

This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, 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 (<http://www.unicode.org/errata/>). For information on more recent versions of the standard, see <http://www.unicode.org/standard/versions/enumeratedversions.html>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and Addison-Wesley was aware of a trademark claim, the designations have been printed in initial capital letters. However, not all words in initial capital letters are trademark designations.

The Unicode® Consortium is a registered trademark, and Unicode™ is a trademark of Unicode, Inc. The Unicode logo is a trademark of Unicode, Inc., and may be registered in some jurisdictions.

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

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

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

Cover and CD-ROM label design: Steve Mehallo, <http://www.mehallo.com>

The publisher offers discounts on this book when ordered in quantity for bulk purchases and special sales. For more information, customers in the U.S. please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsontechgroup.com. For sales outside of the U.S., please contact International Sales, +1 317 581 3793, international@pearsontechgroup.com

Visit Addison-Wesley on the Web: <http://www.awprofessional.com>

Library of Congress Cataloging-in-Publication Data

The Unicode Standard, Version 4.0 : the Unicode Consortium /Joan Aliprand... [et al.].

p. cm.

Includes bibliographical references and index.

ISBN 0-321-18578-1 (alk. paper)

1. Unicode (Computer character set). I. Aliprand, Joan.

QA268.U545 2004

005.7'2—dc21

2003052158

Copyright © 1991–2003 by Unicode, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publisher or Unicode, Inc. Printed in the United States of America. Published simultaneously in Canada.

For information on obtaining permission for use of material from this work, please submit a written request to the Unicode Consortium, Post Office Box 39146, Mountain View, CA 94039-1476, USA, Fax +1 650 693 3010 or to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300 Boston, MA 02116, USA, Fax: +1 617 848 7047.

ISBN 0-321-18578-1

Text printed on recycled paper

1 2 3 4 5 6 7 8 9 10—CRW—0706050403

First printing, August 2003

Chapter 12

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
- Mongolian
- Osmanya
- Cherokee
- Canadian Aboriginal Syllabics
- Deseret
- Shavian

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

The remaining scripts in this chapter have been developed relatively recently. In different ways they show roots in Latin and other letterforms including shorthand, but they are also original creative contributions intended to serve the communities that use them.

12.1 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 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 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 12-1*.

Table 12-1. Labialized Forms in -WAA

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

Table 12-2. Labialized Forms in -WE

“-WE” Form	Encoded as	Not Used
QWE	U+124D ባዊ	1249
QHWE	U+125D ባዊዊ	1259
XWE	U+128D ኣዊ	1289
KWE	U+12B5 ከዊ	12B1
KXWE	U+12C5 ከዊዊ	12C1
GWE	U+1315 ገዊ	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. 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

$$\begin{aligned} 2345 &= (20 + 3) * 100^1 + (40 + 5) * 100^0 \\ &= 20 \quad 3 \quad 100 \quad 40 \quad 5 \\ &= \text{TWENTY THREE HUNDRED FORTY FIVE} \\ &= 1373 \quad 136B \quad 137B \quad 1375 \quad 136D \quad \text{፳፻፹፻፵፮} \end{aligned}$$

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:

$$\begin{aligned} 2,345 &= \text{TWO thousand THREE HUNDRED FORTY FIVE} \\ &= 136A \quad 123A \quad 136B \quad 137B \quad 1375 \quad 136D \quad \text{፳፻፹፻፵፮ ሺ} \end{aligned}$$

12.2 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 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. The complex sequence and shaping rules required for Mongolian are being reviewed and verified by experts, and are merely summarized in the following description. Once the rules are ready for publication, they will be published as a Unicode Technical Report.

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 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+2010 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 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. Seven characters may function as glyph shape format selectors when inserted in a basic encoded Mongolian text sequence:

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

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

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

U+180E MONGOLIAN VOWEL SEPARATOR (MVS)

U+200C ZERO WIDTH NON-JOINER (ZWNJ)

U+200D ZERO WIDTH JOINER (ZWJ)

U+202F NARROW NO-BREAK SPACE (NNBSP)

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 they do in other scripts (see *Chapter 15, 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.

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 (NNBSP). Because it represents a grammatical juncture, this character also affects the contextual forms of the letters preceding and following it. Use of a ZERO WIDTH JOINER next to a NNBSP would be redundant.

Because it uses regular U+0020 SPACE between words, Mongolian may also use a regular U+00A0 NO-BREAK SPACE (NBSP) for its typical functions, such as preventing a line break between a pair of words. NBSP differs in function from NNBSP, just described. While the full details of Mongolian shaping behavior are deferred to a forthcoming description, both NNBSP and NBSP interrupt Mongolian shaping. Therefore, the use of ZERO WIDTH NON-JOINER either preceding or following them would be redundant.

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.

Variation Selectors. The three 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 immediately follows 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.

Like Arabic letters, Mongolian letters have various presentation forms depending on their positions in words. There are additional linguistic constraints that result in variations that must be employed in specific contexts, creating the need for several Mongolian-specific variant selectors, which are encoded at U+180B, U+180C, and U+180D.

Use of these free variation selectors is not the only way that the associated shapes can be selected. Use of an extended context and rules for shape selection can obviate the need for using these variation selectors in many ordinary situations. The precise details are under development.

The table of standardized variants, *StandardizedVariants.txt*, in the Unicode Character Database provides a description of the variant presentation glyphs corresponding to the

use of specified variation selectors with all allowed base Mongolian characters. Only some presentation forms of the base Mongolian characters produce variant presentation glyphs, when immediately followed by the Mongolian free variation selectors. These combinations are exhaustively listed and described in the table. All combinations not listed in the table are unspecified and are reserved for future standardization; no conformant process may interpret them as standardized variants.

Older documentation, external to the Unicode Standard, listed the action of the free variation selectors by using ZWJ and ZWNJ to explicitly indicate the shaping environment affected by the variation selector. The relative order of the ZWJ or ZWNJ and the free variation selector in these documents was different from the one required by *Section 15.6, 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.

12.3 Osmanya

Osmanya: U+10480–U+104AF

The Osmanya script, which in Somali is called *ሒሽገ ገገጌ* *far Soomaali* “Somali writing” or *ሒሽገ ገገጌ Cismaanya*, was devised in 1920–1922 by *ሒሽገ ገገጌ ሐረርሽኝ ሐረርሽኝ* (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 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. 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 ሐ OSMANYA LETTER AA, U+1049C ሐ OSMANYA LETTER EE, and U+1049D ሐ OSMANYA LETTER OO; long *uu* and *ii* are written with the consonants U+10493 ሐ OSMANYA LETTER WAW and U+10495 ሐ OSMANYA LETTER YA.

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 Diringen (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.

12.4 Cherokee

Cherokee: U+13A0–U+13FF

The Cherokee script is used to write the Cherokee language. Cherokee is a member of the Iroquoian 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 **ᎠᎯᏯ** *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.

12.5 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:

∨	∧	>	<	◀
PE	PI	PO	PA	P

Some variations in vowels also occur. For example, in Inuktitut usage, the syllable U+1450 ◻ 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 ◑ 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 ◒ 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.

12.6 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 found. 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 here:

- ɔ AY is written for *eye* or *I*
- ʏ YEE is written for *ye*
- Ɔ BEE is written for *be* or *bee*
- Ɔ GAY is written for *gay*
- ʏ THEE is written for *the* or *thee*

Phonetics. An approximate IPA transcription of the sounds represented by the Deseret Alphabet is shown in *Figure 12-1*.

Figure 12-1. IPA Transcription of Deseret

∂∂	LONG I	i:	Ɔa	BEE	b
ƎƎ	LONG E	eɪ	ʎʎ	TEE	t
ƆƆ	LONG A	a:	Ɔa	DEE	d
ƆƆ	LONG AH	ɔ:	Ɔc	CHEE	tʃ
ƆƆ	LONG O	o:	ʏʏ	JEE	dʒ
ƆƆ	LONG OO	u:	ƆƆ	KAY	k
††	SHORT I	ɪ	ƆƆ	GAY	g
ɹɹ	SHORT E	ɛ	ƆƆ	EF	f
ɹɹ	SHORT A	æ	ƆƆ	VEE	v
ɹɹ	SHORT AH	ɔ	LL	ETH	θ
ɹɹ	SHORT O	ʌ	ʏƆ	THEE	ð
ʏʏ	SHORT OO	ʊ	ʏʏ	ES	s
ɔɔ	AY	aɪ	ƆƆ	ZEE	z
ɹɹ	OI	ɔɪ	ƆƆ	ESH	ʃ
ƆƆ	OW	aʊ	ʏʏ	ZHEE	ʒ
ƆƆ	EW	ju	ʏʏ	ER	r
ʏʏ	WU	w	ʏʏ	EL	l
ʏʏ	YEE	j	ƆƆ	EM	m
ʏʏ	H	h	ʏʏ	EN	n
ɹɹ	PEE	p	ʏʏ	ENG	ŋ

12.7 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 in both 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 } 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.