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 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@pearsontechgroup.com. For sales outside the United States please contact International Sales, international@pearsoned.com

Visit us on the Web: www.awprofessional.com

Chapter 11

Southeast Asian Scripts

The following scripts are discussed in this chapter:

Thai	*Khmer*	*Philippine scripts*
Lao	*Tai Le*	*Buginese*
Myanmar	*New Tai Lue*	*Balinese*

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
ka	กะ	0E01 0E30
ka:	กา	0E01 0E32
ki	กิ	0E01 0E34
ki:	กี	0E01 0E35
ku	กุ	0E01 0E38
ku:	กู	0E01 0E39
kuʾ	กึ	0E01 0E36
kuʾ:	กื	0E01 0E37
ke	เกะ	0E40 0E01 0E30
ke:	เก	0E40 0E01
kae	แกะ	0E41 0E01 0E30
kae:	แก	0E41 0E01
ko	โกะ	0E42 0E01 0E30

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

Syllable	Glyphs	Code Point Sequence
ko:	โก	0E42 0E01
koʾ	เกาะ	0E40 0E01 0E32 0E30
koʾ:	กอ	0E01 0E2D
koe	เกอะ	0E40 0E01 0E2D 0E30
koe:	เกอ	0E40 0E01 0E2D
kia	เกีย	0E40 0E01 0E35 0E22
kuʾa	เกือ	0E40 0E01 0E37 0E2D
kua	กัว	0E01 0E31 0E27
kaw	เกา	0E40 0E01 0E32
koe:y	เกย	0E40 0E01 0E22
kay	ไก	0E44 0E01
kay	ใก	0E43 0E01
kam	กำ	0E01 0E33
kri	กฤ	0E01 0E24

Rendering of Thai Combining Marks. The 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 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ꜰᴏɴɢᴍᴀɴ 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 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ᴍᴀɪʏᴀᴍᴏᴋ is used to mark repetition of preceding letters. U+0E2F ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ᴘᴀɪʏᴀɴɴᴏɪ 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 ᴋʜᴍᴇʀ ꜱɪɢɴ ʙᴇʏʏᴀʟ.

U+0E5A ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ᴀɴɢᴋʜᴀɴᴋʜᴜ is used to mark the end of a long segment of text. It can be combined with a following U+0E30 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ꜱᴀʀᴀ ᴀ to mark a larger segment of text; typically this usage can be seen at the end of a verse in poetry. U+0E5B ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ᴋʜᴏᴍᴜᴛ 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 ᴋʜᴍᴇʀ ꜱɪɢɴ ᴋʜᴀɴ and U+17D5 ᴋʜᴍᴇʀ ꜱɪɢɴ ʙᴀʀɪʏᴏᴏꜱᴀɴ, 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 ᴢᴇʀᴏ ᴡɪᴅᴛʜ ꜱᴘᴀᴄᴇ should be used to place invisible marks for such breaks. The ᴢᴇʀᴏ ᴡɪᴅᴛʜ ꜱᴘᴀᴄᴇ 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 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ᴘʜɪɴᴛʜᴜ (*virama*) to mark the removal of the inherent vowel. There is no conjoining behavior, unlike in other Indic scripts. U+0E4D ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ɴɪᴋʜᴀʜɪᴛ is the Pali *nigghahita* and Sanskrit *anusvara*. U+0E30 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ꜱᴀʀᴀ ᴀ is the Sanskrit *visarga*. U+0E24 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ʀᴜ and U+0E26 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ʟᴜ are vocalic /r/ and /l/, with U+0E45 ᴛʜᴀɪ ᴄʜᴀʀᴀᴄᴛᴇʀ ʟᴀᴋᴋʜᴀɴɢʏᴀᴏ 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 ʟᴀᴏ ᴠᴏᴡᴇʟ ѕɪɢɴ ᴀ.

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
ka	ກະ	0E81 0EB0
ka:	ກາ	0E81 0EB2
ki	ກິ	0E81 0EB4
ki:	ກີ	0E81 0EB5
ku	ກຸ	0E81 0EB8
ku:	ກູ	0E81 0EB9
ku'	ກຶ	0E81 0EB6
ku':	ກື	0E81 0EB7
ke	ເກະ	0EC0 0E81 0EB0
ke:	ເກ	0EC0 0E81
kae	ແກະ	0EC1 0E81 0EB0
kae:	ແກ	0EC1 0E81
ko	ໂກະ	0EC2 0E81 0EB0
ko:	ໂກ	0EC2 0E81
ko'	ເກາະ	0EC0 0E81 0EB2 0EB0
ko':	ກໍ	0E81 0ECD
koe	ເກິ	0EC0 0E81 0EB4
koe:	ເກີ	0EC0 0E81 0EB5
kia	ເກັຽ ເກຍ	0EC0 0E81 0EB1 0EBD 0EC0 0E81 0EA2
ku'a	ເກືອ	0EC0 0E81 0EB7 0EAD
kua	ກົວ	0E81 0EBB 0EA7
kaw	ເກົາ	0EC0 0E81 0EBB 0EB2

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

Syllable	Glyphs	Code Point Sequence
koe:y	ເກິຍ ເກິຍ	0EC0 0E81 0EB5 0EBD 0EC0 0E81 0EB5 0EA2
kay	ໄກ	0EC4 0E81
kay	ໃກ	0EC3 0E81
kam	ກໍາ	0E81 0EB3

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

U+0EBB ʟᴀᴏ ᴠᴏᴡᴇʟ sɪɢɴ ᴍᴀɪ ᴋᴏɴ

U+0EBC ʟᴀᴏ sᴇᴍɪᴠᴏᴡᴇʟ sɪɢɴ ʟᴏ

U+0EBD ʟᴀᴏ sᴇᴍɪᴠᴏᴡᴇʟ sɪɢɴ ɴʏᴏ

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 ʟᴀᴏ ᴛᴏɴᴇ ᴍᴀɪ ᴇᴋ, U+0ECD ʟᴀᴏ ɴɪɢɢᴀʜɪᴛᴀ> should be displayed with the *niggahita* above the *mai ek*, and a sequence containing <U+0ECD ʟᴀᴏ ɴɪɢɢᴀʜɪᴛᴀ, U+0EC8 ʟᴀᴏ ᴛᴏɴᴇ ᴍᴀɪ ᴇᴋ> 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 ʟᴀᴏ ᴠᴏᴡᴇʟ sɪɢɴ ᴀᴍ 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 ʟᴀᴏ ᴛᴏɴᴇ ᴍᴀɪ ᴇᴋ, U+0EB3 ʟᴀᴏ ᴠᴏᴡᴇʟ sɪɢɴ ᴀᴍ> should be displayed with the *mai ek* above the *niggahita*.

Lao Aspirated Nasals. The Unicode character encoding includes two ligatures for Lao: U+0EDC ʟᴀᴏ ʜᴏ ɴᴏ and U+0EDD ʟᴀᴏ ʜᴏ ᴍᴏ. 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 ဝ MYANMAR LETTER WA and U+1040 ၀ MYANMAR DIGIT ZERO, are distinguished by their underlying encoding.

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

myanmar vowel sign o

> U+1000 က *ka* + U+1031 ေ◌ *vowel sign e* + U+102C ◌ာ *vowel sign aa* →
> ကော *kō*

myanmar vowel sign au

> U+1000 က *ka* + U+1031 ေ◌ *vowel sign e* + U+102C ◌ာ *vowel sign aa* +
> U+1039 ◌် *virama* + U+200C ⟨ZWNJ⟩ → ကော် *kau*

myanmar vowel sign ui

> U+1000 က *ka* + U+102F ◌ု *vowel sign u* + U+102D ◌ိ *vowel sign i* → ကိုု
> *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* → င်္က *ṅka*

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* ေကြွ [kjwei] ("to drop off") would be written via the following sequence:

> U+1000 က *ka* + U+1039 ◌် *virama* + U+101B ရ *ra* + U+1039 ◌် *virama*
> + U+101D ဝ *wa* + U+1031 ေ◌ *vowel sign e* → ေကြွ *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
kinzi	<U+1004, U+1039>	ε̇
consonant	[U+1000..U+1021]	က
subscript consonant	<U+1039, [U+1000..U+1019, U+101C, U+101E, U+1020, U+1021]>	္က
medial ya	<U+1039, U+101A>	္ယ
medial ra	<U+1039, U+101B>	္ရ
medial wa	<U+1039, U+101D>	္ဝ
medial ha	<U+1039, U+101F>	္ဟ
vowel sign e	U+1031	ေ
vowel sign u, uu	[U+102F, U+1030]	ု , ူ
vowel sign i, ii, ai	[U+102D, U+102E, U+1032]	ိ , ီ , ဲ
vowel sign aa	U+102C	ာ
anusvara	U+1036	ံ
atha (killer)	<U+1039, U+200C>	္
dot below	U+1037	့
visarga	U+1038	း

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

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

Table 11-4. Independent Khmer Vowel Characters

Name	Independent Vowel	Qa with Vowel Sign
i	ឥ	អិ, អ៊ិ, អ៊ី
ii	ឦ	អី, អ៊ី
u	ឧ	អុ, អ៊ុ
uk	ឩ	អុក
uu	ឩ	អូ, អ៊ូ
uuv	ឩ	អូរ
ry	ឫ	ឫ
ryy	ឬ	ឬ
ly	ឭ	លឹ
lyy	ឮ	លឺ
e	ឯ	េអ, ែអ
ai	ឰ	ែអ
oo	ឱ, ឲ	េអា
au	ឳ	េអៅ

Subscript consonant signs are used to represent any consonant following the first conso-nant 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 consec-utive 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. Theoreti-cally, any consonant cluster composed of any number of consonant sounds without inher-ent vowel sounds in between can be represented systematically by a consonant character and as many subscript consonant signs as necessary.

Examples of subscript consonant signs for a consonant cluster follow:

> ល្ង *lo + coeng + ngo* [lŋɔː] "sesame" (compare លង *lo + ngo* [lɔːŋ] "to haunt")

> លក្ស្មី *lo + ka + coeng + sa + coeng + mo + ii* [lèəksmei] "beauty, luck"

> កាហ្វេ *ka + aa + ha + coeng + vo + e* [kaːfeː] "coffee"

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

Examples of subscript consonant signs for a closing consonant follow:

ទាំង *to + aa + nikahit + coeng + ngo* [tὲəŋ] "both" (= ទាំង) (≠ *ទ្ាំ [tŋɔ̀əm])

បើើ្យ *ha + oe + coeng + yo* [haəi] "already" (= បើើយ) (≠ *បើើ្្ [hyaə])

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

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

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

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

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:

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

ហ្ឫទីយ *ha + coeng + ry + to + samyok sannya + yo* [harɯtey] "heart" (*royal*) (= ហឫទីយ *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	ក *ka* [kɔː] "neck"	គ *ko* [kɔ̀ː] "mute"
2	រ̈ *ro + muusikatoan* [rɔː] "small saw"	រ *ro* [rɔ̀ː] "fence (in the water)"
3	ស̃ក *sa + ka* [sɔːk] "to peel, to shed one's skin"	ស̃̃ក *sa + triisap + ka* [sɔ̀ːk] "to insert"
4	បក *ba + ka* [bɔːk] "to return"	*ប̃̃ក *ba + triisap + ka* [bɔ̀ːk]
5	ប̈ម *ba + muusikatoan + mo* [pɔːm] "blockhouse"	ពម *po + mo* [pɔ̀ːm] "to put into the mouth"
6	ក្ុរ *ka + u + ro* [koː] "to stir"	គ្ុរ *ko + u + ro* [kuː] "to sketch"

Subscript Consonant Signs. In the way that many Cambodians analyze Khmer today, sub-
script consonant signs are considered to be different entities from consonant characters.
The Unicode Standard does not assign independent code points for the subscript conso-
nant signs. Instead, each of these signs is represented by the sequence of two characters: a
special control character (U+17D2 ᴋʜᴍᴇʀ sɪɢɴ ᴄᴏᴇɴɢ) 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 ◌ ᴋʜᴍᴇʀ sɪɢɴ ᴄᴏᴇɴɢ is not actually a *coeng* but a *coeng* generator, because *coeng*
in Khmer refers to the subscript consonant sign. The glyph for U+17D2 ◌ ᴋʜᴍᴇʀ sɪɢɴ
ᴄᴏᴇɴɢ shown in the code charts is arbitrary and is not actually rendered directly; the dot-
ted 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 Uni-
code 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 inde-
pendent vowels also have subscript forms, as shown in *Table 11-8*.

Table 11-6. 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*

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*

The Unicode Standard 5.0 – Electronic edition

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 in the Unicode Standard. The vowel sign *am* is encoded as a nasalization sign, U+17C6 ᴋʜᴍᴇʀ sɪɢɴ ɴɪᴋᴀʜɪᴛ. Two vowel signs, *om* and *aam*, have not been assigned independent code points. They are represented by the sequence of a vowel (U+17BB ᴋʜᴍᴇʀ ᴠᴏᴡᴇʟ sɪɢɴ ᴜ and U+17B6 ᴋʜᴍᴇʀ ᴠᴏᴡᴇʟ sɪɢɴ ᴀᴀ, respectively) and U+17C6 ᴋʜᴍᴇʀ sɪɢɴ ɴɪᴋᴀʜɪᴛ.

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 [ɔəm], not [aːm]. 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:

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

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

Independent Vowel Characters. In Khmer, as in other Brahmic scripts, some independent vowels have their own letterforms, although the sounds they represent may more often be represented with the consonant character for the glottal stop (U+17A2 ᴋʜᴍᴇʀ ʟᴇᴛᴛᴇʀ ǫᴀ) 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 ᴋʜᴍᴇʀ sɪɢɴ ᴄᴏᴇɴɢ and an independent vowel, as shown in *Table 11-8*.

Table 11-8. Khmer Subscript Independent Vowel Signs

Glyph	Code	Name
◌ᬏ	17D2 17A7	*khmer independent vowel sign coeng qu*
◌ᬈ	17D2 17AB	*khmer independent vowel sign coeng ry*
◌ᬈ	17D2 17AC	*khmer independent vowel sign coeng ryy*
◌ᬋ	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 ᴋʜᴍᴇʀ sɪɢɴ ʀᴏʙᴀᴛ, 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:

រាជឫ្សី *ro + aa + co + ro + coeng + sa + ii* [rèəcrsei] "king hermit"

អាយ៌ *qa + aa + yo + robat* [ʔaːrya] "civilized" (= អារ្យ *qa + aa + ro + coeng + yo*)

ពត៌មាន *po + ta + robat + mo + aa + no* [pɔːdɔmèən] "news" (compare Sanskrit वर्तमान *vartamāna* "the present time")

U+17DD ᴋʜᴍᴇʀ sɪɢɴ ᴀᴛᴛʜᴀᴄᴀɴ 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 ᴋʜᴍᴇʀ sɪɢɴ ᴠɪʀɪᴀᴍ. U+17CB ᴋʜᴍᴇʀ sɪɢɴ ʙᴀɴᴛᴏᴄ shortens the vowel sound of the previous ortho-

graphic syllable. U+17C7 ᴋʜᴍᴇʀ sɪɢɴ ʀᴇᴀʜᴍᴜᴋ, U+17C8 ᴋʜᴍᴇʀ sɪɢɴ ʏᴜᴜᴋᴀʟᴇᴀᴘɪɴᴛᴜ, U+17CD ᴋʜᴍᴇʀ sɪɢɴ ᴛᴏᴀɴᴅᴀᴋʜɪᴀᴛ, U+17CE ᴋʜᴍᴇʀ sɪɢɴ ᴋᴀᴋᴀʙᴀᴛ, U+17CF ᴋʜᴍᴇʀ sɪɢɴ ᴀʜsᴅᴀ, and U+17D0 ᴋʜᴍᴇʀ sɪɢɴ sᴀᴍʏᴏᴋ sᴀɴɴʏᴀ are also explicitly encoded signs used to compose an orthographic syllable.

Ligatures. Some vowel signs form ligatures with consonant characters and signs. These ligatures are not encoded separately, but should be presented graphically by the rendering software. Some common ligatures are shown in *Figure 11-1*.

Figure 11-1. Common Ligatures in Khmer

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

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

ប *ba* + េ◌ា *au* = បៅ [baw] "to suck"

ម *mo* + ◌្ឫ *coeng sa* + េ◌ា *au* = ម្បៅ [msaw] "powder"

ស *sa* + ង *ngo* + ◌្ខ *coeng kha* + ◌្យ *coeng yo* + ◌ា *aa* = សង្ខ្យា [sɔŋkʰyaː] "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 ᴋʜᴍᴇʀ sɪɢɴ ᴄᴏᴇɴɢ, U+1789 ᴋʜᴍᴇʀ ʟᴇᴛᴛᴇʀ ɴʏᴏ> 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 ឲ ᴋʜᴍᴇʀ ɪɴᴅᴇᴘᴇɴᴅᴇɴᴛ ᴠᴏᴡᴇʟ ǫᴏᴏ ᴛʏᴘᴇ ᴛᴡᴏ is thought to be a variant of U+17B1 ឱ ᴋʜᴍᴇʀ ɪɴᴅᴇᴘᴇɴᴅᴇɴᴛ ᴠᴏᴡᴇʟ ǫᴏᴏ ᴛʏᴘᴇ ᴏɴᴇ, 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* [ɲɔɲɯm] "to smile"

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

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

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

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

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

The use of U+17A3 ᴋʜᴍᴇʀ ɪɴᴅᴇᴘᴇɴᴅᴇɴᴛ ᴠᴏᴡᴇʟ ǫᴀǫ and U+17A4 ᴋʜᴍᴇʀ ɪɴᴅᴇᴘᴇɴᴅᴇɴᴛ ᴠᴏᴡᴇʟ ǫᴀᴀ is discouraged. One feature of the Khmer script is the introduction of the consonant character for a glottal stop (U+17A2 ᴋʜᴍᴇʀ ʟᴇᴛᴛᴇʀ ǫᴀ). 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 ᴋʜᴍᴇʀ ᴠᴏᴡᴇʟ ɪɴʜᴇʀᴇɴᴛ ᴀǫ and U+17B5 ᴋʜᴍᴇʀ ᴠᴏᴡᴇʟ ɪɴʜᴇʀᴇɴᴛ ᴀᴀ 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 ᴋʜᴍᴇʀ ꜱɪɢɴ ʙᴇʏʏᴀʟ 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:

> ៕ៗ *khan + ba + e + khan*

> –ៗ *en dash + ba + e + en dash*

> ៕ ៗ *khan + lo + khan*

> –ៗ– *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 ខ្ញុំ khnyom "I" is composed of the following three elements: (1) consonant character *kha* as *B*; (2) subscript consonant sign *coeng nyo* as *S*; and

(3) dependent vowel sign *om* as V. In the Unicode Standard, *coeng nyo* and *om* are further decomposed, and the whole word is represented by five coded characters.

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

The order of coded characters does not always match the visual order. For example, some of the dependent vowel signs and their fragments may seem to precede a consonant character, but they are always put after it in the sequence of coded characters. This is also the case with *coeng ro*. Examples of visual reordering and other aspects of syllabic order are shown in *Figure 11-3*.

Figure 11-3. Examples of Syllabic Order in Khmer

តេ *to + e* [tèː] "much"

ប្រើន *ca + coeng + ro + oe + no* [craən] "much"

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

ហើយ *ha + oe + coeng + yo* [haəi] "already"

សញ្ញា *sa + nyo + coeng + nyo + aa* [saɲɲaː] "sign"

ស៊ី *sa + triisap + ii* [siː] "eat"

ប៊ី *ba + muusikatoan + ii* [pei] "a kind of flute"

Consonant Shifters. U+17C9 ᴋʜᴍᴇʀ sɪɢɴ ᴍᴜᴜsɪᴋᴀᴛᴏᴀɴ and U+17CA ᴋʜᴍᴇʀ sɪɢɴ ᴛʀɪ-ɪsᴀᴘ are consonant shifters, also known as register shifters. In the presence of other super-script glyphs, both of these signs are usually rendered with the same glyph shape as that of U+17BB ᴋʜᴍᴇʀ ᴠᴏᴡᴇʟ sɪɢɴ ᴜ, as shown in the last two examples of *Figure 11-3*.

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

ម្ងៃ *mo + muusikatoan + coeng + ngo + ai* [mŋai] "one day"

ម្ហែតៗ *mo + triisap + coeng + ha + ae + ta + lek too* [mhèːtmhèːt] "bland"

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

ប៊ីយែរ *ba* + ⬚ + *triisap* + *ii* + *yo* + *ae* + *ro* [biyɛ̀ː] "beer"

ប្រតឹងអ៊ឺះ *ba* + *coeng* + *ro* + *ta* + *yy* + *ngo* + *qa* + ⬚ + *triisap* + *y* + *reah-muk* [prɔtəːŋʔɯh] "urgent, too busy"

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

Ligature Control. In the *askaa muul* font style, some vowel signs ligate with the consonant characters to which they are applied. The font tables should determine whether they form a ligature; ligature use in *muul* fonts does not affect the meaning. However, U+200C ᴢᴇʀᴏ ᴡɪᴅᴛʜ ɴᴏɴ-ᴊᴏɪɴᴇʀ may be inserted before the vowel sign to explicitly suppress such a ligature, as shown in *Figure 11-4* for the word "savant," pronounced [vituː].

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

វិទ្ធុ	*vo* + *i* + *to* + *uu*	(*aksaa crieng* font)
វិទ្ធុ, វិទ្ធុ	*vo* + *i* + *to* + *uu*	(ligature dependent on the *muul* font)
វិទ្ធុ	*vo* + ⬚ + *i* + *to* + *uu*	(⬚ to prevent the ligature in a *muul* font)
វិទ្ធ	*vo* + ⬚ + *i* + *to* + *uu*	(⬚ to request the ligature in a *muul* font)

Spacing. Khmer does not use whitespace between words, although it does use whitespace between clauses and between parts of a name. If word boundary indications are desired—for example, as part of automatic line layout algorithms—the character U+200B ᴢᴇʀᴏ ᴡɪᴅᴛʜ sᴘᴀᴄᴇ should be used to place invisible marks for such breaks. The ᴢᴇʀᴏ ᴡɪᴅᴛʜ sᴘᴀᴄᴇ 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 ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴘᴀᴛʜᴀᴍᴀsᴀᴛ and U+19F0 ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴛᴜᴛᴇʏᴀsᴀᴛ represent the first and second of August, respectively, in a leap year. The 15 characters from U+19E1 ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴍᴜᴏʏ ᴋᴏᴇᴛ to U+19EF ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴅᴀᴘ-ᴘʀᴀᴍ ᴋᴏᴇᴛ represent the first through the fifteenth lunar waxing days, respectively. The 15 characters from U+19F1 ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴍᴜᴏʏ ʀᴏᴄ through U+19FF ᴋʜᴍᴇʀ sʏᴍʙᴏʟ ᴅᴀᴘ-ᴘʀᴀᴍ ʀᴏᴄ 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, ꡗꡃꡈ ꡆꡯꡂ [tai² lə⁶] (in older orthography ꡗꡘꡈ ꡆꡯꡏ). 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 ꩭ, ꩰ, ꩱ, ꩲ, ꩳ or ꩴ, 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
ta	ꥆ	ꥆ
ta²	ꥆꥍ	ꥆ̈
ta³	ꥆe	ꥆ̌
ta⁴	ꥆꬁ	ꥆ̀
ta⁵	ꥆꬦ	ꥆ̇
ta⁶	ꥆc	ꥆ́
ti	ꥆꩭ	ꥆꩭ
ti²	ꥆꩭꥍ	ꥆꩭ̈
ti³	ꥆꩭe	ꥆꩭ̌
ti⁴	ꥆꩭꬁ	ꥆꩭ̀
ti⁵	ꥆꩭꬦ	ꥆꩭ̇
ti⁶	ꥆꩭc	ꥆꩭ́

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 twenieth 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

ᨠ	ka +	ᨠ e	+	ᨠ6	t1		→ ᨠᨠ6	[ke:²]
ᨠ	ka +	ᨠ e	+	ᨠθ	i		→ ᨠᨠθ	[kə:¹]
ᨠ	ka +	ᨠ e	+	ᨠθᨠ	iy		→ ᨠᨠθᨠ	[kəi¹]
ᨠ	ka +	ᨠ e	+	ᨠθᨠ	iy +	ᨠ6 t1	→ ᨠᨠθᨠ6	[kəi²]
ᨠ	ka +	ᨠ e	+	ᨠθᨠ	iy +	ᨠe t2	→ ᨠᨠθᨠe	[kəi³]

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
ᨠ	ka^h	high		1	[ka¹]
ᨠ6	ka^h + t1	high	t1	2	[ka²]
ᨠe	ka^h + t2	high	t2	3	[ka³]
ᨠ	ka^l	low		4	[ka⁴]
ᨠ6	ka^l + t1	low	t1	5	[ka⁵]
ᨠe	ka^l + t2	low	t2	6	[ka⁶]

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

of normal use by the mid-1700s. The modern Tagalog language—also known as Filipino—is now written in the Latin script.

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

Principles of the Philippine Scripts

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

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

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

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

Virama. Although all languages normally written with the Philippine scripts have syllables ending in consonants, not all of the scripts have a mechanism for expressing the canceled *-a*. As a result, in those orthographies, the final consonants are unexpressed. Francisco Lopez introduced a cross-shaped *virama* in his 1620 catechism in the Ilocano language, but this innovation did not seem to find favor with native users, who seem to have considered the script adequate without it (they preferred ⼆⼆⼆ *kakapi* to ⼆⼆⼆ *kakampi*). A similar reform for the Hanunóo script seems to have been better received. The Hanunóo *pamudpod* was devised by Antoon Postma, who went to the Philippines from the Netherlands in the mid-1950s. In traditional orthography, ⼆⼆⼆⼆ *si apu ba upada* is, with the *pamudpod*, rendered more accurately as ⼆⼆⼆⼆ *si aypud bay upadan*; the Hanunóo pronunciation is *si aypod bay upadan*. The Tagalog *virama* and Hanunóo *pamudpod* cancel only the inherent *-a*. No conjunct consonants are employed in the Philippine scripts.

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, ligatures are often formed, as shown in *Table 11-13*.

Table 11-13. Hanunóo and Buhid Vowel Sign Combinations

Hanunóo			Buhid		
x	*x* + $\bar{\circ}$	*x* + $\underset{\circ}{}$	*x*	*x* + $\bar{\circ}$	*x* + $\underset{\circ}{}$
ᜈ	ᜈ̄	ᜈ	ᝆ	ᝆ̄	ᝆ
ᜈ	ᜈ̄	ᜈ	ᝃ	ᝃ̄	ᝃ
ᜈ	ᜈ	ᜈ	ᝄ	ᝄ̄	ᝄ
ᜃ	ᜃ̄	ᜃ	ᝅ	ᝅ̄	ᝅ
ᜎ	ᜎ̄	ᜎ	ᝆ	ᝆ̄	ᝆ
ᜈ	ᜈ̄	ᜈ	ᝇ	ᝇ̄	ᝇ
ᜏ	ᜏ	ᜏ	ᝈ	ᝈ̄	ᝈ
ᜋ	ᜋ̄	ᜋ	ᝉ	ᝉ̄	ᝉ
ᜌ	ᜌ̄	ᜌ	ᝊ	ᝊ̄	ᝊ
ᜈ	ᜈ̄	ᜈ	ᝋ	ᝋ̄	ᝋ
—	ᜆ	ᜆ	—	ᝌ	ᝌ
ᜈ	ᜈ̄	ᜈ	ᝍ	ᝍ̄	ᝍ
ᜈ	ᜈ̄	ᜈ	ᝎ	ᝎ̄	ᝎ
ᜈ	ᜈ̄	ᜈ	ᝏ	ᝏ̄	ᝏ
ᜐ	ᜐ̄	ᜐ	ᝐ	ᝐ̄	ᝐ

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 marks.

11.8 Buginese

Buginese: U+1A00–U+1A1F

The Buginese script is used on the island of Sulawesi, mainly in the southwest. A variety of traditional literature has been printed in it. As of 1971, as many as 2.3 million speakers of Buginese were reported in the southern part of Sulawesi. The Buginese script is one of the easternmost of the Brahmi scripts and is perhaps related to Javanese. It is attested as early as

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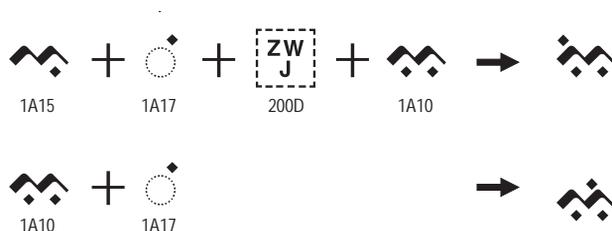
Order. Several orderings are possible for Buginese. The Unicode Standard encodes the Buginese characters in the Matthes order.

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 postfixed, 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
ka	ᬓ	ᬓ
kha	ᬔ	ᬔ
ga	ᬕ	ᬕ
gha	ᬖ	ᬖ
nga	ᬗ	ᬗ
ca	ᬘ	ᬘ
cha	ᬙ	ᬙ
ja	ᬚ	ᬚ
jha	ᬛ	ᬛ
nya	ᬜ	ᬜ
tta	ᬝ	ᬝ
ttha	ᬞ	ᬞ
dda	ᬟ	ᬟ
ddha	ᬠ	ᬠ

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

Consonant	Base Form	Conjunct Form
nna	ᬡ	ᬡ᬴
ta	ᬢ	ᬢ᬴
tha	ᬣ	ᬣ᬴
da	ᬤ	ᬤ᬴
dha	ᬥ	ᬥ᬴
na	ᬦ	ᬦ᬴
pa	ᬧ	ᬧ᬴
pha	ᬨ	ᬨ᬴
ba	ᬩ	ᬩ᬴
bha	ᬪ	ᬪ᬴
ma	ᬫ	ᬫ᬴
ya	ᬬ	ᬬ᬴
ra	ᬭ	ᬭ᬴
la	ᬮ	ᬮ᬴
wa	ᬯ	ᬯ᬴
ssa	ᬰ	ᬰ᬴
sha	ᬱ	ᬱ᬴
sa	ᬲ	ᬲ᬴
ha	ᬳ	ᬳ᬴
ṛ	ᬋ	ᬋ᬴

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
kaf	ꦏ	ꦏ
khot	ꦑ	ꦑ
tzir	ꦡ	ꦡ
ef	ꦍ	ꦍ
ve	ꦮ	ꦮ
zal	ꦗ	ꦗ
asyura	ꦱ	ꦱ

adeg adeg. In Sasak, the vowel letter *akara* can be followed by an explicit *adeg adeg* ꦄ 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 ʙᴀʟɪɴᴇꜱᴇ ʟᴇᴛᴛᴇʀ ʀᴀ 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 ʙᴀʟɪɴᴇꜱᴇ ꜱɪɢɴ ꜱᴜʀᴀɴɢ 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 ʙᴀʟɪɴᴇꜱᴇ ꜱɪɢɴ ᴄᴇᴄᴇᴋ (*anusvara*) is equivalent to *nga*, and U+1B04 ʙᴀʟɪɴᴇꜱᴇ ꜱɪɢɴ ʙɪꜱᴀʜ (*visarga*) is equivalent to *ha*.

***Behavior of* ra repa.** The unique behavior of ʙᴀʟɪɴᴇꜱᴇ ʟᴇᴛᴛᴇʀ ʀᴀ ʀᴇᴘᴀ (*vocalic ṛ*) results from a reanalysis of the independent vowel letter as a consonant. In a compound word in which the first element ends in a consonant and the second element begins with an original *ra* + *pepet*, such as *Pak Rërëh* ꦥꦏꦿꦺꦃ "Mr Rërëh", the postfixed form of ꦽ *ra repa* is used; this particular sequence is encoded *ka* + *adeg adeg* + *ra repa*. However, in other contexts where the *ra repa* represents the original Sanskrit vowel, U+1B3A ʙᴀʟɪɴᴇꜱᴇ ᴠᴏᴡᴇʟ ꜱɪɢɴ ʀᴀ ʀᴇᴘᴀ is used, as in *Krësna* ꦏꦽꦱ .
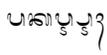
Figure 11-6. Writing *dharma* in Balinese

①	dha	ra	adeg adeg	ma	dha-rma
	1B25	1B2D	1B44	1B2B	(Kawi)

②	dha	ra	adeg adeg	ma	dha-rma
	1B25	1B2D	1B44	1B2B	(Balinese)

③	dha	ma	surang	dha-mar
	1B25	1B2B	1B03	

④	dha	surang	ma	dhar-ma
	1B25	1B03	1B2B	
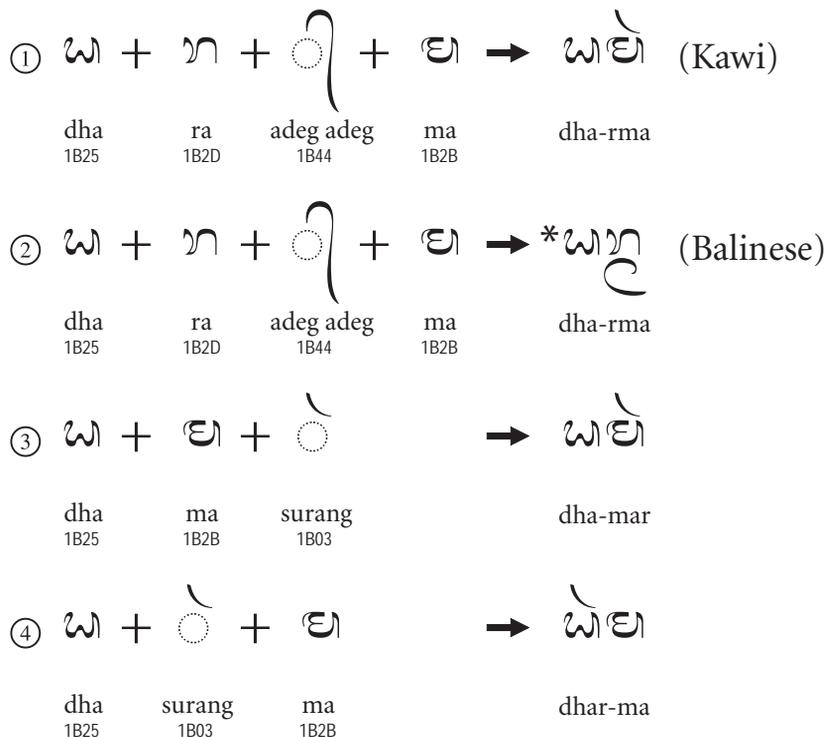
Rendering. The vowel signs /u/ and /u:/ take different forms when combined with sub-scripted consonant clusters, as shown in *Table 11-16*. The upper limit of consonant clusters is three, the last of which can be *-ya*, *-wa*, or *-ra*.

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

Syllable	Glyph
kyu	
kyú	
kwu	
kwú	
kru	
krú	

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

Syllable	Glyph
kyu	
kryu	
kryú	
skru	
skrú	

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.