The Unicode Standard
Version 8.0 – 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 Julie D. Allen ... [et al.]. — Version 8.0
Includes bibliographical references and index.
ISBN 978-1-936213-10-8 (http://www.unicode.org/versions/Unicode8.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2015

ISBN 978-1-936213-10-8
Published in Mountain View, CA
August 2015

Chapter 16
Southeast Asia

This chapter documents the following scripts of Southeast Asia, Indonesia, and the Philippines:

Thai	*Tai Le*	*Kayah Li*
Lao	*New Tai Lue*	*Cham*
Myanmar	*Tai Tham*	*Pahawh Hmong*
Khmer	*Tai Viet*	*Pau Cin Hau*

The scripts of Southeast Asia are written from left to right; many use no interword spacing but use spaces or marks between phrases. They are mostly abugidas, but with various idiosyncrasies that distinguish them from the scripts of South Asia.

Thai and Lao are the official scripts of Thailand and Laos, respectively, and are closely related. These scripts are unusual for Brahmi-derived scripts in the Unicode Standard, because for various implementation reasons they depart from logical order in the representation of consonant-vowel sequences. Vowels that occur to the left side of their consonant are represented in visual order before the consonant in a string, even though they are pronounced afterward.

Myanmar is the official script of Myanmar, and is used to write the Burmese language, as well as many minority languages of Myanmar and Northern Thailand. It has a mixed encoding model, making use of both a virama and a killer character, and having explicitly encoded medial consonants.

The Khmer script is used for the Khmer and related languages in the Kingdom of Cambodia.

The term "Tai" refers to a family of languages spoken in Southeast Asia, including Thai, Lao, and Shan. This term is also part of the name of a number of scripts encoded in the Unicode Standard. The Tai Le script is used to write the language of the same name, which is spoken in south central Yunnan (China). The New Tai Lue script, also known as Xishuangbanna Dai, is unrelated to the Tai Le script, but is also used in south Yunnan. New Tai Lue is a simplified form of the more traditional Tai Tham script, which is also known as Lanna. The Tai Tham script is used for the Northern Thai, Tai Lue, and Khün languages. The Tai Viet script is used for the Tai Dam, Tai Dón, and Thai Song languages of northwestern Vietnam, northern Laos, and central Thailand. Unlike the other Tai scripts, the Tai Viet script makes use of a visual order model, similar to that for the Thai and Lao scripts.

Kayah Li is a relatively recently invented script, used to write the Kayah Li languages of Myanmar and Thailand. Although influenced by the Myanmar script, Kayah Li is basically an alphabet in structure.

Cham is a Brahmi-derived script used by the Austronesian language Cham, spoken in the southern part of Vietnam and in Cambodia. It does not use a virama. Instead, the encoding makes use of medial consonant signs and explicitly encoded final consonants.

Pahawh Hmong is an alphabetic script devised for writing the Hmong language in the latter half of the 20th century. Its development includes several revisions. The script is used by Hmong communities in several countries, including the United States and Australia.

The Pau Cin Hau alphabet is a liturgical script of the Laipian religious tradition, which emerged in the Chin Hills region of present-day Chin State, Myanmar at the turn of the 20th century.

16.1 Thai

Thai: U+0E00–U+0E7F

The Thai script is used to write Thai and other Southeast Asian languages, such as Kuy, Lanna Tai, and Pali. It is a member of the Indic family of scripts descended from Brahmi. Thai modifies the original Brahmi letter shapes and extends the number of letters to accommodate features of the Thai language, including tone marks derived from superscript digits. At the same time, the Thai script lacks the conjunct consonant mechanism and independent vowel letters found in most other Brahmi-derived scripts. As in all scripts of this family, the predominant writing direction is from left to right.

Standards. Thai layout in the Unicode Standard is based on the Thai Industrial Standard 620-2529, and its updated version 620-2533.

Encoding Principles. In common with most Brahmi-derived scripts, each Thai consonant letter represents a syllable possessing an inherent vowel sound. For Thai, that inherent vowel is /o/ in the medial position and /a/ in the final position.

The consonants are divided into classes that historically represented distinct sounds, but in modern Thai indicate tonal differences. The inherent vowel and tone of a syllable are then modified by addition of vowel signs and tone marks attached to the base consonant letter. Some of the vowel signs and all of the tone marks are rendered in the script as diacritics attached above or below the base consonant. These combining signs and marks are encoded after the modified consonant in the memory representation.

Most of the Thai vowel signs are rendered by full letter-sized inline glyphs placed either before (that is, to the left of), after (to the right of), or *around* (on both sides of) the glyph for the base consonant letter. In the Thai encoding, the letter-sized glyphs that are placed before (left of) the base consonant letter, in full or partial representation of a vowel sign, are, in fact, encoded as separate characters that are typed and stored *before* the base consonant character. This encoding for left-side Thai vowel sign glyphs (and similarly in Lao and in Tai Viet) differs from the conventions for all other Indic scripts, which uniformly encode all vowels after the base consonant. The difference is necessitated by the encoding practice commonly employed with Thai character data as represented by the Thai Industrial Standard.

The glyph positions for Thai syllables are summarized in *Table 16-1*.

Table 16-1. Glyph Positions in Thai Syllables

Syllable	Glyphs	Code Point Sequence
ka	กะ	0E01 0E30
ka:	กา	0E01 0E32
ki	กิ	0E01 0E34
ki:	กี	0E01 0E35
ku	กุ	0E01 0E38
ku:	กู	0E01 0E39
ku'	กึ	0E01 0E36
ku':	กื	0E01 0E37
ke	เกะ	0E40 0E01 0E30
ke:	เก	0E40 0E01
kae	แกะ	0E41 0E01 0E30
kae:	แก	0E41 0E01
ko	โกะ	0E42 0E01 0E30
ko:	โก	0E42 0E01
ko'	เกาะ	0E40 0E01 0E32 0E30
ko':	กอ	0E01 0E2D
koe	เกอะ	0E40 0E01 0E2D 0E30
koe:	เกอ	0E40 0E01 0E2D
kia	เกีย	0E40 0E01 0E35 0E22
ku'a	เกือ	0E40 0E01 0E37 0E2D
kua	กัว	0E01 0E31 0E27
kaw	เกา	0E40 0E01 0E32
koe:y	เกย	0E40 0E01 0E22
kay	ไก	0E44 0E01
kay	ใก	0E43 0E01
kam	กำ	0E01 0E33
kri	กฤ	0E01 0E24

Rendering of Thai Combining Marks. The canonical combining classes assigned to tone marks (ccc=107) and to other combining characters displayed above (ccc=0) do not fully account for their typographic interaction.

For the purpose of rendering, the Thai combining marks above (U+0E31, U+0E34..U+0E37, U+0E47..U+0E4E) should be displayed outward from the base character they modify, in the order in which they appear in the text. In particular, a sequence containing <U+0E48 THAI CHARACTER MAI EK, U+0E4D THAI CHARACTER NIKHAHIT> should be displayed with the *nikhahit* above the *mai ek*, and a sequence containing <U+0E4D THAI CHARACTER NIKHAHIT, U+0E48 THAI CHARACTER MAI EK> should be displayed with the *mai ek* above the *nikhahit*.

This does not preclude input processors from helping the user by pointing out or correcting typing mistakes, perhaps taking into account the language. For example, because the string <*mai ek, nikhahit*> is not useful for the Thai language and is likely a typing mistake, an input processor could reject it or correct it to <*nikhahit, mai ek*>.

When the character U+0E33 THAI CHARACTER SARA AM follows one or more tone marks (U+0E48..U+0E4B), the *nikhahit* that is part of the *sara am* should be displayed below those tone marks. In particular, a sequence containing <U+0E48 THAI CHARACTER MAI EK, U+0E33 THAI CHARACTER SARA AM> should be displayed with the *mai ek* above the *nikhahit*.

Thai Punctuation. Thai uses a variety of punctuation marks particular to this script. U+0E4F THAI CHARACTER FONGMAN is the Thai bullet, which is used to mark items in lists or appears at the beginning of a verse, sentence, paragraph, or other textual segment. U+0E46 THAI CHARACTER MAIYAMOK is used to mark repetition of preceding letters. U+0E2F THAI CHARACTER PAIYANNOI is used to indicate elision or abbreviation of letters; it is itself viewed as a kind of letter, however, and is used with considerable frequency because of its appearance in such words as the Thai name for Bangkok. *Paiyannoi* is also used in combination (U+0E2F U+0E25 U+0E2F) to create a construct called *paiyanyai*, which means "et cetera, and so forth." The Thai *paiyanyai* is comparable to its analogue in the Khmer script: U+17D8 KHMER SIGN BEYYAL.

U+0E5A THAI CHARACTER ANGKHANKHU is used to mark the end of a long segment of text. It can be combined with a following U+0E30 THAI CHARACTER SARA A to mark a larger segment of text; typically this usage can be seen at the end of a verse in poetry. U+0E5B THAI CHARACTER KHOMUT marks the end of a chapter or document, where it always follows the *angkhankhu* + *sara a* combination. The Thai *angkhankhu* and its combination with *sara a* to mark breaks in text have analogues in many other Brahmi-derived scripts. For example, they are closely related to U+17D4 KHMER SIGN KHAN and U+17D5 KHMER SIGN BARIYOOSAN, which are themselves ultimately related to the *danda* and *double danda* of Devanagari.

Spacing. Thai words are not separated by spaces. Instead, text is laid out with spaces introduced at text segments where Western typography would typically make use of commas or periods. However, Latin-based punctuation such as comma, period, and colon are also used in text, particularly in conjunction with Latin letters or in formatting numbers, addresses, and so forth. If explicit word break or line break opportunities are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justified. See *Table 23-2*.

Thai Transcription of Pali and Sanskrit. The Thai script is frequently used to write Pali and Sanskrit. When so used, consonant clusters are represented by the explicit use of U+0E3A THAI CHARACTER PHINTHU (*virama*) to mark the removal of the inherent vowel. There is no conjoining behavior, unlike in other Indic scripts. U+0E4D THAI CHARACTER NIKHAHIT is the Pali *nigghahita* and Sanskrit *anusvara*. U+0E30 THAI CHARACTER SARA A is the Sanskrit *visarga*. U+0E24 THAI CHARACTER RU and U+0E26 THAI CHARACTER LU are vocalic /r/ and /l/, with U+0E45 THAI CHARACTER LAKKHANGYAO used to indicate their lengthening.

Patani Malay. The Patani Malay orthography makes use of additional diacritics. A line below a consonant indicates that its sound differs from Thai. The line below is represented using U+0331 COMBINING MACRON BELOW. Nasalization is indicated by U+0303 COMBINING TILDE. Glottalization is marked with the character U+02BC MODIFIER LETTER APOSTROPHE. The character U+02D7 MODIFIER LETTER MINUS SIGN indicates an elision between two vowel sequences.

Use of combining diacritics with the Thai script, such as U+0331 COMBINING MACRON BELOW and U+0303 COMBINING TILDE, imposes additional constraints for rendering systems for Thai. This is because the canonical ordering of these marks with respect to Thai vowels and tone marks may put them in orders which require rearranging during rendering.

16.2 Lao

Lao: U+0E80–U+0EFF

The Lao language and script are closely related to Thai. The Unicode Standard encodes the characters of the Lao script in the same relative order as the Thai characters.

Encoding Principles. Lao contains fewer letters than Thai because by 1960 it was simplified to be fairly phonemic, whereas Thai maintains many etymological spellings that are homonyms. Unlike in Thai, Lao consonant letters are conceived of as simply representing the consonant sound, rather than a syllable with an inherent vowel. The vowel [a] is always represented explicitly with U+0EB0 LAO VOWEL SIGN A.

Punctuation. Regular word spacing is not used in Lao; spaces separate phrases or sentences instead.

Glyph Placement. The glyph placements for Lao syllables are summarized in *Table 16-2*.

Table 16-2. Glyph Positions in Lao Syllables

Syllable	Glyphs	Code Point Sequence
ka	ກະ	0E81 0EB0
ka:	ກາ	0E81 0EB2
ki	ກິ	0E81 0EB4
ki:	ກີ	0E81 0EB5
ku	ກຸ	0E81 0EB8
ku:	ກູ	0E81 0EB9
ku'	ກົ	0E81 0EB6
ku':	ກຶ	0E81 0EB7
ke	ເກະ	0EC0 0E81 0EB0
ke:	ເກ	0EC0 0E81
kae	ແກະ	0EC1 0E81 0EB0
kae:	ແກ	0EC1 0E81
ko	ໂກະ	0EC2 0E81 0EB0
ko:	ໂກ	0EC2 0E81
ko'	ເກາະ	0EC0 0E81 0EB2 0EB0
ko':	ກໍ	0E81 0ECD
koe	ເກີ	0EC0 0E81 0EB4
koe:	ເກີ	0EC0 0E81 0EB5

Table 16-2. Glyph Positions in Lao Syllables (Continued)

Syllable	Glyphs	Code Point Sequence
kia	◌ັ ◌ຽ ◌ັ ◌ຍ	0EC0 0E81 0EB1 0EBD 0EC0 0E81 0EA2
ku'a	◌ົ ◌ວ	0EC0 0E81 0EB7 0EAD
kua	◌ົ ◌ວ	0E81 0EBB 0EA7
kaw	◌ົ ◌ໆ	0EC0 0E81 0EBB 0EB2
koe:y	◌ົ ◌ຽ ◌ົ ◌ຍ	0EC0 0E81 0EB5 0EBD 0EC0 0E81 0EB5 0EA2
kay	◌ ◌ ◌	0EC4 0E81
kay	◌ ◌ ◌	0EC3 0E81
kam	◌ ◌ ◌	0E81 0EB3

Additional Letters. A few additional letters in Lao have no match in Thai:

U+0EBB LAO VOWEL SIGN MAI KON

U+0EBC LAO SEMIVOWEL SIGN LO

U+0EBD LAO SEMIVOWEL SIGN NYO

The preceding two semivowel signs are the last remnants of the system of subscript medials, which in Myanmar retains additional distinctions. Myanmar and Khmer include a full set of subscript consonant forms used for conjuncts. Thai no longer uses any of these forms; Lao has just the two.

Rendering of Lao Combining Marks. The canonical combining classes assigned to tone marks (ccc=122) and to other combining characters displayed above (ccc=0) do not fully account for their typographic interaction.

For the purpose of rendering, the Lao combining marks above (U+0EB1, U+0EB4..U+0EB7, U+0EBB, U+0EC8..U+0ECD) should be displayed outward from the base character they modify, in the order in which they appear in the text. In particular, a sequence containing <U+0EC8 LAO TONE MAI EK, U+0ECD LAO NIGGAHITA> should be displayed with the *niggahita* above the *mai ek*, and a sequence containing <U+0ECD LAO NIGGAHITA, U+0EC8 LAO TONE MAI EK> should be displayed with the *mai ek* above the *niggahita*.

This does not preclude input processors from helping the user by pointing out or correcting typing mistakes, perhaps taking into account the language. For example, because the string <*mai ek, niggahita*> is not useful for the Lao language and is likely a typing mistake, an input processor could reject it or correct it to <*niggahita, mai ek*>.

When the character U+0EB3 LAO VOWEL SIGN AM follows one or more tone marks (U+0EC8..U+0ECB), the *niggahita* that is part of the *sara am* should be displayed below those tone marks. In particular, a sequence containing <U+0EC8 LAO TONE MAI EK, U+0EB3 LAO VOWEL SIGN AM> should be displayed with the *mai ek* above the *niggahita*.

Lao Aspirated Nasals. The Unicode character encoding includes two ligatures for Lao: U+0EDC LAO HO NO and U+0EDD LAO HO MO. They correspond to sequences of [h] plus [n] or [h] plus [m] without ligating. Their function in Lao is to provide versions of the [n] and [m] consonants with a different inherent tonal implication.

16.3 Myanmar

Myanmar: U+1000–U+109F

The Myanmar script is used to write Burmese, the majority language of Myanmar (formerly called Burma). Variations and extensions of the script are used to write other languages of the region, such as Mon, Karen, Kayah, Shan, and Palaung, as well as Pali and Sanskrit. The Myanmar script was formerly known as the Burmese script, but the term "Myanmar" is now preferred.

The Myanmar writing system derives from a Brahmi-related script borrowed from South India in about the eighth century to write the Mon language. The first inscription in the Myanmar script dates from the eleventh century and uses an alphabet almost identical to that of the Mon inscriptions. Aside from rounding of the originally square characters, this script has remained largely unchanged to the present. It is said that the rounder forms were developed to permit writing on palm leaves without tearing the writing surface of the leaf.

The Myanmar script shares structural features with other Brahmi-based scripts such as Khmer: consonant symbols include an inherent "a" vowel; various signs are attached to a consonant to indicate a different vowel; medial consonants are attached to the consonant; and the overall writing direction is from left to right.

Standards. There is not yet an official national standard for the encoding of Myanmar/Burmese. The current encoding was prepared with the consultation of experts from the Myanmar Information Technology Standardization Committee (MITSC) in Yangon (Rangoon). The MITSC, formed by the government in 1997, consists of experts from the Myanmar Computer Scientists' Association, Myanmar Language Commission, and Myanmar Historical Commission.

Encoding Principles. As with Indic scripts, the Myanmar encoding represents only the basic underlying characters; multiple glyphs and rendering transformations are required to assemble the final visual form for each syllable. Characters and combinations that may appear visually identical in some fonts, such as U+101D ဝ MYANMAR LETTER WA and U+1040 ၀ MYANMAR DIGIT ZERO, are distinguished by their underlying encoding.

Composite Characters. As is the case in many other scripts, some Myanmar letters or signs may be analyzed as composites of two or more other characters and are not encoded separately. The following are three examples of Myanmar letters represented by combining character sequences:

U+1000 က *ka* + U+1031 ေ *vowel sign e* + U+102C ာ *vowel sign aa* → ေကာ */kàw/*

U+1000 က *ka* + U+1031 ေ *vowel sign e* + U+102C ာ *vowel sign aa* + U+103A ် *asat* → ေကာ် */kaw/*

U+1000 က *ka* + U+102D ိ *vowel sign i* + U+1031 ု *vowel sign u* → ကို */ko/*

Encoding Subranges. The basic consonants, medials, independent vowels, and dependent vowel signs required for writing the Myanmar language are encoded at the beginning of the Myanmar block. Those are followed by script-specific digits, punctuation, and various signs. The last part of the block contains extensions for consonants, medials, vowels, and tone marks needed to represent historic text and various other languages. These extensions support Pali and Sanskrit, as well as letters and tone marks for Mon, Karen, Kayah, and Shan. The extensions include two tone marks for Khamti Shan and two vowel signs for Aiton and Phake, but the majority of the additional characters needed to support those languages are found in the Myanmar Extended-A block.

Conjuncts. As in other Indic-derived scripts, conjunction of two consonant letters is indicated by the insertion of a virama U+1039 ◌္ MYANMAR SIGN VIRAMA between them. It causes the second consonant to be displayed in a smaller form below the first; the virama is not visibly rendered.

Kinzi. The conjunct form of U+1004 င MYANMAR LETTER NGA is rendered as a superscript sign called *kinzi*. That superscript sign is not encoded as a separate mark, but instead is simply the rendering form of the *nga* in a conjunct context. The *nga* is represented in logical order first in the sequence, before the consonant which actually bears the visible *kinzi* superscript sign in final rendered form. For example, *kinzi* applied to U+1000 က MYANMAR LETTER KA would be written via the following sequence:

U+1004 င *nga* + U+103A ◌် *asat* + U+1039 ◌္ *virama* + U+1000 က *ka*
→ င်္က *ka*

Note that this sequence includes both U+103A *asat* and U+1039 *virama* between the *nga* and the *ka*. Use of the *virama* alone would ordinarily indicate stacking of the consonants, with a small *ka* appearing under the *nga*. Use of the *asat* killer in addition to the *virama* gives a sequence that can be distinguished from normal stacking: the sequence <U+1004, U+103A, U+1039> always maps unambiguously to a visible *kinzi* superscript sign on the following consonant.

Medial Consonants. The Myanmar script traditionally distinguishes a set of "medial" consonants: forms of *ya*, *ra*, *wa*, and *ha* that are considered to be modifiers of the syllable's vowel. Graphically, these medial consonants are sometimes written as subscripts, but sometimes, as in the case of *ra*, they surround the base consonant instead. In the Myanmar encoding, the medial consonants are encoded separately. For example, the word ၵျွဲ [kjwei] ("to drop off") would be written via the following sequence:

U+1000 က *ka* + U+103C ◌ျ *medial ra* + U+103D ◌ွ *medial wa* +
U+1031 ◌ေ *vowel sign e* → ၵျွဲ */kjwei/*

In Pali and Sanskrit texts written in the Myanmar script, as well as in older orthographies of Burmese, the consonants *ya*, *ra*, *wa*, and *ha* are sometimes rendered in subjoined form. In those cases, U+1039 ◌္ MYANMAR SIGN VIRAMA and the regular form of the consonant are used.

Asat. The *asat*, or *killer*, is a visibly displayed sign. In some cases it indicates that the inherent vowel sound of a consonant letter is suppressed. In other cases it combines with other characters to form a vowel letter. Regardless of its function, this visible sign is always represented by the character U+103A ◌် MYANMAR SIGN ASAT.

Contractions. In a few Myanmar words, the repetition of a consonant sound is written with a single occurrence of the letter for the consonant sound together with an *asat* sign. This *asat* sign occurs immediately after the double-acting consonant in the coded representation:

U+101A ယ *ya* + U+1031 ◌ေ *vowel sign e* + U+102C ◌ာ *vowel sign aa* + U+1000 က *ka* + U+103A ◌် *asat* + U+103B ◌ျ *medial ya* + U+102C ◌ာ *vowel sign aa* + U+1038 ◌း *visarga* → ယောကျ်ား man, husband

U+1000 က *ka* + U+103B ◌ျ *medial ya* + U+103D ◌ွ *medial wa* + U+1014 န *na* + U+103A ◌် *asat* + U+102F ◌ု *vowel sign u* + U+1015 ပ *pa* + U+103A ◌် *asat* → ကျွန်ုပ် I (first person singular)

Great sa. The *great sa* is encoded as U+103F ဿ MYANMAR LETTER GREAT SA. This letter should be represented with <U+103F>, while the sequence <U+101E, U+1039, U+101E> should be used for the regular conjunct form of two *sa*, သ္သ, and the sequence <U+101E, U+103A, U+101E> should be used for the form with an *asat sign*, သ်သ.

Tall aa. The two shapes ◌ါ and ◌ာ are both used to write the sound /a/. In Burmese orthography, both shapes are used, depending on the visual context. In S'gaw Karen orthography, only the tall form is used. For this reason, two characters are encoded: U+102B ◌ါ MYANMAR VOWEL SIGN TALL AA and U+102C ◌ာ MYANMAR VOWEL SIGN AA. In Burmese texts, the coded character appropriate to the visual context should be used.

Ordering of Syllable Components. Dependent vowels and other signs are encoded after the consonant to which they apply, except for *kinzi*, which precedes the consonant. Characters occur in the relative order shown in *Table 16-3*.

Table 16-3. Modern Burmese Syllabic Structure

Class	Example	Encoding
kinzi	◌ဳ်	<U+1004, U+103A, U+1039>
consonant and vowel letters	က	[U+1000..U+1021, U+1023..U+1027, U+1029, U+102A, U+103F, U+104E]
subscript consonant	◌္ဿ	<U+1039, [U+1000..U+1008, U+100A..U+1019, U+101B, U+101C, U+101E, U+1020, U+1021]>
asat sign	◌်	U+103A

Table 16-3. Modern Burmese Syllabic Structure (Continued)

Class	Example	Encoding
medial ya (potentially followed by asat sign)	◌ျ	<U+103B, (U+103A)>
medial ra	◌ြ	U+103C
medial wa	◌ွ	U+103D
medial ha	◌ှ	U+103E
vowel sign e	ေ◌	U+1031
vowel sign i, ii, ai	◌ိ, ◌ီ, ◌ဲ	[U+102D, U+102E, U+1032]
vowel sign u, uu	◌ု, ◌ူ	[U+102F, U+1030]
vowel sign tall aa, aa (potentially followed by asat sign)	◌ါ, ◌ော	<[U+102B, U+102C], (U+103A)>
anusvara	◌ံ	U+1036
dot below	◌့	U+1037
visarga	◌း	U+1038

U+1031 ေ◌ MYANMAR VOWEL SIGN E is encoded after its consonant (as in the earlier example), although in visual presentation its glyph appears before (to the left of) the consonant form.

Table 16-3 nominally refers to the character sequences used in representing the syllabic structure of the modern Burmese language proper. Canonical normalization may result in a different ordering, specifically with some occurrences of U+103A ◌် MYANMAR SIGN ASAT reordered after U+1037 ◌့ MYANMAR SIGN DOT BELOW. As such reorderings are canonically equivalent, implementations should support both orders and treat them as fundamentally the same text.

Table 16-3 would require further extensions and modifications to cover various other languages, such as Karen, Mon, Shan, Sanskrit, and Old Burmese, which also use the Myanmar script. For some such extensions and modifications, refer to Unicode Technical Note #11, "Representing Myanmar in Unicode: Details and Examples," or also Microsoft Typography's "Creating and Supporting OpenType Fonts for Myanmar Script." Note that those documents are not normative for the Unicode Standard, and they also differ from each other in some details.

Spacing. Myanmar does not use any whitespace between words. If explicit word break or line break opportunities are desired—for example, for the use of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justi-

fied. Spaces are used to mark phrases. Some phrases are relatively short (two or three syllables).

Myanmar Extended-A: U+AA60–U+AA7F

This block provides additional characters to support Khamti Shan, Aiton and Phake. The block also contains a few additional tone marks for Pa'o Karen and Tai Laing, and two additional letters for Shwe Palaung. Khamti Shan is spoken by approximately 14,000 people in Myanmar and India. Aiton and Phake are smaller language communities of around 2,000 each. Many of the characters needed for these languages are provided by the main Myanmar block. Khamti Shan, Aiton, and Phake writing conventions are based on Shan, and as such follow the general Myanmar model of encoding.

Khamti Shan

The Khamti Shan language has a long literary tradition which has largely been lost, for a variety of reasons. The old script did not mark tones, and it had a scribal tradition that encouraged restriction to a reading elite whose traditions have not been passed on. The script has recently undergone a revival, with plans for it to be taught throughout the Khamti-Shan-speaking regions in Myanmar. A new version of the script has been adopted by the Khamti in Myanmar. The Khamti Shan characters in the Myanmar Extended-A block supplement those in the Myanmar block and provide complete support for the modern Khamti Shan writing system as written in Myanmar. Another revision of the old script was made in India under the leadership of Chau Khouk Manpoong in the 1990s. That revision has not gained significant popularity, although it enjoys some currency today.

Consonants. Approximately half of the consonants used in Khamti Shan are encoded in the Myanmar block. Following the conventions used for Shan, Mon, and other extensions to the Myanmar script, separate consonants are encoded specifically for Khamti Shan in this block when they differ significantly in shape from corresponding letters conveying the same consonant sounds in Myanmar proper. Khamti Shan also uses the three Myanmar medial consonants encoded in the range U+101B..U+101D.

The consonants in this block are displayed in the code charts using a Burmese style, so that glyphs for the entire Myanmar script are harmonized in a single typeface. However, the local style preferred for Khamti Shan is slightly different, typically adding a small dot to each character.

Vowels. The vowels and dependent vowel signs used in Khamti Shan are located in the Myanmar block.

Tones. Khamti Shan has eight tones. Seven of these are written with explicit tone marks; one is unmarked. All of the explicit tone marks are encoded in the Myanmar block. Khamti Shan makes use of four of the Shan tone marks and the *visarga*. In addition, two Khamti Shan-specific tone marks are separately encoded. These tone marks for Khamti Shan are listed in *Table 16-4*.

Table 16-4. Khamti Shan Tone Marks

Tone	Character
1	U+109A MYANMAR SIGN KHAMTI TONE-1
2	U+1089 MYANMAR SIGN SHAN TONE-5
3	U+109B MYANMAR SIGN KHAMTI TONE-3
4	U+1087 MYANMAR SIGN SHAN TONE-2
5	U+1088 MYANMAR SIGN SHAN TONE-3
6	U+1038 MYANMAR SIGN VISARGA
7	*unmarked*
8	U+108A MYANMAR SIGN SHAN TONE-6

The vertical positioning of the small circle in some of these tone marks is considered distinctive. U+109A MYANMAR SIGN KHAMTI TONE-1 (with a high position) is not the same as U+108B MYANMAR SIGN SHAN COUNCIL TONE-2 (with a mid-level position). Neither of those should be confused with U+1089 MYANMAR SIGN SHAN TONE-5 (with a low position).

The tone mark characters in Shan fonts are typically displayed with open circles. However, in Khamti Shan, the circles in the tone marks normally are filled in (black).

Digits. Khamti Shan uses the Shan digits from the range U+1090..U+109A.

Other Symbols. Khamti Shan uses the punctuation marks U+104A MYANMAR SIGN LITTLE SECTION and U+104B MYANMAR SIGN SECTION. The repetition mark U+AA70 MYANMAR MODIFIER LETTER KHAMTI REDUPLICATION is functionally equivalent to U+0E46 THAI CHARACTER MAIYAMOK.

Three logogram characters are also used. These logograms can take tone marks, and their meaning varies according to the tone they take. They are used when transcribing speech rather than in formal writing. For example, U+AA75 MYANMAR LOGOGRAM KHAMTI QN takes three tones and means "negative," "giving" or "yes," according to which tone is applied. The other two logograms are U+AA74 MYANMAR LOGOGRAM KHAMTI OAY and U+AA76 MYANMAR LOGOGRAM KHAMTI HM.

Subjoined Characters. Khamti Shan does not use subjoined characters.

Historical Khamti Shan. The characters of historical Khamti Shan are for the most part identical to those used in the New Khamti Shan orthography. Most variation is merely stylistic. There were no Pali characters. The only significant character difference lies with *ra*—which follows Aiton and Phake in using a *la* with *medial ra* (U+AA7A MYANMAR LETTER AITON RA).

During the development of the New Khamti Shan orthography a few new character shapes were introduced that were subsequently revised. Because materials have been published using these shapes, and these shapes cannot be considered stylistic variants of other characters, these characters are separately encoded in the range U+AA71..U+AA73.

Aiton and Phake

The Aiton and Phake writing systems are very closely related. There are a small number of differences in shape between Aiton and Phake characters, but these are considered only glyphic differences. As for Khamti Shan, most of the characters needed for Aiton and Phake are found in the Myanmar block.

Consonants. U+107A MYANMAR LETTER SHAN NYA is used rather than following the Khamti U+AA65 MYANMAR LETTER KHAMTI NYA because the character shape follows Shan rather than Khamti.

Subjoined Consonants. Aiton and Phake have a system of subjoining consonants to chain syllables in a polysyllabic word. This system follows that of Burmese and is encoded in the same way: with U+1039 MYANMAR SIGN VIRAMA followed by the code of the consonant being subjoined. The following characters may take a subjoined form, which takes the same shape as the base character but smaller: U+1000, U+AA61, U+1010, U+1011, U+1015, U+101A, U+101C. No other subjoined characters are known in Aiton and Phake.

Vowels. The vowels follow Shan for the most part, and are therefore based on the characters in the Myanmar block. In addition to the simple vowels there are a number of diphthongs in Aiton and Phake. One vowel and one diphthong required for these languages were added as extensions at the end of the Myanmar block. A number of the vowel letters and diphthongs in the Aiton and Phake alphabets are composed of a sequence of code points. For example, the vowel *-ue* is represented by the sequence <U+102D, U+102F, U+101D, U+103A>.

Ligatures. The characters in the range U+AA77..U+AA79 are a set of ligature symbols that follow the same principles used for U+109E MYANMAR SYMBOL SHAN ONE and U+109F MYANMAR SYMBOL SHAN EXCLAMATION. They are symbols that constitute a word in their own right and do not take diacritics.

Tones. Traditionally tones are not marked in Aiton and Phake, although U+109C MYANMAR VOWEL SIGN AITON A (*short -a*) can be used as a type of tone marker. All proposed patterns for adding tone marking to Aiton and Phake can be represented with the tone marks used for Shan or Khamti Shan.

Myanmar Extended-B: U+A9E0–U+A9FF

This block contains additional characters for Shan Pali that represent Sanskrit sounds written in Shan. It also contains many characters for Tai Laing, a Tai language related to Khamti and spoken in the Kachin state of Myanmar. Tai Laing has a distinct set of digits that differ in appearance from both the main set of Myanmar digits and the Shan digits encoded in the main Myanmar block.

16.4 Khmer

Khmer: U+1780–U+17FF

Khmer, also known as Cambodian, is the official language of the Kingdom of Cambodia. Mutually intelligible dialects are also spoken in northeastern Thailand and in the Mekong Delta region of Vietnam. Although Khmer is not an Indo-European language, it has borrowed much vocabulary from Sanskrit and Pali, and religious texts in those languages have been both transliterated and translated into Khmer. The Khmer script is also used to render a number of regional minority languages, such as Tampuan, Krung, and Cham.

The Khmer script, called *aksaa khmae* ("Khmer letters"), is also the official script of Cambodia. It is descended from the Brahmi script of South India, as are Thai, Lao, Myanmar, Old Mon, and others. The exact sources have not been determined, but there is a great similarity between the earliest inscriptions in the region and the Pallawa script of the Coromandel coast of India. Khmer has been a unique and independent script for more than 1,400 years. Modern Khmer has two basic styles of script: the *aksaa crieng* ("slanted script") and the *aksaa muul* ("round script"). There is no fundamental structural difference between the two. The slanted script (in its "standing" variant) is chosen as representative in the code charts.

Principles of the Khmer Script

Structurally, the Khmer script has many features in common with other Brahmi-derived scripts, such as Devanagari and Myanmar. Consonant characters bear an inherent vowel sound, with additional signs placed before, above, below, and/or after the consonants to indicate a vowel other than the inherent one. The overall writing direction is left to right.

In comparison with the Devanagari script, explained in detail in *Section 12.1, Devanagari*, the Khmer script has developed several distinctive features during its evolution.

Glottal Consonant. The Khmer script has a consonant character for a glottal stop (*qa*) that bears an inherent vowel sound and can have an optional vowel sign. While Khmer also has independent vowel characters like Devanagari, as shown in *Table 16-5*, in principle many of its sounds can be represented by using *qa* and a vowel sign. This does not mean these representations are always interchangeable in real words. Some words are written with one variant to the exclusion of others.

Subscript Consonants. Subscript consonant signs differ from independent consonant characters and are called *coeng* (literally, "foot, leg") after their subscript position. While a consonant character can constitute an orthographic syllable by itself, a subscript consonant sign cannot. Note that U+17A1 ឡ KHMER LETTER LA does not have a corresponding subscript consonant sign in standard Khmer, but does have a subscript in the Khmer script used in Thailand.

Table 16-5. Independent Khmer Vowel Characters

Name	Independent Vowel	Qa with Vowel Sign
i	ឥ	អិ, អិៈ, អី
ii	ឦ	អី, អិៈ
u	ឧ	អុ, អុៈ
uk	ឪ	អុក
uu	ឩ	អូ, អូៈ
uuv	ឰ	អូរ
ry	ឫ	រឹ
ryy	ឬ	រឺ
ly	ឭ	លឹ
lyy	ឮ	លឺ
e	ឯ	អេ, ែអ
ai	ឰ	ៃអ
oo	ឱ, ឲ	អោ
au	ឳ	អៅ

Subscript consonant signs are used to represent any consonant following the first consonant in an orthographic syllable. They also have an inherent vowel sound, which may be suppressed if the syllable bears a vowel sign or another subscript consonant.

The subscript consonant signs are often used to represent a consonant cluster. Two consecutive consonant characters cannot represent a consonant cluster because the inherent vowel sound in between is retained. To suppress the vowel, a subscript consonant sign (or rarely a subscript independent vowel) replaces the second consonant character. Theoretically, any consonant cluster composed of any number of consonant sounds without inherent vowel sounds in between can be represented systematically by a consonant character and as many subscript consonant signs as necessary.

Examples of subscript consonant signs for a consonant cluster follow:

ល្ង *lo* + *coeng* + *ngo* [lŋɔ̀ː] “sesame” (compare លង *lo* + *ngo* [lɔ̀ːŋ] “to haunt”)

លក្ខ្មី *lo* + *ka* + *coeng* + *sa* + *coeng* + *mo* + *ii* [lɛ̀əksmei] “beauty, luck”

កាហ្វេ *ka* + *aa* + *ha* + *coeng* + *vo* + *e* [kaːfeː] “coffee”

The subscript consonant signs in the Khmer script can be used to denote a final consonant, although this practice is uncommon.

Examples of subscript consonant signs for a closing consonant follow:

ទាំ្ង *to* + *aa* + *nikahit* + *coeng* + *ngo* [tèəŋ] "both" (= ទាំង) (≠ *ទ្ងាំ [tŋɔ̀əm])

ហើ្យ *ha* + *oe* + *coeng* + *yo* [haəi] "already" (= ហើយ) (≠ *ហ្យើ [hyaə])

While these subscript consonant signs are usually attached to a consonant character, they can also be attached to an independent vowel character. Although this practice is relatively rare, it is used in one very common word, meaning "to give."

Examples of subscript consonant signs attached to an independent vowel character follow:

ឲ្យ *qoo-1* + *coeng* + *yo* [ʔaoi] "to give" (= ឱយ and also ឲ្យ)

ឱ្ម *qoo-1* + *coeng* + *mo* [ʔaom] "exclamation of solemn affirmation" (= ឱម)

Subscript Independent Vowel Signs. Some independent vowel characters also have corresponding subscript independent vowel signs, although these are rarely used today.

Examples of subscript independent vowel signs follow:

ផ្ឹម *pha* + *coeng* + *qe* + *mo* [pʰʔaem] "sweet" (= ផែ្ម *pha* + *coeng* + *qa* + *ae* + *mo*)

ហ្ឫទ័យ *ha* + *coeng* + *ry* + *to* + *samyok sannya* + *yo* [harɯtey] "heart" (*royal*) (= ហឫទ័យ *ha* + *ry* + *to* + *samyok sannya* + *yo*)

Consonant Registers. The Khmer language has a richer set of vowels than the languages for which the ancestral script was used, although it has a smaller set of consonant sounds. The Khmer script takes advantage of this situation by assigning different characters to represent the same consonant using different inherent vowels. Khmer consonant characters and signs are organized into two series or registers, whose inherent vowels are nominally *-a* in the first register and *-o* in the second register, as shown in *Table 16-6*.

Table 16-6. Two Registers of Khmer Consonants

Row	First Register	Second Register
1	ក *ka* [kɔː] "neck"	គ *ko* [kɔ̀ː] "mute"
2	រ *ro* + *muusikatoan* [rɔː] "small saw"	រ *ro* [rɔ̀ː] "fence (in the water)"
3	សក *sa* + *ka* [sɔːk] "to peel, to shed one's skin"	ស៊ក *sa* + *triisap* + *ka* [sɔ̀ːk] "to insert"
4	បក *ba* + *ka* [bɔːk] "to return"	*ប៊ក *ba* + *triisap* + *ka* [bɔ̀ːk]
5	ប៉ម *ba* + *muusikatoan* + *mo* [pɔːm] "blockhouse"	ពម *po* + *mo* [pɔ̀ːm] "to put into the mouth"
6	កូរ *ka* + *u* + *ro* [koː] "to stir"	គូរ *ko* + *u* + *ro* [kuː] "to sketch"

The register of a consonant character is generally reflected on the last letter of its transliterated name. Some consonant characters and signs have a counterpart whose consonant sound is the same but whose register is different, as *ka* and *ko* in the first row of the table. For the other consonant characters and signs, two "shifter" signs are available. U+17C9 KHMER SIGN MUUSIKATOAN converts a consonant character and sign from the second to the first register, while U+17CA KHMER SIGN TRIISAP converts a consonant from the first register to the second (rows 2–4). To represent *pa*, however, *muusikatoan* is attached not to *po* but to *ba*, in an exceptional use (row 5). The phonetic value of a dependent vowel sign may also change depending on the context of the consonant(s) to which it is attached (row 6).

Encoding Principles. Like other related scripts, the Khmer encoding represents only the basic underlying characters; multiple glyphs and rendering transformations are required to assemble the final visual form for each orthographic syllable. Individual characters, such as U+1789 KHMER LETTER NYO, may assume variant forms depending on the other characters with which they combine.

Subscript Consonant Signs. In the way that many Cambodians analyze Khmer today, subscript consonant signs are considered to be different entities from consonant characters. The Unicode Standard does not assign independent code points for the subscript consonant signs. Instead, each of these signs is represented by the sequence of two characters: a special control character (U+17D2 KHMER SIGN COENG) and a corresponding consonant character. This is analogous to the virama model employed for representing conjuncts in other related scripts. Subscripted independent vowels are encoded in the same manner. Because the *coeng sign* character does not exist as a letter or sign in the Khmer script, the Unicode model departs from the ordinary way that Khmer is conceived of and taught to native Khmer speakers. Consequently, the encoding may not be intuitive to a native user of the Khmer writing system, although it is able to represent Khmer correctly.

U+17D2 ◌្ KHMER SIGN COENG is not actually a *coeng* but a *coeng* generator, because *coeng* in Khmer refers to the subscript consonant sign. The glyph for U+17D2 ◌្ KHMER SIGN COENG shown in the code charts is arbitrary and is not actually rendered directly; the dotted box around the glyph indicates that special rendering is required. To aid Khmer script users, a listing of typical Khmer subscript consonant letters has been provided in *Table 16-7* together with their descriptive names following preferred Khmer practice. While the Unicode encoding represents both the subscripts and the combined vowel letters with a pair of code points, they should be treated as a unit for most processing purposes. In other words, the sequence functions as if it had been encoded as a single character. A number of independent vowels also have subscript forms, as shown in *Table 16-9*.

Table 16-7. Khmer Subscript Consonant Signs

Glyph	Code	Name
◌្ក	17D2 1780	*khmer consonant sign coeng ka*
◌្ខ	17D2 1781	*khmer consonant sign coeng kha*
◌្គ	17D2 1782	*khmer consonant sign coeng ko*

Table 16-7. Khmer Subscript Consonant Signs (Continued)

Glyph	Code	Name
◌្ឍ	17D2 1783	*khmer consonant sign coeng kho*
◌្ង	17D2 1784	*khmer consonant sign coeng ngo*
◌្ច	17D2 1785	*khmer consonant sign coeng ca*
◌្ឆ	17D2 1786	*khmer consonant sign coeng cha*
◌្ជ	17D2 1787	*khmer consonant sign coeng co*
◌្ឈ	17D2 1788	*khmer consonant sign coeng cho*
◌្ញ	17D2 1789	*khmer consonant sign coeng nyo*
◌្ដ	17D2 178A	*khmer consonant sign coeng da*
◌្ឋ	17D2 178B	*khmer consonant sign coeng ttha*
◌្ឌ	17D2 178C	*khmer consonant sign coeng do*
◌្ឍ	17D2 178D	*khmer consonant sign coeng ttho*
◌្ណ	17D2 178E	*khmer consonant sign coeng na*
◌្ត	17D2 178F	*khmer consonant sign coeng ta*
◌្ថ	17D2 1790	*khmer consonant sign coeng tha*
◌្ទ	17D2 1791	*khmer consonant sign coeng to*
◌្ធ	17D2 1792	*khmer consonant sign coeng tho*
◌្ន	17D2 1793	*khmer consonant sign coeng no*
◌្ប	17D2 1794	*khmer consonant sign coeng ba*
◌្ផ	17D2 1795	*khmer consonant sign coeng pha*
◌្ព	17D2 1796	*khmer consonant sign coeng po*
◌្ភ	17D2 1797	*khmer consonant sign coeng pho*
◌្ម	17D2 1798	*khmer consonant sign coeng mo*
◌្យ	17D2 1799	*khmer consonant sign coeng yo*
◌្រ	17D2 179A	*khmer consonant sign coeng ro*
◌្ល	17D2 179B	*khmer consonant sign coeng lo*
◌្វ	17D2 179C	*khmer consonant sign coeng vo*
◌្ឝ	17D2 179D	*khmer consonant sign coeng sha*
◌្ឞ	17D2 179E	*khmer consonant sign coeng ssa*

Table 16-7. Khmer Subscript Consonant Signs (Continued)

Glyph	Code	Name
◌្ប	17D2 179F	*khmer consonant sign coeng sa*
◌្ហ	17D2 17A0	*khmer consonant sign coeng ha*
◌្ឡ	17D2 17A1	*khmer consonant sign coeng la*
◌្ឣ	17D2 17A2	*khmer vowel sign coeng qa*

As noted earlier, <U+17D2, U+17A1> represents a subscript form of *la* that is not used in Cambodia, although it is employed in Thailand.

Dependent Vowel Signs. Most of the Khmer dependent vowel signs are represented with a single character that is applied after the base consonant character and optional subscript consonant signs. Three of these Khmer vowel signs are not encoded as single characters in in the Unicode Standard. The vowel sign *am* is encoded as a nasalization sign, U+17C6 KHMER SIGN NIKAHIT. Two vowel signs, *om* and *aam*, have not been assigned independent code points. They are represented by the sequence of a vowel (U+17BB KHMER VOWEL SIGN U and U+17B6 KHMER VOWEL SIGN AA, respectively) and U+17C6 KHMER SIGN NIKAHIT.

The *nikahit* is superficially similar to *anusvara*, the nasalization sign in the Devanagari script, although in Khmer it is usually regarded as a vowel sign *am*. *Anusvara* not only represents a special nasal sound, but also can be used in place of one of the five nasal consonants homorganic to the subsequent consonant (velar, palatal, retroflex, dental, or labial, respectively). *Anusvara* can be used concurrently with any vowel sign in the same orthographic syllable. *Nikahit*, in contrast, functions differently. Its final sound is [m], irrespective of the type of the subsequent consonant. It is not used concurrently with the vowels *ii*, *e*, *ua*, *oe*, *oo*, and so on, although it is used with the vowel signs *aa* and *u*. In these cases the combination is sometimes regarded as a unit—*aam* and *om*, respectively. The sound that *aam* represents is [ɔ̀əm], not [aːm]. The sequences used for these combinations are shown in *Table 16-8*.

Table 16-8. Khmer Composite Dependent Vowel Signs with Nikahit

Glyph	Code	Name
◌ុំ	17BB 17C6	*khmer vowel sign om*
◌ាំ	17B6 17C6	*khmer vowel sign aam*

Examples of dependent vowel signs ending with [m] follow:

ដុំ *da* + *nikahit* [dɔm] “to pound” (compare ដម *da* + *mo* [dɔːm] “nectar”)

ពាំ *po* + *aa* + *nikahit* [pɔ̀əm] “to carry in the beak” (compare ពម *po* + *aa* + *mo* [pèəm] “mouth of a river”)

Independent Vowel Characters. In Khmer, as in other Brahmic scripts, some independent vowels have their own letterforms, although the sounds they represent may more often be represented with the consonant character for the glottal stop (U+17A2 KHMER LETTER QA) modified by vowel signs (and optionally a consonant character). These independent vowels are encoded as separate characters in the Unicode Standard.

Subscript Independent Vowel Signs. Some independent vowels have corresponding subscript independent vowel signs, although these are rarely used. Each is represented by the sequence of U+17D2 KHMER SIGN COENG and an independent vowel, as shown in *Table 16-9*.

Table 16-9. Khmer Subscript Independent Vowel Signs

Glyph	Code	Name
◌្ឧ	17D2 17A7	*khmer independent vowel sign coeng qu*
◌្ឫ	17D2 17AB	*khmer independent vowel sign coeng ry*
◌្ឬ	17D2 17AC	*khmer independent vowel sign coeng ryy*
◌្ឯ	17D2 17AF	*khmer independent vowel sign coeng qe*

Other Signs as Syllabic Components. The Khmer sign *robat* historically corresponds to the Devanagari *repha*, a representation of syllable-initial *r-*. However, the Khmer script can treat the initial *r-* in the same way as the other initial consonants—namely, a consonant character *ro* and as many subscript consonant signs as necessary. Some old loan words from Sanskrit and Pali include *robat*, but in some of them the *robat* is not pronounced and is preserved in a fossilized spelling. Because *robat* is a distinct sign from the consonant character *ro*, the Unicode Standard encodes U+17CC KHMER SIGN ROBAT, but it treats the Devanagari *repha* as a part of a ligature without encoding it. The authoritative Chuon Nath dictionary sorts *robat* as if it were a base consonant character, just as the *repha* is sorted in scripts that use it. The consonant over which *robat* resides is then sorted as if it were a subscript.

Examples of consonant clusters beginning with *ro* and *robat* follow:

> រាជឫសី *ro + aa + co + ro + coeng + sa + ii* [rèəcrsei] “king hermit”
>
> អារ្យ៏ *qa + aa + yo + robat* [ʔaːrya] “civilized” (= អារ្យ *qa + aa + ro + coeng + yo*)
>
> ពត៌មាន *po + ta + robat + mo + aa + no* [pɔ̀ːdɔmèən] “news” (compare Sanskrit वर्तमान *vartamāna* “the present time”)

U+17DD KHMER SIGN ATTHACAN is a rarely used sign that denotes that the base consonant character keeps its inherent vowel sound. This use contrasts with U+17D1 KHMER SIGN VIRIAM, which indicates the removal of the inherent vowel sound of a base consonant.

U+17CB KHMER SIGN BANTOC shortens the vowel sound of the previous orthographic syllable. U+17C7 KHMER SIGN REAHMUK, U+17C8 KHMER SIGN YUUKALEAPINTU, U+17CD KHMER SIGN TOANDAKHIAT, U+17CE KHMER SIGN KAKABAT, U+17CF KHMER SIGN AHSDA, and U+17D0 KHMER SIGN SAMYOK SANNYA are also explicitly encoded signs used to compose an orthographic syllable.

Ligatures. Some vowel signs form ligatures with consonant characters and signs. These ligatures are not encoded separately, but should be presented graphically by the rendering software. Some common ligatures are shown in *Figure 16-1*.

Figure 16-1. Common Ligatures in Khmer

ក *ka* + ◌ា *aa* + រ *ro* = ការ [ka:] "job"

ប *ba* + ◌ា *aa* = បា [ba:] "father, male of an animal"; used to prevent confusion with ហ *ha*

ប *ba* + ◌ៅ *au* = បៅ [baw] "to suck"

ម *mo* + ◌្ស *coeng sa* + ◌ៅ *au* = ម្សៅ [msaw] "powder"

ស *sa* + ង *ngo* + ◌្ខ *coeng kha* + ◌្យ *coeng yo* + ◌ា *aa* = សង្ខ្យា [sɔŋkʰya:] "counting"

Multiple Glyphs. A single character may assume different forms according to context. For example, a part of the glyph for *nyo* is omitted when a subscript consonant sign is attached. The implementation must render the correct glyph according to context. *Coeng nyo* also changes its shape when it is attached to *nyo*. The correct glyph for the sequence <U+17D2 KHMER SIGN COENG, U+1789 KHMER LETTER NYO> is rendered according to context, as shown in *Figure 16-2*. This kind of glyph alternation is very common in Khmer. Some spacing subscript consonant signs change their height depending on the orthographic context. Similarly, the vertical position of many signs varies according to context. Their presentation is left to the rendering software.

U+17B2 ឲ KHMER INDEPENDENT VOWEL QOO TYPE TWO is thought to be a variant of U+17B1 ឱ KHMER INDEPENDENT VOWEL QOO TYPE ONE, but it is explicitly encoded in the Unicode Standard. The variant is used in very few words, but these include the very common word *aoi* "to give," as noted in *Figure 16-2*.

Figure 16-2. Common Multiple Forms in Khmer

ញញឹម *nyo* + *nyo* + *y* + *mo* [ɲɔ̀ɲɯm] "to smile"

ចិញ្ចើម *ca* + *i* + *nyo* + *coeng* + *ca* + *oe* + *mo* [ceɲcaəm] "eyebrow"

ស្ងប់ *sa* + *coeng nyo* + *ba* + *bantoc* [sɲɔp] "to respect"

កញ្ញា *ka* + *nyo* + *coeng* + *nyo* + *aa* [kaɲɲa:] "girl, Miss, September"

ឲ្យ *qoo-2* + *coeng* + *yo* (= ឱ្យ *qoo-1* + *coeng* + *yo*) [ʔaoi] "to give"

Characters Whose Use Is Discouraged. Some of the Khmer characters encoded in the Unicode Standard are not recommended for use for various reasons.

U+17A3 KHMER INDEPENDENT VOWEL QAQ and U+17A4 KHMER INDEPENDENT VOWEL QAA are deprecated, and their use is strongly discouraged. One feature of the Khmer script is the introduction of the consonant character for a glottal stop (U+17A2 KHMER LETTER QA). This made it unnecessary for each initial vowel sound to have its own independent vowel character, although some independent vowels exist. Neither U+17A3 nor U+17A4 actually exists in the Khmer script. Other related scripts, including the Devanagari script, have independent vowel characters corresponding to them (*a* and *aa*), but they can be transliterated by *khmer letter qa* and *khmer letter qa* + *khmer vowel aa,* respectively, without ambiguity because these scripts have no consonant character corresponding to the *khmer qa*.

The use of U+17B4 KHMER VOWEL INHERENT AQ and U+17B5 KHMER VOWEL INHERENT AA is discouraged. These newly invented characters do not exist in the Khmer script. They were intended to be used to represent a phonetic difference not expressed by the spelling, so as to assist in phonetic sorting. However, they are insufficient for that purpose and should be considered errors in the encoding. These two characters are ignored by default for collation.

The use of U+17D8 KHMER SIGN BEYYAL is discouraged. It was supposed to represent "et cetera" in Khmer. However, it is a word rather than a symbol. Moreover, it has several different spellings. It should be spelled out fully using normal letters. *Beyyal* can be written as follows:

។ល។ *khan* + *ba* + *e* + *khan*

–ល– *en dash* + *ba* + *e* + *en dash*

។ ល ។ *khan* + *lo* + *khan*

–ល– *en dash* + *lo* + *en dash*

Ordering of Syllable Components. The standard order of components in an orthographic syllable as expressed in BNF is

B {R | C} {S {R}}* {{Z} V} {O} {S}

where

B is a base character (consonant character, independent vowel character, and so on)

R is a *robat*

C is a consonant shifter

S is a subscript consonant or independent vowel sign

V is a dependent vowel sign

Z is a zero width non-joiner or a zero width joiner

O is any other sign

For example, the common word ខ្ញុំ khnyom “I” is composed of the following three elements: (1) consonant character *kha* as B; (2) subscript consonant sign *coeng nyo* as S; and (3) dependent vowel sign *om* as V. In the Unicode Standard, *coeng nyo* and *om* are further decomposed, and the whole word is represented by five coded characters.

ខ្ញុំ *kha* + *coeng* + *nyo* + *u* + *nikahit* [kʰɲom] “I”

The order of coded characters does not always match the visual order. For example, some of the dependent vowel signs and their fragments may seem to precede a consonant character, but they are always put after it in the sequence of coded characters. This is also the case with *coeng ro*. Examples of visual reordering and other aspects of syllabic order are shown in *Figure 16-3*.

Figure 16-3. Examples of Syllabic Order in Khmer

ទេ *to* + *e* [tèː] “much”

ច្រើន *ca* + *coeng* + *ro* + *oe* + *no* [craən] “much”

សង្គ្រាម *sa* + *ngo* + *coeng* + *ko* + *coeng* + *ro* + *aa* + *mo* [sɔŋkrèəm] “war”

ហើយ *ha* + *oe* + *coeng* + *yo* [haəi] “already”

សញ្ញា *sa* + *nyo* + *coeng* + *nyo* + *aa* [saɲɲaː] “sign”

ស៊ី *sa* + *triisap* + *ii* [siː] “eat”

ប៉ី *ba* + *muusikatoan* + *ii* [pei] “a kind of flute”

Consonant Shifters. U+17C9 KHMER SIGN MUUSIKATOAN and U+17CA KHMER SIGN TRIISAP are consonant shifters, also known as register shifters. In the presence of other superscript glyphs, both of these signs are usually rendered with the same glyph shape as that of U+17BB KHMER VOWEL SIGN U, as shown in the last two examples of *Figure 16-3*.

Although the consonant shifter in handwriting may be written after the subscript, the consonant shifter should always be encoded immediately following the base consonant, except when it is preceded by U+200C ZERO WIDTH NON-JOINER. This provides Khmer with a fixed order of character placement, making it easier to search for words in a document.

ម៉ៃ *mo* + *muusikatoan* + *coeng* + *ngo* + *ai* [mŋai] “one day”

ម្ហ៊ែតៗ *mo* + *triisap* + *coeng* + *ha* + *ae* + *ta* + *lek too* [mhèːtmhèːt] “bland”

If either *muusikatoan* or *triisap* needs to keep its superscript shape (as an exception to the general rule that states other superscripts typically force the alternative subscript glyph for either character), U+200C ZERO WIDTH NON-JOINER should be inserted before the consonant shifter to show the normal glyph for a consonant shifter when the general rule requires the alternative glyph. In such cases, U+200C ZERO WIDTH NON-JOINER is inserted before the vowel sign, as shown in the following examples:

បៀរ *ba* + [ZWNJ] + *triisap* + *ii* + *yo* + *ae* + *ro* [biyɛː] “beer”

ប្រតីងអ៊ឹះ *ba* + *coeng* + *ro* + *ta* + *yy* + *ngo* + *qa* + [ZWNJ] + *triisap* + *y* + *reahmuk* [prɔtəːŋʔɯh] “urgent, too busy”

ប្រតីងអ៊ឹះ *ba* + *coeng* + *ro* + *ta* + *yy* + *ngo* + *qa* + *triisap* + *y* + *reahmuk*

Ligature Control. In the *askaa muul* font style, some vowel signs ligate with the consonant characters to which they are applied. The font tables should determine whether they form a ligature; ligature use in *muul* fonts does not affect the meaning. However, U+200C ZERO WIDTH NON-JOINER may be inserted before the vowel sign to explicitly suppress such a ligature, as shown in *Figure 16-4* for the word “savant,” pronounced [vituː].

Figure 16-4. Ligation in *Muul* Style in Khmer

ទ្វ	*vo* + *i* + *to* + *uu*	(*aksaa crieng* font)
ទ្វ, ទ្វ	*vo* + *i* + *to* + *uu*	(ligature dependent on the *muul* font)
ទ្វ	*vo* + [ZWNJ] + *i* + *to* + *uu*	([ZWNJ] to prevent the ligature in a *muul* font)
ទ្វ	*vo* + [ZWJ] + *i* + *to* + *uu*	([ZWJ] to request the ligature in a *muul* font)

Spacing. Khmer does not use whitespace between words, although it does use whitespace between clauses and between parts of a name. If word boundary indications are desired—for example, as part of automatic line layout algorithms—the character U+200B ZERO WIDTH SPACE should be used to place invisible marks for such breaks. The ZERO WIDTH SPACE can grow to have a visible width when justified. See *Table 23-2*.

Khmer Symbols: U+19E0–U+19FF

Symbols. Many symbols for punctuation, digits, and numerals for divination lore are encoded as independent entities. Symbols for the lunar calendar are encoded as single characters that cannot be decomposed even if their appearance might seem to be decomposable. U+19E0 KHMER SYMBOL PATHAMASAT represents the first *ashadha* (eighth month) of the lunar calendar. During the type of leap year in the lunar calendar known as *adhikameas*, there is also a second *ashadha*. U+19F0 KHMER SYMBOL TUTEYASAT represents that second *ashadha*. The 15 characters from U+19E1 KHMER SYMBOL MUOY KOET to U+19EF KHMER SYMBOL DAP-PRAM KOET represent the first through the fifteenth lunar waxing days, respectively. The 15 characters from U+19F1 KHMER SYMBOL MUOY ROC through U+19FF KHMER SYMBOL DAP-PRAM ROC represent the first through the fifteenth waning days, respectively. The typographical form of these lunar dates is a top and bottom section of the same size text. The dividing line between the upper and lower halves of the symbol is the vertical center of the line height.

16.5 Tai Le

Tai Le: U+1950–U+197F

The Tai Le script has a history of 700–800 years, during which time several orthographic conventions were used. The modern form of the script was developed in the years following 1954; it rationalized the older system and added a systematic representation of tones with the use of combining diacritics. The new system was revised again in 1988, when spacing tone marks were introduced to replace the combining diacritics. The Unicode encoding of Tai Le handles both the modern form of the script and its more recent revision.

The Tai Le language is also known as Tai Nüa, Dehong Dai, Tai Mau, Tai Kong, and Chinese Shan. *Tai Le* is a transliteration of the indigenous designation, ᥖᥭᥰ ᥛᥣᥴ [tai^{2} lə^{6}] (in older orthography ᥖᥭᥲ ᥛᥣᥴ). The modern Tai Le orthographies are straightforward: initial consonants precede vowels, vowels precede final consonants, and tone marks, if any, follow the entire syllable. There is a one-to-one correspondence between the tone mark letters now used and existing nonspacing marks in the Unicode Standard. The tone mark is the last character in a syllable string in both orthographies. When one of the combining diacritics follows a tall letter ᥣ, ᥤ, ᥫ, ᥭ, ᥯ or ᥰ, it is displayed to the right of the letter, as shown in *Table 16-10*.

Table 16-10. Tai Le Tone Marks

Syllable	New Orthography	Old Orthography
ta	ᥖ	ᥖ
ta^{2}	ᥖᥰ	ᥖᥲ
ta^{3}	ᥖᥱ	ᥖᥳ
ta^{4}	ᥖᥲ	ᥖᥴ
ta^{5}	ᥖ᥵	ᥖ᥵
ta^{6}	ᥖᥴ	ᥖ᥶
ti	ᥖᥤ	ᥖᥤ
ti^{2}	ᥖᥤᥰ	ᥖᥤᥨ
ti^{3}	ᥖᥤᥱ	ᥖᥤᥬ
ti^{4}	ᥖᥤᥲ	ᥖᥤᥭ
ti^{5}	ᥖᥤ᥵	ᥖᥤᥧ
ti^{6}	ᥖᥤᥴ	ᥖᥤ᥶

Digits. In China, European digits (U+0030..U+0039) are mainly used, although Myanmar digits (U+1040..U+1049) are also used with slight glyph variants, as shown in *Table 16-11*.

Table 16-11. Myanmar Digits

Myanmar-Style Glyphs	Tai Le-Style Glyphs
๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙	๐ ๑ ๒ ๓ ๔ ๕ ๖ ๗ ๘ ๙
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9

Punctuation. Both CJK punctuation and Western punctuation are used. Typographically, European digits are about the same height and depth as the tall characters 𝍐 and 𝍑. In some fonts, the baseline for punctuation is the depth of those characters.

16.6 New Tai Lue

New Tai Lue: U+1980–U+19DF

The New Tai Lue script, also known as Xishuangbanna Dai, is used mainly in southern China. The script was developed in the 20th century as an orthographic simplification of the historic Lanna script used to write the Tai Lue language. "Lanna" refers to a region in present-day northern Thailand as well as to a Tai principality that existed in that region from approximately the late thirteenth century to the early 20th century. The Lanna script grew out of the Mon script and was adapted in various forms in the Lanna kingdom and by Tai-speaking communities in surrounding areas that had close contact with the kingdom, including southern China. The Lanna script, also known as the Tai Tham script (see *Section 16.7, Tai Tham*), is still used to write various languages of the Tai family today, including Tai Lue. The approved orthography for this language uses the New Tai Lue script; however, usage of the older orthography based on a variant of Lanna script can still be found.

New Tai Lue differs from Tai Tham in that it regularizes the consonant repertoire, simplifies the writing of consonant clusters and syllable-final consonants, and uses only spacing vowel signs, which appear before or after the consonants they modify, and which are stored in visual order. By contrast, Lanna uses both spacing vowel signs and nonspacing vowel signs, which appear above or below the consonants they modify, and all of which are stored in logical order.

Structure. New Tai Lue is written left to right. Despite its simplification from the Tai Tham (Lanna) script, it retains an important feature of *abugidas*: the consonant letters have the inherent vowel /a/, which is modified to some other vowel by the addition of an explicit vowel letter.

Visual Order. The New Tai Lue script uses visual ordering—a characteristic it shares with the Thai and Lao scripts. This means that the four New Tai Lue vowels that occur visually on the left side of their associated consonant are stored ahead of those consonants in text. This practice differs from the usual pattern for Brahmi-derived scripts, in which all dependent vowels are stored in logical order after their associated consonants, even when they are displayed to the left of those consonants.

Visual order for New Tai Lue vowels results in simpler rendering for the script and follows current accepted practice for data entry. However, it complicates syllable identification and the processes for searching and sorting. Implementers can take advantage of techniques developed for processing Thai script data to address the issues associated with visual order encoding.

The four New Tai Lue vowel letters that occur in visual order ahead of their associated consonants are given the property value Logical_Order_Exception=True in the Unicode Character Database.

Implementers should note that the visual order model for New Tai Lue was formally introduced as of Unicode 8.0. When New Tai Lue was added to the Unicode Standard in Version 4.1, the text model for the script followed the normal Indic practice: all dependent vowels were intended to follow their consonant, regardless of visual placement. However, in practice, the majority of New Tai Lue text data using Unicode characters prior to Unicode 8.0 already uses visual ordering, and many extant New Tai Lue fonts also assume visual ordering. As a result, the model change for New Tai Lue as of Unicode 8.0 should not pose a substantial migration issue for data or fonts. However, implementations may have glitches in some algorithmic behavior until underlying libraries and platform support catch up to the character property changes for New Tai Lue as of Unicode 8.0.

Two-Part Vowels. Some vowels in New Tai Lue are represented with two vowel letters—one to the left of the consonant letter and one to the right. In these cases, the characters are simply stored in visual order: first the vowel letter on the left, then the consonant letter, and finally the vowel letter on the right. U+19B6 NEW TAI LUE VOWEL SIGN AE is considered a single letter and is displayed to the left of its consonant letter. It is not represented by a sequence of two characters for U+19B5 NEW TAI LUE VOWEL SIGN E. If a tone mark appears in a syllable, it occurs last in the representation, after any right side vowel, again in visual order. *Table 16-12* shows several examples of these ordering relations.

Table 16-12. New Tai Lue Vowel Placement

ᦅ	e	+	ᦀ	ka	+	ᦶ	t1				→	ᦅᦀᦶ	[ke:^{2}]
ᦅ	e	+	ᦀ	ka	+	θ	i				→	ᦅᦀθ	[kə:^{1}]
ᦅ	e	+	ᦀ	ka	+	ᦰ	iy				→	ᦅᦀᦰ	[kəi^{1}]
ᦅ	e	+	ᦀ	ka	+	ᦰ	iy	+	ᦶ	t1	→	ᦅᦀᦰᦶ	[kəi^{2}]
ᦅ	e	+	ᦀ	ka	+	ᦰ	iy	+	e	t2	→	ᦅᦀᦰe	[kəi^{3}]

Final Consonants. A virama or killer character is not used to create conjunct consonants in New Tai Lue, because clusters of consonants do not regularly occur. New Tai Lue has a limited set of final consonants, which are modified with a hook showing that the inherent vowel is killed.

Tones. Similar to the Thai and Lao scripts, New Tai Lue consonant letters come in pairs that denote two tonal registers. The tone of a syllable is indicated by the combination of the tonal register of the consonant letter plus a tone mark written at the end of the syllable, as shown in *Table 16-13*.

Digits. The New Tai Lue script adapted its digits from the Tai Tham (or Lanna) script. Tai Tham used two separate sets of digits, one known as the *hora* set, and one known as the *tham* set. The New Tai Lue digits are adapted from the *hora* set.

The one exception is the additional New Tai Lue digit for one: U+19DA ᧚ NEW TAI LUE THAM DIGIT ONE. The regular *hora* form for the digit, U+19D1 ᧑ NEW TAI LUE DIGIT ONE,

Table 16-13. New Tai Lue Registers and Tones

Display	Sequence	Register	Tone Mark	Tone	Transcription
ᦂ	ka^{h}	high		1	[ka^{1}]
ᦂᧆ	ka^{h} + t1	high	t1	2	[ka^{2}]
ᦂᧈ	ka^{h} + t2	high	t2	3	[ka^{3}]
ᦀ	ka^{l}	low		4	[ka^{4}]
ᦀᧆ	ka^{l} + t1	low	t1	5	[ka^{5}]
ᦀᧈ	ka^{l} + t2	low	t2	6	[ka^{6}]

has the exact same glyph shape as a common New Tai Lue vowel, U+19B1 ᦲ NEW TAI LUE VOWEL SIGN AA. For this reason, U+19DA is often substituted for U+19D1 in contexts which are not obviously numeric, to avoid visual ambiguity. Implementations of New Tai Lue digits need to be aware of this usage, as U+19DA may occur frequently in text.

16.7 Tai Tham

Tai Tham: U+1A20–U+1AAF

The script called Tai Tham is used for three living languages, Lue, Khuen, and Northern Thai, which are spoken in China, Myanmar, Northern Thailand, and surrounding areas. In addition, the script is used for Lao Tham (or Old Lao) and other dialect variants found in Buddhist palm leaves and notebooks. Although the script has no single, commonly recognized name across the region today, it is known by various language-specific and region-specific names, such as Old Xishuangbanna Dai or Old Tai Lue in China, Khün in Myanmar, and Tua Mueang, Lanna, or Yuan in Thailand.

Few of the six million speakers of Northern Thai are literate in the Tai Tham script, although there is some rising interest in the script among the young. There are about 690,000 speakers of Tai Lue. Of those, many people born before 1950 are literate in the Tai Tham script, and newspapers and other literature are regularly produced in the Xishuangbanna region of Yunnan using the script. Younger speakers are taught the New Tai Lue script, instead. (See *Section 16.6, New Tai Lue.*) The Tai Tham script continues to be taught in the Tai Lue monasteries. There are 107,000 speakers of Khün, for which Tai Tham is the only script.

Consonants. Consonants have an inherent *-a* vowel sound. Most consonants have a combining subjoined form, but unlike most other Brahmi-derived scripts, the subjoining of a consonant does not mean that the vowel of the previous consonant is killed. A subjoined consonant may be the first consonant of the following syllable. The encoding model for Tai Tham is more similar to the Khmer *coeng* model than to the usual virama model: the character U+1A60 TAI THAM SIGN SAKOT is entered before a consonant which is to take the subjoined form. A subjoined consonant may be attached to a dependent vowel sign.

U+1A4B TAI THAM LETTER A represents a glottal consonant. Its rendering in Northern Thai differs from that typical for Tai Lue and Khün.

A number of Tai Tham characters did not traditionally take subjoined forms, but modern innovations in borrowed vocabulary suggest that fonts should make provision for subjoining behavior for all of the consonants except the historical *vocalic r* and *l*.

Independent Vowels. Independent vowels are used as in other Brahmi-derived scripts. U+1A52 TAI THAM LETTER OO is not used in Northern Thai.

Dependent Consonant Signs. Seven dependent consonant signs occur. Two of these are used as medials: U+1A55 TAI THAM CONSONANT SIGN MEDIAL RA and U+1A56 TAI THAM CONSONANT SIGN MEDIAL LA form clusters and immediately follow a consonant.

U+1A58 TAI THAM SIGN MAI KANG LAI is used as a final *-ng* in Northern Thai and Tai Lue. Its shape is distinct in Khün. U+1A59 TAI THAM CONSONANT SIGN FINAL NGA is also used as a final *-ng* in Northern Thai.

U+1A5B TAI THAM CONSONANT SIGN HIGH RATHA OR LOW PA represents *high ratha* in *santhān* "shape" and *low pa* in *sappa* "omniscience".

Dependent Vowel Signs. Dependent vowel signs are used in a manner similar to that employed by other Brahmi-derived scripts, although Tai Tham uses many of them in combination.

U+1A63 TAI THAM VOWEL SIGN AA and U+1A64 TAI THAM VOWEL SIGN TALL AA are separately encoded because the choice of which form to use cannot be reliably predicted from context.

The Khün character U+1A6D TAI THAM VOWEL SIGN OY is not used in Northern Thai. Khün vowel order is quite different from that of Northern Thai.

Tone Marks. Tai Tham has two combining tone marks, U+1A75 TAI THAM SIGN TONE-1 and U+1A76 TAI THAM SIGN TONE-2, which are used in Tai Lue and in Northern Thai. These are rendered above the vowel over the base consonant. Three additional tone marks are used in Khün: U+1A77 TAI THAM SIGN KHUEN TONE-3, U+1A78 TAI THAM SIGN KHUEN TONE-4, and U+1A79 TAI THAM SIGN KHUEN TONE-5, which are rendered above and to the right of the vowel over the base consonant. Tone marks are represented in logical order following the vowel over the base consonant or consonant stack. If there is no vowel over a base consonant, then the tone is rendered directly over the consonant; this is the same way tones are treated in the Thai script.

Other Combining Marks. U+1A7A TAI THAM SIGN RA HAAM is used in Northern Thai to indicate that the character or characters it follows are not sounded. The precise range of characters not to be sounded is indeterminant; it is defined instead by reading rules. In Tai Lue, *ra haam* is used as a final *-n*.

The mark U+1A7B TAI THAM SIGN MAI SAM has a range of uses in Northern Thai:

- It is used as a repetition mark, stored as the last character in the word to be repeated: *tang* "be different", *tangtang* "be different in my view".
- It is used to disambiguate the use of a subjoined letters. A subjoined letter may be a medial or final, or it may be the start of a new syllable.
- It is used to mark "double-acting" consonants. It is stored where the consonant would be stored if there were a separate consonant used.

U+1A7F TAI THAM COMBINING CRYPTOGRAMMIC DOT is used singly or multiply beneath letters to give each letter a different value according to some hidden agreement between reader and writer.

Digits. Two sets of digits are in common use: a secular set (Hora) and an ecclesiastical set (Tham). European digits are also found in books.

Punctuation. The four signs U+1AA8 TAI THAM SIGN KAAN, U+1AA9 TAI THAM SIGN KAANKUU, U+1AAA TAI THAM SIGN SATKAAN, and U+1AAB TAI THAM SIGN SATKAANKUU, are used in a variety of ways, with progressive values of finality. U+1AAB TAI THAM SIGN SATKAANKUU is similar to U+0E5A THAI CHARACTER ANGKHANKHU.

At the end of a section, U+1AA9 TAI THAM SIGN KAANKUU and U+1AAC TAI THAM SIGN HANG may be combined with U+1AA6 TAI THAM SIGN REVERSED ROTATED RANA in a number of ways. The symbols U+1AA1 TAI THAM SIGN WIANGWAAK, U+1AA0 TAI THAM SIGN WIANG, and U+1AA2 TAI THAM SIGN SAWAN are logographs for "village," "city," and "heaven," respectively.

The three signs U+1AA3 TAI THAM SIGN KEOW, "courtyard," U+1AA4 TAI THAM SIGN HOY, "oyster," and U+1AA5 TAI THAM SIGN DOKMAI, "flower" are used as dingbats and as section starters. The mark U+1AA7 TAI THAM SIGN MAI YAMOK is used in the same way as its Thai counterpart, U+0E46 THAI CHARACTER MAIYAMOK.

European punctuation like question mark, exclamation mark, parentheses, and quotation marks is also used.

Collating Order. There is no firmly established sorting order for the Tai Tham script. The order in the code charts is based on Northern Thai and Thai. U+1A60 TAI THAM SIGN SAKOT is ignored for sorting purposes.

Linebreaking. Opportunities for linebreaking are lexical, but a linebreak may not be inserted between a base letter and a combining diacritic. There is no line-breaking hyphenation.

16.8 Tai Viet

Tai Viet: U+AA80–U+AADF

The Tai Viet script is used by three Tai languages spoken primarily in northwestern Vietnam, northern Laos, and central Thailand: Tai Dam (also Black Tai or Tai Noir), Tai Dón (White Tai or Tai Blanc), and Thai Song (Lao Song or Lao Song Dam). The Thai Song of Thailand are geographically removed from, but linguistically related to the Tai people of Vietnam and Laos. There are also populations in Australia, China, France, and the United States. The script is related to other Tai scripts used throughout Southeast Asia. The total population using the three languages, across all countries, is estimated to be 1.3 million (Tai Dam 764,000, Tai Dón 490,000, Thai Song 32,000). The script is still used by the Tai people in Vietnam, and there is a desire to introduce it into formal education there. It is unknown whether it is in current use in Laos, Thailand, or China.

Several different spellings have been employed for the name of the script, including Tay Viet. Linguists commonly use "Thai" to indicate the language of central Thailand, and "Tai" to indicate the language family; however, even that usage is inconsistent.

Structure. The Tai Viet script shares many features with other Tai alphabets. It is written left to right and has a double set of initial consonants, one for the low tone class and one for the high tone class. Vowels marks are positioned before, after, above, or below the syllable's initial consonant, depending on the vowel. Some vowels are written with digraphs. The consonants do not carry an implicit vowel. The vowel must always be written explicitly.

The Tai languages are almost exclusively monosyllabic. A very small number of words have an unstressed initial syllable, and loan words may be polysyllabic.

Visual Order. The Tai Viet script uses visual ordering—a characteristic it shares with the Thai and Lao scripts. This means that the five Tai Viet vowels that occur visually on the left side of their associated consonant are stored ahead of those consonants in text. This practice differs from the usual pattern for Brahmi-derived scripts, in which all dependent vowels are stored in logical order after their associated consonants, even when they are displayed to the left of those consonants.

Visual order for Tai Viet vowels results in simpler rendering for the script and follows accepted practice for data entry. However, it complicates syllable identification and the processes for searching and sorting. Implementers can take advantage of techniques developed for processing Thai script data to address the issues associated with visual order encoding.

The five Tai Viet vowels that occur in visual order ahead of their associated consonants are given the property value Logical_Order_Exception=True in the Unicode Character Database.

Tone Classes and Tone Marks. In the Tai Viet script each consonant has two forms. The low form of the initial consonant indicates that the syllable uses tone 1, 2, or 3. The high form of the initial consonant indicates that the syllable uses tone 4, 5, or 6. This is sufficient

to define the tone of closed syllables (those ending /p/, /t/, /k/, or /ʔ/), in that these syllables are restricted to tones 2 and 5.

Traditionally, the Tai Viet script did not use any further marking for tone. The reader had to determine the tone of unchecked syllables from the context. Recently, several groups have introduced tone marks into Tai Viet writing. Tai Dam speakers in the United States began using Lao tone marks with their script about thirty years ago, and those marks are included in SIL's Tai Heritage font. These symbols are written as combining marks above the initial consonant, or above a combining vowel, and are identified by their Laotian names, *mai ek* and *mai tho*. These marks are also used by the Song Petburi font (developed for the Thai Song language), although they were probably borrowed from the Thai alphabet rather than the Lao.

The Tai community in Vietnam invented their own tone marks written on the base line at the end of the syllable, which they call *mai nueng* and *mai song*.

When combined with the consonant class, two tone marks are sufficient to unambiguously mark the tone. No tone is written on loan words or on the unstressed initial syllable of a native word.

Final Consonants. U+AA9A TAI VIET LETTER LOW BO and U+AA92 TAI VIET LETTER LOW DO are used to write syllable-final /p/ and /t/, respectively, as is the practice in many Tai scripts. U+AA80 TAI VIET LETTER LOW KO is used for both final /k/ and final /ʔ/. The high-tone class symbols are used for writing final /j/ and the final nasals, /m/, /n/, and /ŋ/. U+AAAB TAI VIET LETTER HIGH VO is used for final /w/.

There are a number of exceptions to the above rules in the form of vowels which carry an inherent final consonant. These vary from region to region. The ones included in the Tai Viet block are the ones with the broadest usage: /-aj/, /-am/, /-an/, and /-əw/.

Symbols and Punctuation. There are five special symbols in Tai Viet. The meaning and use of these symbols is summarized in *Table 16-14*.

Table 16-14. Tai Viet Symbols and Punctuation

Code	Glyph	Name	Meaning
AADB	ꪀ꪿	*kon*	person
AADC	ꪀꪾ	*nueng*	one
AADD	ꪹ	*sam*	signals repetition of the previous word
AADE	ꪻ	*ho hoi*	beginning of text (used in songs and poems)
AADF	ꪼ	*koi koi*	end of text (used in songs and poems)

U+AADB TAI VIET SYMBOL KON and U+AADC TAI VIET SYMBOL NUENG may be regarded as word ligatures. They are, however, encoded as atomic symbols, without decompositions. In the case of *kon*, the word ligature symbol is used to distinguish the common word "person" from otherwise homophonous words.

Word Spacing. Traditionally, the Tai Viet script was written without spaces between words. In the last thirty years, users in both Vietnam and the United States have started writing spaces between words, in both handwritten and machine produced texts. Most users now use interword spacing. Polysyllabic words may be written without space between the syllables.

Collating Order. The Tai Viet script does not have an established standard for sorting. Sequences have sometimes been borrowed from neighboring languages. Some sources use the Lao order, adjusted for differences between the Tai Dam and Lao character repertoires. Other sources prefer an order based on the Vietnamese alphabet. It is possible that communities in different countries will want to use different orders.

16.9 Kayah Li

Kayah Li: U+A900–U+A92F

The Kayah Li script was invented in 1962 by Htae Bu Phae (also written Hteh Bu Phe), and is used to write the Eastern and Western Kayah Li languages of Myanmar and Thailand. The Kayah Li languages are members of the Karenic branch of the Sino-Tibetan family, and are tonal and mostly monosyllabic. There is no mutual intelligibility with other Karenic languages.

The term *Kayah Li* is an ethnonym referring to a particular Karen people who speak these languages. *Kayah* means "person" and *li* means "red," so *Kayah Li* literally means "red Karen." This use of color terms in ethnonyms and names for languages is a common pattern in this part of Southeast Asia.

Structure. Although Kayah Li is a relatively recently invented script, its structure was clearly influenced by Brahmi-derived scripts, and in particular the Myanmar script, which is used to write other Karenic languages. The order of letters is a variant of the general Brahmic pattern, and the shapes and names of some letters are Brahmi-derived. Other letters are innovations or relate more specifically to Myanmar-based orthographies.

The Kayah Li script resembles an abugida such as the Myanmar script, in terms of the derivation of some vowel forms, but otherwise Kayah Li is closer to a true alphabet. Its consonants have no inherent vowel, and thus no virama is needed to remove an inherent vowel.

Vowels. Four of the Kayah Li vowels (a, o', i, ô) are written as independent spacing letters. Five others (u', e, u, ê, o) are written by means of diacritics applied above the base letter U+A922 KAYAH LI LETTER A, which thus serves as a vowel-carrier. The same vowel diacritics are also written above the base letter U+A923 KAYAH LI LETTER OE to represent sounds found in loanwords.

Tones. Tone marks are indicated by combining marks which subjoin to the four independent vowel letters. The vowel diacritic U+A92A KAYAH LI VOWEL O and the mid-tone mark, U+A92D KAYAH LI TONE CALYA PLOPHU, are each analyzable as composite signs, but encoding of each as a single character in the standard reflects usage in didactic materials produced by the Kayah Li user community.

Digits. The Kayah Li script has its own set of distinctive digits.

Punctuation. Kayah Li text makes use of modern Western punctuation conventions, but the script also has two unique punctuation marks: U+A92E KAYAH LI SIGN CWI and U+A92F KAYAH LI SIGN SHYA. The *shya* is a script-specific form of a *danda* mark.

16.10 Cham

Cham: U+AA00–U+AA5F

Cham is a Austronesian language of the Malayo-Polynesian family. The Cham language has two major dialects: Eastern Cham and Western Cham. Eastern Cham speakers live primarily in the southern part of Vietnam and number about 73,000. Western Cham is spoken mostly in Cambodia, with about 220,000 speakers there and about 25,000 in Vietnam. The Cham script is used more by the Eastern Cham community.

Structure. Cham is a Brahmi-derived script. Consonants have an inherent vowel. The inherent vowel is *-a* in the case of most consonants, but is *-ư* in the case of nasal consonants. There is no virama and hence no killing of the inherent vowel. Dependent vowels (matras) are used to modify the inherent vowel and separately encoded, explicit final consonants are used where there is no inherent vowel. The script does not have productive formation of consonant conjuncts.

Independent Vowel Letters. Six of the initial vowels in Cham are represented with unique, independent vowels. These separately-encoded characters always indicate a syllable-initial vowel, but they may occur word-internally at a syllable break. Other Cham vowels which do not have independent forms are instead represented by dependent vowels (matras) applied to U+AA00 CHAM LETTER A. Four of the other independent vowel letters are also attested bearing matras.

Consonants. Cham consonants can be followed by consonant signs to represent the glides: *-ya*, *-ra*, *-la*, or *-wa*. U+AA33 CHAM CONSONANT SIGN YA, in particular, normally ligates with the base consonant it modifies. When it does so, any dependent vowel is graphically applied to it, rather than to the base consonant.

The independent vowel U+AA00 CHAM LETTER A can cooccur with two of the medial consonant signs: *-ya* or *-wa*. The writing system distinguishes these sequences from single letters which are pronounced the same. Thus, <a, -ya> [ja] contrasts with U+AA22 CHAM LETTER YA, also pronounced [ja], and <a, -wa> [wa] contrasts with U+AA25 CHAM LETTER VA, also pronounced [wa].

Three medial clusters of two consonant signs in a row occur: <-ra, -wa> [-rwa], <-la, -ya> [-lja], and <-la, -wa> [-lwa].

There are three types of final consonants. The majority are simply encoded as separate base characters. Graphically, those final forms appear similar to the corresponding non-final consonants, but typically have a lengthened stroke at the right side of their glyphs. The second type consist of combining marks to represent final *-ng*, *-m*, and *-h*. Finally, U+AA25 CHAM LETTER VA occurs unchanged either in initial or final positions. Final consonants may occur word-internally, in which case they indicate the presence of a syllable boundary.

Ordering of Syllable Components. Dependent vowels and other signs are encoded after the consonant to which they apply. The ordering of elements is shown in more detail in *Table 16-15*.

Table 16-15. Cham Syllabic Structure

Class	Examples	Encoding
consonant or independent vowel	ꨀ	[U+AA00..U+AA28]
consonant sign -ra, -la	◌ꨩ,◌ꨪ	[U+AA34, U+AA35]
consonant sign -ya, -wa	◌ꨫ,◌ꨬ	[U+AA33, U+AA36]
left-side dependent vowel	◌ꨯ,◌ꨰ	[U+AA2F, U+AA30]
other dependent vowel	◌ꨱ	[U+AA2A..U+AA2E, U+AA31..U+AA32]
vowel lengthener -aa	◌ꨳ	U+AA29
final consonant or va	ꨤ,꨷	[U+AA40..U+AA4D, U+AA25]

The left-side dependent vowels U+AA2F CHAM VOWEL SIGN O and U+AA30 CHAM VOWEL SIGN AI occur in logical order after the consonant (and any medial consonant signs), but in visual presentation their glyphs appear *before* (to the left of) the consonant. U+AA2F CHAM VOWEL SIGN O, in particular, may occur together in a sequence with another dependent vowel, the vowel lengthener, or both. In such cases, the glyph for U+AA2F appears to the left of the consonant, but the glyphs for the second dependent vowel and the vowel lengthener are rendered above or to the right of the consonant.

Digits. The Cham script has its own set of digits, which are encoded in this block. However, European digits are also known and occur in Cham texts because of the influence of Vietnamese.

Punctuation. Cham uses *danda* marks to indicate text units. Three levels are recognized, marked respectively with *danda*, *double danda*, and *triple danda*.

U+AA5C CHAM PUNCTUATION SPIRAL often begins a section of text. It can be compared to the usage of Tibetan head marks. The *spiral* may also occur in combination with a *danda*.

Modern Cham text also makes use of European punctuation marks, such as the question mark, hyphen and colon.

Line Breaking. Opportunities for line breaks occur after any full orthographic syllable in Cham. Modern Cham text makes use of spaces between words, and those are also line break opportunities. Line breaks occur after *dandas*.

16.11 Pahawh Hmong

Pahawh Hmong: U+16B00–U+16B8F

The Pahawh Hmong script was originally devised by Shong Lue Yang in 1959 to write the Hmong language. The script was devised in Laos, and taken to refugee camps in northern Thailand. In the late 20th century, it then moved with waves of immigrants to Australia and the United States, where it remains in current use. The Hmong language is also commonly written using the Romanized Popular Alphabet (RPA), which uses the Latin script.

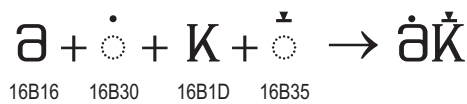
The Pahawh Hmong writing system has had four stages of development: the Source version, the Second Stage Reduced Version, the Third Stage Reduced Version, and the Final Version. Only the Second and Third Stage versions are in current use. The characters in the Pahawh Hmong block support text written in the Second Stage Reduced, Third Stage Reduced, and Final versions.

Character Names. The Pahawh Hmong character names are based on the Third Stage Reduced Version, which provides a one-to-one mapping between the Third Stage tone diacritics and the widely used Romanized Popular Alphabet tone mark letters. The Second Stage names are listed as annotations in the names list.

Structure. The Pahawh Hmong script is written left to right. Text consists of a sequence of syllables that may contain a maximum length of four characters (two base characters and two diacritics). Syllables are separated by spaces.

Unlike other writing systems, Pahawh Hmong writes the vowel of a syllable before the syllable-initial consonant, as illustrated in *Figure 16-5*.

Figure 16-5. Pahawh Hmong Syllable Structure



The example in *Figure 16-5* uses Second Stage Reduced Version conventions. The representation of the syllable is in straightforward visual order. U+16B16 PAHAWH HMONG VOWEL KAB is the base character representing the [a] vowel of the syllable. The combining mark U+16B30 represents the tone mark for the vowel. U+16B1D PAHAWH HMONG CONSONANT NTSAU is the base character representing the initial consonant of the syllable. The combining mark U+16B35 is a diacritical mark which changes the sound of the consonant from [^{n}ts] to [p^{h}]. Altogether, the sequence represents the syllable [p^{h}â].

Because the order of characters in memory matches the visual written order of the text, display rendering does not require any reordering of glyphs. However, implementations such as text-to-speech need to be aware that Pahawh Hmong has unusual reading rules, because initial consonants for syllables graphically follow the vowels which they precede in pronunciation.

Vowels. The characters in the range U+16B00..U+16B1B represent vowels. The addition of a diacritic alters the tone of the vowel. The special characters U+16B1A PAHAWH HMONG VOWEL KAAB and U+16B1B PAHAWH HMONG VOWEL KAAV are atomic characters and do not decompose.

Consonants. U+16B1C..U+16B2F represent consonants. These are phonologically initial in a syllable, but occur after the vowel in written order. .

Combining Marks. The combining marks in the range U+16B30..U+16B36 are used as tone marks. They combine with the vowel letters to indicate particular tones for the syllable. The use for representation of particular tones differs for the two different stages.

U+16B30 PAHAWH HMONG MARK CIM TUB and U+16B35 PAHAWH HMONG MARK CIM HOM also combine with initial consonant letters. When used this way, these marks function as diacritics and indicate a different sound for the consonant letter. Usually the resultant sound is unrelated to that of the unmodified base letter—the particular modification by the diacritic is not predictable.

Punctuation and Other Symbols. Pahawh Hmong makes use of common European punctuation marks as well as script-specific punctuation marks (U+16B37..U+16B3B and U+16B44..U+16B45). The script employs script-specific mathematical operators (U+16B3C..U+16B3F). It also includes a set of modifiers that have various uses: U+16B42..U+16B43 indicate reduplication, U+16B40 identifies the chanting nature of a text, and U+16B41 indicates the following syllable has a non-Hmong pronunciation.

Digits and Numbers. The decimal digits 0-9 are encoded from U+16B50..U+16B59. The representative glyph for U+16B50 PAHAWH HMONG DIGIT ZERO resembles an "I", and is found in the Second Stage Reduced Version orthography. In contrast, the Third Stage Reduced Version orthography has a circular glyph.

A non-decimal system also exists in Pahawh Hmong and is taught today, however, it is not used for arithmetic calculation. The non-decimal numbers are encoded in the range from U+16B5B..U+16B61. The Second Stage Reduced Version glyph for U+16B5B PAHAWH HMONG NUMBER TENS resembles a "W". The Third Stage Reduced Version glyph looks like an "I", and should be distinguished in fonts from U+16B50 PAHAWH HMONG DIGIT ZERO.

Logographs. Characters encoded from U+16B63..U+16B8F are logographs. These include a grammatical classifier (U+16B63). Also included are characters designating periods of time (U+16B64..U+16B6C), correspondence (U+16B6D..U+16B77), and clan names (U+16B7E..U+16B8F). The clan names are encoded for historical reasons, and are not in widespread current use.

16.12 Pau Cin Hau

Pau Cin Hau: U+11AC0–U+11AFF

The Pau Cin Hau alphabet is a liturgical script of the Laipian religious tradition, which emerged in the Chin Hills region of present-day Chin State, Myanmar at the turn of the 20th century. The script is named after Pau Cin Hau (1859–1948), a Tedim Chin, who founded the Laipian tradition and developed the script to convey his teachings. In an account given by J. J. Bennison in the 1931 Census of India report for Burma, Pau Cin Hau stated that the characters of his script were revealed to him in a dream in 1902.

The script was designed to represent Tedim, a language of the northern branch of the Kuki-Chin group of the Tibeto-Burman family, which is spoken in Chin State. Tedim is the modern name for the language previously known as Tiddim; it also refers to the Tedim dialects Kamhau (Kamhow) and Sokte.

While the script was developed for writing Tedim, several letters and tone marks represent sounds that are not attested in Tedim, but which exist in other Chin languages, suggesting that the alphabet may have been created as a universal script for the Chin languages.

There are two distinct writing systems associated with Pau Cin Hau and the Laipian tradition. One is an obsolete syllabary and the other is the alphabetic system encoded in this block. Both are attested in manuscript and printed sources. The alphabetic script is dreived from the syllabary. Neither of these scripts has any genetic relationship with any other writing system.

Structure. The Pau Cin Hau alphabet has 57 characters consisting of 21 consonant letters, 7 vowel letters, 9 final-consonant letters, and 20 tone marks. It is written from left to right. Vowels, consonants, and tone marks are written linearly as independent characters. The syllable canon for Tedim may be described as (C1)V1(V2)(C2)T. The tone (T) is represented using one of the 20 tone marks. These marks are used for indicating vowel length, tone, and glottal stop, as well as punctuation. Of these, 15 represent tones and 5 represent glottal stop. Ten of the tone marks have a dual role and simultaneously denote tone (or glottal stop) and sentence ending.

Digits. Pau Cin Hau uses European digits.

Punctuation. Word boundaries are indicated using spaces. The end of a sentence is marked with final forms of tone marks. Western punctuation is also used. In some cases, sentence-final tone marks may be redundantly followed by a full stop or other Western punctuation mark.

Line breaking should occur at spaces. Words are not broken at end-of-line and no hyphen is used or attested. No breaking may occur between a tone mark and the character that precedes it.