

Electronic Edition

This file is part of the electronic edition of *The Unicode Standard, Version 5.0*, provided for online access, content searching, and accessibility. It may not be printed. Bookmarks linking to specific chapters or sections of the whole Unicode Standard are available at

<http://www.unicode.org/versions/Unicode5.0.0/bookmarks.html>

Purchasing the Book

For convenient access to the full text of the standard as a useful reference book, we recommend purchasing the printed version. The book is available from the Unicode Consortium, the publisher, and booksellers. Purchase of the standard in book format contributes to the ongoing work of the Unicode Consortium. Details about the book publication and ordering information may be found at

<http://www.unicode.org/book/aboutbook.html>

Joining Unicode

You or your organization may benefit by joining the Unicode Consortium: for more information, see [Joining the Unicode Consortium](http://www.unicode.org/consortium/join.html) at

<http://www.unicode.org/consortium/join.html>

This PDF file is an excerpt from *The Unicode Standard, Version 5.0*, issued by the Unicode Consortium and published by Addison-Wesley. The material has been modified slightly for this electronic 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/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 the publisher was aware of a trademark claim, the designations have been printed with initial capital letters or in all capitals.

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 the 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, www.mehallo.com

The publisher offers excellent discounts on this book when ordered in quantity for bulk purchases or special sales, which may include electronic versions and/or custom covers and content particular to your business, training goals, marketing focus, and branding interests. For more information, please contact U.S. Corporate and Government Sales, (800) 382-3419, corpsales@pearsoned.com. For sales outside the United States please contact International Sales, international@pearsoned.com

Visit us on the Web: www.awprofessional.com

Library of Congress Cataloging-in-Publication Data

The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 5.0.
p. cm.

Includes bibliographical references and index.

ISBN 0-321-48091-0 (hardcover : alk. paper)

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

II. Unicode Consortium.

QA268.U545 2007

005.7'22—dc22

2006023526

Copyright © 1991–2007 Unicode, Inc.

All rights reserved. Printed in the United States of America. This publication is protected by copyright, and permission must be obtained from the publisher prior to any prohibited reproduction, storage in a retrieval system, or transmission in any form or by any means, electronic, mechanical, photocopying, recording, or likewise. For information regarding permissions, write to Pearson Education, Inc., Rights and Contracts Department, 75 Arlington Street, Suite 300, Boston, MA 02116. Fax: (617) 848-7047

ISBN 0-321-48091-0

Text printed in the United States on recycled paper at Courier in Westford, Massachusetts.

First printing, October 2006

Chapter 11

Southeast Asian Scripts

The following scripts are discussed in this chapter:

<i>Thai</i>	<i>Khmer</i>	<i>Philippine scripts</i>
<i>Lao</i>	<i>Tai Le</i>	<i>Buginese</i>
<i>Myanmar</i>	<i>New Tai Lue</i>	<i>Balinese</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.

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 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.

Buginese and Balinese are scripts of Indonesia, and both are ultimately related to scripts of South Asia. Buginese is used in Sulawesi; Balinese is used on the island of Bali.

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) 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>	ก๖	0E01 0E30
<i>ka:</i>	ก๗	0E01 0E32
<i>ki</i>	ก๘	0E01 0E34
<i>ki:</i>	ก๙	0E01 0E35
<i>ku</i>	ก๑	0E01 0E38
<i>ku:</i>	ก๒	0E01 0E39
<i>ku'</i>	ก๓	0E01 0E36
<i>ku':</i>	ก๔	0E01 0E37
<i>ke</i>	เก๖	0E40 0E01 0E30
<i>ke:</i>	เก๗	0E40 0E01
<i>kae</i>	แก๖	0E41 0E01 0E30
<i>kae:</i>	แก๗	0E41 0E01
<i>ko</i>	โก๖	0E42 0E01 0E30

Table 11-1. Glyph Positions in Thai Syllables (Continued)

Syllable	Glyphs	Code Point Sequence
<i>ko:</i>	โไก	0E42 0E01
<i>ko'</i>	เกาะะ	0E40 0E01 0E32 0E30
<i>ko':</i>	กอ	0E01 0E2D
<i>koe</i>	เกอะะ	0E40 0E01 0E2D 0E30
<i>koe:</i>	เกอ	0E40 0E01 0E2D
<i>kia</i>	เกีย	0E40 0E01 0E35 0E22
<i>ku'a</i>	เกือ	0E40 0E01 0E37 0E2D
<i>kua</i>	กัว	0E01 0E31 0E27
<i>kaw</i>	เกา	0E40 0E01 0E32
<i>koe:y</i>	เกย	0E40 0E01 0E22
<i>kay</i>	ไไก	0E44 0E01
<i>kay</i>	ไก	0E43 0E01
<i>kam</i>	กำ	0E01 0E33
<i>kri</i>	กฤ	0E01 0E24

Rendering of Thai Combining Marks. The combining classes assigned to tone marks (107) and to other combining characters displayed above (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.

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 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 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>	ກີ'	0E81 0EB6
<i>ku':</i>	ກີ'	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>	ກີ'	0E81 0ECD
<i>koe</i>	ເກີ	0EC0 0E81 0EB4
<i>koe:</i>	ເກີ	0EC0 0E81 0EB5
<i>kia</i>	ເກີ້ຍ 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

Table 11-2. Glyph Positions in Lao Syllables (Continued)

Syllable	Glyphs	Code Point Sequence
<i>koe:y</i>	ເກີ້ໄ ເກີ້ຢ	0EC0 0E81 0EB5 0EBD 0EC0 0E81 0EB5 0EA2
<i>kay</i>	ໄກ	0EC4 0E81
<i>kay</i>	ໄກ່	0EC3 0E81
<i>kam</i>	ກຳ	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 combining classes assigned to tone marks (122) and to other combining characters displayed above (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 Shan and Mon, 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.

Because of its Brahmi origins, the Myanmar script shares the structural features of its Indic relatives: consonant symbols include an inherent “a” vowel; various signs are attached to a consonant to indicate a different vowel; ligatures and conjuncts are used to indicate consonant clusters; and the overall writing direction is from left to right. Thus, despite great differences in appearance and detail, the Myanmar script follows the same basic principles as, for example, Devanagari.

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. Even some single characters, such as U+102C  MYANMAR VOWEL SIGN AA, may assume variant forms (for example, ) depending on the other characters with which they combine. Conversely, 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 examples of Myanmar letters represented by combining character sequences:

myanmar vowel sign o

U+1000 က ka + U+1031 ဧ vowel sign e + U+102C ဣ vowel sign aa →
ကော kō

myanmar vowel sign au

U+1000 က ka + U+1031 ဧ vowel sign e + U+102C ဣ vowel sign aa +
U+1039 ဴ virama + U+200C [ZW] → ကော့ kau

myanmar vowel sign ui

U+1000 က ka + U+102F ့ vowel sign u + U+102D ဲ vowel sign i → ကူ
kui

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 ဴ 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 င 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 င nga + U+1039 ဴ virama + U+1000 က ka → က်း ñka

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* ကြ [kjwei] (“to drop off”) would be written via the following sequence:

U+1000 က ka + U+1039 ဴ virama + U+101B ကြ ra + U+1039 ဴ virama
+ U+101D ဝ wa + U+1031 ဧ vowel sign e → ကြဝ္ဋ krwe

Explicit Virama. The virama U+1039 ဴ 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 9-3*).

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

Name	Encoding	Example
<i>kinzi</i>	<U+1004, U+1039>	ꠊ
<i>consonant</i>	[U+1000..U+1021]	က
<i>subscript consonant</i>	<U+1039, [U+1000..U+1019, U+101C, U+101E, U+1020, U+1021]>	ꠋ
<i>medial ya</i>	<U+1039, U+101A>	ꠌ
<i>medial ra</i>	<U+1039, U+101B>	ꠍ
<i>medial wa</i>	<U+1039, U+101D>	ꠎ
<i>medial ha</i>	<U+1039, U+101F>	ꠏ
<i>vowel sign e</i>	U+1031	ꠐ
<i>vowel sign u, uu</i>	[U+102F, U+1030]	ꠑ ꠒ
<i>vowel sign i, ii, ai</i>	[U+102D, U+102E, U+1032]	ꠓ ꠔ ꠕ
<i>vowel sign aa</i>	U+102C	ꠖ
<i>anusvara</i>	U+1036	ꠗ
<i>atha (killer)</i>	<U+1039, U+200C>	ꠘ
<i>dot below</i>	U+1037	ꠙ
<i>visarga</i>	U+1038	ꠚ

U+1031 ꠐ 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.

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.

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 *akxaa 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 *akxaa crieng* (“slanted script”) and the *akxaa 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 17, 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 11-4*, 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.

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 𑄀 KHMER LETTER LA does not have a corresponding subscript consonant sign in standard Khmer, but does have a subscript in the Khmer script used in Thailand.

Table 11-4. Independent Khmer Vowel Characters

Name	Independent Vowel	Qa with Vowel Sign
<i>i</i>	ឺ	អ៊ិ, អ៊ិ, អ៊ិ
<i>ii</i>	ឺ្គ	អ៊ិ្គ, អ៊ិ្គ
<i>u</i>	ឺ	អ៊ុ, អ៊ុ
<i>uk</i>	ឺ្ក	អ៊ុក
<i>uu</i>	ឺ្គ	អ៊ុ្គ, អ៊ុ្គ
<i>uuv</i>	ឺ្គ្ក	អ៊ុ្គវ
<i>ry</i>	ឺ្យ	អ៊ិ្យ
<i>ryy</i>	ឺ្យ្គ	អ៊ិ្យ្គ
<i>ly</i>	ឺ្រ	អ៊ិ្រ
<i>lyy</i>	ឺ្រ្គ	អ៊ិ្រ្គ
<i>e</i>	ឺ	អ៊ែ, អ៊ែ
<i>ai</i>	ឺ	អ៊ែ
<i>oo</i>	ឺ, ឺ	អ៊ោ
<i>au</i>	ឺ	អ៊ោ

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:

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

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

កាហ្វេ *ka* + *aa* + *ha* + *coeng* + *vo* + *e* [ka:fe:] “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) (\neq *ទាំង [tɛŋəəm])

ហើយ *ha + oe + coeng + yo* [haɛi] “already” (= ហើយ) (\neq *ហើយ [hyaɛ])

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:

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

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

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* [pʰʔaem] “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 11-5*. 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, two “shifter” signs are available. U+17C9 KHMER SIGN MUUSIKATOAN converts a consonant character and sign from the second to the first register, while U+17CA KHMER SIGN TRIISAP converts a consonant from the first register to the second (rows 2–4). To represent *pa*, however, *muusikatoan* is attached not to *po* but to *ba*, in an exceptional use (row 5). The phonetic value of a dependent vowel sign may also change depending on the context of the consonant(s) to which it is attached (row 6).

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

Table 11-5. Two Registers of Khmer Consonants

Row	First Register	Second Register
1	កំ ka [kɔː] “neck”	កំ ko [kɔː] “mute”
2	រំ ro + muusikatoan [rɔː] “small saw”	រំ ro [rɔː] “fence (in the water)”
3	សំកំ sa + ka [sɔːk] “to peel, to shed one’s skin”	សំកំ sa + triisap + ka [sɔːk] “to insert”
4	បំកំ ba + ka [bɔːk] “to return”	*បំកំ ba + triisap + ka [bɔːk]
5	បំមំ ba + muusikatoan + mo [pɔːm] “blockhouse”	បំមំ po + mo [pɔːm] “to put into the mouth”
6	កំរំ ka + u + ro [koː] “to stir”	កំរំ ko + u + ro [kuː] “to sketch”

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



U+17D2  KHMER SIGN COENG is not actually a *coeng* but a *coeng* generator, because *coeng* in Khmer refers to the subscript consonant sign. The glyph for U+17D2  KHMER SIGN COENG shown in the code charts is arbitrary and is not actually rendered directly; the dotted box around the glyph indicates that special rendering is required. To aid Khmer script users, a listing of typical Khmer subscript consonant letters has been provided in *Table 11-6* 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 11-8*.

Table 11-6. Khmer Subscript Consonant Signs




Glyph	Code	Name
	17D2 1780	khmer consonant sign coeng ka
	17D2 1781	khmer consonant sign coeng kha
	17D2 1782	khmer consonant sign coeng ko

Table 11-6. Khmer Subscript Consonant Signs (Continued)







Glyph	Code	Name
	17D2 1783	khmer consonant sign coeng kho
	17D2 1784	khmer consonant sign coeng ngo
	17D2 1785	khmer consonant sign coeng ca
	17D2 1786	khmer consonant sign coeng cha
	17D2 1787	khmer consonant sign coeng co
	17D2 1788	khmer consonant sign coeng cho
	17D2 1789	khmer consonant sign coeng nyo
	17D2 178A	khmer consonant sign coeng da
	17D2 178B	khmer consonant sign coeng ttha
	17D2 178C	khmer consonant sign coeng do
	17D2 178D	khmer consonant sign coeng ttho
	17D2 178E	khmer consonant sign coeng na
	17D2 178F	khmer consonant sign coeng ta
	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

Table 11-6. Khmer Subscript Consonant Signs (Continued)



Glyph	Code	Name
	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 [ɔ̃m], not [a:m]. The sequences used for these combinations are shown in Table 11-7.

Table 11-7. 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:

 *da* + *nikahit* [dɔ̃m] “to pound” (compare  *da* + *mo* [dɔ:m] “nectar”)

 *po* + *aa* + *nikahit* [pɔ̃əm] “to carry in the beak” (compare  *po* + *aa* + *mo* [pɔ̃əm] “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 11-8.

Table 11-8. 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:

រាជវិទ្យា *ro + aa + co + ro + coeng + sa + ii* [rèəɾsei] “king hermit”

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

ព័ត៌មាន *po + ta + robat + mo + aa + no* [pò:dəmèən] “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. In this respect it is similar to U+17D1 KHMER SIGN VIRIAM. U+17CB KHMER SIGN BANTOC shortens the vowel sound of the previous ortho-

graphic 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

កំ ka + ា aa + រ ro = កំរ [ka:] “job”
 បា ba + ា aa = បា [ba:] “father, male of an animal”; used to prevent confusion with ហា ha
 បា ba + ៅ au = បៅ [baw] “to suck”
 មូ mo + ង coeng sa + ៅ au = មូង [msaw] “powder”
 សំ sa + ង ngo + ង coeng kha + ង coeng yo + ា aa = សំងួង [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 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* [ɲɔŋɲɔm] “to smile”
 មីឆើម *ca + i + nyo + coeng + ca + oe + mo* [ceŋcaəm] “eyebrow”
 ស្ងប់ *sa + coeng nyo + ba + bantoc* [sɔɔp] “to respect”
 កញ្ញា *ka + nyo + coeng + nyo + aa* [kaŋna:] “girl, Miss, September”
 ឲ្យ *qoo-2 + coeng + yo* (= ឲ្យ *qoo-1 + coeng + yo*) [faoi] “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:

្ក្ក្ក្ក *khan + ba + e + khan*
 -្ក្ក្ក- *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.

ខ្ញុំ *kha + coeng + nyo + u + nikahit* [kʰŋom] “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

ទេ *to + e* [tè:] “much”

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

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

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

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

ស៊ី *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, 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.

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

ម្ល៉ៃតៗ *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:

ប៊ែរ $ba + \text{[ZW]} + triisap + ii + yo + ae + ro$ [biyè:] “beer”
 ប្រតីងអីន $ba + coeng + ro + ta + yy + ngo + qa + \text{[ZW]} + triisap + y + reah-$
 muk [prətə:ŋʔuh] “urgent, too busy”
 ប្រតីងអីន $ba + coeng + ro + ta + yy + ngo + qa + triisap + y + reahmuk$

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 11-4* for the word “savant,” pronounced [vitu:].

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

វិទូ	$vo + i + to + uu$	(<i>aksaa crieng</i> font)
វិទូ, វិទូ	$vo + i + to + uu$	(ligature dependent on the <i>muul</i> font)
វិទូ	$vo + \text{[ZW]} + i + to + uu$	([ZW] to prevent the ligature in a <i>muul</i> font)
វិទូ	$vo + \text{[ZW]} + i + to + uu$	([ZW] 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 *Figure 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 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.

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, $\text{တၢ်လၢ} \text{ၤ} \text{လၢ} \text{ၤ}$ [tai² lə⁶] (in older orthography $\text{တၢ်} \text{ၤ} \text{လၢ} \text{ၤ}$). 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 ၢ , ၣ , ၤ , ၥ , ၦ or ၧ , it is displayed to the right of the letter, as shown in *Table 11-9*.

Table 11-9. Tai Le Tone Marks

Syllable	New Orthography	Old Orthography
<i>ta</i>	တ	တ
<i>ta</i> ²	တၢ	တ̃
<i>ta</i> ³	တe	တ̂
<i>ta</i> ⁴	တၢ	တ̄
<i>ta</i> ⁵	တၣ	တ̆
<i>ta</i> ⁶	တC	တ̇
<i>ti</i>	တိ	တိ
<i>ti</i> ²	တိၢ	တိ̃
<i>ti</i> ³	တိe	တိ̂
<i>ti</i> ⁴	တိၢ	တိ̄
<i>ti</i> ⁵	တိၣ	တိ̆
<i>ti</i> ⁶	တိC	တိ̇

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 11-10*.

Table 11-10. 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
0 1 2 3 4 5 6 7 8 9	0 1 2 3 4 5 6 7 8 9

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

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 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 Lanna 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 11-11*.

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.

Table 11-11. New Tai Lue Vowel Placement

ꨀ	ka	+	ꨀ	e	+	ꨀ	t1	→	ꨀꨀꨀ	[ke: ²]				
ꨀ	ka	+	ꨀ	e	+	ꨀ	i	→	ꨀꨀꨀ	[kæ: ¹]				
ꨀ	ka	+	ꨀ	e	+	ꨀꨀ	iy	→	ꨀꨀꨀꨀ	[kæi ¹]				
ꨀ	ka	+	ꨀ	e	+	ꨀꨀ	iy	+	ꨀꨀ	t1	→	ꨀꨀꨀꨀꨀ	[kæi ²]	
ꨀ	ka	+	ꨀ	e	+	ꨀꨀ	iy	+	ꨀ	e	t2	→	ꨀꨀꨀꨀꨀ	[kæi ³]

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 11-12*.

Table 11-12. New Tai Lue Registers and Tones

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

11.7 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.

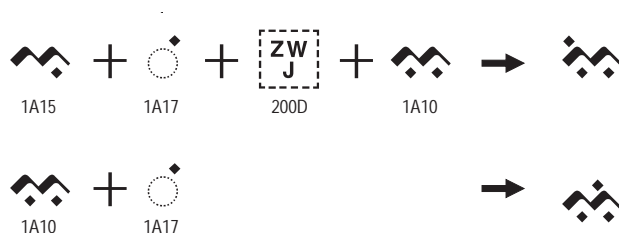
Written accounts of the Tagalog script by Spanish missionaries and documents in Tagalog date from the mid-1500s. The first book in this script was printed in Manila in 1593. While the Tagalog script was used to write Tagalog, Bisaya, Ilocano, and other languages, it fell out

the fourteenth century CE. Buginese bears some affinity to Tagalog and, like Tagalog, does not traditionally record final consonants. The Buginese language, an Austronesian language 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.

11.9 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 11-14* shows the base consonants and their conjunct forms.

Table 11-14. Balinese Base Consonants and Conjunct Forms

Consonant	Base Form	Conjunct Form
<i>ka</i>	ꦏ	ꦏꦲ
<i>kha</i>	ꦏꦲꦲ	ꦏꦲꦲꦲ
<i>ga</i>	ꦒ	ꦒꦲ
<i>gha</i>	ꦒꦲꦲ	ꦒꦲꦲꦲ
<i>nga</i>	ꦒꦤ	ꦒꦤꦲ
<i>ca</i>	ꦕ	ꦕꦲ
<i>cha</i>	ꦕꦲꦲ	ꦕꦲꦲꦲ
<i>ja</i>	ꦗ	ꦗꦲ
<i>jha</i>	ꦗꦲꦲꦲ	ꦗꦲꦲꦲꦲ
<i>nya</i>	ꦚꦤ	ꦚꦤꦲ
<i>tta</i>	ꦠꦠ	ꦠꦠꦲ
<i>ttha</i>	ꦠꦠꦲꦲ	ꦠꦠꦲꦲꦲ
<i>dda</i>	ꦢꦢ	ꦢꦢꦲ
<i>ddha</i>	ꦢꦢꦲꦲ	ꦢꦢꦲꦲꦲ

Table 11-14. Balinese Base Consonants and Conjunct Forms (Continued)

Consonant	Base Form	Conjunct Form
<i>nna</i>	ꦤꦤ	ꦤꦤꦺ
<i>ta</i>	ꦠ	ꦠꦺ
<i>tha</i>	ꦠꦲ	ꦠꦲꦺ
<i>da</i>	ꦢ	ꦢꦺ
<i>dha</i>	ꦢꦲ	ꦢꦲꦺ
<i>na</i>	ꦤ	ꦤꦺ
<i>pa</i>	ꦥ	ꦥꦺ
<i>pha</i>	ꦥꦲ	ꦥꦲꦺ
<i>ba</i>	ꦧ	ꦧꦺ
<i>bha</i>	ꦧꦲ	ꦧꦲꦺ
<i>ma</i>	ꦩ	ꦩꦺ
<i>ya</i>	ꦪ	ꦪꦺ
<i>ra</i>	ꦫ	ꦫꦺ
<i>la</i>	ꦭ	ꦭꦺ
<i>wa</i>	ꦮ	ꦮꦺ
<i>ssa</i>	ꦱꦱ	ꦱꦱꦺ
<i>sha</i>	ꦱꦲ	ꦱꦲꦺ
<i>sa</i>	ꦱ	ꦱꦺ
<i>ha</i>	ꦲ	ꦲꦺ
<i>r</i>	ꦫꦺ	ꦫꦺꦺ

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 11-15*.

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

Table 11-15. Sasak Extensions for Balinese

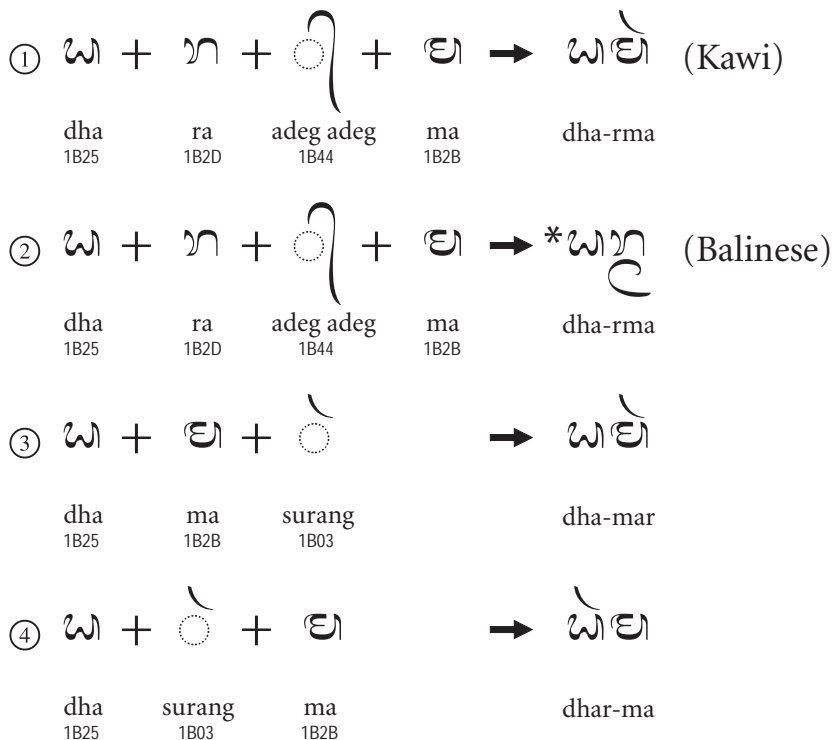
Consonant	Base Form	Conjunct Form
<i>kaf</i>	ꦏꦲ	ꦏꦲꦠ
<i>khot</i>	ꦏꦲꦲ	ꦏꦲꦲꦠ
<i>tzir</i>	ꦠꦶꦂ	ꦠꦶꦂꦠ
<i>ef</i>	ꦺꦴ	ꦺꦴꦠ
<i>ve</i>	ꦺꦴ	ꦺꦴꦠꦺꦴ
<i>zal</i>	ꦠꦶꦲ	ꦠꦶꦲꦠ
<i>asyura</i>	ꦱꦸꦫ	ꦱꦸꦫꦠ

adeg adeg. In Sasak, the vowel letter *akara* can be followed by an explicit *adeg adeg* ^{ꦠꦺꦴꦠꦺꦴ} in word- or syllable-final position, where it indicates the glottal stop; other consonants can also be subjoined to it.

Behavior of ra. The behavior of the U+1B2D BALINESE LETTER RA is unique to Balinese. The inherited Kawi form of the script used a *repha* glyph in the same way as many Brahmic scripts do—it represented *ra* as the first element of a syllable. 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 reanalyzed and would be pronounced *damar*. Furthermore, *damar* would be represented using U+1B03 BALINESE SIGN SURANG for the *-r*, as shown in example 3. The character sequence used in Kawi for spelling *dharma* would in Balinese render as shown in example 2, where the base letter *ra* with a subjoined *ma* is not well formed for the writing system. The correct representation of *dharma* in Balinese is shown in example 4, where the reanalyzed *repha* is represented by the *surang* and is rendered above the first syllable instead of the second.

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 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 Rërèh* ^{ꦠꦏꦫꦺꦫꦺꦲ} “Mr Rërèh”, the postfixed form of ^ꦫ *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* ^{ꦏꦫꦺꦱꦺꦤꦺ}.

Figure 11-6. Writing *dharma* in Balinese

Rendering. The vowel signs /u/ and /u:/ take different forms when combined with subscripted consonant clusters, as shown in *Table 11-16*. The upper limit of consonant clusters is three, the last of which can be *-ya*, *-wa*, or *-ra*.

Table 11-16. Balinese Consonant Clusters with u and u:














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

Table 11-16. Balinese Consonant Clusters with u and u: (Continued)

Syllable	Glyph
<i>kyu</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 11-15, 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 MUSI-

CAL 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. Additional characters, known to be used inline in text (as opposed to decoratively on drawings), are expected to be proposed as Balinese extensions in due course.