The Unicode Standard
Version 6.2 – 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 6.2.
Includes bibliographical references and index.
ISBN 978-1-936213-07-8) (http://www.unicode.org/versions/Unicode6.2.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2012

ISBN 978-1-936213-07-8
Published in Mountain View, CA
September 2012

Chapter 13

Additional Modern Scripts

This chapter contains a collection of additional scripts in modern use that do not fit well into the script categories featured in other chapters:

Ethiopic	*Vai*	*Deseret*
Mongolian	*Bamum*	*Shavian*
Osmanya	*Cherokee*	*Lisu*
Tifinagh	*Canadian Aboriginal Syllabics*	*Miao*
N'Ko		

Ethiopic, Mongolian, and Tifinagh are scripts with long histories. Although their roots can be traced back to the original Semitic and North African writing systems, they would not be classified as Middle Eastern scripts today.

The remaining scripts in this chapter have been developed relatively recently. Some of them show roots in Latin and other letterforms, including shorthand. They are all original creative contributions intended specifically to serve the linguistic communities that use them.

Osmanya is an alphabetic script developed in the early 20th century to write the Somali language. N'Ko is a right-to-left alphabetic script devised in 1949 as a writing system for Manden languages in West Africa. Vai is a syllabic script used for the Vai language in Liberia and Sierra Leone; it was developed in the 1830s, but the standard syllabary was published in 1962. Bamum is a syllabary developed between 1896 and 1910, used for writing the Bamum language in western Cameroon.

The Cherokee script is a syllabary developed between 1815 and 1821, to write the Cherokee language, still spoken by small communities in Oklahoma and North Carolina. Canadian Aboriginal Syllabics were invented in the 1830s for Algonquian languages in Canada. The system has been extended many times, and is now actively used by other communities, including speakers of Inuktitut and Athapascan languages.

Deseret is a phonemic alphabet devised in the 1850s to write English. It saw limited use for a few decades by members of The Church of Jesus Christ of Latter-day Saints. Shavian is another phonemic alphabet, invented in the 1950s to write English. It was used to publish one book in 1962, but remains of some current interest.

The Lisu script was developed in the early 20th century by using a combination of Latin letters, rotated Latin letters, and Latin punctuation repurposed as tone letters, to create a writing system for the Lisu language, spoken by large communities, mostly in Yunnan province in China. It sees considerable use in China, where it has been an official script since 1992.

The Miao script was created in 1904 by adapting Latin letter variants, English shorthand characters, Miao pictographs, and Cree syllable forms. The script was originally developed to write the Northeast Yunnan Miao language of southern China. Today it is also used to write other Miao dialects and the languages of the Yi and Lisu nationalities of southern China.

13.1 Ethiopic

Ethiopic: U+1200–U+137F

The Ethiopic syllabary originally evolved for writing the Semitic language Ge'ez. Indeed, the English noun "Ethiopic" simply means "the Ge'ez language." Ge'ez itself is now limited to liturgical usage, but its script has been adopted for modern use in writing several languages of central east Africa, including Amharic, Tigre, and Oromo.

Basic and Extended Ethiopic. The Ethiopic characters encoded here include the basic set that has become established in common usage for writing major languages. As with other productive scripts, the basic Ethiopic forms are sometimes modified to produce an extended range of characters for writing additional languages.

Encoding Principles. The syllables of the Ethiopic script are traditionally presented as a two-dimensional matrix of consonant-vowel combinations. The encoding follows this structure; in particular, the codespace range U+1200..U+1357 is interpreted as a matrix of 43 consonants crossed with 8 vowels, making 344 conceptual syllables. Most of these consonant-vowel syllables are represented by characters in the script, but some of them happen to be unused, accounting for the blank cells in the matrix.

Variant Glyph Forms. A given Ethiopic syllable may be represented by different glyph forms, analogous to the glyph variants of Latin lowercase "a" or "g", which do not coexist in the same font. Thus the particular glyph shown in the code chart for each position in the matrix is merely one representation of that conceptual syllable, and the glyph itself is not the object that is encoded.

Labialized Subseries. A few Ethiopic consonants have labialized ("W") forms that are traditionally allotted their own consonant series in the syllable matrix, although only a subset of the possible vowel forms are realized. Each of these derivative series is encoded immediately after the corresponding main consonant series. Because the standard vowel series includes both "AA" and "WAA", two different cells of the syllable matrix might represent the "consonant + W + AA" syllable. For example:

U+1257 = QH + WAA: potential but unused version of QHWAA

U+125B = QHW + AA: ETHIOPIC SYLLABLE QHWAA

In these cases, where the two conceptual syllables are equivalent, the entry in the labialized subseries is encoded and not the "consonant + WAA" entry in the main syllable series. The six specific cases are enumerated in *Table 13-1*. In three of these cases, the -WAA position in the syllable matrix has been reanalyzed and used for encoding a syllable in -OA for extended Ethiopic.

Table 13-1. Labialized Forms in Ethiopic -WAA

-WAA Form	Encoded as	Not Used	Contrast
QWAA	U+124B ቋ	1247	U+1247 ቇ QOA
QHWAA	U+125B ቻ	1257	
XWAA	U+128B ኋ	1287	U+1287 ኇ XOA
KWAA	U+12B3 ኳ	12AF	U+12AF ኰ KOA
KXWAA	U+12C3 ዃ	12BF	
GWAA	U+1313 ጓ	130F	

Also, *within* the labialized subseries, the sixth vowel ("-E") forms are sometimes considered to be second vowel ("-U") forms. For example:

U+1249 = QW + U: unused version of QWE

U+124D = QW + E: ETHIOPIC SYLLABLE QWE

In these cases, where the two syllables are nearly equivalent, the "-E" entry is encoded and not the "-U" entry. The six specific cases are enumerated in *Table 13-2*.

Table 13-2. Labialized Forms in Ethiopic -WE

"-WE" Form	Encoded as	Not Used
QWE	U+124D ቍ	1249
QHWE	U+125D ቝ	1259
XWE	U+128D ኍ	1289
KWE	U+12B5 ኵ	12B1
KXWE	U+12C5 ዅ	12C1
GWE	U+1315 ጕ	1311

Keyboard Input. Because the Ethiopic script includes more than 300 characters, the units of keyboard input must constitute some smaller set of entities, typically 43+8 codes interpreted as the coordinates of the syllable matrix. Because these keyboard input codes are expected to be transient entities that are resolved into syllabic characters before they enter stored text, keyboard input codes are not specified in this standard.

Syllable Names. The Ethiopic script often has multiple syllables corresponding to the same Latin letter, making it difficult to assign unique Latin names. Therefore the names list makes use of certain devices (such as doubling a Latin letter in the name) merely to create uniqueness; this device has no relation to the phonetics of these syllables in any particular language.

Encoding Order and Sorting. The order of the consonants in the encoding is based on the traditional alphabetical order. It may differ from the sort order used for one or another language, if only because in many languages various pairs or triplets of syllables are treated as equivalent in the first sorting pass. For example, an Amharic dictionary may start out with a section headed by *three* H-like syllables:

U+1200 ETHIOPIC SYLLABLE HA

U+1210 ETHIOPIC SYLLABLE HHA

U+1280 ETHIOPIC SYLLABLE XA

Thus the encoding order cannot and does not implement a collation procedure for any particular language using this script.

Word Separators. The traditional word separator is U+1361 ETHIOPIC WORDSPACE (፡). In modern usage, a plain white wordspace (U+0020 SPACE) is becoming common.

Section Mark. One or more *section marks* are typically used on a separate line to mark the separation of sections. Commonly, an odd number is used and they are separated by spaces.

Diacritical Marks. The Ethiopic script generally makes no use of diacritical marks, but they are sometimes employed for scholarly or didactic purposes. In particular, U+135F ETHIOPIC COMBINING GEMINATION MARK and U+030E COMBINING DOUBLE VERTICAL LINE ABOVE are sometimes used to indicate emphasis or gemination (consonant doubling).

Numbers. Ethiopic digit glyphs are derived from the Greek alphabet, possibly borrowed from Coptic letterforms. In modern use, European digits are often used. The Ethiopic

number system does not use a zero, nor is it based on digital-positional notation. A number is denoted as a sequence of powers of 100, each preceded by a coefficient (2 through 99). In each term of the series, the power 100^n is indicated by n HUNDRED characters (merged to a digraph when n = 2). The coefficient is indicated by a *tens* digit and a *ones* digit, either of which is absent if its value is zero.

For example, the number 2345 is represented by

2345 = (20 + 3)*100^1 + (40 + 5)*100^0

= 20 3 100 40 5

= TWENTY THREE HUNDRED FORTY FIVE

= 1373 136B 137B 1375 136D ፳፫፻፵፭

A language using the Ethiopic script may have a *word* for "thousand," such as Amharic "SHI" (U+123A), and a quantity such as 2,345 may also be written as it is spoken in that language, which in the case of Amharic happens to parallel English:

2,345 = TWO thousand THREE HUNDRED FORTY FIVE

= 136A 123A 136B 137B 1375 136D ፪ሺ፫፻፵፭

Ethiopic Extensions

The Ethiopic script is used for a large number of languages and dialects in Ethiopia and in some instances has been extended significantly beyond the set of characters used for major languages such as Amharic and Tigre. There are three blocks of extensions to the Ethiopic script: Ethiopic Supplement U+1380..U+139F, Ethiopic Extended U+2D80..U+2DDF, and Ethiopic Extended-A U+AB00..U+AB2F. Those extensions cover such languages as Me'en, Blin, and Sebatbeit, which use many additional characters. The Ethiopic Extended-A block, in particular, includes characters for the Gamo-Gofa-Dawro, Basketo, and Gumuz languages. Several other characters for Ethiopic script extensions can be found in the main Ethiopic script block in the range U+1200..U+137F, including combining diacritic marks used for Basketo.

The Ethiopic Supplement block also contains a set of tonal marks. They are used in multiline scored layout. Like other musical (an)notational systems of this type, these tonal marks require a higher-level protocol to enable proper rendering.

13.2 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 letter forms 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 13-1*. There are many such examples in Mongolian, but not in Todo, Manchu, or Sibe, which have largely eliminated ambiguous letters.

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 13-2*, where the letter *ba* combines with the letter *u* to form a ligature in the Mongolian word *abu* "father."

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

Figure 13-1. Mongolian Glyph Convergence

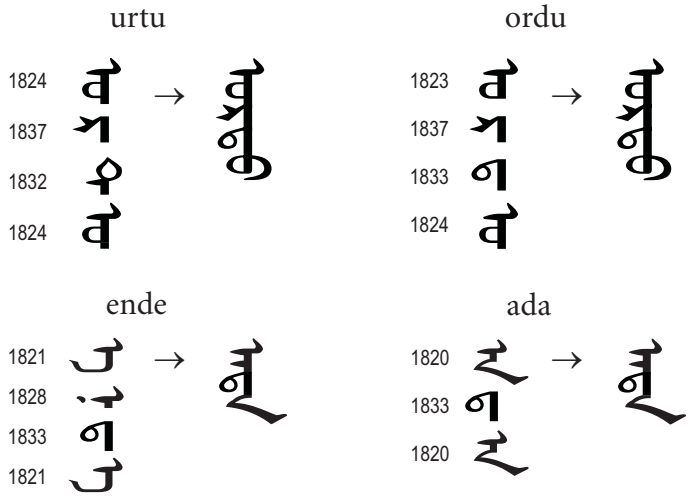
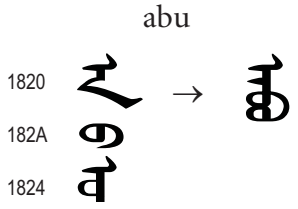


Figure 13-2. Mongolian Consonant Ligation



often the same as the initial form, but the final form is always distinct from the initial or medial form. *Figure 13-3* 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 13-3. 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 16, 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)

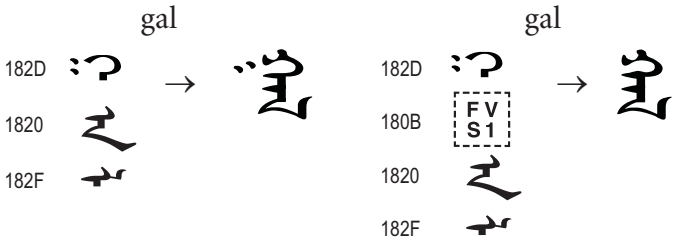
U+180C MONGOLIAN FREE VARIATION SELECTOR TWO (FVS2)

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 13-4 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 13-4. 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 16.4, Variation Selectors*. Older implementations of Mongolian free variation selectors may therefore interpret a sequence such as a base character followed by 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.

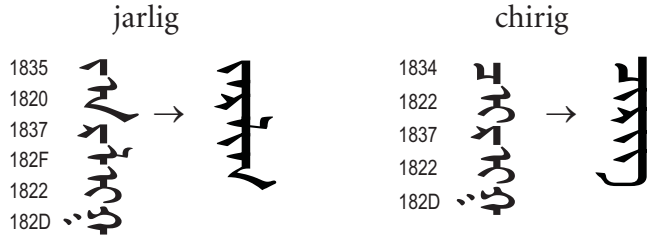
Vowel harmony is an important element of the encoding model, as the gender of a word determines the glyph form of the velar series of consonant letters for Mongolian, Todo, Sibe, and Manchu. In each script, the velar letters have both masculine and feminine forms. For Mongolian and Todo, the masculine 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 13-5*, 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.

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.

Figure 13-5. Mongolian Gender Forms

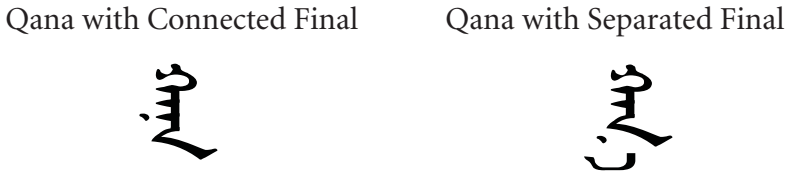


As any suffixes are considered to be an integral part of the word as a whole, a line break opportunity does not occur before a suffix, and the whitespace is represented using U+202F NARROW NO-BREAK SPACE (NNBSP). For a Mongolian font it is recommended that the width of NNBSP should be one-third the width 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 separated from the preceding consonant by a narrow gap, in which case the vowel always takes the "forward tail" form. U+180E MONGOLIAN VOWEL SEPARATOR (MVS) is used to represent the whitespace that separates a final letter *a* or *e* from the rest of the word. MVS is very similar in function to NNBSP, as it divides a word with a narrow non-breaking whitespace. 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. For example, the word *qana* <182C, 1820, 1828, 1820> without a gap before the final letter *a* means "the outer casing of a vein," whereas the word *qana* <182C, 1820, 1828, 180E, 1820> with a gap before the final letter *a* means "the wall of a tent," as shown in *Figure 13-6*.

Figure 13-6. Mongolian Vowel Separator



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.

13.3 Osmanya

Osmanya: U+10480–U+104AF

The Osmanya script, which in Somali is called 𐒍𐒖𐒇 𐒈𐒝𐒑𐒛𐒐𐒘 *far Soomaali* “Somali writing” or 𐒋𐒘𐒈𐒑𐒛𐒒𐒕𐒖 *Cismaanya*, was devised in 1920–1922 by 𐒋𐒘𐒈𐒑𐒛𐒒 𐒕𐒚𐒚𐒈𐒚𐒍 𐒏𐒜𐒒𐒖𐒆𐒘𐒘𐒆 (Cismaan Yuusuf Keenadiid) to represent the Somali language. It replaced an attempt by Sheikh Uweys of the Confraternity Qadiriyyah (died 1909) to devise an Arabic-based orthography for Somali. It has, in turn, been replaced by the Latin orthography of Muuse Xaaji Ismaaciil Galaal (1914–1980). In 1961, both the Latin and the Osmanya scripts were adopted for use in Somalia, but in 1969 there was a coup, with one of its stated aims being the resolution of the debate over the country’s writing system. A Latin orthography was finally adopted in 1973. Gregersen (1977) states that some 20,000 or more people use Osmanya in private correspondence and bookkeeping, and that several books and a biweekly journal *Horseed* (“*Vanguard*”) were published in cyclostyled format.

Structure. Osmanya is an alphabetic script, read from left to right in horizontal lines running from top to bottom. It has 22 consonants and 8 vowels. Unique long vowels are written for U+1049B 𐒛 OSMANYA LETTER AA, U+1049C 𐒜 OSMANYA LETTER EE, and U+1049D 𐒝 OSMANYA LETTER OO; long *uu* and *ii* are written with the consonants U+10493 𐒓 OSMANYA LETTER WAW and U+10495 𐒕 OSMANYA LETTER YA, respectively.

Ordering. Alphabetical ordering is based on the order of the Arabic alphabet, as specified by Osman Abdihalim Yuusuf Osman Keenadiid. This ordering is similar to the ordering given in Diringer (1996).

Names and Glyphs. The character names used in the Unicode Standard are as given by Osman. The glyphs shown in the code charts are taken from *Afkeenna iyo fartysa* ("Our language and its handwriting") 1971.

13.4 Tifinagh

Tifinagh: U+2D30–U+2D7F

The Tifinagh script is used by approximately 20 million people who speak varieties of languages commonly called Berber or Amazigh. The three main varieties in Morocco are known as Tarifite, Tamazighe, and Tachelhite. In Morocco, more than 40% of the population speaks Berber. The Berber language, written in the Tifinagh script, is currently taught to approximately 300,000 pupils in 10,000 schools—mostly primary schools—in Morocco. Three Moroccan universities offer Berber courses in the Tifinagh script leading to a Master's degree.

Tifinagh is an alphabetic writing system. It uses spaces to separate words and makes use of Western punctuation.

History. The earliest variety of the Berber alphabet is Libyan. Two forms exist: a Western form and an Eastern form. The Western variety was used along the Mediterranean coast from Kabylia to Morocco and most probably to the Canary Islands. The Eastern variety, Old Tifinagh, is also called Libyan-Berber or Old Tuareg. It contains signs not found in the Libyan variety and was used to transcribe Old Tuareg. The word *tifinagh* is a feminine plural noun whose singular would be *tafniqt*; it means "the Phoenician (letters)."

Neo-Tifinagh refers to the writing systems that were developed to represent the Maghreb Berber dialects. A number of variants of Neo-Tifinagh exist, the first of which was proposed in the 1960s by the Académie Berbère. That variant has spread in Morocco and Algeria, especially in Kabylia. Other Neo-Tifinagh systems are nearly identical to the Académie Berbère system. The encoding in the Tifinagh block is based on the Neo-Tifinagh systems.

Source Standards. The encoding consists of four Tifinagh character subsets: the basic set of the Institut Royal de la Culture Amazighe (IRCAM), the extended IRCAM set, other Neo-Tifinagh letters in use, and modern Tuareg letters. The first subset represents the set of characters chosen by IRCAM to unify the orthography of the different Moroccan modern-day Berber dialects while using the historical Tifinagh script.

Ordering. The letters are arranged according to the order specified by IRCAM. Other Neo-Tifinagh and Tuareg letters are interspersed according to their pronunciation.

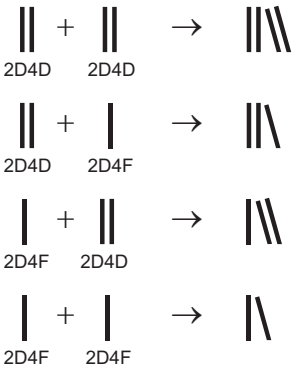
Directionality. Historically, Berber texts did not have a fixed direction. Early inscriptions were written horizontally from left to right, from right to left, vertically (bottom to top, top to bottom); boustrophedon directionality was also known. Modern-day Berber script is most frequently written in horizontal lines from left to right; therefore the bidirectional class for Tifinagh letters is specified as strong left to right. Displaying Berber texts in other

directions can be accomplished by the use of directional overrides or by the use of higher-level protocols.

Diacritical Marks. Modern Tifinagh variants tend to use combining diacritical marks to complement the Tifinagh block. The Hawad notation, for example, uses diacritical marks from the Combining Diacritical Marks block (U+0300–U+036F). These marks are used to represent vowels and foreign consonants. In this notation, <U+2D35, U+0307> represents "a", <U+2D49, U+0309> represents a long "i" /i:/, and <U+2D31, U+0302> represents a "p". Some long vowels are represented using two diacritical marks above. A long "e" /e:/ is thus written <U+2D49, U+0307, U+0304>. These marks are displayed side by side above their base letter in the order in which they are encoded, instead of being stacked.

Contextual Shaping. Contextual shaping of some consonants occurs when U+2D4D TIFINAGH LETTER YAL or U+2D4F TIFINAGH LETTER YAN are doubled or when both characters appear together in various sequences. The shaping distinguishes the characters when they appear next to each other. In contextual shaping, the second character is shifted vertically, or it can be slanted. *Figure 13-7* illustrates the use of contextual shaping.

Figure 13-7. Tifinagh Contextual Shaping



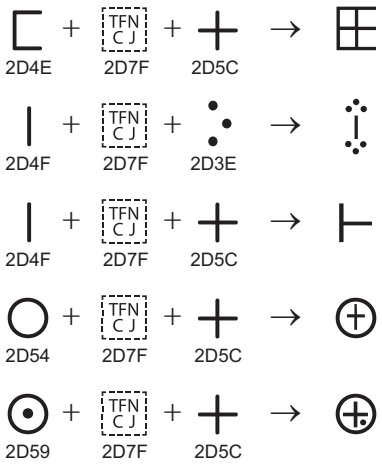
Bi-Consonants. Bi-consonants are additional letterforms used in the Tifinagh script, particularly for Tuareg, to represent a consonant cluster—a sequence of two consonants without an intervening vowel. These bi-consonants, sometimes also referred to as bigraphs, are not directly encoded as single characters in the Unicode Standard. Instead, they are represented as a sequence of the two consonant letters, separated either by U+200D ZERO WIDTH JOINER or by U+2D7F TIFINAGH CONSONANT JOINER.

When a bi-consonant is considered obligatory in text, it is represented by the two consonant letters, with U+2D7F TIFINAGH CONSONANT JOINER inserted between them. This use of U+2D7F is comparable in function to the use of U+0652 ARABIC SUKUN to indicate the absence of a vowel after a consonant, when Tuareg is written in the Arabic script. However, instead of appearing as a visible mark in the text, U+2D7F TIFINAGH CONSONANT JOINER indicates the presence of a bi-consonant, which should then be rendered with a preformed glyph for the sequence. Examples of common Tifinagh bi-consonants and their representation are shown in *Figure 13-8*.

If a rendering system cannot display obligatory bi-consonants with the correct, fully-formed bi-consonant glyphs, a fallback rendering should be used which displays the TIFINAGH CONSONANT JOINER visibly, so that the correct textual distinctions are maintained, even if they cannot be properly displayed.

When a bi-consonant is considered merely an optional, ligated form of two consonant letters, the bi-consonant can be represented by the two consonant letters, with U+200D ZERO WIDTH JOINER inserted between them, as a hint that the ligated form is preferred. If a ren-

Figure 13-8. Tifinagh Consonant Joiner and Bi-consonants



dering system cannot display the optional, ligated form, the fallback display should simply be the sequence of consonants, with no visible display of the ZWJ.

Bi-consonants often have regional glyph variants, so fonts may need to be designed differently for different regional uses of the Tifinagh script.

13.5 N'Ko

N'Ko: U+07C0–U+07FF

N'Ko is a literary dialect used by the Manden (or Manding) people, who live primarily in West Africa. The script was devised by Solomana Kante in 1949 as a writing system for the Manden languages. The Manden language group is known as *Mandenkan*, where the suffix *-kan* means "language of." In addition to the substantial number of Mandens, some non-Mandens speak *Mandenkan* as a second language. There are an estimated 20 million Mandenkan speakers.

The major dialects of the Manden language are Bamanan, Jula, Maninka, and Mandinka. There are a number of other related dialects. When Mandens from different subgroups talk to each other, it is common practice for them to switch—consciously or subconsciously—from their own dialect to the conventional, literary dialect commonly known as *Kangbe*, "the clear language," also known as N'Ko. This dialect switching can occur in conversations between the Bamanan of Mali, the Maninka of Guinea, the Jula of the Ivory Coast, and the Mandinka of Gambia or Senegal, for example. Although there are great similarities between their dialects, speakers sometimes find it necessary to switch to *Kangbe* (N'Ko) by using a common word or phrase, similar to the accommodations Danes, Swedes, and Norwegians sometimes make when speaking to one another. For example, the word for "name" in Bamanan is *togo*, while it is *tooh* in Maninka. Speakers of both dialects will write it as ߒ߬ , although each may pronounce it differently.

Structure. The N'Ko script is written from right to left. It is phonetic in nature (one symbol, one sound). N'Ko has seven vowels, each of which can bear one of seven diacritical marks that modify the tone of the vowel as well as an optional diacritical mark that indicates nasalization. N'Ko has 19 consonants and two "abstract" consonants, U+07E0 NKO LETTER NA WOLOSO and U+07E7 NKO LETTER NYA WOLOSO, which indicate original consonants mutated by a preceding nasal, either word-internally or across word boundaries.

Some consonants can bear one of three diacritical marks to transcribe foreign sounds or to transliterate foreign letters.

U+07D2 NKO LETTER N is considered neither a vowel nor a consonant; it indicates a syllabic alveolar or velar nasal. It can bear a diacritical mark, but cannot bear the nasal diacritic. The letter U+07D1 NKO LETTER DAGBASINNA has a special function in N'Ko orthography. The standard spelling rule is that when two successive syllables have the same vowel, the vowel is written only after the second of the two syllables. For example, ߕ߭ <ba, la, oo> is pronounced [bolo], but in a foreign syllable to be pronounced [blo], the *dagbasinna* is inserted for ߕ߮߭ <ba, dagbasinna, la, oo> to show that a consonant cluster is intended.

Digits. N'Ko uses decimal digits specific to the script. These digits have strong right-to-left directionality. Numbers are stored in text in logical order with most significant digit first; when displayed, numerals are then laid out in right-to-left order, with the most significant digit at the rightmost side, as illustrated for the numeral 144 in *Figure 13-9*. This situation differs from how numerals are handled in Hebrew and Arabic, where numerals are laid out in left-to-right order, even though the overall text direction is right to left.

Diacritical Marks. N'Ko diacritical marks are script-specific, despite superficial resemblances to other diacritical marks encoded for more general use. Some N'Ko diacritics have a wider range of glyph representation than the generic marks do, and are typically drawn rather higher and bolder than the generic marks.

Table 13-3 shows the use of the tone diacritics when applied to vowels.

Table 13-3. N'Ko Tone Diacritics on Vowels

Character		Tone	Applied To
U+07EB	NKO COMBINING SHORT HIGH TONE	high	short vowel
U+07EC	NKO COMBINING SHORT LOW TONE	low	short vowel
U+07ED	NKO COMBINING SHORT RISING TONE	rising-falling	short vowel
U+07EE	NKO COMBINING LONG DESCENDING TONE	descending	long vowel
U+07EF	NKO COMBINING LONG HIGH TONE	high	long vowel
U+07F0	NKO COMBINING LONG LOW TONE	long low	long vowel
U+07F1	NKO COMBINING LONG RISING TONE	rising	long vowel

When applied to a vowel, U+07F2 NKO COMBINING NASALIZATION MARK indicates the nasalization of that vowel. In the text stream, this mark is applied before any of the tone marks because combining marks below precede combining marks above in canonical order.

Two of the tone diacritics, when applied to consonants, indicate specific sounds from other languages—in particular, Arabic or French language sounds. U+07F3 NKO COMBINING DOUBLE DOT ABOVE is also used as a diacritic to represent sounds from other languages. The combinations used are as shown in *Table 13-4*.

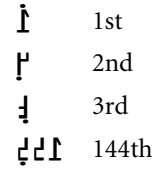
Ordinal Numbers. Diacritical marks are also used to mark ordinal numbers. The first ordinal is indicated by applying U+07ED NKO COMBINING SHORT RISING TONE (a dot above) to U+07C1 NKO DIGIT ONE. All other ordinal numbers are indicated by applying U+07F2 NKO COMBINING NASALIZATION MARK (an oval dot below) to the last digit in any sequence of digits composing the number. Thus the nasalization mark under the digit two would indicate the ordinal value 2nd, while the nasalization mark under the final digit four in the numeral 144 would indicate the ordinal value 144th, as shown in *Figure 13-9*.

Punctuation. N'Ko uses a number of punctuation marks in common with other scripts. U+061F ARABIC QUESTION MARK, U+060C ARABIC COMMA, U+061B ARABIC SEMICOLON, and the paired U+FD3E ORNATE LEFT PARENTHESIS and U+FD3F ORNATE RIGHT PARENTHESIS are used, often with different shapes than are used in Arabic. A script-specific

Table 13-4. Other N'Ko Diacritic Usage

Character	Applied To	Represents
U+07EB NKO COMBINING SHORT HIGH TONE	SA	[s] or Arabic ص SAD
	GBA	[ɣ] or Arabic غ GHAIN
	KA	[q] or Arabic ق QAF
U+07ED NKO COMBINING SHORT RISING TONE	BA	[b^h]
	TA	[ŧ] or Arabic ط TAH
	JA	[z] or Arabic ز ZAIN
	CA	[ð] or Arabic ذ THAL and also French [ʒ]
	DA	[đ] or Arabic ض ZAD
	RA	French [ʀ]
	SA	[ʃ] or Arabic ش SHEEN
	GBA	[g]
	FA	[v]
	KA	[ẖ] or Arabic خ KHAH
	LA	[l^h]
	MA	[m^h]
	NYA	[n^h]
	HA	[ẖ] or Arabic ح HAH
	YA	[y^h]
U+07F3 NKO COMBINING DOUBLE DOT ABOVE	A	[ʕa] or Arabic عَ AIN + A
	EE	French [ə]
	U	French [y]
	JA	[ƶ] or Arabic ظ ZAH
	DA	[d^h]
	SA	[θ] or Arabic ث THEH
	GBA	[kp]

Figure 13-9. Examples of N'Ko Ordinals



U+07F8 NKO COMMA and U+07F9 NKO EXCLAMATION MARK are encoded. The NKO COMMA differs in shape from the ARABIC COMMA, and the two are sometimes used distinctively in the same N'Ko text.

The character U+07F6 NKO SYMBOL OO DENNEN is used as an addition to phrases to indicate remote future placement of the topic under discussion. The decorative U+07F7 ♧ NKO SYMBOL GBAKURUNEN represents the three stones that hold a cooking pot over the fire and is used to end major sections of text.

The two tonal apostrophes, U+07F4 NKO HIGH TONE APOSTROPHE and U+07F5 NKO LOW TONE APOSTROPHE, are used to show the elision of a vowel while preserving the tonal information of the syllable. Their glyph representations can vary in height relative to the baseline. N'Ko also uses a set of paired punctuation, U+2E1C LEFT LOW PARAPHRASE BRACKET and U+2E1D RIGHT LOW PARAPHRASE BRACKET, to indicate indirect quotations.

Character Names and Block Name. Although the traditional name of the N'Ko language and script includes an apostrophe, apostrophes are disallowed in Unicode character and block names. Because of this, the formal block name is "NKo" and the script portion of the Unicode character names is "NKO".

Ordering. The order of N'Ko characters in the code charts reflects the traditional ordering of N'Ko. However, in collation, the three archaic letters U+07E8 NKO LETTER JONA JA, U+07E9 NKO LETTER JONA CHA, and U+07EA NKO LETTER JONA RA should be weighted as variants of U+07D6 NKO LETTER JA, U+07D7 NKO LETTER CHA, and U+07D9 NKO LETTER RA, respectively.

Rendering. N'Ko letters have shaping behavior similar to that of Arabic. Each letter can take one of four possible forms, as shown in *Table 13-5*.

Table 13-5. N'Ko Letter Shaping

Character	X_n	X_r	X_m	X_l
A	߀	߀	߀	߀
EE	߁	߁	߁	߁
I	߂	߂	߂	߂
E	߃	߃	߃	߃
U	߄	߄	߄	߄
OO	߅	߅	߅	߅
O	߆	߆	߆	߆
DAGBASINNA	◌	◌	◌	◌
N	߇	߇	߇	߇
BA	߈	߈	߈	߈
PA	߉	߉	߉	߉
TA	ߊ	ߊ	ߊ	ߊ
JA	ߋ	ߋ	ߋ	ߋ
CHA	ߌ	ߌ	ߌ	ߌ
DA	ߍ	ߍ	ߍ	ߍ
RA	ߎ	ߎ	ߎ	ߎ
RRA	ߏ	ߏ	ߏ	ߏ
SA	ߐ	ߐ	ߐ	ߐ
GBA	ߑ	ߑ	ߑ	ߑ
FA	ߒ	ߒ	ߒ	ߒ
KA	ߓ	ߓ	ߓ	ߓ
LA	ߔ	ߔ	ߔ	ߔ
NA WOLOSO	ߕ	ߕ	ߕ	ߕ
MA	ߖ	ߖ	ߖ	ߖ
NYA	ߗ	ߗ	ߗ	ߗ

Table 13-5. N'Ko Letter Shaping (Continued)

Character	X_n	X_r	X_m	X_l
NA	ߣ	ߣ	ߣ	ߣ
HA	ߤ	ߤ	ߤ	ߤ
WA	ߥ	ߥ	ߥ	ߥ
YA	ߦ	ߦ	ߦ	ߦ
NYA WOLOSO	ߧ	ߧ	ߧ	ߧ
JONA JA	ߨ	ߨ	ߨ	ߨ
JONA CHA	ߩ	ߩ	ߩ	ߩ
JONA RA	ߪ	ߪ	ߪ	ߪ

A noncursive style of N'Ko writing exists where no joining line is used between the letters in a word. This is a font convention, not a dynamic style like bold or italic, both of which are also valid dynamic styles for N'Ko. Noncursive fonts are mostly used as display fonts for the titles of books and articles. U+07FA NKO LAJANYALAN is sometimes used like U+0640 ARABIC TATWEEL to justify lines, although Latin-style justification where space is increased tends to be more common.

13.6 Vai

Vai: U+A500–U+A63F

The Vai script is used for the Vai language, spoken in coastal areas of western Liberia and eastern Sierra Leone. It was developed in the early 1830s primarily by Mọmọlu Duwalu Bukẹlẹ of Jondu, Liberia, who later stated that the inspiration had come to him in a dream. He may have also been aware of, and influenced by, other scripts including Latin, Arabic, and possibly Cherokee, or he may have phoneticized and regularized an earlier pictographic script. In the years afterward, the Vai built an educational infrastructure that enabled the script to flourish; by the late 1800s European traders reported that most Vai were literate in the script. Although there were standardization efforts in 1899 and again at a 1962 conference at the University of Liberia, nowadays the script is learned informally and there is no means to ensure adherence to a standardized version; most Vai literates know only a subset of the standardized characters. The script is primarily used for correspondence and record-keeping, mainly among merchants and traders. Literacy in Vai coexists with literacy in English and Arabic.

Sources. The primary sources for the Vai characters in Unicode are the 1962 Vai Standard Syllabary, modern primers and texts which use the Standard Syllabary (including a few glyph modifications reflecting modern preferences), the 1911 additions of Momolu Massaquoi, and the characters found in *The Book of Ndole*, the longest surviving text from the early period of Vai script usage.

Basic Structure. Vai is a syllabic script written left to right. The Vai language has seven oral vowels [e i a o u ɔ ɛ], five of which also occur in nasal form [ĩ ã ũ ɔ̃ ɛ̃]. The standard syllabary includes standalone vowel characters for the oral vowels and three of the nasal ones, characters for most of the consonant-vowel combinations formed from each of thirty consonants or consonant clusters, and a character for the final velar nasal consonant [ŋ].

The writing system has a *moraic* structure: the weight (or duration) of a syllable determines the number of characters used to write it (as with Japanese kana). A short syllable is written with any single character in the range U+A500..U+A60B. Long syllables are written with two characters, and involve a long vowel, a diphthong, or a syllable ending with U+A60B VAI SYLLABLE NG. Note that the only closed syllables in Vai—that is, those that end with a consonant—are those ending with VAI SYLLABLE NG. The long vowel is generally written using either an additional standalone vowel to double the vowel sound of the preceding character, or using U+A60C VAI SYLLABLE LENGTHENER, while the diphthong is generally written using an additional standalone vowel. In some cases, the second character for a long vowel or diphthong may be written using characters such as U+A54C VAI SYLLABLE HA or U+A54E VAI SYLLABLE WA instead of standalone vowels.

Historic Syllables. In *The Book of Ndole* more than one character may be used to represent a pronounced syllable; they have been separately encoded.

Logograms. The oldest Vai texts used an additional set of symbols called "logograms," representing complete syllables with an associated meaning or range of meanings; these symbols may be remnants from a precursor pictographic script. At least two of these symbols are still used: U+A618 VAI SYMBOL FAA represents the word meaning "die, kill" and is used alongside a person's date of death (the glyph is said to represent a wilting tree); U+A613 VAI SYMBOL FEENG represents the word meaning "thing."

Digits. In the 1920s ten decimal digits were devised for Vai; these digits were "Vai-style" glyph variants of European digits. They never became popular with Vai people, but are encoded in the standard for historical purposes. Modern literature uses European digits.

Punctuation. Vai makes use of European punctuation, although a small number of script-specific punctuation marks commonly occur. U+A60D VAI COMMA rests on or slightly below the baseline; U+A60E VAI FULL STOP rests on the baseline and can be doubled for use as an exclamation mark. U+A60F VAI QUESTION MARK also rests on the baseline; it is rarely used. Some modern primers prefer these Vai punctuation marks; some prefer the European equivalents. Some Vai writers mark the end of a sentence by using U+A502 VAI SYLLABLE HEE instead of punctuation.

Segmentation. Vai is written without spaces between words. Line breaking opportunities can occur between most characters except that line breaks should not occur before U+A60B VAI SYLLABLE NG used as a syllable final, or before U+A60C VAI SYLLABLE LENGTHENER (which is always a syllable final). Line breaks also should not occur before one of the "h-" characters (U+A502, U+A526, U+A54C, U+A573, U+A597, U+A5BD, U+A5E4) when it is used to extend the vowel of the preceding character (that is, when it is a syllable final), and line breaks should not occur before the punctuation characters U+A60D VAI COMMA, U+A60E VAI FULL STOP, and U+A60F VAI QUESTION MARK.

Ordering. There is no evidence of traditional conventions on ordering apart from the order of listings found in syllabary charts. The syllables in the Vai block are arranged in the order recommended by a panel of Vai script experts. Logograms should be sorted by their phonetic values.

13.7 Bamum

Bamum: U+A6A0–U+A6FF

The Bamum script is used for the Bamum language, spoken primarily in western Cameroon. It was developed between 1896 and 1910, mostly by King Ibrahim Njoya of the Bamum Kingdom. Apparently inspired by a dream and by awareness of other writing, his

original idea for the script was to collect and provide approximately 500 logographic symbols (denoting objects and actions) to serve more as a memory aid than as a representation of language.

Using the rebus principle, the script was rapidly simplified through six stages, known as Stage A, Stage B, and so on, into a syllabary known as *A-ka-u-ku*, consisting of 80 syllable characters or letters. These letters are used with two combining diacritics and six punctuation marks. The repertoire in this block covers the *A-ka-u-ku* syllabary, or Phase G form, which remains in modern use.

Structure. Modern Bamum is written left-to-right. One interesting feature is that sometimes more letters than necessary are used to write a given syllable. For example, the word *lam* "wedding" is written using the sequence of syllabic characters, *la* + *a* + *m*. This feature is known as pleonastic syllable representation.

Diacritical Marks. U+A6F0 BAMUM COMBINING MARK KOQNDON may be applied to any of the 80 letters. It usually functions to glottalize the final vowel of a syllable. U+A6F1 BAMUM COMBINING MARK TUKWENTIS is only known to be used with 13 letters—usually to truncate a full syllable to its final consonant.

Punctuation. U+A6F2 BAMUM NJAEMLI was a character used in the original set of logographic symbols to introduce proper names or to change the meaning of a word. The shape of the glyph for *njaemli* has changed, but the character is still in use. The other punctuation marks correspond in function to the similarly-named punctuation marks used in European typography.

Digits. The last ten letters in the syllabary are also used to represent digits. Historically, the last of these was used for 10, but its meaning was changed to represent zero when decimal-based mathematics was introduced.

Bamum Supplement: U+16800–U+16A3F

The Bamum Supplement block contains archaic characters no longer used in the modern Bamum orthography. These historical characters are analogous in some ways to the medievalist characters encoded for the Latin script. Most Bamum writers do not use them, but they are used by specialist linguists and historians.

The main source for the repertoire of Bamum extensions is an analysis in Dugast and Jeffreys 1950. The Bamum script was developed in six phases, labeled with letters. Phase A is the earliest form of the script. Phase G is the modern script encoded in the main Bamum block. The Bamum Supplement block covers distinct characters from the earlier phases which are no longer part of the modern Bamum script.

The character names in this block include a reference to the last phase in which they appear. So, for example, U+16867 BAMUM LETTER PHASE-B PIT was last used during Phase B, while U+168EE BAMUM LETTER PHASE-C PIN continued in use and is attested through Phase C.

Traditional Bamum texts using these historical characters do not use punctuation or digits. Numerical values for digits are written out as words instead.

13.8 Cherokee

Cherokee: U+13A0–U+13FF

The Cherokee script is used to write the Cherokee language. Cherokee is a member of the Iroquioan language family. It is related to Cayuga, Seneca, Onondaga, Wyandot-Huron,

Tuscarora, Oneida, and Mohawk. The relationship is not close because roughly 3,000 years ago the Cherokees migrated southeastward from the Great Lakes region of North America to what is now North Carolina, Tennessee, and Georgia. Cherokee is the native tongue of approximately 20,000 people, although most speakers today use it as a second language. The Cherokee word for both the language and the people is ᏣᎳᎩ *Tsalagi*.

The Cherokee syllabary, as invented by Sequoyah between 1815 and 1821, contained 6 vowels and 17 consonants. Sequoyah avoided copying from other alphabets, but his original letters were modified to make them easier to print. The first font for Cherokee was designed by Dr. Samuel A. Worcester. Using fonts available to him, he assigned a number of Latin letters to the Cherokee syllables. At this time the Cherokee letter "HV" was dropped, and the Cherokee syllabary reached its current size of 85 letters. Dr. Worcester's press printed 13,980,000 pages of Native American-language text, most of it in Cherokee.

Tones. Each Cherokee syllable can be spoken on one of four pitch or tone levels, or can slide from one pitch to one or two others within the same syllable. However, only in certain words does the tone of a syllable change the meaning. Tones are unmarked.

Case and Spelling. The Cherokee script is caseless, although for purposes of emphasis occasionally one letter will be made larger than the others. Cherokee spelling is not standardized: each person spells as the word sounds to him or her.

Numbers. Although Sequoyah invented a Cherokee number system, it was not adopted and is not encoded here. The Cherokee Nation uses European numbers. Cherokee speakers pay careful attention to the use of ordinal and cardinal numbers. When speaking of a numbered series, they will use ordinals. For example, when numbering chapters in a book, Cherokee headings would use First Chapter, Second Chapter, and so on, instead of Chapter One, Chapter Two, and so on.

Rendering and Input. Cherokee is a left-to-right script, which requires no combining characters. Several keyboarding conventions exist for inputting Cherokee. Some involve dead-key input based on Latin transliterations; some are based on sound-mnemonics related to Latin letters on keyboards; and some are ergonomic systems based on frequency of the syllables in the Cherokee language.

Punctuation. Cherokee uses standard Latin punctuation.

Standards. There are no other encoding standards for Cherokee.

13.9 Canadian Aboriginal Syllabics

Canadian Aboriginal Syllabics: U+1400–U+167F

The characters in this block are a unification of various local syllabaries of Canada into a single repertoire based on character appearance. The syllabics were invented in the late 1830s by James Evans for Algonquian languages. As other communities and linguistic groups adopted the script, the main structural principles described in this section were adopted. The primary user community for this script consists of several aboriginal groups throughout Canada, including Algonquian, Inuktitut, and Athapascan language families. The script is also used by governmental agencies and in business, education, and media.

Organization. The repertoire is organized primarily on structural principles found in the CASEC [1994] report, and is essentially a glyphic encoding. The canonical structure of each character series consists of a consonant shape with five variants. Typically the shape points down when the consonant is combined with the vowel /e/, up when combined with the vowel /i/, right when combined with the vowel /o/, and left when combined with the

vowel /a/. It is reduced and superscripted when in syllable-final position, not followed by a vowel. For example:

∨	∧	>	<	˂
PE	PI	PO	PA	P

Some variations in vowels also occur. For example, in Inuktitut usage, the syllable U+1450 ᑐ CANADIAN SYLLABICS TO is transcribed into Latin letters as "TU" rather than "TO", but the structure of the syllabary is generally the same regardless of language.

Arrangement. The arrangement of signs follows the Algonquian ordering (down-pointing, up-pointing, right-pointing, left-pointing), as in the previous example.

Sorted within each series are the variant forms for that series. Algonquian variants appear first, then Inuktitut variants, then Athapascan variants. This arrangement is convenient and consistent with the historical diffusion of Syllabics writing; it does not imply any hierarchy.

Some glyphs do not show the same down/up/right/left directions in the typical fashion—for example, beginning with U+146B ᑫ CANADIAN SYLLABICS KE. These glyphs are variations of the rule because of the shape of the basic glyph; they do not affect the convention.

Vowel length and labialization modify the character series through the addition of various marks (for example, U+143E ᐣ CANADIAN SYLLABICS PWII). Such modified characters are considered unique syllables. They are not decomposed into base characters and one or more diacritics. Some language families have different conventions for placement of the modifying mark. For the sake of consistency and simplicity, and to support multiple North American languages in the same document, each of these variants is assigned a unique code point.

Extensions. A few additional syllables in the range U+166F..U+167F at the end of this block have been added for Inuktitut, Woods Cree, and Blackfoot. Because these extensions were encoded well after the main repertoire in the block, their arrangement in the code charts is outside the framework for the rest of the characters in the block.

Punctuation and Symbols. Languages written using the Canadian Aboriginal Syllabics make use of the common punctuation marks of Western typography. However, a few punctuation marks are specific in form and are separately encoded as script-specific marks for syllabics. These include: U+166E CANADIAN SYLLABICS FULL STOP and U+1400 CANADIAN SYLLABICS HYPHEN.

There is also a special symbol, U+166D CANADIAN SYLLABICS CHI SIGN, used in religious texts as a symbol to denote Christ.

Canadian Aboriginal Syllabics Extended: U+18B0–U+18FF

This block contains many additional syllables attested in various local traditions of syllabics usage in Canada. These additional characters include extensions for several Algonquian communities (Cree, Moose Cree, and Ojibway), and for several Dene communities (Beaver Dene, Hare Dene, Chipewyan, and Carrier).

13.10 Deseret

Deseret: U+10400–U+1044F

Deseret is a phonemic alphabet devised to write the English language. It was originally developed in the 1850s by the regents of the University of Deseret, now the University of Utah. It was promoted by The Church of Jesus Christ of Latter-day Saints, also known as the "Mormon" or LDS Church, under Church President Brigham Young (1801–1877). The name *Deseret* is taken from a word in the Book of Mormon defined to mean "honeybee" and reflects the LDS use of the beehive as a symbol of cooperative industry. Most literature about the script treats the term *Deseret Alphabet* as a proper noun and capitalizes it as such.

Among the designers of the Deseret Alphabet was George D. Watt, who had been trained in shorthand and served as Brigham Young's secretary. It is possible that, under Watt's influence, Sir Isaac Pitman's 1847 English Phonotypic Alphabet was used as the model for the Deseret Alphabet.

The Deseret Alphabet was a work in progress through most of the 1850s, with the set of letters and their shapes changing from time to time. The final version was used for the printed material of the late 1860s, but earlier versions are found in handwritten manuscripts.

The Church commissioned two typefaces and published four books using the Deseret Alphabet. The Church-owned *Deseret News* also published passages of scripture using the alphabet on occasion. In addition, some historical records, diaries, and other materials were handwritten using this script, and it had limited use on coins and signs. There is also one tombstone in Cedar City, Utah, written in the Deseret Alphabet. However, the script failed to gain wide acceptance and was not actively promoted after 1869. Today, the Deseret Alphabet remains of interest primarily to historians and hobbyists.

Letter Names and Shapes. Pedagogical materials produced by the LDS Church gave names to all of the non-vowel letters and indicated the vowel sounds with English examples. In the Unicode Standard, the spelling of the non-vowel letter names has been modified to clarify their pronunciations, and the vowels have been given names that emphasize the parallel structure of the two vowel runs.

The glyphs used in the Unicode Standard are derived from the second typeface commissioned by the LDS Church and represent the shapes most commonly encountered. Alternate glyphs are found in the first typeface and in some instructional material.

Structure. The final version of the script consists of 38 letters, LONG I through ENG. Two additional letters, OI and EW, found only in handwritten materials, are encoded after the first 38. The alphabet is bicameral; capital and small letters differ only in size and not in shape. The order of the letters is phonetic: letters for similar classes of sound are grouped together. In particular, most consonants come in unvoiced/voiced pairs. Forty-letter versions of the alphabet inserted OI after AY and EW after OW.

Sorting. The order of the letters in the Unicode Standard is the one used in all but one of the nineteenth-century descriptions of the alphabet. The exception is one in which the letters WU and YEE are inverted. The order YEE-WU follows the order of the "coalescents" in Pitman's work; the order WU-YEE appears in a greater number of Deseret materials, however, and has been followed here.

Alphabetized material followed the standard order of the Deseret Alphabet in the code charts, except that the short and long vowel pairs are grouped together, in the order long vowel first, and then short vowel.

Typographic Conventions. The Deseret Alphabet is written from left to right. Punctuation, capitalization, and digits are the same as in English. All words are written phonemically with the exception of short words that have pronunciations equivalent to letter names, as shown in *Figure 13-10*.

Figure 13-10. Short Words Equivalent to Deseret Letter Names

- 𐐃 AY is written for *eye* or *I*
- 𐐯 YEE is written for *ye*
- 𐐁 BEE is written for *be* or *bee*
- 𐐘 GAY is written for *gay*
- 𐐾 THEE is written for *the* or *thee*

Phonetics. An approximate IPA transcription of the sounds represented by the Deseret Alphabet is shown in *Table 13-6*.

Table 13-6. IPA Transcription of Deseret

𐐀𐐨	LONG I	i	𐐁𐐩	BEE	b
𐐄𐐪	LONG E	e	𐐂𐐪	TEE	t
𐐆𐐫	LONG A	ɑ	𐐃𐐫	DEE	d
𐐈𐐬	LONG AH	ɒ	𐐅𐐬	CHEE	tʃ
𐐊𐐭	LONG O	o	𐐇𐐭	JEE	dʒ
𐐌𐐮	LONG OO	u	𐐋𐐮	KAY	k
𐐍𐐯	SHORT I	ɪ	𐐏𐐯	GAY	g
𐐎𐐰	SHORT E	ɛ	𐐑𐐰	EF	f
𐐐𐐱	SHORT A	æ	𐐓𐐱	VEE	v
𐐎𐐲	SHORT AH	ɔ	𐐔𐐲	ETH	θ
𐐐𐐳	SHORT O	ʌ	𐐖𐐳	THEE	ð
𐐒𐐴	SHORT OO	ʊ	𐐗𐐴	ES	s
𐐀𐐃	AY	ɑɪ	𐐘𐐵	ZEE	z
𐐎𐐯	OI	ɔɪ	𐐙𐐶	ESH	ʃ
𐐆𐐮	OW	ɑʊ	𐐚𐐷	ZHEE	ʒ
𐐌𐐯	EW	ju	𐐛𐐸	ER	r
𐐋𐐮	WU	w	𐐜𐐹	EL	l
𐐯𐐯	YEE	j	𐐝𐐺	EM	m
𐐑𐐑	H	h	𐐞𐐻	EN	n
𐐂𐐒	PEE	p	𐐟𐐼	ENG	ŋ

13.11 Shavian

Shavian: U+10450–U+1047F

The playwright George Bernard Shaw (1856–1950) was an outspoken critic of the idiosyncrasies of English orthography. In his will, he directed that Britain's Public Trustee seek out and publish an alphabet of no fewer than 40 letters to provide for the phonetic spelling of English. The alphabet finally selected was designed by Kingsley Read and is variously known as Shavian, Shaw's alphabet, and the Proposed British Alphabet. Also in accordance with Shaw's will, an edition of his play, *Androcles and the Lion*, was published and distributed to libraries, containing the text both in the standard Latin alphabet and in Shavian.

As with other attempts at spelling reform in English, the alphabet has met with little success. Nonetheless, it has its advocates and users. The normative version of Shavian is taken to be the version in *Androcles and the Lion*.

Structure. The alphabet consists of 48 letters and 1 punctuation mark. The letters have no case. The digits and other punctuation marks are the same as for the Latin script. The one additional punctuation mark is a "name mark," used to indicate proper nouns. U+00B7 MIDDLE DOT should be used to represent the "name mark." The letter names are intended to be indicative of their sounds; thus the sound /p/ is represented by U+10450 𐑐 SHAVIAN LETTER PEEP.

The first 40 letters are divided into four groups of 10. The first 10 and second 10 are 180-degree rotations of one another; the letters of the third and fourth groups often show a similar relationship of shape.

The first 10 letters are tall letters, which ascend above the x-height and generally represent unvoiced consonants. The next 10 letters are "deep" letters, which descend below the baseline and generally represent voiced consonants. The next 20 are the vowels and liquids. Again, each of these letters usually has a close phonetic relationship to the letter in its matching set of 10.

The remaining 8 letters are technically ligatures, the first 6 involving vowels plus /r/. Because ligation is not optional, these 8 letters are included in the encoding.

Collation. The problem of collation is not addressed by the alphabet's designers.

13.12 Lisu

Lisu: U+A4D0–U+A4FF

Somewhere between 1908 and 1914 a Karen evangelist from Myanmar by the name of Ba Thaw modified the shapes of Latin characters and created the Lisu script. Afterwards, British missionary James Outram Fraser and some Lisu pastors revised and improved the script. The script is commonly known in the West as the Fraser script. It is also sometimes called the Old Lisu script, to distinguish it from newer, Latin-based orthographies for the Lisu language.

There are 630,000 Lisu people in China, mainly in the regions of Nujiang, Diqing, Lijiang, Dehong, Baoshan, Kunming and Chuxiong in the Yunnan Province. Another 350,000 Lisu live in Myanmar, Thailand and India. Other user communities are mostly Christians from the Dulong, the Nu and the Bai nationalities in China.

At present, about 200,000 Lisu in China use the Lisu script and about 160,000 in the other countries are literate in it. The Lisu script is widely used in China in education, publishing, the media and religion. Various schools and universities at the national, provincial and prefectural levels have been offering Lisu courses for many years. Globally, the script is also widely used in a variety of Lisu literature.

Structure. There are 40 letters in the Lisu alphabet. These consist of 30 consonants and 10 vowels. Each letter was originally derived from the capital letters of the Latin alphabet. Twenty-five of them look like sans-serif Latin capital letters (all but "Q") in upright positions; the other 15 are derived from sans-serif Latin capital letters rotated 180 degrees.

Although the letters of the Lisu script clearly derived originally from the Latin alphabet, the Lisu script is distinguished from the Latin script. The Latin script is bicameral, with case mappings between uppercase and lowercase letters. The Lisu script is unicameral; it has no casing, and the letters do not change form. Furthermore, typography for the Lisu script is rather sharply distinguished from typography for the Latin script. There is not the same range of font faces as for the Latin script, and Lisu typography is typically monospaced and heavily influenced by the conventions of Chinese typography.

Consonant letters have an inherent [ɑ] vowel unless followed by an explicit vowel letter. Three letters sometimes represent a vowel and sometimes a consonant: U+A4EA LISU LETTER WA, U+A4EC LISU LETTER YA, and U+A4ED LISU LETTER GHA.

Tone Letters. The Lisu script has six tone letters which are placed after the syllable to mark tones. These tone letters are listed in *Table 13-7*, with the tones identified in terms of their pitch contours.

Table 13-7. Lisu Tone Letters

Code	Glyph	Name	Tone
A4F8	.	mya ti	55
A4F9	,	na po	35
A4FA	..	mya cya	44
A4FB	.,	mya bo	33
A4FC	;	mya na	42
A4FD	:	mya jeu	31

Each of the six tone letters represents one simple tone. Although the tone letters clearly derive from Western punctuation marks (full stop, comma, semicolon, and colon), they do not function as punctuation at all. Rather, they are word-forming modifier letters. Furthermore, each tone letter is typeset on an em-square, including those whose visual appearance consists of two marks.

The first four tone letters can be used in combination with the last two to represent certain combination tones. Of the various possibilities, only ",;" is still in use; the rest are now rarely seen in China.

Other Modifier Letters. Nasalised vowels are denoted by a nasalization mark following the vowel. This word-forming character is not encoded separately in the Lisu script, but is represented by U+02BC MODIFIER LETTER APOSTROPHE, which has the requisite shape and properties (General_Category=Lm) and is used in similar contexts.

A glide based on the vowel *A*, pronounced as [ɑ] without an initial glottal stop (and normally bearing a 31 low falling pitch), is written after a verbal form to mark various aspects. This word-forming modifier letters is represented by U+02CD MODIFIER LETTER LOW MACRON. In a Lisu font, this modifier letter should be rendered on the baseline, to harmonize with the position of the tone letters.

Digits and Separators. There are no unique Lisu digits. The Lisu use European digits for counting. The thousands separator and the decimal point are represented with U+002C COMMA and U+002E FULL STOP, respectively. To separate chapter and verse numbers, U+003A COLON and U+003B SEMI-COLON are used. These can be readily distinguished from the similar-appearing tone letters by their numerical context.

Punctuation. U+A4FE "-." LISU PUNCTUATION COMMA and U+A4FF "=" LISU PUNCTUATION FULL STOP are punctuation marks used respectively to denote a lesser and a greater degree of finality. These characters are similar in appearance to sequences of Latin punctuation marks, but are not unified with them.

Over time various other punctuation marks from European or Chinese traditions have been adopted into Lisu orthography. *Table 13-8* lists all known adopted punctuation, along with the respective contexts of use.

Table 13-8. Punctuation Adopted in Lisu Orthography

Code	Glyph	Name	Context
002D	-	hyphen-minus	syllable separation in names
003F	?	question mark	questions
0021	!	exclamation mark	exclamations
0022	"	quotation mark	quotations
0028/0029	( )	parentheses	parenthetical notes
300A/300B	《 》	double angle brackets	book titles
2026	...	ellipsis	omission of words (always doubled in Chinese usage)

U+2010 HYPHEN may be preferred to U+002D HYPHEN-MINUS for the dash used to separate syllables in names, as its semantics are less ambiguous than U+002D.

The use of the U+003F "?" QUESTION MARK replaced the older Lisu tradition of using a tone letter combination to represent the question prosody, followed by a Lisu full stop: "..:="

Linebreaking. A line break is not allowed within an orthographic syllable in Lisu. A line break is also prohibited before a punctuation mark, even if it is preceded by a space. There is no line-breaking hyphenation of words, except in proper nouns, where a break is allowed after the hyphen used as a syllable separator

Word Separation. The Lisu script separates syllables using a space or, for proper names, a hyphen. In the case of polysyllabic words, it can be ambiguous as to which syllables join together to form a word. Thus for most text processing at the character level, a syllable (starting after a space or punctuation and ending before another space or punctuation) is treated as a word except for proper names—where the occurrence of a hyphen holds the word together.

13.13 Miao

Miao: U+16F00–U+16F9F

The Miao script, also called Lao Miaowen ("Old Miao Script") in Chinese, was created in 1904 by Samuel Pollard and others, to write the Northeast Yunnan Miao language of southern China. The script has also been referred to as the Pollard script, but that usage is no longer preferred. The Miao script was created by an adaptation of Latin letter variants, English shorthand characters, Miao pictographs, and Cree syllable forms. (See *Section 13.9,*

Canadian Aboriginal Syllabics.) Today, the script is used to write various Miao dialects, as well as languages of the Yi and Lisu nationalities in southern China.

The script was reformed in the 1950s by Yang Rongxin and others, and was later adopted as the "Normalized" writing system of Kunming City and Chuxiong Prefecture. The main difference between the pre-reformed and the reformed orthographies is in how they mark tones. Both orthographies can be correctly represented using the Miao characters encoded in the Unicode Standard.

Encoding Principles. The script is written left to right. The basic syllabic structure contains an initial consonant or consonant cluster and a final. The final consists of either a vowel or vowel cluster, an optional final nasal, plus a tone mark. The initial consonant may be preceded by U+16F50 MIAO LETTER NASALIZATION, and can be followed by combining marks for voicing (U+16F52 MIAO SIGN REFORMED VOICING) or aspiration (U+16F51 MIAO SIGN ASPIRATION and U+16F53 MIAO SIGN REFORMED ASPIRATION).

Tone Marks. In the Chuxiong reformed orthography, vowels and final nasals appear on the baseline. If no explicit tone mark is present, this indicates the default tone 3. An additional tone mark, encoded in the range U+16F93..U+16F99, may follow the vowel to indicate other tones. A set of archaic tone marks used in the reformed orthography is encoded in the range U+16F9A..U+16F9F.

In the pre-reformed orthography, such as that used for the language Ahmao (Northern Hmong), the tone marks are represented in a different manner, using one of five shifter characters. These are represented in sequence following the vowel or vowel sequence and indicate where the vowel letter is to be rendered in relation to the consonant. If more than one vowel letter appears before the shifter, all of the vowel glyphs are moved together to the appropriate position.

Rendering of "wart". Several Miao consonants appear in the code charts with a "wart" attached to the glyph, usually on the left-hand side. In the Chuxiong orthography, a dot appears instead of the wart on these consonants. Because the user communities consider the appearance of the wart or dot to be a different way to write the same characters and not a difference of the character's identity, the differences in appearance are a matter of font style.

Ordering. The order of Miao characters in the code charts derives from a reference ordering widely employed in China, based in part on the order of Bopomofo phonetic characters. The expected collation order for Miao strings varies by language and user communities, and requires tailoring. See Unicode Technical Standard #10, "Unicode Collation Algorithm."

Digits. Miao uses European digits.

Punctuation. The Miao script employs a variety of punctuation marks, both from the East Asian typographical tradition and from the Western typographical tradition. There are no script-specific punctuation marks.

