Chapter 9

South and Southeast Asian Scripts

The following scripts are encoded in this group:

- Devanagari
- Bengali
- Gurmukhi
- Gujarati
- Oriya
- Tamil
- Telugu
- Kannada
- Malayalam
- Sinhala
- Thai
- Lao
- Tibetan
- Myanmar
- Khmer

The scripts of South and Southeast 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; many use no interword spacing but use spaces or marks between phrases. They are all essentially alphabetic scripts in which most symbols stand for a consonant plus an inherent vowel (usually the sound "ah"). Word-initial vowels in many of these scripts have special symbols, but 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.

Most of the scripts of South and Southeast Asia, from the Himalayas in the north to Sri Lanka in the south, from Pakistan in the west to the eastern-most islands of Indonesia, are derived from the ancient Brahmi script. The oldest lengthy inscriptions of India, the Asoka edicts of the third century BCE, were written in two scripts, Kharosthi 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. Kharosthi, written from right to left, was supplanted by Brahmi and its derivatives. The descendents 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 and the archetypal representative of the northern branch of the script family. This northern branch includes such modern scripts as Myanmar (Burmese), Gurmukhi, and Tibetan; the southern branch includes scripts such as Malayalam.

The major 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 upon the Indian national standard (ISCII) encoding for these scripts. Sinhala, Myanmar, and Khmer have virama-based models that do not map to ISCII. Thai and Lao share a different model, based on the Thai Industrial Standard encoding for Thai, which uses visual ordering and is not laid out compatibly to ISCII. Tibetan stands apart from all of these models, reflecting its somewhat different structure and usage. 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 Devanagari 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 (Thai, Lao, Khmer, and Myanmar), 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 Standard 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 position as those coded in positions A0–F4$_{16}$ in the ISCII-1988 standard. The same character code layout is followed for eight other Indic scripts in the Unicode Standard: Bengali, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam. This parallel code layout emphasizes the structural similarities of the Brahmi scripts and follows the stated intention of the Indian coding standards to enable one-to-one mappings between analogous coding positions in different scripts in the family. Sinhala, Thai, Lao, Khmer, and Myanmar 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 values and back to their original encoding without loss of information.

Encoding Principles. The writing systems that employ Devanagari and other Indic scripts constitute a cross between syllabic writing systems and phonemic writing systems (alphabets). 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*)*CV*. 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. For example, *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

/a/ + Dependent Vowel		Independent Vowel
$A_n + I_{vs} \rightarrow I_{vs} + A_n$	≈	I_n
अ + ◌ि → िअ	≈	इ
$A_n + U_{vs} \rightarrow A_n + U_{vs}$	≈	U_n
अ + ◌ु → अु	≈	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. Devanagari and other Indic scripts employ a sign known as the *virama*, *halant*, or vowel omission sign. A virama sign (for example, U+094D DEVANAGARI SIGN VIRAMA) nominally serves to cancel (or kill) the inherent vowel of the consonant to which it is applied. The virama functions as a combining character, with its shape varying from script to script. 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*.

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

Figure 9-2. Dead Consonants

$$\mathrm{TA}_n + \mathrm{VIRAMA}_n \rightarrow \mathrm{TA}_d$$

त + ◌् → त्

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) or an independent vowel letter.

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) $\mathrm{GA}_d + \mathrm{DHA}_l \rightarrow \mathrm{GA}_h + \mathrm{DHA}_n$

ग् + ध → ग्‍ध

(3) $\mathrm{KA}_d + \mathrm{SSHA}_l \rightarrow \mathrm{K.SSHA}_n$

क् + ष → क्ष

(2) $\mathrm{KA}_d + \mathrm{KA}_l \rightarrow \mathrm{K.KA}_n$

क् + क → क्क

(4) $\mathrm{RA}_d + \mathrm{RI}_n \rightarrow \mathrm{RI}_n + \mathrm{RA}_{sup}$

र् + ऋ → र्ऋ

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*; (3) a fully ligated conjunct *K.SSHA,* in which the components are no longer distinct; and (4) a rare conjunct formed with an independent vowel letter, in this case the vowel letter *RI* (also known as *vocalic r*). 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.

When an independent vowel appears as the terminal element of a consonant cluster, as in example (4) in *Figure 9-3*, the independent vowel should not be depicted as a dependent vowel sign, but as an independent vowel letterform.

Explicit Virama. 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, however, 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 ($\mathrm{K.SSHA}_n$).

Figure 9-4. Preventing Conjunct Forms

$$\mathrm{KA}_d + \mathrm{ZWNJ} + \mathrm{SSHA}_l \rightarrow \mathrm{KA}_d + \mathrm{SSHA}_n$$

क् + ZWNJ + ष → क्‌ष

Explicit Half-Consonants. When a dead consonant participates in forming a conjunct, the dead consonant form is often absorbed into the conjunct form, such that it is no longer distinctly visible. In other contexts, however, 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

$$\mathrm{KA}_d + \mathrm{ZWJ} + \mathrm{SSHA}_l \rightarrow \mathrm{KA}_h + \mathrm{SSHA}_n$$

क् + ZWJ + ष → क्‍ष

- In the absence of the zero width joiner, this sequence would normally produce the full conjunct form क्ष ($\mathrm{K.SSHA}_n$).

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*.

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 9-6. Independent Half-Forms

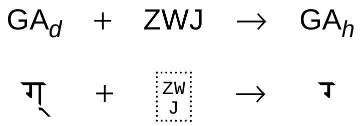
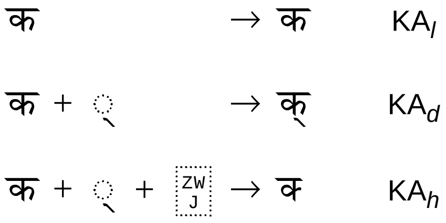


Figure 9-7. Consonant Forms



Rendering

Rules for Rendering. The following 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 set of glyphs is greater than the number of Devanagari Unicode 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 the U+0930 DEVANAGARI LETTER RA positioned above or attached to the upper part of a base glyph form. This form is also known as *repha*.

RA_{sub} A nonspacing combining mark glyph form of the 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 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.*

$$\mathrm{TA}_n + \mathrm{VIRAMA}_n \rightarrow \mathrm{TA}_d$$

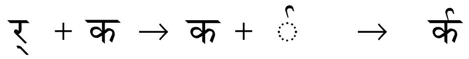
त + ◌् → त्

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

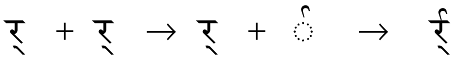
R2 *If the dead consonant RA_d precedes either a consonant or an independent vowel, then it is replaced by the superscript nonspacing mark RA_{sup}, which is positioned so that it applies to the logically subsequent element in the memory representation.*

$$\mathrm{RA}_d + \mathrm{KA}_l \rightarrow \mathrm{KA}_l + \mathrm{RA}_{sup}$$

Displayed Output



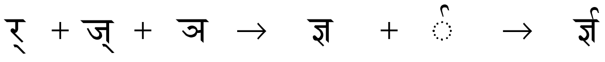
$$\overset{1}{\mathrm{RA}}_d + \overset{2}{\mathrm{RA}}_d \rightarrow \overset{2}{\mathrm{RA}}_d + \overset{1}{\mathrm{RA}}_{sup}$$



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.*

$$\mathrm{RA}_d + \mathrm{JA}_d + \mathrm{NYA}_l \rightarrow \mathrm{J.NYA}_n + \mathrm{RA}_{sup}$$

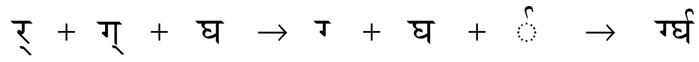
Displayed Output



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.*

$$\mathrm{RA}_d + \mathrm{GA}_d + \mathrm{GHA}_l \rightarrow \mathrm{GA}_h + \mathrm{GHA}_l + \mathrm{RA}_{sup}$$

Displayed Output



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.*

$$\mathrm{RRA}_n + \mathrm{VIRAMA}_n \rightarrow \mathrm{RRA}_h$$

ऱ़् + ◌् → ऱ्‍

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} .

$$\mathrm{RA}_d + \mathrm{ZWJ} \rightarrow \mathrm{RA}_h$$

र् + ZWJ → ऱ्‍

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 .

$$\mathrm{THA}_d + \mathrm{RA}_l \rightarrow \mathrm{THA}_n + \mathrm{RA}_{sub}$$

Displayed Output

ठ् + र → ठ + ◌्र → ठ्र

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

$$\mathrm{PHA}_d + \mathrm{RA}_l \rightarrow \mathrm{PHA}_n + \mathrm{RA}_{sub}$$

Displayed Output

फ् + र → फ + ◌्र → 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.*

$$\mathrm{TA}_d + \mathrm{RA}_d \rightarrow \mathrm{TA}_n + \mathrm{RA}_{sub} + \mathrm{VIRAMA}_n \rightarrow \mathrm{T.RA}_d$$

त् + र् → त + ◌्र + ◌् → त्र्

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

$$\mathrm{T.RA}_d + \mathrm{YA}_l \rightarrow \mathrm{T.R.YA}_n$$

त्र् + य → त्र्य

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

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

$$\mathrm{KA}_n + \mathrm{NUKTA}_n + \mathrm{VIRAMA}_n \rightarrow \mathrm{QA}_d$$

क + ◌़ + ◌् → क़्

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.*

$$\mathrm{KA}_n + \mathrm{AA}_{vs} + \mathrm{CANDRABINDU}_n$$

क + ◌ा + ◌ँ → काँ

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

$\mathrm{JA}_d + \mathrm{NYA}_l \rightarrow \mathrm{J.NYA}_n$ $\mathrm{TTA}_d + \mathrm{TTHA}_l \rightarrow \mathrm{TT.TTHA}_n$

ज् + ञ → ज्ञ ट् + ठ → ट्ठ

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

$$\mathrm{SA}_d + \mathrm{TA}_d + \mathrm{RA}_n \rightarrow \mathrm{SA}_d + \mathrm{T.R.A}_n \rightarrow \mathrm{S.T.RA}_n$$

स् + त् + र → स् + त्र → स्त्र

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

$$\mathrm{K.SSHA}_d + \mathrm{YA}_l \rightarrow \mathrm{K.SSH}_h + \mathrm{YA}_n$$

क्ष् + य → क्ष्य

R13 *If a nominal consonant or conjunct ligature form precedes* RA_{sub} *as a result of the application of rule R2, then the consonant or ligature form may join with* RA_{sub} *to form a multipart conjunct ligature (see rule R2 for more information).*

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

क + ◌्र → क्र फ + ◌्र → फ्र

R14 *In some cases, other combining marks will also 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} .

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

र + ◌ु → रु र + ◌ू → 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 with both the phonetic and keying order of textual data (see *Figure 9-8*).

Figure 9-8. Rendering Order

Character Order Glyph Order

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

क + ि◌ → कि

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

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

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

त् + र + ि◌ → त्र + ि◌ → 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	◌्	ZWJ	क्‍
ख	◌्	ZWJ	ख्‍
ग	◌्	ZWJ	ग्‍
घ	◌्	ZWJ	घ्‍
च	◌्	ZWJ	च्‍
ज	◌्	ZWJ	ज्‍
झ	◌्	ZWJ	झ्‍
ञ	◌्	ZWJ	ञ्‍
ण	◌्	ZWJ	ण्‍
त	◌्	ZWJ	त्‍
थ	◌्	ZWJ	थ्‍
ध	◌्	ZWJ	ध्‍

न	◌्	ZWJ	न्‍
प	◌्	ZWJ	प्‍
फ	◌्	ZWJ	फ्‍
ब	◌्	ZWJ	ब्‍
भ	◌्	ZWJ	भ्‍
म	◌्	ZWJ	म्‍
य	◌्	ZWJ	य्‍
ल	◌्	ZWJ	ल्‍
व	◌्	ZWJ	व्‍
श	◌्	ZWJ	श्‍
ष	◌्	ZWJ	ष्‍
स	◌्	ZWJ	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
ङ	◌्	क	ङ्क
ङ	◌्	ख	ङ्ख
ङ	◌्	ग	ङ्ग
ङ	◌्	घ	ङ्घ
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	ह्म
ह	◌्	य	ह्य
ह	◌्	ल	ह्ल

Table 9-2. Sample Ligatures (Continued)

द	◌्	ब	द्ब
द	◌्	भ	द्भ
द	◌्	म	द्म
द	◌्	य	द्य
द	◌्	व	द्व
ट	◌्	ट	ट्ट

ह	◌्	व	ह्व
ह		◌ृ	हृ
र		◌ु	रु
र		◌ू	रू
स	◌्	त्र	स्त्र

Sample Half-Ligature Forms. In addition to half-form glyphs of individual consonants, half-forms are also 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	◌्	ZWJ	क्ष्‍
ज	◌्	ञ	◌्	ZWJ	ज्ञ्‍
त	◌्	त	◌्	ZWJ	त्त्‍
त	◌्	र	◌्	ZWJ	त्र्‍
श	◌्	र	◌्	ZWJ	श्र्‍

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.

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.

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 West Bengal state (India) and in the nation of Bangladesh. It is also used to write Assamese in Assam (India) and a number of other minority languages (Daphla, Garo, Hallam, Khasi, Manipuri, Mizo, Munda, Naga, Rian, and Santali) in northeastern India.

Two-Part Vowel Signs. The Bengali script, along with a number of other Indic scripts, makes use of two-part vowel signs; in these vowels one-half of the vowel is placed on each side of a consonant letter or cluster—for example, U+09CB BENGALI VOWEL SIGN O and U+09CC BENGALI VOWEL SIGN AU. The vowel signs are coded in each case in the position in the charts isomorphic with the corresponding vowel in Devanagari. Hence U+09CC BENGALI VOWEL SIGN AU is isomorphic with U+094C DEVANAGARI VOWEL SIGN AU. To provide compatibility with existing implementations of the scripts that use two-part vowel signs, the Unicode Standard explicitly encodes the right half of these vowel signs; for example, U+09D7 BENGALI AU LENGTH MARK represents the right-half glyph component of U+09CC BENGALI VOWEL SIGN AU.

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

Rendering Behavior. For rendering of the Bengali script, see the rules for rendering in *Section 9.1, Devanagari.*

9.3 Gurmukhi

Gurmukhi: U+0A00–U+0A7F

The Gurmukhi script is a North Indian script historically derived from an older script called Lahnda. It is quite closely related to Devanagari structurally. Gurmukhi is used to write the Punjabi language in the Punjab in India.

Rendering Behavior. For rendering of the Gurmukhi script, see the rules for rendering in *Section 9.1, Devanagari*.

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.*

9.5 Oriya

Oriya: U+0B00–U+0B7F

The Oriya script is a North Indian script 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 Orissa state, 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*.

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. This difference is partly a result of the fact that the South Indian scripts were originally carved with needles on palm leaves, a technology that apparently favored rounded letter shapes rather than square, blocklike shapes.

The Tamil script is used to write the Tamil language of Tamil Nadu state in India as well as minority languages such as Badaga. Tamil is also used in Sri Lanka, Singapore, and parts of Malaysia. This script has fewer consonants than the other Indic scripts. It also lacks conjunct consonant forms. Instead of conjunct consonant forms, the virama (U+0BCD) is normally fully depicted in Tamil text.

Naming Conventions for Mid Vowels. The Unicode character encoding for Tamil uses a distinct set of naming conventions for mid vowels in the South Indian (Dravidian) scripts. These conventions are illustrated by U+0B8E TAMIL LETTER E and U+0B8F TAMIL LETTER EE, to be contrasted with the isomorphic positions in Devanagari: U+090E DEVANAGARI LETTER SHORT E and U+090F DEVANAGARI LETTER E. The Dravidian languages have a regular length distinction in the mid vowels that is not reflected in normal Devanagari. U+090E DEVANAGARI LETTER SHORT E is an addition to Devanagari to enable transcription of the Dravidian short vowel forms. The naming conventions are chosen to best reflect the actual nature of the vowels in question in the Dravidian scripts, as well as in Devanagari and the other North Indian scripts.

Special Characters. U+0BD7 TAMIL AU LENGTH MARK is provided as an encoding for the right side of the surroundrant (or two-part) vowel U+0BCC TAMIL VOWEL SIGN AU.

Rendering of Tamil Script. The South Indic scripts function in much the same way as Devanagari, with the additional feature of two-part vowels. As in the Devanagari example, the words "TAMIL LETTER" and "TAMIL VOWEL SIGN" will be omitted where this deletion does not cause ambiguity.

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

Table 9-4 is a summary of the Tamil letters.

Table 9-4. Tamil Letter Summary

க KA	ங NGA	ச CA	ஜ JA	ஞ NYA	ட TTA	ண NNA	த TA	ந NA	ன NNNA	ப PA	
ம MA	ய YA	ர RA	ற RRA	ல LA	ள LLA	ழ LLLA	வ VA	ஷ SSA	ஸ SA	ஹ HA	
அ A	ஆ AA	இ I	ஈ II	உ U	ஊ UU	எ E	ஏ EE	ஐ AI	ஒ O	ஓ OO	ஔ AU
A	ா AA	ி I	◌ீ II	◌ு U	◌ூ UU	ெ◌ E	ே◌ EE	ை◌ AI	ெ◌ா O	ே◌ா OO	ெ◌ள AU
◌் VIRAMA	ௗ AU LENGTH										

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

ன + ி ≠ ன + ◌் + இ

As in the case of Devanagari, a consonant cluster is any sequence of one or more consonants separated by viramas, possibly terminated with a virama.

Two-Part Vowels. Certain Indic vowels consist of two discontiguous elements. As in other cases of discontiguous elements, two sequences of Unicode values can be used to express equivalent spellings. This representation is similar to the case of letters such as “â”, which can be spelled either with “a” followed by a nonspacing “ˆ” or with a single Unicode character “â”.

ெ◌ா (0BCA) ≈ ெ◌+ ா (0BC6 + 0BBE)
ே◌ா (0BCB) ≈ ே◌+ ா (0BC7 + 0BBE)
ெ◌ள (0BCC) ≈ ெ◌+ ௗ (0BC7 + 0BD7)

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

If the precomposed forms are used in the memory representation instead of the separate characters, then a similar transformation occurs in the rendering process. The precomposed form on the left is transformed into the two separate forms equivalent to those on the right, which are then subject to vowel reordering, as shown below. Thus in rendering:

ெ◌ா → ெ◌ + ா
ே◌ா → ே◌ + ா
ெ◌ள → ெ◌ + ௗ

Vowel Reordering. As shown in *Table 9-5*, the following vowels are always reordered in front of the previous consonant cluster, similar to the rendering behavior of the DEVANAGARI VOWEL SIGN I:

◌ெ(0BC6) ◌ே(0BC7) ◌ை(0BC8)

Table 9-5. Vowel Reordering

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

The same effect occurs with the results of vowel splitting (see *Table 9-6*).

Table 9-6. Vowel Splitting and Reordering

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

In both cases, the ordering of the elements is *unambiguous*: the consonant (cluster) occurs *first* in the memory representation. The vowel ஔ also has two discontinuous parts and can be composed using the 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. Unlike Devanagari, Tamil includes very few conjunct consonants; most ligatures are located between a vowel and a neighboring consonant.

1. Conjunct consonants.

   க + ◌் + ஷ → க்ஷ

   As with Devanagari, vowel reordering occurs around conjunct consonants. For example:

   க + ◌் + ஷ + ◌ெ + ா → க்ஷொ

2. The vowel ா optionally ligates with ண, ன, or ற on its left:

ண + ா → ணா
ன + ா → னா
ற + ா → றா

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

Separate Vowels	*Precomposed Vowels*
ண + ெ◌ + ா → ணொ	ண + ◌ொ → ணொ
ண + ே◌ + ா → ணோ	ண + ◌ோ → ணோ
ன + ெ◌ + ா → னொ	ன + ◌ொ → னொ
ன + ே◌ + ா → னோ	ன + ◌ோ → னோ
ற + ெ◌ + ா → றொ	ற + ◌ொ → றொ
ற + ே◌ + ா → றோ	ற + ◌ோ → றோ

3. The vowel signs ி and ◌ீ form ligatures with ட on their left.

ட + ி → டி
ட + ◌ீ → டீ

These vowels often change shape or position slightly to link up with the appropriate shape of the consonant on their left:

ல + ி → லி
ல + ◌ீ → லீ

4. The vowel signs ◌ு and ◌ூ typically change form or ligate (see *Table 9-7*).

Table 9-7. Ligating Vowel Signs

x	*x* + ◌ு	*x* + ◌ூ
க	கு	கூ
ங	ஙு	ஙூ
ச	சு	சூ
ஞ	ஞு	ஞூ
ட	டு	டூ
ண	ணு	ணூ
த	து	தூ
ப	பு	பூ
ம	மு	மூ
ய	யு	யூ
ர	ரு	ரூ
ற	று	றூ
ல	லு	லூ
ள	ளு	ளூ

Table 9-7. Ligating Vowel Signs (Continued)

x	*x* + ◌ு	*x* + ◌ூ
ந	நு	நூ
ன	னு	னூ

x	*x* + ◌ு	*x* + ◌ூ
ழ	ழு	ழூ
வ	வு	வூ

To the right of ஜ, ஷ, ஸ, ஹ, or க்ஷ, these forms have a spacing form (see *Figure 9-9*).

Figure 9-9. Spacing Forms of Vowels

ஜ + ◌ு → ஜு

ஜ + ◌ூ → ஜூ

5. The vowel sign ை◌ changes to ை◌ to the left of ண, ன, ல, or ள.

ை◌ + ண → ணை

ை◌ + ன → னை

ை◌ + ல → லை

ை◌ + ள → ளை

Remember that this change takes place after the vowel reordering. In the first example, the vowel ை◌ 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 + ை◌ → ை◌ + ண → ணை

6. The consonant ர changes shape to ா.

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

ர + ◌் → ா + ◌்

ர + ி → ா + ி

ர + ◌ீ → ா + ◌ீ

9.7 Telugu

Telugu: U+0C00–U+0C7F

The Telugu script is a South Indian script used to write the Telugu language of Andhra Pradesh state in India, as well as minority languages such as Gondi (Adilabad and Koi dialects) and Lambadi.

Rendering Behavior. For rendering of the Telugu script, see the rules for rendering in *Section 9.6, Tamil*. Take note that, unlike Tamil, the Telugu script writes conjunct consonants with subscripted letters. There are also numerous consonant letters with contextual shape changes when used in conjuncts. Some vowel signs also change their shape in specified combinations.

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.

9.8 Kannada

Kannada: U+0C80–U+0CFF

The Kannada script is a South Indian script used to write the Kannada (or Kanarese) language of Karnataka state, as well as minority languages such as Tulu. It is very closely related to the Telugu script both with regard to the shapes of the letters and in the way that conjunct consonants behave.

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. 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 letters to the right of the consonant letter.

Kannada Letter LLLA. U+0CDE KANNADA LETTER FA is actually an obsolete Kannada letter that is transliterated in Dravidian scholarship as an "r" marked with the diacritic U+0324 COMBINING DIAERESIS BELOW. This form ought to have been given the letter name "LLLA", rather than "FA", so the Unicode name is simply a mistake. The letter in question 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.

9.9 Malayalam

Malayalam: U+0D00–U+0D7F

The Malayalam script is a South Indian script used to write the Malayalam language of Kerala state.

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

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.

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, formerly Ceylon. 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: අඬ[U+0D85 U+0DAC] / a.Nda/ “sound” versus අණ්ඩ [U+0D85 U+0DAB U+0DCA U+0DA9] /a.n.da/ “egg”. In addition, Sinhala has separate distinct signs for both a short and 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 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 the Sinhala numerals nor the punctuation sign Kunddaliya (U+0DF4) is in general use today, having been replaced by Western digits and Western-style punctuation. The Kunddaliya was formerly used as a full stop; it is included for scholarly use. The Sinhala numerals are not presently encoded.

9.11 Thai

Thai: U+0E00–U+0E7F

The Thai script is used to write Thai and other Southeast Asian languages, such as Kuy, Lavna, and Pali. It is a member of the Indic family of scripts descended from Brahmi. Thai modifies the original Brahmi letter shapes and extends the number of letters to accommodate features of the Thai language, including tone marks derived from superscript digits. On the other hand, Thai script lacks the conjunct consonant mechanism and independent vowel letters found in most other Bhrami-derived scripts. As in all scripts of this family, the predominant writing direction is left to right.

The Lao script is closely related to Thai, and the encoding principles described in this section apply to the Lao encoding as well.

Standards. Thai layout in the Unicode Standard is based on the Thai Industrial Standard 620-2529, and its updated version 620-2533.

Encoding Principles. In common with the Indic scripts, each Thai letter is a consonant possessing an inherent vowel sound. Thai letters further feature inherent tones. The inherent vowel and tone can be modified by means of vowel signs and tone marks attached to the base consonant letter. Some of the vowel signs and all of the tone marks are rendered in the script as diacritics attached above or below the base consonant. These combining signs and marks are encoded after the modified consonant in the memory representation.

Most of the Thai vowel signs are rendered by full letter-sized in-line glyphs placed either before (that is, to the left of) or after (to the right of) or *around* (on both sides of) the glyph for the base consonant letter. In the Thai encoding, the letter-sized glyphs that are placed before (left of) the base consonant letter, in full or partial representation of a vowel sign, are in fact encoded as separate characters that are typed and stored *before* the base consonant character. This encoding for left-side Thai vowel sign glyphs (and similarly in Lao) differs from the conventions for all other Indic scripts, which uniformly encode all vowels after the base consonant. The difference is necessitated by the encoding practice commonly employed with Thai character data as represented by the Thai Industrial Standard.

Thai Punctuation. Thai uses a variety of punctuation marks particular to this script. U+0E4F THAI CHARACTER FONGMAN is the Thai bullet, used to mark items in lists, or appearing at the beginning of a verse, sentence, paragraph, or other textual segment. U+0E46 THAI CHARACTER MAIYAMOK is used to mark repetition of preceding letters. U+0E2F THAI CHARACTER PAIYANNOI is used to indicate ellision or abbreviation of letters; it is itself viewed as a kind of letter, however, and is used with considerable frequency because of its appearance in such words as the Thai name for Bangkok. *Paiyannoi* is also used in combination (U+0E2F U+0E25 U+0E2F) to create a construct called *paiyanyai*, which means “et cetera, and so forth.” The Thai *paiyanyai* is comparable to the isomorphic analogue in the Khmer script: U+17D8 KHMER SIGN BEYYAL.

U+0E5A THAI CHARACTER ANGKHANKHU is used to mark the end of a long segment of text. It can be combined with a following U+0E30 THAI CHARACTER SARA A to mark a larger segment of text; typically this usage can be seen at the end of a verse in poetry. U+0E5B THAI CHARACTER KHOMUT marks the end of a chapter or document, where it always follows the *angkhankhu + sara a* combination. The Thai angkhankhu and its combination with sara a to mark breaks in text have analogues in many other Brahmi-derived scripts. For example, they are closely related to U+17D4 KHMER SIGN KHAN and U+17D5 KHMER SIGN

BARIYOOSAN, which are themselves ultimately related to the *danda* and *double danda* of Devanagari.

Thai words are not separated by spaces. Text is laid out with spaces introduced at text segments where Western typography would typically make use of commas or periods. However, Latin-based punctuation such as comma, period, and colon are also used in text, particularly in conjunction with Latin letters, or in formatting numbers, addresses, and so forth. If word boundary indications are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justified.

Thai Transcription of Pali and Sanskrit. The Thai script is frequently used to write Pali and Sanskrit. When so used, consonant clusters are represented by the explicit use of U+0E3A THAI CHARACTER PHINTHU (*virama*) to mark the removal of the inherent vowel. There is no conjoining behavior, unlike other Indic scripts. U+0E4D THAI CHARACTER NIKHAHIT is the Pali *nigghahita* and Sanskrit *anusvara*. U+0E30 THAI CHARACTER SARA A is the Sanskrit visarga. U+0E24 THAI CHARACTER RU and U+0E26 THAI CHARACTER LU are vocalic /r/ and /l/, with U+0E45 THAI CHARACTER LAKKHANGYAO used to indicate their lengthening.

9.12 Lao

Lao: U+0E80–U+0EFF

The Lao language and script are closely related to Thai. The Unicode Standard encodes the Lao script in the same relative order as Thai.

A few additional letters in Lao have no match in Thai:

U+0EBB LAO VOWEL SIGH MAI KON

U+0EBC LAO SEMIVOWEL SIGN LO

U+0EBD LAO SEMIVOWEL SIGN NYO

The preceding two semivowel signs are the last remnants of the system of subscript medials, which in Myanmar also includes original “rw”. Myanmar and Khmer include a full set of subscript consonant forms used for conjuncts. Thai no longer uses any of these forms; Lao has just the two.

There are also two ligatures in the Unicode character encoding for Lao: U+0EDC LAO HO NO and U+0EDD LAO HO MO. They correspond to sequences of [h] plus [n] or [h] plus [m] without ligating. Their function in Lao is to provide versions of the [n] and [m] consonants with a different inherent tonal implication.

9.13 Tibetan

Tibetan: U+0F00–U+0FBF

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. Tibetan is also 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 Devanagri. 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 Wartu 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.

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 topmost position is "head," anything not in 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. 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 codepoints 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) at the top of a stack in the normal Tibetan-defined lettering set, the shape of the consonant can change. 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 + sub-joined 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. In these cases, *only* the code U+0F6A TIBETAN LETTER FIXED FORM-RA should be used. 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.

Line-Breaking Considerations. Tibetan text separates units called natively “tsheg-bar”, an inexact translation of which is “syllable.” Tsheg-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 principle 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-spung-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 Tibetan tradition gives it 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-go 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 3.6, 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 tshig-grub to have justifying tseks added to the shay at the end of the tshig-grub. That is,

xxxx| |

is not usually padded like this (though it is allowable):

xxxx| |.....

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, the following may not be written with the padding,

xxxxxxx| ......

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 following is not acceptable:

xxxxxga ......

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, 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-spung-shad (U+0F11). (The reason for this rule is that the presence of the rin-chen-spung-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-spung-shad is also replaced sometimes with a rin-chen-spung-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 breaks the rule. Not only is rin-chen-spung-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-spung-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 4.2, 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 constitues 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.

9.14 Myanmar

Myanmar: U+1000–U+109F

The Myanmar script is used to write Burmese, the majority language of Myanmar (formerly called Burma). Variations and extensions of the script are used to write other languages of the region, such as Shan and Mon, as well as Pali and Sanskrit. The Myanmar script was formerly known as the Burmese script, but the term "Myanmar" is now preferred.

The Myanmar writing system derives from a Brahmi-related script borrowed from South India in about the eighth century for the Mon language. The first inscription in the Myanmar script dates from the eleventh century, using an alphabet almost identical to that of the Mon inscriptions. Aside from rounding of the originally square characters, this script has remained largely unchanged to the present. It is said that the rounder forms were developed to permit writing on palm leaves without tearing the writing surface of the leaf.

Because of its Brahmi origins, the Myanmar script shares the structural features of its Indic relatives: consonant symbols include an inherent "a" vowel; various signs are attached to a consonant to indicate a different vowel; ligatures and conjuncts are used to indicate consonant clusters; and the overall writing direction is left to right. Thus, despite great differences in appearance and detail, the Myanmar script follows the same basic principles as, for example, Devanagari.

Standards. There is not yet an official national standard for the encoding of Myanmar/Burmese. The current encoding was prepared with the consultation of experts from the Myanmar Information Technology Standardization Committee (MITSC) in Yangon (Rangoon). The MITSC, formed by the government in 1997, consists of experts from the Myanmar Computer Scientists' Association, Myanmar Language Commission, and Myanmar Historical Commission.

Encoding Principles. As with Indic scripts, the Myanmar encoding represents only the basic underlying characters; multiple glyphs and rendering transformations are required to assemble the final visual form for each syllable. Even some single characters, such as U+102C MYANMAR VOWEL SIGN AA, may assume variant forms depending on the other characters with which they combine. Conversely, characters or combinations that may appear visually identical in some fonts, such as U+101D MYANMAR LETTER WA and U+1040 MYANMAR DIGIT ZERO, are distinguished by their underlying encoding.

Composite Characters. As is the case in Extended Latin and many other scripts, some Myanmar letters or signs may be analyzed as composites of two or more other characters, and are not encoded separately. The following is an example of a Myanmar letter represented by a combining character sequence:

myanmar vowel sign o	= U+1031 MYANMAR VOWEL SIGN E
	+ U+102C MYANMAR VOWEL SIGN AA

Encoding Subranges. The basic consonants, independent vowels, and dependent vowel signs required for writing the Myanmar language are encoded at the beginning of the Myanmar range. Extensions of each of these categories for use in writing other languages, such as Pali and Sanskrit, are appended at the end of the range. In between these two sets lie the script-specific signs, punctuation, and digits.

Conjunct and Medial Consonants. As in other Indic-derived scripts, conjunction of two consonant letters is indicated by the insertion of a virama U+1039 MYANMAR SIGN VIRAMA between them; it causes ligation or other rendered combination of the consonants, although the virama itself is not rendered visibly.

The conjunct form of U+1004 MYANMAR LETTER NGA is rendered as a superscript sign called *kinzi*. Kinzi is encoded in logical order as a conjunct consonant *before* the syllable to which the mark applies, such as the Devanagari *ra*. (See *Section 9.1*, *Devanagari*, Rule R2.) For example, kinzi applied to U+1000 MYANMAR LETTER KA would be written via the following sequence:

U+1004 MYANMAR LETTER NGA

U+1039 MYANMAR SIGN VIRAMA

U+1000 MYANMAR LETTER KA

The Myanmar script traditionally distinguishes a set of subscript “medial” consonants: forms of YA, RA, WA, and HA that are considered to be modifiers of the syllable’s vowel. In the Myanmar encoding, the medial consonants are treated as conjuncts; that is, they are coded using the virama. So, for example, the word “kywe” (“to drop off”) would be written via the following sequence:

U+1000 MYANMAR LETTER KA

U+1039 MYANMAR SIGN VIRAMA

U+101A MYANMAR LETTER YA

U+1039 MYANMAR SIGN VIRAMA

U+101D MYANMAR LETTER WA

U+1031 MYANMAR VOWEL SIGN E

Explicit Virama. The virama U+1039 MYANMAR SIGN VIRAMA also participates in some common constructions where it appears as a *visible* sign, commonly termed *killer.* In this usage where it appears as a visible diacritic, U+1039 is followed by a U+200C ZERO WIDTH NON-JOINER, as with Devanagari (see *Figure 9-7*).

Signs After Consonants. Dependent vowels and other signs (except kinzi, as noted above) are encoded in logical order after the consonant to which they apply, regardless of where the glyph for the sign happens to be rendered relative to the glyph for the consonant. In particular, U+1031 MYANMAR VOWEL SIGN E is encoded *after* its consonant (as in the previous example), although in visual presentation it is reordered to appear *before* (to the left of) the consonant form.

Spacing. Myanmar does not use any whitespace between words. If word boundary indications are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justified.

9.15 Khmer

Khmer: U+1780–U+17FF

Khmer, also known as Cambodian, is the official language of Kampuchea. Mutually intelligible dialects are also spoken in northeastern Thailand and the Mekong Delta region of Vietnam. Although Khmer is not an Indo-European language, it has borrowed much vocabulary from Sanskrit and Pali, and religious texts in those languages have been transliterated as well as translated into Khmer.

The Khmer script, called *a'saa kmae* ("Khmer letters"), is descended from the Brahmi script of South India, as are Thai, Lao, Myanmar, Old Mon, and others. The exact sources have not been determined, but there is a great similarity between the earliest inscriptions in the region and the Pallawa script of the Coromandel coast of India. Modern Khmer has two basic styles of script: the *a'saa criang* ("slanted script") and the *a'saa muul* ("round script"). There is no fundamental structural difference between the two. The slanted script ("standing" variant) is chosen as representative in *Chapter 14, Code Charts*.

Structurally, the Khmer script stays close to its southern Brahmi origins; in modern terms, it shares features with both Thai and Myanmar. Consonant letters bear an inherent vowel sound, with additional signs placed before, above, below, and after the consonants to indicate vowels other than the inherent one. Consonant clusters are represented by conjuncts, where the first consonant of the cluster maintains its full form and succeeding consonants are written as subscripts. The overall writing direction is left to right.

The Khmer language has a richer set of vowels than the languages for which the ancestral script was used, but it has a smaller set of consonant sounds. The Khmer script takes advantage of this situation by assigning different symbols to the same consonant but with different inherent vowels. This usage is analogous to the situation in Thai, where different consonant symbols have the same sound but encode different tones. The Khmer consonant letters are organized into two series or registers, whose inherent vowels are nominally "a" and "o". Two "shifter" signs convert a consonant from one series to the other. The dependent vowel signs then do not have a single phonetic value, but rather are interpreted in the context of the consonant to which they are attached.

Encoding Principles. As with other related scripts, the Khmer encoding represents only the basic underlying characters; multiple glyphs and rendering transformations are required to assemble the final visual form for each syllable. Even some single characters, such as U+1789 KHMER LETTER NYO, may assume variant forms depending on the other characters with which they combine. Conversely, characters or combinations that may appear visually identical in some fonts, such as U+17A2 KHMER LETTER QA and U+17A3 KHMER INDEPENDENT VOWEL QAQ, are distinguished by their underlying encoding.

Independent Vowels. In Khmer, as in most related scripts (but in contrast to Thai), independent vowels have their own letterforms.

Conjunct Consonants. U+17D2 KHMER SIGN COENG plays the conjunct formation role of Indic virama, killing the vowel of its preceding consonant, and indicating that the following consonant should be treated as a subscript. *Sign coeng* should not be confused with U+17D1 KHMER SIGN VIRIAM, which has a name similar to virama but an unrelated function—indicating that the base character is part of the previous word.

Note that a subjoined U+179A KHMER LETTER RO is typed normally—that is, after the preceding consonant and *sign coeng*, even though the glyph for the subjoint *letter ro* is rendered before (to the left of) the full-size consonant. U+17CC KHMER SIGN ROBAT historically corresponds to the Devanagari *repha* (that is, an initial /r/), but it has lost this function in Khmer and instead is treated as a simple diacritical mark rather than as part of a conjunct.

Consonant Shifters. The marks U+17C9 KHMER SIGN MUUSIKATOAN and U+17CA KHMER SIGN TRIISAP are used to shift the base consonant between registers. In the presence of other superscript glyphs, both of these signs may be rendered via the same glyph shape as that of U+17BB KHMER VOWEL SIGN U. Selection of the proper rendering form is left to the display software.

Dependent Vowel Signs. Structurally, the Khmer dependent vowel signs are close to those of Thai, but the encoding principle is different. Khmer follows the general model for Brahmi-derived scripts, in which each dependent vowel sign is represented by a single code that occurs after the code for the base consonant. This principle reflects the fact that a vowel sign has its own identity, regardless of the number and placement of glyph fragments that may be used to render the sign, and regardless of its contextual phonetic interpretation. Each Khmer consonant either expresses its implicit vowel or is followed by a single dependent vowel sign character.

Ordering of Syllable Components. The standard order of components in consonantal syllables as expressed in BNF is

C (SC C)* { S } { V } { O }

where

C is a consonant

SC is a sign coeng (= virama)

S is a consonant shift

V is a dependent vowel sign

O is any other sign

For example, the word /khnyom/, “I”, would represented by the following sequence:

U+1781 KHMER LETTER KHA

U+17D2 KHMER SIGN COENG

U+1789 KHMER LETTER NYO

U+17B9 KHMER VOWEL SIGN Y

U+17C6 KHMER SIGN NIKAHIT

Except for the presence of consonant shifters, this structure is analogous to Devanagari (see *Section 9.1, Devanagari*).

Special Character Usage. Although U+17A3 INDEPENDENT VOWEL QAQ and U+17A2 KHMER LETTER QA have identical glyphs, only U+17A2 should be used for modern text. U+17A3 and U+17A4 INDEPENDENT VOWEL QAA are solely for Pali/Sanskrit transliteration.

U+1789 KHMER LETTER NYO may have two different subscript forms. Furthermore, the glyph part of the base letter U+1789, which normally occurs under the baseline, is omitted whenever a subscript is conjoined. U+17B1 KHMER INDEPENDENT VOWEL QOO TYPE ONE bears a close relationship to U+17B2 KHMER INDEPENDENT VOWEL QOO TYPE TWO. The very commonly used word meaning “give” should be spelled U+17B2 U+17D2 U+1799, with rendering software optionally replacing the glyph for U+17B1 with the glyph for U+17B2. U+17D3 KHMER SIGN BATHAMASAT is an exceedingly rare sign used in historic lunar dates; it should not be confused with U+17C6 KHMER SIGN NIKAHIT, which carries an “m” sound and occurs frequently in modern Khmer text. U+17C6 KHMER SIGN NIKAHIT, U+17C7 KHMER SIGN REAHMUK, and U+17C8 KHMER SIGN YUUKALEAPINTU are signs that occur in many permutations with vowels; they are not vowels themselves.

Spacing. Khmer does not use any whitespace between words. If word boundary indications are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justified.

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, 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 (see http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html). More recent versions of the Unicode standard exist (see http://www.unicode.org/unicode/standard/versions/).

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.

ISBN 0-201-61633-5

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: http://www.awl.com/cseng/

First printing, January 2000.