13.5 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 the country of Mongolia (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 the traditional writing system for the Mongolian language and the three derivative writing systems Todo, Manchu, and Sibe. The traditional writing system is also known as "Hudum Mongolian," and is explicitly referred to as "Hudum" in the following text. Each of the three derivative writing systems shares some common letters with Hudum, and these letters are encoded only once. Each derivative writing system also has a number of modified letterforms or new letters, which are encoded separately. The letters typically required by each writing system's modern usage are encoded as shown in *Table 13-4*.

Table 13-4. Letter Usage in Mongolian Writing Systems

Hudum	Todo	Manchu	Sibe	Note
1820..1842	1820 1828 182F..1831 1834 1837..1838 1840	1820 1823 1828..182A 182E..1830 1834..1836 1838 183A	1820 1823 1828 182A 182E..1830 1834 1836..1838 183A	Unified Mongolian letters
	1843..185A 185C	185D 185F..1861 1864..1869 186C..1871 1873..1877	185D..1872	Letters specific to the derivative writing systems

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 created around the beginning of the thirteenth century, during the reign of Genghis Khan. It derives from the Old Uyghur script, which was in use from about the eighth to the fifteenth century. Old Uyghur itself was an adaptation of Sogdian Aramaic, a Semitic script written horizontally from right to left. Probably under the influence of the Chinese script, the Old Uyghur script became rotated ninety degrees coun-

terclockwise so that the lines of text read vertically in columns running from left to right. The Mongolian script inherited this directionality from the Old Uyghur script.

The Mongolian script has remained in continuous use for writing Mongolian within the Inner Mongolia Autonomous Region of China and elsewhere in China. However, in Mongolia (Outer 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 dis-

persed 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 ninety 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. If such text is viewed sideways, the usual Mongolian column order appears reversed, but this orientation can be workable for short stretches of text. There are no bidirectional effects in such a layout because all text is horizontal left to right.

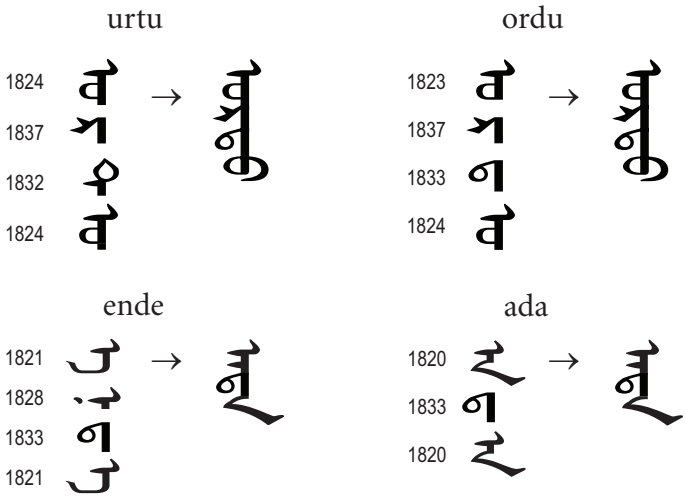
Encoding Principles. The encoding model for Mongolian is somewhat different from that for any other script within Unicode, and in many respects it is the most complicated. For this reason, only the essential features of Mongolian shaping behavior are presented here.

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

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

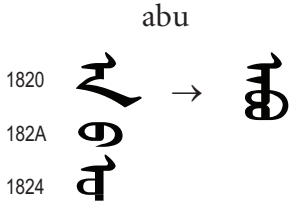
Cursive Joining. The Mongolian script is cursive, and the letters constituting a word are normally joined together. In most cases the letters join together naturally along a vertical stem, but in the case of certain "bowed" consonants (for example, U+182A MONGOLIAN LETTER BA and the feminine form of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

Figure 13-3. Mongolian Glyph Convergence



ing vertical stem, they may form ligatures with a following vowel. This is illustrated in *Figure 13-4*, where the letter *ba* combines with the letter *u* to form a ligature in the Mongolian word *abu* "father."

Figure 13-4. Mongolian Consonant Ligation



The Joining_Type values for Mongolian characters are defined in ArabicShaping.txt in the Unicode Character Database. For a discussion of the meaning of Joining_Type values in the context of a vertically rendered script, see "Cursive Joining" in *Section 14.4, Phags-pa*. Most Mongolian characters are Dual_Joining, as they may join on both top and bottom.

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

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

Figure 13-5. Mongolian Positional Forms

expected positional form within a word. Basically, they evoke the same contextual selection effects in neighboring letters as do non-joining or joining regular letters, but are themselves invisible (see *Chapter 23, Special Areas and Format Characters*). For example, the various positional forms of U+1820 MONGOLIAN LETTER A may be selected by means of the following character sequences:

<1820> selects the isolate form.

<1820 200D> selects the initial form.

<200D 1820> selects the final form.

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

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

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

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

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

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

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

These format characters normally have no visual appearance. When required, a free variation selector immediately follows the base character it modifies. This combination of base character and variation selector is known as a standardized variant. The table of standardized variants, StandardizedVariants.txt, in the Unicode Character Database exhaustively lists all currently defined standardized variants. All combinations not listed in the table are

unspecified and are reserved for future standardization; no conformant process may interpret them as standardized variants. Therefore, any free variation selector not immediately preceded by one of their defined base characters will be ignored.

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

Figure 13-6. Mongolian Free Variation Selector

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

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

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

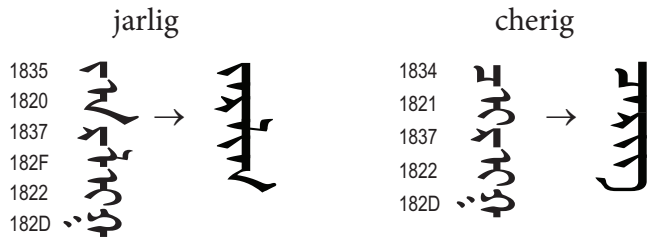
Vowel Harmony. Mongolian has a system of vowel harmony, whereby the vowels in a word are either all "masculine" and "neuter" vowels (that is, back vowels plus /i/) or all "feminine" and "neuter" vowels (that is, front vowels plus /i/). Words that are written with masculine/neuter vowels are considered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

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

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

When one of these consonants precedes another consonant or is the final letter in a word, it may take either a masculine or feminine glyph form, depending on its context. The rendering system should automatically select the correct gender form for these letters based on the gender of the word (in Mongolian and Todo) or the gender of the preceding vowel (in Manchu and Sibe). This is illustrated by *Figure 13-7*, where U+182D MONGOLIAN LETTER GA takes a masculine glyph form when it occurs finally in the masculine word *jarlig* "order," but takes a feminine glyph form when it occurs finally in the feminine word *cherig* "soldier." In this example, the gender form of the final letter *ga* depends on whether the first vowel in the word is a back (masculine) vowel or a front (feminine or neuter) vowel. Where the gender is ambiguous or a form not derivable from the context is required, the user needs to specify which form is required by means of the appropriate free variation selector.

Figure 13-7. Mongolian Gender Forms



Narrow No-Break Space. In Mongolian, Todo, Manchu, and Sibe, certain grammatical suffixes are separated from the word stem or from other suffixes by a gap. Many separated

suffixes exhibit shapes that are distinct from ordinary words, and thus require special shaping.

There are many separated suffixes in Mongolian, usually occurring in masculine and feminine pairs (for example, the dative suffixes *-dur* and *-dür*), most of which require special shaping; a stem may have multiple separated suffixes. In contrast, there are only six separated suffixes for Manchu and Sibe, only one of which (*-i*) requires special shaping; stems do not have more than one separated suffix at a time.

Because separated suffixes are usually considered an integral part of the word as a whole, a line break opportunity does not normally occur before a separated suffix. The whitespace preceding the suffix is often narrower than an ordinary space, although the width may expand during justification. U+202F NARROW NO-BREAK SPACE (NNBSP) is used to represent this small whitespace; it not only prevents word breaking and line breaking, but also triggers special shaping for the following separated suffix. The resulting shape depends on the particular separated suffix. Note that NNBSP may be preceded by another separated suffix, and NNBSP may also appear between non-Mongolian characters and a separated suffix.

Normally, NNBSP does not provide a line breaking opportunity. However, in situations where a line is broken before a separated suffix, such as in narrow columns, it is important not to disable the special shaping triggered by NNBSP. This behavior may be achieved by placing the break so that the NNBSP character is at the start of the new line. At the beginning of the line, the NNBSP should affect only the shaping of the following Mongolian characters, and should display with no advance width.

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

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

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.

Figure 13-8. Mongolian Vowel Separator

Qana with Connected Final Qana with Separated Final

Baluda. The two Mongolian *baluda* characters, U+1885 MONGOLIAN LETTER ALI GALI BALUDA and U+1886 MONGOLIAN LETTER ALI GALI THREE BALUDA, are historically related to U+0F85 TIBETAN MARK PALUTA. When used in Mongolian text rendered vertically, a *baluda* or *three baluda* character appears to the right side of the first character in a word. To simplify rendering implementations for Mongolian Ali Gali texts, the *baluda* characters have been categorized as nonspacing combining marks, rather than as letters.

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 Mongolian and 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 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.

Mongolian Supplement: U+11660–U+1167F

The Mongolian Supplement block contains a supplemental collection of *birga* head mark signs of various shapes and orientations. These complement the basic U+1800 MONGOLIAN BIRGA.