# The Unicode Standard
# Version 7.0 – Core Specification

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



The Unicode Standard / the Unicode Consortium ; edited by Julie D. Allen ... [et al.]. — Version 7.0
Includes bibliographical references and index.
ISBN 978-1-936213-09-2) (http://www.unicode.org/versions/Unicode7.0.0/)
1. Unicode (Computer character set) I. Allen, Julie D. II. Unicode Consortium.
QA268.U545 2014

ISBN 978-1-936213-09-2
Published in Mountain View, CA
October 2014

# Chapter 15

# *South and Central Asia-IV*

## *Other Historic Scripts*

This chapter documents scripts of South Asia aside from the major official scripts of India, which are described in *Chapter 12, South and Central Asia-I*.

The following South Asian scripts are described in this chapter:

| | | |
|---|---|---|
| ***Syloti Nagri*** | ***Siddham*** | ***Tirhuta*** |
| ***Kaithi*** | ***Mahajani*** | ***Modi*** |
| ***Sharada*** | ***Khojki*** | ***Grantha*** |
| ***Takri*** | ***Khudawadi*** | ***Sora Sompeng*** |

Most of these scripts are historically related to the other scripts of India, and most are ultimately derived from the Brahmi script. None of them were standardized in ISCII. The encoding for each script is done on its own terms, and the blocks do not make use of a common pattern for the layout of code points.

This introduction briefly identifies each script, occasionally highlighting the most salient distinctive attributes of the script. Details are provided in the individual block descriptions that follow.

Syloti Nagri is used to write the modern Sylheti language of northeast Bangladesh and southeast Assam in India.

Kaithi is a historic North Indian script, closedly related to the Devanagari and Gujarati scripts. It was used in the area of the present-day states of Bihar and Uttar Pradesh in northern India, from the 16th century until the early 20th century.

Sharada is a historical script that was used to write Sanskrit, Kashmiri, and other languages of northern South Asia; it was the principal inscriptional and literary script of Kashmir from the 8th century CE until the 20th century. It has limited and specialized modern use.

Takri, descended from Sharada, is used in northern India and surrounding countries. It is the traditional writing system for the Chambeali and Dogri languages, as well as several "Pahari" languages. In addition to popular usage for commercial and informal purposes, Takri served as the official script of several princely states of northern and northwestern India from the 17th century until the middle of the 20th century.

Siddham is another Brahmi-based writing system related to Sharada, and structurally similar to Devanagari. It originated in India, and was used across South, Central, and East Asia,

and is presently predominantly used in East Asia. Originally used for writing Buddhist manuscripts, the script is still used by Japanese Buddist communities.

Mahajani is a Brahmi-based alphabet commonly used by bankers and money lenders across northern India until the middle of the 20th century. It is a specialized commercial script used for writing accounts and financial records. Mahajani has similarities to Landa, Kaithi, and Devanagari.

Khojki is a writing system used by the Nizari Ismaili community of South Asia for recording religious literature. It is one of two Landa scripts—the other being Gurmuhki—that were developed into formal liturgical scripts for use by religious communities. It is still used today.

Khudawadi is a Landa-based script that was used to write the Sindhi language spoken in India and Pakistan. It is related to Sharada. Known as the shopkeeper and merchant script, it was used for routine writing, accounting, and other commercial purposes.

Tirhuta, another Brahmi-based script, is related to the Bengali, Newari, and Oriya scripts. Tirhuta is the traditional writing system for the Maithili language, which is spoken by more than 35 million people in parts of India and Nepal. Maithili is an official regional language of India and the second most spoken language in Nepal.

Modi is another Brahmi-based script mainly used to write Marathi, a language spoken in western and central India. It emerged in the 16th century and derives from the Nagari scripts. It is still used some today.

Grantha, a script with a long history, is used to write the Sanskrit language in parts of South India, Sri Lanka and elsewhere. It is in daily use by Vedic scholars and Hindu temple priests.

Sora Sompeng is used to write the Sora language spoken by the Sora people, who live in eastern India between the Oriya- and Telugu-speaking populations. The script was created in 1936 and is used in religious contexts.

# 15.1 Syloti Nagri

## *Syloti Nagri: U+A800–U+A82F*

Syloti Nagri is a lesser-known Brahmi-derived script used for writing the Sylheti language. Sylheti is an Indo-European language spoken by some 5 million speakers in the Barak Valley region of northeast Bangladesh and southeast Assam in India. Worldwide there may be as many as 10 million speakers. Sylheti has commonly been regarded as a dialect of Bengali, with which it shares a high proportion of vocabulary.

The Syloti Nagri script has 27 consonant letters with an inherent vowel of /o/ and 5 independent vowel letters. There are 5 dependent vowel signs that are attached to a consonant letter. Unlike Devanagari, there are no vowel signs that appear to the left of their associated consonant.

Only two proper diacritics are encoded to support Syloti Nagri: *anusvara* and *hasanta*. Aside from its traditional Indic designation, *anusvara* can also be considered a final form for the sequence /-ng/, which does not have a base glyph in Syloti Nagri because it does not occur in other positions. *Anusvara* can also occur with the vowels U+A824 ◌ꠤ SYLOTI NAGRI VOWEL SIGN I and U+A826 ◌ꠦ SYLOTI NAGRI VOWEL SIGN E, creating a potential problem with the display of both items. It is recommended that *anusvara* always occur in sequence after any vowel signs, as a final character.

***Virama and Conjuncts.*** Syloti Nagri is atypical of Indic scripts in use of the *virama* (*hasanta*) and conjuncts. Conjuncts are not strictly correlated with the phonology being represented. They are neither necessary in contexts involving a dead consonant, nor are they limited to such contexts. *Hasanta* was only recently introduced into the script and is used only in limited contexts. Conjuncts are not limited to sequences involving dead consonants but can be formed from pairs of characters of almost any type (consonant, independent vowel, dependent vowel) and can represent a wide variety of syllables. It is generally unnecessary to overtly indicate dead consonants with a conjunct or explicit *hasanta*. The only restriction is that an overtly rendered *hasanta* cannot occur in connection with the first element of a conjunct. The absence of *hasanta* does not imply a live consonant and has no bearing on the occurrence of conjuncts. Similarly, the absence of a conjunct does not imply a live consonant and has no bearing on the occurrence of *hasanta*.

***Digits.*** There are no unique Syloti Nagri digits. When digits do appear in Syloti Nagri texts, they are generally Bengali forms. Any font designed to support Syloti Nagri should include the Bengali digits because there is no guarantee that they would otherwise exist in a user's computing environment. They should use the corresponding Bengali block code points, U+09E6..U+09EF.

***Punctuation.*** With the advent of digital type and the modernization of the Syloti Nagri script, one can expect to find all of the traditional punctuation marks borrowed from the Latin typography: *period*, *comma*, *colon*, *semicolon*, *question mark*, and so on. In addition, the Devanagari *single danda* and *double danda* are used with great frequency.

***Poetry Marks.*** Four native poetry marks are included in the Syloti Nagri block. The script also makes use of U+2055 ❋ FLOWER PUNCTUATION MARK (in the General Punctuation block) as a poetry mark.

# 15.2 Kaithi

## *Kaithi: U+11080–U+110CF*

Kaithi, properly transliterated Kaithī, is a North Indian script, related to the Devanagari and Gujarati scripts. It was used in the area of the present-day states of Bihar and Uttar Pradesh in northern India.

Kaithi was employed for administrative purposes, commercial transactions, correspondence, and personal records, as well as to write religious and literary materials. As a means of administrative communication, the script was in use at least from the 16th century until the early 20th century, when it was eventually eclipsed by Devanagari. Kaithi was used to write Bhojpuri, Magahi, Awadhi, Maithili, Urdu, and other languages related to Hindi.

***Standards.*** There is no preexisting character encoding standard for the Kaithi script. The repertoire encoded in this block is based on the standard form of Kaithi developed by the British government of Bihar and the British provinces of northwest India in the 19th century. A few additional Kaithi characters found in manuscripts, printed books, alphabet charts, and other inventories of the script are also included.

***Styles.*** There are three presentation styles of the Kaithi script, each generally associated with a different language: Bhojpuri, Magahi, or Maithili. The Magahi style was adopted for official purposes in the state of Bihar, and is the basis for the representative glyphs in the code charts.

***Rendering Behavior.*** Kaithi is a Brahmi-derived script closely related to Devanagari. In general, the rules for Devanagari rendering apply to Kaithi as well. For more information, see *Section 12.1, Devanagari*.

***Vowel Letters.*** An independent Kaithi letter for *vocalic r* is represented by the consonant-vowel combination: U+110A9 KAITHI LETTER RA and U+110B2 KAITHI VOWEL SIGN II.

In print, the distinction between short and long forms of *i* and *u* is maintained. However, in handwritten text, there is a tendency to use the long vowels for both lengths.

***Consonant Conjuncts.*** Consonant clusters were handled in various ways in Kaithi. Some spoken languages that used the Kaithi script simplified clusters by inserting a vowel between the consonants, or through metathesis. When no such simplification occurred, conjuncts were represented in different ways: by ligatures, as the combination of the half-form of the first consonant and the following consonant, with an explicit virama (U+110B9 KAITHI SIGN VIRAMA) between two consonants, or as two consonants without a virama.

Consonant conjuncts in Kaithi are represented with a virama between the two consonants in the conjunct. For example, the ordinary representation of the conjunct *mba* would be by the sequence:

U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA + U+110A5 KAITHI LETTER BA

Consonant conjuncts may be rendered in distinct ways. Where there is a need to render conjuncts in the exact form as they appear in a particular source document, U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER can be used to request the appropriate presentation by the rendering system. For example, to display the explicitly ligated glyph 𑂦 for the conjunct *mba*, U+200D ZERO WIDTH JOINER is inserted after the virama:

U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA + U+200D ZERO WIDTH JOINER + U+110A5 KAITHI LETTER BA

To block use of a ligated glyph for the conjunct, and instead to display the conjunct with an explicit virama, U+200C ZERO WIDTH NON-JOINER is inserted after the virama:

U+110A7 KAITHI LETTER MA + U+110B9 KAITHI SIGN VIRAMA + U+200C ZERO WIDTH NON-JOINER + U+110A5 KAITHI LETTER BA

Conjuncts composed of a nasal and a consonant may be written either as a ligature with the half-form of the appropriate class nasal letter, or the full form of the nasal letter with an explicit virama (U+110B9 KAITHI SIGN VIRAMA) and consonant. In Grierson's *Linguistic Survey of India*, however, U+110A2 KAITHI LETTER NA is used for all articulation classes, both in ligatures and when the full form of the nasal appears with the virama.

***Ruled Lines.*** Kaithi, unlike Devanagari, does not employ a headstroke. While several manuscripts and books show a headstroke similar to that of Devanagari, the line is actually a ruled line used for emphasis, titling or sectioning, and is not broken between individual letters. Some Kaithi fonts, however, were designed with a headstroke, but the line is not broken between individual letters, as would occur in Devanagari.

***Nukta.*** Kaithi includes a nukta sign, U+110BA KAITHI SIGN NUKTA, a dot which is used as a diacritic below various consonants to form new letters. For example, the nukta is used to distinguish the sound *va* from *ba*. The precomposed character U+110AB KAITHI LETTER VA is separately encoded, and has a canonical decomposition into the sequence of U+110A5 KAITHI LETTER BA plus U+110BA KAITHI SIGN NUKTA. Precomposed characters are also encoded for two other Kaithi letters, *rha* and *dddha*.

The glyph for U+110A8 KAITHI LETTER YA may appear with or without a nukta. Because the form without the nukta is considered a glyph variant, it is not separately encoded as a character. The representative glyph used in the chart contains the dot. The nukta diacritic also marks letters representing some sounds in Urdu or sounds not native to Hindi. No precomposed characters are encoded in those cases, and such letters must be represented by a base character followed by the nukta.

***Punctuation.*** A number of Kaithi-specific punctuation marks are encoded. Two marks designate the ends of text sections: U+110BE KAITHI SECTION MARK, which generally indicates the end of a sentence, and U+110BF KAITHI DOUBLE SECTION MARK, which delimits larger blocks of text, such as paragraphs. Both section marks are generally drawn so that their glyphs extend to the edge of the text margins, particularly in manuscripts.

The character U+110BD KAITHI NUMBER SIGN is a format control character that interacts with digits, occurring either above or below a digit. The position of the KAITHI NUMBER

SIGN indicates its usage: when the mark occurs above a digit, it indicates a number in an itemized list, similar to U+2116 NUMERO SIGN. If it occurs below a digit, it indicates a numerical reference. Like U+0600 ARABIC NUMBER SIGN and the other Arabic signs that span numbers (see *Section 9.2, Arabic*), the KAITHI NUMBER SIGN precedes the numbers they graphically interact with, rather than following them, as would combining characters. The U+110BC KAITHI ENUMERATION SIGN is the spacing version of the KAITHI NUMBER SIGN, and is used for inline usage.

U+110BB KAITHI ABBREVIATION SIGN, shaped like a small circle, is used in Kaithi to indicate abbreviations. This mark is placed at the point of elision or after a ligature to indicate common words or phrases that are abbreviated, in a similar way to U+0970 DEVANAGARI ABBREVIATION SIGN.

Kaithi makes use of two script-specific dandas: U+110C0 KAITHI DANDA and U+110C1 KAITHI DOUBLE DANDA.

For other marks of punctuation occurring in Kaithi texts, available Unicode characters may be used. A cross-shaped character, used to mark phrase boundaries, can be represented by U+002B PLUS SIGN. For hyphenation, users should follow whatever is the recommended practice found in similar Indic script traditions, which might be U+2010 HYPHEN or U+002D HYPHEN-MINUS. For dot-like marks that appear as word-separators, U+2E31 WORD SEPARATOR MIDDLE DOT, or, if the word boundary is more like a dash, U+2010 HYPHEN can be used.

***Digits.*** The digits in Kaithi are considered to be stylistic variants of those used in Devanagari. Hence the Devanagari digits located at U+0966..096F should be employed. To indicate fractions and unit marks, Kaithi makes use of the numbers encoded in the Common Indic Number Forms block, U+A830..A839.

# 15.3 Sharada

## *Sharada: U+11180–U+111DF*

Sharada is a historical script that was used to write Sanskrit, Kashmiri, and other languages of northern South Asia. It served as the principal inscriptional and literary script of Kashmir from the 8th century CE until the 20th century. In the 19th century, expanded use of the Arabic script to write Kashmiri and the growth of Devanagari contributed to the marginalization of Sharada. Today the script is employed in a limited capacity by Kashmiri pandits for horoscopes and ritual purposes.

***Rendering Behavior.*** Sharada is a Brahmi-based script, closely related to Devanagari. In general, the rules for Devanagari rendering apply to Sharada as well. For more information, see *Section 12.1, Devanagari*.

***Ruled Lines.*** While the headstroke is an important structural feature of a character's glyph in Sharada, there is no rule governing the joining of headstrokes of characters to other characters. The variation was probably due to scribal preference, and should be handled at the font level.

***Virama.*** The U+111C0 ◌𑇀 SHARADA SIGN VIRAMA is a spacing mark, written to the right of the consonant letter it modifies. Semantically, it is identical to the Devanagari *virama* and other similar Indic scripts.

***Candrabindu and Avagraha.*** U+11180 ◌𑆀 SHARADA SIGN CANDRABINDU indicates nasalization of a vowel. It may appear in manuscripts in an inverted form but with no semantic difference. Such glyph variants should be handled in the font. U+111C1 𑇁 SHARADA AVAGRAHA represents the elision of a word-initial *a*. Unlike the usual practice in Devanagari in which the *avagraha* is written at the normal letter height and attaches to the top stroke of the following character, the *avagraha* in Sharada is written at or below the baseline and does not connect to the neighboring letter.

***Jihvamuliya and Upadhmaniya.*** The velar and labial allophones of /h/, followed by voiceless velar and labial stops respectively, are written in Sharada with separate signs, U+111C2 𑇂 SHARADA SIGN JIHVAMULIYA and U+111C3 𑇃 SHARADA SIGN UPADHMANIYA. These two signs have the properties of a letter and appear only in stacked conjuncts without the use of *virama*. *Jihvamuliya* is used to represent the velar fricative [x] in the context of following voiceless velar stops:

U+111C2 𑇂 *jihvamuliya* + U+11191 𑆑 *ka* → 𑆑𑇀𑇂

U+111C2 𑇂 *jihvamuliya* + U+11192 𑆒 *kha* → 𑆒𑇀𑇂

*Upadhmaniya* is used to represent the bilabial fricative [ɸ] in the context of following voiceless labial stops:

U+111C3 𑇃 *upadhmaniya* + U+111A5 𑆥 *pa* → 𑆥𑇀𑇃

U+111C3 𑇃 *upadhmaniya* + U+111A6 𑆦 *pha* → 𑆦𑇀𑇃

***Punctuation.*** U+111C7 𑇇 SHARADA ABBREVIATION SIGN appears after letters or combinations of letters. It marks the sequence as an abbreviation. A word separator, U+111C8 𑇈 SHARADA SEPARATOR, indicates word and other boundaries. Sharada also makes use of two script-specific dandas: U+111C5 𑇅 SHARADA DANDA and U+111C6 𑇆 SHARADA DOUBLE DANDA.

***Digits.*** Sharada has a distinctive set of digits encoded in the range U+111D0..U+111D9.

# 15.4 Takri

## *Takri: U+11680–U+116CF*

Takri is a script used in northern India and surrounding countries in South Asia, including the areas that comprise present-day Jammu and Kashmir, Himachal Pradesh, Punjab, and Uttarakhand. It is the traditional writing system for the Chambeali and Dogri languages, as well as several "Pahari" languages, such as Jaunsari, Kulvi, and Mandeali. It is related to the Gurmukhi, Landa, and Sharada scripts. Like other Brahmi-derived scripts, Takri is an *abugida*, with consonants taking an inherent vowel unless accompanied by a vowel marker or the *virama* (vowel killer).

Takri is descended from Sharada through an intermediate form known as Devāśeṣa, which emerged in the 14th century. Devāśeṣa was a script used for religious and official purposes, while its popular form, known as Takri, was used for commercial and informal purposes. Takri became differentiated from Devāśeṣa during the 16th century. In its various regional manifestations, Takri served as the official script of several princely states of northern and northwestern India from the 17th century until the middle of the 20th century. Until the late 19th century, Takri was used concurrently with Devanagari, but it was gradually replaced by the latter.

Owing to its use as both an official and a popular script, Takri appears in numerous records, from manuscripts to inscriptions to postage stamps. There are efforts to revive the use of Takri for languages such as Dogri, Kishtwari, and Kulvi as a means of preserving access to these language's literatures.

There is no universal, standard form of Takri. Where Takri was standardized, the reformed script was limited to a particular polity, such as a kingdom or a princely state. The representative glyphs shown in the code charts are taken mainly from the forms used in a variant established as the official script for writing the Chambeali language in the former Chamba State, now in Himachal Pradesh, India. There are a number of other regional varieties of Takri that have varying letterforms, sometimes quite different from the representative forms shown in the code charts. Such regional forms are considered glyphic variants and should be handled at the font level.

***Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 15-1* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

***Consonant Conjuncts.*** Conjuncts in Takri are infrequent and, when written, consist of two consonants, the second of which is always *ya*, *ra*, or *ha*. Takri *ya* is written as a subjoining form; Takri *ra* can be written as a ligature or a subjoining form; and Takri *ha* is written as a half-form.

**Table 15-1.** Takri Vowel Letters

| For | Use | Do Not Use |
|---|---|---|
| 𑚅 | 11681 | <11680, 116AD> |
| 𑚇 | 11687 | <11686, 116B2> |
| 𑚈 | 11688 | <11680, 116B4> |
| 𑚉 | 11689 | <11680, 116B5> |

***Nukta.*** A combining *nukta* character is encoded as U+116B7 TAKRI SIGN NUKTA. Characters that use this sound, mainly loan words and words from other languages, may be represented using the base character plus *nukta*.

***Headlines.*** Unlike Devanagari, headlines are not generally used in Takri. However, headlines do appear in the glyph shapes of certain Takri letters. The headline is an intrinsic feature of glyph shapes in some regional varieties such as Dogra Akkhar, where it appears to be inspired by the design of Devanagari characters. There are no fixed rules for the joining of headlines. For example, the headlines of two sequential characters possessing headlines are left unjoined in Chambeali, while the headlines of a letter and a vowel sign are joined in printed Dogra Akkhar.

***Punctuation.*** Takri uses U+0964 DEVANAGARI DANDA and U+0965 DEVANAGARI DOUBLE DANDA from Devanagari.

***Fractions.*** Fraction signs and currency marks found in Takri documents use the characters in the Common Indic Number Forms block (U+A830..U+A83F).

# 15.5 Siddham

## *Siddham: U+11580–U+115FF*

Siddham is a Brahmi-based writing system that originated in India, and is presently used primarily in East Asia. The script is also known as Siddhamātṛkā and Kuṭila. The name Siddhamatrika has broad historic and regional usage throughout India and East Asia. However, modern usage is most strongly associated with the Shingon and Tendai Buddhist traditions in Japan, where the script is also known as Bonji. The representative glyphs in the code charts are based upon Japanese forms of Siddham characters.

The historical record shows the use of Siddham in Central Asia, but the predominant examples are of its use for writing Sanskrit in China, Japan, and Korea, notably for Buddhist manuscripts. Today, it is mainly used for ceremonial and ritualistic purposes associated with esoteric Buddhist practices.

Siddham is most closely related to Sharada, another Brahmi-based script that originated in Kashmir.

***Nukta.*** The sign U+115C0 𑖿 SIDDHAM SIGN NUKTA is used for transcribing sounds that are not native to the writing system. The *nukta sign* is not a traditional Siddham character, but it is part of modern Siddham, so that it can accommodate the writing of Japanese and English.

***Virama and Conjuncts.*** The virama, U+115BF 𑖿 SIDDHAM SIGN VIRAMA, is identical to the corresponding character in Devanagari and silences the inherent vowel of a consonant. The default rendering of the Siddham *virama* is as a visible sign.

Consonant clusters in Siddham are written as conjuncts and follow the same model as conjuncts in Devanagari. Conjuncts are represented using the Siddham *virama*, which is written between each consonant in the cluster. Conjuncts may be written vertically, horizontally, or as independent ligatures. There are traditional Chinese and Japanese tabulations for Siddham conjuncts.

Siddham conjuncts may represent clusters with a large number of consonants. For example, *rkṣvrya* is a conjunct cluster produced by a sequence of six conjuncts, as shown in *Figure 15-1*.

**Figure 15-1.** Siddham Consonant Cluster


***Head Marks.*** The mark U+115C1 𑖁 SIDDHAM SIGN SIDDHAM is written at the beginning of a text. Palaeographically, the sign corresponds to characters used in other scripts, such as

U+0FD3 ྌ TIBETAN MARK INITIAL BRDA RNYING YIG MGO MDUN MA. It represents the Sanskrit word *siddham*, "accomplished," and the phrase *siddhirastu*, "may there be success." A vertically-oriented glyph variant is used for vertical text layout.

***Repetition Marks.*** Three marks, U+115C6 𑇆 SIDDHAM REPETITION MARK-1, U+115C7 𑇇 SIDDHAM REPETITION MARK-2, and U+115C8 𑇈 SIDDHAM REPETITION MARK-3 are used to indicate the text repetition. They are written after the text that is to be repeated.

***Punctuation.*** There are five punctuation characters encoded for Siddham, as shown in *Table 15-2*. Both Siddham *danda* and Siddham *double danda* have graphical variants used in informal Japanese writing of Siddham.

**Table 15-2.** Siddham Punctuation Characters

| Name | | | Purpose |
|---|---|---|---|
| U+115C2 | 𑇂 | SIDDHAM DANDA | marks the end of sentences and other short text sections |
| U+115C3 | 𑇃 | SIDDHAM DOUBLE DANDA | used at the end of paragraphs and larger text blocks |
| U+115C4 | 𑇄 | SIDDHAM SEPARATOR DOT | marks boundaries between syllables, words, and phrases; written at the head-height. |
| U+115C5 | 𑇅 | SIDDHAM SEPARATOR BAR | marks boundaries between syllables, words, and phrases |
| U+115C9 | 𑇉 | SIDDHAM END OF TEXT MARK | indicates the end or completion of a text |

# 15.6 Mahajani

## *Mahajani: U+11150–U+1117F*

Mahajani is a Brahmi-based writing system that was commonly used across northern India until the middle of the 20th century. It is a specialized commercial script used for writing accounts and financial records. It was used for recording several languages: Hindi, Marwari, and Punjabi. Mahajani was taught and used as a medium of education in Punjab, Rajasthan, Uttar Pradesh, Bihar, and Madhya Pradesh in schools where students from merchant and trading communities learned the script and other writing skills required for business. The name "Mahajani" refers to bankers and money lenders, who were the primary users of the script. The majority of Mahajani records are account books. Although the Mahajani script is no longer in general use, it is an important key to the historical financial records of northern India.

Mahajani has similarities to Landa, Kaithi, and Devanagari. In structure and orthography, Mahajani resembles scripts of the Landa family used in Punjab and Sind, which are related to Sharada.

***Structure.*** Mahajani is written from left to right. It is based upon the Brahmi model, but it is structurally simpler and behaves as an alphabet. Vowel signs are not used, and there is no virama. Consonant clusters are not written in Mahajani using half-forms or ligatures, or even a visible virama. The elements of a consonant cluster are written sequentially using regular consonant letters.

Vowel signs are not written. Consonant letters theoretically bear the inherent vowel /a/, but the glyph for *ka* for example represents not only *ka*, but also any one of the syllables *ka*, *kā*, *ki*, *kī*, *ke*, and so on. In cases where greater precision is required, a vowel letter may be written after a consonant to convey the intended vocalic context. In general, the value of a consonant letter must be inferred at the morphological level.

Nasalization is not represented using special signs, such as *anusvara*. Instead U+11167 MAHAJANI LETTER NA is used in cases where nasalization is explicitly recorded. In several cases, words are written simply with nasalization deleted.

U+11173 MAHAJANI SIGN NUKTA is used for writing sounds that are not represented by a unique character, such as allophonic variants and sounds that occur in local dialects or in loanwords. It has limited use in Mahajani.

Several letters have glyphic variants. Those variants are not separately encoded.

***Digits.*** Mahajani does not have distinctive script-specific digits. Digits similar to those used in Devanagari and Gujarati are attested.

***Other Symbols.*** Fraction signs and unit marks are found in Mahajani documents, and may be represented using the characters encoded in the "Common Indic Number Forms" block.

***Punctuation.*** Mahajani employs a dash, middle dot, and colon, which should be represented by the corresponding Latin characters. For the *dandas*, Mahajani employs U+0964

DEVANAGARI DANDA and U+0965 DEVANAGARI DOUBLE DANDA. Mahajani also contains two other script-specific punctuation signs, U+11174 ABBREVIATION SIGN and U+11175 SECTION MARK. There are no formal rules for punctuation and word spacing is not generally observed.

# 15.7 Khojki

## *Khojki: U+11200–U+1124F*

Khojki is a writing system used by the Nizari Ismaili community of South Asia for recording religious literature. It was developed in Sindh, now in Pakistan, for representing the Sindhi language. The script spread to surrounding regions and was used for writing Gujarati, Punjabi, and Siraiki, as well as several languages related to Hindi. It was also used for writing Arabic and Persian. Popular Nizari Ismaili tradition states that Khojki was invented and propagated by Pir Sadruddin, an Ismaili missionary.

Khojki is one of two Landa scripts that were developed into formal liturgical scripts for use by religious communities; the other is Gurmukhi, which was developed for writing the sacred literature of the Sikh tradition.

Khojki is also called "Sindhi" and "Khwajah Sindhi." Khojki was in use by the 16th century CE, as attested by manuscript evidence. The printing of Khojki books flourished after Laljibhai Devraj produced metal types for Khojki in Germany for use at his Khoja Sindhi Printing Press in Mumbai.

While usage of Khojki has declined over the past century, it is used wherever Nizari Ismaili Muslims of South Asian origin reside. The largest communities are found in Pakistan, India, Canada, United States, the United Kingdom, Kenya, Tanzania, and Uganda. Khojki primers continue to be published in Pakistan for teaching the script. Khojki manuscripts and books are used in Ismaili ceremonies not only in South Asia, but in east and south Africa, where large diaspora communities formed by the 19th century. The script was also used by communities related to the Nizari Ismailis, such as the Imamshahis of Gujarat.

***Structure.*** The general structure of Khojki is similar to that of other Brahmi-derived Indic scripts. It is written from left-to-right.

Khojki has a smaller repertoire of independent vowel letters than other Brahmi-derived scripts. The letters U+11202 KHOJKI LETTER I and U+11203 KHOJKI LETTER U are used for writing both short and long forms of *i* and *u*, respectively. The letters U+11205 KHOJKI LETTER AI and U+11207 KHOJKI LETTER AU represent diphthongs. Although they are attested in manuscripts and books, Khojki originally did not have unique letters for these vowels. In early Khojki records, diphthongs are generally represented as diagraphs. Several variant forms of vowel letters are also attested.

The repertoire of dependent vowel signs is larger than that of independent vowel letters. There are separate signs for U+1122D KHOJKI VOWEL SIGN I and U+1122E KHOJKI VOWEL SIGN II, but no form for *uu*. Instead, the single sign U+1122F KHOJKI VOWEL SIGN U is used for both short and long forms. U+11232 KHOJKI VOWEL SIGN O is often written by placing the U+11230 KHOJKI VOWEL SIGN E element above the consonant letter.

Geminate consonants are marked by the U+11237 KHOJKI SIGN SHADDA, written above the consonant letter that is doubled. The positioning may change in relation to vowel signs.

Nasalization is indicated by the sign U+11234 KHOJKI SIGN ANUSVARA. It is written to the right of the letter or sign with which it combines.

U+11235 KHOJKI SIGN VIRAMA is identical in function to corresponding characters in other Indic scripts. It is written to the right of a consonant letter.

U+11236 KHOJKI SIGN NUKTA is used for producing characters to represent sounds not native to Sindhi. The sign may be written with vowel letters, vowel signs, and consonant letters. The *nukta* is written above a letter.

***Punctuation.*** Khojki separates words using U+1123A KHOJKI WORD SEPARATOR. U+11238 KHOJKI DANDA and U+11239 KHOJKI DOUBLE DANDA are used to mark the end of sentences. The DOUBLE DANDA is also used to mark verse sections. Typically, DOUBLE DANDA is written with U+1123A KHOJKI WORD SEPARATOR to the left and right of verse numbers.

Section marks appear frequently in Khojki manuscripts as punctuation that delimits the end of a section or another larger block of text. The U+1123B KHOJKI SECTION MARK is generally used to mark the end of a sentence, while U+1123C DOUBLE SECTION MARK is used to delimit larger blocks of text, such as paragraphs. Both generally extend to the margin of the text-block.

Latin punctuation marks are also used in printed Khojki.

U+1123D KHOJKI ABBREVIATION SIGN is used for marking abbreviations.

***Digits.*** Khojki makes use of Gujarati digits U+0AE6 through U+0AEF.

# 15.8 Khudawadi

## *Khudawadi: U+112B0–U+112FF*

Khudawadi is a script used historically for writing the Sindhi language, which is spoken in India and Pakistan. Official forms of Khudawadi are known as "Hindi Sindhi," "Hindu Sindhi," and "Standard Sindhi." Khudawadi is a Landa-based script and related to Sharada. Like other Landa writing systems, Khudawadi is a mercantile script used for routine writing, accounting, and other commercial purposes and was known as the shopkeeper and merchant script. It is associated with the merchant communities of Hyderabad, Sindh. In addition to mercantile records, Khudawadi was used in education, book printing, and for court records.

In the 1860s, Khudawadi was chosen as the basis for a written standard for education and administration in Sindh and was developed as an official language. Official Khudawadi possesses unique characters for each vowel and consonant sound of the Sindhi language, as well as vowel signs. In the late 19th century, an Arabic-based script became the official writing system for Sindhi in Pakistan and India. Sindhi is also written in the Devanagari script in India. Khudawadi is now obsolete.

***Structure.*** The general structure of Khudawadi is similar to that of other Brahmi-based Indic scripts. It is written from left-to-right.

***Vowel Letters.*** Some independent vowel letters may be represented using a combination of a base vowel letter and a dependent vowel sign. This practice is not recommended. The atomic character for the independent vowel letter should always be used.

**Table 15-3.** Khudawadi Vowel Letters

| For | Use | Do Not Use |
|---|---|---|
| 𑊱 | 112B1 | 112B0 + 112E0 |
| 𑊶 | 112B6 | 112B0 + 112E5 |
| 𑊷 | 112B7 | 112B0 + 112E6 |
| 𑊸 | 112B8 | 112B0 + 112E7 |
| 𑊹 | 112B9 | 112B0 + 112E8 |

***Consonant Conjuncts.*** Consonant clusters generally consist of two consonants. These are written using a visible *virama*. The encoded representation is <C1 + *virama* + C2>. Half-forms and ligated conjunct forms are not attested.

***Nasalization.*** U+112DF ◌𑋟 KHUDAWADI SIGN ANUSVARA is used for indicating nasalization.

***Nukta.*** U+112E9 ◌𑋩 KHUDAWADI SIGN NUKTA is used for representing sounds not native to Sindhi, such as those that may occur in Persian and Arabic loanwords. Attested Khudawadi letters with *nukta* are shown in *Table 15-4*, along with the Arabic letters for which they substitute. JA + NUKTA, pronounced *za*, corresponds to a number of distinct Arabic letters.

**Table 15-4.** Representation of Arabic Sounds in Khudawadi

| Sound | Khudawadi | Arabic |
|---|---|---|
| *kha* | 𑊵𑊺 KHA + NUKTA | U+062E ARABIC LETTER KHAH |
| *ġa* | 𑊶𑊺 GA + NUKTA | U+063A ARABIC LETTER GHAIN |
| *za* | 𑊸𑊺 JA + NUKTA | U+0630 ARABIC LETTER THAL<br>U+0632 ARABIC LETTER ZAIN<br>U+0636 ARABIC LETTER DAD<br>U+0638 ARABIC LETTER ZAH |
| *fa* | 𑊷𑊺 PHA + NUKTA | U+0641 ARABIC LETTER FEH |

In principle, the *nukta* may be written with any Khudawadi vowel or consonant letter. If other combining marks, such as a dependent vowel sign or *anusvara*, also occur in a combining sequence applied to that base character, then the convention is to represent the *nukta* first in the combining sequence.

***Punctuation.*** The Khudawadi uses *dandas* and European punctuation, such as periods, dashes, colons, and semi-colons. Khudawadi *dandas* are unified with those of Devanagari. Line breaking for Khudawadi characters follows the rules for Devanagari.

***Digits.*** Khudawadi has a full set of decimal digits. Fraction signs and currency marks are attested in Khudawadi records. These may be represented using characters in the Common Indic Number Forms block found at U+A830..U+A83F.

# 15.9 Tirhuta

## *Tirhuta: U+11480–U+114DF*

Tirhuta is the traditional writing system for the Maithili language, which is spoken by more than 35 million people in the state of Bihar in India, and in the Narayani and Janakpur zones of Nepal. Maithili is an official regional language of India and the second most spoken language in Nepal. Tirhuta is a Brahmi-based script derived from Gauḍī, or "Proto-Bengali," which evolved from the Kuṭila branch of Brahmi by the 10th century. It is related to the Bengali, Newari, and Oriya scripts, which are also descended from Gauḍī, and became differentiated from them by the 14th century.

Tirhuta remained the primary writing system for Maithili until the late 20th century, when it was replaced by Devanagari. The Tirhuta script forms the basis of scholarly and religious scribal traditions that have been associated with the Maithili and Sanskrit languages since the 14th century. Tirhuta continues to be used for writing manuscripts of religious and literary texts, as well as personal correspondence. Since the 1950s, various literary societies, such as the Maithili Akademi and Chetna Samiti, have been publishing literary, educational, and linguistic materials in Tirhuta. The script is also used in signage in Darbhanga and other districts of north Bihar, and as an optional script for writing the civil services examination in Bihar.

Although several Tirhuta characters, ligatures or combined shapes bear resemblance to those of Bengali, these similarities are superficial.

***Structure.*** The general structure (phonetic order, matra reordering, use of *virama*, and so on) of Tirhuta is similar to that of other Brahmi-based Indic scripts. The script is written from left-to-right.

***Vowels.*** Tirhuta uses independent vowel letters and corresponding combining vowel signs. The signs U+114BA TIRHUTA VOWEL SIGN SHORT E and U+114BD TIRHUTA VOWEL SIGN SHORT O do not have corresponding independent forms, because the sounds they represent do not occur in word initial position.

Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 15-5* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

***Consonants.*** Some of the 33 consonants look like Bengali consonants, but represent different sounds. For example, U+114A9 TIRHUTA LETTER RA has the same form as U+09AC BENGALI LETTER BA, and U+09B2 BENGALI LETTER RA has the same shape as U+114AB TIRHUTA LETTER VA.

Consonants combined with vowel signs, combined in conjuncts, or appearing at the end of a word commonly use context-dependent ligatures or glyph combinations. These shapes also contrast with usage in Bengali. For example, the consonant-vowel combination <U+1149E TIRHUTA LETTER TA, U+114B3 TIRHUTA VOWEL SIGN U> in Tirhuta produces

**Table 15-5.** Tirhuta Vowel Letters

| For | Use | Do Not Use |
|---|---|---|
| 𑒋 | 11482 | <11481, 114B0> |
| 𑒌 | 11489 | <114AA, 114B5> |
| 𑒍 | 1148A | <114AA, 114B6> |
| 𑒏 | 1148C | <1148B, 114BA> |
| 𑒑 | 1148E | <1148D, 114BA> |

the same shape as the conjunct <U+09A4 BENGALI LETTER TA, U+09CD BENGALI SIGN VIRAMA, U+09A4 BENGALI LETTER TA> in the Bengali script.

All variant forms for letters, character elements and conjuncts in Tirhuta should be managed at the font level.

***Virama.*** U+114C2 TIRHUTA SIGN VIRAMA is identical in function to the corresponding character in other Indic scripts.

***Nasalization.*** Nasalization is indicated by U+114BF TIRHUTA SIGN CANDRABINDU and U+114C0 TIRHUTA SIGN ANUSVARA. These signs are written centered above the base. If written with an above-base sign or a letter with a graphical element that extends past the headstroke, they are placed to the right of such signs and elements.

***Characters for Representing Sanskrit.*** Two characters are attested in Vedic and classical Sanskrit manuscripts written in Tirhuta. U+114C1 TIRHUTA SIGN VISARGA represents an allophone of *ra* or *sa* at word-final position in Sanskrit orthography. U+114C5 TIRHUTA GVANG represents nasalization. It belongs to the same class of characters as U+1CE9 VEDIC SIGN ANUSVARA ANTARGOMUKHA, U+1CEA VEDIC SIGN ANUSVARA BAHIRGOMUKHA, and so on.

Tihuta also uses U+1CF2 VEDIC SIGN ARDHAVISARGA which can be found in the Vedic Extensions block.

***Nukta.*** U+114C3 TIRHUTA SIGN NUKTA is used for writing sounds that are not represented by a unique character, such as allophonic variants and sounds that occur in local dialects or in loanwords. The *nukta* may be written with any vowel or consonant letter. If other combining marks, such as a vowel sign or *anusvara*, also appear with the base character, then the *nukta* is written first.

U+114A5 TIRHUTA LETTER BA and U+114AB TIRHUTA LETTER VA have shapes that include a dot, but this is not semantically equivalent to a *nukta*. These letters do not decompose to *nukta*, and are treated as atomic characters.

***Punctuation.*** Tirhuta uses U+0964 DEVANAGARI DANDA and U+0965 DEVANAGARI DOUBLE DANDA from the Devanagari block.

***Special Signs.*** U+114C6 TIRHUTA ABBREVIATION SIGN denotes abbreviations. There are also two special script-specific signs in Tirhuta. The first, U+11480 TIRHUTA ANJI, is used

in the invocations of letters, manuscripts, books, and charts of the script. The sign *anji* is said to represent the tusk of the deity Ganesa, patron of learning. The second, U+114C7 TIRHUTA OM, contrasts with the Bengali sign for *om*, the latter being a simple combination of U+0993 BENGALI LETTER O plus U+0981 BENGALI SIGN CANDRABINDU.

***Numbers.*** Tirhuta has a full set of decimal digits.

Number forms and unit marks are also found in Tirhuta documents. The most common of these are signs for writing fractions and currency, and they are represented using characters in the Common Indic Number Forms block (U+A830..U+A83F). They include U+A831 NORTH INDIC FRACTION ONE HALF, U+A832 NORTH INDIC FRACTION THREE QUARTERS, and so on, as well as U+A838 NORTH INDIC RUPEE MARK. Tirhuta also uses Bengali "currency numerators," such as U+09F4 BENGALI CURRENCY NUMERATOR ONE.

# 15.10 Modi

## *Modi: U+11600–U+1165F*

Modi is a Brahmi-based script used mainly for writing Marathi. Modi was also used to write other regional languages such as Hindi, Gujarati, Kannada, Konkani, Persian, Tamil, and Telugu. According to an old legend, the Modi script was brought to India from Sri Lanka by Hemadri Pandit, known also as Hemadpant, who was the chief minister of Ramacandra, the last king of the Yadava dynasty, who reigned from 1271 to about 1309. Another tradition credits the creation of the script to Balaji Avaji, secretary of state to the late 17th-century Maratha king Shivaji Raje Bhonsle, also known as Chhatrapati Shivaji Maharaj. While the veracity of such accounts is difficult to ascertain, it is clear that Modi derives from the Nagari family of scripts and is a modification of the Nagari model intended for continous writing.

Modi emerged as an administrative writing system in the 16th century before the rise of the Maratha dynasties. It was adopted by the Marathas as an official script beginning in the 17th century and was used in such a capacity in Maharashtra until the middle of the 20th century. In the 1950s the use of Modi was formally discontinued and the Devanagari script, known as "Balbodh," was promoted as the standard writing system for Marathi.

There are thousands of Modi documents preserved in South Asia and Europe. The majority of these are in various archives in Maharashtra, while smaller collections are kept in Denmark and other countries, because of European presence in Tanjore, Pondicherry, and other regions in South Asia through the 19th century. The earliest extant Modi document dates from the early 17th century. While the majority of Modi documents are official letters, land records, and other administrative documents, the script was also used in education, journalism, and other routine activities before the 1950s. Printing in Modi began in the early 19th century after Charles Wilkins cut the first metal fonts for the script in Kolkata. Newspapers were published in Modi; primers were produced to teach the script in schools, and various personal papers and diaries were kept in the script.

***Structure.*** Modi is a Brahmi-based script related to Devanagari. It is written from left-to-right. In general, the rules for Devanagari rendering also apply to Modi (see *Section 12.1, Devanagari*). However, one characteristic feature of Modi is a large number of context-dependent forms of consonants and vowel-signs. Shaping and glyph substitutions for these contextual forms are managed in the font.

***Vowel Letters.*** Generally, the distinction between regular and long forms of *i* and *u* is not preserved in Modi. The letter U+11603 MODI LETTER II may represent both *i* and *ī*, and U+11604 MODI LETTER U may be used for writing both *u* and *ū*. The same can be said of the corresponding dependent vowel signs. Both regular and long forms appear in the Modi block, because they are attested in documentation about Modi.

The vocalic letters in the range U+11635..U+11638 are included in the encoding, but are not in modern use, as is the case in other Indic scripts. Modi *vocalic r* may alternatively be written as the sequence <U+11628 MODI LETTER RA, U+11632 MODI VOWEL SIGN II> *rī* .

Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. *Table 15-6* shows the letters that can be analyzed, the single code point that should be used to represent them in text, and the sequence of code points resulting from analysis that should not be used.

**Table 15-6.** Modi Vowel Letters

| For | Use | Do Not Use |
|---|---|---|
| 𑘎 | 1160A | <11600, 11639> |
| 𑘏 | 1160B | <11600, 1163A> |
| 𑘐 | 1160C | <11601, 11639> |
| 𑘑 | 1160D | <11601, 1163A> |

***Rendering.*** Many of the consonant-vowel and consonant-consonant combinations in Modi involve special contextual forms of the consonant or vowel-sign or both. These are rendered by means of contextual rules in the font, using specially shaped and positioned glyph pieces or preformed ligatures.

***Consonant Clusters Involving RA.*** A number of contextual forms are used for U+11628 𑘨 MODI LETTER RA. Some of these are similar to the use of *ra* in Devanagari. As the first consonant in a cluster it is generally rendered as a *repha*; however, Modi also uses the *eyelash ra* in place of *repha* in certain native Marathi contexts. As in Devanagari, the *eyelash ra* is produced using the sequence <U+11628 𑘨 MODI LETTER RA, U+1163F 𑘿 MODI SIGN VIRAMA, U+200D [ZWJ] ZERO WIDTH JOINER>.

Non-initial *ra* in conjuncts is typically rendered using one of two subjoined forms; however, some conjuncts with *ra* are represented as distinct ligatures. The most common of these is the conjunct represented by the sequence <U+1161D 𑘝 MODI LETTER TA, U+1163F 𑘿 MODI SIGN VIRAMA, U+11628 𑘨 MODI LETTER RA>.

Unusually, the shape of *ra* is also influenced at the word level depending upon the characters in the preceding syllable, or even preceding punctuation. In certain environments, *ra* following a *danda* or *double danda* is written using a special contexual form. For example:

U+11642 𑙂 *double danda* + U+11628 𑘨 *ra* → 𑙂𑘯

To produce this behavior, the *danda* and *double danda* characters in the Modi block should be used instead of the ones in the Devanagari block.

***Punctuation and Word Boundaries.*** Traditionally, word boundaries are not marked in Modi because it is an administrative script, characterized by the practice of rapid writing without lifting the pen. Paragraph and other section boundaries are, however, indicated in some Modi documents through the use of whitespace. Modern practice uses spaces and various punctuation conventions, including *danda* and Western punctuation marks. Some printed books use a period instead of a *danda* to indicate a sentence boundary.

***Various Signs.*** Nasalization is indicated by U+1163D MODI SIGN ANUSVARA, and abbreviations are indicated using U+11643 MODI ABBREVIATION SIGN. U+1163E MODI SIGN

VISARGA represents an allophone of *ra* or *sa* at word-final position in Sanskrit orthography. U+11640 MODI SIGN ARDHACANDRA is used for transcribing sounds used in English names and loanwords.

U+11644 MODI SIGN HUVA is written as an invocation in several Modi documents. It is derived from the Arabic *huwa*.

Currency values are written using U+A838 NORTH INDIC RUPEE MARK.

***Numbers.*** Modi has a full set of decimal digits. Several number forms and unit marks are used for writing Modi and are represented using characters in the Common Indic Number Forms block. They include the base-16 fraction signs U+A830..U+A835. The absence of intermediate units is indicated by U+A837 NORTH INDIC PLACEHOLDER MARK, which is called *ali* in Marathi. U+A836 NORTH INDIC QUARTER MARK is used for representing *anna* values.

# 15.11 Grantha

## *Grantha: U+11300–U+1137F*

The Grantha script descends from Brahmi. The modern form is chiefly used to write the Sanskrit language, including Vedic Sanskrit. It is used primarily in Tamil Nadu, and to a lesser extent in Sri Lanka and other parts of South India.

The Grantha script is frequently mixed with the Tamil script to write Sanskrit words. Grantha has also been used to write the Sanskrit words of Tamil Manipravalam—a mixed Sanskrit-Tamil language—though this usage has become rare.

Historically, intermediate forms which gave rise to the Grantha script are attested as of the fourth century CE. The earliest examples are found in inscriptions of the early Pallava kings who ruled over parts of what is currently northern Tamil Nadu and southern Andhra Pradesh. Modern Grantha, which this encoding represents, belongs to the period after the thirteenth century CE.

Modern Grantha is frequently used by Tamil speakers to represent Sanskrit because Grantha's large set of letters can represent all the sounds of Sanskrit without the use of diacritical marks. The Tamil script has a smaller repertoire of letters that requires diacritical marks to represent Sanskrit directly. This use of diacritical marks often leads to confusion regarding the pronunciation of Sanskrit when written in the Tamil script.

## *Rendering Behavior*

Although the Grantha script is visually similar to Tamil, its structure is similar to other Indic scripts that are used to write Sanskrit. Written Sanskrit requires support for stacked consonant structures.

***Consonant Clusters.*** Some consonant clusters are stacks, some consonant structures are a combination of ligatures and stacks, and some are just ligatures. Ligatures are often used instead of stacks, and consonant clusters are frequently written as a combination of ligatures and stacking.

The typical stack height found in print in non-Vedic Sanskrit is two elements, but it is three in Vedic Sanskrit. Stacks, like ligatures, are equivalent to single consonants for the purpose of application of vowel signs.

Instances requiring more than three elements in a stack require special handling. In these cases, the initial elements are pushed out of the consonant stack and may form their own stacks. Such special cases are illustrated in *Figure 15-2*. In this situation, a single phonological consonant cluster followed by a vowel may be represented by more than one orthographic cluster.

***Virama.*** Grantha follows the same *virama* model as Telugu and Kannada, in which the sequence *consonant* + *virama* should be rendered as the vowelless form of the consonant in the desired orthographic style. For example, in the prevalent orthographic style used in

**Figure 15-2.** Splitting Large Conjunct Stacks in Grantha

| | | |
|---|---|---|
| two elements | → | two-level stack |
| three elements | → | three-level stack |
| four elements | → | vowelless element + three-level stack |
| five elements | → | vowelless two-level stack + three-level stack |
| six elements | → | vowelless three-level stack + three-level stack |

modern printing, *ta*, *na*, and *ma* consistently fuse with the virama; *ra* and *la* superficially connect with it, and the virama stands apart for all other consonants, as shown in *Table 15-7*.

**Table 15-7.** Rendering of Explicit Virama Forms in Grantha

| | | | | |
|---|---|---|---|---|
| **Fused** | | | | |
| *ta + virama* | த | + ◌் | → | ஃ |
| *na + virama* | ந | + ◌் | → | ஙு |
| *ma + virama* | ம | + ◌் | → | ஶ |
| **Connected** | | | | |
| *ra + virama* | ற | + ◌் | → | ற் |
| *la + virama* | ல | + ◌் | → | ல் |
| **Unconnected** | | | | |
| *ka + virama* | கூ | + ◌் | → | கூ் |
| *tta + virama* | ட | + ◌் | → | ட் |

These visual distinctions in the rendering of explicit viramas also apply to the various ligated conjuncts of Grantha.

***Vowels.*** There are two forms of the *au* vowel sign: U+11357 GRANTHA AU LENGTH MARK is the modern one-part form, while the two-part form U+1134C GRANTHA VOWEL SIGN AU, is somewhat archaic, but is found in manuscripts.

Only two vowel signs touch their base consonant in printed Grantha: U+1133F GRANTHA VOWEL SIGN I and U+11340 GRANTHA VOWEL SIGN II. U+11347 GRANTHA VOWEL SIGN EE and U+11348 GRANTHA VOWEL SIGN AI are rendered to the left of their base. U+1134B GRANTHA VOWEL SIGN OO and the archaic U+1134C GRANTHA VOWEL SIGN AU are two-part vowels with one part placed to the left of the base and one part to the right. All other vowel signs are placed to the right of the base.

Manuscripts written in Grantha will show archaic ligatures of consonants with vowel signs. The vowel signs U+11362 GRANTHA VOWEL SIGN VOCALLIC L and U+11363 GRANTHA VOWEL SIGN VOCALLIC LL are sometimes placed below and sometimes placed to the right of the base consonant. In contemporary printing practice, vowel signs are placed to the right.

***Signs.*** Grantha uses the *pluta* sign to denote vowel lengthening. The *pluta* is not in current use, but it is found in Vedic manuscripts. The *nukta* is not used to write Sanskrit, but it is used to transcribe words from other languages.

***Cantillation Marks.*** Grantha uses a number of cantillation marks to represent tone, stress, and breathing in Vedic texts. These marks include the twelve marks encoded in the Grantha block in the range from U+11366..U+11374, and many encoded in other blocks as well, including those listed in *Table 15-8*.

**Table 15-8.** Additional *Svara* Marks used in Grantha

| **Generic Vedic Accents** |
|---|
| 0951 DEVANAGARI STRESS SIGN UDATTA |
| 0952 DEVANAGARI STRESS SIGN ANUDATTA |
| **Samavedic Marks** |
| 1CD0 VEDIC TONE KARSHANA |
| 1CD2 VEDIC TONE PRENKHA |
| 1CD3 VEDIC SIGN NIHSHVASA |
| 20F0 COMBINING ASTERISK ABOVE |
| **Additional Marks** |
| 1CF2 VEDIC SIGN ARDHAVISARGA |
| 1CF3 VEDIC SIGN ROTATED ARDHAVISARGA |
| 1CF4 VEDIC TONE CANDRA ABOVE |
| 1CF8 VEDIC TONE RING ABOVE |
| 1CF9 VEDIC TONE DOUBLE RING ABOVE |

These nonspacing marks are normally applied to independent vowels, to consonants with an inherent vowel, and to consonants with vowel signs. Sometimes they are also applied to dead consonants which are displayed with a visible *virama*.

The preferred placement of *svara* marks in Grantha is horizontally centered relative to the syllable. These marks should not extend beyond the horizontal span of the base syllable. The *svara* marks can be applied to either syllables or digits, and used in combination with each other.

***Punctuation.*** *Danda* and *double danda* marks used with Grantha are found in the Devanagari block; see *Section 12.1, Devanagari*.

# 15.12 Sora Sompeng

## *Sora Sompeng: U+110D0–U+110FF*

The Sora Sompeng script is used to write the Sora language. Sora is a member of the Munda family of languages, which, together with the Mon-Khmer languages, makes up Austro-Asiatic.

The Sora people live between the Oriya- and Telugu-speaking populations in what is now the Odisha-Andhra border area.

Sora Sompeng was devised in 1936 by Mangei Gomango, who was inspired by the vision he had of the 24 letters. The script was promulgated as part of a comprehensive cultural program, and was offered as an improvement over IPA-based scripts used by linguists and missionaries, and the Telugu and Oriya scripts used by Hindus. Sora Sompeng is used in religious contexts, and is published in a variety of printed materials.

***Encoding Structure.*** The Sora Sompeng script is structured as an abugida. The consonant letters contain an inherent vowel. There are no conjunct characters for consonant clusters, and there is no visible vowel killer to show the deletion of the inherent vowel. The reader must determine the presence or absence of the inherent schwa based on recognition of each word. The character repertoire does not match the phonemic repertoire of Sora very well.

U+110E4 SORA SOMPENG LETTER IH is used for both [i] and [ɨ], and U+110E6 SORA SOMPENG LETTER OH is used for both [o] and [ɔ], for instance. The glottal stop is written with U+110DE SORA SOMPENG LETTER HAH, and the sequence of U+110DD SORA SOMPENG LETTER RAH and U+110D4 SORA SOMPENG LETTER DAH is used to write retroflex [ɽ]. There is also an additional "auxiliary" U+110E8 SORA SOMPENG LETTER MAE used to transcribe foreign sounds.

***Character Names.*** Consonant letter names for Sora Sompeng are derived by adding [aʔa] (written *ah*) to the consonant.

***Punctuation.*** Sora Sompeng uses Western-style punctuation.

***Linebreaking.*** Letters and digits behave as in Latin and other alphabetic scripts.