

This PDF file is an excerpt from *The Unicode Standard, Version 4.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this online edition, however the PDF files have not been modified to reflect the corrections found on the Updates and Errata page (<http://www.unicode.org/errata/>). For information on more recent versions of the standard, see <http://www.unicode.org/standard/versions/enumeratedversions.html>.

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

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

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

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

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

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

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

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

Library of Congress Cataloging-in-Publication Data

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

p. cm.

Includes bibliographical references and index.

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

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

QA268.U545 2004

005.7'2—dc21

2003052158

Copyright © 1991–2003 by Unicode, Inc.

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

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

ISBN 0-321-18578-1

Text printed on recycled paper

1 2 3 4 5 6 7 8 9 10—CRW—0706050403

First printing, August 2003

Chapter 10

Southeast Asian Scripts

The following scripts are discussed in this chapter:

- Thai
- Lao
- Myanmar
- Khmer
- Tai Le
- Philippine scripts

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.

The four Philippine scripts included here operate on similar principles; each uses non-spacing vowel signs. In addition, the Tagalog script has a virama.

The Tai Le script is encoded alphabetically.

10.1 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. On the other hand, 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 left to right.

The Lao script is closely related to Thai, and the encoding principles described in this section apply to the Lao encoding as well.

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 the Indic scripts, each Thai letter is a consonant possessing an inherent vowel sound. Thai letters further feature inherent tones. The inherent vowel and tone can be modified by means 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 in-line glyphs placed either before (that is, to the left of) or 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) 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 10-1*.

Table 10-1. Glyph Positions in Thai Syllables

Syllable	Glyphs	Code Point Sequence
<i>ka</i>	กะ	U+0E01 U+0E30
<i>ka:</i>	กา	U+0E01 U+0E32
<i>ki</i>	กิ	U+0E01 U+0E34
<i>ki:</i>	กี	U+0E01 U+0E35
<i>ku</i>	กู	U+0E01 U+0E38
<i>ku:</i>	กุ	U+0E01 U+0E39
<i>ku'</i>	กิ๋	U+0E01 U+0E35
<i>ku':</i>	กิ๊	U+0E01 U+0E36
<i>ke</i>	กะะ	U+0E40 U+0E01 U+0E30
<i>ke:</i>	เก	U+0E40 U+0E01
<i>kae</i>	แกะะ	U+0E41 U+0E01 U+0E30
<i>kae:</i>	แก	U+0E41 U+0E01
<i>ko</i>	โกะ	U+0E42 U+0E01 U+0E30
<i>ko:</i>	โก	U+0E42 U+0E01
<i>ko'</i>	เกาะะ	U+0E40 U+0E01 U+0E32 U+0E30
<i>ko':</i>	กอ	U+0E01 U+0E2D
<i>koe</i>	เกอะะ	U+0E40 U+0E01 U+0E2D U+0E30
<i>koe:</i>	เกอ	U+0E40 U+0E01 U+0E2D
<i>kia</i>	เกีย	U+0E40 U+0E01 U+0E35 U+0E22
<i>ku'a</i>	เกือ	U+0E40 U+0E01 U+0E37 U+0E2D
<i>kua</i>	กิว	U+0E01 U+0E31 U+0E27
<i>kaw</i>	เกา	U+0E40 U+0E01 U+0E32
<i>koe:y</i>	เกย	U+0E40 U+0E01 U+0E22
<i>kay</i>	ไก	U+0E44 U+0E01
<i>kay</i>	ไ	U+0E43 U+0E01
<i>kam</i>	กำ	U+0E01 U+0E33
<i>kri</i>	กฤ	U+0E01 U+0E24

Thai Punctuation. Thai uses a variety of punctuation marks particular to this script. U+0E4F THAI CHARACTER FONGMAN is the Thai bullet, used to mark items in lists, or appearing 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 the 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.

Thai words are not separated by spaces. 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 word boundary indications 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 *Figure 15-1*.

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.

10.2 Lao

Lao: U+0E80–U+0EFF

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

Lao contains fewer letters than Thai because by 1960 it was simplified to be fairly phonemic, while Thai maintains many etymological spellings that are homonyms. Regular word spacing is not used in Lao; spaces separate phrases or sentences instead. The glyph placements for Lao syllables are summarized in *Table 10-2*.

Table 10-2. Glyph Positions in Lao Syllables

Syllable	Glyphs	Code Point Sequence
<i>ka</i>	ກະ	U+0E81 U+0EB0
<i>ka:</i>	ກາ	U+0E81 U+0EB2
<i>ki</i>	ກີ	U+0E81 U+0EB4
<i>ki:</i>	ກື	U+0E81 U+0EB5
<i>ku</i>	ກຸ	U+0E81 U+0EB8
<i>ku:</i>	ກູ	U+0E81 U+0EB9
<i>ku'</i>	ກື້	U+0E81 U+0EB5
<i>ku':</i>	ກືື	U+0E81 U+0EB6
<i>ke</i>	ເກະ	U+0EC0 U+0E81 U+0EB0
<i>ke:</i>	ເກ	U+0EC0 U+0E81
<i>kae</i>	ແກະ	U+0EC1 U+0E81 U+0EB0
<i>kae:</i>	ແກ	U+0EC1 U+0E81
<i>ko</i>	ໂກະ	U+0EC2 U+0E81 U+0EB0
<i>ko:</i>	ໂກ	U+0EC2 U+0E81
<i>ko'</i>	ເກາະ	U+0EC0 U+0E81 U+0EB2 U+0EB0
<i>ko':</i>	ກີ້	U+0E81 U+0ECD
<i>koe</i>	ເກີ້	U+0EC0 U+0E81 U+0EB4
<i>koe:</i>	ເກີ	U+0EC0 U+0E81 U+0EB5
<i>kia</i>	ເກີ້ວ, ເກຢ	U+0EC0 U+0E81 U+0EB1 U+0EBD, U+0EC0 U+0E81 U+0EA2
<i>ku'a</i>	ເກືອ	U+0EC0 U+0E81 U+0EB7 U+0EAD
<i>kua</i>	ກົວ	U+0E81 U+0EBB U+0EA7
<i>kaw</i>	ເກົາ	U+0EC0 U+0E81 U+0EBB U+0EB2
<i>koe:y</i>	ເກີ້ວ, ເກືຢ	U+0EC0 U+0E81 U+0EB5 U+0EBD, U+0EC0 U+0E81 U+0EB5 U+0EA2
<i>kay</i>	ໄກ	U+0EC4 U+0E81
<i>kay</i>	ໃກ	U+0EC3 U+0E81
<i>kam</i>	ກຳ	U+0E81 U+0EB3

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.

There are also two ligatures in the Unicode character encoding 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.

myanmar vowel sign ui

$$U+1000 \text{ က } ka + U+102F \text{ ျ } \text{vowel sign u} + U+102D \text{ ျ } \text{vowel sign i} \rightarrow \text{ကိ}$$

Encoding Subranges. The basic consonants, independent vowels, and dependent vowel signs required for writing the Myanmar language are encoded at the beginning of the Myanmar range. Extensions of each of these categories for use in writing other languages, such as Pali and Sanskrit, are appended at the end of the range. In between these two sets lie the script-specific signs, punctuation, and digits.

Conjunct and Medial Consonants. As in other Indic-derived scripts, conjunction of two consonant letters is indicated by the insertion of a virama U+1039 မြာမာ MYANMAR SIGN VIRAMA between them; it causes ligation or other rendered combination of the consonants, although the virama itself is not rendered visibly.

The conjunct form of U+1004 က MYANMAR LETTER NGA is rendered as a superscript sign called *kinzi*. *Kinzi* is encoded in logical order as a conjunct consonant *before* the syllable to which it applies; this is similar to the treatment of the Devanagari *ra*. (See *Section 9.1, Devanagari*, rule R2.) For example, *kinzi* applied to U+1000 က MYANMAR LETTER KA would be written via the following sequence:

$$U+1004 \text{ က } nga + U+1039 \text{ မြာမာ } \text{virama} + U+1000 \text{ က } ka \rightarrow \text{ကိနာ}$$

The Myanmar script traditionally distinguishes a set of subscript “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 treated as conjuncts; that is, they are coded using the virama. For example, the word *krwe* ကြွေ [kjwei] (“to drop off”) would be written via the following sequence:

$$U+1000 \text{ က } ka + U+1039 \text{ မြာမာ } \text{virama} + U+101B \text{ ျ } ra + U+1039 \text{ မြာမာ } \text{virama} \\ + U+101D \text{ ျ } wa + U+1031 \text{ ျ } \text{vowel sign e} \rightarrow \text{ကြွေ}$$

Explicit Virama. The virama U+1039 မြာမာ MYANMAR SIGN VIRAMA also participates in some common constructions where it appears as a *visible* sign, commonly termed *killer*. In this usage where it appears as a visible diacritic, U+1039 is followed by a U+200C ZERO WIDTH NON-JOINER, as with Devanagari (see *Figure 9-4*).

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 10-3*.

Table 10-3. Myanmar Syllabic Structure

Name	Encoding	Example
<i>kinzi</i>	<U+1004, U+1039>	ꨀ
<i>consonant</i>	[U+1000–U+1021]	ꨁ
<i>subscript consonant</i>	<U+1039, [U+1000–U+1019, U+101C, U+101E, U+1020, U+1021]>	ꨂ
<i>medial ya</i>	<U+1039, U+101A>	ꨃ
<i>medial ra</i>	<U+1039, U+101B>	ꨄ
<i>medial wa</i>	<U+1039, U+101D>	ꨅ
<i>medial ha</i>	<U+1039, U+101F>	ꨆ
<i>vowel sign e</i>	U+1031	ꨇ
<i>vowel sign u, uu</i>	[U+102F, U+1030]	ꨈ ꨉ
<i>vowel sign i, ii, ai</i>	[U+102D, U+102E, U+1032]	ꨊ ꨋ ꨌ
<i>vowel sign aa</i>	U+102C	ꨍ
<i>anusvara</i>	U+1036	ꨎ
<i>atha (killer)</i>	<U+1039, U+200C>	ꨏ
<i>dot below</i>	U+1037	ꨐ
<i>visarga</i>	U+1038	ꨑ

Note that U+1031 MYANMAR VOWEL SIGN E is encoded *after* its consonant (as in the earlier example), although in visual presentation it is reordered to appear *before* (to the left of) the consonant form.

Spacing. Myanmar does not use any whitespace between words. If word boundary indications 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.

10.4 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 transliterated, as well as 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 *akṣaa 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 *akṣaa crieng* (“slanted script”) and the *akṣaa muul* (“round script”). There is no fundamental structural difference between the two. The slanted script (in its “standing” variant) is chosen as representative in *Chapter 16, Code Charts*.

Principles of the 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 10-4*, 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 10-4. Independent Vowel Characters

Name	Independent Vowel	Qa with Vowel Sign
<i>i</i>	ឺ	ឺ, ឺ, ឺ
<i>ii</i>	ឺ	ឺ, ឺ
<i>u</i>	្ម	្ម, ្ម
<i>uk</i>	្ម	្មក
<i>uu</i>	្ម	្ម, ្ម
<i>uuv</i>	្ម	្ម
<i>ry</i>	្រ	្រ

Table 10-4. Independent Vowel Characters (Continued)

Name	Independent Vowel	Qa with Vowel Sign
<i>ryy</i>	រ្យ	រ្យ
<i>ly</i>	ល្យ	ល្យ
<i>lyy</i>	ល្យ្យ	ល្យ្យ
<i>e</i>	ឯ	អេ, ែ
<i>ai</i>	ឿ	ៃ
<i>oo</i>	ឺ, ឺ	អោ
<i>au</i>	ឺ	អោ

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:

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

លក្ខី *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 is uncommon.

Examples of subscript consonant signs for a closing consonant:

ទាំង *to + a + nikahit + coeng + ngo* [tɛəŋ] “both” (= ទាំង) (≠ *ទាំង [tɲəŋ])

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 uncommon, it is used in one very common word, meaning “to give.”

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

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

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

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

Examples of subscript independent vowel signs:

ផ្អែម *pha + coeng + qe + mo* [pʰʔaem] “sweet” (= ផ្អែម *pha + coeng + qa + ae + mo*)

ហ្វូទ័យ *ha + coeng + ry + to + samyok sannya + yo* [harutey] “heart”
(*royal*) (= ហ្វូទ័យ *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, though 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 10-5*. 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, there are two “shifter” signs. 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 10-5. Two Registers of Khmer Consonants

Row	First Register	Second Register
1	កី <i>ka</i> [kɔ:] “neck”	កី <i>ko</i> [kɔ:] “mute”
2	រី <i>ro + muusikatoan</i> [rɔ:] “small saw”	រី <i>ro</i> [rɔ:] “fence (in the water)”
3	សីកី <i>sa + ka</i> [sɔ:k] “to peel, to shed one’s skin”	សីកី <i>sa + triisap + ka</i> [sɔ:k] “to insert”
4	បីកី <i>ba + ka</i> [bɔ:k] “to return”	*បីកី <i>ba + triisap + ka</i> [bɔ:k]
5	បីមី <i>ba + muusikatoan + mo</i> [pɔ:m] “blockhouse”	ពីមី <i>po + mo</i> [pɔ:m] “to put into the mouth”
6	ក្តីរី <i>ka + u + ro</i> [ko:] “to stir”	ក្តីរី <i>ko + u + ro</i> [ku:] “to sketch”

Examples of dependent vowel signs ending with [m]:

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

ពី *po + aa + nikahit* [pɔəm] “to carry in the beak” (compare ពីមី *po + aa + mo* [pɔəm] “mouth of a river”)

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 them is represented by the sequence of two characters: a special control character (U+17D2 KHMER SIGN COENG) and a corresponding consonant character. This is analogous to the virama model employed for representing conjuncts in other related scripts. Subscripted independent vowels are encoded in the same manner. Because the *coeng sign* character does not exist as a letter or sign in the Khmer script, the Unicode model departs from the ordinary way that Khmer is conceived of and taught to native Khmer speakers. Consequently, the encoding may not be intuitive to a native user of the Khmer writing system, although it is able to represent Khmer correctly.

U+17D2 KHMER SIGN COENG is not actually a *coeng* but a *coeng* generator, because *coeng* in Khmer refers to the subscript consonant sign. To aid Khmer script users, a listing of typical Khmer subscript consonant letters has been provided in *Table 10-6* 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 10-8*.

Table 10-6. Khmer Subscript Consonant Signs

Glyph	Code	Name
្ក	17D2 1780	khmer consonant sign coeng ka
្ខ	17D2 1781	khmer consonant sign coeng kha
្គ	17D2 1782	khmer consonant sign coeng ko
្ឃ	17D2 1783	khmer consonant sign coeng kho
្ង	17D2 1784	khmer consonant sign coeng ngo
្ច	17D2 1785	khmer consonant sign coeng ca
្ឆ	17D2 1786	khmer consonant sign coeng cha
្ជ	17D2 1787	khmer consonant sign coeng co
្ឈ	17D2 1788	khmer consonant sign coeng cho
្ញ	17D2 1789	khmer consonant sign coeng nyo
្ដ	17D2 178A	khmer consonant sign coeng da
្ឋ	17D2 178B	khmer consonant sign coeng tha
្ឌ	17D2 178C	khmer consonant sign coeng do
្ឍ	17D2 178D	khmer consonant sign coeng tho
្ណ	17D2 178E	khmer consonant sign coeng na
្ត	17D2 178F	khmer consonant sign coeng ta

Table 10-6. Khmer Subscript Consonant Signs (Continued)

Glyph	Code	Name
	17D2 1790	khmer consonant sign coeng tha
	17D2 1791	khmer consonant sign coeng to
	17D2 1792	khmer consonant sign coeng tho
	17D2 1793	khmer consonant sign coeng no
	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
	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 attested 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 [ɔ̃m], not [a:m]. The sequences used for these combinations are shown in Table 10-7.

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 10-1*.

Figure 10-1. Common Ligatures

ក *ka* + ា *aa* + រ *ro* = កាវ [ka:] “job”

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

ប *ba* + ៅ *au* = បៅ [baw] “to suck”

ម *mo* + ្យ *coeng sa* + ៅ *au* = ម្យៅ [msaw] “powder”

ស *sa* + ង *ngo* + ្យ *coeng kha* + ្យ *coeng yo* + ា *aa* = សង្យា [sɔŋkʰya:] “counting”

Multiple Glyphs. A single character may assume different forms according to context. For example, a part of the glyph for *nyo* is omitted when a subscript consonant sign is attached. The implementation must render the correct glyph according to context. *Coeng nyo* also changes its shape when it is attached to *nyo*. The correct glyph for the sequence <U+17D2 KHMER SIGN COENG, U+1789 KHMER LETTER NYO> is rendered according to context, as shown in *Figure 10-2*. 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 10-2*.

Figure 10-2. Common Multiple Forms

ញញឹម *nyo + nyo + y + mo* [nɔ̃nɔ̃m] “to smile”

មីមីម *ca + i + nyo + coeng + ca + oe + mo* [cẽcãm] “eyebrow”

ស្តប់ *sa + coeng nyo + ba + bantoc* [sɔ̃ɔ̃p] “to respect”

កញ្ញា *ka + nyo + coeng + nyo + aa* [kãna:] “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.

The use of U+17A3 KHMER INDEPENDENT VOWEL QAA and U+17A4 KHMER INDEPENDENT VOWEL QAA is 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.

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:

១៤៤១ *khan + ba + e + khan*
 -៤៤- *en dash + ba + e + en dash*
 ១ ៤៤ ១ *khan + lo + khan*
 -៤៤- *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 the zero width non-joiner

O is any other sign

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

ខ្ញុំ *kha + coeng + nyo + u + nikahit* [kʰnom] “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 10-3*.

Figure 10-3. Examples of Syllabic Order

្រ to + e [tè:] “much”
 ្រើន ca + coeng + ro + oe + no [craən] “much”
 ស្រ្តាម sa + ngo + coeng + ko + coeng + ro + aa + mo [sɔŋkrèəm] “war”
 ហើយ ha + oe + coeng + yo [haəi] “already”
 ស្រ្តា sa + nyo + coeng + nyo + aa [saŋna:] “sign”
 ស្រី sa + triisap + ii [si:] “eat”
 ប៊ី 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.

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.

្រៃ mo + muusikatoan + coeng + ngo + ai [mɿjai] “one day”
 ្រៃតិ្រ 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 where 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:

ប៊ី ្រែ រ ba + ZW + triisap + ii + yo + ae + ro [biyè:] “beer”
 ប្រតិ្រងអ៊ី្រ ba + coeng + ro + ta + yy + ngo + qa + ZW + triisap + y + reah-
 muk [prɔtə:ŋtuh] “urgent, too busy”
 ប្រតិ្រងអ៊ី្រ ba + coeng + ro + ta + yy + ngo + qa + triisap + y + reahmuk

Ligature Control. In the *aska 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 10-4*.

Figure 10-4. Ligation in *Muul* Style

វិទូ *vo + i + to + uu* [vitu:] “savant” (*akxaa crieng* font)

វិទូ, វិទូ *vo + i + to + uu* [vitu:] “savant” (ligature dependent on the *muul* font)

វិទូ *vo + [ZW] + i + to + uu* [vitu:] “savant” (*[ZW]*) to prevent the ligature in a *muul* font)

វិទូ *vo + [ZW] + i + to + uu* [vitu:] “savant” (*[ZW]*) to request the ligature in a *muul* font)

Spacing. Khmer does not use whitespace between words, although it does use whitespace between clauses and between parts of a name. If word boundary indications are desired—for example, 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 *Figure 15-1*.

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 and U+19F0 KHMER SYMBOL TUTEYASAT represent the first and second of August, respectively, in a leap year. 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.

10.5 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, and spacing tone marks were introduced to replace the combining diacritics. The Unicode encoding of Tai Le handles both orthographies.

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, တၢ်လၢ [tai² lə⁶] (in older orthography တၢ်လဲ [təi² lə⁶]). Tai Le orthography is 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 ᩁ, ᩃ, ᩅ, ᩇ, ᩉ or ᩊ, it is displayed to the right of the letter, as shown in *Table 10-9*.

Table 10-9. Tai Le Tone Marks

Syllable	New Orthography	Old Orthography
<i>ta</i>	တ	တ
<i>ta</i> ²	တၢ	တံ
<i>ta</i> ³	တe	တံ
<i>ta</i> ⁴	တၢ	တံ
<i>ta</i> ⁵	တၢ	တံ
<i>ta</i> ⁶	တc	တံ
<i>ti</i>	တ	တ
<i>ti</i> ²	တၢၢ	တံ
<i>ti</i> ³	တeၢ	တံ
<i>ti</i> ⁴	တၢၢ	တံ
<i>ti</i> ⁵	တၢၢ	တံ
<i>ti</i> ⁶	တcၢ	တံ

Digits. In China, European digits (U+0030..U+0039) are mainly used, though Myanmar digits, (U+1040..U+1049) are also used with slight glyph variants, as shown in *Table 10-10*.

Table 10-10. 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 [and]; in some fonts, the baseline for punctuation is the depth of those characters.

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.

Directionality. The Philippine scripts are read from left to right in horizontal lines running from top to bottom. They may be written or carved either in that manner or in vertical lines running from bottom to top, moving from left to right. In the latter case, the letters are written sideways so they may be read horizontally. This method of writing is probably due to the medium and writing implements used. Text is often scratched with a sharp instrument onto beaten strips of bamboo, which are held pointing away from the body and worked from the proximal to distal ends, in columns from left to right.

Rendering. In Tagalog and Tagbanwa, the vowel signs simply rest over or under the consonants. In Hanunóo and Buhid, however, special ligatures are often formed, as shown in Table 10-11.

Table 10-11. Hanunóo and Buhid Vowel Sign Combinations

Hanunóo			Buhid		
<i>x</i>	<i>x + \bar{o}</i>	<i>x + \bar{u}</i>	<i>x</i>	<i>x + \bar{o}</i>	<i>x + \bar{u}</i>
ϕ	ϕ̄	ϕ̇	≡	+	≡
Ϡ	Ϡ̄	Ϡ̇	Ϡ	Ϡ̄	Ϡ̇
ϡ	ϡ̄	ϡ̇	ϡ	ϡ̄	ϡ̇
Ϣ	Ϣ̄	Ϣ̇	Ϣ	Ϣ̄	Ϣ̇
ϣ	ϣ̄	ϣ̇	ϣ	ϣ̄	ϣ̇
Ϥ	Ϥ̄	Ϥ̇	Ϥ	Ϥ̄	Ϥ̇
ϥ	ϥ̄	ϥ̇	ϥ	ϥ̄	ϥ̇
Ϧ	Ϧ̄	Ϧ̇	Ϧ	Ϧ̄	Ϧ̇
ϧ	ϧ̄	ϧ̇	ϧ	ϧ̄	ϧ̇
Ϩ	Ϩ̄	Ϩ̇	Ϩ	Ϩ̄	Ϩ̇
ϩ	ϩ̄	ϩ̇	ϩ	ϩ̄	ϩ̇
Ϫ	Ϫ̄	Ϫ̇	Ϫ	Ϫ̄	Ϫ̇
ϫ	ϫ̄	ϫ̇	ϫ	ϫ̄	ϫ̇
Ϭ	Ϭ̄	Ϭ̇	Ϭ	Ϭ̄	Ϭ̇
ϭ	ϭ̄	ϭ̇	ϭ	ϭ̄	ϭ̇
Ϯ	Ϯ̄	Ϯ̇	Ϯ	Ϯ̄	Ϯ̇
ϯ	ϯ̄	ϯ̇	ϯ	ϯ̄	ϯ̇
ϰ	ϰ̄	ϰ̇	ϰ	ϰ̄	ϰ̇
ϱ	ϱ̄	ϱ̇	ϱ	ϱ̄	ϱ̇
ϲ	ϲ̄	ϲ̇	ϲ	ϲ̄	ϲ̇
ϳ	ϳ̄	ϳ̇	ϳ	ϳ̄	ϳ̇
ϴ	ϴ̄	ϴ̇	ϴ	ϴ̄	ϴ̇
ϵ	ϵ̄	ϵ̇	ϵ	ϵ̄	ϵ̇
϶	϶̄	϶̇	϶	϶̄	϶̇
Ϸ	Ϸ̄	Ϸ̇	Ϸ	Ϸ̄	Ϸ̇
ϸ	ϸ̄	ϸ̇	ϸ	ϸ̄	ϸ̇
Ϲ	Ϲ̄	Ϲ̇	Ϲ	Ϲ̄	Ϲ̇
Ϻ	Ϻ̄	Ϻ̇	Ϻ	Ϻ̄	Ϻ̇
ϻ	ϻ̄	ϻ̇	ϻ	ϻ̄	ϻ̇
ϼ	ϼ̄	ϼ̇	ϼ	ϼ̄	ϼ̇
Ͻ	Ͻ̄	Ͻ̇	Ͻ	Ͻ̄	Ͻ̇
Ͼ	Ͼ̄	Ͼ̇	Ͼ	Ͼ̄	Ͼ̇
Ͽ	Ͽ̄	Ͽ̇	Ͽ	Ͽ̄	Ͽ̇
Ͽ	Ͽ̄	Ͽ̇	Ͽ	Ͽ̄	Ͽ̇

Punctuation. Punctuation has been unified for the Philippine scripts. In the Hanunóo block, U+1735 PHILIPPINE SINGLE PUNCTUATION and U+1736 PHILIPPINE DOUBLE PUNCTUATION are encoded. Tagalog makes use of only the latter; Hanunóo, Buhid, and Tagbanwa make use of both of them.