

The Unicode® Standard

Version 14.0 – Core Specification

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

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

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

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

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

© 2021 Unicode, Inc.

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

The Unicode Standard / the Unicode Consortium; edited by the Unicode Consortium. — Version 14.0.

Includes index.

ISBN 978-1-936213-29-0 (<https://www.unicode.org/versions/Unicode14.0.0/>)

1. Unicode (Computer character set) I. Unicode Consortium.

QA268.U545 2021

ISBN 978-1-936213-29-0

Published in Mountain View, CA

September 2021

Chapter 14

South and Central Asia-III

Ancient Scripts

The following scripts are described in this chapter:

<i>Brahmi</i>	<i>Marchen</i>	<i>Old Sogdian</i>
<i>Kharoshthi</i>	<i>Old Turkic</i>	<i>Sogdian</i>
<i>Bhaiksuki</i>	<i>Soyombo</i>	<i>Old Uyghur</i>
<i>Phags-pa</i>	<i>Zanabazar Square</i>	

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

The Bhaiksuki script is a Brahmi-derived script used around 1000 CE, primarily in the area of the present-day states of Bihar and West Bengal in India and northern Bangladesh. Surviving Bhaiksuki texts are limited to a few Buddhist manuscripts and inscriptions.

Phags-pa is an historical script related to Tibetan that was created as the national script of the Mongol empire. Phags-pa was used mostly in Eastern and Central Asia for writing text in the Mongolian and Chinese languages.

The Marchen script (Tibetan *sMar-chen*) is a Brahmi-derived script used in the Tibetan Bön liturgical tradition. Marchen is used to write Tibetan and the historic Zhang-zhung language. Although few historical examples of the script have been found, Marchen appears in modern-day inscriptions and in modern Bön literature.

The Old Turkic script is known from eighth-century Siberian stone inscriptions, and is the oldest known form of writing for a Turkic language. Also referred to as Turkic Runes due to its superficial resemblance to Germanic Runes, it appears to have evolved from the Sogdian script, which is in turn derived from Aramaic.

Both the Soyombo script and the Zanabazar Square script are historic scripts used to write Mongolian, Sanskrit, and Tibetan. These two scripts were both invented by Zanabazar (1635–1723), one of the most important Buddhist leaders in Mongolia. Each script is an *abugida*. Soyombo appears primarily in Buddhist texts in Central Asia. Zanabazar Square has also been called “Horizontal Square” script, “Mongolian Horizontal Square” script and “Xewtee Dörböljin Bicig.”

Old Sogdian and Sogdian are related scripts used in Central Asia. The Old Sogdian script was used for a group of related writing systems dating from the third to the sixth century CE. These writing systems were all used to write Sogdian, an eastern Iranian language. Old Sogdian is a non-joining abjad. Its basic repertoire consists of 20 of the 22 letters of the Aramaic alphabet.

The Sogdian script, which derives from Old Sogdian, is also an abjad, and was used from the seventh to the fourteenth century CE, also to write Sogdian. Its repertoire corresponds to that of Old Sogdian, but has a number of differences in the glyphs and also has additional characters. The script was also used to write Chinese, Sanskrit, and Uyghur. Sogdian is the ancestor of the Old Uyghur and Mongolian scripts.

The Old Uyghur script flourished between the 8th and 17th centuries in northwest China and other parts of Asia. Originally used to write medieval Turkish languages, its use later expanded to write other languages, including Chinese, Mongolian, Tibetan and Arabic. Old Uyghur is a cursive joining alphabet, and developed from the cursive style of the Sogdian script. The default orientation of the script is horizontal, with the script being read right-to-left.

14.1 Brahmi

Brahmi: U+11000–U+1106F

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

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

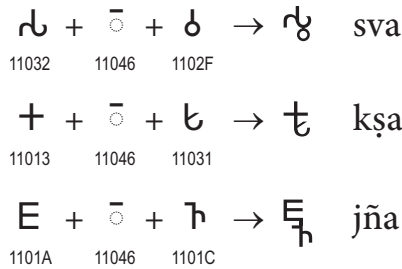
Vowel Letters. Vowel letters are encoded atomically in Brahmi, even if they can be analyzed visually as consisting of multiple parts. *Table 14-1* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

Table 14-1. Brahmi Vowel Letters

To Represent	Use	Do Not Use
𑀓	11006	<11005, 11038>
𑀘	1100C	<1100B, 1103E>
𑀡	11010	<1100F, 11042>

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

The elements of consonant ligatures are laid out from top left to bottom right, as shown for *sva* in *Figure 14-1*. Preconsonantal *r*, postconsonantal *r* and postconsonantal *y* assume special reduced shapes in all except the earliest varieties of Brahmi. The *kṣa* and *jña* ligatures, however, are often transparent, as also shown in *Figure 14-1*.

Figure 14-1. Consonant Ligatures in Brahmi

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

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

Old Tamil Brahmi. Brahmi was used to write the Tamil language starting from the second century BCE. The different orthographies used to write Tamil in the Brahmi script are covered by the Unicode encoding of Brahmi.

In one of these orthographies the inherent vowel of Brahmi consonant letters is dropped, and U+11038 BRAHMI VOWEL SIGN AA is used to represent both short and long [a] / [a:]. In this orthography, consonant signs without a vowel sign always represent the bare consonant without an inherent vowel.

Some orthographies employ U+11070 BRAHMI SIGN OLD TAMIL VIRAMA to cancel the inherent vowel of the consonants, but the virama does not form conjuncts. The glyph for Old Tamil virama is a dot, called a *pulli*, which may appear identical to U+11001 BRAHMI SIGN ANUSVARA. Fonts may differentiate the Old Tamil virama from the Brahmi *anusvara*

by placing the dots at different positions according to the style of the font. These orthographies also use Old Tamil short vowels [e] and [o], which are atomically encoded at U+11071..U+11074. The glyphs for these vowels appear with a *pulli*, but the short vowels [e] and [o] are not decomposed.

Distinct Old Tamil consonants not found in Prakrit and Sanskrit are encoded at U+11035..U+11037 and U+11075. When U+1102B BRAHMI LETTER MA occurs in Old Tamil text, it may be shown with a glyphic variant distinct from the form shown in the Brahmi code charts.

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

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

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

The set of additive-multiplicative numerals of the Brahmi script contains separate signs for the digits from 1 to 9, the tens from 10 to 90, as well as signs for 100 and 1000. Numbers are written additively, with the higher-valued signs preceding the lower-valued ones. Multiples of 100 and of 1000 are expressed multiplicatively with character sequences consisting of the sign for 100 or 1000, followed by U+1107F BRAHMI NUMBER JOINER and then the multiplier. The component parts of additive numbers are rendered unligated, whereas multiples are rendered in ligated form.

For example, the sequence <U+11064 BRAHMI NUMBER ONE HUNDRED, U+11055 BRAHMI NUMBER FOUR> represents the number $100 + 4 = 104$ and is rendered unligated, whereas the sequence <U+11064 BRAHMI NUMBER ONE HUNDRED, U+1107F BRAHMI NUMBER JOINER, U+11055 BRAHMI NUMBER FOUR> represents the number $100 \times 4 = 400$ and is rendered as a ligature.

U+1107F BRAHMI NUMBER JOINER forms a ligature between the two numeral characters surrounding it. It functions similarly to U+2D7F TIFINAGH CONSONANT JOINER, but is intended to be used only with Brahmi numerals in the range U+11052 BRAHMI NUMBER ONE through U+11065 BRAHMI NUMBER ONE THOUSAND, and not with consonants or other

characters. Because U+1107F BRAHMI NUMBER JOINER marks a semantic distinction between additive numbers and multiples, it should be rendered with a visible fallback glyph to indicate its presence in the text when it cannot be displayed by normal rendering.

In addition to the ligated forms of the multiples of 100 and 1000, other examples from the middle and late Brahmi periods show the signs for 200, 300, and 2000 in special forms not obviously connected with a ligature of the component parts. Such forms may be enabled in fonts using a ligature substitution.

A special sign for zero was invented later, and the positional system came into use. This system is the ancestor of modern decimal number systems. Due to the different systemic features and shapes, the signs in this set are separately encoded in the range from U+11066 BRAHMI DIGIT ZERO through U+1106F BRAHMI DIGIT NINE. These signs have the same properties as the modern Indic digits. Examples are shown in *Table 14-2*. Brahmi decimal digits are categorized as regular bases and can act as vowel carriers, whereas the numerals U+11052 BRAHMI NUMBER ONE through U+11065 BRAHMI NUMBER ONE THOUSAND and their ligatures formed with U+1107F BRAHMI NUMBER JOINER are not used as vowel carriers.

Table 14-2. Brahmi Positional Digits

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

14.2 Kharoshthi

Kharoshthi: U+10A00–U+10A5F

The Kharoshthi script, properly spelled as Kharoṣṭhī, was used historically to write Gāndhārī and Sanskrit as well as various mixed dialects. Kharoshthi is an Indic script of the *abugida* type. However, unlike other Indic scripts, it is written from right to left. The Kharoshthi script was initially deciphered around the middle of the 19th century by James Prinsep and others who worked from short Greek and Kharoshthi inscriptions on the coins of the Indo-Greek and Indo-Scythian kings. The decipherment has been refined over the last 150 years as more material has come to light.

The Kharoshthi script is one of the two ancient writing systems of India. Unlike the pan-Indian Brāhmī script, Kharoshthi was confined to the northwest of India centered on the region of *Gandhāra* (modern northern Pakistan and eastern Afghanistan, as shown in *Figure 14-2*). Gandhara proper is shown on the map as the dark gray area near Peshawar. The lighter gray areas represent places where the Kharoshthi script was used and where manuscripts and inscriptions have been found.

Figure 14-2. Geographical Extent of the Kharoshthi Script



The exact details of the origin of the Kharoshthi script remain obscure, but it is almost certainly related to Aramaic. The Kharoshthi script first appears in a fully developed form in the Aśokan inscriptions at Shahbazarhi and Mansehra which have been dated to around 250 BCE. The script continued to be used in Gandhara and neighboring regions, sometimes alongside Brahmi, until around the third century CE, when it disappeared from its homeland. Kharoshthi was also used for official documents and epigraphs in the Central Asian cities of Khotan and Niya in the third and fourth centuries CE, and it appears to have survived in

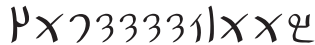
Kucha and neighboring areas along the Northern Silk Road until the seventh century. The Central Asian form of the script used during these later centuries is termed *Formal Kharoshthi* and was used to write both Gandhari and Tocharian B. Representation of Kharoshthi in the Unicode code charts uses forms based on manuscripts of the first century CE.

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

Diacritical Marks and Vowels. All vowels other than *a* are written with diacritical marks in Kharoshthi. In addition, there are six vowel modifiers and three consonant modifiers that are written with combining diacritics. In general, only one combining vowel sign is applied to each syllable (*aksara*). However, there are some examples of two vowel signs on *aksaras* in the Kharoshthi of Central Asia.

Numerals. Kharoshthi employs a set of eight numeral signs unique to the script. Like the letters, the numerals are written from right to left. Numbers in Kharoshthi are based on an additive system. There is no zero, nor separate signs for the numbers five through nine. The number 1996, for example, would logically be represented as 1000 4 4 1 100 20 20 20 20 10 4 2 and would appear as shown in *Figure 14-3*. The numerals are encoded in the range U+10A40..U+10A47.

Figure 14-3. Kharoshthi Number 1996



Punctuation. Nine different punctuation marks are used in manuscripts and inscriptions. The punctuation marks are encoded in the range U+10A50..U+10A58.

Word Breaks, Line Breaks, and Hyphenation. Most Kharoshthi manuscripts are written as continuous text with no indication of word boundaries. Only a few examples are known where spaces have been used to separate words or verse quarters. Most scribes tried to finish a word before starting a new line. There are no examples of anything akin to hyphenation in Kharoshthi manuscripts. In cases where a word would not completely fit into a line, its continuation appears at the start of the next line. Modern scholarly practice uses spaces and hyphenation. When necessary, hyphenation should follow Sanskrit practice.

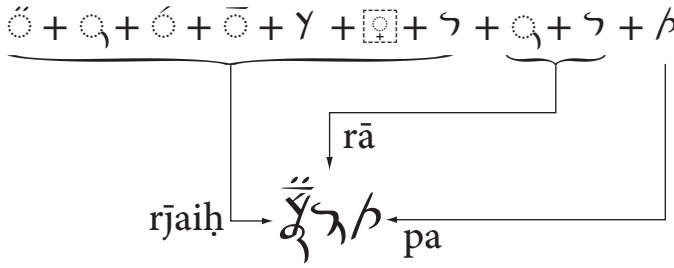
Sorting. There is an ancient ordering connected with Kharoshthi called *Arapacana*, named after the first five *aksaras*. However, there is no evidence that words were sorted in this order, and there is no record of the complete *Arapacana* sequence. In modern scholarly practice, Gandhari is sorted in much the same order as Sanskrit. Vowel length, even when marked, is ignored when sorting Kharoshthi.

Rendering Kharoshthi

Rendering requirements for Kharoshthi are similar to those for Devanagari. This section specifies a minimum set of combining rules that provide legible Kharoshthi diacritic and ligature substitution behavior.

All unmarked consonants include the inherent vowel *a*. Other vowels are indicated by one of the combining vowel diacritics. Some letters may take more than one diacritical mark. In these cases the preferred sequence is Letter + {Consonant Modifier} + {Vowel Sign} + {Vowel Modifier}. For example the Sanskrit word *parārdhyaiḥ* might be rendered in Kharoshthi script as **parārjaiḥ*, written from right to left, as shown in Figure 14-4.

Figure 14-4. Kharoshthi Rendering Example



Combining Vowels. The various combining vowels attach to characters in different ways. A number of groupings have been determined on the basis of their visual types, such as horizontal or vertical, as shown in Table 14-3.

Table 14-3. Kharoshthi Vowel Signs

Type	Example	Group Members
Vowel sign i		
Horizontal	a + -i → i 𑀓 + 𑀓 → 𑀓	A, NA, HA
Vertical	tha + -i → thi 𑀧 + 𑀓 → 𑀧	THA, PA, PHA, MA, LA, SHA
Diagonal	ka + -i → ki 𑀕 + 𑀓 → 𑀕	All other letters
Vowel sign u		
Independent	ha + -u → hu 𑀇 + 𑀓 → 𑀇	TTA, HA
Ligated	ma + -u → mu 𑀛 + 𑀓 → 𑀛	MA
Attached	a + -u → u 𑀓 + 𑀓 → 𑀓	All other letters

Table 14-3. Kharoshthi Vowel Signs (Continued)

Type	Example	Group Members
Vowel sign vocalic r		
Attached	a + -ṛ → ṛ 𑀅 + 𑀇 → 𑀆	A, KA, KKA, KHA, GA, GHA, CA, CHA, JA, TA, DA, DHA, NA, PA, PHA, BA, BHA, VA, SHA, SA
Independent	ma + -ṛ → mṛ 𑀢 + 𑀇 → 𑀣	MA, HA
Vowel sign e		
Horizontal	a + -e → e 𑀅 + 𑀈 → 𑀆	A, NA, HA
Vertical	tha + -e → the 𑀢 + 𑀈 → 𑀣	THA, PA, PHA, LA, SSA
Ligated	da + -e → de 𑀢 + 𑀈 → 𑀣	DA, MA
Diagonal	ka + -e → ke 𑀅 + 𑀈 → 𑀆	All other letters
Vowel sign o		
Vertical	pa + -o → po 𑀢 + 𑀉 → 𑀣	PA, PHA, YA, SHA
Diagonal	a + -o → o 𑀅 + 𑀉 → 𑀆	All other letters

Combining Vowel Modifiers. U+10A0C 𑀇 KHAROSHTHI VOWEL LENGTH MARK indicates equivalent long vowels and, when used in combination with -e and -o, indicates the diphthongs -ai and -au. U+10A0D 𑀈 KHAROSHTHI SIGN DOUBLE RING BELOW appears in some Central Asian documents, but its precise phonetic value has not yet been established. These two modifiers have been found only in manuscripts and inscriptions from the first century CE onward. U+10A0E 𑀉 KHAROSHTHI SIGN ANUSVARA indicates nasalization, and U+10A0F 𑀊 KHAROSHTHI SIGN VISARGA is generally used to indicate unvoiced syllable-final [h], but has a secondary use as a vowel length marker. *Visarga* is found only in Sanskritized forms of the language and is not known to occur in a single *aksara* with *anusvara*. The modifiers and the vowels they modify are given in *Table 14-4*.

Table 14-4. Kharoshthi Vowel Modifiers

Type	Example	Group Members
Vowel length mark	ma + ◌̄ → mā 𑀢 + ◌̄ → 𑀣	A, I, U, R, E, O
Double ring below	sa + ◌̣ → sā 𑀢 + ◌̣ → 𑀤	A, U
Anusvara	a + -ṃ → aṃ 𑀢 + ◌̣̣ → 𑀥	A, I, U, R, E, O
Visarga	ka + -ḥ → kaḥ 𑀢 + ◌̣̣̣ → 𑀦	A, I, U, R, E, O

Combining Consonant Modifiers. U+10A38 ◌̄ KHAROSHTHI SIGN BAR ABOVE indicates various modified pronunciations depending on the consonants involved, such as nasalization or aspiration. U+10A39 ◌̣ KHAROSHTHI SIGN CAUDA indicates various modified pronunciations of consonants, particularly fricativization. The precise value of U+10A3A ◌̣̣ KHAROSHTHI SIGN DOT BELOW has not yet been determined. Usually only one consonant modifier can be applied to a single consonant. The resulting combined form may also combine with vowel diacritics, one of the vowel modifiers, or anusvara or visarga. The modifiers and the consonants they modify are given in *Table 14-5*.

Table 14-5. Kharoshthi Consonant Modifiers

Type	Example	Group Members
Bar above	ja + ◌̄ → jā 𑀢 + ◌̄ → 𑀧	GA, CA, JA, NA, MA, SHA, SSA, SA, HA
Cauda	ga + ◌̣ → gā 𑀢 + ◌̣ → 𑀨	GA, JA, DDA, TA, DA, PA, YA, VA, SHA, SA
Dot below	ma + ◌̣̣ → mā 𑀢 + ◌̣̣ → 𑀩	MA, HA

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

The virama can follow only a consonant or a consonant modifier. It cannot follow a space, a vowel, a vowel modifier, a number, a punctuation sign, or another virama. Examples of the use of the Kharoshthi virama are given in *Table 14-6*.

Table 14-6. Examples of Kharoshthi Virama

Type	Example
Pure virama	dha + i + k + virama → dhik 𑖅 + 𑖇 + 𑖆 + 𑖇 → 𑖅𑖆
Ligatures	ka + virama + ṣa → kṣa 𑖅 + 𑖇 + 𑖱 → 𑖅𑖱
Consonants with special combining forms	sa + virama + ya → sya 𑖱 + 𑖇 + 𑖅 → 𑖱𑖅
Consonants with full combined form	ka + virama + ta → kta 𑖅 + 𑖇 + 𑖱 → 𑖅𑖱

Subjoined ya. A special form of subjoined *ya* appears in the Kharoshthi documents from Niya. In most cases this sign occurs in loan words into Gandhari. The most common source for these loans is presumed to be Tocharian A, where the sequence *-ly-* is normal. This special form resembles the full form of *ya* (𑖅), attached cursively to the stem of the preceding consonant sign. This contrasts with the common form of subjoined *ya* which is a looped flourish extension of the stem. The special form of *ya* can be requested using U+200D ZERO WIDTH JOINER as shown in *Figure 14-5*.

Figure 14-5. Subjoined Forms of *ya*

la + virama + ya → lya
𑖇 + 𑖇 + 𑖅 → 𑖇𑖅

la + zwj + virama + ya → lya
𑖇 + 𑖇 + 𑖇 + 𑖅 → 𑖇𑖅

14.3 Bhaiksuki

Bhaiksuki: U+11C00–U+11C6F

The Bhaiksuki script is a Brahmi-derived script used from about the 10th to the 13th centuries CE, primarily in the area of the present-day states of Bihar and West Bengal in India and northern Bangladesh. The original name of the script was Saindhavī (that is, the Sindhu or Indus script), but after its discovery in the late 19th century, scholars called it Bhaiksuki or they used a descriptive name, the Arrow-headed script. Surviving Bhaiksuki texts are limited to a few Buddhist manuscripts and inscriptions.

Structure. The structure of Bhaiksuki script is similar to that of other Brahmi-based Indic scripts. It is an *abugida* that makes use of a virama. The script is written from left to right.

Rendering. Many of the vowel signs have contextual variants when they occur with certain consonants. The consonants U+11C22 BHAIKSUKI LETTER PA, U+11C27 BHAIKSUKI LETTER YA, and U+11C28 BHAIKSUKI LETTER RA have special combining forms when they occur with certain vowel signs.

Virama and Conjuncts. The script includes a virama, U+11C3F BHAIKSUKI SIGN VIRAMA, which functions to suppress the inherent vowel and to form conjuncts. Consonant clusters are generally rendered as vertically stacked ligatures, with non-initial consonants attached below the initial letter. Above-base vowel signs and consonant letters attach to the glyph of the initial consonant, while below-base vowel signs attach to the glyph of the final consonant. The letters *ka*, *pa*, *ra*, and *ya* take special forms when they occur in conjuncts.

The Bhaiksuki dependent vowel signs in the range U+11C38..U+11C3B, *e*, *ai*, *o*, and *au*, are simply treated as above-base vowel signs. Although the historically cognate vowel signs may be treated as having left-side parts, or as two- or three-part vowels in many other scripts of India, the peculiarities of rendering for these vowel signs in the Bhaiksuki script can be handled more easily with the above-base designations. The dependent vowel signs *ai*, *o*, and *au* are not given formal canonical decompositions, but are encoded instead as atomic characters.

The sequence <C, *virama*> is rendered using a visible virama by default. The combinations <*ta*, *virama*>, <*na*, *virama*>, and <*ma*, *virama*> may also be displayed with special ligatures; there is no apparent semantic distinction between sequences containing the visible virama and sequences displayed as ligatures.

Various Signs. Nasalization is represented by U+11C3C BHAIKSUKI SIGN CANDRABINDU and U+11C3D BHAIKSUKI SIGN ANUSVARA. Post-vocalic aspiration in Sanskrit is indicated by U+11C3E BHAIKSUKI SIGN VISARGA. Use of U+11C40 BHAIKSUKI SIGN AVAGRAHA indicates elision of a word-initial *a* in Sanskrit as a result of sandhi.

Digits and Numbers. Bhaiksuki has a script-specific set of decimal digits. Because the glyphs for zero and three have not been yet identified in the Bhaiksuki corpus, representative glyphs for U+11C50 BHAIKSUKI DIGIT ZERO and U+11C53 BHAIKSUKI DIGIT THREE

are based upon corresponding digits in other scripts that are contemporaneous with Bhaiksuki.

In addition to the decimal digits, the script has a distinct numerical notation system. Bhaiksuki contains numbers for primary and tens units, and U+11C6C BHAIKSUKI HUNDREDS UNIT MARK. The numbers are written vertically, with the largest number written above smaller units. Control of vertical orientation is managed at the font level, but the default rendering is horizontal left to right.

Punctuation. The script employs script-specific dandas, U+11C41 BHAIKSUKI DANDA and U+11C42 BHAIKSUKI DOUBLE DANDA. Words are separated by U+11C43 BHAIKSUKI WORD SEPARATOR. Two characters, U+11C44 BHAIKSUKI GAP FILLER-1 and U+11C45 BHAIKSUKI GAP FILLER-2, are used as spacing or completion marks, especially to indicate the end of a line. They also can indicate a deliberate elision or an otherwise missing portion of text.

14.4 Phags-pa

Phags-pa: U+A840–U+A87F

The Phags-pa script is an historic script with some limited modern use. It bears some similarity to Tibetan and has no case distinctions. It is written vertically in columns running from left to right, like Mongolian. Units are often composed of several syllables and may be separated by whitespace.

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

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

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

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

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

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

Basic Structure. The Phags-pa script is based on Tibetan, but unlike any other Brahmic script Phags-pa is written vertically from top to bottom in columns advancing from left to right across the writing surface. This unusual directionality is borrowed from Mongolian, as is the way in which Phags-pa letters are ligated together along a vertical stem axis. In modern contexts, when embedded in horizontally oriented scripts, short sections of Phags-pa text may be laid out horizontally from left to right.

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

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

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

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

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

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

Candrabindu. U+A873 PHAGS-PA LETTER CANDRABINDU is used in writing Sanskrit mantras, where it represents a final nasal sound. However, although it represents the final sound in a syllable unit, it is always written as the first glyph in the sequence of letters, above the initial consonant or vowel of the syllable, but not ligated to the following letter. For example, *om* is written as a *candrabindu* followed by the letter *o*. To simplify cursor placement, text selection, and so on, the *candrabindu* is encoded in visual order rather than logical order. Thus *om* would be represented by the sequence <U+A873, U+A861>, rendered as shown in *Figure 14-6*.

Figure 14-6. Phags-pa Syllable Om



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

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

U+A86D PHAGS-PA LETTER ALTERNATE YA

U+A86E PHAGS-PA LETTER VOICELESS SHA

U+A86F PHAGS-PA LETTER VOICED HA

U+A870 PHAGS-PA LETTER ASPIRATED FA

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

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

LETTER ALTERNATE YA. U+A857 is used to represent historic null initials, whereas U+A86D is used to represent historic palatal initials.

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

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

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

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

Positional Variants. The four vowel letters U+A85E PHAGS-PA LETTER I, U+A85F PHAGS-PA LETTER U, U+A860 PHAGS-PA LETTER E, and U+A861 PHAGS-PA LETTER O have different isolate, initial, medial, and final glyph forms depending on whether they are immediately preceded or followed by another Phags-pa letter (other than U+A873 PHAGS-PA LETTER CANDRABINDU, which does not affect the shaping of adjacent letters). The code charts show these four characters in their isolate form. The various positional forms of these letters are shown in *Table 14-7*.

Table 14-7. Phags-pa Positional Forms of I, U, E, and O

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

Consonant letters and the vowel letter U+A866 PHAGS-PA LETTER EE do not have distinct positional forms, although initial, medial, final, and isolate forms of these letters may be distinguished by the presence or absence of a stem extender that is used to ligate to the following letter.

The invisible format characters U+200D ZERO WIDTH JOINER (ZWJ) and U+200C ZERO WIDTH NON-JOINER (ZWNJ) may be used to override the expected shaping behavior, in the same way that they do for Mongolian and other scripts (see *Chapter 23, Special Areas and*

Format Characters). For example, ZWJ may be used to select the initial, medial, or final form of a letter in isolation:

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

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

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

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

Mirrored Variants. The four characters U+A869 PHAGS-PA LETTER TTA, U+A86A PHAGS-PA LETTER TTHA, U+A86B PHAGS-PA LETTER DDA, and U+A86C PHAGS-PA LETTER NNA are mirrored forms of the letters U+A848 PHAGS-PA LETTER TA, U+A849 PHAGS-PA LETTER THA, U+A84A PHAGS-PA LETTER DA, and U+A84B PHAGS-PA LETTER NA, respectively, and are used to represent the Sanskrit retroflex dental series of letters. Because these letters are mirrored, their stem axis is on the lefthand side rather than the righthand side, as is the case for all other consonant letters. This means that when the letters *tta*, *ttha*, *dda*, and *nna* occur at the start of a syllable unit, to correctly ligate with them any following letters normally take a mirrored glyph form. Because only a limited number of words use these letters, only the letters U+A856 PHAGS-PA LETTER SMALL A, U+A85C PHAGS-PA LETTER HA, U+A85E PHAGS-PA LETTER I, U+A85F PHAGS-PA LETTER U, U+A860 PHAGS-PA LETTER E, and U+A868 PHAGS-PA SUBJOINED LETTER YA are affected by this glyph mirroring behavior. The Sanskrit syllables that exhibit glyph mirroring after *tta*, *ttha*, *dda*, and *nna* are shown in Table 14-8.

Table 14-8. Contextual Glyph Mirroring in Phags-pa

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

Glyph mirroring is not consistently applied to the letters U+A856 PHAGS-PA LETTER SMALL A and U+A85E PHAGS-PA LETTER I in the extant Sanskrit Phags-pa inscriptions. The letter *i* may occur both mirrored and unmirrored after the letter *ttha*, although it always occurs mirrored after the letter *nna*. *Small a* is not normally mirrored after the letters *tta* and *ttha* as its mirrored glyph is identical in shape to U+A85A PHAGS-PA LETTER SHA. Nevertheless, *small a* does sometimes occur in a mirrored form after the letter *ttha*, in which case context indicates that this is a mirrored letter *small a* and not the letter *sha*.

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

To override the default mirroring behavior of the letters *small a*, *ha*, *i*, *u*, *e*, and *subjoined ya*, U+FE00 VARIATION SELECTOR-1 (VS1) may be applied to the appropriate character, as shown in Table 14-9. Note that only the variation sequences shown in Table 14-9 are valid; any other sequence of a Phags-pa letter and VS1 is unspecified.

Table 14-9. Phags-pa Standardized Variants

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

In Table 14-9, “reversed shaping” means that the appearance of the character is reversed with respect to its expected appearance. Thus, if no mirroring would be expected for the character in the given context, applying VS1 would cause the rendering engine to select a mirrored glyph form. Similarly, if context would dictate glyph mirroring, application of VS1 would inhibit the expected glyph mirroring. This mechanism will typically be used to select a mirrored glyph for the letters *small a*, *ha*, *i*, *u*, *e*, or *subjoined ya* in isolation (for example, in discussion of the Phags-pa script) or to inhibit mirroring of the letters *small a* and *i* when they are not mirrored after the letters *tta* and *ttha*, as shown in Figure 14-7.

Figure 14-7. Phags-pa Reversed Shaping



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

Cursive Joining. Joining types are defined for Phags-pa characters in the file ArabicShaping.txt. Joining types identify the joining behavior of characters in cursive joining scripts

and were originally introduced for the Arabic script. Because the Phags-pa script is typically rendered from top to bottom, `Joining_Type = L (Left_Joining)` conventionally refers to bottom joining that is, joining to a character which follows (is below) it. `Joining_Type = R (Right_Joining)` is not used for the Phags-pa script, but would refer to top joining, that is, joining to a character which precedes (is above) it. Most Phags-pa characters are `Dual_Joining`, as they may join on both top and bottom.

The L and R designations of the `Joining_Type` property should not be confused with the left-hand and right-hand placement of stem axes in the Phags-pa script in vertical layout. Whether a Phags-pa character joins on the left-hand or right-hand side in its stem axis is not defined in `ArabicShaping.txt`.

14.5 Marchen

Marchen: U+11C70–U+11CBF

The Marchen script (Tibetan *sMar-chen*) is a Brahmi-derived script used in the Tibetan Bön liturgical tradition. Marchen is used to write Tibetan and also the historic Zhang-zhung language. The script is said to originate in the ancient kingdom of Zhang-zhung, which flourished in western and northern Tibet before Buddhism was introduced in the area in the seventh century. Although few historical examples of the script have been found, Marchen appears in modern-day inscriptions and is widely used in modern Bön literature.

Encoding Model. The encoding model for Marchen follows that of Tibetan. Marchen contains thirty base consonants and thirty subjoined consonants, which can be used to form vertical stacks of two or more consonants. Although not all subjoined consonants have been identified in extant texts, the full set of subjoined forms is encoded, so that all possible stack combinations can be represented.

Vowels and Consonants. As in Tibetan, two or more Marchen consonants can stack vertically. Vowel signs are placed above, below, or alongside a stack of one or more consonants.

Other Signs. Marchen includes a vowel lengthener, U+11CB0 MARCHEN VOWEL SIGN AA, known as *a-chung*. Nasalization is represented by U+11CB6 MARCHEN SIGN CANDRABINDU and U+11CB5 MARCHEN SIGN ANUSVARA.

Punctuation. There are two script-specific punctuation marks encoded. U+11C70 MARCHEN HEAD MARK corresponds to U+0F04 TIBETAN MARK INITIAL YIG MGO MDUN MA. The sentence-final shad mark, U+11C71 MARCHEN MARK SHAD, corresponds to U+0F0D TIBETAN MARK SHAD. Marchen does not use an explicit mark to separate syllables; this differs from the use of the Tibetan *tsek* (tsheg) mark.

14.6 Zanabazar Square

Zanabazar Square: U+11A00–U+11A4F

The Zanabazar Square script is an *abugida* based upon Tibetan and inspired by the Brahmi model. The script has some similarities with both Tibetan and Phags-pa. It was used to write Mongolian, Sanskrit, and Tibetan, and has also been called “Horizontal Square” script, “Mongolian Horizontal Square” script and “Xewtee Dörböljin Bicig.”

The script was invented by Zanabazar (1635–1723), one of the most important Buddhist leaders in Mongolia, who also developed the Soyombo script. Its creation likely preceded that of Soyombo.

Structure. The Zanabazar Square script is written from left to right. The script is generally written horizontally, but in some instances occurs in vertical environments. Consonant letters possess the inherent vowel /a/.

The phonetic value of a consonant letter is changed by the attachment of a vowel sign. In Mongolian, the inherent vowel is suppressed by a final-consonant mark, which indicates both a syllable-final consonant and a syllabic boundary. In Sanskrit or Tibetan, the virama silences the inherent vowel of a consonant, but does not mark syllable boundaries.

Vowels and Diphthongs. The Zanabazar Square script has one vowel letter, nine dependent vowel marks, and one vowel length mark. The *letter a* vowel, U+11A00 ZANABAZAR SQUARE LETTER A, has the value /a/ when it occurs independently. It can also assume the value of a combined vowel sign.

A long vowel is represented by placing the vowel length mark, U+11A0A ZANABAZAR SQUARE VOWEL LENGTH MARK, after a consonant or vowel sign. When combined with the *letter a* vowel or a consonant letter, the length mark lengthens the inherent vowel /a/ to /ā/. Vowel signs are used with the *letter a* vowel and with consonants. Multiple vowel signs may combine with a single base letter. Independent vowels are represented by attaching vowel signs to the *letter a* vowel, except for U+11A09 ZANABAZAR SQUARE VOWEL SIGN REVERSED I. The vowel sign *reversed i* is used for writing four Sanskrit vocalic letters.

U+11A07 ZANABAZAR SQUARE VOWEL SIGN AI and U+11A08 ZANABAZAR SQUARE VOWEL SIGN AU represent the diphthongs *ai* and *au*. They also function as secondary vowel signs for *i* and *u* to produce additional diphthongs in Mongolian.

Consonants. There are 40 consonants, including the following:

- U+11A26 ZANABAZAR SQUARE LETTER DZHA represents Sanskrit *jha*
- U+11A29 ZANABAZAR SQUARE LETTER -A represents Tibetan *’a chung*
- U+11A32 ZANABAZAR SQUARE LETTER KSSA represents Sanskrit cluster *kṣa* (/kṣa/)

Consonant clusters are written as conjuncts, which are generally rendered as vertical stacks, with each non-initial letter subjoined sequentially beneath the initial letter of the cluster.

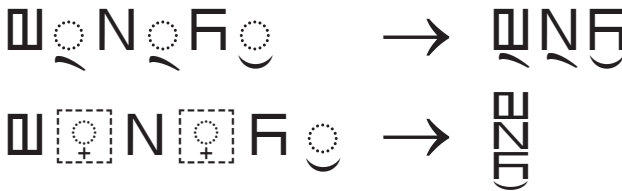
The consonants *ya*, *ra*, *la*, *va* have different representations when they occur in Sanskrit and Tibetan conjuncts. Therefore, contextual forms of these letters are encoded as separate characters.

Virama and Subjoiner. U+11A34 ZANABAZAR SQUARE SIGN VIRAMA is used to silence the inherent vowel of a consonant for writing Sanskrit and Tibetan. The virama is used only with a consonant and behaves as other combining marks in the script, always with a visible display.

Vowel-silencing characters in Brahmi-based scripts often have a secondary function of controlling conjunct formation, however, the Zanabazar Square script does not follow this pattern. A separate character, U+11A47 ZANABAZAR SQUARE SUBJOINER, is used to control conjunct formation.

The representation of a vertical conjunct stack uses the *subjoiner* character between each consonant of the cluster. For example, the syllable *mstu* is represented with the sequence <ma, subjoiner, sa, subjoiner, ta, vowel sign ue>, as shown in the second line of *Figure 14-8*. To suppress the visual stacking of a cluster, the *virama* character is used instead, which kills the vowel and results in a visual marking of the dead consonant which does not stack. For example, if the syllable *mstu* is represented with the sequence <ma, virama, sa, virama, ta, vowel sign ue>, the rendering is as shown in the first row of *Figure 14-8*.

Figure 14-8. Conjunct Stacking in Zanabazar Square



Head Marks. There are four head marks in the Zanabazar Square script. These four head marks are used in transliterations of Tibetan texts when written with the Zanabazar Square script. They occur at the beginning of texts.

- U+11A3F ZANABAZAR SQUARE INITIAL HEAD MARK
- U+11A40 ZANABAZAR SQUARE CLOSING HEAD MARK
- U+11A45 ZANABAZAR SQUARE INITIAL DOUBLE-LINED HEAD MARK
- U+11A46 ZANABAZAR SQUARE CLOSING DOUBLE-LINED HEAD MARK

Both U+11A3F ZANABAZAR SQUARE INITIAL HEAD MARK and U+11A45 ZANABAZAR SQUARE INITIAL DOUBLE-LINED HEAD MARK are used as a base for candrabindu and anusvara signs.

The U+11A40 ZANABAZAR SQUARE CLOSING HEAD MARK and U+11A46 ZANABAZAR SQUARE CLOSING DOUBLE-LINED HEAD MARK may be used for producing extended head marks, similar to usage in Tibetan.

Other Marks. Two vowel modifiers are used to transliterate words of Sanskrit origin:

- U+11A38 ZANABAZAR SQUARE SIGN ANUSVARA indicates nasalization
- U+11A39 ZANABAZAR SQUARE SIGN VISARGA indicates post-vocalic aspiration

In addition, three combining signs are used as nasalization marks and ornaments for the head mark:

- U+11A35 ZANABAZAR SQUARE SIGN CANDRABINDU
- U+11A36 ZANABAZAR SQUARE SIGN CANDRABINDU WITH ORNAMENT
- U+11A37 ZANABAZAR SQUARE SIGN CANDRA WITH ORNAMENT

The U+11A33 ZANABAZAR SQUARE FINAL CONSONANT MARK marks syllable-final consonants when writing Mongolian.

Numerals. There are no known script-specific numerals.

Punctuation. The Zanabazar Square script includes four punctuation marks used for writing Tibetan:

- U+11A41 ZANABAZAR SQUARE MARK TSHEG indicates the end of a syllable
- U+11A42 ZANABAZAR SQUARE MARK SHAD indicates the end of the phrase or sentence
- U+11A43 ZANABAZAR SQUARE MARK DOUBLE SHAD marks the end of a text section
- U+11A44 ZANABAZAR SQUARE MARK LONG TSHEG behaves as a comma

14.7 Soyombo

Soyombo: U+11A50–U+11AAF

The Soyombo script is an historic script used to write Mongolian, Sanskrit, and Tibetan. It was created in 1686 by Zanabazar (1635–1723), who also developed the Zanabazar Square script. The script appears primarily in Buddhist texts in Central Asia. Most of these texts consist of either handwritten manuscripts or inscriptions.

Structure. Soyombo is an *abugida*. Consonants generally include an inherent vowel /a/, as is the case with many other Brahmi-derived scripts. The script also includes final consonant signs and four cluster-initial letters. A special subjoiner is employed to create conjuncts.

Soyombo text is typically written horizontally left-to-right. In vertically written text, characters are oriented in columns laid out left-to-right, with upright glyphs.

The graphical structure of Soyombo letters consists of two parts: a frame, made up of a vertical bar with a triangle at the top, and a nucleus that represents a phoneme. Together the frame and the nucleus represent the atomic letter. Vowel signs, final consonants, and other phonetic features appear as dependent signs attached to the letters. The signs may appear above or to the right of the frame, or below the nucleus.

Vowels and Diphthongs. The vowel *a* is represented by U+11A50 SOYOMBO LETTER A. When it occurs with a vowel sign, SOYOMBO LETTER A serves as a vowel-carrier, indicating an independent vowel. Long vowels are represented by appending U+11A5B SOYOMBO VOWEL LENGTH MARK. When used to write Mongolian, U+11A57 SOYOMBO VOWEL SIGN AI and U+11A58 SOYOMBO VOWEL SIGN AU are used with other vowel signs to represent diphthongs.

Consonants. Mongolian syllable-final consonants are represented by U+11A50 SOYOMBO LETTER A followed by a final consonant sign. To indicate geminated consonants, U+11A98 SOYOMBO GEMINATION MARK is stacked above the triangle of the frame. In the backing store, it occurs immediately after the base letter, but before any other combining mark. Other above-base signs are shown above the gemination mark.

Generally, consonant clusters are written as a conjunct forms. Because Soyombo does not have a native virama, a special subjoiner character, U+11A99 SOYOMBO SUBJOINER, is used. Conjuncts are represented by using a subjoiner between each pair of consonants in a cluster. A conjunct is rendered as a vertical stack of the regular form of the initial letter and the nucleus of each non-initial letter. Four cluster-initial letters have special forms: *la*, *sha*, *sa* and *ra*. Depending upon the context, clusters involving these four letters may be rendered using the stacked or prefixed forms. The consonant cluster *kssa* has the structure of an atomic letter, and is separately encoded as U+11A83 SOYOMBO LETTER KSSA.

Character Names. The character names are based on their values for writing Tibetan, with the exception of the final consonant signs, which reflect their Mongolian usage. The order

of the consonant letters follows the alphabetical order of the Tibetan script. This also matches the order of letters in the Zanabazar Square script.

Other Marks. Two vowel modifiers are used to transliterate words of Sanskrit origin, U+11A96 SOYOMBO SIGN ANUSVARA, which indicates nasalization, and U+11A97 SOYOMBO SIGN VISARGA, which is used to indicate post-vocalic aspiration. Independent forms of these modifiers are represented by combining them with U+11A50 SOYOMBO LETTER A.

Numerals. There are no known script-specific numerals.

Punctuation. The Soyombo script includes a number of punctuation marks. U+11A9A SOYOMBO MARK TSHEG indicates the end of a syllable, and corresponds to U+0F0B TIBETAN MARK INTERSYLLABIC TSHEG. To indicate the end of a phrase or syllable, U+11A9B SOYOMBO MARK SHAD may be employed. It corresponds to U+0F0D TIBETAN MARK SHAD and U+0964 DEVANAGARI DANDA. The end of a section is marked by U+11A9C SOYOMBO MARK DOUBLE SHAD, corresponding to U+0F0E TIBETAN MARK NYIS SHAD and U+0965 DEVANAGARI DOUBLE DANDA.

The script also contains three head marks, similar to those used in Mongolian and Tibetan. The Soyombo marks may be followed by a *shad* or *double shad*. The U+11A9E SOYOMBO HEAD MARK WITH MOON AND SUN AND TRIPLE FLAME, also known as the Svayambhu or “Soyombo” sign, is the official symbol of Mongolia. In addition, the script includes terminal marks, which appear at the end of text.

14.8 Old Turkic

Old Turkic: U+10C00–U+10C4F

The origins of the Old Turkic script are unclear, but it seems to have evolved from a non-cursive form of the Sogdian script, one of the Aramaic-derived scripts used to write Iranian languages, in order to write the Old Turkish language. Old Turkic is attested in stone inscriptions from the early eighth century CE found around the Orkhon River in Mongolia, and in a slightly different version in stone inscriptions of the later eighth century found in Siberia near the Yenisei River and elsewhere. These inscriptions are the earliest written examples of a Turkic language. By the ninth century the Old Turkic script had been supplanted by the Uyghur script.

Because Old Turkic characters superficially resemble Germanic runes, the script is also known as Turkic Runes and Turkic Runiform, in addition to the names Orkhon script, Yenisei script, and Siberian script.

Where the Orkhon and Yenisei versions of a given Old Turkic letter differ significantly, each is separately encoded.

Structure. Old Turkish vowels can be classified into two groups based on their front or back articulation. A given word uses vowels from only one of these groups; the group is indicated by the form of the consonants in the word, because most consonants have separate forms to match the two vowel types. Other phonetic rules permit prediction of rounded and unrounded vowels, and high, medium or low vowels within a word. Some consonants also indicate that the preceding vowel is a high vowel. Thus, most initial and medial vowels are not explicitly written; only vowels that end a word are always written, and there is sometimes ambiguity about whether a vowel precedes a given consonant.

Ligature. Old Turkic includes one ligature, which is used to represent [tʃi]. It should be represented as:

$$\begin{array}{ccccccc} \lambda & + & \boxed{\begin{array}{c} ZW \\ J \end{array}} & + & \uparrow & \rightarrow & \lambda \\ 10C32 & & 200D & & 10C03 & & \end{array}$$

Directionality. For horizontal writing, the Old Turkic script is written from right to left within a row, with rows running from bottom to top. Conformant implementations of Old Turkic script must use the Unicode Bidirectional Algorithm (see Unicode Standard Annex #9, “Unicode Bidirectional Algorithm”).

In some cases, under Chinese influence, the layout was rotated ninety degrees counter-clockwise to produce vertical columns of text in which the characters are read top to bottom within a column, and the columns are read right to left.

Punctuation. Word division and some other punctuation functions are usually indicated by a two-dot mark similar to a colon; U+205A TWO DOT PUNCTUATION may be used to represent this punctuation mark. In some cases a mark such as U+2E30 RING POINT is used instead.

14.9 Old Sogdian

Old Sogdian: U+10F00–U+10F2F

The Old Sogdian script is used to represent a group of related writing systems of Central Asia dating from the third to the sixth century CE. These writing systems were all used to write Sogdian, an eastern Iranian language. Old Sogdian is based on four sets of written materials: the Kultobe inscriptions in modern Kazakhstan; the preserved epistles called the “Ancient Letters,” which are the earliest attested Sogdian manuscripts found in Dunhuang, China; inscriptions from the Upper Indus area of Pakistan; and inscriptions found on coins and vessels around Tashkent, Uzbekistan.

Repertoire. The basic repertoire consists of 20 of the 22 letters of the Aramaic alphabet. However, some of the original Aramaic letters ceased to be distinct in Old Sogdian. In the Ancient Letters, the usual glyph for *resh* is identical to the glyph for *daleth* and for *ayin*. As a result, *resh*, *ayin* and *daleth* are unified as a single character, U+10F18 OLD SOGDIAN LETTER RESH-AYIN-DALETH. In addition, the Old Sogdian repertoire includes six final letters, three final letters with vertical tail, and one alternate letter, U+10F13 OLD SOGDIAN LETTER ALTERNATE AYIN. The script also includes one heterogram, U+10F27 OLD SOGDIAN LIGATURE AYIN-DALETH, meaning “to,” used in salutations in the Ancient Letters.

Structure. Old Sogdian is a non-joining *abjad*, like Hebrew. The letters retain their shape within a word, and six letters, *aleph*, *beth*, *he*, *nun*, *sadhe*, and *taw*, have distinctive word-final forms. Adjacent letters may connect or overlap due to cursive writing, but unlike the later Sogdian script, letters do not change their shape based on word position.

Orientation. Most Old Sogdian text is written right to left, in lines running from top to bottom. Some Upper Indus inscriptions are written vertically, with the letters rotated ninety degrees counter-clockwise, in columns running from left to right. As a result of this behavior in vertical writing, Old Sogdian characters are given the Vertical_Orientation property value R.

Numbers. Ten Sogdian-specific numbers and fractions are encoded in the range U+10F1D..U+10F26.

Punctuation. No script-specific punctuation marks have been attested.

14.10 Sogdian

Sogdian: U+10F30–U+10F6F

Derived from Old Sogdian, the Sogdian script was used from the seventh to the fourteenth century CE in Central Asia to write the eastern Iranian language Sogdian. It was also used to write Chinese, Sanskrit, and Uyghur. Sogdian is the ancestor of the Mongolian and Old Uyghur scripts. It is attested in manuscripts and inscribed on coins, stone, pottery, and other media. The script has two major styles: “formal,” used in Buddhist *sutra* manuscripts, and a simplified, “cursive” style. The Old Uyghur script is believed to have derived from the Sogdian cursive style in the eighth or ninth century CE.

Structure. Sogdian is an *abjad* that can be written horizontally from right to left, or vertically from top to bottom, in columns running from left to right. When the script appears in vertical orientation, the glyphs are rotated ninety degrees counter-clockwise. Unlike Old Sogdian, Sogdian is a cursive joining script. Eleven combining signs in the range U+10F46..U+10F50 are used for disambiguation and transcription.

The Sogdian repertoire corresponds to that of Old Sogdian, but has a number of differences in the glyphs and also has additional characters. Sogdian has a special form of *ayin* for an Aramaic heterogram, and includes two characters not found in Old Sogdian, *feth* and *lesh*. The letter *feth* is used to represent [f]. *Lesh* or “hooked resh” is an extension of *resh-ayin* with a below-base hook that has become an intrinsic part of the letter. The repertoire includes one phonogram, U+10F45 SOGDIAN INDEPENDENT SHIN, an alternate form of isolated *shin*, used to transcribe one Chinese character, U+6240 所. The glyph for *ayin* is identical to the glyph for *resh*; therefore the two letters have been unified as a single character, U+10F40 SOGDIAN LETTER RESH-AYIN.

Glyphs. The representative glyphs are generally based on the isolated or independent form of letters found in the formal style of Sogdian. Fonts may be used to show the formal or cursive style of a text. As in other *abjads*, the letters connect and change shape based on their position within a word. In the later Sogdian styles, some letters, such as *nun*, *gimel* and *beth*, remain unconnected from a following letter to distinguish them from similar shapes.

Numbers. The Sogdian script includes script-specific numbers encoded in the range U+10F51..U+10F54.

Punctuation. Five script-specific punctuation characters are included in the repertoire. The four Sogdian punctuation characters, U+10F55 SOGDIAN PUNCTUATION TWO VERTICAL BARS, U+10F56 SOGDIAN PUNCTUATION TWO VERTICAL BARS WITH DOTS, U+10F57 SOGDIAN PUNCTUATION CIRCLE WITH DOT and U+10F58 SOGDIAN PUNCTUATION TWO CIRCLES WITH DOTS, delimit text segments and may vary in shape. U+10F59 SOGDIAN PUNCTUATION HALF CIRCLE WITH DOT generally indicates the completion of a text. Various other punctuation marks occur in Sogdian texts, and in some cases may be represented by punctuation characters from other blocks, such as General Punctuation.

14.11 Old Uyghur

Old Uyghur: U+10F70–U+10FAF

The historical Old Uyghur script flourished between the 8th and 17th centuries, primarily in the Tarim Basin of northwest China and other parts of Asia. The script was originally used to write medieval Turkish languages, but was later expanded to write other languages, including Chinese, Mongolian, Tibetan and Arabic. Old Uyghur developed from the cursive style of the Sogdian script (see *Section 14.10, Sogdian*) and is the ancestor of the Mongolian script (see *Section 13.5, Mongolian*).

The script has two main styles. “Square” style is a formal, book style where the letters are carefully written out. The square style is found in manuscripts, official documents, and in block printing for religious and literary texts. The second main style is “cursive,” used for rapid writing, particularly for administrative documents, as well as religious and literary texts. Other styles also developed, such as “post-Mongolic,” which was employed for literary and civil documents after the 14th century.

Structure. Old Uyghur is a cursive joining alphabet. The default orientation of the script is horizontal, with the script being read right-to-left. Although the script is traditionally laid out vertically in columns that run left to right, horizontal orientation facilitates the handling of Old Uyghur in multilingual contexts. Texts with vertical orientation should be handled by vertical text layout.

Repertoire. Based on evidence from 9th century documents, the Old Uyghur repertoire contained 15 consonants and three additional letters—*aleph*, *waw* and *yodh*—used to mark long vowels. The letters *aleph*, *waw* and *yodh* combine as digraphs and trigraphs to represent vowels of the Turkic languages.

Over time, some Old Uyghur letters fell together. For example, in the 11th century *samekh* and *shin* were both represented by *shin*. Diacritics were used to distinguish the merged letters: *samekh* was written using U+10F7F OLD UYGHUR LETTER SHIN, and *shin* was written with <U+10F7F OLD UYGHUR LETTER SHIN, U+10F85 OLD UYGHUR COMBINING TWO DOTS BELOW>. The reading of Old Uyghur text may be ambiguous due to the merger of letters and the nature of rapid, cursive writing. This ambiguity can be addressed using markup.

Representative Glyphs. The representative glyphs are based on the isolated form of the square style letters. Contextual forms of the letters are based on normalized shapes of the square style and from block prints. The square and cursive styles are not encoded separately. Fonts should handle the different styles, which can vary across regions and time. The terminals of many Old Uyghur letters, such as *aleph* and *beth*, may have different orientations and should be treated as glyph variants.

Shaping Behavior. Most Old Uyghur characters are dual-joining, except *zayin* and *heth*, which are right-joining.

Punctuation. U+10F86 OLD UYGHUR PUNCTUATION BAR and U+10F87 OLD UYGHUR PUNCTUATION TWO BARS delimit text sections for shorter and longer sections, respectively. In a

similar way, U+10F88 OLD UYGHUR PUNCTUATION TWO DOTS separates shorter text units and U+10F89 OLD UYGHUR PUNCTUATION FOUR DOTS separates longer sections. The script also uses a sign that is unified with U+10AF2 MANICHAEAN PUNCTUATION DOUBLE DOT WITHIN DOT.

Word boundaries are indicated by spaces. In documents with the square script, letters with extended horizontal terminals may be stretched to touch the initial letter of the following word. However, this behavior reflects no semantic distinction, and in plain text spaces should be used between words. To represent the joining of the two words calligraphically, U+200C ZERO WIDTH NON-JOINER may be used. Some texts extend the initial baseline to fill out the space on a line. For example, the space between the last word in the line and the margin may be filled by using U+0640 ARABIC TATWEEL between the last two letters of a word.

Other Signs. Four diacritics with dots encoded in the range U+10F82..U+10F85 differentiate merged letters and indicate sounds for which no distinct letter exists. The diacritics, whose shapes may vary across different script styles, commonly occur with *nun*, *gimel*, *heth*, *samekh*, and *shin*. No script-specific digits have been encoded.