The Unicode® Standard
Version 10.0 – 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 the Unicode Consortium. — Version 10.0.
Includes bibliographical references and index.
ISBN 978-1-936213-16-0 (http://www.unicode.org/versions/Unicode10.0.0/)
1. Unicode (Computer character set) I. Unicode Consortium.
QA268.U545 2017

ISBN 978-1-936213-16-0
Published in Mountain View, CA
June 2017

Chapter 20
Americas

The following scripts from the Americas are discussed in this chapter:

Cherokee	*Osage*
Canadian Aboriginal Syllabics	*Deseret*

The Cherokee script is a syllabary developed between 1815 and 1821, to write the Cherokee language. The Cherokee script is still used 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.

The Osage script is an alphabet used to write the Osage language spoken by a Native American tribe in the United States. The script was written with a variety of ad-hoc orthographies and transcriptions for two centuries until the Osage Nation recently developed its standard orthography in 2014.

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.

20.1 Cherokee

Cherokee: U+13A0–U+13FF
Cherokee Supplement: U+AB70–U+ABBF

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 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. Samuel Worcester worked in conjunction with Sequoyah, Chief Charles Hicks, and Charles Thompson (first cousin of Sequoyah) in the design of the Cherokee type which was finalized in 1827. Using fonts available to him, Worcester assigned a number of Latin letters to the Cherokee syllables. At this time the Cherokee letter "MV" was dropped, and the Cherokee syllabary reached the size of 85 letters. Worcester's press printed 13,980,000 pages of Native American-language text, most of it in Cherokee.

Structure. Cherokee is a left-to-right script. It has no Cherokee-specific combining characters.

Casing. Most existing Cherokee text is caseless. Traditionally, the forms of the syllable letters were designed as caps height—and in fact, a number of the Cherokee syllables are visually indistinguishable from Latin uppercase letters. As a result, most Cherokee text has the visual appearance of all caps. The characters used for representing such unicameral Cherokee text are the basic syllables in the Cherokee block: U+13A0 CHEROKEE LETTER A, and so forth.

In some old printed material, such as the Cherokee New Testament, case conventions adapted from the Latin script were used. Sentence-initial letters and initial letters for personal and place names, for example, were typeset using a larger size font. Furthermore, systematic distinction in casing has become more prevalent in modern typeset materials, as well.

Starting with Version 8.0, the Unicode Standard includes a set of lowercase Cherokee syllables to accommodate the need to represent casing distinctions in Cherokee text. The Cherokee script is now encoded as a fully bicameral script, with case mapping.

The lowercase syllable letters are mostly encoded in the Cherokee Supplement block. A few are encoded at the end of the Cherokee block, after the basic Cherokee syllable letters, which are now treated as the uppercase of the case pairs.

The usual way for a script originally encoded in the Unicode Standard as a unicameral script to later gain casing is by adding a new set of uppercase letters for it. The Cherokee script is an important exception because the previously encoded Cherokee unicameral set is treated as the uppercase as of Version 8.0, and the new set of letters are the lowercase. The reason for this exception has to do with Cherokee typography and the status of existing fonts. Because all existing fonts already treated Cherokee syllable letters as cap height, attempting to extend them by changing the existing letters to less than cap height and adding new uppercase letters to the fonts would have destabilized the layout of all existing Cherokee text. On the other hand, innovating in the fonts by adding new lowercase forms with a smaller size and less than cap height allows a graceful introduction of casing without invalidating the layout of existing text.

This exceptional introduction of a lowercase set to change a unicameral encoding into a bicameral encoding has important implications that implementers of the Cherokee script need to keep in mind. First, in order to preserve case folding stability, Cherokee case folds to the previously encoded uppercase letters, rather than to the newly encoded lowercase letters. This exceptional case folding behavior impacts identifiers, and so can trip up implementations if they are not prepared for it. Second, representation of cased Cherokee text requires using the new lowercase letters for most of the body text, instead of just changing a few initial letters to uppercase. That means that representation of traditional text such as the Cherokee New Testament requires substantial re-encoding of the text. Third, the fact that *uppercase* Cherokee still represents the default and is most widely supported in fonts means that input systems which are extended to support the new lowercase letters face unusual design choices.

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.

Input. 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

Numbers. Although Sequoyah invented a Cherokee number system, it was not adopted and is not encoded in the Unicode Standard. 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.

Punctuation. Cherokee uses standard Latin punctuation.

Standards. There are no other encoding standards for Cherokee. Cherokee spelling is not standardized: each person spells as the word sounds to him or her.

20.2 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).

20.3 Osage

Osage: U+104B0–U+104FF

The Osage script is used to write the Osage language. This language is spoken by a Native American tribe of the Great Plains that originated in the Ohio River valley area of the present-day United States. By the 17th century, the Osage people had migrated to their current locations in Missouri, Kansas, Arkansas, Oklahoma, and Texas. The term "Osage" roughly translates to "mid-waters."

For two centuries, the Osage language was written with a variety of ad-hoc Latin orthographies and transcription systems. In 2004, the Osage Nation initiated a program to develop a standard orthography to write the language. By 2006, a practical orthography had been designed based on modifications or fusions of the shapes of Latin letters. Use of the Osage orthography led to further improvements, culminating in the adoption of the current set of letters in 2014.

Structure. Osage is a left-to-right alphabetic script. It has no Osage-specific combining characters, but makes use of common diacritical marks.

Casing. Casing is used in the standard Osage orthography.

Vowels. Diacritical marks are used in Osage to distinguish length, nasalization, and accents. The particular diacritical marks used to make these distinctions are shown in *Table 20-1*.

Table 20-1. Combining Marks used in Osage

Nasal vowels	U+0358 ◌͘ COMBINING DOT ABOVE RIGHT
Long vowels	U+0304 ◌̄ COMBINING MACRON ABOVE
Pitch accents	U+0301 ◌́ COMBINING ACUTE ACCENT
Pitch accent with vowel length	U+030B ◌̋ COMBINING DOUBLE ACUTE ACCENT

Numbers and Punctuation. Osage uses European numbers and standard Latin punctuation.

20.4 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 let-

ters 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 20-1*.

Figure 20-1. 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*
- 𐐍 GAY is written for *gay*
- 𐐰 THEE is written for *the* or *thee*

Phonetics. An approximate IPA transcription of the sounds represented by the Deseret Alphabet is shown in *Table 20-2*.

Table 20-2. IPA Transcription of Deseret

𐐀𐐨	LONG I	i	𐐙𐐩	BEE	b
𐐁𐐩	LONG E	e	𐐚𐐪	TEE	t
𐐂𐐨	LONG A	ɑ	𐐛𐐩	DEE	d
𐐃𐐨	LONG AH	ɒ	𐐜𐐩	CHEE	tʃ
𐐄𐐨	LONG O	o	𐐝𐐩	JEE	dʒ
𐐅𐐨	LONG OO	u	𐐞𐐩	KAY	k
𐐆𐐩	SHORT I	ɪ	𐐟𐐩	GAY	g
𐐇𐐩	SHORT E	ɛ	𐐠𐐩	EF	f
𐐈𐐩	SHORT A	æ	𐐡𐐩	VEE	v
𐐉𐐩	SHORT AH	ɔ	𐐢𐐩	ETH	θ
𐐊𐐩	SHORT O	ʌ	𐐣𐐩	THEE	ð
𐐋𐐩	SHORT OO	ʊ	𐐤𐐩	ES	s
𐐌𐐨	AY	ɑɪ	𐐥𐐩	ZEE	z
𐐍𐐨	OI	ɔɪ	𐐦𐐩	ESH	ʃ
𐐎𐐨	OW	ɑʊ	𐐧𐐩	ZHEE	ʒ
𐐏𐐨	EW	ju	𐐨𐐩	ER	r
𐐐𐐩	WU	w	𐐩𐐩	EL	l
𐐑𐐩	YEE	j	𐐪𐐩	EM	m
𐐒𐐩	H	h	𐐫𐐩	EN	n
𐐓𐐩	PEE	p	𐐬𐐩	ENG	ŋ