The Unicode Standard
Version 8.0 – Core Specification

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

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Chapter 13

South and Central Asia-II

Other Modern Scripts

This chapter describes the following other modern scripts in South and Central Asia:

Thaana	*Limbu*	*Ol Chiki*
Sinhala	*Meetei Mayek*	*Chakma*
Tibetan	*Mro*	*Lepcha*
Mongolian	*Warang Citi*	*Saurashtra*

The Thaana script is used to write Dhivehi, the language of the Republic of Maldives, an island nation in the middle of the Indian Ocean.

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

The Mongolian script was derived from the Uighur script around the beginning of the thirteenth century, during the reign of Genghis Khan. It is used in both China and Mongolia.

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.

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.

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.

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.

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.

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.

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 Odisha (Orissa).

Mro is a Tibeto-Burman language spoken primarily in Bangladesh. The Mro script is a left-to-right alphabet used to write the Mro language. It was invented in the 1980s and is unrelated to existing scripts.

The Warang Citi script is a recently devised left-to-right alphabet. The script is used to write the Ho language, a North Munda language which has an emergent literary tradition. The Ho people live in eastern India.

13.1 Thaana

Thaana: U+0780–U+07BF

The Thaana script is used to write the modern Dhivehi language of the Republic of Maldives, a group of atolls in the Indian Ocean. Like the Arabic script, Thaana is written from right to left and uses vowel signs, but it is not cursive. The basic Thaana letters have been extended by a small set of dotted letters used to transcribe Arabic. The use of modified Thaana letters to write Arabic began in the middle of the 20th century. Loan words from Arabic may be written in the Arabic script, although this custom is not very prevalent today. (See *Section 9.2, Arabic.*)

While Thaana's glyphs were borrowed in part from Arabic (letters *haa* through *vaavu* were based on the Arabic-Indic digits, for example), and while vowels and *sukun* are marked with combining characters as in Arabic, Thaana is properly considered an alphabet, rather than an abjad, because writing the vowels is obligatory.

Directionality. The Thaana script is written from right to left. Conformant implementations of Thaana script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, "Unicode Bidirectional Algorithm").

Vowels. Consonants are always written with either a vowel sign (U+07A6..U+07AF) or the null vowel sign (U+07B0 THAANA SUKUN). U+0787 THAANA LETTER ALIFU with the null vowel sign denotes a glottal stop. The placement of the Thaana vowel signs is shown in *Table 13-1*.

Table 13-1. Thaana Glyph Placement

Syllable	Display
tha	◌ަ
thaa	◌ާ
thi	◌ި
thee	◌ީ
thu	◌ު
thoo	◌ޫ
the	◌ެ
they	◌ޭ
tho	◌ޮ
thoa	◌ޯ
th	◌ް

Numerals. Both European (U+0030..U+0039) and Arabic digits (U+0660..U+0669) are used. European numbers are used more commonly and have left-to-right display direc-

tionality in Thaana. Arabic numeric punctuation is used with digits, whether Arabic or European.

Punctuation. The Thaana script uses spaces between words. It makes use of a mixture of Arabic and European punctuation, though rules of usage are not clearly defined. Sentence-final punctuation is now generally shown with a single period (U+002E "." FULL STOP) but may also use a sequence of two periods (U+002E followed by U+002E). Phrases may be separated with a comma (usually U+060C ARABIC COMMA) or with a single period (U+002E). Colons, dashes, and double quotation marks are also used in the Thaana script. In addition, Thaana makes use of U+061F ARABIC QUESTION MARK and U+061B ARABIC SEMICOLON.

Character Names and Arrangement. The character names are based on the names used in the Republic of Maldives. The character name at U+0794, *yaa,* is found in some sources as *yaviyani*, but the former name is more common today. Characters are listed in Thaana alphabetical order from *haa* to *ttaa* for the Thaana letters, followed by the extended characters in Arabic alphabetical order from *hhaa* to *waavu.*

13.2 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, අඬ <U+0D85, U+0DAC> *aňḍa* [aⁿḍa] "sound" versus අණ්ඩ <U+0D85, U+0DAB, U+0DCA, U+0DA9> *aṇḍa* [aṇḍa] "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 13-2* 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 13-2. Sinhala Vowel Letters

To Represent	Use	Do Not Use
අා	0D86	<0D85, 0DCF>
අැ	0D87	<0D85, 0DD0>
අෑ	0D88	<0D85, 0DD1>
උෟ	0D8C	<0D8B, 0DDF>
ඍෘ	0D8E	<0D8D, 0DD8>
ඏෟ	0D90	<0D8F, 0DDF>
එ්	0D92	<0D91, 0DCA>
එේ	0D93	<0D91, 0DD9>
ඔෟ	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.

Punctuation. Sinhala currently uses Western-style punctuation marks. U+0DF4 SINHALA PUNCTUATION KUNDDALIYA was used historically as a full stop. U+0964 DEVANAGARI DANDA is used to represent the dandas which occur in historic Sanskrit or Pali texts written in the Sinhala script.

Digits. Modern Sinhala text uses Western digits. The set of digits in the range U+0DE6 to U+0DEF was used into the twentieth century, primarily to write horoscopes. That set of astrological digits is known as Sinhala Lith Illakkam, and includes a form for zero.

Sinhala Archaic Numbers: U+111E0–U+111FF

The Sinhala Archaic Numbers block contains characters used in a historic number system called Sinhala Illakkam, which was in use prior to 1815. Sinhala Illakkam was not a positional notation, and lacks a digit for zero. It is distinct from the set of Sinhala astrological digits called Sinhala Lith Illakkam (U+0DE6..U+0DEF).

13.3 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 Wartu script. Numerous styles of *u-mey* have evolved since then, including both formal calligraphic styles used in manuscripts and running handwriting styles. All Tibetan scripts follow the same lettering rules, though there is a slight difference in the way that certain compound stacks are formed in *uchen* and *u-mey*.

General Principles of the Tibetan Script. Tibetan grammar divides letters into consonants and vowels. There are 30 consonants, and each consonant is represented by a discrete written character. There are five vowel sounds, only four of which are represented by written marks. The four vowels that are explicitly represented in writing are each represented with a single mark that is applied above or below a consonant to indicate the application of that vowel to that consonant. The absence of one of the four marks implies that the first vowel

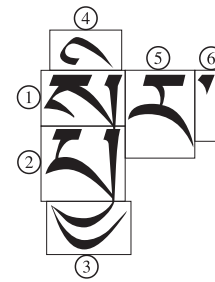
sound (like a short "ah" in English) is present and is not modified to one of the four other possibilities. Three of the four marks are written above the consonants; one is written below.

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

For example, the syllable *spyir* "general," pronounced [tʃiː], is a typical example of a Tibetan syllable that includes a stack comprising a head letter, two subscript letters, and a vowel sign. *Figure 13-1* shows the characters in the order in which they appear in the backing store.

Figure 13-1. 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



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

opers 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+0FBE and U+0FCF.

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

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

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 used in normal Tibetan. Precomposed forms of aspirate letters (and the conjunct "kssa") are explicitly coded, along with their corresponding subjoined forms: for example, U+0F43 TIBETAN LETTER GHA, and U+0F93 TIBETAN SUBJOINED LETTER GHA, or U+0F69 TIBETAN LETTER KSSA, and U+0FB9 TIBETAN SUBJOINED LETTER KSSA. However, these characters, including the subjoined forms, decompose to stacked sequences involving subjoined "ha" (or "reversed sha") in all Unicode normalization forms.

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 12.1, 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.

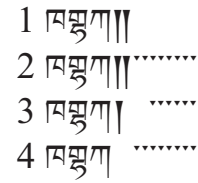
U+0F3A TIBETAN MARK GUG RTAGS GYON and U+0F3B TIBETAN MARK GUG RTAGS GYAS are paired punctuation marks (brackets).

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 13-2* is not usually padded as in the second line of *Figure 13-2*, 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 13-2* 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 13-2* is not acceptable.

Figure 13-2. Justifying Tibetan Tseks



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 breaks 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 arrange-

ment. These stacks are highly unusual and are considered beyond the scope of plain text rendering. They may be handled by higher-level mechanisms.

13.4 Mongolian

Mongolian: U+1800–U+18AF

The Mongolians are key representatives of a cultural-linguistic group known as Altaic, after the Altai mountains of central Asia. In the past, these peoples have dominated the vast expanses of Asia and beyond, from the Baltic to the Sea of Japan. Echoes of Altaic languages remain from Finland, Hungary, and Turkey, across central Asia, to Korea and Japan. Today the Mongolians are represented politically in Mongolia proper (formally the Mongolian People's Republic, also known as Outer Mongolia) and Inner Mongolia (formally the Inner Mongolia Autonomous Region, China), with Mongolian populations also living in other areas of China.

The Mongolian block unifies Mongolian and the three derivative scripts Todo, Manchu, and Sibe. Each of the three derivative scripts shares some common letters with Mongolian, and these letters are encoded only once. Each derivative script also has a number of modified letterforms or new letters, which are encoded separately.

Mongolian, Todo, and Manchu also have a number of special "Ali Gali" letters that are used for transcribing Tibetan and Sanskrit in Buddhist texts.

History. The Mongolian script was derived from the Uighur script around the beginning of the thirteenth century, during the reign of Genghis Khan. The Uighur script, which was in use from about the eighth to the fifteenth centuries, was derived from Sogdian Aramaic, a Semitic script written horizontally from right to left. Probably under the influence of the Chinese script, the Uighur script became rotated 90 degrees counterclockwise so that the lines of text read vertically in columns running from left to right. The Mongolian script inherited this directionality from the Uighur script.

The Mongolian script has remained in continuous use for writing Mongolian within the Inner Mongolia Autonomous Region of the People's Republic of China and elsewhere in China. However, in the Mongolian People's Republic (present-day Mongolia), the traditional script was replaced by a Cyrillic orthography in the early 1940s. The traditional script was revived in the early 1990s, so that now both the Cyrillic and the Mongolian scripts are used. The spelling used with the traditional Mongolian script represents the literary language of the seventeenth and early eighteenth centuries, whereas the Cyrillic script is used to represent the modern, colloquial pronunciation of words. As a consequence, there is no one-to-one relationship between the traditional Mongolian orthography and Cyrillic orthography. Approximate correspondence mappings are indicated in the code charts, but are not necessarily unique in either direction. All of the Cyrillic characters needed to write Mongolian are included in the Cyrillic block of the Unicode Standard.

In addition to the traditional Mongolian script of Mongolia, several historical modifications and adaptations of the Mongolian script have emerged elsewhere. These adaptations are often referred to as scripts in their own right, although for the purposes of character encoding in the Unicode Standard they are treated as styles of the Mongolian script and share encoding of their basic letters.

The Todo script is a modified and improved version of the Mongolian script, devised in 1648 by Zaya Pandita for use by the Kalmyk Mongolians, who had migrated to Russia in the sixteenth century, and who now inhabit the Republic of Kalmykia in the Russian Federation. The name *Todo* means "clear" in Mongolian; it refers to the fact that the new script eliminates the ambiguities inherent in the original Mongolian script. The orthography of the Todo script also reflects the Oirat-Kalmyk dialects of Mongolian rather than literary Mongolian. In Kalmykia, the Todo script was replaced by a succession of Cyrillic and Latin orthographies from the mid-1920s and is no longer in active use. Until very recently the Todo script was still used by speakers of the Oirat and Kalmyk dialects within Xinjiang and Qinghai in China.

The Manchu script is an adaptation of the Mongolian script used to write Manchu, a Tungusic language that is not closely related to Mongolian. The Mongolian script was first adapted for writing Manchu in 1599 under the orders of the Manchu leader Nurhachi, but few examples of this early form of the Manchu script survive. In 1632, the Manchu scholar Dahai reformed the script by adding circles and dots to certain letters in an effort to distinguish their different sounds and by devising new letters to represent the sounds of the Chinese language. When the Manchu people conquered China to rule as the Qing dynasty (1644–1911), Manchu become the language of state. The ensuing systematic program of translation from Chinese created a large and important corpus of books written in Manchu. Over time the Manchu people became completely sinified, and as a spoken language Manchu is now almost extinct.

The Sibe (also spelled Sibo, Xibe, or Xibo) people are closely related to the Manchus, and their language is often classified as a dialect of Manchu. The Sibe people are widely dispersed across northwest and northeast China due to deliberate programs of ethnic dispersal during the Qing dynasty. The majority have become assimilated into the local population and no longer speak the Sibe language. However, there is a substantial Sibe population in the Sibe Autonomous County in the Ili River valley in Western Xinjiang, the descendants of border guards posted to Xinjiang in 1764, who still speak and write the Sibe language. The Sibe script is based on the Manchu script, with a few modified letters.

Directionality. The Mongolian script is written vertically from top to bottom in columns running from left to right. In modern contexts, words or phrases may be embedded in horizontal scripts. In such a case, the Mongolian text will be rotated 90 degrees counterclockwise so that it reads from left to right.

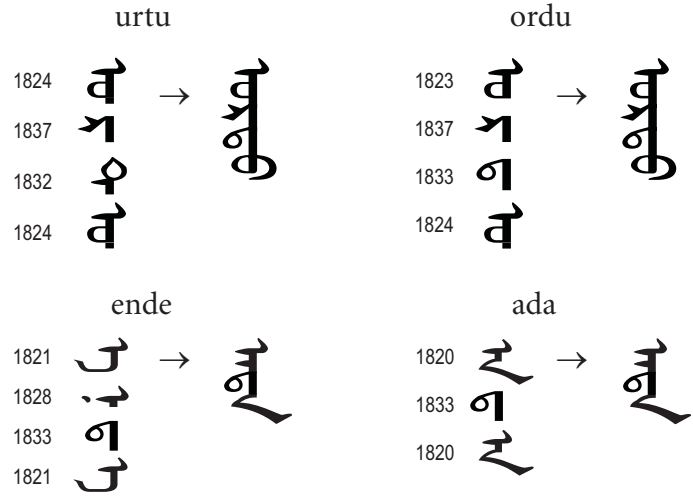
When rendering Mongolian text in a system that does not support vertical layout, the text should be laid out in horizontal lines running left to right, with the glyphs rotated 90 degrees counterclockwise with respect to their orientation in the code charts. If such text is viewed sideways, the usual Mongolian column order appears reversed, but this orientation can be workable for short stretches of text. There are no bidirectional effects in such a layout because all text is horizontal left to right.

Encoding Principles. The encoding model for Mongolian is somewhat different from that for any other script within Unicode, and in many respects it is the most complicated. For this reason, only the essential features of Mongolian shaping behavior are presented here.

The Semitic alphabet from which the Mongolian script was ultimately derived is fundamentally inadequate for representing the sounds of the Mongolian language. As a result, many of the Mongolian letters are used to represent two different sounds, and the correct pronunciation of a letter may be known only from the context. In this respect, Mongolian orthography is similar to English spelling, in which the pronunciation of a letter such as *c* may be known only from the context.

Unlike in the Latin script, in which *c* /k/ and *c* /s/ are treated as the same letter and encoded as a single character, in the Mongolian script different phonetic values of the same glyph may be encoded as distinct characters. Modern Mongolian grammars consider the phonetic value of a letter to be its distinguishing feature, rather than its glyph shape. For example, the four Mongolian vowels *o*, *u*, *ö,* and *ü* are considered four distinct letters and are encoded as four characters (U+1823, U+1824, U+1825, and U+1826, respectively), even though *o* is written identically to *u* in all positional forms, *ö* is written identically to *ü* in all positional forms, *o* and *u* are normally distinguished from *ö* and *ü* only in the first syllable of a word. Likewise, the letters *t* (U+1832) and *d* (U+1833) are often indistinguishable. For example, pairs of Mongolian words such as *urtu* "long" and *ordu* "palace, camp, horde" or *ende* "here" and *ada* "devil" are written identically, but are represented using different sequences of Unicode characters, as shown in *Figure 13-3*. There are many such examples in Mongolian, but not in Todo, Manchu, or Sibe, which have largely eliminated ambiguous letters.

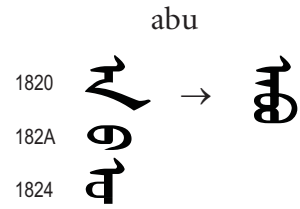
Figure 13-3. Mongolian Glyph Convergence



Cursive Joining. The Mongolian script is cursive, and the letters constituting a word are normally joined together. In most cases the letters join together naturally along a vertical stem, but in the case of certain "bowed" consonants (for example, U+182A MONGOLIAN LETTER BA and the feminine form of U+182C MONGOLIAN LETTER QA), which lack a trailing vertical stem, they may form ligatures with a following vowel. This is illustrated in

Figure 13-4, where the letter *ba* combines with the letter *u* to form a ligature in the Mongolian word *abu* "father."

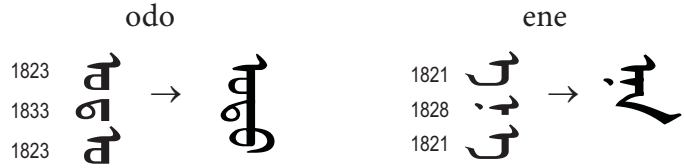
Figure 13-4. Mongolian Consonant Ligation



The Joining_Type values for Mongolian characters are defined in ArabicShaping.txt in the Unicode Character Database. For a discussion of the meaning of Joining_Type values in the context of a vertically rendered script, see "Cursive Joining" in *Section 14.3, Phags-pa*. Most Mongolian characters are Dual_Joining, as they may join on both top and bottom.

Many letters also have distinct glyph forms depending on their position within a word. These positional forms are classified as initial, medial, final, or isolate. The medial form is often the same as the initial form, but the final form is always distinct from the initial or medial form. *Figure 13-5* shows the Mongolian letters U+1823 *o* and U+1821 *e*, rendered with distinct positional forms initially and finally in the Mongolian words *odo* "now" and *ene* "this."

Figure 13-5. Mongolian Positional Forms



U+200C ZERO WIDTH NON-JOINER (ZWNJ) and U+200D ZERO WIDTH JOINER (ZWJ) may be used to select a particular positional form of a letter in isolation or to override the expected positional form within a word. Basically, they evoke the same contextual selection effects in neighboring letters as do non-joining or joining regular letters, but are themselves invisible (see *Chapter 23, Special Areas and Format Characters*). For example, the various positional forms of U+1820 MONGOLIAN LETTER A may be selected by means of the following character sequences:

<1820> selects the isolate form.

<1820 200D> selects the initial form.

<200D 1820> selects the final form.

<200D 1820 200D> selects the medial form.

Some letters have additional variant forms that do not depend on their position within a word, but instead reflect differences between modern versus traditional orthographic practice or lexical considerations—for example, special forms used for writing foreign words. On occasion, other contextual rules may condition a variant form selection. For example, a certain variant of a letter may be required when it occurs in the first syllable of a word or when it occurs immediately after a particular letter.

The various positional and variant glyph forms of a letter are considered presentation forms and are not encoded separately. It is the responsibility of the rendering system to select the correct glyph form for a letter according to its context.

Free Variation Selectors. When a glyph form that cannot be predicted algorithmically is required (for example, when writing a foreign word), the user needs to append an appropriate variation selector to the letter to indicate to the rendering system which glyph form is required. The following free variation selectors are provided for use specifically with the Mongolian block:

U+180B MONGOLIAN FREE VARIATION SELECTOR ONE (FVS1)

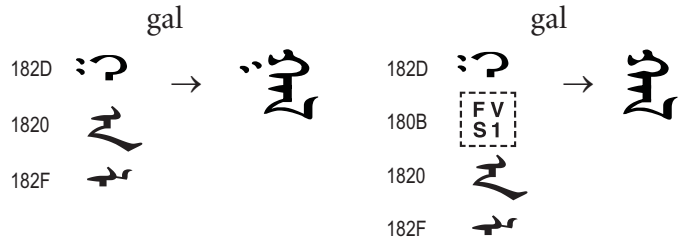
U+180C MONGOLIAN FREE VARIATION SELECTOR TWO (FVS2)

U+180D MONGOLIAN FREE VARIATION SELECTOR THREE (FVS3)

These format characters normally have no visual appearance. When required, a free variation selector immediately follows the base character it modifies. This combination of base character and variation selector is known as a standardized variant. The table of standardized variants, StandardizedVariants.txt, in the Unicode Character Database exhaustively lists all currently defined standardized variants. All combinations not listed in the table are unspecified and are reserved for future standardization; no conformant process may interpret them as standardized variants. Therefore, any free variation selector not immediately preceded by one of their defined base characters will be ignored.

Figure 13-6 gives an example of how a free variation selector may be used to select a particular glyph variant. In modern orthography, the initial letter *ga* in the Mongolian word *gal* "fire" is written with two dots; in traditional orthography, the letter *ga* is written without any dots. By default, the dotted form of the letter *ga* is selected, but this behavior may be overridden by means of FVS1, so that *ga* plus FVS1 selects the undotted form of the letter *ga*.

Figure 13-6. Mongolian Free Variation Selector



It is important to appreciate that even though a particular standardized variant may be defined for a letter, the user needs to apply the appropriate free variation selector only if the correct glyph form cannot be predicted automatically by the rendering system. In most cases, in running text, there will be few occasions when a free variation selector is required to disambiguate the glyph form.

Older documentation, external to the Unicode Standard, listed the action of the free variation selectors by using ZWJ to explicitly indicate the shaping environment affected by the variation selector. The relative order of the ZWJ and the free variation selector in these documents was different from the one required by *Section 23.4, Variation Selectors*. Older implementations of Mongolian free variation selectors may therefore interpret a sequence such as a base character followed first by ZWJ and then by FVS1 as if it were a base character followed first by FVS1 and then by ZWJ.

Representative Glyphs. The representative glyph in the code charts is generally the isolate form for the vowels and the initial form for the consonants. Letters that share the same glyph forms are distinguished by using different positional forms for the representative glyph. For example, the representative glyph for U+1823 MONGOLIAN LETTER O is the isolate form, whereas the representative glyph for U+1824 MONGOLIAN LETTER U is the initial form. However, this distinction is only nominal, as the glyphs for the two characters are identical for the same positional form. Likewise, the representative glyphs for U+1863 MONGOLIAN LETTER SIBE KA and U+1874 MONGOLIAN LETTER MANCHU KA both take the final form, as their initial forms are identical to the representative glyph for U+182C MONGOLIAN LETTER QA (the initial form).

Vowel Harmony. Mongolian has a system of vowel harmony, whereby the vowels in a word are either all "masculine" and "neuter" vowels (that is, back vowels plus /i/) or all "feminine" and "neuter" vowels (that is, front vowels plus /i/). Words that are written with masculine/neuter vowels are considered to be masculine, and words that are written with feminine/neuter vowels are considered to be feminine. Words with only neuter vowels behave as feminine words (for example, take feminine suffixes). Manchu and Sibe have a similar system of vowel harmony, although it is not so strict. Some words in these two scripts may include both masculine and feminine vowels, and separated suffixes with masculine or feminine vowels may be applied to a stem irrespective of its gender.

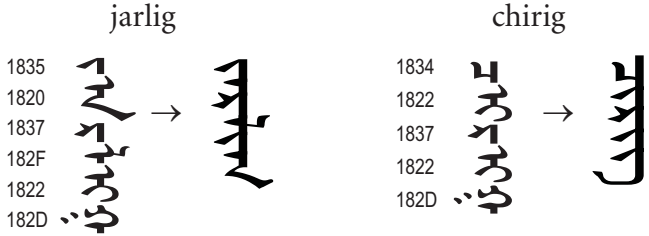
Vowel harmony is an important element of the encoding model, as the gender of a word determines the glyph form of the velar series of consonant letters for Mongolian, Todo, Sibe, and Manchu. In each script, the velar letters have both masculine and feminine forms. For Mongolian and Todo, the masculine and feminine forms of these letters have different pronunciations.

When one of the velar consonants precedes a vowel, it takes the masculine form before masculine vowels, and the feminine form before feminine or neuter vowels. In the latter case, a ligature of the consonant and vowel is required.

When one of these consonants precedes another consonant or is the final letter in a word, it may take either a masculine or feminine glyph form, depending on its context. The rendering system should automatically select the correct gender form for these letters based on

the gender of the word (in Mongolian and Todo) or the gender of the preceding vowel (in Manchu and Sibe). This is illustrated by *Figure 13-7*, where U+182D MONGOLIAN LETTER GA takes a masculine glyph form when it occurs finally in the masculine word *jarlig* "order," but takes a feminine glyph form when it occurs finally in the feminine word *chirig* "soldier." In this example, the gender form of the final letter *ga* depends on whether the first vowel in the word is a back (masculine) vowel or a front (feminine or neuter) vowel. Where the gender is ambiguous or a form not derivable from the context is required, the user needs to specify which form is required by means of the appropriate free variation selector.

Figure 13-7. Mongolian Gender Forms



Narrow No-Break Space. In Mongolian, Todo, Manchu, and Sibe, certain grammatical suffixes are separated from the stem of a word or from other suffixes by a narrow gap. There are many such suffixes in Mongolian, usually occurring in masculine and feminine pairs (for example, the dative suffixes *-dur* and *-dür*), and a stem may take multiple suffixes. In contrast, there are only six separated suffixes for Manchu and Sibe, and stems do not take more than one suffix at a time.

As any suffixes are considered to be an integral part of the word as a whole, a line break opportunity does not occur before a suffix, and the whitespace is represented using U+202F NARROW NO-BREAK SPACE (NNBSP). For a Mongolian font it is recommended that the width of NNBSP should be one-third the width of an ordinary space (U+0020 SPACE).

NNBSP affects the form of the preceding and following letters. The final letter of the stem or suffix preceding the NNBSP takes the final positional form, whereas the first letter of the suffix following NNBSP may take the normal initial form, a variant initial form, a medial form, or a final form, depending on the particular suffix.

Mongolian Vowel Separator. In Mongolian, the letters *a* (U+1820) and *e* (U+1821) in a word-final position may take a "forward tail" form or a "backward tail" form depending on the preceding consonant that they are attached to. In some words, a final letter *a* or *e* is disconnected from the preceding consonant, in which case the vowel always takes the "forward tail" form. U+180E MONGOLIAN VOWEL SEPARATOR (MVS) is used to represent the break between a final letter *a* or *e* and the rest of the word. MVS is similar in function to NNBSP, as it divides a word and disconnects the two parts. Whereas NNBSP marks off a grammatical suffix, however, the *a* or *e* following MVS is not a suffix but an integral part of the word stem.

Whether a final letter *a* or *e* is joined or separated is purely lexical and is not a question of varying orthography. This distinction is shown in *Figure 13-8*. The example on the left shows the word *qana* <182C, 1820, 1828, 1820> without a break before the final letter *a,* which means "the outer casing of a vein." The example on the right shows the word *qana* <182C, 1820, 1828, 180E, 1820> with a break before the final letter *a,* which means "the wall of a tent.".

Figure 13-8. Mongolian Vowel Separator

Qana with Connected Final

Qana with Separated Final

The MVS has a twofold effect on shaping. On the one hand, it always selects the forward tail form of a following letter *a* or *e*. On the other hand, it may affect the form of the preceding letter. The particular form that is taken by a letter preceding an MVS depends on the particular letter and in some cases on whether traditional or modern orthography is being used. The MVS is not needed for writing Todo, Manchu, or Sibe.

Numbers. The Mongolian and Todo scripts use a set of ten digits derived from the Tibetan digits. In vertical text, numbers are traditionally written from left to right across the width of the column. In modern contexts, they are frequently rotated so that they follow the vertical flow of the text.

The Manchu and Sibe scripts do not use any special digits, although Chinese number ideographs may be employed—for example, for page numbering in traditional books.

Punctuation. Traditional punctuation marks used for Mongolian and Todo include the U+1800 MONGOLIAN BIRGA (marks the start of a passage or the recto side of a folio), U+1802 MONGOLIAN COMMA, U+1803 MONGOLIAN FULL STOP, and U+1805 MONGOLIAN FOUR DOTS (marks the end of a passage). The *birga* occurs in several different glyph forms.

In writing Todo, U+1806 MONGOLIAN TODO SOFT HYPHEN is used at the beginning of the second line to indicate resumption of a broken word. It functions like U+2010 HYPHEN, except that U+1806 appears at the beginning of a line rather than at the end.

The Manchu script normally uses only two punctuation marks: U+1808 MONGOLIAN MANCHU COMMA and U+1809 MONGOLIAN MANCHU FULL STOP.

In modern contexts, Mongolian, Todo, and Sibe may use a variety of Western punctuation marks, such as parentheses, quotation marks, question marks, and exclamation marks. U+2048 QUESTION EXCLAMATION MARK and U+2049 EXCLAMATION QUESTION MARK are used for side-by-side display of a question mark and an exclamation mark together in vertical text. Todo and Sibe may additionally use punctuation marks borrowed from Chinese,

such as U+3001 IDEOGRAPHIC COMMA, U+3002 IDEOGRAPHIC FULL STOP, U+300A LEFT DOUBLE ANGLE BRACKET, and U+300B RIGHT DOUBLE ANGLE BRACKET.

Nirugu. U+180A MONGOLIAN NIRUGU acts as a stem extender. In traditional Mongolian typography, it is used to physically extend the stem joining letters, so as to increase the separation between all letters in a word. This stretching behavior should preferably be carried out in the font rather than by the user manually inserting U+180A.

The *nirugu* may also be used to separate two parts of a compound word. For example, *altan-agula* "The Golden Mountains" may be written with the words *altan,* "golden," and *agula*, "mountains," joined together using the *nirugu*. In this usage the *nirugu* is similar to the use of hyphen in Latin scripts, but it is nonbreaking.

Syllable Boundary Marker. U+1807 MONGOLIAN SIBE SYLLABLE BOUNDARY MARKER, which is derived from the medial form of the letter *a* (U+1820), is used to disambiguate syllable boundaries within a word. It is mainly used for writing Sibe, but may also occur in Manchu texts. In native Manchu or Sibe words, syllable boundaries are never ambiguous; when transcribing Chinese proper names in the Manchu or Sibe script, however, the syllable boundary may be ambiguous. In such cases, U+1807 may be inserted into the character sequence at the syllable boundary.

13.5 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 *yakthuŋ*. 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 ◌ᤨ 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 ᤀ 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 ᤁ [kɔ], ᤂ [kʰɔ], and so on, and do not represent the Limbu syllables ᤁᤠ [ka], ᤂᤠ [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 ᤀᤶ [ɔm].

Vowel Length. Vowel length is phonologically distinctive in many contexts. Length in open syllables is indicated by writing U+193A ◌᤺ LIMBU SIGN KEMPHRENG, which looks like the diaeresis sign, over the initial consonant or cluster: ᤈ᤺ᤠ *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 ◌᤻ LIMBU SIGN SA-I: ᤐᤠᤏ᤻ *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*: ᤐ᤺ᤠᤴ *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: ᤏᤧᤛ᤻ᤛᤧ *nesse* "it lay," ᤗ᤺ᤠᤒ᤻ *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 ◌᤻ 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 ᤆ *na* is placed before ᤃ *ta*, and the letter ᤇ *gha* is placed at the end of the alphabet.

Glyph Placement. The glyph positions for Limbu combining characters are summarized in *Table 13-3*.

Table 13-3. Positions of Limbu Combining Characters

Syllable	Glyphs	Code Point Sequence
ta	ᤃᤰ	190B 1920
ti	ᤃᤱ	190B 1921
tu	ᤃᤲ	190B 1922
tee	ᤃᤳ	190B 1923
tai	ᤃᤴ	190B 1924
too	ᤃᤵ	190B 1925
tau	ᤃᤶ	190B 1926
te	ᤃᤷ	190B 1927
to	ᤃᤸ	190B 1928
tya	ᤃ᤹	190B 1929
tra	ᤃ᤺	190B 192A
twa	ᤃ᤻	190B 192B
tak	ᤃ᤼	190B 1930
taŋ	ᤃ᤽	190B 1931
taṁ	ᤃ᤾	190B 1932
tat	ᤃ᤿	190B 1933
tan	ᤃ᥀	190B 1934
tap	ᤃ᥁	190B 1935
tam	ᤃ᥂	190B 1936
tar	ᤃᥑ	190B 1937
tal	ᤃᥒ	190B 1938
tā	ᤃᤰᤢ	190B 1920 193A
tī	ᤃᤱᤢ	190B 1921 193A

Punctuation. The main punctuation mark used is the double vertical line, U+0965 DEVANAGARI DOUBLE DANDA. U+1945 ᥅ LIMBU QUESTION MARK and U+1944 ᥄ 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 in the same documents. U+1940 ᥀ LIMBU SIGN LOO is used for the exclamatory particle *lo*. This particle is also often simply spelled out ᤕᤰ.

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.

13.6 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 ꯀꯔ may be read *kara* or *kra*, ꯀ꯭ꯔ 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+AAF6 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 Brahmi-derived scripts. For example, the conjunct /ṅha/ is represented by the sequence <ABC9, AAF6, 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 REPETITION MARK.

13.7 Mro

Mro: U+16A40–U+16A6F

The Mro script was invented in the 1980s. It is used to write the Mro (or Mru) language, a language of the Mruic branch of Tibeto-Burman spoken in Southeastern Bangladesh and neighboring areas of Myanmar. (This language is distinct from Mro-Khimi, a language of the Kukish branch of Tibeto-Burman spoken in Myanmar.)

The Mro script is unrelated to any other script. Some of the letters of the Mro alphabet have a visual similarity to letters from other alphabets, but such similarities are coincidental.

Structure. The Mro script is a left-to-right alphabet with no combining characters or tone marks. Some sounds are represented by more than one letter.

Character Names. Consonant letter names are traditional, based on phonetic transcription.

Digits. Mro has a script-specific set of digits.

Punctuation. There are two script-specific punctuation characters, U+16A6E MRO DANDA and U+16A6F MRO DOUBLE DANDA. Words are separated by spaces.

Two of the Mro letters are used as abbreviations. U+16A5E MRO LETTER TEK can be used instead of the word "tek," meaning "quote." U+16A5C MRO LETTER HAI can be used for various groups of letters.

13.8 Warang Citi

Warang Citi: U+118A0–U+118FF

The Warang Citi script is used to write the Ho language. Ho is a North Munda language. Warang Citi was devised by community leader Lako Bodra as an improvement over scripts used by Christian missionary linguists. Speakers of Ho live in the Indian states of Odisha (formerly Orissa) and Jharkhand. There are at present two publications in the script: a yearly magazine and a biweekly publication.

The Ho community is primarily an oral community, with an emergent literary tradition. Many Ho speakers do not write their language in any form. In some areas, Ho speakers use the Devanagari script or Warang Citi, in other locations they use the Oriya (now officially known as Odia) script or Warang Citi. There are also people who use Latin letters to write Ho on an ad-hoc basis.

Structure. Warang Citi is an alphabet, written from left to right. Unlike many other Indic scripts, vowels are written as full letters, with no vowel-modifiers. However, consonants may have an inherent vowel; it typically is pronounced [a] or [ɔ], and less often [ɛ], but this vowel does not occur in final position in a word. Because these inherent vowels are not written explicitly, there can be ambiguity in the reading of certain words.

Warang Citi has no regular system of conjuncts nor an explicit virama. However, a small number of conjunct forms are used; most of these represent doubled consonants. The choice of a conjunct form does not appear to be predictable. The recommended mechanism for representing these conjuncts is to make use of U+200D ZERO WIDTH JOINER.

Warang Citi uses case distinctions, so both uppercase and lowercase letters are encoded.

The script does not include a diacritical mark for *anusvara* as in Devanagari, but rather has a separate character, U+118C0 WARANG CITI SMALL LETTER NGAA.

Digits and Numbers. Warang Citi has a set of digits and numbers, but the orthographic conventions for writing numbers have not yet stabilized. European digits are also used, though not consistently.

Punctuation. Warang Citi uses Latin punctuation. There is no script-specific punctuation.

13.9 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 Cemeť, 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 Odishan Mayurbhañj district. The script has received some official recognition by the Odishan 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+1C5F, 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-GAAHLAA 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 [k'] 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.

13.10 Chakma

Chakma: U+11100–U+1114F

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: 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 *Cāṅmā 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.

13.11 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 letterforms 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 letterforms, not clearly related to the corresponding consonants without *-la*.

Retroflex Consonants. The Lepcha language contains three retroflex consonants: [ṭ], [th], and [ḍ]. 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 LEPCHA 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 13-4*.

Table 13-4. Lepcha Syllabic Structure

Class	Example	Encoding
consonant, letter a	ᰀ	[U+1C00..U+1C23, U+1C4D..U+1C4F]
nukta	◌᰷	U+1C37
medial -ra	◌᰻	U+1C25
medial -ya	◌ᰵ	U+1C24
dependent vowel	◌ᰦ	[U+1C26..U+1C2C]
final consonant sign	◌᰸	[U+1C2D..U+1C35]
syllabic modifier	◌᰼	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 PUNCTUATION 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.

13.12 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, Saurashtra 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+A8C4 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.