Chapter 11

Additional Scripts

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

- Ethiopic
- Cherokee
- Canadian Aboriginal Syllabics
- Mongolian

Although the Mongolian and Ethiopic scripts can be traced to roots in the original Semitic writing systems, they would not be classified as Middle Eastern scripts today. The Cherokee and Canadian Aboriginal Syllabic scripts have roots in Latin letterforms and shorthand, but their application and evolution is a native North American development.

11.1 Ethiopic

Ethiopic: U+1200–U+137F

The Ethiopic syllabary originally evolved for writing the Semitic language Ge'ez, and 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 are 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. Research is ongoing to identify these extended Ethiopic forms, though many are rare and have scant typographic tradition.

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 voweled 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 might represent the "consonant + W + AA" syllable. For example:

U+1247 = Q + WAA: unused version of QWAA

U+124B = QW + AA: ETHIOPIC SYLLABLE QWAA

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 11-1*.

Table 11-1. Labialized Forms in -WAA

-WAA Form	Encoded as	Not Used
QWAA	U+124B	1247
QHWAA	U+125B	1257
XWAA	U+128B	1287
KWAA	U+12B3	12AF
KXWAA	U+12C3	12BF
GWAA	U+1313	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 11-2*.

Table 11-2. Labialized Forms in -WE

"-WE" Form	Encoded as	Not Used
QWE	U+124D	1249
QHWE	U+125D	1259
XWE	U+128D	1289
KWE	U+12B5	12B1
KXWE	U+12C5	12C1
GWE	U+1315	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 ETHIOPIAN WORDSPACE (:). In modern usage, a plain white wordspace (U+0020 SPACE) is becoming common.

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+0308 COMBINING DIAERESIS and U+030E COMBINING DOUBLE VERTICAL LINE ABOVE are sometimes used to indicate emphasis or gemination (consonant doubling).

Numbers. 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

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

11.2 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 "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,… .

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.

11.3 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 below 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:

∨	∧	>	<	ᐧ
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 of each character 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 value.

11.4 Mongolian

Mongolian: U+1800–U+18AF

The Mongolians are key representatives of a cultural-linguistic group known as Altaic, after the Altai mountains of central Asia. In the past, these peoples have dominated the vast expanses of Asia and beyond, from the Baltic to the Sea of Japan. Echoes of Altaic languages remain from Finland, Hungary, and Turkey, across central Asia, to Korea and Japan. Today the Mongolians themselves are represented politically in Mongolia proper (formerly 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 traditional Mongolian script encoded here has been in continuous use since the times of Genghis Khan. In the Mongolian People's Republic, it was replaced in general use by a Cyrillic orthography in the early 1940s, but the traditional script was restored by law in 1992. As a practical matter, both traditional and Cyrillic texts are still seen in Mongolia and in Western scholarship. There is no one-to-one transcription between the two scripts; approximate correspondence mappings are indicated in the Mongolian character names list, but are not necessarily unique in either direction. All of the Cyrillic characters needed to write Mongolian are included in the Cyrillic section of the Unicode Standard.

The traditional Mongolian script is used to write a Mongolian literary language of classical origin, which is distinct from spoken Mongolian dialects. In addition, the Manchu, Sibe, and Oirats formed their own writing systems based on the Mongolian script. To convey Buddhist classics accurately, the Ali Gali letters were added to traditional Mongolian, Todo, and Manchu for the transcription of Tibetan and Sanskrit.

Directionality. The Mongolian script was derived around the twelfth century from the Uighur script, which originated from Aramaic, a right-to-left Semitic script. At some point the writing was transformed as though the whole page had been rotated 90 degrees counterclockwise, with the result that Mongolian is traditionally written vertically top to bottom in columns advancing from left to right. This directional pattern is unique among scripts for living languages.

In modern contexts, the Mongolian script frequently occurs together with scripts of different directionality. Ideally, each script should appear separately in its own orientation, but in cases where this separation cannot be achieved, one of the scripts can be adopted into the directionality of the other.

If the Mongolian script is adopted into horizontal text, its lines are rotated *another* 90 degrees counterclockwise so that the letters join left to right, and the columns are transcribed to the equivalent lines (first column becomes first line, and so on). If such text is viewed sideways, the usual Mongolian column order appears reversed, but this orientation can be workable for short stretches of text. Note that there are no bidirectional effects in such a layout, because all text is horizontal left to right.

Standards. The encoding of Mongolian presented here has been cooperatively developed over years of careful research by a group of experts from Mongolia, China, and the West. These authorities are to publish a document called "Users' Convention for System Implementation of the International Standard on Mongolian Encoding," explicitly defining Mongolian character shaping behavior in full. The complex sequence and shaping rules detailed in that document are only summarized in the following description.

Encoding Principle. The relationship between language and script in Mongolian is in some ways reminiscent of that in English: the same letter may represent different sounds; different letters may represent the same sound; and sometimes spellings are purely historical with little relation to modern pronunciation. Additionally, the Mongolian script is in some ways reminiscent of Arabic: letters usually join cursively together, assuming contextual forms according to a variety of rules, sometimes having up to 10 different presentation forms for the same letter.

Despite these variations, the basic encoding principle for Mongolian is simple—the same principle as that for English and Arabic: there is a basic underlying alphabet whose letters are well agreed upon by the user community, apart from any question of their sounds or forms. The elements that are encoded are just the letters of the conventional Mongolian, Todo, Sibe, and Manchu alphabets, along with their associated numerals and punctuation.

Punctuation. Punctuation symbols specific to Mongolian include U+1800 MONGOLIAN BIRGA and U+1805 MONGOLIAN FOUR DOTS, which are used to mark, respectively, the beginning and end of a unit of text, such as a section or paragraph. In Mongolian Todo text, U+1806 MONGOLIAN TODO SOFT HYPHEN is used at the beginning of the second line to indicate resumption of the broken word. It functions like a U+00AD SOFT HYPHEN, except that this version appears at the beginning of a line rather than at the end. Some punctuation symbols used in Mongolian are coded in the General Punctuation block, such as U+2048 "?!" QUESTION EXCLAMATION MARK and U+2049 "!?" EXCLAMATION QUESTION MARK, used for side-by-side display in vertical text. Mongolian employs the usual U+0020 SPACE to separate words, plus distinctive narrow gap spaces *within* words, as discussed below.

Letterforms. Mongolian is a cursive script, so the visual form of each letter generally depends on its position within a word. There are forms for the beginning (initial form), the middle (medial form), and the end of a word (final form). Vowels have an additional isolated form, used when the vowel appears between whitespaces. The positional forms are not all unique: in some cases, a character can have the same form in different positions (for example, the initial and medial forms of U+182A MONGOLIAN LETTER BA are the same); in other cases, two different characters can look the same (for example, the isolated form of U+1824 MONGOLIAN LETTER U is the same as that of U+1823 MONGOLIAN LETTER O).

Besides positional forms, many letters also have free variants. These variants are not determined by the position of a character within a word, but are prescribed by rules of spelling and grammar. Mongolian also employs some ligatures, particularly those formed by a vowel following a "bowed" consonant, which is a consonant like U+182A MONGOLIAN LETTER BA that lacks a trailing vertical stem.

The representative glyph in the code charts is generally the isolated form for the vowels and the initial form for the consonants, with the most common variant being chosen in case of alternatives. Other forms are occasionally chosen to avoid having different letters with the same glyph in the code charts: for example, U+1824 MONGOLIAN LETTER U is represented by its initial form, because its isolated form is the same as that of U+1823 MONGOLIAN LETTER O.

Shaping Format Characters. For cases in which the contextual sequence of basic letters is not sufficient for a rendering engine to uniquely determine the appropriate glyph for a particular letter, additional format characters are provided so that the typist may specify the desired rendering. There are seven characters that may function as glyph shape format selectors when inserted in a basic encoded Mongolian text sequence:

U+180B MONGOLIAN FREE VARIATION SELECTOR ONE

U+180C MONGOLIAN FREE VARIATION SELECTOR TWO

U+180D MONGOLIAN FREE VARIATION SELECTOR THREE

U+180E MONGOLIAN VOWEL SEPARATOR

U+200C ZERO WIDTH NON-JOINER

U+200D ZERO WIDTH JOINER

U+202F NARROW NO-BREAK SPACE

Except for U+202F NARROW NO-BREAK SPACE and U+180E MONGOLIAN VOWEL SEPARATOR, these characters normally have no visual appearance. Their sole purpose is to guide the rendering process in selecting the appropriate glyphs to represent base Mongolian letters in a particular context. A fully detailed specification of the shape selection rules is under development. The following are summaries of the effects of the shaping format characters.

U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER play the same role in Mongolian as in other scripts (see *Chapter 13, Special Areas and Format Characters*). Basically, they evoke the same contextual selection effects in neighboring letters as do non-joining or joining regular letters, but are themselves invisible.

U+202F NARROW NO-BREAK SPACE: Many Mongolian words are formed by the addition of one or more grammatical suffixes to a basic stem word. The stem together with its suffixes is considered to be a single word, but each suffix is separated from the stem or from the preceding suffix by a small gap that breaks the visual connectivity of the word. This word-internal whitespace is coded by U+202F NARROW NO-BREAK SPACE. Because it represents a grammatical juncture, this character also functions as a variant selector that affects the contextual forms of the letters preceding and following it.

Because it uses regular U+0020 SPACE between words, Mongolian may also use regular U+00A0 NO-BREAK SPACE for its typical functions, such as preventing a line break between a pair of words. U+00A0 NO-BREAK SPACE does not serve the function of U+202F NARROW NO-BREAK SPACE.

U+180E MONGOLIAN VOWEL SEPARATOR is a word-internal thin whitespace that may occur only before the word-final vowels U+1820 MONGOLIAN LETTER A and U+1821 MONGOLIAN LETTER E. It determines the specific form of the character preceding it, selects a special variant shape of these vowels, and produces a small gap within the word. If it erroneously occurs before any character other than "A" or "E", it behaves like a U+202F NARROW NO-BREAK SPACE.

The MONGOLIAN FREE VARIATION SELECTOR characters are used to distinguish different variants of the same letter appearing under the same conditions—that is, where more than one rendered shape is possible and the selection must be indicated by human intervention rather than derived by algorithm. A free variant selector is encoded *after* the base character it modifies. Free variant selectors are not productive and are therefore ignored when not immediately preceded by one of their listed base characters. Tables specifying the precise shape-selection behavior of these characters are in the midst of the standardization process; the details will be published as a Unicode Technical Report and posted on the Unicode Web site when they become available. Conformant usage of these characters requires adherence to the shape-selection tables.

This PDF file is an excerpt from *The Unicode Standard, Version 3.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (see http://www.unicode.org/unicode/uni2errata/UnicodeErrata.html). More recent versions of the Unicode standard exist (see http://www.unicode.org/unicode/standard/versions/).

Dai Kan-Wa Jiten used as the source of reference Kanji codes was written by Tetsuji Morohashi and published by Taishukan Shoten.

ISBN 0-201-61633-5

This book is set in Minion, designed by Rob Slimbach at Adobe Systems, Inc. It was typeset using FrameMaker 5.5 running under Windows NT. ASMUS, Inc. created custom software for chart layout. The Han radical-stroke index was typeset by Apple Computer, Inc. The following companies and organizations supplied fonts:

Apple Computer, Inc.
Atelier Fluxus Virus
Beijing Zhong Yi (Zheng Code) Electronics Company
DecoType, Inc.
IBM Corporation
Monotype Typography, Inc.
Microsoft Corporation
Peking University Founder Group Corporation
Production First Software

Additional fonts were supplied by individuals as listed in the *Acknowledgments*.

The publisher offers discounts on this book when ordered in quantity for special sales. For more information please contact:

Corporate, Government, and Special Sales
Addison Wesley Longman, Inc.
One Jacob Way
Reading, Massachusetts 01867

Visit A-W on the Web: http://www.awl.com/cseng/

First printing, January 2000.