# The Unicode® Standard
# Version 11.0 – Core Specification

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

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

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

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

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

# Chapter 17

# *Indonesia and Oceania*

The scripts described in this chapter are:

| | | |
|---|---|---|
| *Philippine scripts* | *Javanese* | *Sundanese* |
| *Buginese* | *Rejang* | *Makasar* |
| *Balinese* | *Batak* | |

Four traditional Philippine scripts are described here: Tagalog, Hanunóo, Buhid, and Tagbanwa. They have limited current use. Each is a very simplified abugida which makes use of two nonspacing vowel signs.

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

Like the other scripts in this chapter, the Makasar script is a Brahmi-derived abugida. Makasar is thought to have evolved from Rejang, and was used in South Sulawesi, Indonesia for writing the Makasar language. It has some similarities to the Buginese script, which superseded it in the 19th century.

# 17.1  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 consid-

ered the script adequate without it (they preferred ⟨ᜃᜃᜒ⟩ *kakapi* to ⟨ᜃᜃᜈ᜔ᜉᜒ⟩ *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, ⟨ᜰᜲ ᜪᜥ ᜲ ᜪᜯᜮ⟩ *si apu ba upada* is, with the *pamudpod*, rendered more accurately as ⟨ᜰᜲ ᜪᜮᜥᜮ ᜲᜮ ᜪᜯᜮᜲ⟩ *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 17-1*.

### Table 17-1. Hanunóo and Buhid Vowel Sign Combinations

| Hanunóo | | | | Buhid | | |
|---|---|---|---|---|---|---|
| *x* | *x* + ō | *x* + o̦ | | *x* | *x* + ō | *x* + o̦ |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| — | (glyph) | (glyph) | | — | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |
| (glyph) | (glyph) | (glyph) | | (glyph) | (glyph) | (glyph) |

**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.
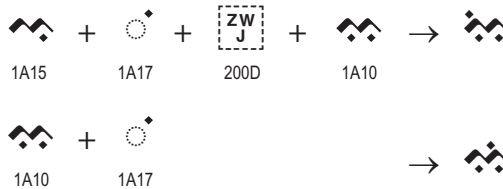
# 17.2  Buginese

## *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.

*Repertoire.* The repertoire contained in the Buginese block is intended to represent the core set of Buginese characters in standard printing fonts developed in the mid 19th century for the Bugis and Makassarese languages. Variant letterforms and other extensions seen in palm leaf manuscripts or additional letters used in some languages are not yet encoded in this block. A visible virama symbol has also been attested, but is not needed for this core repertoire for Buginese.

*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.

*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 17-1*. 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 17-1*. The ligature for *îya* is not obligatory; it would be requested by inserting a *zero width joiner*.

**Figure 17-1.**  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+A9CF JAVANESE PANGRANGKEP or a doubling of the vowel sign (especially U+1A19 BUGINESE VOWEL SIGN E and U+1A1A BUGINESE VOWEL SIGN O) is sometimes used to denote word reduplication. The shape of the Buginese reduplication sign is based on the Arabic digit two. The functionally similar U+A9CF JAVANESE PANGRANGKEP which has the same shape, is recommended for this sign in Buginese, rather than U+0662 ARABIC-INDIC DIGIT TWO, to avoid potential problems for text layout.

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

# 17.3 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 17-2* shows the base consonants and their conjunct forms.

**Table 17-2.** Balinese Base Consonants and Conjunct Forms

| Consonant | Base Form | Conjunct Form |
|---|---|---|
| *ka* | | |
| *kha* | | |
| *ga* | | |
| *gha* | | |
| *nga* | | |
| *ca* | | |
| *cha* | | |
| *ja* | | |
| *jha* | | |
| *nya* | | |
| *tta* | | |
| *ttha* | | |
| *dda* | | |
| *ddha* | | |

**Table 17-2.**  Balinese Base Consonants and Conjunct Forms (Continued)

| Consonant | Base Form | Conjunct Form |
|-----------|-----------|---------------|
| *nna* | ᭙ | ᭙ |
| *ta* | ᬢ | ᬢ |
| *tha* | ᬣ | ᬣ |
| *da* | ᬤ | ᬤ |
| *dha* | ᬥ | ᬥ |
| *na* | ᬦ | ᬦ |
| *pa* | ᬧ | ᬧ |
| *pha* | ᬨ | ᬨ |
| *ba* | ᬩ | ᬩ |
| *bha* | ᬪ | ᬪ |
| *ma* | ᬫ | ᬫ |
| *ya* | ᬬ | ᬬ |
| *ra* | ᬭ | ᬭ |
| *la* | ᬮ | ᬮ |
| *wa* | ᬯ | ᬯ |
| *ssa* | ᬰ | ᬰ |
| *sha* | ᬱ | ᬱ |
| *sa* | ᬲ | ᬲ |
| *ha* | ᬳ | ᬳ |
| *ṛ* | ᬀ | ᬀ |

The seven letters U+1B45 ʙᴀʟɪɴᴇsᴇ ʟᴇᴛᴛᴇʀ ᴋᴀf sᴀsᴀᴋ through U+1B4B ʙᴀʟɪɴᴇsᴇ ʟᴇᴛᴛᴇʀ ᴀsʏᴜʀᴀ sᴀsᴀᴋ are base consonant extensions for the Sasak language. Their base forms and conjunct forms are shown in *Table 17-3*.

**Table 17-3.** Sasak Extensions for Balinese

| Consonant | Base Form | Conjunct Form |
|---|---|---|
| *kaf* | ᬓ | ᬓ |
| *khot* | ᬔ | ᬔ |
| *tzir* | ᬘ | ᬘ |
| *ef* | ᬧ | ᬧ |
| *ve* | ᬯ | ᬯ |
| *zal* | ᬚ | ᬚ |
| *asyura* | ᬰ | ᬰ |

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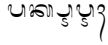
Independent vowels are used in a manner similar to that seen in other Brahmic scripts, with a few differences. For example, U+1B05 BALINESE LETTER AKARA and U+1B0B BALI-NESE LETTER RA REPA can be treated as consonants; that is, they can be followed by *adeg adeg*. In Sasak, the vowel letter *akara* can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

***Behavior of* ra.** Unlike most Brahmi-derived scripts, a Balinese *ra* that starts a sequence of consonants without intervening vowels is represented by U+1B03 BALINESE SIGN SURANG over the preceding syllable, as shown in the fourth example in *Figure 17-2*. The inherited Kawi form of the script used a *repha* glyph in the same way as many Brahmic scripts do. This is seen in the first example in *Figure 17-2*, where the sequence <*ra, virama, ma*> is rendered with the *repha* glyph. However, because many syllables end in -*r* in the Balinese language, this written form was historically reanalyzed, and is now pronounced *damar* in Balinese, as shown in the third example. In Balinese, the character sequence used in Kawi to spell *dharma* would render as shown in the second example, where the base letter *ra* with a subjoined *ma* is not well formed for the writing system.

Because of its relationship to *ra*, *surang* should be treated as equivalent to *ra* for searching and sorting purposes. Two other combining signs are also equivalent to base letters for searching and sorting: U+1B02 BALINESE SIGN CECEK (*anusvara*) is equivalent to *nga*, and U+1B04 BALINESE SIGN BISAH (*visarga*) is equivalent to *ha*.

**Figure 17-2.** Writing *dharma* in Balinese
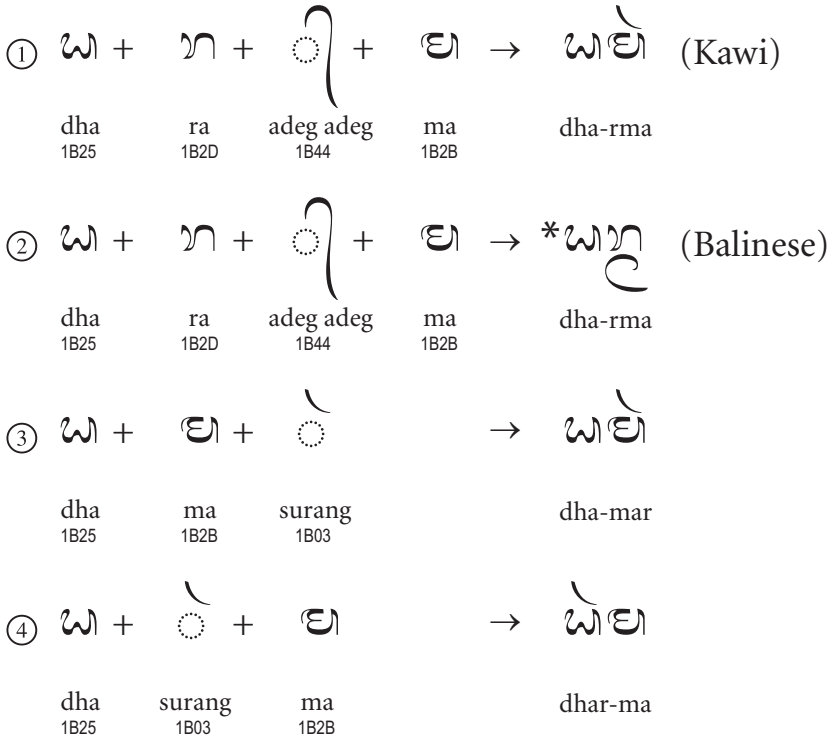


**Behavior of ra repa.** The unique behavior of BALINESE LETTER RA REPA (*vocalic ṛ*) results from a reanalysis of the independent vowel letter as a consonant. In a compound word in which the first element ends in a consonant and the second element begins with an original *ra + pepet*, such as *Pak Rëréh* ᬧᬓ᭄ᬭᬾᬭᬾᬄ "Mr Rëréh", the postfixed form of ᬭ *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* ᬓᬺᬲ᭄ᬡ .

**Rendering.** The vowel signs /u/ and /u:/ take different forms when combined with subscripted consonant clusters, as shown in *Table 17-4*. The upper limit of consonant clusters is three, the last of which can be *-ya*, *-wa*, or *-ra*.

**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 17-3*, also making use of some unused Kawi letters for these Arabic sounds.

**Table 17-4.** Balinese Consonant Clusters with *u* and *u:*

| Syllable | Glyph |
|----------|-------|
| *kyu* | |
| *kyú* | |
| *kwu* | |
| *kwú* | |
| *kru* | |
| *krú* | |
| *kryu* | |
| *kryú* | |
| *skru* | |
| *skrú* | |

***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 sys-

tems use other spacing letters, such as U+1B09 ʙᴀʟɪɴᴇsᴇ ʟᴇᴛᴛᴇʀ ᴜᴋᴀʀᴀ and U+1B27 ʙᴀʟɪ-ɴᴇsᴇ ʟᴇᴛᴛᴇʀ ᴘᴀ, which are not separately encoded for musical use. The U+1B01 ʙᴀʟɪɴᴇsᴇ sɪɢɴ ᴜʟᴜ ᴄᴀɴᴅʀᴀ (*candrabindu*) can also be used with U+1B62 ʙᴀʟɪɴᴇsᴇ ᴍᴜsɪᴄᴀʟ sʏᴍ-ʙᴏʟ ᴅᴇɴɢ and U+1B68 ʙᴀʟɪɴᴇsᴇ ᴍᴜsɪᴄᴀʟ sʏᴍʙᴏʟ ᴅᴇᴜɴɢ, and possibly others. ʙᴀʟɪɴᴇsᴇ sɪɢɴ ᴜʟᴜ ᴄᴀɴᴅʀᴀ 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 per-formance 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 ʙᴀʟɪɴᴇsᴇ sɪɢɴ ᴜʟᴜ ᴄᴀɴᴅʀᴀ.

# 17.4 Javanese

## *Javanese: U+A980–U+A9DF*

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

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

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

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

U+A9B3 JAVANESE SIGN CECAK TELU is a diacritic used with various consonantal base letters to represent foreign sounds. Typically these diacritic-marked consonants are used for sounds borrowed from Arabic.

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

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

**Dependent Vowels.** Javanese—unlike Balinese—represents multi-part dependent vowels with sequences of characters, in a manner similar to the Myanmar script. The Balinese community considers it important to be able to directly transliterate Sanskrit into Balinese, so multi-part dependent vowels are encoded as single, composite forms in Balinese, as is done in Devanagari. In contrast, for the Javanese script, the correspondence with Sanskrit letters is not so critical, and a different approach to the encoding has been taken. Similar to

the treatment of long versions of Javanese independent vowels, the two-part dependent vowels are explicitly represented with a sequence of two characters, using U+A9B4 JAVA-NESE VOWEL SIGN TARUNG, as shown in *Figure 17-3.*

**Figure 17-3.** Representation of Javanese Two-Part Vowels

| ka | pepet | tarung | keu |
| A98F | A9BC | A9B4 | |

| ka | taling | tarung | ko |
| A98F | A9BA | A9B4 | |

| ka | dirga mure | tarung | kau |
| A98F | A9BB | A9B4 | |
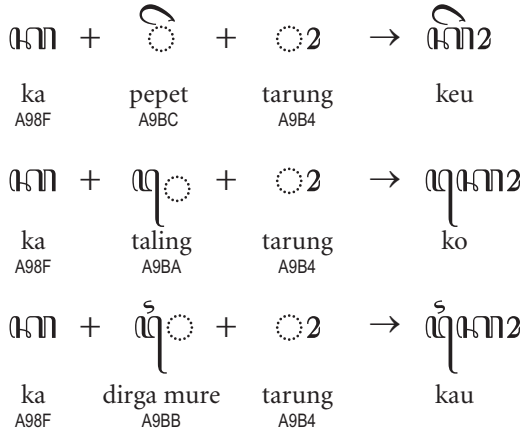
**Consonant Signs.** The characters U+A980 JAVANESE SIGN PANYANGGA, U+A981 JAVANESE SIGN CECAK, and U+A983 JAVANESE SIGN WIGNYAN are analoguous to U+0901 DEVANAGARI SIGN CANDRABINDU, U+0902 DEVANAGARI SIGN ANUSVARA, and U+0903 DEVANAGARI SIGN VISARGA and behave in much the same way.

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

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

**Digits.** The Javanese script has its own set of digits, seven of which (1, 2, 3, 6, 7, 8, 9) look just like letters of the alphabet. Implementations with concerns about security issues need to take this into account. The punctuation mark U+A9C7 JAVANESE PADA PANGKAT is often used with digits in order to help to distinguish numbers from sequences of letters.

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

U+A9C8 Javanese pada lingsa is a danda mark that corresponds functionally to the use of a comma. The doubled form, U+A9C9 Javanese pada lungsi, corresponds functionally to the use of a full stop. It is also used as a "ditto" mark in vertical lists. U+A9C7 Javanese pada pangkat is used much like the European colon.

U+A9C7 Javanese pada pangkat is used to abbreviate personal names and is placed at the end of the abbreviation.

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

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

***Reduplication.*** U+A9CF Javanese pangrangkep is used to show the reduplication of a syllable. The character derives from U+0662 arabic-indic digit two but in Javanese it does not have a numeric use. The Javanese reduplication mark is encoded as a separate character from the Arabic digit, because it differs in its Bidi_Class property value.

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

{C F} C {{R}Y} {V{A}} {Z}

where

> C is a letter (consonant or independent vowel), or a consonant followed by the diacritic U+A9B3 Javanese sign cecak telu
>
> F is the virama, U+A9C0 Javanese pangkon
>
> R is the medial -ra, U+A9BF Javanese consonant sign cakra
>
> Y is the medial -ya, U+A9BE Javanese consonant sign pengkal
>
> V is a dependent vowel sign
>
> A is the dependent vowel sign *-aa*, U+A9B4 Javanese vowel sign tarung
>
> Z is a consonant sign: U+A980, U+A981, U+A982, or U+A983

***Line Breaking.*** Opportunities for line breaking occur after any full orthographic syllable. Hyphens are not used.

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

# 17.5 Rejang

## *Rejang: U+A930–U+A95F*

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

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

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

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

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

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

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

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

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

# 17.6 Batak

### *Batak: U+1BC0–U+1BFF*

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

*Structure.* Batak is a Brahmi-derived script. It is written left to right.

Consonants in Batak have an inherent /a/ vowel sound. Batak uses a vowel killer which is called *pangolat* in Mandailing, Pakpak, and Toba. In Karo the killer is called *penengen*, and in Simalungen it is known as *panongonan*. The appearance of the killer differs between some of the dialects.

Batak has three independent vowels and makes use of a number of vowel signs and two consonant signs. Some vowel signs are only used by certain language communities. There are no consonant conjuncts. The basic syllabic structure is $C(V)(C_s|C_d)$: a consonant, followed by an optional vowel sign, which may be followed either by a consonant sign $C_s$ (*-ng* or *-h*) or a killed final consonant $C_d$.

*Rendering.* Most vowel signs and the two killers, U+1BF2 ʙᴀᴛᴀᴋ ᴘᴀɴɢᴏʟᴀᴛ and U+1BF3 ʙᴀᴛᴀᴋ ᴘᴀɴᴏɴɢᴏɴᴀɴ, are spacing marks. U+1BEE ʙᴀᴛᴀᴋ ᴠᴏᴡᴇʟ ꜱɪɢɴ ᴜ can ligate with its base consonant.

The two consonant signs, U+1BF0 ʙᴀᴛᴀᴋ ᴄᴏɴꜱᴏɴᴀɴᴛ ꜱɪɢɴ ɴɢ and U+1BF1 ʙᴀᴛᴀᴋ ᴄᴏɴꜱᴏɴᴀɴᴛ ꜱɪɢɴ ʜ, are nonspacing marks, usually rendered above the spacing vowel signs. When U+1BF0 ʙᴀᴛᴀᴋ ᴄᴏɴꜱᴏɴᴀɴᴛ ꜱɪɢɴ ɴɢ occurs together with the nonspacing mark, U+1BE9 ʙᴀᴛᴀᴋ ᴠᴏᴡᴇʟ ꜱɪɢɴ ᴇᴇ, both are rendered above the base consonant, with the glyph for the *ee* at the top left and the glyph for the *ng* at the top right.

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

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

The minor mark used to begin paragraphs and stanzas is U+1BFC ʙᴀᴛᴀᴋ ꜱʏᴍʙᴏʟ ʙɪɴᴅᴜ ɴᴀ ᴍᴇᴛᴇᴋ, which means "small bindu." It has a shape-based variant, U+1BFD ʙᴀᴛᴀᴋ ꜱʏᴍʙᴏʟ ʙɪɴᴅᴜ ᴘɪɴᴀʀʙᴏʀᴀꜱ ("rice-shaped bindu"), which is likewise used to separate sections

of text. U+1BFE ʙᴀᴛᴀᴋ sʏᴍʙᴏʟ ʙɪɴᴅᴜ ᴊᴜᴅᴜʟ ("title bindu") is sometimes used to separate a title from the main text, which normally begins on the same line.

**_Line Breaking._** Opportunities for a line break occur after any full orthographic syllable.

# 17.7 Sundanese

### *Sundanese: U+1B80–U+1BBF*

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

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

*Structure.* Sundanese consonants have an inherent vowel /a/. This inherent vowel can be modified by the addition of dependent vowel signs (*matras*). The script also has independent vowels.

In the modern orthography, an explicit vowel killer character, U+1BAA SUNDANESE SIGN PAMAAEH, is used to indicate the absence, or "killing," of the inherent vowel, but does not build consonant conjuncts. In Old Sundanese, however, consonant conjuncts do appear, and are formed with U+1BAB SUNDANESE SIGN VIRAMA.

*Medials.* In the modern orthography, initial Sundanese consonants can be followed by one of the three consonant signs for medial consonants, *-ya*, *-ra*, or *-la*. These medial consonants are graphically displayed as subjoined elements to their base consonants, and are not considered conjuncts proper, because they are not formed using a *virama*. In Old Sundanese, a subjoined *ma*, U+1BAC SUNDANESE CONSONANT SIGN PASANGAN MA, and a subjoined *wa*, U+1BAD SUNDANESE CONSONANT SIGN PASANGAN WA, occur. They contrast with the conjunct forms created with the *virama*.

*Final Consonants.* Sundanese historical texts employ two final consonants, U+1BBE SUNDANESE LETTER FINAL K and U+1BBF SUNDANESE LETTER FINAL M, which are distinct from the modern representation of these final consonants with the explicit killer U+1BAA SUNDANESE SIGN PAMAAEH.

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

*Historic Characters.* Additional historic consonants appear only in old texts: *reu*, *leu*, and *bha*. Another historic character, U+1BBA SUNDANESE AVAGRAHA, kills the vowel of the preceding consonant, and causes hiatus before an initial *a*.

***Additional Consonants.*** Two supplemental consonant letters are used in the modern script: U+1BAE ꮪ sundanese letter kha and U+1BAF sundanese letter sya. These are used to represent the borrowed sounds denoted by the Arabic letters *kha* and *sheen*, respectively.

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

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

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

***Ordering of Syllable Components.*** Dependent vowels and other signs are encoded after the consonant to which they apply. The ordering of elements for the modern Sundanese orthography is shown in more detail in *Table 17-5*.

**Table 17-5.** Modern Sundanese Syllabic Structure

| Class | Examples | Encoding |
|---|---|---|
| *consonant or independent vowel* | ꮐ | [U+1B83..U+1BA0, U+1BAE, U+1BAF] |
| *consonant sign -ya, -ra, -la* | ◌ꮪ , ◌ꮪ , ◌ꮪ | [U+1BA1..U+1BA3] |
| *dependent vowel, killer* | ◌ꮪ , ◌ꮪ | [U+1BA4..U+1BA9, U+1BAA] |
| *final consonant* | ◌ꮪ | [U+1B80..U+1B82] |

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

The left-side dependent vowel U+1BA6 sundanese vowel sign panaelaeng occurs in logical order after the consonant (and any medial consonant sign), but in visual presentation its glyph appears *before* (to the left of) the consonant.

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

## Sundanese Supplement: U+1CC0–U+1CCF

The Sundanese Supplement block contains eight *bindu* punctuation marks found in historical materials.

# 17.8 Makasar

## *Makasar: U+11EE0–U+11EFF*

The Makasar script was used historically in South Sulawesi, Indonesia for writing the Makasar language. It is sometimes spelled "Makassar," and is also referred to as "Old Makassarese" or "Makassarese bird script." The script was maintained for official purposes in the kingdoms of Makasar in the 17th century, and it was used for writing a number of historical accounts, such as the "Chronicles of Gowa and Tallo," but it was superseded by the Buginese script in the 19th century and is no longer used. Although Makasar is thought to have evolved from Rejang, it shares several similarities with Buginese.

**Structure.** Makasar is a Brahmi-derived *abugida.* It is written horizontally, from left to right. Consonant signs carry an inherent /a/ vowel sign. Alternative vowel sounds are expressed by applying one of four combining characters to a consonant. Each vowel sign appears on a different side of the base consonant: right, left, top, and bottom. They are all encoded as combining characters following the consonant.

Like Buginese, geminated and clustered consonants are not indicated, nor are syllable-final consonants. However, Makasar differs from the Buginese script in that it does not have the pre-nasalized clusters, such as /ŋka/, that occur in Buginese, and it includes special features for consonant repetition.

There is only one independent vowel sign, U+11EF1 ꜱ MAKASAR LETTER A. Vowel signs can be attached to this character to produce other vowel sounds when a syllable has no consonant, such as at the beginning of a word.

**Consonant Repetition.** Adjacent syllables that use the same consonant can be written by appending two vowel signs to a single consonant, as shown in the following example. Usually both vowels are the same in this case, and a consonant can take a maximum of two vowel signs.

> U+11EE7 ꜩ *da* + U+11EF4 ◌ *vowel sign u* + U+11EF4 ◌ *vowel sign u* →
> ꜩ [dudu]

U+11EF2 Ꜳ MAKASAR ANGKA can also be used to repeat the consonant used in the previous syllable. This is particularly useful when one or both syllables use the inherent vowel, but *angka* may also be followed by a different vowel sound from that of the previous syllable. *Angka* is associated with the inherent vowel or a vowel sign in the same way as any normal consonant character. For example:

> U+11EED ꜭ *ra* + U+11EF4 ◌ *vowel sign u* + U+11EF2 Ꜳ *angka* → ꜭꜲ
> [rura]

> U+11EE5 ꜳ *ma* + U+11EF2 Ꜳ *angka* + U+11EF3 ◌ *vowel sign i* → ꜳꜲ
> [mami]

***Letter* va.** U+11EEF MAKASAR LETTER VA is named "VA" even though the consonant is pronounced /w/ in the Makasar language. The name for this character aligns with the name for the related letter U+1A13 BUGINESE LETTER VA.

***Digits.*** The available Makasar manuscript sources show two distinct sets of digits. The first set strongly resembles European digits and can be represented with U+0030..U+0039. The second set strongly resembles Arabic-Indic digits, and can be represented with U+0660..U+0669. Therefore, script-specific digits for Makasar are not separately encoded. Digits are frequently used, and both sets occur concurrently in the sources.

The Arabic-Indic digits are restricted to Arabic-language environments—particularly for expressing dates of the Hijri era. The European digits are used for general purposes, but occur within Arabic-language contexts for writing non-Hijri dates, specifically those of the Gregorian calendar.

Digits may occur above U+0600 ـ ARABIC NUMBER SIGN or U+0601 �ح ARABIC SIGN SANAH, see *Figure 9-6* for an example.

***Punctuation.*** Sentences are delimited with U+11EF7 ⁝ MAKASAR PASSIMBANG, and sections are terminated with U+11EF8 ⁑ MAKASAR END OF SECTION. Words are often, but not always, separated by spaces. Line breaks normally appear after syllable boundaries. Hyphens or other marks indicating continuance are not used.

The end of a text is often marked using a stylized rendering of the Arabic word *tammat* تمّت, meaning "it is complete." There is no atomic character encoded for this symbol, so the sequence should be represented using Arabic letters *<ta + meem + shadda + ta>*, where the *shadda* is optional.