

# The Unicode Standard

## Version 6.2 – Core Specification

To learn about the latest version of the Unicode Standard, see <http://www.unicode.org/versions/latest/>.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

Unicode and the Unicode Logo are registered trademarks of Unicode, Inc., in the United States and other countries.

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

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

Copyright © 1991–2012 Unicode, Inc.

All rights reserved. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction. For information regarding permissions, inquire at <http://www.unicode.org/reporting.html>. For information about the Unicode terms of use, please see <http://www.unicode.org/copyright.html>.

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 6.2.

Includes bibliographical references and index.

ISBN 978-1-936213-07-8 (<http://www.unicode.org/versions/Unicode6.2.0/>)

1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.

QA268.U545 2012

ISBN 978-1-936213-07-8

Published in Mountain View, CA

September 2012

## Chapter 11

# *Southeast Asian Scripts*

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

<i>Thai</i>	<i>Tai Tham</i>	<i>Balinese</i>
<i>Lao</i>	<i>Tai Viet</i>	<i>Javanese</i>
<i>Myanmar</i>	<i>Kayah Li</i>	<i>Rejang</i>
<i>Khmer</i>	<i>Cham</i>	<i>Batak</i>
<i>Tai Le</i>	<i>Philippine scripts</i>	<i>Sundanese</i>
<i>New Tai Lue</i>	<i>Buginese</i>	

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.

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.

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 Xishuang Banna 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 northwest-

ern 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.

There are four traditional Philippine scripts: Tagalog, Hanunóo, Buhid, and Tagbanwa. They have limited current use. They are discussed together, because each is structured quite similarly. Each is a very simplified abugida which makes use of two nonspacing vowel signs.

Although the official language of Indonesia, Bahasa Indonesia, is written in the Latin script, Indonesia has many local, traditional scripts, most of which are ultimately derived from Brahmi. Five of these scripts are documented in this chapter. Buginese is used for several different languages on the island of Sulawesi. Balinese and Javanese are closely related, highly ornate scripts; Balinese is used for the Balinese language on the island of Bali, and Javanese for the Javanese language on the island of Java. Sundanese is used to write the Sundanese language on the island of Java. The Rejang script is used to write the Rejang language in southwest Sumatra, and the Batak script is used to write several Batak dialects, also on the island of Sumatra.

---

## 11.1 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 11-1*.

Table 11-1. Glyph Positions in Thai Syllables

Syllable	Glyphs	Code Point Sequence
<i>ka</i>	กะ	0E01 0E30
<i>ka:</i>	กา	0E01 0E32
<i>ki</i>	กิ	0E01 0E34
<i>ki:</i>	กี	0E01 0E35
<i>ku</i>	กุ	0E01 0E38
<i>ku:</i>	กู	0E01 0E39
<i>ku'</i>	กิ	0E01 0E36
<i>ku':</i>	กี	0E01 0E37
<i>ke</i>	กะ	0E40 0E01 0E30
<i>ke:</i>	เก	0E40 0E01
<i>kae</i>	กะ	0E41 0E01 0E30
<i>kae:</i>	แก	0E41 0E01
<i>ko</i>	โกะ	0E42 0E01 0E30
<i>ko:</i>	โก	0E42 0E01
<i>ko'</i>	กะ	0E40 0E01 0E32 0E30
<i>ko':</i>	ก	0E01 0E2D
<i>koe</i>	กะ	0E40 0E01 0E2D 0E30
<i>koe:</i>	ก	0E40 0E01 0E2D
<i>kia</i>	กีย	0E40 0E01 0E35 0E22
<i>ku'a</i>	กือ	0E40 0E01 0E37 0E2D
<i>kua</i>	กัว	0E01 0E31 0E27
<i>kaw</i>	กา	0E40 0E01 0E32
<i>koe:y</i>	เกย	0E40 0E01 0E22
<i>kay</i>	ไก	0E44 0E01
<i>kay</i>	ไ	0E43 0E01
<i>kam</i>	กำ	0E01 0E33
<i>kri</i>	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 16-2*.

**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.

---

## 11.2 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 11-2*.

**Table 11-2.** Glyph Positions in Lao Syllables

Syllable	Glyphs	Code Point Sequence
<i>ka</i>	ກະ	0E81 0EB0
<i>ka:</i>	ກາ	0E81 0EB2
<i>ki</i>	ກີ	0E81 0EB4
<i>ki:</i>	ກີ	0E81 0EB5
<i>ku</i>	ກຸ	0E81 0EB8
<i>ku:</i>	ກູ	0E81 0EB9
<i>ku'</i>	ກີ້	0E81 0EB6
<i>ku':</i>	ກີ້	0E81 0EB7
<i>ke</i>	ເກະ	0EC0 0E81 0EB0
<i>ke:</i>	ເກ	0EC0 0E81
<i>kae</i>	ແກະ	0EC1 0E81 0EB0
<i>kae:</i>	ແກ	0EC1 0E81
<i>ko</i>	ໂກະ	0EC2 0E81 0EB0
<i>ko:</i>	ໂກ	0EC2 0E81
<i>ko'</i>	ເກາະ	0EC0 0E81 0EB2 0EB0
<i>ko':</i>	ກໍ	0E81 0ECD
<i>koe</i>	ເກີ	0EC0 0E81 0EB4
<i>koe:</i>	ເກີ	0EC0 0E81 0EB5
<i>kia</i>	ເກີ້ຮ ເກຢ	0EC0 0E81 0EB1 0EBD 0EC0 0E81 0EA2
<i>ku'a</i>	ເກີ້ອ	0EC0 0E81 0EB7 0EAD
<i>kua</i>	ກົວ	0E81 0EBB 0EA7
<i>kaw</i>	ເກົາ	0EC0 0E81 0EBB 0EB2
<i>koe:y</i>	ເກີ້ຮ ເກຢ	0EC0 0E81 0EB5 0EBD 0EC0 0E81 0EB5 0EA2
<i>kay</i>	ໄກ	0EC4 0E81
<i>kay</i>	ໄກ	0EC3 0E81
<i>kam</i>	ກໍ່າ	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.

## 11.3 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 ◯ MYANMAR LETTER WA and U+1040 ◯ 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 ◯ ka + U+1031 ◯ vowel sign e + U+102C ◯ vowel sign aa →  
 ◯◯◯ /kàw/

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

U+1000 ◯ ka + U+102D ◯ vowel sign i + U+102F ◯ vowel sign u → ◯◯  
 /kol

**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 ◯ 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 ◯ MYANMAR LETTER KA would be written via the following sequence:

U+1004 ◯ nga + U+103A ◯ asat + U+1039 ◯ virama + U+1000 ◯ ka  
 → ◯<sup>ṅ</sup>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 [kjwei] (“to drop off”) would be written via the following sequence:

U+1000 *ka* + U+103C *medial ra* + U+103D *medial wa* +  
U+1031 *vowel sign e* → /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 *ya* + U+1031 *vowel sign e* + U+102C *vowel sign aa* +  
U+1000 *ka* + U+103A *asat* + U+103B *medial ya* + U+102C *vowel sign aa* + U+1038 *visarga* → *man, husband*

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

**Great sa.** The *great sa* is encoded as U+103F 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*, *sa sa*, and the sequence <U+101E, U+103A, U+101E> should be used for the form with an *asat sign*, *sa sa*.

**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 11-3*.

Table 11-3. Myanmar Syllabic Structure

Class	Example	Encoding
<i>kinzi</i>	ꨀ	<U+1004, U+103A, U+1039>
<i>consonant and vowel letters</i>	ꨁ	[U+1000..U+102A, U+103F, U+104E]
<i>asat sign (for contractions)</i>	ꨂ	U+103A
<i>subscript consonant</i>	ꨃ	<U+1039, [U+1000..U+1019, U+101C, U+101E, U+1020, U+1021]>
<i>medial ya</i>	ꨄ	U+103B
<i>medial ra</i>	ꨅ	U+103C
<i>medial wa</i>	ꨆ	U+103D
<i>medial ha</i>	ꨇ	U+103E
<i>vowel sign e</i>	ꨈ	U+1031
<i>vowel sign i, ii, ai</i>	ꨉ, ꨊ, ꨋ	[U+102D, U+102E, U+1032]
<i>vowel sign u, uu</i>	ꨌ, ꨍ	[U+102F, U+1030]
<i>vowel sign tall aa, aa</i>	ꨎ, ꨏ	[U+102B, U+102C]
<i>anusvara</i>	ꨐ	U+1036
<i>asat sign</i>	ꨑ	U+103A
<i>dot below</i>	ꨒ	U+1037
<i>visarga</i>	ꨓ	U+1038

U+1031 ꨈ 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 11-3 nominally refers to the character sequences used in representing the syllabic structure of the Burmese language proper. It would require further extensions and modifications to cover the various other languages, such as Karen, Mon, and Shan, which also use the Myanmar script.

**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 justified. 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. 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 11-4*.

**Table 11-4.** 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	<i>unmarked</i>
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 MAIMAYOK.

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+102E, 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.

## 11.4 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 9.1, 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 11-5*, 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.

Table 11-5. Independent Khmer Vowel Characters

Name	Independent Vowel	Qa with Vowel Sign
<i>i</i>	ឺ	ឺ, ឺ, ឺ
<i>ii</i>	ឺ	ឺ, ឺ
<i>u</i>	្ហ	្ហ, ្ហ
<i>uk</i>	្ហ	្ហ
<i>uu</i>	្ហ	្ហ, ្ហ
<i>uuv</i>	្ហ	្ហ
<i>ry</i>	្ហ	្ហ
<i>ryy</i>	្ហ	្ហ

Table 11-5. Independent Khmer Vowel Characters (Continued)

Name	Independent Vowel	Qa with Vowel Sign
<i>ly</i>	ឺ	ឺ
<i>lyy</i>	ឺ	ឺ
<i>e</i>	ឺ	េ, ែ
<i>ai</i>	ឺ	ៃ
<i>oo</i>	ឺ, ឺ	ោ
<i>au</i>	ឺ	ោ

**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.

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:

លូ *lo* + *coeng* + *ngo* [lɲɔː] “sesame” (compare លង *lo* + *ngo* [lɔːŋ] “to haunt”)

លក្សី *lo* + *ka* + *coeng* + *sa* + *coeng* + *mo* + *ii* [ləksmei] “beauty, luck”

កាហ្វេ *ka* + *aa* + *ha* + *coeng* + *vo* + *e* [ka:feː] “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:

ទាំង *to* + *aa* + *nikahit* + *coeng* + *ngo* [tɛəŋ] “both” (= ទាំង) (≠ \*ទាំង [tɲəəm])

ហើយ *ha* + *oe* + *coeng* + *yo* [haəi] “already” (= ហើយ) (≠ \*ហើយ [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:

ឡើង *qoo-1* + *coeng* + *yo* [ʔaoi] “to give” (= ឡើង and also ឡើង)

ឡើង *qoo-1* + *coeng* + *mo* [ʔaom] “exclamation of solemn affirmation” (= ឡើង)

**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:

ផ្អែម *pha + coeng + qe + mo* [p<sup>h</sup>ʔaem] “sweet” (= ផ្អែម *pha + coeng + qa + ae + mo*)

ហ្វូង *ha + coeng + ry + to + samyok sannya + yo* [harutey] “heart”  
(*royal*) (= ហ្វូង *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 11-6. 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).

Table 11-6. Two Registers of Khmer Consonants

Row	First Register	Second Register
1	ក <i>ka</i> [kɔː] “neck”	ក <i>ko</i> [kɔː] “mute”
2	រ <i>ro + muusikatoan</i> [rɔː] “small saw”	រ <i>ro</i> [rɔː] “fence (in the water)”
3	សក <i>sa + ka</i> [sɔːk] “to peel, to shed one’s skin”	សក <i>sa + triisap + ka</i> [sɔːk] “to insert”
4	បក <i>ba + ka</i> [bɔːk] “to return”	*បក <i>ba + triisap + ka</i> [bɔːk]
5	បម <i>ba + muusikatoan + mo</i> [pɔːm] “blockhouse”	ពម <i>po + mo</i> [pɔːm] “to put into the mouth”
6	កូរ <i>ka + u + ro</i> [koː] “to stir”	កូរ <i>ko + u + ro</i> [kuː] “to sketch”

**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 𑄒 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 𑄒 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 11-7* 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 11-9*.

Table 11-7. Khmer Subscript Consonant Signs

Glyph	Code	Name
𑄒	17D2 1780	khmer consonant sign coeng ka
𑄓	17D2 1781	khmer consonant sign coeng kha
𑄔	17D2 1782	khmer consonant sign coeng ko
𑄕	17D2 1783	khmer consonant sign coeng kho
𑄖	17D2 1784	khmer consonant sign coeng ngo
𑄗	17D2 1785	khmer consonant sign coeng ca
𑄘	17D2 1786	khmer consonant sign coeng cha
𑄙	17D2 1787	khmer consonant sign coeng co
𑄚	17D2 1788	khmer consonant sign coeng cho
𑄛	17D2 1789	khmer consonant sign coeng nyo
𑄜	17D2 178A	khmer consonant sign coeng da
𑄝	17D2 178B	khmer consonant sign coeng ttha
𑄞	17D2 178C	khmer consonant sign coeng do
𑄟	17D2 178D	khmer consonant sign coeng ttho
𑄠	17D2 178E	khmer consonant sign coeng na
𑄡	17D2 178F	khmer consonant sign coeng ta
𑄢	17D2 1790	khmer consonant sign coeng tha
𑄣	17D2 1791	khmer consonant sign coeng to
𑄤	17D2 1792	khmer consonant sign coeng tho
𑄥	17D2 1793	khmer consonant sign coeng no
𑄦	17D2 1794	khmer consonant sign coeng ba
𑄧	17D2 1795	khmer consonant sign coeng pha
𑄨	17D2 1796	khmer consonant sign coeng po
𑄩	17D2 1797	khmer consonant sign coeng pho



Table 11-7. Khmer Subscript Consonant Signs (Continued)



Glyph	Code	Name
	17D2 1798	khmer consonant sign coeng mo
	17D2 1799	khmer consonant sign coeng yo
	17D2 179A	khmer consonant sign coeng ro
	17D2 179B	khmer consonant sign coeng lo
	17D2 179C	khmer consonant sign coeng vo
	17D2 179D	khmer consonant sign coeng sha
	17D2 179E	khmer consonant sign coeng ssa
	17D2 179F	khmer consonant sign coeng sa
	17D2 17A0	khmer consonant sign coeng ha
	17D2 17A1	khmer consonant sign coeng la
	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 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 [ɔ̃əm], not [a:m]. The sequences used for these combinations are shown in Table 11-8.

Table 11-8. 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:

ដំ *da + nikahit* [dɔm] “to pound” (compare ដំ *da + mo* [dɔ:m] “nec-tar”)

ព័ *po + aa + nikahit* [pɔəm] “to carry in the beak” (compare ព័ *po + aa + mo* [pɔəm] “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 11-9.

Table 11-9. Khmer Subscript Independent Vowel Signs

Glyph	Code	Name
្ក	17D2 17A7	khmer independent vowel sign coeng qu
្ខ	17D2 17AB	khmer independent vowel sign coeng ry
្គ	17D2 17AC	khmer independent vowel sign coeng ryy
្ឃ	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:

រាជ្ជី *ro + aa + co + ro + coeng + sa + ii* [rɛ̀ɛrsei] “king hermit”

អាឃី *qa + aa + yo + robat* [ʔa:rya] “civilized” (= អាឃ្យ *qa + aa + ro + coeng + yo*)

ព័ត៌មាន *po + ta + robat + mo + aa + no* [pɔ̀:ɔmə̀ɛn] “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 11-1*.

**Figure 11-1. Common Ligatures in Khmer**

ក ka + ា aa + រ ro = កា រ [ka:] “job”  
 ប ba + ា aa = បា [ba:] “father, male of an animal”; used to prevent confusion with ហ ha  
 ប ba + ោ au = បោ [baw] “to suck”  
 ម mo + ួ coeng sa + ោ au = មួ ោ [msaw] “powder”  
 ស sa + ង ngo + ួ coeng kha + ួ coeng yo + ា aa = សង្ឃួ ោ [sɔŋk<sup>h</sup>ya:] “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 11-2*. 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 ឺ KHMER INDEPENDENT VOWEL QOO TYPE TWO is thought to be a variant of U+17B1 ឺ 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 11-2*.

**Figure 11-2. Common Multiple Forms in Khmer**

ញញឹម *nyo + nyo + y + mo* [ɲɔ̃ɲum] “to smile”  
 មីណើម *ca + i + nyo + coeng + ca + oe + mo* [ceŋcaəm] “eyebrow”  
 ស្ងប់ *sa + coeng nyo + ba + bantoc* [sɔ̃ɔp] “to respect”  
 កញ្ញា *ka + nyo + coeng + nyo + aa* [kaɲna:] “girl, Miss, September”  
 ឲ្យ *qoo-2 + coeng + yo* (= ឲ្យ *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 QAA 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:

្ក្ក្ក្ក *khan + ba + e + khan*  
 -្ក្ក- *en dash + ba + e + en dash*  
 ្ក្ក ្ក្ក *khan + lo + khan*  
 -្ក្ក- *en dash + lo + en dash*

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

$$B \{R \mid C\} \{S \{R\}^* \{Z\} V\} \{O\} \{S\}$$

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 ខ្មែរ *khnyom* “I” is composed of the following three elements: (1) consonant character *khā* 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.

ខ្មែរ *kha + coeng + nyo + u + nikahit* [k<sup>h</sup>nom] “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 11-3*.

Figure 11-3. Examples of Syllabic Order in Khmer

្រ to + e [tè:] “much”  
 ្រ្រ្រ ca + coeng + ro + oe + no [craən] “much”  
 ស្រ្រ្រាម sa + ngo + coeng + ko + coeng + ro + aa + mo [səŋkrèəm] “war”  
 ្រ្រ្រ ha + oe + coeng + yo [haəi] “already”  
 ស្រ្រ្រ sa + nyo + coeng + nyo + aa [səŋna:] “sign”  
 ស្រ្រ sa + triisap + ii [si:] “eat”  
 ្រ 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 11-3*.

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.

្រ្រ mo + muusikatoan + coeng + ngo + ai [mŋai] “one day”  
 ្រ្រ្រ្រ mo + triisap + coeng + ha + ae + ta + lek too [mhè:tmhè:t]  
 “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:

្រ្រ ្រ ba + <sup>[ZW]</sup> + triisap + ii + yo + ae + ro [biyè:] “beer”  
 ្រ្រ្រ្រ ba + coeng + ro + ta + yy + ngo + qa + <sup>[ZW]</sup> + triisap + y +  
 reahmuk [prətə:ŋtuh] “urgent, too busy”  
 ្រ្រ្រ្រ 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 11-4* for the word “savant,” pronounced [vitu:].

Figure 11-4. Ligation in *Muul* Style in Khmer

វិទូ	vo + i + to + uu	( <i>askaa crieng</i> font)
វិទូ, វិទូ	vo + i + to + uu	(ligature dependent on the <i>muul</i> font)
វិទូ	vo + <sup>[ZW]</sup> + i + to + uu	( <sup>[ZW]</sup> to prevent the ligature in a <i>muul</i> font)
វិទូ	vo + <sup>[ZW]</sup> + i + to + uu	( <sup>[ZW]</sup> to request the ligature in a <i>muul</i> 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 16-2*.

### **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.

---

## 11.5 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,  $\text{ᩉ᩠ᩅᩃᩁᩣ᩠ᨾᩮᩥᩣ᩠ᨾᩮᩥ}$  [tai<sup>2</sup> la<sup>6</sup>] (in older orthography  $\text{ᩉ᩠ᩅᩃᩁᩣ᩠ᨾᩮᩥ}$ ). 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 ᩉ, ᩊ, ᩋ, ᩌ, ᩍ or ᩎ, it is displayed to the right of the letter, as shown in *Table 11-10*.

**Table 11-10. Tai Le Tone Marks**

Syllable	New Orthography	Old Orthography
<i>ta</i>	ᩉ	ᩉ
<i>ta</i> <sup>2</sup>	ᩉ᩠	ᩉ᩠
<i>ta</i> <sup>3</sup>	ᩉᩣ	ᩉᩣ
<i>ta</i> <sup>4</sup>	ᩉ᩠ᩣ	ᩉ᩠ᩣ

Table 11-10. Tai Le Tone Marks (Continued)

Syllable	New Orthography	Old Orthography
<i>ta</i> <sup>5</sup>	တၢ	တံ
<i>ta</i> <sup>6</sup>	တၢင	တံင
<i>ti</i>	တံ	တံ
<i>ti</i> <sup>2</sup>	တံၣ်	တံး
<i>ti</i> <sup>3</sup>	တံၣ်း	တံး
<i>ti</i> <sup>4</sup>	တံၣ်ၤ	တံး
<i>ti</i> <sup>5</sup>	တံၣ်ၤ	တံး
<i>ti</i> <sup>6</sup>	တံၣ်င	တံး

**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 11-11*.

Table 11-11. Myanmar Digits

Myanmar-Style Glyphs	Tai Le-Style Glyphs
၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉
၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉	၀ ၁ ၂ ၃ ၄ ၅ ၆ ၇ ၈ ၉

**Punctuation.** Both CJK punctuation and Western punctuation are used. Typographically, European digits are about the same height and depth as the tall characters [၂] and [၃]. In some fonts, the baseline for punctuation is the depth of those characters.

---

## 11.6 New Tai Lue

### *New Tai Lue: U+1980–U+19DF*

The New Tai Lue script, also known as Xishuang Banna Dai, is used mainly in southern China. The script was developed in the twentieth 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 twentieth 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 11.7, 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. By contrast, Lanna uses both spacing vowel signs and nonspacing vowel signs, which appear above or below the consonants they modify.

**Syllabic Structure.** All vowel signs in New Tai Lue are considered combining characters and follow their base consonants in the text stream. Where a syllable is composed of a vowel sign to the left and a vowel or tone mark on the right of the consonant, a sequence of characters is used, in the order *consonant + vowel + tone mark*, as shown in Table 11-12.

Table 11-12. New Tai Lue Vowel Placement

ꨀ	ka +	ꨀ	e +	ꨀ	t1	→	ꨀꨀꨀ	[ke: <sup>2</sup> ]		
ꨀ	ka +	ꨀ	e +	ꨀ	i	→	ꨀꨀꨀ	[kə: <sup>1</sup> ]		
ꨀ	ka +	ꨀ	e +	ꨀ	iy	→	ꨀꨀꨀ	[kəi <sup>1</sup> ]		
ꨀ	ka +	ꨀ	e +	ꨀ	iy +	ꨀ	t1	→	ꨀꨀꨀꨀ	[kəi <sup>2</sup> ]
ꨀ	ka +	ꨀ	e +	ꨀ	iy +	ꨀ	t2	→	ꨀꨀꨀꨀ	[kəi <sup>3</sup> ]

**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 11-13.

Table 11-13. New Tai Lue Registers and Tones

Display	Sequence	Register	Tone Mark	Tone	Transcription
ꨀ	ka <sup>h</sup>	high		1	[ka <sup>1</sup> ]
ꨀꨀ	ka <sup>h</sup> + t1	high	t1	2	[ka <sup>2</sup> ]
ꨀꨀ	ka <sup>h</sup> + t2	high	t2	3	[ka <sup>3</sup> ]
ꨀ	ka <sup>l</sup>	low		4	[ka <sup>4</sup> ]
ꨀꨀ	ka <sup>l</sup> + t1	low	t1	5	[ka <sup>5</sup> ]
ꨀꨀ	ka <sup>l</sup> + t2	low	t2	6	[ka <sup>6</sup> ]

**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 ꨀ NEW TAI LUE THAM DIGIT ONE. The regular *hora* form for the digit, U+19D1 ꨀ NEW TAI LUE DIGIT ONE, has the exact same glyph shape as a common New Tai Lue vowel, U+19B3 ꨀ 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.

## 11.7 Tai Tham

### *Tai Tham: U+1A20–U+1AAF*

The Tai Tham (or Lanna) script is used for three living languages: Northern Thai (that is, Kam Mu'ang), Tai Lue, and Khün. In addition, the script is also used for Lao Tham (or Old



Lao) and other dialect variants in Buddhist palm leaves and notebooks. The script is also known as the Tham or Yuan script. 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 670,000 speakers of Tai Lue. Of those, people born before 1950 may be literate in the Tai Tham script. Younger speakers are taught the New Tai Lue script, instead. (See *Section 11.6, New Tai Lue.*) The Tai Tham script continues to be taught in the Tai Lue monasteries. There are 120,000 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 HAM 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 KAAAN, 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.

## 11.8 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. Vowel 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 11-14.

Table 11-14. Tai Viet Symbols and Punctuation

Code	Glyph	Name	Meaning
AADB	ꨀ	<i>kon</i>	person
AADC	ꨁ	<i>nueng</i>	one
AADD	ꨂ	<i>sam</i>	signals repetition of the previous word
AADE	ꨃ	<i>ho hoi</i>	beginning of text (used in songs and poems)
AADF	ꨄ	<i>koi koi</i>	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.

---

## 11.9 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.

## 11.10 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 *-u* 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 11-15.

Table 11-15. Cham Syllabic Structure

Class	Examples	Encoding
consonant or independent vowel		[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		[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*.

## 11.11 Philippine Scripts

**Tagalog:** U+1700–U+171F

**Hanunóo:** U+1720–U+173F

**Buhid:** U+1740–U+175F

**Tagbanwa:** U+1760–U+177F

The first of these four scripts—Tagalog—is no longer used, whereas the other three—Hanunóo, Buhid, and Tagbanwa—are living scripts of the Philippines. South Indian scripts of the Pallava dynasty made their way to the Philippines, although the exact route is uncertain. They may have been transported by way of the Kavi scripts of Western Java between the tenth and fourteenth centuries CE.

Written accounts of the Tagalog script by Spanish missionaries and documents in Tagalog date from the mid-1500s. The first book in this script was printed in Manila in 1593. While the Tagalog script was used to write Tagalog, Bisaya, Ilocano, and other languages, it fell out of normal use by the mid-1700s. The modern Tagalog language—also known as Filipino—is now written in the Latin script.

The three living scripts—Hanunóo, Buhid, and Tagbanwa—are related to Tagalog but may not be directly descended from it. The Hanunóo and the Buhid peoples live in Mindoro, while the Tagbanwa live in Palawan. Hanunóo enjoys the most use; it is widely used to write love poetry, a popular pastime among the Hanunóo. Tagbanwa is used less often.

### Principles of the Philippine Scripts

The Philippine scripts share features with the other Brahmi-derived scripts to which they are related.

**Consonant Letters.** Philippine scripts have consonants containing an inherent *-a* vowel, which may be modified by the addition of vowel signs or canceled (killed) by the use of a virama-type mark.

**Independent Vowel Letters.** Philippine scripts have null consonants, which are used to write syllables that start with a vowel.

**Dependent Vowel Signs.** The vowel *-i* is written with a mark above the associated consonant, and the vowel *-u* with an identical mark below. The mark is known as *kudlit* “diacritic,” *tuldik* “accent,” or *tuldok* “dot” in Tagalog, and as *ulitan* “diacritic” in Tagbanwa. The Philippine scripts employ only the two vowel signs *i* and *u*, which are also used to stand for the vowels *e* and *o*, respectively.

**Virama.** Although all languages normally written with the Philippine scripts have syllables ending in consonants, not all of the scripts have a mechanism for expressing the canceled *-a*. As a result, in those orthographies, the final consonants are unexpressed. Francisco Lopez introduced a cross-shaped *virama* in his 1620 catechism in the Ilocano language, but this innovation did not seem to find favor with native users, who seem to have considered the script adequate without it (they preferred 𑄆𑄆𑄇 *kakapi* to 𑄆𑄆𑄇𑄇 *kakampi*). A similar reform for the Hanunóo script seems to have been better received. The Hanunóo *pamudpod* was devised by Antoon Postma, who went to the Philippines from the Netherlands in the mid-1950s. In traditional orthography, 𑄆𑄇 𑄆𑄇 𑄇 𑄆𑄇𑄆 *si apu ba upada* is, with the *pamudpod*, rendered more accurately as 𑄆𑄇 𑄆𑄇𑄆𑄆𑄆𑄆 𑄇𑄆𑄆 𑄆𑄇𑄆𑄆𑄆 *si aypud bay upadan*; the Hanunóo pronunciation is *si aypod bay upadan*. The Tagalog *virama* and Hanunóo *pamudpod* cancel only the inherent *-a*. No conjunct consonants are employed in the Philippine scripts.



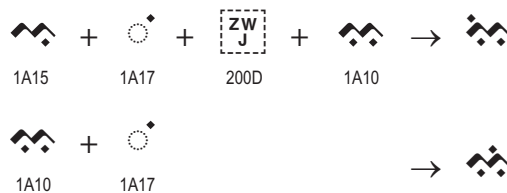


guage with a rich traditional literature, is one of the foremost languages of Indonesia. The script was previously also used to write the Makassar, Bimanese, and Madurese languages.

**Structure.** Buginese vowel signs are used in a manner similar to that seen in other Brahmi-derived scripts. Consonants have an inherent /a/ vowel sound. Consonant conjuncts are not formed. Traditionally, a virama does not exist, but is included for modern usage in transcribing many non-Buginese words. This innovation is paralleled by a similar innovation in Hanunóo and Tagalog. The virama is always a visible sign. Because conjuncts are not formed in Buginese, U+200C ZERO WIDTH NON-JOINER is not necessary to force the display of the virama.

**Ligature.** One ligature is found in the Buginese script. It is formed by the ligation of <a, -i> + *ya* to represent *îya*, as shown in the first line of *Figure 11-5*. The ligature takes the shape of the Buginese letter *ya*, but with a dot applied at the far left side. Contrast that with the normal representation of the syllable *yi*, in which the dot indicating the vowel sign occurs in a centered position, as shown in the second line of *Figure 11-5*. The ligature for *îya* is not obligatory; it would be requested by inserting a *zero width joiner*.

Figure 11-5. Buginese Ligature



**Order.** Several orderings are possible for Buginese. The Unicode Standard encodes the Buginese characters in the Matthes order.

**Punctuation.** Buginese uses spaces between certain units. One punctuation symbol, U+1A1E BUGINESE PALLAWA, is functionally similar to the full stop and comma of the Latin script. There is also another separation mark, U+1A1F BUGINESE END OF SECTION.

U+0662 ARABIC-INDIC DIGIT TWO or a doubling of the vowel sign (especially U+1A19 BUGINESE VOWEL SIGN E and U+1A1A BUGINESE VOWEL SIGN O) is used sometimes to denote word reduplication.

**Numerals.** There are no known digits specific to the Buginese script.

## 11.13 Balinese

### **Balinese:** U+1B00–U+1B7F

The Balinese script, or *aksara Bali*, is used for writing the Balinese language, the native language of the people of Bali, known locally as *basa Bali*. It is a descendant of the ancient Brahmi script of India, and therefore it has many similarities with modern scripts of South Asia and Southeast Asia, which are also members of that family. The Balinese script is used to write Kawi, or Old Javanese, which strongly influenced the Balinese language in the eleventh century CE. A slightly modified version of the script is used to write the Sasak language, which is spoken on the island of Lombok to the east of Bali. Some Balinese words have been borrowed from Sanskrit, which may also be written in the Balinese script.

**Structure.** Balinese consonants have an inherent -a vowel sound. Consonants combine with following consonants in the usual Brahmic fashion: the inherent vowel is “killed” by

U+1B44 BALINESE ADEG ADEG (*virama*), and the following consonant is subjoined or post-fixed, often with a change in shape. Table 11-17 shows the base consonants and their conjunct forms.

Table 11-17. Balinese Base Consonants and Conjunct Forms

Consonant	Base Form	Conjunct Form
ka	ꦏ	ꦏꦲ
kha	ꦏꦲꦲ	ꦏꦲꦲꦲ
ga	ꦒ	ꦒꦲ
gha	ꦒꦲꦲ	ꦒꦲꦲꦲ
nga	ꦒ	ꦒꦲ
ca	ꦚ	ꦚꦲ
cha	ꦚꦲꦲ	ꦚꦲꦲꦲ
ja	ꦗ	ꦗꦲ
jha	ꦗꦲꦲ	ꦗꦲꦲꦲ
nya	ꦚꦺ	ꦚꦺꦲ
tta	ꦠꦲ	ꦠꦲꦲ
ttha	ꦠꦲꦲꦲ	ꦠꦲꦲꦲꦲ
dda	ꦢ	ꦢꦲ
ddha	ꦢꦲꦲ	ꦢꦲꦲꦲ
nna	ꦚꦺ	ꦚꦺꦲ
ta	ꦠ	ꦠꦲ
tha	ꦠꦲꦲ	ꦠꦲꦲꦲ
da	ꦢ	ꦢꦲ
dha	ꦢꦲꦲ	ꦢꦲꦲꦲ
na	ꦚꦺ	ꦚꦺꦲ
pa	ꦥ	ꦥꦲ
pha	ꦥꦲꦲ	ꦥꦲꦲꦲ
ba	ꦨ	ꦨꦲ
bha	ꦨꦲꦲ	ꦨꦲꦲꦲ
ma	ꦩ	ꦩꦲ
ya	ꦚꦺ	ꦚꦺꦲ
ra	ꦫ	ꦫꦲ
la	ꦭ	ꦭꦲ
wa	ꦮ	ꦮꦲ

Table 11-17. Balinese Base Consonants and Conjunct Forms (Continued)

Consonant	Base Form	Conjunct Form
<i>ssa</i>	ꦱꦱ	ꦱꦱꦱ
<i>sha</i>	ꦱ	ꦱꦱ
<i>sa</i>	ꦱ	ꦱꦱꦱ
<i>ha</i>	ꦱ	ꦱꦱ
<i>r</i>	ꦱ	ꦱꦱ

The seven letters U+1B45 BALINESE LETTER KAF SASAK through U+1B4B BALINESE LETTER ASYURA SASAK are base consonant extensions for the Sasak language. Their base forms and conjunct forms are shown in *Table 11-18*.

Table 11-18. Sasak Extensions for Balinese

Consonant	Base Form	Conjunct Form
<i>kaf</i>	ꦱꦱ	ꦱꦱꦱ
<i>khot</i>	ꦱꦱꦱ	ꦱꦱꦱꦱ
<i>tzir</i>	ꦱꦱ	ꦱꦱꦱ
<i>ef</i>	ꦱ	ꦱꦱ
<i>ve</i>	ꦱ	ꦱꦱ
<i>zal</i>	ꦱꦱ	ꦱꦱꦱ
<i>asyura</i>	ꦱꦱꦱ	ꦱꦱꦱꦱ

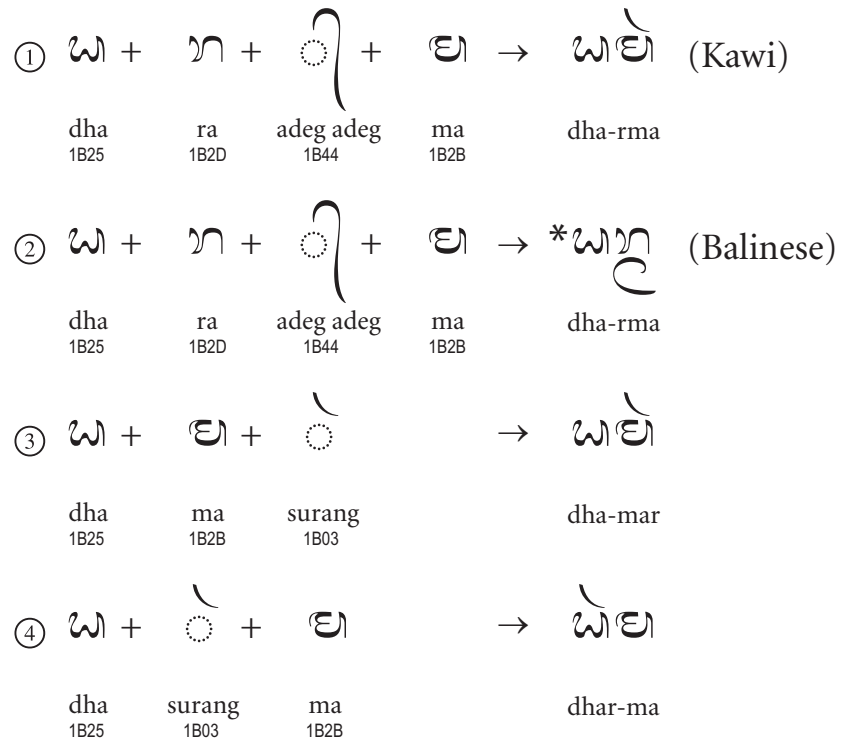
Balinese dependent vowel signs are used in a manner similar to that employed by other Brahmic scripts.

Independent vowels are used in a manner similar to that seen in other Brahmic scripts, with a few differences. For example, U+1B05 BALINESE LETTER AKARA and U+1B0B BALINESE LETTER RA REPA can be treated as consonants; that is, they can be followed by *adeg* *adeg*. In Sasak, the vowel letter *akara* can be followed by an explicit *adeg adeg* <sup>ꦱꦱ</sup> in word- or syllable-final position, where it indicates the glottal stop; other consonants can also be subjoined to it.




**Behavior of ra.** Unlike most Brahmi-derived scripts, a Balinese *ra* that starts a sequence of consonants without intervening vowels is represented by U+1B03 BALINESE SIGN SURANG over the preceding syllable, as shown in the fourth example in *Figure 11-6*. The inherited Kawi form of the script used a *repha* glyph in the same way as many Brahmic scripts do. This is seen in the first example in *Figure 11-6*, where the sequence <*ra*, *virama*, *ma*> is rendered with the *repha* glyph. However, because many syllables end in *-r* in the Balinese language, this written form was historically reanalyzed, and is now pronounced *damar* in Balinese, as shown in the third example. In Balinese, the character sequence used in Kawi to spell *dharma* would render as shown in the second example, where the base letter *ra* with a subjoined *ma* is not well formed for the writing system.

Because of its relationship to *ra*, *surang* should be treated as equivalent to *ra* for searching and sorting purposes. Two other combining signs are also equivalent to base letters for

Figure 11-6. Writing *dharma* in Balinese



searching and sorting: U+1B02 BALINESE SIGN CECEK (*anusvara*) is equivalent to *nga*, and U+1B04 BALINESE SIGN BISAH (*visarga*) is equivalent to *ha*.

**Behavior of ra repa.** The unique behavior of BALINESE LETTER RA REPA (*vocalic r*) results from a reanalysis of the independent vowel letter as a consonant. In a compound word in which the first element ends in a consonant and the second element begins with an original *ra + pepet*, such as *Pak Rërëh*  “Mr Rërëh”, the postfixed form of  *ra repa* is used; this particular sequence is encoded *ka + adeg adeg + ra repa*. However, in other contexts where the *ra repa* represents the original Sanskrit vowel, U+1B3A BALINESE VOWEL SIGN RA REPA is used, as in *Krësna* .

**Rendering.** The vowel signs /u/ and /u:/ take different forms when combined with subscripted consonant clusters, as shown in *Table 11-19*. The upper limit of consonant clusters is three, the last of which can be *-ya*, *-wa*, or *-ra*.

Table 11-19. Balinese Consonant Clusters with u and u:













Syllable	Glyph
<i>kyu</i>	
<i>kyú</i>	
<i>kwu</i>	
<i>kwú</i>	
<i>kru</i>	

Table 11-19. Balinese Consonant Clusters with u and u: (Continued)

Syllable	Glyph
<i>krú</i>	
<i>kryu</i>	
<i>kryú</i>	
<i>skru</i>	
<i>skrú</i>	

**Nukta.** The combining mark U+1B34 BALINESE SIGN REREKAN (*nukta*) and a similar sign in Javanese are used to extend the character repertoire for foreign sounds. In recent times, Sasak users have abandoned the Javanese-influenced *rerekan* in favor of the series of modified letters shown in Table 11-18, also making use of some unused Kawi letters for these Arabic sounds.

**Ordering.** The traditional order *ha na ca ra ka | da ta sa wa la | ma ga ba nga | pa ja ya nya* is taught in schools, although van der Tuuk followed the Javanese order *pa ja ya nya | ma ga ba nga* for the second half. The arrangement of characters in the code charts follows the Brahmic ordering.

**Punctuation.** Both U+1B5A BALINESE PANTI and U+1B5B BALINESE PAMADA are used to begin a section in text. U+1B5D BALINESE CARIK PAMUNGKAH is used as a colon. U+1B5E BALINESE CARIK SIKI and U+1B5F BALINESE CARIK PAREREN are used as comma and full stop, respectively. At the end of a section,  *pasalinan* and  *carik agung* may be used (depending on which sign began the section). They are encoded using the punctuation ring U+1B5C BALINESE WINDU together with *carik pareren* and *pamada*.

**Hyphenation.** Traditional Balinese texts are written on palm leaves; books of these bound leaves together are called *lontar*. U+1B60 BALINESE PAMENENG is inserted in *lontar* texts where a word must be broken at the end of a line (always after a full syllable). This sign is not used as a word-joining hyphen—it is used only in line breaking.

**Musical Symbols.** Bali is well known for its rich musical heritage. A number of related notation systems are used to write music. To represent degrees of a scale, the syllables *ding dong dang deng dung* are used (encoded at U+1B61..U+1B64, U+1B66), in the same way that *do re mi fa so la ti* is used in Western tradition. The symbols representing these syllables are based on the vowel matras, together with some other symbols. However, unlike the regular vowel matras, these stand-alone spacing characters take diacritical marks. They also have different positions and sizes relative to the baseline. These matra-like symbols are encoded in the range U+1B61..U+1B6A, along with a modified *aikara*. Some notation systems use other spacing letters, such as U+1B09 BALINESE LETTER UKARA and U+1B27 BALINESE LETTER PA, which are not separately encoded for musical use. The U+1B01 BALINESE SIGN ULU CANDRA (*candrabindu*) can also be used with U+1B62 BALINESE MUSICAL SYMBOL DENG and U+1B68 BALINESE MUSICAL SYMBOL DEUNG, and possibly others. BALINESE SIGN ULU CANDRA can be used to indicate modre symbols as well.

A range of diacritical marks is used with these musical notation base characters to indicate metrical information. Some additional combining marks indicate the instruments used; this set is encoded at U+1B6B..U+1B73. A set of symbols describing certain features of performance are encoded at U+1B74..U+1B7C. These symbols describe the use of the right or left hand, the open or closed hand position, the “male” or “female” drum (of the pair) which is struck, and the quality of the striking.

**Modre Symbols.** The Balinese script also includes a range of “holy letters” called modre symbols. Most of these letters can be composed from the constituent parts currently encoded, including U+1B01 BALINESE SIGN ULU CANDRA.

---

## 11.14 Javanese

### *Javanese: U+A980–U+A9DF*

The Javanese script, or *aksara Jawa*, is used for writing the Javanese language, known locally as *basa Jawa*. The script is a descendent of the ancient Brahmi script of India, and so has many similarities with the modern scripts of South Asia and Southeast Asia which are also members of that family. The Javanese script is also used for writing Sanskrit, Jawa Kuna (a kind of Sanskritized Javanese), and transcriptions of Kawi, as well as the Sundanese language, also spoken on the island of Java, and the Sasak language, spoken on the island of Lombok.

The Javanese script was in current use in Java until about 1945; in 1928 Bahasa Indonesia was made the national language of Indonesia and its influence eclipsed that of other languages and their scripts. Traditional Javanese texts are written on palm leaves; books of these bound together are called *lontar*, a word which derives from ron “leaf” and tal “palm”.

**Consonants.** Consonants have an inherent *-a* vowel sound. Consonants combine with following consonants in the usual Brahmic fashion: the inherent vowel is “killed” by U+A9C0 JAVANESE PANGKON, and the following consonant is subjoined or postfixed, often with a change in shape.

Vocalic liquids (*r* and *l*) are treated as consonant letters in Javanese; they are not independent vowels with dependent vowel equivalents, as is the case in Balinese or Devanagari. Short and long versions of the *vocalic-l* are separately encoded, as U+A98A JAVANESE LETTER NGA LELET and U+A98B JAVANESE LETTER NGA LELELT RASWADI. In contrast, the long version of the *vocalic-r* is represented by a sequence of the short vowel U+A989 JAVANESE LETTER PA CERK followed by the dependent vowel sign *-aa*, U+A9B4 JAVANESE SIGN TARUNG, serving as a length mark in this case.

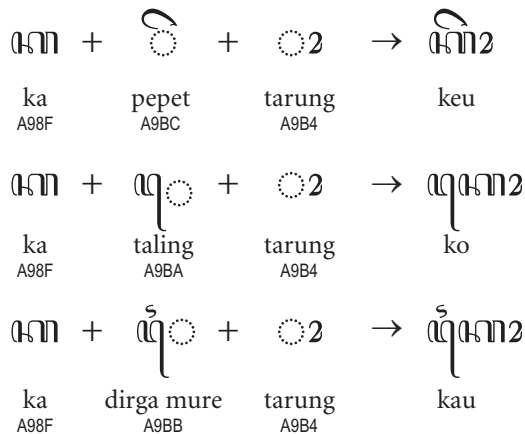
U+A983 JAVANESE SIGN CECAK TELU is a diacritic used with various consonantal base letters to represent foreign sounds. Typically these diacritic-marked consonants are used for sounds borrowed from Arabic.

**Independent Vowels.** Independent vowel letters are used essentially as in other Brahmic scripts. Modern Javanese uses U+A986 JAVANESE LETTER I and U+A987 JAVANESE LETTER II for short and long *i*, but the Kawi orthography instead uses U+A985 JAVANESE LETTER I KAWI and U+A986 JAVANESE LETTER I for short and long *i*, respectively.

The long versions of the *u* and *o* vowels are written as sequences, using U+A9B4 JAVANESE SIGN TARUNG as a length mark.

**Dependent Vowels.** Javanese—unlike Balinese—represents multi-part dependent vowels with sequences of characters, in a manner similar to the Myanmar script. The Balinese community considers it important to be able to directly transliterate Sanskrit into Balinese, so multi-part dependent vowels are encoded as single, composite forms in Balinese, as is done in Devanagari. In contrast, for the Javanese script, the correspondence with Sanskrit letters is not so critical, and a different approach to the encoding has been taken. Similar to the treatment of long versions of Javanese independent vowels, the two-part dependent vowels are explicitly represented with a sequence of two characters, using U+A9B4 JAVANESE VOWEL SIGN TARUNG, as shown in *Figure 11-7*.

Figure 11-7. Representation of Javanese Two-Part Vowels



**Consonant Signs.** The characters U+A980 JAVANESE SIGN PANYANGGA, U+A981 JAVANESE SIGN CECAK, and U+A983 JAVANESE SIGN WIGNYAN are analogous to U+0901 DEVANAGARI SIGN CANDRABINDU, U+0902 DEVANAGARI SIGN ANUSVARA, and U+0903 DEVANAGARI SIGN VISARGA and behave in much the same way.

There are two medial consonant signs, U+A9BE JAVANESE CONSONANT SIGN PENGKAL and U+A9BF JAVANESE CONSONANT SIGN CAKRA, which represent *-y-* and *-r-* respectively. These medial consonant signs contrast with the subjoined forms of the letters *ya* and *ra*. The subjoined forms may indicate a syllabic boundary, whereas *pengkal* and *cakra* are used in ordinary consonant clusters.

**Rendering.** There are many conjunct forms in Javanese, though most are fairly regular and easy to identify. Subjoined consonants and vowel signs rendered below them usually interact typographically. For example, the vowel signs [u] and [u:] take different forms when combined with subscripted consonant clusters. Consonant clusters may have up to three elements. In three-element clusters, the last element is always one of the medial glides: *-ya*, *-wa*, or *-ra*.

**Digits.** The Javanese script has its own set of digits, seven of which (1, 2, 3, 6, 7, 8, 9) look just like letters of the alphabet. Implementations with concerns about security issues need to take this into account. The punctuation mark U+A9C8 JAVANESE PADA LINGSA is often used with digits in order to help to distinguish numbers from sequences of letters. When Javanese personal names are abbreviated, the letters are followed, not preceded, by PADA LINGSA.

**Punctuation.** A large number of punctuation marks are used in Javanese. Titles may be flanked by the pair of ornamental characters, U+A9C1 JAVANESE LEFT RERENGGAN and U+A9C2 JAVANESE RIGHT RERENGGAN; glyphs used for these may vary widely.

U+A9C8 JAVANESE PADA LINGSA is a danda mark that corresponds functionally to the use of a comma. The doubled form, U+A9C9 JAVANESE PADA LUNGSI, corresponds functionally to the use of a full stop. It is also used as a “ditto” mark in vertical lists. U+A9C7 JAVANESE PADA PANGKAT is used much like the European colon.

The doubled U+A9CB JAVANESE PADA ADEG ADEG typically begins a paragraph or section, while the simple U+A9CA JAVANESE PADA ADEG is used as a common divider though it can be used in pairs marking text for attention. The two characters, U+A9CC JAVANESE PADA PISELEH and U+A9CD JAVANESE TURNED PADA PISELEH, are used similarly, either both together or with U+A9CC JAVANESE PADA PISELEH simply repeated.

The punctuation ring, U+A9C6 JAVANESE PADA WINDU, is not used alone, a situation similar to the pattern of use for its Balinese counterpart U+1B5C BALINESE WINDU. When used with U+A9CB JAVANESE PADA ADEG ADEG this *windu* sign is called *pada guru*, *pada bab*, or *uger-uger*, and is used to begin correspondence where the writer does not desire to indicate a rank distinction as compared to his audience. More formal letters may begin with one of the three signs: U+A9C3 JAVANESE PADA ANDAP (for addressing a higher-ranked person), U+A9C4 JAVANESE PADA MADYA (for addressing an equally-ranked person), or U+A9C5 JAVANESE PADA LUHUR (for addressing a lower-ranked person).

**Reduplication.** U+A9CF JAVANESE PADA PANGRANGKEP is used to show the reduplication of a syllable. The character derives from U+0662 ARABIC-INDIC DIGIT TWO but in Javanese it does not have a numeric use. The Javanese reduplication mark is encoded as a separate character from the Arabic digit, because it differs in its Bidi\_Class property value.

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

$$\{C\} C \{\{R\}Y\} \{V\{A\}\} \{Z\}$$

where

C is a letter (consonant or independent vowel), or a consonant followed by the diacritic U+A9B3 JAVANESE SIGN CECAK TELU

F is the virama, U+A9C0 JAVANESE PANGKON

R is the medial -ra, U+A9BF JAVANESE CONSONANT SIGN CAKRA

Y is the medial -ya, U+A9BE JAVANESE CONSONANT SIGN PENGKAL

V is a dependent vowel sign

A is the dependent vowel sign -aa, U+A9B4 JAVANESE VOWEL SIGN TARUNG

Z is a consonant sign: U+A980, U+A981, U+A982, or U+A983

**Linebreaking.** Opportunities for linebreaking occur after any full orthographic syllable. Hyphens are not used.

In some printed texts, an epenthetic spacing U+A9BA JAVANESE VOWEL SIGN TALING is placed at the end of a line when the next line begins with the glyph for U+A9BA JAVANESE VOWEL SIGN TALING, which is reminiscent of a specialized hyphenation (or of quire marking). This practice is nearly impossible to implement in a free-flowing text environment. Typographers wishing to duplicate a printed page may manually insert U+00A0 NO-BREAK SPACE before U+A9BA JAVANESE VOWEL SIGN TALING at the end of a line, but this would not be orthographically correct.

---

## 11.15 Rejang

### **Rejang: U+A930–U+A95F**

The Rejang language is spoken by about 200,000 people living on the Indonesian island of Sumatra, mainly in the southwest. There are five major dialects: Lebong, Musi, Kebanagun, Pesisir (all in Bengkulu Province), and Rawas (in South Sumatra Province). Most Rejang speakers live in fairly remote rural areas, and slightly less than half of them are literate.



The Rejang script was in use prior to the introduction of Islam to the Rejang area. The earliest attested document appears to date from the mid-18th century CE. The traditional Rejang corpus consists chiefly of ritual texts, medical incantations, and poetry.

**Structure.** Rejang is a Brahmi-derived script. It is related to other scripts of the Indonesian region, such as Batak and Buginese.

Consonants in Rejang have an inherent /a/ vowel sound. Vowel signs are used in a manner similar to that employed by other Brahmi-derived scripts. There are no consonant conjuncts. The basic syllabic structure is C(V)(F): a consonant, followed by an optional vowel sign and an optional final consonant sign or virama.

**Rendering.** Rejang texts tend to have a slanted appearance typified by the appearance of U+A937 REJANG LETTER BA. This sense that the script is tilted to the right affects the placement of the combining marks for vowel signs. Vowel signs above a letter are offset to the right, and vowel signs below a letter are offset to the left, as the “above” and “below” positions for letters are perceived in terms of the overall slant of the letters.

**Ordering.** The ordering of the consonants and vowel signs for Rejang in the code charts follows a generic Brahmic script pattern. The Brahmic ordering of Rejang consonants is attested in numerous sources. There is little evidence one way or the other for preferences in the relative order of Rejang vowel signs and consonant signs.

**Digits.** There are no known script-specific digits for the Rejang script.

**Punctuation.** European punctuation marks such as comma, full stop, and colon, are used in modern writing. U+A95F REJANG SECTION MARK may be used at the beginning and end of paragraphs.

Traditional Rejang texts tend not to use spaces between words, but their use does occur in more recent texts. There is no known use of hyphenation.

## 11.16 Batak

### **Batak:** U+1BC0–U+1BFF

The Batak script is used on the island of Sumatra to write the five Batak dialects: Karo, Mandailing, Pakpak, Simalungun, and Toba. The script is called *si-sia-sia* or *surat na sampulu sia*, which means “the nineteen letters.” The script is taught in schools mainly for cultural purposes, and is used on some signs for shops and government offices.

**Structure.** Batak is a Brahmi-derived script. It is written left to right. Batak uses a vowel killer which is called *pangolat* in Mandailing, Pakpak, and Toba. In Karo the killer is called *penengen*, and in Simalungun it is known as *panongonan*. The appearance of the killer differs between some of the dialects. Consonant conjuncts are not formed. Batak has three independent vowels and makes use of a number of vowel signs and two consonant signs. Some vowel signs are only used by certain language communities.

**Rendering.** Most vowel signs and the two killers, U+1BF2 BATAK PANGOLAT and U+1BF3 BATAK PANONGONAN, are spacing marks. U+1BEE BATAK VOWEL SIGN U can ligate with its base consonant.

The two consonant signs, U+1BF0 BATAK CONSONANT SIGN NG and U+1BF1 BATAK CONSONANT SIGN H, are nonspacing marks, usually rendered above the spacing vowel signs. When U+1BF0 BATAK CONSONANT SIGN NG occurs together with the nonspacing mark, U+1BE9 BATAK VOWEL SIGN EE, both are rendered above the base consonant, with the glyph for the *ee* at the top left and the glyph for the *ng* at the top right.

The main peculiarity of Batak rendering concerns the reordering of the glyphs for vowel signs when one of the two killers, *pangolat* or *panongonan*, is used to close the syllable by killing the inherent vowel of a final consonant. This reordering for display is entirely regular. So, while the representation of the syllable /tip/ is done in logical order: <ta, vowel sign i, pa, pangolat>, when rendered for display the glyph for the vowel sign is visually applied to the final consonant, *pa*, rather than to the *ta*. The glyph for the *pangolat* always stays at the end of the syllable.

**Punctuation.** Punctuation is not normally used; instead all letters simply run together. However, a number of *bindu* characters are occasionally used to disambiguate similar words or phrases. U+1BFF BATAK SYMBOL BINDU PANGOLAT is trailing punctuation, following a word, surrounding the previous character somewhat.

The minor mark used to begin paragraphs and stanzas is U+1BFC BATAK SYMBOL BINDU NA METEK, which means “small bindu.” It has a shape-based variant, U+1BFD BATAK SYMBOL BINDU PINARBORAS (“rice-shaped bindu”), which is likewise used to separate sections of text. U+1BFE BATAK SYMBOL BINDU JUDUL (“title bindu”) is sometimes used to separate a title from the main text, which normally begins on the same line.

**Linebreaking.** Opportunities for a linebreak occur after any full orthographic syllable, defined as C(V(C<sub>s</sub>|C<sub>d</sub>)) where a consonant C may be followed by a vowel sign V which may be followed either by a consonant sign C<sub>s</sub> (-ng or -h) or a killed final consonant C<sub>d</sub>.

---

## 11.17 Sundanese

### **Sundanese: U+1B80–U+1BBF**

The Sundanese script, or *aksara Sunda*, is used for writing the Sundanese language, one of the languages of the island of Java in Indonesia. It is a descendent of the ancient Brahmi script of India, and so has similarities with the modern scripts of South Asia and Southeast Asia which are also members of that family. The script has official support. It is taught in schools and used on road signs.

The Sundanese language has been written using a number of different scripts over the years. Pallawa or Pra-Nagari was first used in West Java to write Sanskrit from the fifth to the eighth centuries CE. *Sunda Kuna* or Old Sundanese was derived from Pallawa and was used in the Sunda Kingdom from the 14th to the 18th centuries. The earliest example of Old Sundanese is the Prasasti Kawali stone. The Javanese script was used to write Sundanese from the 17th to the 19th centuries, and the Arabic script was used from the 17th to the 20th centuries. The Latin script has been in wide use since the 20th century. The modern Sundanese script, called *Sunda Baku* or Official Sundanese, became official in 1996. This modern script was derived from Old Sundanese.

**Structure.** Sundanese consonants have an inherent vowel /a/. This inherent vowel can be modified by the addition of dependent vowel signs (matras). An explicit *virama*, U+1BAA SUNDANESE SIGN PAMAAEH, is used to indicate the absence, or “killing,” of the inherent vowel. Sundanese does not use the *virama* to cluster consonants or build consonant conjuncts.

Initial Sundanese consonants can be followed by one of the three consonant signs for medial consonants: *-ya*, *-ra*, or *-la*. These medial consonants are graphically displayed as subjoined elements to their base consonants. The script also has independent vowel letters.

Three final consonants are separately encoded as combining marks: *-ng*, *-r*, and *-h*. These are analogues of Brahmic anusvara, repha, and visarga, respectively.

**Consonant Additions.** Two supplemental consonant letters have been added to the script recently: U+1BAE SUNDANESE LETTER KHA and U+1BAF SUNDANESE LETTER SYA. These are used to represent the borrowed sounds denoted by the Arabic letters *kha* and *sheen*, respectively.

**Digits.** Sundanese has its own script-specific digits, which are separately encoded in this block.

**Punctuation.** Sundanese uses European punctuation marks, such as comma, full stop, question mark, and quotation marks. Spaces are used in text. Opportunities for hyphenation occur after any full orthographic syllable.

**Ordering.** The order of characters in the code charts follows the Brahmic ordering. The ha-na-ca-ra-ka order found in Javanese and Balinese does not seem to be used in Sundanese.

**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 11-20.

Table 11-20. Sundanese Syllabic Structure

Class	Examples	Encoding
consonant or independent vowel	ᮊ	[U+1B83..U+1BA0, U+1BAE, U+1BAF]
consonant sign -ya, -ra, -la	ᮊᮒ, ᮊᮓ, ᮊᮔ	[U+1BA1..U+1BA3]
dependent vowel, virama	ᮊᮕ, ᮊᮖ	[U+1BA4..U+1BA9, U+1BAA]
final consonant	ᮊᮗ	[U+1B80..U+1B82]

The *virama* occupies the same logical position as a dependent vowel, but indicates the absence, rather than the presence of a vowel. It cannot be followed by a combining mark for a final consonant, nor can it be preceded by a consonant sign.

The left-side dependent vowel U+1BA6 SUNDANESE VOWEL SIGN PANAE LAENG OCCURS in logical order after the consonant (and any medial consonant sign), but in visual presentation its glyph appears *before* (to the left of) the consonant.

**Rendering.** When more than one sign appears above or below a consonant, the two are rendered side-by-side, rather than being stacked vertically.