-91 does not consider U+0A36 to be equivalent to <U+0A38, U+0A3C>, or U+0A33 to be equivalent to <U+0A32, U+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 after AA, II, EE, AI, OO, AU, and the independent vowels U and UU; *tippi* is used in the other contexts. Older texts may depart from this requirement. Note that ISCII-91 uses only one encoding point for both marks.

U+0A71 GURMUKHI ADDAK is a special sign to indicate that the following consonant is geminate. Note that ISCII-91 does not have a specific code point for addak, and encodes it as a cluster. For example, the word ਪੱਗ *pagg*, “turban”, is encoded as <U+0A2A, U+0A71, U+0A17> /pa/addak/ga/ in Unicode, while in ISCII-91 it would instead 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 written <U+0A17, U+0A41, U+0A30, U+0A2E, U+0A41, U+0A16. U+0A40>, which could be transliterated as *gurmukhii*; 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.

Ordering. U+0A72 GURMUKHI IRI and U+0A73 GURMUKHI URA 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+0A72 *iri*, U+0A05 *a*, U+0A73 *ura*, 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 the South and Southeast Asian scripts.
- The semi-consonants follow in the seventh row: U+0A2F *ya*, U+0A30 *ra*, U+0A32 *la*, U+0A35 *va*, U+0A5C *rri*.

- 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. On the other hand, 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. Other letters behaved similarly in old inscriptions. When U+0A2F GURMUKHI LETTER YA follows a dead consonant, it assumes a different form, without the leftmost part, and the dead consonant returns to the nominal form, as shown in *Table 9-4*.

Table 9-4. Gurmukhi Conjuncts

ਮ	+	ੜ	+	ਹ	→	ਮੁ	(mha)
ਪ	+	ੜ	+	ਰ	→	ਪੁ	(pra)
ਦ	+	ੜ	+	ਵ	→	ਦੁ	(dwa)
ਦ	+	ੜ	+	ਯ	→	ਯੁ	(dya)

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 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.

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 *khodo*, 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 9-5*.

Table 9-5. Gujarati Conjuncts

ક	+	◌̣	+	૫	→	ક્ષ	(<i>kṣa</i>)
જ	+	◌̣	+	ઞ	→	જ્ઞ	(<i>jña</i>)
ત	+	◌̣	+	૫	→	ત્પ	(<i>tpa</i>)
ટ	+	◌̣	+	ટ	→	ટ્ટ	(<i>ṭṭa</i>)
ર	+	◌̣	+	ક	→	ર્ક	(<i>rka</i>)
ક	+	◌̣	+	ર	→	ક્ર	(<i>kra</i>)

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.

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 + ୠ → କ (k)

The virama is also used to form conjunct consonants, as shown in *Table 9-6*.

Table 9-6. Oriya Conjuncts

କ + ୠ + ଷ → କ୍ଷ (kṣa)

କ + ୠ + ଡ → କ୍ତ (kta)

ଡ + ୠ + କ → ଢକ (ṭka)

ଡ + ୠ + ଯ → ଢ୍ୟ (ṭya)

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 + ୠ + ପ → ର୍ପ (rpa)

ପ + ୠ + ର → ପ୍ର (pra)

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

ଅ + ଓ + ୠ + କ → ଅକ୍ (aṅka)

ଅ + ୠ + କ → ଅକ (aṅka)

As with other scripts, some dependent vowels are rendered in front of their consonant, some after it, and some are placed above, some below, and some both in front of and after their consonant. A few of the dependent vowels fuse with their consonants. See *Table 9-7*.

Table 9-7. Oriya Vowel Placement

କ	+	ା	→	କା	(kā)
କ	+	ି	→	କି	(ki)
କ	+	ି	→	କି	(kī)
କ	+	ୁ	→	କୁ	(ku)
କ	+	ୁ	→	କୁ	(kū)
କ	+	ୃ	→	କୃ	(kr)
କ	+	େ	→	କେ	(ke)
କ	+	ୈ	→	କୈ	(kai)
କ	+	ୋ	→	କୋ	(ko)
କ	+	ୌ	→	କୌ	(kau)

U+0B02 ORIYA CANDRABINDU is used for nasal vowels:

କ + ୠ → କିଁ (kain)

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 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 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 ବ is sometimes found for *va*, and in others ଶ and ଣ have been shown, which in a more modern type style would be ଶ and ଣ. 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]:

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 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 ୱ, ୱ, and ୱ.

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 a South Indian script. South Indian scripts are structurally related to the North Indian scripts, but they are used to write the Dravidian languages of southern India and of Sri Lanka, which are genetically unrelated to the North Indian languages such as Hindi, Bengali, and Gujarati. The shapes of letters in the South Indian scripts are generally quite distinct from the shapes of letters in Devanagari and its related scripts.

The Tamil script 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 $\text{ᱚ}^2 = pha$, $\text{ᱚ}^3 = ba$, and $\text{ᱚ}^4 = bha$. The characters U+00B2, U+00B3, and U+2074 can be used to preserve this distinction in plain text. The Tamil script also lacks conjunct consonant forms.

Virama (Pulḷi). The Tamil script uses the Unicode virama model to form conjunct consonants. In Tamil the virama (U+0BCD) is known as *pulḷi*. The virama is normally fully depicted in Tamil text.

Rendering of Tamil Script. The South Indic scripts function in much the same way as Devanagari, with the additional feature of two-part vowels. This description provides minimal requirements for legibly rendering interchanged Tamil text. As with any script, a more complex procedure can add rendering characteristics, depending on the font and application.

It is important to emphasize that in a font that is capable of rendering Tamil, the number of glyphs is greater than the number of Tamil characters.

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

ஊ + ி ≠ ஊ + ி + ஐ

In the Tamil script, a consonant cluster is any sequence of one or more consonants separated by viramas, possibly terminated with a virama.

Two-Part Vowels. Certain Tamil vowels consist of two discontinuous elements. A sequence of two Unicode code points can be used to express equivalent spellings for these vowels. This representation is similar to the case of letters such as “â”, which can be spelled either with “a” followed by a nonspacing “̂” or with a single Unicode character “â”. In the following examples, the representation on the left, which is a single code point, is the preferred form and the form in common use.

ஊ (0BCA) ≈ ஊ + ி (0BC6 + 0BBE)

஋ (0BCB) ≈ ஋ + ி (0BC7 + 0BBE)

஌ (0BCC) ≈ ஌ + ி (0BC6 + 0BD7)

Note that the ெள in the third example is *not* U+0BB3 TAMIL LETTER LLA; it is U+0BD7 TAMIL AU LENGTH MARK.

In the rendering process, the precomposed form on the left is transformed into the two separate glyphs equivalent to those on the right, which are then subject to vowel reordering (see Table 9-9 below).

Vowel Reordering. As shown in Table 9-8, the following vowels are always reordered in front of the previous consonant cluster:

ெ (0BC6) ே (0BC7) ை (0BC8)

Table 9-8. Vowel Reordering

Memory Representation		Display
க	ெ	கெ
க	ே	கே
க	ை	கை

The same effect occurs with the results of vowel splitting (see Table 9-9).

Table 9-9. Vowel Splitting and Reordering

Memory Representation			Display
க	ெ	ா	கொ
க	ெ	ா	கொ
க	ே	ா	கோ
க	ே	ா	கோ
க	ெ	ள	கொள
க	ெ	ள	கொள

In both cases, the ordering of the elements is *unambiguous*: the consonant (cluster) occurs *first* in the memory representation. The vowel *au* ெள also has two discontinuous parts and can be composed using the TAMIL AU LENGTH MARK.

Ligatures. The following examples illustrate the range of ligatures available in Tamil. These changes take place after vowel reordering and vowel splitting. Tamil includes very few conjunct consonants; most ligatures are located between a vowel and a neighboring consonant.

1. Conjunct consonants:

$$\text{க} + \text{ஃ} + \text{ஷ} \rightarrow \text{கஷ} \text{ (kṣa)}$$

Vowel reordering occurs around conjunct consonants. For example:

$$\text{க} + \text{ஃ} + \text{ஷ} + \text{ஶ} + \text{ஶ} \rightarrow \text{கஷஶ} \text{ (kṣo)}$$

2. In older Tamil orthography, the vowel *aa* ஶ optionally ligates with ண, ன, or ற on its left:

$$\text{ண} + \text{ஶ} \rightarrow \text{ணஶ} \text{ (ṅā)}$$

$$\text{ன} + \text{ஶ} \rightarrow \text{னஶ} \text{ (ṅā)}$$

$$\text{ற} + \text{ஶ} \rightarrow \text{றஶ} \text{ (rā)}$$

Because this process takes place after reordering and splitting, the following ligatures may also occur in older Tamil orthography:

$$\text{ண} + \text{ஶ} \rightarrow \text{ணஶ} \text{ (ṅo)}$$

$$\text{ண} + \text{ஶ} \rightarrow \text{ணஶ} \text{ (ṅō)}$$

$$\text{ன} + \text{ஶ} \rightarrow \text{னஶ} \text{ (ṅo)}$$

$$\text{ன} + \text{ஶ} \rightarrow \text{னஶ} \text{ (ṅō)}$$

$$\text{ற} + \text{ஶ} \rightarrow \text{றஶ} \text{ (ro)}$$

$$\text{ற} + \text{ஶ} \rightarrow \text{றஶ} \text{ (rō)}$$

3. The vowel signs *i* ஶ and *ii* ஶ form ligatures with the consonant *tta* ழ on their left.

$$\text{ழ} + \text{ஶ} \rightarrow \text{ழஶ} \text{ (ti)}$$

$$\text{ழ} + \text{ஶ} \rightarrow \text{ழஶ} \text{ (tī)}$$

These vowels also often change shape or position slightly to link up with the appropriate shape of other consonants on their left. For example:

$$\text{ல} + \text{ஶ} \rightarrow \text{லி} \text{ (li)}$$

$$\text{ல} + \text{ஶ} \rightarrow \text{லி} \text{ (lī)}$$

4. The vowel signs *u* ஶ and *uu* ஶ typically change form or ligate (see Table 9-10).

Table 9-10. Ligating Vowel Signs

x	$x + \overset{\circ}{\underset{ }{i}}$	$x + \overset{\circ}{\underset{u}{i}}$
க	கூ	கூ
ங	ங்	ங்
ச	சு	சூ
ஞ	ஞு	ஞூ
ட	டு	டூ
ண	ணு	ணூ
த	து	து
ந	நு	நூ
ன	னு	னூ

x	$x + \overset{\circ}{\underset{ }{i}}$	$x + \overset{\circ}{\underset{u}{i}}$
ப	பு	பூ
ம	மு	மூ
ய	யு	யூ
ர	ரு	ரூ
ற	று	றூ
ல	லு	லூ
ள	ளு	ளூ
ழ	ழு	ழூ
வ	வு	வூ

5. To the right of ஐ, ஓ, ஸ, ஹ, or சை, the vowel signs $u \overset{\circ}{\underset{|}{i}}$ and $uu \overset{\circ}{\underset{u}{i}}$ have a spacing form (see Figure 9-11).

Figure 9-11. Spacing Forms of Vowels

ஐ + $\overset{\circ}{\underset{|}{i}}$ → ஐ[□] (*ju*)

ஐ + $\overset{\circ}{\underset{u}{i}}$ → ஐ[□] (*jū*)

6. In older Tamil orthography, the vowel sign *ai* $\overset{\circ}{\underset{|}{i}}$ changes to $\overset{\circ}{\underset{|}{i}}$ to the left of ண, ன, ல, or ள.

$\overset{\circ}{\underset{|}{i}}$ + ண → $\overset{\circ}{\underset{|}{i}}$ ண (*ṅai*)

$\overset{\circ}{\underset{|}{i}}$ + ன → $\overset{\circ}{\underset{|}{i}}$ ன (*ṅai*)

$\overset{\circ}{\underset{|}{i}}$ + ல → $\overset{\circ}{\underset{|}{i}}$ ல (*lai*)

$\overset{\circ}{\underset{|}{i}}$ + ள → $\overset{\circ}{\underset{|}{i}}$ ள (*lai*)

Remember that this change takes place after the vowel reordering. In the first example, the vowel sign *ai* $\overset{\circ}{\underset{|}{i}}$ follows ண in the memory representation. After vowel reordering, it is on the left of ண, and thus changes form. The complete process is:

ண + $\overset{\circ}{\underset{|}{i}}$ → $\overset{\circ}{\underset{|}{i}}$ + ண → $\overset{\circ}{\underset{|}{i}}$ ண (*ṅai*)

In modern Tamil orthography, the last step is omitted:

ண + $\overset{\circ}{\underset{|}{i}}$ → $\overset{\circ}{\underset{|}{i}}$ + ண → ணை (*ṅai*)

7. The consonant *ra* ர changes shape to ரீ.

This change occurs when the ர form of U+0BB0 ர TAMIL LETTER RA would not be confused with the nominal form ரீ of U+0BBE TAMIL VOWEL SIGN AA (for example, when ர is combined with ி, ி, or ி).

$$\text{ர} + \text{ி} \rightarrow \text{ரீ} \text{ (r)}$$

$$\text{ர} + \text{ி} \rightarrow \text{ரீ} \text{ (ri)}$$

$$\text{ர} + \text{ி} \rightarrow \text{ரீ} \text{ (rī)}$$

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.

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

U+0C15 క *ka* + U+0C4D ీ *virama* + U+200C ZW ZERO WIDTH NON-JOINER → క̣ (*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. The Kannada 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 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.

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 consonant cluster is rendered with the implicit vowel or a different dependent vowel appearing as the terminal element of the consonant 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. Note that these two length marks have no independent existence in the Kannada writing system and do not play any part as independent codes in the traditional collation order.

Kannada Letter LLLA. U+0CDE ೃ KANNADA LETTER FA is actually an obsolete Kannada letter that is transliterated in Dravidian scholarship as *z*, *l*, or *r*. This form should have been named “LLLA”, rather than “FA”, so the name in this standard is simply a mistake. 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

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 Devanagari and some other Indian scripts, all 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 ೄ 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 ೄ *ra* + U+0CCD ು halant + U+0C95 ಕ *ka* → ಕ಼ (rka)

U+0CB0 ೄ *ra* + U+0CCD ು halant + U+200D zw + U+0C95 ಕ *ka* → ರ್ಕ
(rka)

U+0C95 ಕ *ka* + U+0CCD ು halant + U+0CB0 ೄ *ra* → ಕ್ಕ (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 ೪ 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.

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 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 9-11*, an initial consonant plus the vowel sign yields a syllable. Both the older orthography and the newer orthography are shown on the right.

Table 9-11. Malayalam Orthographic Reform

Syllable		Older Orthography	Newer Orthography
<i>ku</i>	ക + ു	ക	കു
<i>gu</i>	ഗ + ു	ഗ	ഗു
<i>chu</i>	ച + ു	ച	ചു
<i>ju</i>	ജ + ു	ജ	ജു
<i>ṅu</i>	ണ + ു	ണ	ണു
<i>tu</i>	ത + ു	ത	തു
<i>nu</i>	ന + ു	ന	നു
<i>bhu</i>	ഭ + ു	ഭ	ഭു
<i>ru</i>	ര + ു	ര	രു
<i>śu</i>	ശ + ു	ശ	ശു
<i>hu</i>	ഹ + ു	ഹ	ഹു
<i>kū</i>	ക + ൂ	ക	കു
<i>gū</i>	ഗ + ൂ	ഗ	ഗു
<i>chū</i>	ച + ൂ	ച	ചു
<i>jū</i>	ജ + ൂ	ജ	ജു
<i>ṅū</i>	ണ + ൂ	ണ	ണു
<i>tū</i>	ത + ൂ	ത	തു
<i>nū</i>	ന + ൂ	ന	നു
<i>bhū</i>	ഭ + ൂ	ഭ	ഭു
<i>rū</i>	ര + ൂ	ര	രു
<i>śū</i>	ശ + ൂ	ശ	ശു
<i>hū</i>	ഹ + ൂ	ഹ	ഹു

Like other Brahmic scripts in the Unicode Standard, Malayalam uses the virama to form conjunct characters; 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 9-12. Malayalam Conjuncts

ക	+	്	+	ഷ	→	കഷ	(kṣa)
ക	+	്	+	ക	→	കക	(kka)
ജ	+	്	+	ഞ	→	ജഞ	(jña)
ട	+	്	+	ട	→	ട്ട	(ṭṭa)
പ	+	്	+	പ	→	പ്പ	(ppa)
ച	+	്	+	ച	→	ച്ഛ	(ccha)
ബ	+	്	+	ബ	→	ബ്ബ	(bba)
ന	+	്	+	യ	→	ന്യ	(nya)
പ	+	്	+	ര	→	പ്ര	(pra)
ര	+	്	+	പ	→	രപ	(rpa)
ശ	+	്	+	വ	→	ശവ	(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 *cillaks.aram*:

ണ	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	→	നമ	(nma)		
ന	+	്	+	ZW NJ	+	മ	→	ന [̣] മ	(nma)
ന	+	്	+	ZW J	+	മ	→	ൻമ	(nma)

Special Characters. U+0D57 MALAYALAM AU LENGTH MARK is provided as an encoding for the right side of the two-part vowel U+0D4C MALAYALAM VOWEL SIGN AU. 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 marks as well as some other unified punctuation used with Malayalam are found in the Devanagari block; see Section 9.1, *Devanagari*.

9.10 Sinhala

Sinhala: U+0D80–U+0DFF

The Sinhala script, also known as Sinhalese, is used to write the Sinhala language, the majority language of Sri Lanka. It is also used to write the Pali and Sanskrit languages. The script is a descendant of Brahmi and resembles the scripts of South India in form and structure.

Sinhala differs from other languages of the region in that it has a series of prenasalized stops that are distinguished from the combination of a nasal followed by a stop. In other words, both forms occur and are written differently—for example, අඳ <U+0D85, U+0DAC> *aṅḍa* [a^ṅḍa] “sound” versus අඳ්ද <U+0D85, U+0DAB, U+0DCA, U+0DA9> *aṅḍa* [aṅḍa] “egg”. In addition, Sinhala has separate distinct signs for both a short and a long low front vowel sounding similar to the initial vowel of the English word “apple,” usually represented in IPA as U+00E6 æ (*ash*). The independent forms of these vowels are encoded at U+0D87 and U+0D88; the corresponding dependent forms are U+0DD0 and U+0DD1.

Because of these extra letters, the encoding for Sinhala does not precisely follow the pattern established for the other Indic scripts (for example, Devanagari), but does use the same general structure, making use of phonetic order, matra reordering, and use of the virama (U+0DCA SINHALA SIGN AL-LAKUNA) to indicate conjunct consonant clusters. Sinhala does not use half-forms in the Devanagari manner, but does use many ligatures.

Other Letters for Tamil. The Sinhala script may also be used to write Tamil. In this case, some additional combinations may be required. Some letters, such as U+0DBB SINHALA LETTER RAYANNA and U+0DB1 SINHALA LETTER DANTAJA NAYANNA, may be modified by adding the equivalent of a nukta. There is, however, no nukta presently encoded in the Sinhala block.

Historical Symbols. Neither U+0DF4 SINHALA PUNCTUATION KUNDDALIYA nor the Sinhala numerals are in general use today, having been replaced by Western-style punctuation and Western digits. The *kunddaliya* was formerly used as a full stop or period. It is included for scholarly use. The Sinhala numerals are not presently encoded.

9.11 Tibetan

Tibetan: U+0F00–U+0FFF

The Tibetan script is used for writing Tibetan in several countries and regions throughout the Himalayas. Aside from Tibet itself, the script is used in Ladakh, Nepal, and northern areas of India bordering Tibet where large Tibetan-speaking populations now reside. The Tibetan script is also used in Bhutan to write Dzongkha, the official language of that country. In addition, Tibetan is used as the language of philosophy and liturgy by Buddhist traditions spread from Tibet into the Mongolian cultural area that encompasses Mongolia, Buriatia, Kalmykia, and Tuva.

The Tibetan scripting and grammatical systems were originally defined together in the sixth century by royal decree when the Tibetan King Songtsen Gampo sent 16 men to India to study Indian languages. One of those men, Thumi Sambhota, is credited with creating the Tibetan writing system upon his return, having studied various Indic scripts and grammars. The king's primary purpose was to bring Buddhism from India to Tibet. The new script system was therefore designed with compatibility extensions for Indic (principally Sanskrit) transliteration so that Buddhist texts could be properly represented. Because of this origin, over the last 1,500 years the Tibetan script has also been widely used to represent Indic words, a number of which have been adopted into the Tibetan language retaining their original spelling.

A note on Latin transliteration: Tibetan spelling is traditional, and does not generally reflect modern pronunciation. Throughout this section, Tibetan words are represented in italics when transcribed as spoken, followed at first occurrence by a parenthetical transliteration; in these transliterations, presence of the *tsek* (tsheg) character is expressed with a hyphen.

Thumi Sambhota's original grammar treatise defined two script styles. The first, called *uchen* (dbu-can, "with head"), is a formal "inscriptional capitals" style said to be based on an old form of Devanagari. It is the script used in Tibetan xylograph books and the one used in the coding tables. The second style, called *u-mey* (dbu-med, or "headless"), is more cursive and said to be based on the Warty script. Numerous styles of *u-mey* have evolved since then, including both formal calligraphic styles used in manuscripts and running handwriting styles. All Tibetan scripts follow the same lettering rules, though there is a slight difference in the way that certain compound stacks are formed in *uchen* and *u-mey*.

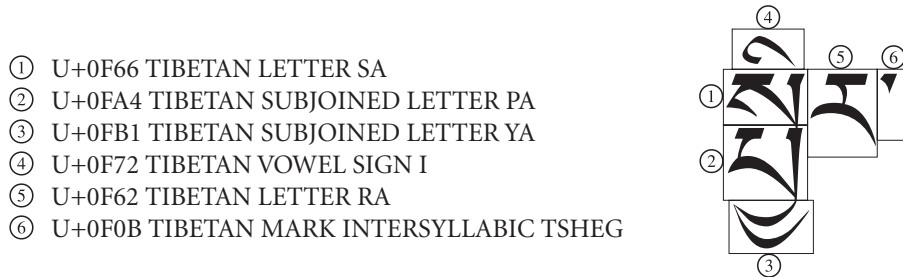
General Principles of the Tibetan Script. Tibetan grammar divides letters into consonants and vowels. There are 30 consonants, and each consonant is represented by a discrete written character. There are five vowel sounds, only four of which are represented by written marks. The four vowels that are explicitly represented in writing are each represented with a single mark that is applied above or below a consonant to indicate the application of that vowel to that consonant. The absence of one of the four marks implies that the first vowel sound (like a short "ah" in English) is present and is not modified to one of the four other possibilities. Three of the four marks are written above the consonants; one is written below.

Each word in Tibetan has a base or root consonant. The base consonant can be written singly or it can have other consonants added above or below it to make a vertically "stacked" letter. Tibetan grammar contains a very complete set of rules regarding letter gender, and these rules dictate which letters can be written in adjacent positions. The rules therefore dictate which combinations of consonants can be joined to make stacks. Any combination not allowed by the gender rules does not occur in native Tibetan words. However, when

transcribing other languages (for example, Sanskrit, Chinese) into Tibetan, these rules do not operate. In certain instances, other than transliteration, any consonant may be combined with any other subjoined consonant. Implementations should therefore be prepared to accept and display any combinations.

For example, the syllable *spyir* “general,” pronounced [tʃi:], is a typical example of a Tibetan syllable that includes a stack comprising a head letter, two subscript letters, and a vowel sign. *Figure 9-12* shows the characters in the order in which they appear in the backing store.

Figure 9-12. Tibetan Syllable Structure



The model adopted to encode the Tibetan lettering set described above contains the following groups of items: Tibetan consonants, vowels, numerals, punctuation, ornamental signs and marks, and Tibetan-transliterated Sanskrit consonants and vowels. Each of these will be described below.

Both in this description and in Tibetan, the terms “subjoined” (-btags) and “head” (-mgo) are used in different senses. In the structural sense, they indicate specific slots defined in native Tibetan orthography. In spatial terms, they refer to the position in the stack; anything in the topmost position is “head,” anything not in the topmost position is “subjoined.” Unless explicitly qualified, the terms “subjoined” and “head” are used here in their spatial sense. For example, in a conjunct like “rka” the letter in the root slot is “KA,” but, because it is not the topmost letter of the stack, it is expressed with a subjoined character code, while “RA,” which is structurally in the head slot, is expressed with a nominal character code. On the other hand, in a conjunct “kra” in which the root slot is also occupied with “KA,” the “KA” is encoded with a nominal character code, because it is in the topmost position in the stack.

The Tibetan script has its own system of formatting and details of that system relevant to the characters encoded in this standard are explained herein. However, an increasing number of publications in Tibetan do not strictly adhere to this original formatting system. This change is due to the partial move from publishing on long, horizontal, loose-leaf folios, to publishing in vertically oriented, bound books. The Tibetan script also has a punctuation set designed to meet needs quite different from the punctuation that has evolved for Western scripts. With the appearance of Tibetan newspapers, magazines, school textbooks, and Western-style reference books in the last 20 or 30 years, Tibetans have begun using things like columns, indented blocks of text, Western-style headings, and footnotes. Some Western punctuation marks, including brackets, parentheses, and quotation marks, are also becoming commonplace in these kinds of publication. With the introduction of more sophisticated electronic publishing systems, there is also a renaissance in the publication of voluminous religious and philosophical works in the traditional horizontal, loose-leaf format—many set in digital typefaces closely conforming to the proportions of traditional hand-lettered text.

Consonants. The system devised to encode the Tibetan system of writing consonants in both single and stacked forms as described above is as follows:

All of the consonants are encoded a first time from U+0F40 through U+0F69. There are the basic Tibetan consonants and, in addition, six compound consonants used to represent the Indic consonants *gha*, *jha*, *d.ha*, *dha*, *bha*, and *ksh.a*. These codes are used to represent occurrences of either a stand-alone consonant or a consonant in the head position of a vertical stack. Glyphs generated from these codes will always sit in the normal position starting at and dropping down from the design baseline. All of the consonants are then encoded a second time. These second encodings from U+0F90 through U+0FB9 represent consonants in subjoined stack position.

To represent a single consonant in a text stream, one of the first, “nominal,” set of codes is placed. To represent a stack of consonants in the text stream, a “nominal” consonant code is followed directly by one or more of the subjoined consonant codes. The stack so formed continues for as long as subjoined consonant codes are contiguously placed.

This encoding method was chosen over an alternative method that would have involved a virama-based encoding, like Devanagari. There were two main reasons for this choice. First, the virama is not normally used in the Tibetan writing system to create letter combinations. There is a virama in the Tibetan script, but only because of the need to represent Devanagari; it is called “srog-med” and encoded at U+0F84 TIBETAN MARK HALANTA. The virama is never used in writing Tibetan words and can be—but is almost never—used as a substitute for stacking in writing Sanskrit mantras in the Tibetan script. Second, there is a prevalence of stacking in native Tibetan, and the model chosen specifically results in decreased data storage requirements. Furthermore, in languages other than Tibetan, there are many cases where stacks occur that do not occur in Tibetan-language texts; it is thus imperative to have a model that allows for any consonant to be stacked with any subjoined consonant(s). Thus a model for stack building was chosen that follows the Tibetan approach to creating letter combinations, but is not limited to a specific set of the possible combinations.

Vowels. Each of the four basic Tibetan vowel marks mentioned above is coded as a separate entity. They are U+0F72, U+0F74, U+0F7A, and U+0F7C. For compatibility, a set of several compound vowels for Sanskrit transcription is also provided in the other code points between U+0F71 and U+0F7D. Most Tibetan users do not view these compound vowels as single characters, and their use is limited to Sanskrit words. It is acceptable for users to enter these compounds as a series of simpler elements and have software render them appropriately. Canonical equivalences are specified for all except U+0F77 and U+0F79. All vowel signs are nonspacing marks above or below a stack of consonants, sometimes on both sides.

A stand-alone consonant or a stack of consonants can have a vowel sign applied to it. In accordance with the rules of Tibetan writing, a code for a vowel sign applied to a consonant should always be placed after the bare consonant or the stack of consonants formed by the method just described.

All of the symbols and punctuation marks have straightforward encodings. Further information about many of them is included below.

Coding Order. In general, the correct coding order for a stream of text will be the same as the order in which Tibetans spell and in which the characters of the text would be written by hand. For example, the correct coding order for the most complex Tibetan stack would be:

head position consonant
 first subjoined consonant
 ... (intermediate subjoined consonants, if any)
 last subjoined consonant
 subjoined vowel a-chung (U+0F71)
 standard or compound vowel sign, or virama

Where used, the character U+0F39 TIBETAN MARK TSA -PHRU occurs immediately after the consonant it modifies.

Allographical Considerations. When consonants are combined to form a stack, one of them retains the status of being the principal consonant in the stack. The principal consonant always retains its stand-alone form. However, consonants placed in the “head” and “subjoined” positions to the main consonant sometimes retain their stand-alone form and sometimes are given a new, special form. Because of this fact, certain of the consonants are given a further, special encoding treatment. The affected consonants are “wa” (U+0F5D), “ya” (U+0F61), and “ra” (U+0F62).

Head Position “ra”. When the consonant “ra” is written in the “head” position (ra-mgo, pronounced *ra-go*) at the top of a stack in the normal Tibetan-defined lettering set, the shape of the consonant can change. This is called *ra-go* (ra-mgo). It can either be a full-form shape or the full-form shape but with the bottom stroke removed (looking like a short-stemmed letter “T”). This requirement of “ra” in the head position where the glyph representing it can change shape is correctly coded by using the stand-alone “ra” consonant (U+0F62) followed by the appropriate subjoined consonant(s). For example, in the normal Tibetan ra-mgo combinations, the “ra” in the head position is mostly written as the half-ra but in the case of “ra + subjoined nya” must be written as the full-form “ra”. Thus the normal Tibetan ra-mgo combinations are correctly encoded with the normal “ra” consonant (U+0F62) because it can change shape as required. It is the responsibility of the font developer to provide the correct glyphs for representing the characters where the “ra” in the head position will change shape—for example, as in “ra + subjoined nya”.

Full-Form “ra” in Head Position. Some instances of “ra” in the head position require that the consonant be represented as a full-form “ra” that never changes. This is *not* standard usage for the Tibetan language itself, but occurs in transliteration and transcription. Only in these cases should the character U+0F6A TIBETAN LETTER FIXED FORM-RA be used instead of U+0F62 TIBETAN LETTER RA. This “ra” will always be represented as a full-form “ra consonant” and will never change shape to the form where the lower stroke has been cut off. For example, the letter combination “ra + ya” when appearing in transliterated Sanskrit works is correctly written with a full-form “ra” followed by either a modified subjoined “ya” form or a full-form subjoined “ya” form. Note that the fixed-form “ra” should be used *only* in combinations where “ra” would normally transform into a short form but the user specifically wants to prevent that change. For example, the combination “ra + subjoined nya” never requires the use of fixed-form “ra”, because “ra” normally retains its full glyph form over “nya”. It is the responsibility of the font developer to provide the appropriate glyphs to represent the encodings.

Subjoined Position “wa”, “ya”, and “ra”. All three of these consonants can be written in subjoined position to the main consonant according to normal Tibetan grammar. In this position, *all* of them change to a new shape. The “wa” consonant when written in subjoined position is not a full “wa” letter any longer but is literally the bottom-right corner of the “wa” letter cut off and appended below. For that reason it is called a *wazur* (wa-zur, or “corner of a wa”) or less frequently, but just as validly, *wa-ta* (wa-btags) to indicate that it is a subjoined “wa”. The consonants “ya” and “ra” when in the subjoined position are called *ya-ta* (ya-btags) and *ra-ta* (ra-btags), respectively. To encode these subjoined consonants

that follow the rules of normal Tibetan grammar, the shape-changed, subjoined forms U+0F5D TIBETAN LETTER WA, U+0F61 TIBETAN LETTER YA, and U+0F62 TIBETAN LETTER RA should be used.

All three of these subjoined consonants also have full-form non-shape-changing counterparts for the needs of transliterated and transcribed text. For this purpose, the full subjoined consonants that do not change shape (encoded at U+0FBA, U+0FBB, and U+0FBC, respectively) are used where necessary. The combinations of “ra + ya” are a good example because they include instances of “ra” taking a short (ya-btags) form and “ra” taking a full-form subjoined “ya”.

U+0FB0 TIBETAN SUBJOINED LETTER -A (*a-chung*) should be used only in the very rare cases where a full-sized subjoined a-chung letter is required. The small vowel lengthening a-chung encoded as U+0F71 TIBETAN VOWEL SIGN AA is *far* more frequently used in Tibetan text, and it is therefore recommended that implementations treat this character rather than U+0FB0 as the normal subjoined a-chung.

Halanta (Srog-Med). Because there are two sets of consonants encoded for Tibetan, with the second set providing explicit ligature formation, there is no need for a “dead character” in Tibetan. When a *halanta* (srog-med) is used in Tibetan, its purpose is to suppress the inherent vowel “a”. If anything, the *halanta* should *prevent* any vowel or consonant from forming a ligature with the consonant preceding the *halanta*. In Tibetan text, this character should be displayed beneath the base character as a combining glyph and not used as a (purposeless) dead character.

Line Breaking Considerations. Tibetan text separates units called natively *tsek-bar* (“tsheg-bar”), an inexact translation of which is “syllable.” *Tsek-bar* is literally the unit of text between *tseks*, and is generally a consonant cluster with all of its prefixes, suffixes, and vowel signs. It is not a “syllable” in the English sense.

Tibetan script has two break characters only. The primary break character is the standard interword *tsek* (tsheg), which is encoded at U+0F0B. The other break character is the space. Space or *tsek* characters in a stream of Tibetan text are not always break characters and so need proper contextual handling. Issues surrounding these two potential break characters will now be discussed.

The primary delimiter character in Tibetan text is the *tsek* (U+0F0B TIBETAN MARK INTERSYLLABIC TSHEG). In general, automatic line breaking processes may break after any occurrence of this *tsek*, except where it follows a U+0F44 TIBETAN LETTER NGA (with or without a vowel sign) and precedes a *shay* (U+0F0D), or where Tibetan grammatical rules do not permit a break. (Normally, *tsek* is not written before *shay* except after “nga”. This type of tsek-after-nga is called “nga-phye-tsheg”, and may be expressed by U+0F0B, or by the special character U+0F0C, a nonbreaking form of *tsek*.) The Unicode names for these two types of *tsek* are misnomers, retained for compatibility. The standard *tsek* U+0F0B TIBETAN MARK INTERSYLLABIC TSHEG is always required to be a potentially breaking character, whereas the “nga-phye-tsheg” is always required to be a nonbreaking *tsek*. U+0F0C TIBETAN MARK DELIMITER TSHEG BSTAR is specifically not a “delimiter” and is not for general use.

There are no other break characters in Tibetan text. Unlike English, Tibetan has no system for hyphenating or otherwise breaking a word within the group of letters making up the word. Tibetan text formatting does not allow text to be broken within a word.

Whitespace appears in Tibetan text, although it should be represented by U+00A0 NO-BREAK SPACE instead of U+0020 SPACE. Tibetan text breaks lines after *tsek* instead of at whitespace.

Complete Tibetan text formatting is best handled by a formatter in the application and not just by the code stream. If the interword and nonbreaking *tseks* are properly employed as breaking and nonbreaking characters, respectively, and if all spaces are nonbreaking spaces, then any application will still wrap lines correctly on that basis, even though the breaks might be sometimes inelegant.

Tibetan Punctuation. The punctuation apparatus of Tibetan is relatively limited. The principal punctuation characters are the *tsek* already mentioned, the *shay* (transliterated “shad”), which is a vertical stroke used to mark the end of a section of text, the space used sparingly as a space, and two of several variant forms of the *shay* that are used in specialized situations requiring a *shay*. There are also several other marks and signs but they are sparingly used.

The *shay* at U+0F0D marks the end of a piece of text called “tshig-grub”. The mode of marking bears no commonality with English phrases or sentences and should not be described as a delimiter of phrases. In Tibetan grammatical terms, a *shay* is used to mark the end of an expression (“brjod-pa”) and a complete expression. Two *shays* are used at the end of whole topics (“don-tshan”). Because some writers use the double *shay* with a different spacing than would be obtained by coding two adjacent occurrences of U+0F0D, the double *shay* has been coded at U+0F0E with the intent that it would have a larger spacing between component *shays* than if two *shays* were simply written together. However, most writers do not use an unusual spacing between the double *shay*, so the application should allow the user to write two U+0F0D codes one after the other. Additionally, font designers will have to decide whether to implement these *shays* with a larger than normal gap.

The U+0F11 *rin-chen-pung-shay* (rin-chen-spungs-shad) is a variant *shay* used in a specific “new-line” situation. Its use was not defined in the original grammars but Tibetan tradition gives it a highly defined use. The *drul-shay* (“sbrul-shad”) is likewise not defined by the original grammars but has a highly defined use; it is used for separating sections of meaning that are equivalent to topics (“don-tshan”) and subtopics. A *drul-shay* is usually surrounded on both sides by the equivalent of about three spaces (though there is no rule specified). Hard spaces will be needed for these because the *drul-shay* should not appear at the beginning of a new line and the whole structure of spacing-plus-*shay* should not be broken up, if possible.

Tibetan texts use a *yig-go* (“head mark,” *yig-mgo*) to indicate the beginning of the front of a folio, there being no other certain way, in the loose-leaf style of traditional Tibetan books, to tell which is the front of a page. The head mark can and does vary from text to text; there are many different ways to write it. The common type of head mark has been provided for with U+0F04 TIBETAN MARK INITIAL YIG MGO MDUN MA and its extension U+0F05 TIBETAN MARK CLOSING YIG MGO. An initial mark *yig-mgo* can be written alone or combined with as many as three closing marks following it. When the initial mark is written in combination with one or more closing marks, the individual parts of the whole must stay in proper registration with each other to appear authentic. Therefore, it is strongly recommended that font developers create precomposed ligature glyphs to represent the various combinations of these two characters. The less common head marks mainly appear in Nyingmapa and Bonpo literature. Three of these head marks have been provided for with U+0F01, U+0F02, and U+0F03; however, many others have not been encoded. Font developers will have to deal with the fact that many types of head marks in use in this literature have not been encoded, cannot be represented by a replacement that has been encoded, and will be required by some users.

Two characters, U+0F3C TIBETAN MARK ANG KHANG GYON and U+0F3D TIBETAN MARK ANG KHANG GYAS, are paired punctuation, typically used together forming a roof over one or more digits or words. In this case, kerning or special ligatures may be required for proper rendering. The right *ang khang* may also be used much as a single closing parenthesis is

used in forming lists; again, special kerning may be required for proper rendering. The marks U+0F3E TIBETAN SIGN YAR TSHES and U+0F3F TIBETAN SIGN MAR TSHES are paired signs used to combine with digits; special glyphs or compositional metrics are required for their use.

A set of frequently occurring astrological and religious signs specific to Tibetan is encoded between U+0FBE and U+0FCF.

U+0F34 means “et cetera” or “and so on” and is used after the first few *tsek-bar* of a recurring phrase. U+0FBE (often three times) indicates a refrain.

U+0F36 and U+0FBF are used to indicate where text should be inserted within other text or as references to footnotes or marginal notes.

Other Characters. The Wheel of Dharma, which occurs sometimes in Tibetan texts, is encoded in the Miscellaneous Symbols block at U+2638.

Left-facing and right-facing *swastika* symbols are likewise used. They are found among the Chinese ideographs at U+534D (“yung-drung-chi-khor”) and U+5350 (“yung-drung-nang-khor”).

The marks U+0F35 TIBETAN MARK NGAS BZUNG NYI ZLA and U+0F37 TIBETAN MARK NGAS BZUNG SGOR TAGS conceptually attach to a *tsek-bar* rather than to an individual character and function more like attributes than characters—like underlining to mark or emphasize text. In Tibetan interspersed commentaries, they may be used to tag the *tsek-bar* belonging to the root text that is being commented on. The same thing is often accomplished by setting the *tsek-bar* belonging to the root text in large type and the commentary in small type. Correct placement of these glyphs may be problematic. If they are treated as normal combining marks, they can be entered into the text following the vowel signs in a stack; if used, their presence will need to be accounted for by searching algorithms, and so forth.

Tibetan Half-Numbers. The half-number forms (U+0F2A..U+0F33) are peculiar to Tibetan, though other scripts (for example, Bengali) have similar fractional concepts. The value of each half-number is 0.5 less than the number within which it appears. These forms are used only in some traditional contexts and appear as the *last* digit of a multidigit number. The sequence of digits “U+0F24 U+0F2C” represents the number 42.5 or forty-two and one-half.

Tibetan Transliteration and Transcription of Other Languages. Tibetan traditions are in place for transliterating other languages. Most commonly, Sanskrit has been the language being transliterated, though Chinese has become more common in modern times. Additionally, Mongolian has a transliterated form. There are even some conventions for transliterating English. One feature of Tibetan script/grammar is that it allows for totally accurate transliteration of Sanskrit. The basic Tibetan letterforms and punctuation marks contain most of what is needed, although a few extra things are required. With these additions, Sanskrit can be transliterated perfectly into Tibetan, and the Tibetan transliteration can be rendered backward perfectly into Sanskrit with no ambiguities or difficulties.

The six Sanskrit retroflex letters are interleaved among the other consonants.

The compound Sanskrit consonants are not included in normal Tibetan. They could be made using the method described earlier for Tibetan stacked consonants, generally by subjoining “ha”. However, to maintain consistency in transliterated texts and for ease in transmission and searching, it is recommended that implementations of Sanskrit in the Tibetan script use the precomposed forms of aspirated letters (and U+0F69 “ka + reversed sha”) whenever possible, rather than implementing these consonants as completely decomposed stacks. Note that implementations must ensure that decomposed stacks and precomposed forms are interpreted equivalently (see *Section 3.7, Decomposition*). The compound consonants are explicitly coded as follows: U+0F93 TIBETAN SUBJOINED LETTER GHA, U+0F9D

TIBETAN SUBJOINED LETTER DDHA, U+0FA2 TIBETAN SUBJOINED LETTER DHA, U+0FA7 TIBETAN SUBJOINED LETTER BHA, U+0FAC TIBETAN SUBJOINED LETTER DZHA, and U+0FB9 TIBETAN SUBJOINED LETTER KSSA.

The vowel signs of Sanskrit not included in Tibetan are encoded with other vowel signs between U+0F70 and U+0F7D. U+0F7F TIBETAN SIGN RNAM BCAD (*nam chay*) is the visarga, and U+0F7E TIBETAN SIGN RJES SU NGA RO (*ngaro*) is the anusvara. See Section 9.1, *Devanagari*, for more information on these two characters.

The characters encoded in the range U+0F88..U+0F8B are used in transliterated text and are most commonly found in Kalachakra literature.

When the Tibetan script is used to transliterate Sanskrit, consonants are sometimes stacked in ways that are not allowed in native Tibetan stacks. Even complex forms of this stacking behavior are catered for properly by the method described earlier for coding Tibetan stacks.

Other Signs. U+0F09 TIBETAN MARK BSKUR YIG MGO is a list enumerator used at the start of administrative letters in Bhutan, as is the petition honorific U+0F0A TIBETAN MARK BKA- SHOG YIG MGO.

U+0F3A TIBETAN MARK GUG RTAGS GYON and U+0F3B TIBETAN MARK GUG RTAGS GYAS are paired punctuation marks (brackets).

The sign U+0F39 TIBETAN MARK TSA -PHRU (*tsa-’phru*, which is a lenition mark) is the ornamental flaglike mark that is an integral part of the three consonants U+0F59 TIBETAN LETTER TSA, U+0F5A TIBETAN LETTER TSHA, and U+0F5B TIBETAN LETTER DZA. Although those consonants are not decomposable, this mark has been abstracted and may by itself be applied to “pha” and other consonants to make new letters for use in transliteration and transcription of other languages. For example, in modern literary Tibetan, it is one of the ways used to transcribe the Chinese “fa” and “va” sounds not represented by the normal Tibetan consonants. *Tsa-’phru* is also used to represent *tsa*, *tsha*, or *dza* in abbreviations.

Traditional Text Formatting and Line Justification. Native Tibetan texts (“pecha”) are written and printed using a justification system that is, strictly speaking, right-ragged but with an attempt to right-justify. Each page has a margin. That margin is usually demarcated with visible border lines required of a pecha. In modern times, as Tibetan text is produced in Western-style books, the margin lines may be dropped and an invisible margin used. When writing the text within the margins, an attempt is made to have the lines of text justified up to the right margin. To do so, writers keep an eye on the overall line length as they fill lines with text and try manually to justify to the right margin. Even then, there is often a gap at the right margin that cannot be filled. If the gap is short, it will be left as is and the line will be said to be justified enough, even though by machine-justification standards the line is not truly flush on the right. If the gap is large, the intervening space will be filled with as many *tseks* as are required to justify the line. Again, the justification is not done perfectly in the way that English text might be perfectly right-justified; as long as the last *tsek* is more or less at the right margin, that will do. The net result is that of a right-justified, blocklike look to the text, but the actual lines are always a little right-ragged.

Justifying *tseks* are nearly always used to pad the end of a line when the preceding character is a *tsek*—in other words, when the end of a line arrives in the middle of tshig-grub (see the previous definition under “Tibetan Punctuation”). However, it is unusual for a line that ends at the end of a tshig-grub to have justifying *tseks* added to the *shay* at the end of the tshig-grub. That is, a sequence like that in the first line of Figure 9-13 is not usually padded as in the second line of Figure 9-13, though it is allowable. In this case, instead of justifying the line with *tseks*, the space between *shays* is enlarged and/or the whitespace following the final *shay* is usually left as is. Padding is *never* applied following an actual space character. For example, given the existence of a space after a *shay*, a line such as the third line of Figure 9-13 may not be written with the padding as shown because the final *shay* should

have a space after it, and padding is never applied after spaces. The same applies where the final *consonant* of a tshig-grub that ends a line is a “ka” or “ga”. In that case, the ending *shay* is dropped but a space is still required after the consonant and that space must not be padded. For example, the appearance shown in the fourth line of *Figure 9-13* is not acceptable.

Figure 9-13. Justifying Tseks

```

1 འགྲུག།
2 འགྲུག།.....
3 འགྲུག། .....
4 འགྲུག .....

```

Tibetan text has two rules regarding the formatting of text at the beginning of a new line. There are severe constraints on which characters can start a new line, and the rule is traditionally stated as follows: A *shay* of any description may never start a new line. Nothing but actual words of text can start a new line, with the only exception being a *go-yig* (yig-mgo) at the head of a front page or a *da-tshe* (zla-tshe, meaning “crescent moon”—for example, U+0F05) or one of its variations, which is effectively an “in-line” go-yig (yig-mgo), on any other line. One of two or three ornamental *shays* is also commonly used in short pieces of prose in place of the more formal *da-tshe*. This rule also means that a space may not start a new line in the flow of text. If there is a major break in a text, a new line might be indented.

A syllable (tsheg-bar) that comes at the end of a tshig-grub and that starts a new line must have the *shay* that would normally follow it replaced by a rin-chen-spungs-shad (U+0F11). (The reason for this rule is that the presence of the rin-chen-spungs-shad makes the end of tshig-grub more visible and hence makes the text easier to read.)

In verse, the second *shay* following the first rin-chen-spungs-shad is also replaced sometimes with a rin-chen-spungs-shad, though the practice is formally incorrect. It is a writer’s trick done to make a particular scribing of a text more elegant. It is moderately popular device but does break the rule. Not only is rin-chen-spungs-shad used as the replacement for the *shay* but a whole class of “ornamental *shays*” are used for the same purpose. All are scribal variants on a rin-chen-spungs-shad, which is correctly written with three dots above it.

Tibetan Shorthand Abbreviations (bskungs-yig) and Limitations of the Encoding. A consonant functioning as the word-base (ming-gzhi) is allowed to take only one vowel sign according to Tibetan grammar. The Tibetan shorthand writing technique called bskungs-yig does allow one or more words to be contracted into a single, very unusual combination of consonants and vowels. This construction frequently entails the application of more than one vowel sign to a single consonant or stack, and the composition of the stacks themselves can break the rules of normal Tibetan grammar. For this reason, vowel signs do sometimes interact typographically, which accounts for their particular combining classes (see *Section 4.3, Combining Classes—Normative*).

The Unicode Standard accounts for plain text compounds of Tibetan that contain at most one base consonant, any number of subjoined consonants, followed by any number of vowel signs. This coverage constitutes the vast majority of Tibetan text. Rarely, stacks are seen that contain more than one such consonant-vowel combination in a vertical arrangement. These stacks are highly unusual and are considered beyond the scope of plain text rendering. They may be handled by higher-level mechanisms.

9.12 Limbu

Limbu: U+1900–U+194F

The Limbu script is a Brahmic script primarily used to write the Limbu language. Limbu is a Tibeto-Burman language of the East Himalayish group, spoken by about 200,000 persons mainly in eastern Nepal, but also in the neighboring Indian states of Sikkim and West Bengal (Darjeeling district). Its close relatives are the languages of the East Himalayish or “Kiranti” group in Eastern Nepal. It is distantly related to the Lepcha (Róng) language of Sikkim, and to Tibetan. Limbu was recognized as an official language in Sikkim in 1981.

The Nepali name *Limbu* is of uncertain origin. In Limbu, the Limbu call themselves *yak-thuy*. Individual Limbus often take the surname “Subba,” a Nepali term of Arabic origin meaning headman. The Limbu script is often called “Sirijanga” after the Limbu culture-hero Sirijanga, who is credited with its invention. It is also sometimes called Kirat, *kirāta* being a Sanskrit term probably referring to some variety of non-Aryan hill-dwellers.

The oldest known writings in the Limbu script, most of which are held in the India Office Library, London, were collected in Darjeeling district in the 1850s. The modern script was developed beginning in 1925 in Kalimpong (Darjeeling district) in an effort to revive writing in Limbu, which had fallen into disuse. The encoding in the Unicode Standard supports the three versions of the Limbu script: the nineteenth-century script, found in manuscript documents; the early modern script, used in a few, mainly mimeographed, publications between 1928 and the 1970s; and the current script, used in Nepal and India (especially Sikkim) since the 1970s. There are significant differences, particularly between some of the glyphs required for the nineteenth-century and modern scripts.

Virtually all Limbu speakers are bilingual in Nepali, and far more Limbus are literate in Nepali than in Limbu. For this reason, many Limbu publications contain material both in Nepali and in Limbu, and in some cases Limbu appears in both the Limbu script and the Devanagari script. In some publications, literary coinages are glossed in Nepali or in English.

Consonants. Consonant letters and clusters represent syllable initial consonants and clusters followed by the inherent vowel, short open o ([ɔ]). Subjoined consonant letters are joined to the bottom of the consonant letters, extending to the right to indicate “medials” in syllable-initial consonant clusters. There are very few of these clusters in native Limbu words. The script provides for subjoined ு -ya, ூ -ra, and ௃ -wa. Small letters are used to indicate syllable-final consonants. (See the following information on vowel length for further details.) The small letter consonants are found in the range U+1930..U+1938, corresponding to the syllable finals of native Limbu words. These letters are independent forms that, unlike the conjoined or half-letter forms of Indian scripts, may appear alone as word-final consonants (where Indian scripts use full consonant letters and a virama). The syllable finals are pronounced without a following vowel.

Limbu is a language with a well-defined syllable structure, in which syllable-initial stops are pronounced differently from finals. Syllable initials may be voiced following a vowel, whereas finals are never voiced but are pronounced unreleased with a simultaneous glottal closure, and geminated before a vowel. Therefore, the Limbu block encodes an explicit set of ten syllable-final consonants. These are called LIMBU SMALL LETTER KA, and so on.

Vowels. The Limbu vowel system has seven phonologically distinct timbres, [i, e, ε, a, ɔ, o, u]. The vowel [ɔ] functions as the inherent vowel in the modern Limbu script. To indicate a syllable with a vowel other than the inherent vowel, a *vowel sign* is added over, under, or to

the right of the initial consonant letter or cluster. Although the vowel [ɔ] is the inherent vowel, the Limbu script has a combining vowel sign 𑄛 that may optionally be used to represent it. Many writers avoid using this sign because they consider it redundant.

Syllable-initial vowels are represented by a vowel-carrier character, U+1900 𑄀 LIMBU VOWEL-CARRIER LETTER, together with the appropriate vowel sign. Used without a following vowel sound, the vowel-carrier letter represents syllable-initial [ɔ], the inherent vowel. The initial consonant letters have been named *ka*, *kha*, and so on, in this encoding, although they are in fact pronounced 𑄀 [kɔ], 𑄁 [kʰɔ], and so on, and do not represent the Limbu syllables 𑄀 [ka], 𑄁 [kʰa], and so on. This is in keeping with the practice of educated Limbus in writing the letter-names in Devanagari. It would have been confusing to call the vowel-carrier letter A, however, so an artificial name is used in the Unicode Standard. The native name is 𑄀𑄀 [ɔm].

Vowel Length. Vowel length is phonologically distinctive in many contexts. Length in open syllables is indicated by writing U+193A 𑄛 LIMBU SIGN KEMPHRENG, which looks like the diaeresis sign, over the initial consonant or cluster: 𑄛 𑄀 *tā*.

In closed syllables, there are two different methods of indicating vowel length. In the first method, vowel length is not indicated by *kemphreng*. The syllable-final consonant is written as a full form (that is, like a syllable-initial), marked by U+193B 𑄛 LIMBU SIGN SA-I: 𑄛𑄀 *pān* “speech.” This sign thus marks vowel length, in addition to functioning as a virama by suppressing the inherent vowel of the syllable-final consonant. This method is widely used in Sikkim.

In the second method in use in Nepal, vowel length is indicated by *kemphreng*, as for open syllables, and the syllable-final consonant appears in “small” form without *sa-i*: 𑄛𑄀 *pān* “speech.” Writers who consistently follow this practice reserve the use of *sa-i* for syllable-final consonants that do not have small forms, regardless of the length of the syllable vowel: 𑄛𑄀 *nesse* “it lay,” 𑄛𑄀 *lāb* “moon.” Because almost all of the syllable finals that normally occur in native Limbu words have small forms, *sa-i* is used only for consonant combinations in loan words, and for some indications of rapid speech.

U+193B 𑄛 LIMBU SIGN SA-I is based on the Indic virama, but for a majority of current writers it has a different semantics because it indicates the length of the preceding vowel in addition to “killing” the inherent vowel of consonants functioning as syllable finals. It is therefore not suitable for use as a general virama as used in other Brahmic scripts in the Unicode Standard.

Glottalization. U+1939 LIMBU SIGN MUKPHRENG represents glottalization. *Mukphreng* never appears as a syllable-initial. Although some linguists consider that word-final nasal consonants may be glottalized, this is never indicated in the script; *mukphreng* is not currently written after final consonants. No other syllable-final consonant clusters occur in Limbu.

Collating Order. There is no universally accepted alphabetical order for Limbu script. One ordering is based on the Limbu dictionary edited by Bairagi Kainla, with the addition of the obsolete letters, whose position is not problematic. In Sikkim, a somewhat different order is used: the letter 𑄀 *na* is placed before 𑄀 *ta*, and the letter 𑄀 *gha* is placed at the end of the alphabet.

Glyph Placement. The glyph positions for Limbu combining characters are summarized in Table 9-13.

Punctuation. The main punctuation mark used is the double vertical line, U+0965 DEVANAGARI DOUBLE DANDA. U+1945 𑄛 LIMBU QUESTION MARK and U+1944 𑄛 LIMBU EXCLAMATION MARK have shapes peculiar to Limbu, especially in Sikkimese typography. They are encoded in the Unicode Standard to facilitate the use of both Limbu and Devana-

Table 9-13. Positions of Limbu Combining Marks

Syllable	Glyphs	Code Point Sequence
<i>ta</i>	ᱠ	U+190B U+1920
<i>ti</i>	ᱡ	U+190B U+1921
<i>tu</i>	ᱢ	U+190B U+1922
<i>tee</i>	ᱣ	U+190B U+1923
<i>tai</i>	ᱤ	U+190B U+1924
<i>too</i>	ᱥ	U+190B U+1925
<i>tau</i>	ᱦ	U+190B U+1926
<i>te</i>	ᱧ	U+190B U+1927
<i>to</i>	ᱨ	U+190B U+1928
<i>tya</i>	ᱩ	U+190B U+1929
<i>tra</i>	ᱪ	U+190B U+192A
<i>twa</i>	ᱫ	U+190B U+192B
<i>tak</i>	ᱬ	U+190B U+1930
<i>taŋ</i>	ᱭ	U+190B U+1931
<i>taŋh</i>	ᱮ	U+190B U+1932
<i>tat</i>	ᱯ	U+190B U+1933
<i>tan</i>	ᱰ	U+190B U+1934
<i>tap</i>	ᱱ	U+190B U+1935
<i>tam</i>	ᱲ	U+190B U+1936
<i>tar</i>	ᱳ	U+190B U+1937
<i>tal</i>	ᱴ	U+190B U+1938
<i>tā</i>	ᱵ	U+190B U+1920 U+193A
<i>tī</i>	ᱶ	U+190B U+1921 U+193A

gari scripts in the same documents. U+1940 ᱮ LIMBU SIGN LOO is used for the exclamatory particle *lo*. This particle is also often simply spelled out ᱮᱱᱟ.

Digits. Limbu digits have distinctive forms and are assigned code points because Limbu and Devanagari (or Limbu and Arabic-Indic) numbers are often used in the same document.