

The Unicode Standard

Version 6.1 – Core Specification

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

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

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

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

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

Copyright © 1991–2012 Unicode, Inc.

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

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.1.

Includes bibliographical references and index.

ISBN 978-1-936213-02-3 (<http://www.unicode.org/versions/Unicode6.1.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2012

ISBN 978-1-936213-02-3

Published in Mountain View, CA

April 2012

Chapter 10

South Asian Scripts-II

This chapter documents scripts of South Asia aside from the major official scripts of India, which are described in *Chapter 9, South Asian Scripts-I*.

The following South Asian scripts are described in this chapter:

| | | |
|---------------------|-------------------|---------------------|
| <i>Sinhala</i> | <i>Kaithi</i> | <i>Meetei Mayek</i> |
| <i>Tibetan</i> | <i>Saurashtra</i> | <i>Ol Chiki</i> |
| <i>Lepcha</i> | <i>Sharada</i> | <i>Sora Sompeng</i> |
| <i>Phags-pa</i> | <i>Takri</i> | <i>Kharoshthi</i> |
| <i>Limbu</i> | <i>Chakma</i> | <i>Brahmi</i> |
| <i>Syloti Nagri</i> | | |

Most of these scripts are historically related to the other scripts of India, and most are ultimately derived from the Brahmi script. None of them were standardized in ISCII. The encoding for each script is done on its own terms, and the blocks do not make use of a common pattern for the layout of code points.

This introduction briefly identifies each script, occasionally highlighting the most salient distinctive attributes of the script. Details are provided in the individual block descriptions that follow.

Sinhala is an official script of Sri Lanka, where it is used to write the majority language, also known as Sinhala.

The Tibetan script is used for writing the Tibetan language in several countries and regions throughout the Himalayas. The approach to the encoding of Tibetan in the Unicode Standard differs from that for most Brahmi-derived scripts. Instead of using a virama-based model for consonant conjuncts, it uses a subjoined consonant model.

Lepcha is the writing system for the Lepcha language, spoken in Sikkim and in the Darjeeling district of the West Bengal state of India. Lepcha is directly derived from the Tibetan script, but all of the letters were rotated by ninety degrees.

Phags-pa is a historical script related to Tibetan that was created as the national script of the Mongol empire. Even though Phags-pa was used mostly in Eastern and Central Asia for writing text in the Mongolian and Chinese languages, it is discussed in this chapter because of its close historical connection to the Tibetan script.

Limbu is a Brahmi-derived script primarily used to write the Limbu language, spoken mainly in eastern Nepal, Sikkim, and in the Darjeeling district of West Bengal. Its encoding follows a variant of the Tibetan model, making use of subjoined medial consonants, but also explicitly encoded syllable-final consonants.

Syloti Nagri is used to write the modern Sylheti language of northeast Bangladesh and southeast Assam in India.

Kaithi is a historic North Indian script, closely related to the Devanagari and Gujarati scripts. It was used in the area of the present-day states of Bihar and Uttar Pradesh in northern India, from the 16th century until the early 20th century.

Saurashtra is used to write the Saurashtra language, related to Gujarati, but spoken in southern India. The Saurashtra language is most often written using the Tamil script, instead.

Sharada is a historical script that was used to write Sanskrit, Kashmiri, and other languages of northern South Asia; it was the principal inscriptional and literary script of Kashmir from the 8th century CE until the 20th century. It has limited and specialized modern use.

Takri, descended from Sharada, is used in northern India and surrounding countries. It is the traditional writing system for the Chambeali and Dogri languages, as well as several “Pahari” languages. In addition to popular usage for commercial and informal purposes, Takri served as the official script of several princely states of northern and northwestern India from the 17th century until the middle of the 20th century. There are efforts to revive its use for Dogri and other languages.

Chakma is used to write the language of the Chakma people of southeastern Bangladesh and surrounding areas. The language, spoken by about half a million people, is related to other eastern Indo-European languages such as Bengali.

Meetei Mayek is used to write Meetei, a Tibeto-Burman language spoken primarily in Manipur, India. Like Limbu, it makes use of explicitly encoded syllable-final consonants.

Ol Chiki is an alphabetic script invented in the 20th century to write Santali, a Munda language of India. It is used primarily for the southern dialect of Santali spoken in the state of Orissa.

Sora Sompeng is used to write the Sora language spoken by the Sora people, who live in eastern India between the Oriya- and Telugu-speaking populations. The script was created in 1936 and is used in religious contexts.

The oldest lengthy inscriptions of India, the edicts of Ashoka from the third century BCE, were written in two scripts, Kharoshthi and Brahmi. These are both ultimately of Semitic origin, probably deriving from Aramaic, which was an important administrative language of the Middle East at that time. Kharoshthi, which was written from right to left, was supplanted by Brahmi and its derivatives.

10.1 Sinhala

***Sinhala:* U+0D80–U+0DFF**

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

Sinhala differs from other languages of the region in that it has a series of prenasalized stops that are distinguished from the combination of a nasal followed by a stop. In other words, both forms occur and are written differently—for example, අඳ <U+0D85, U+0DAC> *añḍa* [aⁿḍa] “sound” versus අඳ්ඳ <U+0D85, U+0DAB, U+0DCA, U+0DA9> *aṅḍa* [aṅḍa] “egg.” In addition, Sinhala has separate distinct signs for both a short and a long low front vowel sounding similar to the initial vowel of the English word “apple,” usually represented in IPA as U+00E6 æ LATIN SMALL LETTER AE (*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). It 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.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. Table 10-1 shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 10-1. Sinhala Vowel Letters

| To Represent | Use | Do Not Use |
|--------------|------|--------------|
| ඒ | 0D86 | <0D85, 0DCF> |
| ඒෆ | 0D87 | <0D85, 0DD0> |
| ඒෆ් | 0D88 | <0D85, 0DD1> |
| ඌ | 0D8C | <0D8B, 0DDF> |
| ඌෆෆ | 0D8E | <0D8D, 0DD8> |
| ඌෆ | 0D90 | <0D8F, 0DDF> |
| ඌෆ් | 0D92 | <0D91, 0DCA> |
| ඌෆ් | 0D93 | <0D91, 0DD9> |
| ඌෆ් | 0D96 | <0D94, 0DDF> |

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

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

10.2 Tibetan

Tibetan: U+0F00–U+0FFF

The Tibetan script is used for writing Tibetan in several countries and regions throughout the Himalayas. Aside from Tibet itself, the script is used in Ladakh, Nepal, and northern areas of India bordering Tibet where large Tibetan-speaking populations now reside. The Tibetan script is also used in Bhutan to write Dzongkha, the official language of that country. In Bhutan, as well as in some scholarly traditions, the Tibetan script is called the Bodhi script, and the particular version written in Bhutan is known as Joyi (mgyogs yig). In addition, Tibetan is used as the language of philosophy and liturgy by Buddhist traditions

spread from Tibet into the Mongolian cultural area that encompasses Mongolia, Buriatia, Kalmykia, and Tuva.

The Tibetan scripting and grammatical systems were originally defined together in the sixth century by royal decree when the Tibetan King Songtsen Gampo sent 16 men to India to study Indian languages. One of those men, Thumi Sambhota, is credited with creating the Tibetan writing system upon his return, having studied various Indic scripts and grammars. The king's primary purpose was to bring Buddhism from India to Tibet. The new script system was therefore designed with compatibility extensions for Indic (principally Sanskrit) transliteration so that Buddhist texts could be represented properly. Because of this origin, over the last 1,500 years the Tibetan script has 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, the presence of the *tsek* (tsheg) character is expressed with a hyphen.

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

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

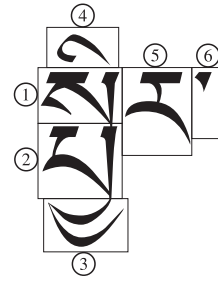
Each word in Tibetan has a base or root consonant. The base consonant can be written singly or it can have other consonants added above or below it to make a vertically "stacked" letter. Tibetan grammar contains a very complete set of rules regarding letter gender, and these rules dictate which letters can be written in adjacent positions. The rules therefore dictate which combinations of consonants can be joined to make stacks. Any combination not allowed by the gender rules does not occur in native Tibetan words. However, when transcribing other languages (for example, Sanskrit, Chinese) into Tibetan, these rules do not operate. In certain instances other than transliteration, any consonant may be combined with any other subjoined consonant. Implementations should therefore be prepared to accept and display any combinations.

For example, the syllable *spyir* "general," pronounced [tʃi:], is a typical example of a Tibetan syllable that includes a stack comprising a head letter, two subscript letters, and a vowel sign. *Figure 10-1* shows the characters in the order in which they appear in the backing store.

The model adopted to encode the Tibetan lettering set described above contains the following groups of items: Tibetan consonants, vowels, numerals, punctuation, ornamental signs

Figure 10-1. Tibetan Syllable Structure

- ① U+0F66 TIBETAN LETTER SA
- ② U+0FA4 TIBETAN SUBJOINED LETTER PA
- ③ U+0FB1 TIBETAN SUBJOINED LETTER YA
- ④ U+0F72 TIBETAN VOWEL SIGN I
- ⑤ U+0F62 TIBETAN LETTER RA
- ⑥ U+0F0B TIBETAN MARK INTERSYLLABIC TSHEG



and marks, and Tibetan-transliterated Sanskrit consonants and vowels. Each of these will be described in this section.

Both in this description and in Tibetan, the terms “subjoined” (-btags) and “head” (-mgo) are used in different senses. In the structural sense, they indicate specific slots defined in native Tibetan orthography. In spatial terms, they refer to the position in the stack; anything in the topmost position is “head,” anything not in the topmost position is “subjoined.” Unless explicitly qualified, the terms “subjoined” and “head” are used here in their spatial sense. For example, in a conjunct like “rka,” the letter in the root slot is “KA.” Because it is not the topmost letter of the stack, however, it is expressed with a subjoined character code, while “RA,” which is structurally in the head slot, is expressed with a nominal character code. 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 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 described here has been devised to encode the Tibetan system of writing consonants in both single and stacked forms.

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, such as 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; called “srog-med”, it is encoded at U+0F84 TIBETAN MARK HALANTA. The virama is never used in writing Tibetan words and can be—but almost never is—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 appear 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 is coded as a separate entity. These code points are U+0F72, U+0F74, U+0F7A, and U+0F7C. For compatibility, a set of several compound vowels for Sanskrit transcription is also provided in the other code points between U+0F71 and U+0F7D. Most Tibetan users do not view these compound vowels as single characters, and their use is limited to Sanskrit words. It is acceptable for users to enter these compounds as a series of simpler elements and have software render them appropriately. Canonical equivalences are specified for all of these compound vowels, with the exception of U+0F77 TIBETAN VOWEL SIGN VOCALIC RR and U+0F79 TIBETAN VOWEL SIGN VOCALIC LL, which for historic reasons have only compatibility equivalences specified. These last two characters are deprecated, and their use is strongly discouraged.

A vowel sign may be applied either to a stand-alone consonant or to a stack of consonants. The vowel sign occurs in logical order after the consonant (or stack of consonants). Each of the vowel signs is a nonspacing combining mark. The four basic vowel marks are rendered either above or below the consonant. The compound vowel marks also appear either above or below the consonant, but in some cases have one part displayed above and one part displayed below the consonant.

All of the symbols and punctuation marks have straightforward encodings. Further information about many of them appears later in this section.

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 forms and sometimes are given new, special forms. Because of this fact, certain consonants are given a further, special encoding treatment—namely, “wa” (U+0F5D), “ya” (U+0F61), and “ra” (U+0F62).

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

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

Subjoined Position “wa”, “ya”, and “ra”. All three of these consonants can be written in subjoined position to the main consonant according to normal Tibetan grammar. In this position, *all* of them change to a new shape. The “wa” consonant when written in subjoined position is not a full “wa” letter any longer but is literally the bottom-right corner of the “wa” letter cut off and appended below it. 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+0FAD TIBETAN SUBJOINED LETTER WA, U+0FB1 TIBETAN SUBJOINED LETTER YA, and U+0FB2 TIBETAN SUBJOINED LETTER RA should be used.

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

U+0FB0 TIBETAN SUBJOINED LETTER -A (*a-chung*) should be used only in the very rare cases where a full-sized subjoined a-chung letter is required. The small vowel lengthening a-chung encoded as U+0F71 TIBETAN VOWEL SIGN AA is *far* more frequently used in

Tibetan text, and it is therefore recommended that implementations treat this character (rather than U+0FB0) as the normal subjoined a-chung.

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

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

Tibetan script has only two break characters. The primary break character is the standard interword *tsek* (tsheg), which is encoded at U+0F0B. The second 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.

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

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

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

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

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

The *shay* at U+0F0D marks the end of a piece of text called “tshig-grub”. The mode of marking bears no commonality with English phrases or sentences and should not be described as a delimiter of phrases. In Tibetan grammatical terms, a *shay* is used to mark the end of an expression (“brjod-pa”) and a complete expression. Two *shays* are used at the

end of whole topics (“don-tshan”). Because some writers use the double *shay* with a different spacing than would be obtained by coding two adjacent occurrences of U+0F0D, the double *shay* has been coded at U+0F0E with the intent that it would have a larger spacing between component *shays* than if two *shays* were simply written together. However, most writers do not use an unusual spacing between the double *shay*, so the application should allow the user to write two U+0F0D codes one after the other. Additionally, font designers will have to decide whether to implement these *shays* with a larger than normal gap.

The U+0F11 *rin-chen-pung-shay* (rin-chen-spungs-shad) is a variant *shay* used in a specific “new-line” situation. Its use was not defined in the original grammars but Tibetan tradition gives it a highly defined use. The *drul-shay* (“sbrul-shad”) is likewise not defined by the original grammars but has a highly defined use; it is used for separating sections of meaning that are equivalent to topics (“don-tshan”) and subtopics. A *drul-shay* is usually surrounded on both sides by the equivalent of about three spaces (though no rule is specified). Hard spaces will be needed for these instances 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 SGAB MA. An initial mark *yig-mgo* can be written alone or combined with as many as three closing marks following it. When the initial mark is written in combination with one or more closing marks, the individual parts of the whole must stay in proper registration with each other to appear authentic. Therefore, it is strongly recommended that font developers create precomposed ligature glyphs to represent the various combinations of these two characters. The less common head marks mainly appear in Nyingmapa and Bonpo literature. Three of these head marks have been provided for with U+0F01, U+0F02, and U+0F03; however, many others have not been encoded. Font developers will have to deal with the fact that many types of head marks in use in this literature have not been encoded, cannot be represented by a replacement that has been encoded, and will be required by some users.

Two characters, U+0F3C TIBETAN MARK ANG KHANG GYON and U+0F3D TIBETAN MARK ANG KHANG GYAS, are paired punctuation; they are typically used together to form 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+0FB E and U+0FC F.

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

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

Svasti Signs. The *svasti* signs encoded in the range U+0FD5..U+0FD8 are widely used sacred symbols associated with Hinduism, Buddhism, and Jainism. They are often printed in religious texts, marriage invitations, and decorations, and are considered symbols of good luck and well-being. In the Hindu tradition in India, the dotted forms are often used.

The *svasti* signs are used to mark religious flags in Jainism and also appear on Buddhist temples, or as map symbols to indicate the location of Buddhist temples throughout Asia. These signs are encoded in the Tibetan block, but are intended for general use; they occur with many other scripts in Asia.

In the Tibetan language, the right-facing *svasti* sign is referred to as *gyung drung nang -khor* and the left-facing *svasti* sign as *gyung drung phyi -khor*. U+0FCC TIBETAN SYMBOL NOR BU BZHI -KHYIL, or quadruple body symbol, is a Tibetan-specific version of the left-facing *svasti* sign.

The *svasti* signs have also been borrowed into the Han script and adapted as CJK ideographs. The CJK unified ideographs U+534D and U+5350 correspond to the left-facing and right-facing *svasti* signs, respectively. These CJK unified ideographs have adopted Han script-specific features and properties: they share metrics and type style characteristics with other ideographs, and are given radicals and stroke counts like those for other ideographs.

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

The marks U+0F35 TIBETAN MARK NGAS BZUNG NYI ZLA and U+0F37 TIBETAN MARK NGAS BZUNG SGOR RTAGS conceptually attach to a *tsek-bar* rather than to an individual character and function more like attributes than characters—for example, as 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, among other things.

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. For example, 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, although 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. Implementations must ensure that decomposed stacks and precomposed forms are interpreted equivalently (see Section 3.7, *Decomposition*). The compound consonants are

explicitly coded as follows: U+0F93 TIBETAN SUBJOINED LETTER GHA, U+0F9D TIBETAN SUBJOINED LETTER DDHA, U+0FA2 TIBETAN SUBJOINED LETTER DHA, U+0FA7 TIBETAN SUBJOINED LETTER BHA, U+0FAC TIBETAN SUBJOINED LETTER DZHA, and U+0FB9 TIBETAN SUBJOINED LETTER KSSA.

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

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

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

Other Signs. U+0F09 TIBETAN MARK BSKUR YIG MGO is a list enumerator used at the beginning 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*, and *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, when 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, a gap at the right margin often cannot be filled. If the gap is short, it will be left as is and the line will be said to be justified enough, even though by machine-justification standards the line is not truly flush on the right. If the gap is large, the intervening space will be filled with as many *tseks* as are required to justify the line. Again, the justification is not done perfectly in the way that English text might be perfectly right-justified; as long as the last *tsek* is more or less at the right margin, that will do. The net result is that of a right-justified, blocklike look to the text, but the actual lines are always a little right-ragged.

Justifying *tseks* are nearly always used to pad the end of a line when the preceding character is a *tsek*—in other words, when the end of a line arrives in the middle of tshig-grub (see the previous definition under “Tibetan Punctuation”). However, it is unusual for a line that ends at the end of a tshig-grub to have justifying *tseks* added to the *shay* at the end of the tshig-grub. That is, a sequence like that shown in the first line of *Figure 10-2* is not usually padded as in the second line of *Figure 10-2*, though it is allowable. In this case, instead of justifying the line with *tseks*, the space between *shays* is enlarged and/or the whitespace following the final *shay* is usually left as is. Padding is *never* applied following an actual space character. For example, given the existence of a space after a *shay*, a line such as the third

line of *Figure 10-2* may not be written with the padding as shown because the final *shay* should have a space after it, and padding is never applied after spaces. The same rule applies where the final *consonant* of a tshig-grub that ends a line is a “ka” or “ga”. In that case, the ending *shay* is dropped but a space is still required after the consonant and that space must not be padded. For example, the appearance shown in the fourth line of *Figure 10-2* is not acceptable.

Figure 10-2. Justifying Tibetan Tseks

```

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

```

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

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

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

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

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

10.3 Lepcha

Lepcha: U+1C00–U+1C4F

Lepcha is a Sino-Tibetan language spoken by people in Sikkim and in the West Bengal state of India, especially in the Darjeeling district, which borders Sikkim. The Lepcha script is a writing system thought to have been invented around 1720 CE by the Sikkim king Phyag-rdor rNam-rgyal (“Chakdor Namgyal,” born 1686). Both the language and the script are also commonly known by the term *Rong*.

Structure. The Lepcha script was based directly on the Tibetan script. The letter forms are obviously related to corresponding Tibetan letters. However, the *dbu-med* Tibetan precursors to Lepcha were originally written in vertical columns, possibly influenced by Chinese conventions. When Lepcha was invented it changed the *dbu-med* text to a left-to-right, horizontal orientation. In the process, the entire script was effectively rotated ninety degrees counter-clockwise, so that the letters resemble Tibetan letters turned on their sides. This reorientation resulted in some letters which are nonspacing marks in Tibetan becoming spacing letters in Lepcha. Lepcha also introduced its own innovations, such as the use of diacritical marks to represent final consonants.

The Lepcha script is an abugida: the consonant letters have an inherent vowel, and dependent vowels (*matras*) are used to modify the inherent vowel of the consonant. No virama (or vowel killer) is used to remove the inherent vowel. Instead, the script has a separate set of explicit final consonants which are used to represent a consonant with no inherent vowel.

Vowels. Initial vowels are represented by the neutral letter U+1C23 LEPCHA LETTER A, followed by the appropriate dependent vowel. U+1C23 LEPCHA LETTER A thus functions as a vowel carrier.

The dependent vowel signs in Lepcha always follow the base consonant in logical order. However, in rendering, three of these dependent vowel signs, *-i*, *-o*, and *-oo*, reorder to the left side of their base consonant. One of the dependent vowel signs, *-e*, is a nonspacing mark which renders below its base consonant.

Medials. There are three medial consonants, or glides: *-ya*, *-ra*, and *-la*. The first two are represented by separate characters, U+1C24 LEPCHA SUBJOINED LETTER YA and U+1C25 LEPCHA SUBJOINED LETTER RA. These are called “subjoined”, by analogy with the corresponding letters in Tibetan, which actually do join below a Tibetan consonant, but in Lepcha these are spacing forms which occur to the right of a consonant letter and then ligate with it. These two medials can also occur in sequence to form a composite medial, *-rya*. In that case both medials ligate with the preceding consonant.

On the other hand, Lepcha does not have a separate character to represent the medial *-la*. Phonological consonant clusters of the form *kla*, *gla*, *pla*, and so on simply have separate, atomic characters encoded for them. With few exceptions, these letters for phonological clusters with the medial *-la* are independent letter forms, not clearly related to the corresponding consonants without *-la*.

Retroflex Consonants. The Lepcha language contains three retroflex consonants: [ʈ], [ʈʰ], and [ɖ]. Traditionally, these retroflex consonants have been written in the Lepcha script with the syllables *kra*, *hra*, and *gra*, respectively. In other words, the retroflex *t* would be represented as <U+1C00 LEPCHA LETTER KA, U+1C25 LEPCHA SUBJOINED LETTER RA>. To distinguish such a sequence representing a retroflex *t* from a sequence representing the actual syllable [kra], it is common to use the *nukta* diacritic sign, U+1C37 LEPCHA SIGN NUKTA. In that case, the retroflex *t* would be visually distinct, and would be represented by the sequence <U+1C00 LEPCHA LETTER KA, U+1C37 LEPCHA SIGN NUKTA, U+1C25 LEP-

CHA SUBJOINED LETTER RA>. Recently, three newly invented letters have been added to the script to unambiguously represent the retroflex consonants: U+1C4D LEPCHA LETTER TTA, U+1C4E LEPCHA LETTER TTHA, and U+1C4F LEPCHA LETTER DDA.

Ordering of Syllable Components. Dependent vowels and other signs are encoded after the consonant to which they apply. The ordering of elements is shown in more detail in Table 10-2.

Table 10-2. Lepcha Syllabic Structure

| Class | Example | Encoding |
|----------------------|---------|----------------------------------|
| consonant, letter a | ᱠ | [U+1C00..U+1C23, U+1C4D..U+1C4F] |
| nukta | ᱡ | U+1C37 |
| medial -ra | ᱢ | U+1C25 |
| medial -ya | ᱣ | U+1C24 |
| dependent vowel | ᱤ | [U+1C26..U+1C2C] |
| final consonant sign | ᱥ | [U+1C2D..U+1C35] |
| syllabic modifier | ᱦ | U+1C36 |

Rendering. Most final consonants consist of nonspacing marks rendered above the base consonant of a syllable.

The combining mark U+1C36 LEPCHA SIGN RAN occurs only after the inherent vowel *-a* or the dependent vowels *-aa* and *-i*. When it occurs together with a final consonant sign, the *ran* sign renders above the sign for that final consonant.

The two final consonants representing the velar nasal occur in complementary contexts. U+1C34 LEPCHA CONSONANT SIGN NYIN-DO is only used when there is no dependent vowel in the syllable. U+1C35 LEPCHA CONSONANT SIGN KANG is used instead when there is a dependent vowel. These two consonant signs are rendered to the left of the base consonant. If used with a left-side dependent vowel, the glyph for the *kang* is rendered to the left of the dependent vowel. This behavior is understandable because these two marks are derived from the Tibetan analogues of the Brahmic *bindu* and *candrabindu*, which normally stand above a Brahmic *aksara*.

Digits. The Lepcha script has its own, distinctive set of digits.

Punctuation. Currently the Lepchas use traditional punctuation marks only when copying the old books. In everyday writing they use common Western punctuation marks such as comma, full stop, and question mark.

The traditional punctuation marks include a script-specific *danda* mark, U+1C3B LEPCHA PUNCTUATION TA-ROL, and a *double danda*, U+1C3C LEPCHA NYET THYOOM TA-ROL. Depending on style and hand, the Lepcha *ta-rol* may have a glyph appearance more like its Tibetan analogue, U+0F0D TIBETAN MARK SHAD.

10.4 Phags-pa

Phags-pa: U+A840–U+A87F

The Phags-pa script is an historic script with some limited modern use. It bears some similarity to Tibetan and has no case distinctions. It is written vertically in columns running

from left to right, like Mongolian. Units are often composed of several syllables and may be separated by whitespace.

The term *Phags-pa* is often written with an initial apostrophe: *'Phags-pa*. The Unicode Standard makes use of the alternative spelling without an initial apostrophe because apostrophes are not allowed in the normative character and block names.

History. The Phags-pa script was devised by the Tibetan lama Blo-gros rGyal-mtshan [lodoi jaltsan] (1235–1280 CE), commonly known by the title *Phags-pa Lama* (“exalted monk”), at the behest of Khubilai Khan (reigned 1260–1294) when he assumed leadership of the Mongol tribes in 1260. In 1269, the “new Mongolian script,” as it was called, was promulgated by imperial edict for use as the national script of the Mongol empire, which from 1279 to 1368, as the Yuan dynasty, encompassed all of China.

The new script was not only intended to replace the Uighur-derived script that had been used to write Mongolian since the time of Genghis Khan (reigned 1206–1227), but was also intended to be used to write all the diverse languages spoken throughout the empire. Although the Phags-pa script never succeeded in replacing the earlier Mongolian script and had only very limited usage in writing languages other than Mongolian and Chinese, it was used quite extensively during the Yuan dynasty for a variety of purposes. There are many monumental inscriptions and manuscript copies of imperial edicts written in Mongolian or Chinese using the Phags-pa script. The script can also be found on a wide range of artifacts, including seals, official passes, coins, and banknotes. It was even used for engraving the inscriptions on Christian tombstones. A number of books are known to have been printed in the Phags-pa script, but all that has survived are some fragments from a printed edition of the Mongolian translation of a religious treatise by the Phags-pa Lama’s uncle, Sakya Pandita. Of particular interest to scholars of Chinese historical linguistics is a rhyming dictionary of Chinese with phonetic readings for Chinese ideographs given in the Phags-pa script.

An ornate, pseudo-archaic “seal script” version of the Phags-pa script was developed specifically for engraving inscriptions on seals. The letters of the seal script form of Phags-pa mimic the labyrinthine strokes of Chinese seal script characters. A great many official seals and seal impressions from the Yuan dynasty are known. The seal script was also sometimes used for carving the title inscription on stone stelae, but never for writing ordinary running text.

Although the vast majority of extant Phags-pa texts and inscriptions from the thirteenth and fourteenth centuries are written in the Mongolian or Chinese languages, there are also examples of the script being used for writing Uighur, Tibetan, and Sanskrit, including two long Buddhist inscriptions in Sanskrit carved in 1345.

After the fall of the Yuan dynasty in 1368, the Phags-pa script was no longer used for writing Chinese or Mongolian. However, the script continued to be used on a limited scale in Tibet for special purposes such as engraving seals. By the late sixteenth century, a distinctive, stylized variety of Phags-pa script had developed in Tibet, and this Tibetan-style Phags-pa script, known as *hor-yig*, “Mongolian writing” in Tibetan, is still used today as a decorative script. In addition to being used for engraving seals, the Tibetan-style Phags-pa script is used for writing book titles on the covers of traditional style books, for architectural inscriptions such as those found on temple columns and doorways, and for calligraphic samplers.

Basic Structure. The Phags-pa script is based on Tibetan, but unlike any other Brahmic script Phags-pa is written vertically from top to bottom in columns advancing from left to right across the writing surface. This unusual directionality is borrowed from Mongolian, as is the way in which Phags-pa letters are ligated together along a vertical stem axis. In

modern contexts, when embedded in horizontally oriented scripts, short sections of Phags-pa text may be laid out horizontally from left to right.

Despite the difference in directionality, the Phags-pa script fundamentally follows the Tibetan model of writing, and consonant letters have an inherent /a/ vowel sound. However, Phags-pa vowels are independent letters, not vowel signs as is the case with Tibetan, so they may start a syllable without being attached to a null consonant. Nevertheless, a null consonant (U+A85D PHAGS-PA LETTER A) is still needed to write an initial /a/ and is orthographically required before a diphthong or the semivowel U+A867 PHAGS-PA SUBJOINED LETTER WA. Only when writing Tibetan in the Phags-pa script is the null consonant required before an initial pure vowel sound.

Except for the *candrabinu* (which is discussed later in this section), Phags-pa letters read from top to bottom in logical order, so the vowel letters *i*, *e*, and *o* are placed below the preceding consonant—unlike in Tibetan, where they are placed above the consonant they modify.

Syllable Division. Text written in the Phags-pa script is broken into discrete syllabic units separated by whitespace. When used for writing Chinese, each Phags-pa syllabic unit corresponds to a single Han ideograph. For Mongolian and other polysyllabic languages, a single word is typically written as several syllabic units, each separated from each other by whitespace.

For example, the Mongolian word *tengri*, “heaven,” which is written as a single ligated unit in the Mongolian script, is written as two separate syllabic units, *deng ri*, in the Phags-pa script. Syllable division does not necessarily correspond directly to grammatical structure. For instance, the Mongolian word *usun*, “water,” is written *u sun* in the Phags-pa script, but its genitive form *usunu* is written *u su nu*.

Within a single syllabic unit, the Phags-pa letters are normally ligated together. Most letters ligate along a righthand stem axis, although reversed-form letters may instead ligate along a lefthand stem axis. The letter U+A861 PHAGS-PA LETTER O ligates along a central stem axis.

In traditional Phags-pa texts, normally no distinction is made between the whitespace used in between syllables belonging to the same word and the whitespace used in between syllables belonging to different words. Line breaks may occur between any syllable, regardless of word status. In contrast, in modern contexts, influenced by practices used in the processing of Mongolian text, U+202F NARROW NO-BREAK SPACE (NNBSP) may be used to separate syllables within a word, whereas U+0020 SPACE is used between words—and line breaking would be affected accordingly.

Candrabinu. U+A873 PHAGS-PA LETTER CANDRABINDU is used in writing Sanskrit mantras, where it represents a final nasal sound. However, although it represents the final sound in a syllable unit, it is always written as the first glyph in the sequence of letters, above the initial consonant or vowel of the syllable, but not ligated to the following letter. For example, *om* is written as a *candrabinu* followed by the letter *o*. To simplify cursor placement, text selection, and so on, the *candrabinu* is encoded in visual order rather than logical order. Thus *om* would be represented by the sequence <U+A873, U+A861>, rendered as shown in *Figure 10-3*.

Figure 10-3. Phags-pa Syllable Om



As the *candrabindu* is separated from the following letter, it does not take part in the shaping behavior of the syllable unit. Thus, in the syllable *om*, the letter *o* (U+A861) takes the isolate positional form.

Alternate Letters. Four alternate forms of the letters *ya*, *sha*, *ha*, and *fa* are encoded for use in writing Chinese under certain circumstances:

U+A86D PHAGS-PA LETTER ALTERNATE YA

U+A86E PHAGS-PA LETTER VOICELESS SHA

U+A86F PHAGS-PA LETTER VOICED HA

U+A870 PHAGS-PA LETTER ASPIRATED FA

These letters are used in the early-fourteenth-century Phags-pa rhyming dictionary of Chinese, *Menggu ziyun*, to represent historical phonetic differences between Chinese syllables that were no longer reflected in the contemporary Chinese language. This dictionary follows the standard phonetic classification of Chinese syllables into 36 initials, but as these had been defined many centuries previously, by the fourteenth century some of the initials had merged together or diverged into separate sounds. To distinguish historical phonetic characteristics, the dictionary uses two slightly different forms of the letters *ya*, *sha*, *ha*, and *fa*.

The historical phonetic values that U+A86E, U+A86F, and U+A870 represent are indicated by their character names, but this is not the case for U+A86D, so there may be some confusion as to when to use U+A857 PHAGS-PA LETTER YA and when to use U+A86D PHAGS-PA LETTER ALTERNATE YA. U+A857 is used to represent historic null initials, whereas U+A86D is used to represent historic palatal initials.

Numbers. There are no special characters for numbers in the Phags-pa script, so numbers are spelled out in full in the appropriate language.

Punctuation. The vast majority of traditional Phags-pa texts do not make use of any punctuation marks. However, some Mongolian inscriptions borrow the Mongolian punctuation marks U+1802 MONGOLIAN COMMA, U+1803 MONGOLIAN FULL STOP, and U+1805 MONGOLIAN FOUR DOTS.

Additionally, a small circle punctuation mark is used in some printed Phags-pa texts. This mark can be represented by U+3002 IDEOGRAPHIC FULL STOP, but for Phags-pa the *ideographic full stop* should be centered, not positioned to one side of the column. This follows traditional, historic practice for rendering the ideographic full stop in Chinese text, rather than more modern typography.

Tibetan Phags-pa texts also use head marks, U+A874 PHAGS-PA SINGLE HEAD MARK U+A875 PHAGS-PA DOUBLE HEAD MARK, to mark the start of an inscription, and *shad* marks, U+A876 PHAGS-PA MARK SHAD and U+A877 PHAGS-PA MARK DOUBLE SHAD, to mark the end of a section of text.

Positional Variants. The four vowel letters U+A85E PHAGS-PA LETTER I, U+A85F PHAGS-PA LETTER U, U+A860 PHAGS-PA LETTER E, and U+A861 PHAGS-PA LETTER O have different isolate, initial, medial, and final glyph forms depending on whether they are immediately preceded or followed by another Phags-pa letter (other than U+A873 PHAGS-PA LETTER CANDRABINDU, which does not affect the shaping of adjacent letters). The code charts show these four characters in their isolate form. The various positional forms of these letters are shown in *Table 10-3*.

Consonant letters and the vowel letter U+A866 PHAGS-PA LETTER EE do not have distinct positional forms, although initial, medial, final, and isolate forms of these letters may be

Table 10-3. Phags-pa Positional Forms of I, U, E, and O

| Letter | Isolate | Initial | Medial | Final |
|--------------------------|---------|---------|--------|-------|
| U+A85E PHAGS-PA LETTER I | ᠶ | ᠶ | ᠶ | ᠶ |
| U+A85F PHAGS-PA LETTER U | ᠸ | ᠸ | ᠸ | ᠸ |
| U+A860 PHAGS-PA LETTER E | ᠺ | ᠺ | ᠺ | ᠺ |
| U+A861 PHAGS-PA LETTER O | ᠻ | ᠻ | ᠻ | ᠻ |

distinguished by the presence or absence of a stem extender that is used to ligate to the following letter.

The invisible format characters U+200D ZERO WIDTH JOINER (ZWJ) and U+200C ZERO WIDTH NON-JOINER (ZWNJ) may be used to override the expected shaping behavior, in the same way that they do for Mongolian and other scripts (see *Chapter 16, Special Areas and Format Characters*). For example, ZWJ may be used to select the initial, medial, or final form of a letter in isolation:

<U+200D, U+A861, U+200D> selects the medial form of the letter *o*

<U+200D, U+A861> selects the final form of the letter *o*

<U+A861, U+200D> selects the initial form of the letter *o*

Conversely, ZWNJ may be used to inhibit expected shaping. For example, the sequence <U+A85E, U+200C, U+A85F, U+200C, U+A860, U+200C, U+A861> selects the isolate forms of the letters *i*, *u*, *e*, and *o*.

Mirrored Variants. The four characters U+A869 PHAGS-PA LETTER TTA, U+A86A PHAGS-PA LETTER TTTHA, U+A86B PHAGS-PA LETTER DDA, and U+A86C PHAGS-PA LETTER NNA are mirrored forms of the letters U+A848 PHAGS-PA LETTER TA, U+A849 PHAGS-PA LETTER THA, U+A84A PHAGS-PA LETTER DA, and U+A84B PHAGS-PA LETTER NA, respectively, and are used to represent the Sanskrit retroflex dental series of letters. Because these letters are mirrored, their stem axis is on the lefthand side rather than the righthand side, as is the case for all other consonant letters. This means that when the letters *tta*, *ttha*, *dda*, and *nna* occur at the start of a syllable unit, to correctly ligate with them any following letters normally take a mirrored glyph form. Because only a limited number of words use these letters, only the letters U+A856 PHAGS-PA LETTER SMALL A, U+A85C PHAGS-PA LETTER HA, U+A85E PHAGS-PA LETTER I, U+A85F PHAGS-PA LETTER U, U+A860 PHAGS-PA LETTER E, and U+A868 PHAGS-PA SUBJOINED LETTER YA are affected by this glyph mirroring behavior. The Sanskrit syllables that exhibit glyph mirroring after *tta*, *ttha*, *dda*, and *nna* are shown in *Table 10-4*.

Table 10-4. Contextual Glyph Mirroring in Phags-pa

| Character | Syllables with Glyph Mirroring | Syllables without Glyph Mirroring |
|-------------------------------------|--------------------------------|-----------------------------------|
| U+A856 PHAGS-PA LETTER SMALL A | <i>tthā</i> | <i>ttā, tthā</i> |
| U+A85E PHAGS-PA LETTER I | <i>tthi, nni</i> | <i>tthi</i> |
| U+A85F PHAGS-PA LETTER U | <i>nnu</i> | |
| U+A860 PHAGS-PA LETTER E | <i>tthe, dde, nne</i> | |
| U+A85C PHAGS-PA LETTER HA | <i>ddha</i> | |
| U+A868 PHAGS-PA SUBJOINED LETTER YA | <i>nnya</i> | |

Glyph mirroring is not consistently applied to the letters U+A856 PHAGS-PA LETTER SMALL A and U+A85E PHAGS-PA LETTER I in the extant Sanskrit Phags-pa inscriptions. The letter *i* may occur both mirrored and unmirrored after the letter *ttha*, although it always occurs

mirrored after the letter *nna*. *Small a* is not normally mirrored after the letters *tta* and *ttha* as its mirrored glyph is identical in shape to U+A85A PHAGS-PA LETTER SHA. Nevertheless, *small a* does sometimes occur in a mirrored form after the letter *ttha*, in which case context indicates that this is a mirrored letter *small a* and not the letter *sha*.

When any of the letters *small a*, *i*, *u*, *e*, *ha*, or *subjoined ya* immediately follow either *tta*, *ttha*, *dda*, or *nna* directly or another mirrored letter, then a mirrored glyph form of the letter should be selected automatically by the rendering system. Although *small a* is not normally mirrored in extant inscriptions, for consistency it is mirrored by default after *tta*, *ttha*, *dda*, and *nna* in the rendering model for Phags-pa.

To override the default mirroring behavior of the letters *small a*, *ha*, *i*, *u*, *e*, and *subjoined ya*, U+FE00 VARIATION SELECTOR-1 (VS1) may be applied to the appropriate character, as shown in Table 10-5. Note that only the variation sequences shown in Table 10-5 are valid; any other sequence of a Phags-pa letter and VS1 is unspecified.

Table 10-5. Phags-pa Standardized Variants

| Character Sequence | Description of Variant Appearance |
|--------------------|---|
| <U+A856, U+FE00> | <i>phags-pa letter reversed shaping small a</i> |
| <U+A85C, U+FE00> | <i>phags-pa letter reversed shaping ha</i> |
| <U+A85E, U+FE00> | <i>phags-pa letter reversed shaping i</i> |
| <U+A85F, U+FE00> | <i>phags-pa letter reversed shaping u</i> |
| <U+A860, U+FE00> | <i>phags-pa letter reversed shaping e</i> |
| <U+A868, U+FE00> | <i>phags-pa letter reversed shaping ya</i> |

In Table 10-5, “reversed shaping” means that the appearance of the character is reversed with respect to its expected appearance. Thus, if no mirroring would be expected for the character in the given context, applying VS1 would cause the rendering engine to select a mirrored glyph form. Similarly, if context would dictate glyph mirroring, application of VS1 would inhibit the expected glyph mirroring. This mechanism will typically be used to select a mirrored glyph for the letters *small a*, *ha*, *i*, *u*, *e*, or *subjoined ya* in isolation (for example, in discussion of the Phags-pa script) or to inhibit mirroring of the letters *small a* and *i* when they are not mirrored after the letters *tta* and *ttha*, as shown in Figure 10-4.

Figure 10-4. Phags-pa Reversed Shaping



The first example illustrates the normal shaping for the syllable *thi*. The second example shows the reversed shaping for *i* in that syllable and would be represented by a standardized variation sequence: <U+A849, U+A85E, U+FE00>. Example 3 illustrates the normal shaping for the Sanskrit syllable *tthi*, where the reversal of the glyph for the letter *i* is automatically conditioned by the lefthand stem placement of the Sanskrit letter *ttha*. Example 4 shows reversed shaping for *i* in the syllable *tthi* and would be represented by a standardized variation sequence: <U+A86A, U+A85E, U+FE00>.

10.5 Limbu

Limbu: U+1900–U+194F

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Glyph Placement. The glyph positions for Limbu combining characters are summarized in Table 10-6.

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

Table 10-6. Positions of Limbu Combining Characters

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

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

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

10.6 Syloti Nagri

Syloti Nagri: U+A800–U+A82F

Syloti Nagri is a lesser-known Brahmi-derived script used for writing the Sylheti language. Sylheti is an Indo-European language spoken by some 5 million speakers in the Barak Valley region of northeast Bangladesh and southeast Assam in India. Worldwide there may be as many as 10 million speakers. Sylheti has commonly been regarded as a dialect of Bengali, with which it shares a high proportion of vocabulary.

The Syloti Nagri script has 27 consonant letters with an inherent vowel of /o/ and 5 independent vowel letters. There are 5 dependent vowel signs that are attached to a consonant letter. Unlike Devanagari, there are no vowel signs that appear to the left of their associated consonant.

Only two proper diacritics are encoded to support Syloti Nagri: *anusvara* and *hasanta*. Aside from its traditional Indic designation, *anusvara* can also be considered a final form for the sequence /-ng/, which does not have a base glyph in Syloti Nagri because it does not occur in other positions. *Anusvara* can also occur with the vowels U+A824 𑒠 SYLOTI NAGRI VOWEL SIGN I and U+A826 𑒡 SYLOTI NAGRI VOWEL SIGN E, creating a potential problem with the display of both items. It is recommended that *anusvara* always occur in sequence after any vowel signs, as a final character.

Virama and Conjuncts. Syloti Nagri is atypical of Indic scripts in use of the *virama* (*hasanta*) and conjuncts. Conjuncts are not strictly correlated with the phonology being represented. They are neither necessary in contexts involving a dead consonant, nor are they limited to such contexts. *Hasanta* was only recently introduced into the script and is used only in limited contexts. Conjuncts are not limited to sequences involving dead consonants but can be formed from pairs of characters of almost any type (consonant, independent vowel, dependent vowel) and can represent a wide variety of syllables. It is generally unnecessary to overtly indicate dead consonants with a conjunct or explicit *hasanta*. The only restriction is that an overtly rendered *hasanta* cannot occur in connection with the first element of a conjunct. The absence of *hasanta* does not imply a live consonant and has no bearing on the occurrence of conjuncts. Similarly, the absence of a conjunct does not imply a live consonant and has no bearing on the occurrence of *hasanta*.

Digits. There are no unique Syloti Nagri digits. When digits do appear in Syloti Nagri texts, they are generally Bengali forms. Any font designed to support Syloti Nagri should include the Bengali digits because there is no guarantee that they would otherwise exist in a user's computing environment. They should use the corresponding Bengali block code points, U+09E6..U+09EF.

Punctuation. With the advent of digital type and the modernization of the Syloti Nagri script, one can expect to find all of the traditional punctuation marks borrowed from the Latin typography: *period*, *comma*, *colon*, *semicolon*, *question mark*, and so on. In addition, the Devanagari *single danda* and *double danda* are used with great frequency.

Poetry Marks. Four native poetry marks are included in the Syloti Nagri block. The script also makes use of U+2055 * FLOWER PUNCTUATION MARK (in the General Punctuation block) as a poetry mark.

10.7 Kaithi

Kaithi: U+11080–U+110CF

Kaithi, properly transliterated Kaithī, is a North Indian script, related to the Devanagari and Gujarati scripts. It was used in the area of the present-day states of Bihar and Uttar Pradesh in northern India.

Kaithi was employed for administrative purposes, commercial transactions, correspondence, and personal records, as well as to write religious and literary materials. As a means of administrative communication, the script was in use at least from the 16th century until the early 20th century, when it was eventually eclipsed by Devanagari. Kaithi was used to write Bhojpuri, Magahi, Awadhi, Maithili, Urdu, and other languages related to Hindi.

Standards. There is no preexisting character encoding standard for the Kaithi script. The repertoire encoded in this block is based on the standard form of Kaithi developed by the British government of Bihar and the British provinces of northwest India in the nineteenth century. A few additional Kaithi characters found in manuscripts, printed books, alphabet charts, and other inventories of the script are also included.

Styles. There are three presentation styles of the Kaithi script, each generally associated with a different language: Bhojpuri, Magahi, or Maithili. The Magahi style was adopted for official purposes in the state of Bihar, and is the basis for the representative glyphs in the code charts.

Rendering Behavior. Kaithi is a Brahmi-derived script closely related to Devanagari. In general, the rules for Devanagari rendering apply to Kaithi as well. For more information, see *Section 9.1, Devanagari*.


Vowel Letters. An independent Kaithi letter for *vocalic r* is represented by the consonant-vowel combination: U+110A9 KAITHI LETTER RA and U+110B2 KAITHI VOWEL SIGN II.

In print, the distinction between short and long forms of *i* and *u* is maintained. However, in handwritten text, there is a tendency to use the long vowels for both lengths.

Consonant Conjuncts. Consonant clusters were handled in various ways in Kaithi. Some spoken languages that used the Kaithi script simplified clusters by inserting a vowel between the consonants, or through metathesis. When no such simplification occurred, conjuncts were represented in different ways: by ligatures, as the combination of the half-form of the first consonant and the following consonant, with an explicit virama (U+110B9 KAITHI SIGN VIRAMA) between two consonants, or as two consonants without a virama.

Consonant conjuncts in Kaithi are represented with a virama between the two consonants in the conjunct. For example, the ordinary representation of the conjunct *mba* would be by the sequence:

```
U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA +
U+110A5 KAITHI LETTER BA
```

Consonant conjuncts may be rendered in distinct ways. Where there is a need to render conjuncts in the exact form as they appear in a particular source document, U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER can be used to request the appropriate presentation by the rendering system. For example, to display the explicitly ligated glyph  for the conjunct *mba*, U+200D ZERO WIDTH JOINER is inserted after the virama:

```
U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA +
U+200D ZERO WIDTH JOINER + U+110A5 KAITHI LETTER BA
```

To block use of a ligated glyph for the conjunct, and instead to display the conjunct with an explicit virama, U+200C ZERO WIDTH NON-JOINER is inserted after the virama:

```
U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA +
U+200C ZERO WIDTH NON-JOINER + U+110A5 KAITHI LETTER BA
```

Conjuncts composed of a nasal and a consonant may be written either as a ligature with the half-form of the appropriate class nasal letter, or the full form of the nasal letter with an explicit virama (U+110B9 KAITHI SIGN VIRAMA) and consonant. In Grierson's *Linguistic Survey of India*, however, U+110A2 KAITHI LETTER NA is used for all articulation classes, both in ligatures and when the full form of the nasal appears with the virama.

Ruled Lines. Kaithi, unlike Devanagari, does not employ a headstroke. While several manuscripts and books show a headstroke similar to that of Devanagari, the line is actually a ruled line used for emphasis, titling or sectioning, and is not broken between individual letters. Some Kaithi fonts, however, were designed with a headstroke, but the line is not broken between individual letters, as would occur in Devanagari.

Nukta. Kaithi includes a nukta sign, U+110BA KAITHI SIGN NUKTA, a dot which is used as a diacritic below various consonants to form new letters. For example, the nukta is used to distinguish the sound *va* from *ba*. The precomposed character U+110AB KAITHI LETTER

va is separately encoded, and has a canonical decomposition into the sequence of U+110A5 KAITHI LETTER BA plus U+110BA KAITHI SIGN NUKTA. Precomposed characters are also encoded for two other Kaithi letters, *rha* and *dddha*.

The glyph for U+110A8 KAITHI LETTER YA may appear with or without a nukta. Because the form without the nukta is considered a glyph variant, it is not separately encoded as a character. The representative glyph used in the chart contains the dot. The nukta diacritic also marks letters representing some sounds in Urdu or sounds not native to Hindi. No precomposed characters are encoded in those cases, and such letters must be represented by a base character followed by the nukta.

Punctuation. A number of Kaithi-specific punctuation marks are encoded. Two marks designate the ends of text sections: U+110BE KAITHI SECTION MARK, which generally indicates the end of a sentence, and U+110BF KAITHI DOUBLE SECTION MARK, which delimits larger blocks of text, such as paragraphs. Both section marks are generally drawn so that their glyphs extend to the edge of the text margins, particularly in manuscripts.

The character U+110BD KAITHI NUMBER SIGN is a format control character that interacts with digits, occurring either above or below a digit. The position of the KAITHI NUMBER SIGN indicates its usage: when the mark occurs above a digit, it indicates a number in an itemized list, similar to U+2116 NUMERO SIGN. If it occurs below a digit, it indicates a numerical reference. Like U+0600 ARABIC NUMBER SIGN and the other Arabic signs that span numbers (see *Section 8.2, Arabic*), the KAITHI NUMBER SIGN precedes the numbers they graphically interact with, rather than following them, as would combining characters. The U+110BC KAITHI ENUMERATION SIGN is the spacing version of the KAITHI NUMBER SIGN, and is used for inline usage.

U+110BB KAITHI ABBREVIATION SIGN, shaped like a small circle, is used in Kaithi to indicate abbreviations. This mark is placed at the point of elision or after a ligature to indicate common words or phrases that are abbreviated, in a similar way to U+0970 DEVANAGARI ABBREVIATION SIGN.

Kaithi makes use of two script-specific dandas: U+110C0 KAITHI DANDA and U+110C1 KAITHI DOUBLE DANDA.

For other marks of punctuation occurring in Kaithi texts, available Unicode characters may be used. A cross-shaped character, used to mark phrase boundaries, can be represented by U+002B PLUS SIGN. For hyphenation, users should follow whatever is the recommended practice found in similar Indic script traditions, which might be U+2010 HYPHEN or U+002D HYPHEN-MINUS. For dot-like marks that appear as word-separators, U+2E31 WORD SEPARATOR MIDDLE DOT, or, if the word boundary is more like a dash, U+2010 HYPHEN can be used.

Digits. The digits in Kaithi are considered to be stylistic variants of those used in Devanagari. Hence the Devanagari digits located at U+0966..096F should be employed. To indicate fractions and unit marks, Kaithi makes use of the numbers encoded in the Common Indic Number Forms block, U+A830..A839.

10.8 Saurashtra

Saurashtra: U+A880–U+A8DF

Saurashtra is an Indo-European language, related to Gujarati and spoken by about 310,000 people in southern India. The Telugu, Tamil, Devanagari, and Saurashtra scripts have been used to publish books in Saurashtra since the end of the 19th century. At present, Saurash-

tra is most often written in the Tamil script, augmented with the use of superscript digits and a colon to indicate sounds not available in the Tamil script.

The Saurashtra script is of the Brahmic type. Early Saurashtra text made use of conjuncts, which can be handled with the usual Brahmic shaping rules. The modernized script, developed in the 1880s, has undergone some simplification. Modern Saurashtra does not use complex consonant clusters, but instead marks a killed vowel with a visible virama, U+A8CF SAURASHTRA SIGN VIRAMA. An exception to the non-occurrence of complex consonant clusters is the conjunct *ksa*, formed by the sequence <U+A892, U+A8C4, U+200D, U+A8B0>. This conjunct is sorted as a unique letter in older dictionaries. Apart from its use to form *ksa*, the virama is always visible by default in modern Saurashtra. If necessary, U+200D ZERO WIDTH JOINER may be used to force conjunct behavior.

The Unicode encoding of the Saurashtra script supports both older and newer conventions for writing Saurashtra text.

Glyph Placement. The vowel signs (*matras*) in Saurashtra follow the consonant to which they are applied. The long and short -i vowels, however, are typographically joined to the top right corner of their consonant. Vowel signs are also applied to U+A8B4 SAURASHTRA CONSONANT SIGN HAARU.

Digits. The Saurashtra script has its own set of digits. These are separately encoded in the Saurashtra block.

Punctuation. Western-style punctuation, such as comma, full stop, and the question mark are used in modern Saurashtra text. U+A8CE SAURASHTRA DANDA is used as a text delimiter in traditional prose. U+A8CE SAURASHTRA DANDA and U+A8CF SAURASHTRA DOUBLE DANDA are used in poetic text.

Saurashtra Consonant Sign Haaru. The character U+A8B4 SAURASHTRA CONSONANT SIGN HAARU, transliterated as “H”, is unique to Saurashtra, and does not have an equivalent in the Devanagari, Tamil, or Telugu scripts. It functions in some regards like the Tamil *aytam*, modifying other letters to represent sounds not found in the basic Brahmic alphabet. It is a dependent consonant and is thus classified as a consonant sign in the encoding.

10.9 Sharada

Sharada is a historical script that was used to write Sanskrit, Kashmiri, and other languages of northern South Asia. It served as the principal inscriptional and literary script of Kashmir from the 8th century CE until the 20th century. In the 19th century, expanded use of the Arabic script to write Kashmiri and the growth of Devanagari contributed to the marginalization of Sharada. Today the script is employed in a limited capacity by Kashmiri pandits for horoscopes and ritual purposes.

Rendering Behavior. Sharada is a Brahmi-based script, closely related to Devanagari. In general, the rules for Devanagari rendering apply to Sharada as well. For more information, see *Section 9.1, Devanagari*.

Ruled Lines. While the headstroke is an important structural feature of a character’s glyph in Sharada, there is no rule governing the joining of headstrokes of characters to other characters. The variation was probably due to scribal preference, and should be handled at the font level.

Virama. The U+111C0 𑆀 SHARADA SIGN VIRAMA is a spacing mark, written to the right of the consonant letter it modifies. Semantically, it is identical to the Devanagari *virama* and other similar Indic scripts.

Candrabindu and Avagraha. U+11180 𑀓 SHARADA SIGN CANDRABINDU indicates nasalization of a vowel. It may appear in manuscripts in an inverted form but with no semantic difference. Such glyph variants should be handled in the font. U+111C1 𑀛 SHARADA AVAGRAHA represents the elision of a word-initial *a*. Unlike the usual practice in Devanagari in which the *avagraha* is written at the normal letter height and attaches to the top stroke of the following character, the *avagraha* in Sharada is written at or below the baseline and does not connect to the neighboring letter.

Jihvamuliya and Upadhmaniya. The velar and labial allophones of /h/, followed by voiceless velar and labial stops respectively, are written in Sharada with separate signs, U+111C2 𑀛 SHARADA SIGN JIHVAMULIYA and U+111C3 𑀜 SHARADA SIGN UPADHMANIYA. These two signs have the properties of a letter and appear only in stacked conjuncts without the use of *virama*. *Jihvamuliya* is used to represent the velar fricative [x] in the context of following voiceless velar stops:

U+111C2 𑀛 *jihvamuliya* + U+11191 𑀅 *ka* → 𑀛𑀅

U+111C2 𑀛 *jihvamuliya* + U+11192 𑀆 *kha* → 𑀛𑀆

Upadhmaniya is used to represent the bilabial fricative [ɸ] in the context of following voiceless labial stops:

U+111C3 𑀜 *upadhmaniya* + U+111A5 𑀅 *pa* → 𑀜𑀅

U+111C3 𑀜 *upadhmaniya* + U+111A6 𑀆 *pha* → 𑀜𑀆

Punctuation. U+111C7 𑀟 SHARADA ABBREVIATION SIGN appears after letters or combinations of letters. It marks the sequence as an abbreviation. A word separator, U+111C8 𑀠 SHARADA SEPARATOR, indicates word and other boundaries. Sharada also makes use of two script-specific dandas: U+111C5 𑀡 SHARADA DANDA and U+111C6 𑀢 SHARADA DOUBLE DANDA.

Digits. Sharada has a distinctive set of digits encoded in the range U+111D0..U+111D9.

10.10 Takri

Takri is a script used in northern India and surrounding countries in South Asia, including the areas that comprise present-day Jammu and Kashmir, Himachal Pradesh, Punjab, and Uttarakhand. It is the traditional writing system for the Chambeali and Dogri languages, as well as several “Pahari” languages, such as Jaunsari, Kulvi, and Mandeali. It is related to the Gurmukhi, Landa, and Sharada scripts. Like other Brahmi-derived scripts, Takri is an *abugida*, with consonants taking an inherent vowel unless accompanied by a vowel marker or the *virama* (vowel killer).

Takri is descended from Sharada through an intermediate form known as Devāṣeṣa, which emerged in the 14th century. Devāṣeṣa was a script used for religious and official purposes, while its popular form, known as Takri, was used for commercial and informal purposes. Takri became differentiated from Devāṣeṣa during the 16th century. In its various regional manifestations, Takri served as the official script of several princely states of northern and northwestern India from the 17th century until the middle of the 20th century. Until the late 19th century, Takri was used concurrently with Devanagari, but it was gradually replaced by the latter.

Owing to its use as both an official and a popular script, Takri appears in numerous records, from manuscripts to inscriptions to postage stamps. There are efforts to revive the use of Takri for languages such as Dogri, Kishtwari, and Kulvi as a means of preserving access to these language’s literatures.

There is no universal, standard form of Takri. Where Takri was standardized, the reformed script was limited to a particular polity, such as a kingdom or a princely state. The representative glyphs shown in the code charts are taken mainly from the forms used in a variant established as the official script for writing the Chambeali language in the former Chamba State, now in Himachal Pradesh, India. There are a number of other regional varieties of Takri that have varying letter forms, sometimes quite different from the representative forms shown in the code charts. Such regional forms are considered glyphic variants and should be handled at the font level.

Vowel Letters. Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 10-7* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 10-7. Takri Vowel Letters

| For | Use | Do Not Use |
|-----|-------|----------------|
| 𑆚 | 11681 | <11680, 116AD> |
| 𑆛 | 11687 | <11686, 116B2> |
| 𑆜 | 11688 | <11680, 116B4> |
| 𑆝 | 11689 | <11680, 116B5> |

Consonant Conjuncts. Conjuncts in Takri are infrequent and, when written, consist of two consonants, the second of which is always *ya*, *ra*, or *ha*. Takri *ya* is written as a subjoining form; Takri *ra* can be written as a ligature or a subjoining form; and Takri *ha* is written as a half-form.

Nukta. A combining *nukta* character is encoded as U+116B7 TAKRI SIGN NUKTA. Characters that use this sound, mainly loan words and words from other languages, may be represented using the base character plus *nukta*.

Headlines. Unlike Devanagari, headlines are not generally used in Takri. However, headlines do appear in the glyph shapes of certain Takri letters. The headline is an intrinsic feature of glyph shapes in some regional varieties such as Dogra Akkhar, where it appears to be inspired by the design of Devanagari characters. There are no fixed rules for the joining of headlines. For example, the headlines of two sequential characters possessing headlines are left unjoined in Chambeali, while the headlines of a letter and a vowel sign are joined in printed Dogra Akkhar.

Punctuation. Takri uses U+0964 DEVANAGARI DANDA and U+0965 DEVANAGARI DOUBLE DANDA from Devanagari.

Fractions. Fraction signs and currency marks found in Takri documents use the characters in the Common Indic Number Forms block (U+A830..U+A83F).

10.11 Chakma

The Chakma people, who live in southeast Bangladesh near Chittagong City, as well as in parts of India such as Mizoram, Assam, Tripura, and Arunachal Pradesh, speak an Indo-European language also called Chakma. The language, spoken by about 500,000 people, is related to the Assamese, Bengali, Chittagonian, and Sylheti languages.

The Chakma script is Brahmi-derived, and is sometimes also called *Ajhā pāṭh* or *Ojhopath*. There are some efforts to adapt the Chakma script to write the closely related Tanchangya language.

One of the interesting features of Chakma writing is that *candrabindu* (*cānaphupudā*) can be used together with *anusvara* (*ekaphudā*) and *visarga* (*dviphudā*).

Independent Vowels. Like other Brahmi-derived scripts, Chakma uses consonant letters that contain an inherent vowel. Consonant clusters are written with conjunct characters, while a visible “vowel killer” (called the *maayyaa*) shows the deletion of the inherent vowel when there is no conjunct. There are four independent vowels in the script: U+11103 CHAKMA LETTER AA /ā/, U+11104 CHAKMA LETTER I /i/, U+11105 CHAKMA LETTER U /u/, and U+11106 CHAKMA LETTER E /e/. Other vowels in the initial position are formed by adding a dependent vowel sign to the independent vowel /ā/, to form vowels such as /ī/, /ō/, /ai/, and /oi/.

Vowel Killer and Virama. Like the Myanmar script and the characters used to write historic Meetei Mayek, Chakma is encoded with two vowel-killing characters to conform to modern user expectations. Chakma uses the *maayyaa* (killer) to invoke conjoined consonants. Most letters have their vowels killed with the use of the explicit *maayyaa* character. In addition to the visible killer, there is an explicit conjunct-forming character (*virama*), permitting the user to choose between the subjoining style and the ligating style. Whether a conjunct is required or not is part of the spelling of a word.

In principle, nothing prevents the visible killer from appearing together with a subjoining sequence formed with *virama*. However, in practice, combinations of *virama* and *maayyaa* following a consonant are not meaningful, as both kill the inherent vowel.

In 2001, an orthographic reform was recommended in the book *Cānmā pattham pāt*, limiting the standard repertoire of conjuncts to those composed with the five letters U+11121 CHAKMA LETTER YAA /yā/, U+11122 CHAKMA LETTER RAA /rā/, U+11123 CHAKMA LETTER LAA /lā/, U+11124 CHAKMA LETTER WAA /wā/, and U+1111A CHAKMA LETTER NAA /nā/.

Chakma Fonts. Chakma fonts by default should display the subjoined form of letters that follow virama to ensure legibility.

Punctuation. Chakma has a single and double danda. There is also a unique question mark and a section mark, *phulacihna*.

Digits. A distinct set of digits is encoded for Chakma. Bengali digits are also used with Chakma. Myanmar digits are used with the Chakma script when writing Tanchangya.



10.12 Meetei Mayek

Meetei Mayek: U+ABC0–U+ABFF

Meetei Mayek is a script used for Meetei, a Tibeto-Burman language spoken primarily in Manipur, India. The script originates from the Tibetan group of scripts, which in turn derive from Gupta Brahmi. The script has experienced a recent resurgence in use. The modern-day Meetei Mayek script is made up of a core repertoire of 27 letters, alongside letters and symbols for final consonants, dependent vowel signs, punctuation, and digits.

The name “Meetei Mayek” is used in official documentation in Manipur. The script may also appear with other spellings and names, such as “Meitei Mayek,” “Methei,” “Meetei,” or the older “Manipuri.”

Structure. Meetei Mayek is a Brahmic script with consonants bearing the inherent vowel and vowel matras modifying it. However, unlike most other Brahmi-derived scripts, Meetei Mayek employs explicit final consonants which contain no final vowels.

Meetei Mayek has a killer character, U+ABED MEETEI MAYEK APUN IYEK, which may be used to indicate the lack of an inherent vowel when no explicit consonant letter exists. In modern orthography, the killer does not cause conjunct formation and is always visible. The use of the killer is optional in spelling; for example, while  may be read *kara* or *kra*,  must be read *kra*. In the medial position, the glyph of the killer usually extends below the killed letter and the following letter.

Vowel Letters. In modern use, only three vowel characters, U+ABD1 MEETEI MAYEK LETTER ATIYA, U+ABCF MEETEI MAYEK LETTER I, and U+ABCE MEETEI MAYEK LETTER UN (= *u*), may appear initially or word-internally. Other vowels without independent forms are represented by vowel matras applied to U+ABD1 MEETEI MAYEK LETTER ATIYA. In modern orthography, the seven dependent vowel signs and the *anusvara*, U+ABEA MEETEI MAYEK VOWEL SIGN NUNG, located from U+ABE3..U+ABEA, are used with consonants.

Syllable initial combinations for vowels can occur in modern usage to represent diphthongs.

Final Consonants. There are three ways to indicate final consonants in Meetei Mayek: by the eight explicit final consonant letters, by U+ABEA MEETEI MAYEK VOWEL SIGN NUNG, which acts as an *anusvara*, or by U+ABCE MEETEI MAYEK LETTER UN, which may act as a final consonant without modification.

Abbreviations. Unusual abbreviations composed of a single consonant and more than one matra may occur in a manner similar that found in Tibetan. In such cases, the vowel matra may occur at the end of a word.

Order. The order of the first 18 Meetei letters is based upon the parts of the body. This system is discussed in a religious manuscript, the *Wakoklon hilel thilel salai amailon pukok puya* (commonly referred to as the *Wakoklon puya*), which describes the letters, and relates them to the corresponding body part. The Meetei Mayek letter *kok*, for example, means “head,” *sam* designates “hair-parting,” and *lai* is “forehead.” The last 9 letters, *gok*, *jham*, *rai*, and so forth, derive from a subset of the original 18. The ordering system employed today differs from the Brahmi-based order, which relies on the point of articulation.

Punctuation. The modern Meetei Mayek script uses two punctuation marks in addition to the killer. U+ABEB MEETEI MAYEK CHEIKHEI functions as a double danda mark. U+ABEC MEETEI MAYEK LUM IYEK is a heavy tone mark, used to orthographically distinguish words which would otherwise not be differentiated.

Digits. Meetei Mayek has a unique set of ten digits for zero to nine encoded in the range at U+ABF0..U+ABF9.

Meetei Mayak Extensions: U+AAE0–U+AAF6

The Meetei Mayak Extensions block contains additional characters needed to represent the historical orthographies of Meetei. The block includes nine consonants, encoded in the range U+AAE2..U+AAEA, two independent vowel signs (U+AAE0 MEETEI MAYEK LETTER E and U+AAE1 MEETEI MAYEK LETTER O), and five dependent vowels signs in the range U+AAEB..U+AAEF.

U+AAF5 MEETEI MAYEK VIRAMA should be used to represent conjuncts that may occur in historical texts. The *virama* is not visibly rendered, but it behaves as in other Brahmic-derived scripts. For example, the conjunct /ñha/ is represented by the sequence <ABC9, AAF5, ABCD>.

This block also includes two punctuation marks, U+AAF0 MEETEI MAYEK CHEIKHAN and U+AAF1 MEETEI MAYEK AHANG KHUDAM. The *cheikhan* is a single *danda*, and *ahang khudam* is a question mark. U+AAF2 MEETEI MAYEK ANJI is a philosophical sign indicating auspiciousness. Finally, two repetition marks are included in the block: U+AAF3 MEETEI MAYEK SYLLABLE REPETITION MARK and U+AAF4 MEETEI MAYEK WORD REPRESENTATION MARK.

10.13 Ol Chiki

Ol Chiki: U+1C50–U+1C7F

The Ol Chiki script was invented by Pandit Raghunath Murmu in the first half of the 20th century CE to write Santali, a Munda language of India. The script is also called Ol Cemet, Ol Ciki, or simply Ol. Santali has also been written with the Devanagari, Bengali, and Oriya scripts, as well as the Latin alphabet.

Various dialects of Santali are spoken by 5.8 million people, with 25% to 50% literacy rates, mostly in India, with a few in Nepal or Bangladesh. The Ol Chiki script is used primarily for the southern dialect of Santali as spoken in the Orissan Mayurbhañj district. The script has received some official recognition by the Orissan government.

Ol Chiki has recently been promoted by some Santal organizations, with uncertain success, for use in writing certain other Munda languages in the Chota Nagpur area, as well as for the Dravidian Dhangar-Kudux language.

Structure. Ol Chiki is alphabetic and has none of the structural properties of the abugidas typical for other Indic scripts. There are separate letters representing consonants and vowels. A number of modifier letters are used to indicate tone, nasalization, vowel length, and deglottalization. There are no combining characters in the script.

Ol Chiki is written from left to right.

Digits. The Ol Chiki script has its own set of digits. These are separately encoded in the Ol Chiki block.

Punctuation. Western-style punctuation, such as the comma, exclamation mark, question mark, and quotation marks are used in Ol Chiki text. U+002E “.” FULL STOP is not used, because it is visually confusable with the modifier letter U+1C79 OL CHIKI GAAHLAA TTUDDAAG.

The *danda*, U+1C7E OL CHIKI PUNCTUATION MUCAAD, is used as a text delimiter in prose. The *danda* and the *double danda*, U+1C7F OL CHIKI PUNCTUATION DOUBLE MUCAAD, are both used in poetic text.

Modifier Letters. The southern dialect of Santali has only six vowels, each represented by a single vowel letter. The Santal Parganas dialect, on the other hand, has eight or nine vowels. The extra vowels for Santal Parganas are represented by a sequence of one of the vowel letters U+1C5A, U+1C5E, or U+1C6E followed by the diacritic modifier letter, U+1C79 OL CHIKI GAAHLAA TTUDDAAG, displayed as a baseline dot.

Nasalization is indicated by the modifier letter, U+1C78 OL CHIKI MU TTUDDAG, displayed as a raised dot. This mark can follow any vowel, long or short.

When the vowel diacritic and nasalization occur together, the combination is represented by a separate modifier letter, U+1C7A OL CHIKI MU-GAHLAA TTUDDAAG, displayed as both a baseline and a raised dot. The combination is treated as a separate character and is entered using a separate key on Ol Chiki keyboards.

U+1C7B OL CHIKI RELAA is a length mark, which can be used with any oral or nasalized vowel.

Glottalization. U+1C7D OL CHIKI AHAD is a special letter indicating the deglottalization of an Ol Chiki consonant in final position. This unique feature of the writing system preserves the morphophonemic relationship between the glottalized (ejective) and voiced equivalents of consonants. For example, U+1C5C OL CHIKI LETTER AG represents an ejective [kʰ] when written in word-final position, but voiced [g] when written word-initially. A voiced [g] in word-final position is written with the deglottalization mark as a sequence: <U+1C5C OL CHIKI LETTER AG, U+1C7D OL CHIKI AHAD>.

U+1C7C OL CHIKI PHAARKAA serves the opposite function. It is a “glottal protector.” When it follows one of the four ejective consonants, it preserves the ejective sound, even in word-initial position followed by a vowel.

Aspiration. Aspirated consonants are written as digraphs, with U+1C77 OL CHIKI LETTER OH as the second element, indicating the aspiration.

Ligatures. Ligatures are not a normal feature of printed Ol Chiki. However, in handwriting and script fonts, letters form cursive ligatures with the deglottalization mark, U+1C7D OL CHIKI AHAD.

10.14 Sora Sompeng

Sora Sompeng: U+110D0–U+110FF

The Sora Sompeng script is used to write the Sora language. Sora is a member of the Munda family of languages, which, together with the Mon-Khmer languages, makes up Austro-Asiatic.

The Sora people live between the Oriya- and Telugu-speaking populations in what is now the Orissa-Andhra border area.

Sora Sompeng was devised in 1936 by Mangei Gomango, who was inspired by the vision he had of the 24 letters. The script was promulgated as part of a comprehensive cultural program, and was offered as an improvement over IPA-based scripts used by linguists and missionaries, and the Telugu and Oriya scripts used by Hindus. Sora Sompeng is used in religious contexts, and is published in a variety of printed materials.

Encoding Structure. The Sora Sompeng script is structured as an abugida. The consonant letters contain an inherent vowel. There are no conjunct characters for consonant clusters, and there is no visible vowel killer to show the deletion of the inherent vowel. The reader must determine the presence or absence of the inherent schwa based on recognition of each word. The character repertoire does not match the phonemic repertoire of Sora very well.

U+110E4 SORA SOMPENG LETTER IH is used for both [i] and [ɨ], and U+110E6 SORA SOMPENG LETTER OH is used for both [o] and [ɔ], for instance. The glottal stop is written with U+110DE SORA SOMPENG LETTER HAH, and the sequence of U+110DD SORA SOMPENG LETTER RAH and U+110D4 SORA SOMPENG LETTER DAH is used to write retroflex [ɽ]. There is also an additional “auxiliary” U+110E8 SORA SOMPENG LETTER MAE used to transcribe foreign sounds.

Character Names. Consonant letter names for Sora Sompeng are derived by adding [aʔa] (written *ah*) to the consonant.

Punctuation. Sora Sompeng uses Western-style punctuation.

Linebreaking. Letters and digits behave as in Latin and other alphabetic scripts.

10.15 Kharoshthi

Kharoshthi: U+10A00–U+10A5F

The Kharoshthi script, properly spelled as Kharoṣṭhī, was used historically to write Gāndhārī and Sanskrit as well as various mixed dialects. Kharoshthi is an Indic script of the *abugida* type. However, unlike other Indic scripts, it is written from right to left. The Kharoshthi script was initially deciphered around the middle of the nineteenth century by James Prinsep and others who worked from short Greek and Kharoshthi inscriptions on the coins of the Indo-Greek and Indo-Scythian kings. The decipherment has been refined over the last 150 years as more material has come to light.

The Kharoshthi script is one of the two ancient writing systems of India. Unlike the pan-Indian Brāhmī script, Kharoshthi was confined to the northwest of India centered on the region of *Gandhāra* (modern northern Pakistan and eastern Afghanistan, as shown in *Figure 10-5*). Gandhara proper is shown on the map as the dark gray area near Peshawar. The lighter gray areas represent places where the Kharoshthi script was used and where manuscripts and inscriptions have been found.

The exact details of the origin of the Kharoshthi script remain obscure, but it is almost certainly related to Aramaic. The Kharoshthi script first appears in a fully developed form in the Aśokan inscriptions at Shahbazgarhi and Mansehra which have been dated to around 250 BCE. The script continued to be used in Gandhara and neighboring regions, sometimes alongside Brahmi, until around the third century CE, when it disappeared from its homeland. Kharoshthi was also used for official documents and epigraphs in the Central Asian cities of Khotan and Niya in the third and fourth centuries CE, and it appears to have survived in Kucha and neighboring areas along the Northern Silk Road until the seventh century. The Central Asian form of the script used during these later centuries is termed *Formal Kharoshthi* and was used to write both Gandhari and Tocharian B. Representation of Kharoshthi in the Unicode code charts uses forms based on manuscripts of the first century CE.

Figure 10-5. Geographical Extent of the Kharoshthi Script



Directionality. Kharoshthi can be implemented using the rules of the Unicode Bidirectional Algorithm. Both letters and digits are written from right to left. Kharoshthi letters do not have positional variants.

Combining Vowels. The various combining vowels attach to characters in different ways. A number of groupings have been determined on the basis of their visual types, such as horizontal or vertical, as shown in *Table 10-8*.

Table 10-8. Kharoshthi Vowel Signs

| Type | Example | Group Members |
|-----------------------------|-----------------------------|---|
| Vowel sign i | | |
| Horizontal | a + -i → i 𑀓 + 𑀓 → 𑀓 | A, NA, HA |
| Vertical | tha + -i → thi 𑀓 + 𑀓 → 𑀓 | THA, PA, PHA, MA, LA, SHA |
| Diagonal | ka + -i → ki 𑀓 + 𑀓 → 𑀓 | All other letters |
| Vowel sign u | | |
| Independent | ha + -u → hu 𑀓 + 𑀓 → 𑀓 | TTA, HA |
| Ligated | ma + -u → mu 𑀓 + 𑀓 → 𑀓 | MA |
| Attached | a + -u → u 𑀓 + 𑀓 → 𑀓 | All other letters |
| Vowel sign vocalic r | | |
| Attached | a + -r → r 𑀓 + 𑀓 → 𑀓 | A, KA, KKA, KHA, GA, GHA, CA, CHA, JA, TA, DA, DHA, NA, PA, PHA, BA, BHA, VA, SHA, SA |
| Independent | ma + -r → mr 𑀓 + 𑀓 → 𑀓 | MA, HA |
| Vowel sign e | | |
| Horizontal | a + -e → e 𑀓 + 𑀓 → 𑀓 | A, NA, HA |
| Vertical | tha + -e → the 𑀓 + 𑀓 → 𑀓 | THA, PA, PHA, LA, SSA |
| Ligated | da + -e → de 𑀓 + 𑀓 → 𑀓 | DA, MA |
| Diagonal | ka + -e → ke 𑀓 + 𑀓 → 𑀓 | All other letters |
| Vowel sign o | | |
| Vertical | pa + -o → po 𑀓 + 𑀓 → 𑀓 | PA, PHA, YA, SHA |
| Diagonal | a + -o → o 𑀓 + 𑀓 → 𑀓 | All other letters |

Combining Vowel Modifiers. U+10A0C 𑀓 K HAROSHTHI VOWEL LENGTH MARK indicates equivalent long vowels and, when used in combination with -e and -o, indicates the diphthongs -ai and -au. U+10A0D 𑀓 K HAROSHTHI SIGN DOUBLE RING BELOW appears in some Central Asian documents, but its precise phonetic value has not yet been established. These two modifiers have been found only in manuscripts and inscriptions from the first century CE onward. U+10A0E 𑀓 K HAROSHTHI SIGN ANUSVARA indicates nasalization, and

U+10A0F ̄̄ KHAROSHTHI SIGN VISARGA is generally used to indicate unvoiced syllable-final [h], but has a secondary use as a vowel length marker. *Visarga* is found only in Sanskritized forms of the language and is not known to occur in a single *aksara* with *anusvara*. The modifiers and the vowels they modify are given in *Table 10-9*.

Table 10-9. Kharoshthi Vowel Modifiers

| Type | Example | Group Members |
|-------------------|-------------------------------|------------------|
| Vowel length mark | ma + ̄̄ → mā 𑖞 + ̄̄ → 𑖞̄̄ | A, I, U, R, E, O |
| Double ring below | sa + 𑖞 → sā 𑖞 + 𑖞 → 𑖞̄̄ | A, U |
| Anusvara | a + -ṃ → aṃ 𑖞 + 𑖞 → 𑖞̣ | A, I, U, R, E, O |
| Visarga | ka + -h → kaḥ 𑖞 + ̄̄ → 𑖞̄̄ | A, I, U, R, E, O |

Combining Consonant Modifiers. U+10A38 ̄ KHAROSHTHI SIGN BAR ABOVE indicates various modified pronunciations depending on the consonants involved, such as nasalization or aspiration. U+10A39 𑖞 KHAROSHTHI SIGN CAUDA indicates various modified pronunciations of consonants, particularly fricativization. The precise value of U+10A3A 𑖞 KHAROSHTHI SIGN DOT BELOW has not yet been determined. Usually only one consonant modifier can be applied to a single consonant. The resulting combined form may also combine with vowel diacritics, one of the vowel modifiers, or anusvara or visarga. The modifiers and the consonants they modify are given in *Table 10-10*.

Table 10-10. Kharoshthi Consonant Modifiers

| Type | Example | Group Members |
|-----------|---------------------------|---|
| Bar above | ja + ̄ → jā 𑖞 + ̄ → 𑖞̄ | GA, CA, JA, NA, MA, SHA, SSA, SA, HA |
| Cauda | ga + 𑖞 → gā 𑖞 + 𑖞 → 𑖞̣ | GA, JA, DDA, TA, DA, PA, YA, VA, SHA, SA |
| Dot below | ma + 𑖞 → mā 𑖞 + 𑖞 → 𑖞̣ | MA, HA |

Virama. The virama is used to indicate the suppression of the inherent vowel. The glyph for U+10A3F 𑖞 KHAROSHTHI VIRAMA shown in the code charts is arbitrary and is not actually rendered directly; the dotted box around the glyph indicates that special rendering is required. When not followed by a consonant, the virama causes the preceding consonant to be written as subscript to the left of the letter preceding it. If followed by another consonant, the virama will trigger a combined form consisting of two or more consonants. The resulting form may also be subject to combinations with the previously noted combining diacritics.

The virama can follow only a consonant or a consonant modifier. It cannot follow a space, a vowel, a vowel modifier, a number, a punctuation sign, or another virama. Examples of the use of the Kharoshthi virama are given in *Table 10-11*.

Table 10-11. Examples of Kharoshthi Virama

| Type | Example |
|---|---|
| Pure virama | $dha + i + k + \text{VIRAMA} \rightarrow dhik$ 𑀢 + 𑀣 + 𑀤 + 𑀥 → 𑀦 |
| Ligatures | $ka + \text{VIRAMA} + sa \rightarrow ksa$ 𑀤 + 𑀥 + 𑀧 → 𑀨 |
| Consonants with special combining forms | $sa + \text{VIRAMA} + ya \rightarrow sya$ 𑀧 + 𑀥 + 𑀩 → 𑀪 |
| Consonants with full combined form | $ka + \text{VIRAMA} + ta \rightarrow kta$ 𑀤 + 𑀥 + 𑀫 → 𑀬 |

10.16 Brahmi

Brahmi: U+11000–U+1106F

The Brahmi script is an historical script of India attested from the third century BCE until the late first millennium CE. Over the centuries Brahmi developed many regional varieties, which ultimately became the modern Indian writing systems, including Devanagari, Tamil and so on. The encoding of the Brahmi script in the Unicode Standard supports the representation of texts in Indian languages from this historical period. For texts written in historically transitional scripts—that is, between Brahmi and its modern derivatives—there may be alternative choices to represent the text. In some cases, there may be a separate encoding for a regional medieval script, whose use would be appropriate. In other cases, users should consider whether the use of Brahmi or a particular modern script best suits their needs.

Encoding Model. The Brahmi script is an *abugida* and is encoded using the Unicode *virama* model. Consonants have an inherent vowel /a/. A separate character is encoded for the virama: U+11046 BRAHMI VIRAMA. The *virama* is used between consonants to form conjunct consonants. It is also used as an explicit killer to indicate a vowelless consonant.

Vowel Letters. Vowel letters are encoded atomically in Brahmi, even if they can be analyzed visually as consisting of multiple parts. Table 10-12 shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

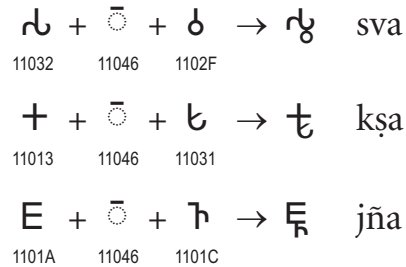
Table 10-12. Brahmi Vowel Letters

| To Represent | Use | Do Not Use |
|--------------|-------|----------------|
| 𑀢 | 11006 | <11005, 11038> |
| 𑀣 | 1100C | <1100B, 1103E> |
| 𑀤 | 11010 | <1100F, 11042> |

Rendering Behavior. Consonant conjuncts are represented by a sequence including virama: <C, virama, C>. In Brahmi these consonant conjuncts are rendered as consonant ligatures. Up to a very late date, Brahmi used vertical conjuncts exclusively, in which the ligation involves stacking of the consonant glyphs vertically. The Brahmi script does not have a parallel series of half-consonants, as developed in Devanagari and some other modern Indic scripts.

The elements of consonant ligatures are laid out from top left to bottom right, as shown for *sva* in *Figure 10-8*. Preconsonantal *r*, postconsonantal *r* and postconsonantal *y* assume special reduced shapes in all except the earliest varieties of Brahmi. The *kṣa* and *jña* ligatures, however, are often transparent, as also shown in *Figure 10-8*.

Figure 10-8. Consonant Ligatures in Brahmi



A vowelless consonant is represented in text by following the consonant with a *virama*: <C, virama>. The presence of the *virama* “kills” the vowel. Such vowelless consonants have visible distinctions from regular consonants, and are rendered in one of two major styles. In the first style, the vowelless consonant is written smaller and lower than regular consonants, and often has a connecting line drawn from the vowelless consonant to the preceding *aksara*. In the second style, a horizontal line is drawn above the vowelless consonant. The second style is the basis for the representative glyph for U+10146 BRAHMI VIRAMA in the code charts. These differences in presentation are purely stylistic; it is up to the font developers and rendering systems to render Brahmi vowelless consonants in the appropriate style.

Vowel Modifiers. U+11000 BRAHMI SIGN CANDRABINDU indicates nasalization of a vowel. U+11001 BRAHMI SIGN ANUSVARA is used to indicate that a vowel is nasalized (when the next syllable starts with a fricative), or that it is followed by a nasal segment (when the next syllable starts with a stop). U+11002 BRAHMI SIGN VISARGA is used to write syllable-final voiceless /h/; that is, [x] and [f]. The velar and labial allophones of /h/, followed by voiceless velar and labial stops respectively, are sometimes written with separate signs U+11003 BRAHMI SIGN JIHVAMULIYA and U+11004 BRAHMI SIGN UPADHMANIYA. Unlike *visarga*, these two signs have the properties of a letter, and are not considered combining marks. They enter into ligatures with the following homorganic voiceless stop consonant, without the use of a *virama*.

Old Tamil Brahmi. Brahmi was used to write the Tamil language starting from the second century BCE. The different orthographies used to write Tamil Brahmi are covered by the Unicode encoding of Brahmi. For example, in one Tamil Brahmi system the inherent vowel of Brahmi consonant signs is dropped, and U+11038 BRAHMI VOWEL SIGN AA is used to represent both short and long [a] / [a:]. In this orthography consonant signs without a vowel sign always represent the bare consonant without an inherent vowel. Three consonant letters are encoded to represent sounds particular to Dravidian. These are U+11035 BRAHMI LETTER OLD TAMIL LLLA, U+11036 BRAHMI LETTER OLD TAMIL RRA, and U+11037 BRAHMI LETTER OLD TAMIL NNA.

Tamil Brahmi *pulli* (*virama*) had two functions: to cancel the inherent vowel of consonants; and to indicate the short vowels [e] and [o] in contrast to the long vowels [e:] and [o:] in Prakrit and Sanskrit. As a consequence, in Tamil Brahmi text, the *virama* is used not only after consonants, but also after the vowels *e* (U+1100E, U+11042) and *o* (U+11011, U+11044). This *pulli* is represented using U+11046 BRAHMI SIGN VIRAMA.

Bhattiprolu Brahmi. Ten short Middle Indo-Aryan inscriptions from the second century BCE found at Bhattiprolu in Andhra Pradesh show an orthography that seems to be derived from the Tamil Brahmi system. To avoid the phonetic ambiguity of the Tamil Brahmi U+11038 BRAHMI VOWEL SIGN AA (standing for either [a] or [a:]), the Bhattiprolu inscriptions introduced a separate vowel sign for long [a:] by adding a vertical stroke to the end of the earlier sign. This is encoded as U+11039 BRAHMI VOWEL SIGN BHATTIPROLU AA.

Punctuation. There are seven punctuation marks in the encoded repertoire for Brahmi. The single and double dandas, U+11047 BRAHMI DANDA and U+11048 BRAHMI DOUBLE DANDA, delimit clauses and verses. U+11049 BRAHMI PUNCTUATION DOT, U+1104A BRAHMI PUNCTUATION DOUBLE DOT, and U+1104B BRAHMI PUNCTUATION LINE delimit smaller textual units, while U+1104C BRAHMI PUNCTUATION CRESCENT BAR and U+1104D BRAHMI PUNCTUATION LOTUS separate larger textual units.

Numerals. Two sets of numbers, used for different numbering systems, are attested in Brahmi documents. The first set is the old additive-multiplicative system that goes back to the beginning of the Brahmi script. The second is a set of decimal numbers that occurs side by side with the earlier numbering system in manuscripts and inscriptions during the late Brahmi period.

The set of additive-multiplicative numbers of the Brahmi script contains separate number signs for the digits from 1 to 9, the decades from 10 to 90, as well as signs for 100 and 1000. Numbers are written additively, with higher number signs preceding lower ones. Multiples of 100 and of 1000 are expressed multiplicatively, with the multiplier following and forming a ligature with 100 or 1000. There are examples from the middle and late Brahmi periods in which the signs for 200, 300, and 2000 appear in special forms and are not obviously connected with a ligature of the component parts. Such forms may be enabled in fonts using a ligature substitution.

A special sign for zero was invented later, and the positional system came into use. This system is the ancestor of the modern decimal number system. Due to the different systemic features and shapes, the signs in this set have been encoding separately. These signs have the same properties as the modern Indian digits. Examples are shown in *Table 10-13*.

Table 10-13. Brahmi Positional Digits

| Display | Value | Code Points |
|---------|-------|-----------------------|
| · | 0 | 11066 |
| ↘ | 1 | 11067 |
| २ | 2 | 11068 |
| ३ | 3 | 11069 |
| ४ | 4 | 1106A |
| ↘· | 10 | <11067, 11066> |
| २३४ | 234 | <11066, 11069, 1106A> |

