Indic Scripts in Unicode

L2/12-289
Eric Muller

Summary

¶3 This document is a *very rough draft* of a description of the encoding of the Indic scripts in Unicode. It probably contains many mistakes. It is not a publication of the Unicode Consortium, nor has it been endorsed in any way by the Unicode Technical Committee.

Contents

¶12 The traditional repertoire starts with the short and long vowels:

short		long	
अ	a	आ	ā
इ	i	ई	ī
उ	u	ऊ	ū
ए	ē	ऐ	ai
ओ	ō	औ	au

¶13 Nasalized vowels are written with the diacritic sign ◌ँ known as the *candrabindu* in Hindi.

¶14 Next come the vocalic letters:

short		long	
ऋ	ṛ	ॠ	ṝ
ऌ	ḷ	ॡ	ḹ

¶15 Next come consonant letters. The first group covers the occlusive consonants, organized by point of articulation and other phonetic features:

	Voiceless				Voiced				Nasals
	unaspirated		aspirated		unaspirated		aspirated		
velar	क	ka	ख	kha	ग	ga	घ	gha	ङ ṅa
palatal	च	ca	छ	cha	ज	ja	झ	jha	ञ ña
retroflex	ट	ṭa	ठ	ṭha	ड	ḍa	ढ	ḍha	ण ṇa
dental	त	ta	थ	tha	द	da	ध	dha	न na
labial	प	pa	फ	pha	ब	ba	भ	bha	म ma

¶16 Next come the sonorants and fricatives:

	Sonorants		Fricatives	
palatal	य	ya	श	śa
retroflex	र	ra	ष	ṣa
dental	ल	la	स	sa
labial	व	va		

¶17 The remaining consonants are not classified by articulation:

ह	ha		
ळ	ḷa		

¶18 Nasalized consonants are written with the diacritic sign ◌ं known as the *anusvara* in Hindi.

¶19 The unvoiced aspiration is written with the diacritic sign ◌ः known as the *visarga* in Hindi.

¶20 As in most writing systems, the traditional inventory of letters often needs to be extended to transcribe words of other languages. A common device in the Indic scripts is the addition of the diacritic sign ◌ known as the *nukta* in Hindi.

¶21 The Dravidian languages, spoken in the south of India, have additional letters to write sounds which are not found in the Indo-European languages of the north, illustrated here by their form in Tamil:

ெ e
ொ o
ன n̲a
ற r̲a
ள l̲a

2.2 Digits and Numbers

¶22 Many of the Indic scripts have digits, which can often be used in decimal notation. In some cases, these script-specific digits are still in modern use; in other cases, the Latin digits are used instead in modern texts.

¶23 A number of scripts also have distinct signs which are assembled to write fractions.

2.3 Punctuation and Other Signs

¶24 Traditional punctuation is achieved by the sign ।, called *danda* for a full stop, and by the sign ॥, called *double danda*, to mark the end of a verse in traditional texts. The term *danda* is from Sanskrit, and those signs sometime have another name in the Indic languages (for example *viram* in Hindi). As with the letters, the appearance of those signs varies across scripts.

¶25 Many modern languages written in the Indic scripts intersperse punctuation derived from the Latin script. Thus U+002C ‘,’ COMMA and U+002E ‘.’ FULL STOP are freely used in writing Hindi, and the *danda* is usually restricted to more traditional texts. The *danda* may be preserved when such traditional texts are transliterated into the Latin script.

¶26 The *avagraha* symbol ऽ is used todo . It is not present in every script.

¶27 The *om* symbol ॐ 'om is a religious todo . It is not present in every script.

¶28 The *rupee sign* ૱ (Gujarati) is used todo . It is not present in every script.

2.4 Typographic Clusters

¶29 In many scripts, text is essentially written by the juxtaposition of letters, with limited typographic interaction between successive letters; furthermore, those interactions, such as an “fi” ligature, serve mostly an aesthetic purpose and are not used to indicate distinctions of meaning.

¶30 The Indic scripts, on the other hand, use typographic interactions to indicate differences of meaning. For example, the letter क ka followed by the letter र ra, without typographic interaction (कर) is usually read /kara/; with typographic interaction (क्र), the same sequence of letters is usually read /kra/. We call *typographic cluster* a succession of letters which interact typographically. The visible result of the interaction is that the shapes of the component letters are modified or even merged and spatially rearranged, both horizontally and vertically.

¶31 We use the notation $l_1 \bullet l_2 \ldots \bullet l_n$ to indicate that the letters l_1 , l_2 ... l_n form a typographic cluster. Thus <क ka • र ra> → क्र kra, while <क ka, र ra> → कर kara.

¶32 A consonantal typographic cluster corresponds roughly to a sequence of consonant sounds followed by a vowel sound. Graphically, this translates in *core consonants, satellite consonants, vowel signs* and *other signs* for secondary phonetic features. The vast majority of consonantal typographic clusters are made of a single core consonant.

¶33 A vowel typographic cluster corresponds roughly to a vowel sound alone. Graphically, this translates in a *vowel letter, vowel signs* and *other signs* for secondary phonetic features. The vast majority of vowel typographic clusters are made of a single vowel letter.

¶34 ***Core Consonants.*** The core of a typographic cluster writes a sequence of consonant sounds. Its written appearance is often a shape unique to that sequence:

क ka • ष ṣa → क्ष kṣa

श śa • र ra → श्र śra

द da • म ma → द्म dma

¶35 In other cases, it is more straightforward to recognize a combination of the basic letters; typically, one letter retains its normal shape, and the others take a reduced form:

क ka • र ra → क्र kra

क ka • न na → क्न kna

¶36 Yet in other cases (rare in the scripts of the North and essentially the norm in Tamil), the core is rendered by the juxtaposition of the basic letters, with the addition of the diacritic sign ◌् known as *halant* in Hindi:

ङ ṅa • ग ga → ङ्ग ṅga

¶37 Those examples illustrate more a continuum in the assembly of basic letters than strict categories in which every typographic cluster core falls neatly.

¶38 ***Satellite Consonants.*** Depending on the script, some consonants in initial or final position within a cluster may behave graphically more as satellites to the rest of the core of the typographic cluster. For example, in Devanagari, an initial र ra is depicted by a small hook, or *repha* at the extreme right of the core; in Myanmar, a final ရ ra will be depicted by an enclosing sign on the left:

र ra • क ka → र्क rka

र ra • क ka • न na → र्क्न rkna

ကra • ရ ra → ကြ kra

¶39 ***Vowel signs.*** Consonantal typographic clusters can include one or more *vowel sign* to denote their vowel sound. Those signs are diacritic forms of the vowel letters. We illustrate those on a few cores:

	क ka	र ra	क्न kna	र्क rka
आ ā	का kā	रा rā	क्ना knā	र्का rkā
इ i	कि ki	रि ri	क्नि kni	र्कि rki
ई ī	की kī	री rī	क्नी knī	र्की rkī
उ u	कु ku	रु ru	क्नु knu	र्कु rku
ऊ ū	कू kū	रू rū	क्नू knū	र्कू rkū
ए ē	के kē	रे rē	क्ने knē	र्के rkē

ऐ ai कै kai रै rai क्नै knai कैं rkai

ओ ō को kō रो rō क्नो knō कों rkō

औ au कौ kau रौ rau क्नौ knau कौं rkau

¶40 The normal form of a vowel letter is called an *independent vowel* or a *vowel letter*, while its diacritic form is called a *dependent vowel* or *vowel sign*.

¶41 As can be seen in those examples, the vowel signs can be placed all around the core, even though they write a vowel sound pronounced after that core. In some cases, they actually fuse into the core, as in रु ru or रू rū.

¶42 In some scripts, a vowel sign can be rendered by fragments on multiple sides of the core. For example, the vowel sign of Bengla ৌ au attaches to the simple core ক ka on both sides: কৌ kau; the vowel sign of Oriya ୌ au attaches to the simple core କ ka on the left, top and right sides: ଜୌ kau. In Unicode terminology, those vowels are called *split vowels*.

¶43 In some cases, vowel signs are also present in vowel typographic clusters; they attach to the vowel letter that form the core of the typographic cluster.

¶44 ***Clusters without vowels.*** Sometimes clusters represent only a sequence of consonant sounds without a vowel sound. This is most common at the end of words of Sanskrit or foreign origin. This situation is sometimes reflected in writing, by the addition of the halant sign to the cluster.

Hindi ् todo need example

Telugu ్ అప్ ap (water)

¶45 ***Other Signs.*** Finally, the signs ँ m̐, ं ṁ and ः ḥ can be added to a cluster. Their position may vary across scripts. For example, the Bengla ং ṁ attaches to the right of the core, rather than to the top.

¶46 ***Vowel Clusters.*** When a word starts with a vowel sound, there is no consonant to which to attach the diacritic form of the vowel letter. Instead, the non-diacritic form of the vowel acts as the core of the cluster, as in Hindi अगर agara. This core can be decorated with other signs, for example Hindi ऊँचा ūm̐cā.

¶47 ***Clusters and Syllables.*** The typograhic clusters need not correspond exactly with phonological syllables, especially when a consonant cluster is involved. For example, Hindi पुर्ति purti is made of two typographic clusters पु pu and र्ति rti, while the syllables are /pur/ and /ti/.

3 Encoding Models

3.1 Three Models

¶48 Three distinct models have been used in the Unicode standard for encoding the Indic scripts. They differ principally in the scheme used to indicate typographic clusters.

¶49 The *consonant linking model* (also known as the *virama model*) encodes one coded character for each consonant; one *consonant linker* coded character; one coded character for each vowel which can appear as a *vowel letter*; and one coded character for each vowel which can appear as a *vowel sign*. A typographic cluster is formed by placing the consonant linker between the consonants of the cluster.

¶50 The *subjoined consonant model* encodes two or more characters for each consonant, one non-combining and the others combining; one coded character for each vowel which can appear as a *vowel letter*; and one coded character for each vowel which can appear as a *vowel sign*. A typographic cluster is formed by using the non-combining coded character for the first consonant, and combining coded

characters for the remaining consonants.

¶51 In those two models, coded characters appear in logical order, that is in the same order as the sounds they write. This is particularly visible for vowels rendered on the left side of their typographic cluster, or the left part of a two-part vowel: the corresponding characters appear toward the end of the typographic cluster.

¶52 The *visual model* encodes a non-combining character for each spacing sign, and a combining character for each non-spacing sign. Furthermore, the coded characters appear in textual representation in visual order. This model was motivated by simple rendering technologies (such as typewriters).

¶53 For a given script, the choice of model in Unicode is guided by a number of factors, including legacy practices, closeness to other scripts, and perception of the script by its users.

¶54 In all cases, Unicode does not encode a character for each typographic cluster, but one or two for each letter. On the other hand, the glyph complement of fonts is more likely to include separate glyphs for many of the typographic clusters, or at least for their cores. Thus, the number of glyphs in an Indic font can far exceed the number of corresponding characters. It is up to the rendering software to select the appropriate glyph sequence for a given character sequence.

3.2 The Consonant Linking Model

¶55 The consonant linking model uses a *consonant linker* coded character to link together the consonants of a typographic cluster.

¶56 The consonant linking model is used for scripts covered by ISCII (which also uses that model) and closely related scripts: Devanagari, Bengla, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, Malayalam, Sinhala.

¶57 ***Coded Characters.*** In the consonant linking model, we have

- ¶58 one coded character for each consonant; for example:

  क ka → U+0915 क DEVANAGARI LETTER KA

  र ra → U+0930 र DEVANAGARI LETTER RA

- ¶59 one coded character which acts as a *consonant linker*; for example:

  U+094D ◌् DEVANAGARI SIGN VIRAMA

- ¶60 one coded character for each vowel which can appear as a *vowel letter*; for example:

  इ i → U+0907 इ DEVANAGARI LETTER I

- ¶61 one coded character for each vowel which can appear as a *vowel sign*; for example:

  इ i → U+093F ◌ि DEVANAGARI VOWEL SIGN I

- ¶62 one coded character for each diacritic sign; for example:

  ◌ँ m̐ → U+0901 ◌ँ DEVANAGARI SIGN CANDRABINDU

¶63 A cluster such as क ka • र ra • इ i → क्रि kri is represented by the sequence of coded characters:

U+0915 क DEVANAGARI LETTER KA

U+094D ◌् DEVANAGARI SIGN VIRAMA

U+0930 र DEVANAGARI LETTER RA

U+093F ि DEVANAGARI VOWEL SIGN I

¶64 That is: each consonant is represented by the corresponding coded character; successive consonants in the same cluster are linked by the consonant linker coded character; vowels signs in the cluster are represented by the corresponding vowel sign coded character.

¶65 A cluster such as इ i • ँ m̐ → ईँ im̐ is represented by the sequence of coded characters:

U+0907 इ DEVANAGARI LETTER I

U+0901 ँ DEVANAGARI SIGN CANDRABINDU

¶66 That is: each vowel letter is represented by the corresponding vowel letter coded character; diacritic signs are represented by the corresponding coded character.

¶67 The ordering of the coded characters matches the pronunciation order. In the cluster क ka • र ra • इ i → क्रि kri, the coded character for the vowel sign is last in the sequence because it is pronounced last, eventhough it is written first.

¶68 The representative glyph for the consonant linker is often the halant sign of the script: U+094D ् DEVANAGARI SIGN VIRAMA. However, the coded character needs to be understood more as a control character, which denotes the formation of typographic clusters, rather than as a graphic sign. The rendering of a cluster may result in the halant sign, but this caused only indirectly by the consonant linker. This situation is very similar to hyphenation opportunities which are denoted by U+00AD SOFT HYPHEN and can result in the - sign.

¶69 ***Atomicity of Vowel Letters.*** It is sometimes possible to analyze the visual appearance of vowel letters, and notice that they are formed by assembling the rendering of other vowel letters and signs. For example, the Devanagari letter ओ ō is clearly composed graphically of an आ ā and a े -ē. This characteristic is more than a graphical accident, and leads to two possible approaches for the encoding of vowel letters. The approach chosen by Unicode is to encode each vowel letter separately, without any decomposition: ओ is represented by the single coded character U+0913 ओ DEVANAGARI LETTER O and not by the sequence <U+0906 आ DEVANAGARI LETTER AA, U+0947 े DEVANAGARI VOWEL SIGN E>.

¶70 Similarly, vowel signs can be sometimes be analyzed. For example, Bengla ো is composed graphically of a ে and and a া. The approach chosen by Unicode is to encode the constituent parts: ো is represented by the sequence of coded characters <U+09C7 ে BENGALI VOWEL SIGN E, U+09BE া BENGALI VOWEL SIGN AA> rather than by a single coded character that has no decomposition. For compatibility with existing standards, some of these sequences are also encoded as a single code point (U+09CB ো BENGALI VOWEL SIGN O in this case), but that code point canonically decomposes in the sequence.

¶71 In the descriptions of the individual scripts, we will list the vowel letters and signs which can be confused graphically.

¶72 ***Use of Joiners for Consonants.*** As we have seen before, there is generally a high degree of typographic interaction within a cluster, and the amount of interaction is better understood as a continuum. Nevertheless, it is possible to identify three broad forms for the rendering of the core of a cluster:

- ¶73 fully-joined forms, where the consonants merge into a shape where both original shapes are modified
- ¶74 conjoined forms, where one of the consonant retains its full shape, and the others take reduced forms that attach to the unmodified consonant
- ¶75 non-joined forms, where each consonant retains its full shape, and a *vowel killer* sign is added to all but the last consonant

¶76 The following illustrates the three forms:

	fully-joined	conjoined	non-joined
क ka • ष ṣa	क्ष	क्ष	क्‌ष
क ka • र ra	क्र	क्र	क्‌र
क ka • न na		क्न	क्‌न
ङ ṅa • ग ga	ङ्ग		ङ्‌ग

¶77 As shown above, a given cluster can generally be rendered with more than one form, although not all forms need to exist. While all the forms are equivalent and can be used intercheangly, the context (point in time, nature of the text, etc) usually dictates one preferred form. For example, the forms क्ष and क्‌ष can be understood by all readers, but the form क्ष is pretty much the norm. On the other hand, the form ङ्ग is mostly found in older texts, and the form ङ्‌ग is more commonly found in modern texts. The historical trend is to abandon the use of fully-joined forms for the rare clusters.

¶78 A rendering system can render a cluster with any of the possible forms, and should normally use what is considered the preferred form. Some degree of control is achieved by the selection of a font appropriate for the nature of the text, much like one may wish to select a font with “ct” and “st” ligatures for an 18th century text.

¶79 The author of a text can exert some degree of control over which form should be used by the rendering system, by inserting the characters U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER appropriately. The general rules are as follows.

¶80 Without any joiner, no preference is expressed, and the rendering system should select the fully-joined form if it exists in the font.

क ka • ष ṣa <0915, 094D, 0937> क्ष preferred, then क्‌ष, then क्‌ष

क ka • र ra <0915, 094D, 0930> क्र preferred, then क्‌र, then क्‌र

क ka • न na <0915, 094D, 0928> क्न preferred, then क्‌न

ङ ṅa • ग ga <0919, 094D, 0917> ङ्ग preferred, then ङ्‌ग

¶81 With a ZWJ between a consonant and a consonant linker, the fully-joined form of the cluster is discouraged. The cluster should be rendered preferrably with a conjoined form, with the consonant before the ZWJ in full form and the consonant after the consonant linker in the a reduced form. If no such form is possible, then a non-joined form should be used.

क ka • ष ṣa <0915, 200D, 094D, 0937> क्‌ष

क ka • र ra <0915, 200D, 094D, 0930> क्र preferred, then क्‌र

क ka • न na <0915, 200D, 094D, 0928> क्‌न

ङ ṅa • ग ga <0919, 200D, 094D, 0917> ङ्‌ग

¶82 With a ZWJ between a consonant linker and a consonant, the fully-joined form of the cluster is discouraged. The cluster should be rendered preferrably with a conjoined form, with the consonant after the ZWJ in full form and the consonant before the consonant linker in the a reduced form. If no such form is possible, then a non-joined form should be used.

क ka • ष ṣa <0915, 094D, 200D, 0937> क्ष preferred, then क्‌ष

क ka • र ra <0915, 094D, 200D, 0930> क्र preferred, then क्‌र

क ka • न na <0915, 094D, 200D, 0928> क्न preferred, then क्‌न

ঙ ṅa • গ ga <0919, 094D, 200D, 0917> ঙ্গ

¶83 With a ZWNJ between a consonant linker and a consonant, the fully-joined form and the conjoined form of the cluster are discouraged. The cluster should be rendered with both consonants in full form, and a consonant killer sign.

ক ka • ষ ṣa <0915, 094D, 200C, 0937> ক্ষ

ক ka • র ra <0915, 094D, 200C, 0930> ক্র

ক ka • ন na <0915, 094D, 200C, 0928> ক্ন

ঙ ṅa • গ ga <0919, 094D, 200C, 0917> ঙ্গ

¶84 The rules above are generally applicable to all the scripts covered in this document with the notable exception of Sinhala.

¶85 ***Use of Joiners for Vowel Signs.*** A similar situtation exists for the rendering of vowel signs in a cluster. Two broad forms can be identified:

- ¶86 fully-joined forms, where the vowel sign and the rest of the cluster merge into a unique shape
- ¶87 conjoined forms, where the general diacrictic form is used

¶88 Here are illustrations of the two forms:

	fully-joined	conjoined
গ ga • ু u	গু	গু
ক ka • ু u		কু

¶89 Again, the joiners characers can be used to encourage one or another rendering:

¶90 Without any joiner, no preference is expressed, and the rendering system should select the most common form.

গ ga • ু u <0997, 09C1> গু fallback to গু

ক ka • ু u <0995, 09C1> কু

¶91 With a ZWJ between the consonant and the vowel sign, the cluster should be rendered in a fully-joined form.

গ ga • ু u <0997, 200D, 09C1> গু fallback to গু

ক ka • ু u <0995, 200D, 09C1> কু

¶92 With a ZWNJ between the consonant and the vowel sign, the cluster should be rendered in a conjoined form.

গ ga • ু u <0997, 200C, 09C1> গু fallback to গু

ক ka • ু u <0995, 200C, 09C1> কু

¶93 ***Use of Joiners for Typographic Fragments.*** Unicode also supports the representation of typographic fragments, such as the half form ক of the letter ক ka. This can be useful in didactic materials. This support is typically limited to the half forms of consonant letters and is achieved using the joiners. In general, the approach is to use sequences of the form <consonant, consonant linker, ZWJ> for consonant which take a reduced form when first in a cluster, and <U+0020 SPACE, ZWJ, consonant linker, consonant> for consonants which take a reduced form when second in a cluster. The script descriptions below will give the precise list of sequences which are recognized by the Unicode standard.

¶94 ***Other Uses of Joiners.*** In a few situations, U+200C ZERO WIDTH NON-JOINER and U+200D ZERO WIDTH JOINER are used for other purposes than those listed above. Unlike the cases listed so far, some of those situations correspond to semantic differences.

3.3 The Subjoined Model

¶95 The subjoined model is used for Tibetan and to a large extent for Myanmar.

¶96 Most of the scripts using the subjoined model do not have vowel letters, and only vowel signs.

¶97 ***Coded Characters.*** In the subjoined model, we have:

- ¶98 two or more coded characters for each consonant; for example:

  ཀ ka → U+0F40 ཀ TIBETAN LETTER KA

  → U+0F90 ྐ TIBETAN SUBJOINED LETTER KA

  ར ra → U+0F62 ར TIBETAN LETTER RA

  → U+0FB2 ྲ TIBETAN SUBJOINED LETTER RA

- ¶99 one coded character for each vowel which can appear at a *vowel sign*; for example:

  ི i → U+0F72 ི TIBETAN VOWEL SIGN I

- ¶100 one coded character for each diacritic sign; for example:

  ྃ → U+0F83 ྃ TIBETAN SIGN SNA LDAN

¶101 A cluster such as ཀ ka • ར ra • ི i → kri is represented by the sequence of coded characters:

U+0F40 ཀ TIBETAN LETTER KA

U+0FB2 ྲ TIBETAN SUBJOINED LETTER RA

U+0F72 ི TIBETAN VOWEL SIGN I

3.4 The Visual Model

¶102 The visual model is used where there is a well-established legacy practice: Thai, Lao.

4 Commonalities in Encoding

4.1 The ISCII Standard

¶103 A number of blocks in the Unicode Standard are based on ISCII-1988 (Indian Script Code for Information Interchange). The ISCII standard of 1988 differs from and is an update of earlier ISCII standards issued in 1983 and 1986.

¶104 The Unicode Standard encodes the characters in the same relative positions as those coded in positions A0-F4$_{16}$ in the ISCII-1988 standard. The same character code layout is followed for nine Indic scripts in the Unicode Standard: Devanagari, Bengla, Gurmukhi, Gujarati, Oriya, Tamil, Telugu, Kannada, and Malayalam. This parallel code layout emphasizes the structural similarities of the Brahmi

scripts and follows the stated intention of the Indian coding standards to enable one-to-one mappings between analogous coding positions in different scripts in the family. Sinhala, Tibetan, Thai, Lao, Khmer, Myanmar, and other scripts depart to a greater extent from the Devanagari structural pattern, so the Unicode Standard does not attempt to provide any direct mappings for these scripts to the Devanagari order.

¶105 In November 1991, at the time *The Unicode Standard, Version 1.0*, was published, the Bureau of Indian Standards published a new version of ISCII in Indian Standard (IS) 13194:1991. This new version partially modified the layout and repertoire of the ISCII-1988 standard. Because of these events, the Unicode Standard does not precisely follow the layout of the current version of ISCII. Nevertheless, the Unicode Standard remains a superset of the ISCII-1991 repertoire. Modern texts encoded with ISCII-1991 may be automatically converted to Unicode code points and back to their original encoding without loss of information.

4.2 Letters with Nukta

¶106 The scripts which use a *nukta* to create new letters encode their own nukta character.

¶107 Some, but not all, combinations of a base letter with a nukta which are actually used are also encoded as individual coded characters. All those coded characters have a canonical decomposition into the coded character for the base letter and the coded character for the nukta. Either the individual coded characters or their decompositions can be used.

4.3 Dandas

¶108 The *danda* and *double danda* are encoded only once at U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA. Those two coded characters, despite their location in the Devanagari block and the word DEVANAGARI in their names, are used regardless of the script.

4.4 Spoken Versus Written

¶109 While the basic inventory is strongly aligned with pronunciation, there are numerous cases of a disconnect between the spoken and the written. In Hindi, many words do not end with a final /a/ sound, yet they are written as if they did: राम rāma is pronounced /rām/; the same can occur inside words: सरकार sarakāra is pronounced /sarkār/.

¶110 The Unicode standard is first and foremost an encoding of the written signs rather than an encoding of the sounds. Therefore, सरकार is represented by <स sa, र ra, क ka, ा -ā, र ra>, eventhough there is no /a/ sound after the first /r/. The more exact representation of the sounds, <स sa, र ra, • क ka, ा -ā, र ra> would actually look like सर्कार sarkāra, but is not the common way to write the word.

¶111 Another manifestation of the predominance of the written can be seen with the nasal consonants and the anusvara. The Oriya word ଅଙ୍କ aṅka can also be written with an anusvara: ଅଂକ aṅka. Those two written forms are represented in Unicode by different sequences: <ଅ a, ଙ ṅa • କ ka> for the first, and <ଅ a, ଂ ṁ, କ ka> for the second.

¶112 This is very similar to the way “color” and “colour” are two different spellings of the same English word, and have different representations matching their written form.

5 ?

5.1 Collation

¶113 todo

5.2 Security

¶114 todo

5.3 Identifiers

¶115 todo

6 South Asian Scripts

6.1 Devanagari

¶116 The Devanagari script is used for writing classical Sanskrit and its modern historical derivative, Hindi. Extensions to the Sanskrit repertoire are used to write other related languages of India (such as Marathi) and of Nepal (Nepali). In addition, the Devanagari script is used to write the following languages: Awadhi, Bagheli, Bhatneri, Bhili, Bihari, Braj Bhasha, Chhattisgarhi, Garhwali, Gondi (Betul, Chhindwara, and Mandla dialects), Harauti, Ho, Jaipuri, Kachchhi, Kanauji, Konkani, Kului, Kumaoni, Kurku, Kurukh, Marwari, Mundari, Newari, Palpa, and Santali.

¶117 Devanagari is written horizontally from left to right, and words are separated by space.

¶118 ***Standards.*** The Devanagari block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶119 ***Encoding.*** The encoding of the Devanagari script uses the consonant linking model. The consonant linker coded character is U+094D ◌् DEVANAGARI SIGN VIRAMA.

6.1.1 Letters

¶120 Devanagari has the full complement of basic letters presented in section *2.1, Letters*.

¶121 The vowels:

short			long		
letter		sign	letter		sign
अ a 0905			आ ā 0906	◌ा -ā 093E
इ i 0907	◌ि -i 093F		ई ī 0908	◌ी -ī 0940
उ u 0909	◌ु -u 0941		ऊ ū 090A	◌ू -ū 0942
ऐ e 090E	◌ॅ -e 0946		ए ē 090F	◌े -ē 0947
ओ o 0912	◌ो -o 094A		ओ ō 0913	◌ो -ō 094B

¶122 Other vowels:

letter	sign	
ऐ ai 0910	◌ै -ai 0948	
औ au 0914	◌ौ -au 094C	
ऍ ĕ 090D	◌ॆ -ĕ 0945	todo usage?
ऑ ŏ 0911	◌ॉ -ŏ 0949	todo usage?
ॲ ॲ 0972		todo usage?
ऄ 0904		todo usage?
	◌ॎ ◌ॏ 0955	for Avestan

ॎ ॎ	094E	Prishthamatra; see below	

¶123 Sanskrit:

short		long	
letter	sign	letter	sign
ऋ ṛ 090B	◌ृ -ṛ 0943	ॠ ṝ 0960	◌ॢ -ṝ 0962
ऌ ḷ 090C	◌ॄ -ḷ 0944	ॡ ḹ 0961	◌ॣ -ḹ 0963

¶124 The plosive consonants:

	Voiceless		Voiced		Nasals
	unaspirated	aspirated	unaspirated	aspirated	
velar	क ka 0915	ख kha 0916	ग ga 0917	घ gha 0918	ङ ṅa 0919
palatal	च ca 091A	छ cha 091B	ज ja 091C	झ jha 091D	ञ ña 091E
retroflex	ट ṭa 091F	ठ ṭha 0920	ड ḍa 0921	ढ ḍha 0922	ण ṇa 0923
dental	त ta 0924	थ tha 0925	द da 0926	ध dha 0927	न na 0928
labial	प pa 092A	फ pha 092B	ब ba 092C	भ bha 092D	म ma 092E

¶125 The sonorant and fricative consonants:

	Sonorants	Fricatives
palatal	य ya 092F	श śa 0936
retroflex	र ra 0930	ष ṣa 0937
dental	ल la 0932	स sa 0938
labial	व va 0935	

¶126 Other consonants:

ॺ ॺ	097A	HEAVY YA	todo
ॹ ॹ	0979	ZHA	for Avestan
ॻ ॻ	097B	GGA	for Sindhi
ॼ ॼ	097C	JJA	for Sindhi
ॾ ॾ	097E	DDDA	for Sindhi
ॿ ॿ	097F	BBA	for Sindhi
ॽ ॽ	097D	GLOTTAL STOP	for Limbu

¶127 Signs:

	Signs
vowel nasalization	◌ँ m̐ 0901 candrabindu
consonant nasalization	◌ं ṁ 0902 anusvara

unvoiced aspiration ◌ः ḥ 0903 visarga

¶128 Forms with nukta:

◌ 093C nukta

क़ qa 0958 = 0915, 093C

ख़ k̲h̲a 0959 = 0916, 093C

ग़ ġa 095A = 0917, 093C

ज़ za 095B = 091C, 093C

ड़ ṛa 095C = 0921, 093C

ढ़ ṛha 095D = 0922, 093C

फ़ fa 095E = 092B, 093C

य़ ẏa 095F = 092F, 093C

¶129 Nepali and Marathi use an additional letter ॱ called the *eyelash ra*, to represent an /r/ sound. This letter is used contrastively with the letter र ra: for example Marathi दर्या daryā "ocean" vs. दऱ्या daṟyā "valleys". For legacy reasons, Unicode does not encode a separate character for the eyelash ra; instead, it is represented by the sequence <U+0931 ऱ DEVANAGARI LETTER RRA, U+094D ◌् DEVANAGARI SIGN VIRAMA>. For compatibility with Unicode 2.0, the sequence <U+0930 र DEVANAGARI LETTER RA, U+094D ◌् DEVANAGARI SIGN VIRAMA, U+200D ZERO WIDTH JOINER> can also be supported. Although the eyelash ra is represented as if it were a half form of either र or ऱ, it is actually a separate letter.

¶130 The Sindhi language uses implosive sounds which are written using the following four characters: U+097B ग़ DEVANAGARI LETTER GGA, U+097C ज़ DEVANAGARI LETTER JJA, U+097E ड़ DEVANAGARI LETTER DDDA and U+097F ब़ DEVANAGARI LETTER BBA. Implementations may recognize the previously recommended sequences using U+093C ◌़ DEVANAGARI SIGN NUKTA or U+0952 ◌॒ DEVANAGARI STRESS SIGN ANUDATTA, but authors are strongly encouraged to use the specific characters listed above.

¶131 In older texts, alternate rendering for the vowel signs e, ai, o and au are sometime used. Those are represented using U+094E ◌ॎ DEVANAGARI VOWEL SIGN PRISHTHAMATRA E in combination with other vowel signs:

e ◌ॎ <094E>

ai ◌ॎे <094E 0947>

o ◌ॎा <094E 093E>

au ◌ॎो <094E 094B>

¶132 Note that this is an exception to the general rule that vowel signs are represented atomically. Because of the combining class of the characters involved, to avoid ambiguity in representation U+094E should be first.

¶133 ***Marathi Allographs.*** In Marathi and some South Indian orthographies, variant glyphs are preferred for U+0932 ल DEVANAGARI LETTER LA and U+0936 श DEVANAGARI LETTER SHA:

¶134 todo figure 9.9

6.1.2 Digits and Numbers

¶135 Devanagari has a full set of decimal digits, encoded at U+0966 ० DEVANAGARI DIGIT ZERO .. U+096F ९ DEVANAGARI DIGIT NINE.

० 0 0966 ५ 5 096B

१ 1 0967 ६ 6 096C

२ 2 0968 ७ 7 096D

३ 3 0969 ८ 8 096E

४ 4 096A ९ 9 096F

¶136 There are no additional number characters.

6.1.3 Punctuation and Signs

¶137 The danda । is known as *purna viram* in Hindi, and the double danda ॥ is known as *deergh viram* in Hindi. Devanagari uses the common characters encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶138 The *avagraha* is encoded as U+093D ऽ DEVANAGARI SIGN AVAGRAHA.

¶139 The *om sign* is encoded as U+0950 ॐ DEVANAGARI OM.

¶140 ॰ appears after letters or combinations of letters and marks the sequence as an abbreviation. It is encoded as U+0970 ॰ DEVANAGARI ABBREVIATION SIGN.

¶141 ◌॑ is used for todo and is encoded as U+0951 ◌॑ DEVANAGARI STRESS SIGN UDATTA.

¶142 ◌॒ is used for todo and is encoded as U+0952 ◌॒ DEVANAGARI STRESS SIGN ANUDATTA.

¶143 U+0953 ◌॓ DEVANAGARI GRAVE ACCENT and U+0954 ◌॔ DEVANAGARI ACUTE ACCENT were originally encoded for todo . Instead, U+xx and U+yy are recommended for those uses.

¶144 todo U+0971 ॱ DEVANAGARI SIGN HIGH SPACING DOT

¶145 ***Bodo.*** The orthography of Bodo uses ’ as a tone mark. This sign is called गोजौ कमा (gojau kamaa) and is represented using U+02BC ’ MODIFIER LETTER APOSTROPHE. For example: खर’ (head), दख’ना (dress commonly worn by Bodo women).

¶146 ***Dogri.*** The orthography of Dogri uses ’ as a tone mark. This sign is called सुर-चि’न्ह (sur chinha) and is represented using U+02BC ’ MODIFIER LETTER APOSTROPHE. It occurs after short vowels (may be inherent) and indicates a high-falling tone: ख’ल्ल (down), कु’न (who), ति’लकना (to slip). After long vowels, a high-falling tone is written using U+0939 ह DEVANAGARI LETTER HA.

¶147 The avagraha ऽ is used to indicate extra-long vowels; for example, तला (sole) versus तलाऽ (pond).

¶148 ***Maithili.*** The orthography of Maithili uses ’ to denote the prolongation of a short अ a and to denote the truncation of words. This sign is called बिकारी कामा (bikari kaamaa) and is represented using U+02BC ’ MODIFIER LETTER APOSTROPHE. For example:

कहाँ जहे हो - कत’ जा रहल छें (prolongation); कत’ पड़ा’ गेल’, abbreviation for कतए पड़ाए गेलह (where did you go away?).

¶149 The avagraha ऽ, called बिकारी (bikari), is used to indicate extra-long vowels.

¶150 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ॲ	0904	<0905, 0946>
आ	0906	<0905, 093E>
ऊ	090A	<0909, 0941>
ऍ	090D	<090F, 0945>
ऐ	090E	<090F, 0946>
ऐ	0910	<090F, 0947>
ऑ	0911	<0905, 0949>
ओ	0912	<0905, 094A>
ओ	0913	<0905, 094B>
औ	0914	<0905, 094C>
ॅ	0972	<0905, 0945>

6.1.4 Typographic clusters

¶151 ***Typographic Cluster Cores.*** Here are some of the common typographic clusters made of two consonants which exhibit a fully-joined form:

क ka • क ka → क्क kka

क ka • त ta → क्त kta

क ka • र ra → क्र kra

क ka • ष ṣa → क्ष kṣa

ङ ṅa • क ka → ङ्क ṅka

ङ ṅa • ख kha → ङ्ख ṅkha

ङ ṅa • ग ga → ङ्ग ṅga

ङ ṅa • घ gha → ङ्घ ṅgha

ञ ña • ज ja → ञ्ज ñja

ज ja • ञ ña → ज्ञ jña

द da • घ gha → द्ध dgha

द da • द da → द्द dda

द da • ध dha → द्ध d'dha

द da • ब ba → द्ब dba

द da • भ bha → द्भ dbha

द da • म ma → द्म dma

द da • य ya → द्य dya

द da • व va → द्व dva

ट ṭa • ट ṭa → ट्ट ṭṭa

ट ṭa • ठ ṭha → ट्ठ ṭṭha

ठ ṭha • ठ ṭha → ठ्ठ ṭhṭha

ड ḍa • ग ga → ड्ग ḍga

ड ḍa • ड ḍa → ड्ड ḍḍa

ड ḍa • ढ ḍha → ड्ढ ḍḍha

त ta • त ta → त्त tta

त ta • र ra → त्र tra

न na • न na → न्न nna

फ pha • र ra → फ्र phra

श śa • र ra → श्र śra

ह ha • म ma → ह्म hma

ह ha • य ya → ह्य hya

ह ha • ल la → ह्ल hla

ह ha • व va → ह्व hva

¶152 In the absence of fully-joined form, the core of a typographic cluster is displayed in a conjoined form, where all but the last consonant are in *half form* and the last consonant is its full form.

¶153 The half form of the consonants क ka and फ pha is obtained by dropping the fragment on the right side of the stem: क् k and फ् ph.

क ka • ख kha → क्ख kkha

क ka • ळ ḷa → क्ळ kḷa

¶154 The half form of the consonants with a vertical stem on the right (ख ग घ च ज झ ञ ण त थ ध न ऩ प ब भ म य ल व श ष स) is obtained by dropping their vertical stem: ख् ग् घ् च् ज् झ् ञ् ण् त् थ् ध् न् ऩ् प् ब् भ् म् य् ल् व् श् ष् स्.

ग ga • ट ṭa → ग्ट gṭa

ग ga • ळ ḷa → ग्ळ gḷa

ण ṇa • ट ṭa → ण्ट ṇṭa

ण ṇa • ठ ṭha → ण्ठ ṇṭha

¶155 The remaining consonants (ङ छ ट ठ ड ढ द ळ ह) do not have half form. When they are present in non-final position in a cluster core,

the non-joined form of the cluster is displayed, with a *halant* sign below all but the last consonant.

ट ṭa • स sa → ट्स ṭsa

¶156 ***Satellite Consonants.*** र ra in initial position is a satellite consonant. It is then displayed as a small hook, called a *repha*, on the top right side of the cluster, including any satellite vowels:

र ra • त ta → र्त rta

र ra • क ka • ख kha → र्क्ख rkkha

र ra • म ma • आ ā → र्मा rmā

र ra • थ tha • ई ī → र्थी rthī

र ra • म ma • ओ ō • ं ṁ → र्मों rmōṁ

¶157 However, when a cluster is displayed in a non-conjoined form (either because such form is the most natural form or because it was requested by the use of a ZWNJ), the *repha* is displayed on the first consonant: र ra • क ka • क ka → ‍र्क्क rkka. todo wrong rendering

¶158 र ra in final position is a satellite consonant. It is displayed as small stroke attached to the vertical stem, or as a circumflex-like sign, called a *vattu*, below the consonants which do not have a stem.

ग ga • र ra → ग्र gra

प pa • र ra → प्र pra

ङ ṅa • र ra → ङ्र dra

ट ṭa • र ra → ट्र ṭra

¶159 ***Satellite Vowels.*** Some typographic clusters with vowel signs have full-joined forms:

र ra • उ u → रु ru

र ra • ऊ ū → रू rū

ह ha • ॠ ṛ → ह्ॠ hṛ

¶160 Otherwise, the vowel sign attaches around the cluster core.

			on क ka	on र्क rka
आ ā	◌ा	-ā	का kā	र्का rkā
इ i	◌ि	-i	कि ki	र्कि rki
ई ī	◌ी	-ī	की kī	र्की rkī
उ u	◌ु	-u	कु ku	र्कु rku
ऊ ū	◌ू	-ū	कू kū	र्कू rkū
ऍ ĕ	◌ॅ	-ĕ	कॅ kĕ	र्कॅ rkĕ
ऐ e	◌ै	-e	के ke	र्के rke
ए ē	◌े	-ē	के kē	र्के rkē

ऐ ai ै -ai कै kai र्कै rkai

ऑ ŏ ॅ -ŏ कॉ kŏ र्कॉ rkŏ

ओ o ो -o को ko र्को rko

ओ ō ो -ō को kō र्को rkō

औ au ौ -au कौ kau र्कौ rkau

ऋ ṛ ृ -ṛ कृ kṛ र्कृ rkṛ

ॠ ṝ ॄ -ṝ कॄ kṝ र्कॄ rkṝ

ऌ ḷ ॢ -ḷ कॢ kḷ र्कॢ rkḷ

ॡ ḹ ॣ -ḹ कॣ kḹ र्कॣ rkḹ

¶161 There is no vowel sign for ऄ and अ a.

¶162 The vowel sign for इ i is generally displayed on the left side of its typographic cluster: क ka • इ i → कि ki; क ka • ट ṭa • इ i → क्टि kṭi. When a cluster is displayed in non-joined form (either because such form is the most natural form or because it was requested by the use of a ZWNJ), this vowel sign is displayed just before the last consonant of the cluster: क ka • क ka • इ i → क्कि kki.

¶163 ***Special Clusters.*** In a cluster formed of र ra and a vocalic vowel, the र ra is displayed as a *repha* on the vowel letter:

र ra • ऋ ṛ → र्ऋ rṛ <0930, 0943>

र ra • ॠ ṝ → र्ॠ rṝ <0930, 0944>

र ra • ऌ ḷ → र्ॡ rḹ <0930, 0962>

र ra • ॡ ḹ → र्ॡ rḹ <0930, 0963>

¶164 ***Conjoined and Non-Joined Forms of Clusters.*** The display of a cluster can be encouraged to be in conjoined form (with fallback to a non-joined form if necessary) by inserting a U+200D ZERO WIDTH JOINER after the consonant linker coded character.

¶165 The display of a cluster can be encouraged to be in non-joined form by inserting a U+200C ZERO WIDTH NON-JOINER after the consonant linker coded character.

x • *y*	*x*, 094D, *y*	*x*, 094D, 200D, *y*	*x*, 094D, 200C, *y*
क ka • क ka	→ क्क kka	क्क kka	क्क kka
क ka • त ta	→ क्त kta	क्त kta	क्त kta

¶166 The display of vowel signs can be encouraged to be in full-joined form by inserting a U+200D ZERO WIDTH JOINER before them, and encouraged to be in conjoined form by inserting a U+200C ZERO WIDTH NON-JOINER before them.

x • *y*	*x*, y_{vs}	*x*, 200D, y_{vs}	*x*, 200C, y_{vs}
र ra • उ u	→ रु rau or ऱु rau	रु rau	ऱु rau
र ra • ऊ ū	→ रू raū or ऱू raū	रू raū	ऱू raū
ह ha • ऋ ṛ	→ ह्र haṛ or ह्ऱ haṛ	ह्र haṛ	ह्ऱ haṛ

¶167 ***Typographic Fragments.*** The following typographic fragments can be represented as follows:

Full form	Half form	Represented by
क	क्	<0915, 094D, 200D>
ख	ख्	<0916, 094D, 200D>
ग	ग्	<0917, 094D, 200D>
घ	घ्	<0918, 094D, 200D>
च	च्	<091A, 094D, 200D>
ज	ज्	<091C, 094D, 200D>
झ	झ्	<091D, 094D, 200D>
ञ	ञ्	<091E, 094D, 200D>
ण	ण्	<0923, 094D, 200D>
त	त्	<0924, 094D, 200D>
थ	थ्	<0925, 094D, 200D>
ध	ध्	<0927, 094D, 200D>
न	न्	<0928, 094D, 200D>
ऩ	ऩ्	<0929, 094D, 200D>
प	प्	<092A, 094D, 200D>
फ	फ्	<092B, 094D, 200D>
ब	ब्	<092C, 094D, 200D>
भ	भ्	<092D, 094D, 200D>
म	म्	<092E, 094D, 200D>
य	य्	<092F, 094D, 200D>
ल	ल्	<0932, 094D, 200D>
व	व्	<0935, 094D, 200D>
श	श्	<0936, 094D, 200D>
ष	ष्	<0937, 094D, 200D>
स	स्	<0938, 094D, 200D>
क्ष	क्ष्	<0915, 094D, 0937, 094D, 200D>
ज्ञ	ज्ञ्	<091c, 094D, 091e, 094D, 200D>
त्त	त्त्	<0924, 094D, 0924, 094D, 200D>
त्र	त्र्	<0924, 094D, 0930, 094D, 200D>
श्र	श्र्	<0936, 094D, 0930, 094D, 200D>

¶168 do we want the "half YA in post form"? subjoined tta, ttha, ga, dha, bha?

¶169 the last five entries are those in Table 9.4; however, that table is not meant to be exhaustive; should we have more?

¶170 the last two entries actually show some ambiguity: we say that <930, 94D, 200D> is to be rendered by an eyelash ra (rule R5a, for compatibility with Unicode 2.0); so which takes precedence? The eyelash ra or the half form of the conjunct?

6.2 Bengla

¶171 The Bengla script is a North Indian script closely related to Devanagari. It is used to write the Bengla language primarily in the West Bengal state and in the nation of Bangladesh. It is also used to write Assamese in Assam and a number of other minority languages, such as Bishnupriya Manipuri, Daphla, Garo, Hallam, Khasi, Mizo, Munda, Naga, Rian, and Santali, in northeastern India.

¶172 Bengla is written horizontally from left to right, and words are separated by space.

¶173 In English, the term "Bengali" is commonly used instead of "Bengla", but the later term is preferred. This older term explains the formal name of the script and the character names in Unicode.

¶174 ***Standards.*** The Bengla block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶175 ***Encoding.*** The encoding of the Bengla script follows the consonant linking model. The consonant linker coded character is U+09CD ◌্ BENGALI SIGN VIRAMA.

¶176 ***Letters.*** Bengla has all the basic letters presented in section *2.1, Letters*, except the consonants ৱ va and ৳ ḷa.

short			long		
letter		sign	letter		sign
অ a 0985			আ ā 0986	◌া -ā	09BE
ই i 0987	◌ি -i	09BF	ঈ ī 0988	◌ী -ī	09C0
উ u 0989	◌ু -u	09C1	ঊ ū 098A	◌ূ -ū	09C2
এ ē 098F	◌ে -ē	09C7	ঐ ai 0990	◌ৈ -ai	09C8
ও ō 0993	◌ো -ō	09CB	ঔ au 0994	◌ৌ -au	09CC

¶177

short			long		
letter		sign	letter		sign
ঋ ṛ 098B	◌ৃ -ṛ	09C3	ৠ ṝ 09E0	◌ৄ -ḝ	09C4
ঌ ḷ 098C	◌ু -ṝ	09E2	ৡ ḹ 09E1	◌ু -ḝ	09E3

¶178

	Voiceless				Voiced				Nasals	
	unaspirated		aspirated		unaspirated		aspirated			
velar	ক ka	0995	খ kha	0996	গ ga	0997	ঘ gha	0998	ঙ ṅa	0999
palatal	চ ca	099A	ছ cha	099B	জ ja	099C	ঝ jha	099D	ঞ ña	099E
retroflex	ট ṭa	099F	ঠ ṭha	09A0	ড ḍa	09A1	ঢ ḍha	09A2	ণ ṇa	09A3
dental	ত ta	09A4	থ tha	09A5	দ da	09A6	ধ dha	09A7	ন na	09A8
labial	প pa	09AA	ফ pha	09AB	ব ba	09AC	ভ bha	09AD	ম ma	09AE

¶179

	Sonorants		Fricatives	
palatal	য ya	09AF	শ śa	09B6
retroflex	র ra	09B0	ষ ṣa	09B7
dental	ল la	09B2	স sa	09B8

¶180

Others	
হ ha	09B9

¶181

Signs			
ঁ m̐	0981	চন্দ্রবিন্দু	candrabindu
ং ṁ	0982	অনুস্বার	anusbāra
ঃ ḥ	0983	বিসর্গ	bisarga

¶182 The vowel nasalisation sign ঁ is known as the *candrabindu* (Bengla চন্দ্রবিন্দু candrabindu), and is encoded as U+0981 ঁ BENGALI SIGN CANDRABINDU.

¶183 The consonant nasalization sign ং is known as the *anusvara* (Bengla অনুস্বার anusbāra), and is encoded as U+0982 ং BENGALI SIGN ANUSVARA.

¶184 The unvoiced aspiration sign ঃ is known as the *visarga* (Bengla বিসর্গ bisarga), and is encoded as U+0983 ঃ BENGALI SIGN VISARGA.

¶185 ৎ t̲, called *khanda ta*, was originally a reduced form of ত ta. This letterform has been pressed in service as a letter on its own. For example, it can be seen in the Bengla word মৎ mat̲ "by me", and is in contrast with the letter ত ta: the Bengla word মত mata means "opinion". To accomodate that case, Unicode encodes U+09CE ৎ BENGALI LETTER KHANDA TA. In all known uses, ৎ t̲ appears as the only letter of a typographic cluster.

¶186 ৰ ra and ৱ ra are used to write Assamese, and are encoded as U+09F0 ৰ BENGALI LETTER RA WITH MIDDLE DIAGONAL and U+09F1 ৱ BENGALI LETTER RA WITH LOWER DIAGONAL respectively.

¶187 ***Digits and Numbers.*** Bengla has a full set of decimal digits, encoded at U+09E6 ০ BENGALI DIGIT ZERO .. U+09EF ৯ BENGALI DIGIT NINE.

¶188 todo something about the currency fractions 09F4..09F9

¶189 ***Punctuation and Signs.*** Bengla uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶190 The *avagraha* is encoded as U+09BD ঽ BENGALI SIGN AVAGRAHA.

¶191 The rupee mark ৲ is used todo and is encoded as U+09F2 ৲ BENGALI RUPEE MARK. It is in addition to the rupee sign ৳

which is encoded as U+09F3 ৳ BENGALI RUPEE SIGN.

¶192 ৺ is used todo and is encoded as U+09FA ৺ BENGALI ISSHAR.

¶193 ☐ is use todo and is encoded as U+09FB ☐ BENGALI GANDA MARK.

¶194 The coded character U+09D7 ◌ৗ BENGALI AU LENGTH MARK does not correspond to an abstract character, i.e., it is not used for the representation of text. It is encoded to represent the right side of the vowel sign for ঔ au in isolation, and to provide a canonical decomposition for U+09CC ◌ৌ BENGALI VOWEL SIGN AU.

¶195 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
আ	0986	<0985, 09BE>

¶196 ***Typographic Cluster Cores.*** Here are some of the common typographic clusters made of two consonants which exhibit a fully-joined form:

- ক ka • ক ka → ক্ক kka
- ক ka • ত ta → ক্ত kta
- ক ka • র ra → ক্র kra
- ক ka • ষ ṣa → ক্ষ kṣa
- গ ga • ধ dha → গ্ধ gdha
- ঙ ṅa • ক ka → ঙ্ক ṅka
- ঙ ṅa • গ ga → ঙ্গ ṅga
- জ ja • ঞ ña → জ্ঞ jña
- ঞ ña • চ ca → ঞ্চ ñca
- ঞ ña • ছ cha → ঞ্ছ ñcha
- ঞ ña • জ ja → ঞ্জ ñja
- ণ ṇa • ড ḍa → ণ্ড ṇḍa
- ত ta • ত ta → ত্ত tta
- ত ta • থ tha → ত্থ t'tha
- ত ta • র ra → ত্র tra
- দ da • দ da → দ্দ dda
- দ da • ধ dha → দ্ধ d'dha
- ন na • ত ta → ন্ত nta
- ব ba • দ da → ব্দ bda
- ব ba • ধ dha → ব্ধ bdha
- শ śa • চ ca → শ্চ śca

শ śa • ছ cha → শ্ছ ścha

ষ ṣa • ক ka → ষ্ক ṣka

ষ ṣa • ট ṭa → ষ্ট ṣṭa

ষ ṣa • প pa → ষ্প ṣpa

স sa • ক ka → স্ক ska

স sa • ত ta → স্ত sta

স sa • থ tha → স্থ stha

স sa • প pa → স্প spa

হ ha • ণ ṇa → হ্ণ hṇa

হ ha • ন na → হ্ন hna

হ ha • ম ma → হ্ম hma

¶197 In the absence of a fully-joined form, the core of a typographic cluster is displayed in a conjoined form, where all the but the last consonant are in *half form* and the last consonant is in its full form.

¶198 The reduced form of the consonants খ গ ণ ন প শ স is obtained by dropping the vertical stem and slightly reducing their size: খ্ গ্ ণ্ ন্ প্ শ্ স্.

¶199 The reduced form of the other consonants is a smaller version of them.

ক ka • খ kha → ক্খ kkha

ক ka • গ ga → ক্গ kga

গ ga • ট ṭa → গ্ট gṭa

গ ga • ড ḍa → গ্ড gḍa

¶200 Because every consonant has a half form, it is rare to encounter clusters in non-joined form, although they are in principle possible. This is displayed with a *hasant* sign below all but the last consonant:

ট ṭa • স sa → ট্স ṭsa

¶201 Not dealt with correctly yet: "clusters" of the form cons + hasant. May be this is the right place?

¶202 ***Satellite Consonants.*** র ra in initial position is a satellite consonant. It is then displayed as a slanted stoke on the top right side of the cluster core:

র ra • ত ta → র্ত rta

র ra • ক ka • খ kha → র্ক্খ rkkha

র ra • ম ma • আ ā → র্মা rmā

র ra • থ tha • ঈ ī → র্থী rthī

¶203 A number of consonants in final position are satellite consonants:

য ya ক্য kya খ্য khya গ্য gya ঘ্য ghya চ্য cya

র ra ক্র kra গ্র gra ঘ্র ghra জ্র jra ত্র tra
ল la ক্ল kla গ্ল gla ম্ল mla প্ল pla ল্ল lla

ব ba ক্ব kba জ্ব jba ত্ব tba দ্ব dba ন্ব nba

ন na গ্ন gna ন্ন nna ত্ন tna স্ন sna হ্ন hna

ম ma ক্ম kma ত্ম tma দ্ম dma ন্ম nma স্ম sma

¶204 In a cluster formed of র ra and one of য ya, র ra, ল la, ব ba, ন na, ম ma, the initial র ra adopts its satellite form and the second consonant adopts its full form. To encourage the initial র ra to have its full form and the second consonant to have its satellite form, one inserts a U+200D ZERO WIDTH JOINER between the র ra and the consonant linker.

র ra, 09CD, *y* র ra, 200D, 09CD, *y*

র্য rya র‍্য raya

র্র rra র‍্র rara

র্ল rla র‍্ল rala

র্ব rba র‍্ব raba

র্ন rna র‍্ন rana

র্ম rma র‍্ম rama

¶205 ***Satellite Vowels.*** Some typographic clusters with vowel signs have full-joined forms:

গ ga • ু u → গু gu

র ra • ু u → রু ru

র ra • ূ ū → রূ rū

শ śa • ু u → শু śu

হ ha • ু u → হু hu

হ ha • ৃ ṛ → হৃ hṛ

ৰ ra • ু u → ৰু rau

ৰ ra • ূ ū → ৰূ raū

ৱ ra • ু u → ৱু rau

ৱ ra • ূ ū → ৱূ raū

ত ta • র ra • ু u → ত্রু trau

ন na • ত ta • ু u → ন্তু ntau

স sa • ত ta • ু u → স্তু stau

¶206 Otherwise, the vowel sign attaches around the cluster core.

	on ক ka	on র্ক rka
আ ā া -ā	কা kā	র্কা rkā
ই i ি -i	কি ki	র্কি rki

ঈ ī ী -ī কী kī র্কী rkī

উ u ু -u কু ku র্কু rku

ঊ ū ূ -ū কূ kū র্কূ rkū

এ ē ে -ē কে kē র্কে rkē

ঐ ai ৈ -ai কৈ kai র্কৈ rkai

ও ō ো -ō কো kō র্কো rkō

ঔ au ৌ -au কৌ kau র্কৌ rkau

ঋ ṛ ৃ -ṛ কৃ kṛ র্কৃ rkṛ

ঋ̄ ṝ ৄ -ḷ কৄ kḷ র্কৄ rkḷ

ঌ ḷ ৢ -ṝ কৢ kṝ র্কৢ rkṝ

ঌ̄ ḹ ৣ -ḹ কৣ kḹ র্কৣ rkḹ

¶207 There is no vowel sign for অ a.

¶208 The vowel sign for ই i is generally displayed on the left side of its typographic cluster: ক ka • ই i → কি ki; ন na • গ ga • ই i → ন্‌গই n'ga'i. todo where when there is an hasant?

¶209 The vowel signs for এ ē and ঐ ai are similarly displayed.

¶210 The vowel sign for ও ō is made of two parts, one that is generally displayed on the left side of the typographic cluster and one that is display on the right side: ক ka • ও ō → কো kō; ন na • গ ga • ও ō → ন্‌গো n'gō. todo where when there is an hasant?

¶211 The vowel sign for ঔ au is similarly displayed in two parts.

¶212 ***Ya-phalaa.*** The form that য ya adopts when last in a conjunct, য, is known as *ya-phalaa*. This form is also used, followed by the vowel sign া ◌া ā, as a diacritic sign on the vowel letters অ a and এ ē; those combinations are used to transcribe accurately the vowel sounds [æ] and todo in loan words. Those combinations are represented by:

To represent	Use
অ্যা	<0985, 09CD, 09AF, 09BE>
এ্যা	<098F, 09CD, 09AF, 09BE>

¶213 ***Interaction of Repha and Ya-phalaa..*** todo

¶214 ***Conjoined and Non-Joined Forms of Clusters.*** The display of a cluster can be encouraged to be in conjoined form (with fallback to a non-joined form if necessary) by inserting a U+200D ZERO WIDTH JOINER after the consonant linker coded character.

¶215 The display of a cluster can be encouraged to be in non-joined form by inserting a U+200C ZERO WIDTH NON-JOINER after the consonant linker coded character.

x • *y*		*x*, 09CD, *y*	*x*, 09CD, 200D, *y*	*x*, 09CD, 200C, *y*
ক ka • ক ka	→ ক্ক kka	ক্‌ক kka	ক্‌ক kka
ক ka • ত ta	→ ক্ত kta	ক্‌ত kta	ক্‌ত kta
	→			

ক ka • খ kha ক্খ kkha ক্খ kkha ক্খ kkha

¶216 The display of vowel signs can be encouraged to be in full-joined form by inserting a U+200D ZERO WIDTH JOINER before them, and encouraged to be in conjoined form by inserting a U+200C ZERO WIDTH NON-JOINER before them.

x • *y*	*x*, y_{vs}	*x*, 200D, y_{vs}	*x*, 200C, y_{vs}
গ ga • ঊ u	→ গু gau or গু gau	গু gau	গু gau
র ra • ঊ u	→ রু rau or রু rau	রু rau	রু rau
র ra • ঊ ū	→ রূ raū or রূ raū	রূ raū	রূ raū
শ śa • ঊ u	→ শু śau or শু śau	শু śau	শু śau
হ ha • ঊ u	→ হু hau or হু hau	হু hau	হু hau
হ ha • ঋ ṛ	→ হৃ haṛ or হৃ haṛ	হৃ haṛ	হৃ haṛ
ৰ ra • ঊ u	→ রু rau or রু rau	রু rau	রু rau
ৰ ra • ঊ ū	→ রূ raū or রূ raū	রূ raū	রূ raū
ৱ ra • ঊ u	→ রু rau or রু rau	রু rau	রু rau
ৱ ra • ঊ ū	→ রূ raū or রূ raū	রূ raū	রূ raū
ত ta • র ra • ঊ u	→ ত্রু trau or ত্রু trau	ত্রু trau	ত্রু trau
ন na • ত ta • ঊ u	→ ন্তু ntau or ন্তু ntau	ন্তু ntau	ন্তু ntau
স sa • ত ta • ঊ u	→ স্তু stau or স্তু stau	স্তু stau	স্তু stau

¶217 ***Typographic Fragments.*** The following typographic fragments be represented as follows:

Full form Half form Represented by

খ খ্ <0996, 09CD, 200D>

more...

¶218 The left side of the vowel signs ো -ō and ৌ -au are represented in isolation by <0020, 09C7> → ে

¶219 The right side of the vowel sign ো -ō is represented in isolation by <0020, 09BE> → া.

¶220 The right side of the vowel sign ৌ -au is represented in isolation by <0020, 09D7> → ৗ. The coded character U+09D7 ৌ BENGALI AU LENGTH MARK does not correspond to an abstract character and is encoded solely to provide a canonical decomposition for U+09CC ৌ BENGALI VOWEL SIGN AU, and to represent its right side in isolation.

6.3 Gurmukhi

¶221 The Gurmukhi script is a North Indian script used to write the Punjabi (or Panjabi) language of the Punjab state of India. Gurmukhi, which literally means "proceeding from the mouth of the Guru," is attributed to Angad, the second Sikh Guru (1504 – 1552 CE). It is derived from an older script called Landa, and is closely related to Devanagari structurally. The script is closely associated with Sikh and Sikhism, but it is used on an everyday basis in East Punjab. (West Punjab, now in Pakistan, uses the Arabic script.)

¶222 Gurmukhi is written horizontally from left to right, and words are separated by space.

¶223 ***Standards.*** The Gurmukhi block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶224 ***Encoding.*** The encoding of the Gurmukhi script follows the consonant linking model. The consonant linker coded character of Gurmukhi is U+0A4D ◌੍ GURMUKHI SIGN VIRAMA.

¶225 ***Letters.*** Gurmukhi has all the basic letters presented in section *2.1, Letters* , except the consonant ਖ਼ ṣa, and the vocalic vowels ਋ r̥, ਋ੲ r̥̄, ਌ l̥ and ਌ੲ l̥̄.

short			long		
letter		sign	letter		sign
ਅ a 0A05			ਆ ā 0A06	◌ਾ -ā 0A3E	
ਇ i 0A07	◌ਿ -i 0A3F		ਈ ī 0A08	◌ੀ -ī 0A40	
ਉ u 0A09	◌ੁ -u 0A41		ਊ ū 0A0A	◌ੂ -ū 0A42	
ਏ ē 0A0F	◌ੇ -ē 0A3E		ਐ ai 0A10	◌ੈ -ai 0A48	
ਓ ō 0A13	◌ੋ -ō 0A4B		ਔ au 0A14	◌ੌ -au 0A4C	

	Voiceless		Voiced		Nasals	
	unaspirated	aspirated	unaspirated	aspirated		
velar	ਕ ka 0A15	ਖ kha 0A16	ਗ ga 0A17	ਘ gha 0A18	ਙ ṅa	0A19
palatal	ਚ ca 0A1A	ਛ cha 0A1B	ਜ ja 0A1E	ਝ jha 0A1D	ਞ ña	0A1E
retroflex	ਟ ṭa 0A1F	ਠ ṭha 0A20	ਡ ḍa 0A21	ਢ ḍha 0A22	ਣ ṇa	0A23
dental	ਤ ta 0A24	ਥ tha 0A25	ਦ da 0A26	ਧ dha 0A27	ਨ na	0A28
labial	ਪ pa 0A2A	ਫ pha 0A2B	ਬ ba 0A2C	ਭ bha 0A2D	ਮ ma	0A2E

¶226 ੜ ṛa is encoded as U+0A5C ੜ GURMUKHI LETTER RRA.

¶227 The vowel nasalisation sign ◌ੱ is known as the *adak bindi*, and is encoded as U+0A01 ◌ੱ GURMUKHI SIGN ADAK BINDI

¶228 Two different marks are used for the consonant nasalization sign. ◌ं is known as the *bindi* and is encoded as U+0A02 ◌ं GURMUKHI SIGN BINDI. ◌ं is known as the *tippi* and is encoded as U+0A70 ◌ं GURMUKHI TIPPI. Present practice is to use bindi only after the vowel signs ◌ਾ -ā, ◌ੀ -ī,◌ੇ -ē, ◌ੈ -ai, ◌ੋ -ō, ◌ੌ -au, and after the vowel letters ਉ u and ਊ ū, and to use tippi in the other cases. However, older texts may follow different conventions.

¶229 The unvoiced aspiration sign ◌ः is known as the *visarga*, and is encoded as U+0A03 ◌ः GURMUKHI SIGN VISARGA.

¶230 The additional consonants with nukta (called *pairin bindi*, literally, "with a dot in the foot," in Punjabi) are primarily used to differentiate Urdu or Persian loan words. They include ਸ਼ śa and ਲ਼ l̤a, but do not include ੜ ṛa which is genuinely Punjabi.

¶231 The *yakash* sign ◌ੵ probably originated from a subjoined form of ਯ ya. However, because its usage is relatively rare and is not entirely predictable, it is encoded as a separate character, U+0A75 ◌ੵ GURMUKHI SIGN YAKASH. This character should occur after the consonant to which it attaches and before any vowel sign.

¶232 ***Digits and Numbers.*** Gurmukhi has a full set of decimal digits, encoded at U+0A66 ੦ GURMUKHI DIGIT ZERO .. U+0A6F ੯ GURMUKHI DIGIT NINE.

¶233 ***Punctuation and Signs.*** Gurmukhi uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶234 The sign ੱ indicates that the following consonant is geminate. It is encoded as U+0A71 ੱ GURMUKHI ADDAK. For example, the word ਪੱਗ paga, “turban”, can be represented with the sequence <U+0A2A, U+0A71, U+0A17>. todo

¶235 In dictionnaries and similar materials, words starting with ੁ u or ੂ ū are sorted together, under the heading ੳ. For the purpose of representing this heading, Unicode encodes U+0A73 ੳ GURMUKHI URA. This character should not be used to represent ੁ u and ੂ ū. Similarly, words starting with ਿ i or ੀ ī are sorted together, under the heading ੲ. Again, only for the purpose of representing that heading, Unicode encodes U+0A72 ੲ GURMUKHI IRI.

¶236 The religious symbol *khanda* sometimes used in Gurmukhi texts is encoded at U+262C ☬ ADI SHAKTI in the Miscellaneous Symbols block. U+0A74 ੴ GURMUKHI EK ONKAR, which is also a religious symbol, can have different presentation forms, which do not change its meaning. The font used in the code charts shows a highly stylized form; simpler forms look like the digit one, followed by a sign based on ura, along with a long upper tail.

¶237 In older texts, the *udaat* sign ◌ऀ indicates a high tone. It is encoded as U+0A51 ◌ੑ GURMUKHI SIGN UDAAT. This character should occur after the consonant to which is attaches and before any vowel sign.

¶238 ***Tones.*** Tones. The Punjabi language is tonal, but the Gurmukhi script does not contain any specific signs to indicate tones. Instead, the voiced aspirates (ਘ gha, ਝ jha, ਢ ḍha, ਧ dha, ਭ bha) and the letter ਹ ha combine consonantal and tonal functions.

¶239 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ਆ	0A06	<0A05, 0A3E>
ਇ	0A07	<0A72, 0A3F>
ਈ	0A08	<0A72, 0A40>
ਉ	0A09	<0A73, 0A42>
ਊ	0A0A	<0A73, 0A42>
ਏ	0A0F	<0A72, 0A47>
ਐ	0A10	<0A05, 0A48>
ਓ	0A13	<0A73, 0A4B>
ਔ	0A14	<0A05, 0A4C>

¶240 ***Typographic Cluster Cores.*** Gurmukhi is unusual in that there are no fully joined forms. In addition, in modern usage, the core of a typographic cluster is always a single consonant.

¶241 ***Satellite Consonants.*** In modern Punjabi, the consonants ਰ ra, ਹ ha and ਵ va are satellite consonants in final position, and take a subjoined (*pairin*) form.

ਪ pa • ਰ ra → ਪ੍ਰ pra

ਮ ma • ਹ ha → ਮ੍ਹ mha

ਦ da • ਵ va → ਦ੍ਵ dva

¶242 In modern Punjabi, the consonant ਯ ya is a satellite consonant in final position, and take a postjoined (*addha*) form.

ਦ da • ਯ ya → ਦ੍ਯ dya

¶243 In older texts, one can find more consonants with a *pairin* form:

ਸ sa • ਗ ga → ਸ੍ਗ sga
ਸ sa • ਚ ca → ਸ੍ਚ sca
ਸ sa • ਟ ṭa → ਸ੍ਟ sṭa
ਸ sa • ਠ ṭha → ਸ੍ਠ sṭha
ਸ sa • ਤ ta → ਸ੍ਤ sta
ਸ sa • ਦ da → ਸ੍ਦ sda
ਸ sa • ਨ na → ਸ੍ਨ sna
ਸ sa • ਥ tha → ਸ੍ਥ stha
ਸ sa • ਯ ya → ਸ੍ਯ sya

¶244 In older texts, one can find more consonants with an *addha* form:

ਸ sa • ਥ tha → ਸ੍ਥ stha
ਸ sa • ਮ ma → ਸ੍ਮ sma

¶245 Older texts also exhibit another feature which is not found in modern Punjabi: instead of showing the second consonant in satellite form, one can see the first consonant in a satellite form. For ਰ ra, this satellite form is similar to the Devanagari *reph*. For the other consonants, this is is a half form, generally formed by dropping the vertical stem. This conjoined form can be encouraged by inserting a U+200D ZERO WIDTH JOINER after the consonant linker. Inserting a U+200C ZERO WIDTH NON-JOINER encourages a non-joined form:

x • *y*	*x*, 0A4D, *y*	*x*, 0A4D, ZWJ, *y*	*x*, 0A4D, ZWNJ, *y*
ਰ ra • ਵ va → ਰ੍ਵ rva	ਰ੍ਵ rva	ਰ੍ਵ rva	ਰ੍ਵ rva
ਸ sa • ਵ va → ਸ੍ਵ sva	ਸ੍ਵ sva	ਸ੍ਵ sva	ਸ੍ਵ sva

¶246 ***Satellite Vowels.*** The vowel signs attach around the cluster core.

	on ਕ ka
ਆ ā	◌ਾ -ā	ਕਾ kā
ਇ i	◌ਿ -i	ਕਿ ki
ਈ ī	◌ੀ -ī	ਕੀ kī
ਉ u	◌ੁ -u	ਕੁ ku
ਊ ū	◌ੂ -ū	ਕੂ kū
ਏ ē	◌ੇ -ē	ਕੇ kē
ਐ ai	◌ੈ -ai	ਕੈ kai
ਓ ō	◌ੋ -ō	ਕੋ kō
ਔ au	◌ੌ -au	ਕੌ kau

¶247 There is no vowel sign for ਅ a.

¶248 In older texts, such as the *Sri Guru Granth Sahib* (the Sikh holy book), one can find typographic clusters with a vowel sign attached to a vowel letter, or with two vowel signs attached to a consonant. The most common cases are ੁ -u attached to ੳ ō, as in ਓਮਾਹਾ ōumāhā, and both the vowel signs ੋ -ō and ੁ -u attached to a consonant, as in ਗੌਬਿਂਦ gōubida; this is used to indicate the metrical shortening of /ō/ or the lengthening of /u/ depending on the context. Other combinations are attested as well, such as ਗ੍ਯਾਨ g'hiāna.

¶249 Because of the combining classes of the characters U+0A4B ੋ GURMUKHI VOWEL SIGN OO and U+0A41 ੁ GURMUKHI VOWEL SIGN U, the sequences <*consonant*, U+0A4B, U+0A41> and <*consonant*, U+0A41, U+0A4B> are not canonically equivalent. To avoid ambiguity in representation, the first sequence, with U+0A4B before U+0A41, should be used in such cases. More generally, when a consonant or independent vowel is modified by multiple vowel signs, the sequence ov vowel signs in the underlying representation of the text should be: left, top, bottom, right.

¶250 ***Conjoined and Non-Joined Forms of Clusters.*** The display of a cluster can be encouraged to be in non-joined form by inserting a U+200C ZERO WIDTH NON-JOINER after the consonant linker coded character.

x • *y*	*x*, 0A4D, *y*	*x*, 0A4D, 200C, *y*
ਕ ka • ਰ ra	→ ਕ੍ਰ kra	ਕ੍‌ਰ kra
ਕ ka • ਯ ya	→ ਕ੍ਯ kya	ਕ੍‌ਯ kya

6.4 Gujarati

¶251 The Gujarati script is a North Indian script closely related to Devanagari. It is most obviously distinguished from Devanagari by not having a horizontal bar for its letterforms, a characteristic of the older Kaithi script to which Gujarati is related. The Gujarati script is used to write the Gujarati language of the Gujarat state in India.

¶252 Gujarati is written horizontally from left to right, and words are separated by space. The consonant letters often have a vertical stem on their right side, which is graphically and historically related to the sign denoting the inherent /a/ vowel.

¶253 ***Standards.*** The Gujarati block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶254 ***Encoding.*** The encoding of the Gujarati script follows the consonant linking model. The consonant linker coded character of Gujarati is U+0ACD ◌્ GUJARATI SIGN VIRAMA.

¶255 ***Letters.*** Gujarati has all the basic letters presented in section *2.1, Letters*.

¶256 The vowel nasalisation sign ◌ઁ is encoded as U+0A81 ◌ઁ GUJARATI SIGN CANDRABINDU.

¶257 The consonant nasalization sign ◌ં is known as the *ansuvara* in Gujarati, and is--> encoded as U+0A82 ◌ં GUJARATI SIGN ANUSVARA.

¶258 The unvoiced aspiration sign ◌ઃ is encoded as U+0A83 ◌ઃ GUJARATI SIGN VISARGA.

¶259 ઍ ĕ is used todo and is encoded as U+0A8D ઍ GUJARATI VOWEL CANDRA E and U+0AC5 ◌ૅ GUJARATI VOWEL SIGN CANDRA E.

¶260 ઑ ŏ is used todo and is encoded as U+0A91 ઑ GUJARATI VOWEL CANDRA O and U+0AC9 ◌ૉ GUJARATI VOWEL SIGN CANDRA O.

¶261 ***Digits and Numbers.*** Gujarati has a full set of digits, used to write decimal numbers. Those are encoded at U+0AE6 ૦ GUJARATI DIGIT ZERO .. U+0AEF ૯ GUJARATI DIGIT NINE.

¶262 ***Punctuation and Signs.*** Gujarati uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶263 The *avagraha* is encoded as U+0ABD ઽ GUJARATI SIGN AVAGRAHA.

¶264 The *rupee sign* is encoded as U+0AF1 ૱ GUJARATI RUPEE SIGN.

¶265 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
આ	0A86	<0A85, 0ABE>
ઍ	0A8D	<0A85, 0AC5>
એ	0A8F	<0A85, 0AC7>
ઐ	0A90	<0A85, 0AC8>
ઑ	0A91	<0A85, 0AC9>
ઓ	0A93	<0A85, 0ACB>
ઔ	0A94	<0A85, 0ACC>

¶266 ***Typographic Cluster Cores.*** Here are some of the common typographic clusters made of two consonants which exhibit a fully-joined form:

ક ka • ક ka → ક્ક kka
ક ka • ષ ṣa → ક્ષ kṣa
ઙ ṅa • ક ka → ઙ્ક ṅka
ઙ ṅa • ગ ga → ઙ્ગ ṅga
ઙ ṅa • ઘ gha → ઙ્ઘ ṅgha
ઙ ṅa • મ ma → ઙ્મ ṅma
જ ja • ઞ ña → જ્ઞ jña
ઞ ña • ચ ca → ઞ્ચ ñca
ઞ ña • જ ja → ઞ્જ ñja
ટ ṭa • ટ ṭa → ટ્ટ ṭṭa
ટ ṭa • ઠ ṭha → ટ્ઠ ṭṭha
ડ ḍa • ડ ḍa → ડ્ડ ḍḍa
ડ ḍa • ઢ ḍha → ડ્ઢ ḍḍha
ઢ ḍha • ઢ ḍha → ઢ્ઢ ḍhḍha
ત ta • ત ta → ત્ત tta
દ da • ગ ga → દ્ગ dga
દ da • ઘ gha → દ્ઘ dgha
દ da • દ da → દ્દ dda
દ da • ધ dha → દ્ધ d'dha

ક્ષ da • ન na → ક્ષ્ન dna

ક્ષ da • બ ba → ક્ષ્બ dba

ક્ષ da • ભ bha → ક્ષ્ભ dbha

ક્ષ da • મ ma → ક્ષ્મ dma

ન na • ન na → ન્ન nna

¶267 In the absence of a fully-joined form, the core of a typographic cluster is displayed in a conjoined form, where all the but last consoant are in half form and the last consonant is in full form.

¶268 The half form of ક ka is ક્; the half form of જ ja is જ્; the half form of ફ pha is ફ્. The half form of the consonants with a vertical stem on the right is obtained by dropping this vertical stem.

ક ka • ઠ ṭha → ક્ઠ kṭha

જ ja • ક ka → જ્ક jka

ફ pha • ક ka → ફ્ક phka

ખ kha • ક ka → ખ્ક khka

ખ kha • ક્ષ da → ખ્ક્ષ khda

ટ ta • ક ka → ટ્ક tka

ટ ta • મ ma → ટ્મ tma

¶269 The remaining consonants do not have a half form. When they are present in non-final position in a cluster core, they non-joined form of the cluster is displayed, with a *halant* sign below all but the last consonant.

છ cha • ક ka → છ્ક chka

¶270 ***Satellite Consonants.*** ર ra in initial position is a satellite consonant. It is then displayed as a small hook, called a *repha*, on the top right side of the cluster, including any satellite vowels:

ર ra • ક ka → ર્ક rka

ર ra • ક ka • ખ kha → ર્ક્ખ rkkha

¶271 Where should the reph be placed if there is an a sign? clean up the examples above.

¶272 ર ra in final position is a satellite consonant. It is displayed as a small stroked attached to the vertical stem, or a circumfles-like sign, called a *vattu*, below the consonants which do not have a stem.

ખ kha • ર ra → ખ્ર khra

ખ kha • ર ra → ખ્ર khra

ઙ ṅa • ર ra → ઙ્ર ṅra

ઝ jha • ર ra → ઝ્ર jhra

ઠ ṭha • ર ra → ઠ્ર ṭhra

¶273 ***Satellite Vowels.*** The vowel signs attach around the cluster core.

on ક ka on ર્ક rka

આ ā	◌ા -ā	કા kā	કૃા rkā	
ઈ i	◌િ -i	કિ ki	કૃિ rki	
ઈ ī	◌ી -ī	કી kī	કૃી rkī	
ઉ u	◌ુ -u	કુ ku	કૃુ rku	
ઊ ū	◌ૂ -ū	કૂ kū	કૃૂ rkū	
એ ĕ	◌ૅ -ĕ	કૅ kĕ	કૃૅ rkĕ	
એ ē	◌े -ē	કે kē	કૃે rkē	
ઐ ai	◌ૈ -ai	કૈ kai	કૃૈ rkai	
ઓ ŏ	◌ૉ -ŏ	કૉ kŏ	કૃૉ rkŏ	
ઓ ō	◌ો -ō	કો kō	કૃો rkō	
ઔ au	◌ૌ -au	કૌ kau	કૃૌ rkau	
ઋ ṛ	◌ૃ -ṛ	કૃ kṛ	કૃૃ rkṛ	
ૠ ṝ	◌ૄ -ḷ	કૄ kḷ	કૃૄ rkḷ	
ઌ ḷ	◌ૢ ◌ૣ	કૢ kaḷ	કૃૢ rkaḷ	
ૡ ḹ	◌ૣ ◌ૣ	કૣ kaḷ	કૃૣ rkaḷ	

¶274 There is no vowel sign for અ a.

¶275 The vowel sign for ઈ i is generally displayed on the left side of its typographic cluster: ક ka • ઈ i → કિ ki; ન na • ગ ga • ઈ i → ન્ગિ n'gi.

¶276 todo where when there is hasant?.

¶277 ***Conjoined and Non-Joined Forms of Clusters.*** The display of a cluster can be encouraged to be in conjoined form (with fallback to a non-joined form if necessary) by inserting a U+200D ZERO WIDTH JOINER after the consonant linker coded character.

¶278 The display of a cluster can be encouraged to be in non-joined form by inserting a U+200C ZERO WIDTH NON-JOINER after the consonant linker coded character.

x • *y*	*x*, 0ACD, *y*	*x*, 0ACD, 200D, *y*	*x*, 0ACD, 200C, *y*
ઞ ña • ચ ca → ઞ્ચ ñca	ઞ્ચ ñca	ઞ‌્ચ ñca	
દ da • મ ma → દ્મ dma	દ્મ dma	દ‌્મ dma	

¶279 ***Typographic Fragments.*** The following typographic fragments can be represented as follows:

Full form	Half form	Represented by
ક	ક	<0A95, 0ACD, 200D>
ખ	ખ	<0A96, 0ACD, 200D>
ગ	ગ	<0A97, 0ACD, 200D>
ઘ	ઘ	<0A98, 0ACD, 200D>
ચ	ચ	<0A9A, 0ACD, 200D>
જ	જ	<0A9C, 0ACD, 200D>
ઞ	ઞ	<0A9E, 0ACD, 200D>
ઠા	ઠ	<0AA3, 0ACD, 200D>

ત	૮	<0AA4, 0ACD, 200D>
થ	૯	<0AA5, 0ACD, 200D>
દ	૯	<0AA7, 0ACD, 200D>
ન	૭	<0AA8, 0ACD, 200D>
પ	૫	<0AAA, 0ACD, 200D>
ફ	૬	<0AAB, 0ACD, 200D>
બ	૭	<0AAC, 0ACD, 200D>
ભ	૮	<0AAD, 0ACD, 200D>
મ	૯	<0AAE, 0ACD, 200D>
ય	૨	<0AAF, 0ACD, 200D>
ર	૯	<0AB2, 0ACD, 200D>
લ	૭	<0AB3, 0ACD, 200D>
વ	૮	<0AB5, 0ACD, 200D>
શ	૨	<0AB6, 0ACD, 200D>
ષ	૬	<0AB7, 0ACD, 200D>
સ	૨	<0AB8, 0ACD, 200D>
ક્ષ	૬	<0A95, 0ACD, 0AB7, 0ACD, 200D>
જ્ઞ	૬	<0A9C, 0ACD, 0A9e, 0ACD, 200D>
ત્ર	૨	<0AA4, 0ACD, 0AB0, 0ACD, 200D>
શ્ર	૨	<0AB6, 0ACD, 0AB0, 0ACD, 200D>

6.5 Oriya

¶280 The Oriya script is a North Indian script that is structurally similar to Devanagari, but with semicircular lines at the top of most letters instead of the straight horizontal bars of Devanagari. The actual shapes of the letters, particularly for vowel signs, show similarities to Tamil. The Oriya script is used to write the Oriya language of the Orissa state in India, as well as minority languages such as Khondi and Santali.

¶281 ***Standards.*** The Oriya block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard* , for more details on the relation to ISCII.

¶282 ***Encoding.*** The encoding of the Oriya script follows the consonant linking model. The consonant linker coded character of Oriya is U+0B4D ◌୍ ORIYA SIGN VIRAMA.

¶283 ***Letters.*** Oriya has the full complement of basic letters presented in section *2.1, Letters* .

¶284 The vowel nasalisation sign ◌ଁ is known as the *candrabindu* in Oriya, and is encoded as U+0B01 ◌ଁ ORIYA SIGN CANDRABINDU.

¶285 The consonant nasalization sign ◌ଂ is known as the *ansuvara* in Oriya, and is encoded as U+0B02 ◌ଂ ORIYA SIGN ANUSVARA.

¶286 The unvoiced aspiration sign ◌ଃ is known as the *visarga* in Oriya, and is encoded as U+0B03 ◌ଃ ORIYA SIGN VISARGA.

¶287 When Sanskrit words written in Devanagari using the letter व va are borrowed in the Oriya language, they are most often written with the Oriya letter ବ ba, to match the corresponding change in pronunciation: Sanskrit वन vana becomes Oriya ବନ bana. For the purpose of accurate transliteration, the letter ୱ va emerged, and is used to transliterate Devanagari व va, while the letter ବ ba is used only for the transliteration of Devanagari ब ba. This letter is not in common use today, but can be found in academic and technical texts. It is encoded as U+0B35 ୱ ORIYA LETTER VA. In older sources, this character is rendered by a wide range glyphs: ୱ, ୱ, ୱ, ୱ, ୱ (the latter being a modern rendering of the previous two).

¶288 the representative glyph uses a dot inside the circle and the text gives a variant with a ring above the circle. The font utkal.ttf, which is relatively recent uses a dot above, and does not even have a glyph matching the representative glyph. Is there is a disconnect somewhere?

¶289 To write Perso-Arabic or English loan words with a [w], the letter ୵ is sometimes used. It appears to have originally been devised as a ligature of ଓ ō and ବ ba, but because ligatures of independent vowels and consonants are not normally used in Oriya, this letter has been encoded as U+0B71 ୵ ORIYA LETTER WA, without a canonical decomposition. It is used initially in words or orthographic syllables to represent the foreign consonant; as a native semivowel, virama + ba is used because that is historically accurate. This character has a wide range of glyphs: ୵, ୵, ୵.

¶290 I can't make complete sense of the "as a native semivowel" part. Does it mean to that in native words, the *sound* [w] does not occur initially, and is written by BA? Also, this business of not doing vowel letter + virama + consonant is no longer accurate (bengla vowel + virama + ya), right? if so, how do we position U+B71?

¶291 ୟ ẏa is used for [...] and is encoded as U+0B5F ୟ ORIYA LETTER YYA.

¶292 ***Digits and Numbers.*** Oriya has a full set of decimal digits, encoded at U+0B66 ୦ ORIYA DIGIT ZERO .. U+0B6F ୯ ORIYA DIGIT NINE.

¶293 ***Punctuation and Signs.*** Oriya uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶294 The *avagraha* is encoded as U+0B3D ଽ ORIYA SIGN AVAGRAHA.

¶295 ୰ is placed before names of persons who are deceased and is encoded as U+0B70 ୰ ORIYA ISSHAR.

¶296 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ଆ	0B06	<0B05, 0B3E>
ଐ	0B10	<0B0F, 0B57>
ଔ	0B14	<0B13, 0B57>

¶297 ***Typographic Cluster Cores.*** Here are some of the common typographic clusters made of two consonants which exhibi a fully joined form:

କ ka • ଷ ṣa → କ୍ଷ kṣa

ଙ ṅa • କ ka → ଙ୍କ ṅka

ଙ ṅa • ଖ kha → ଙ୍ଖ ṅkha

ଙ ṅa • ଗ ga → ଙ୍ଗ ṅga

ଚ ca • ଚ ca → ଚ୍ଚ cca

ଚ ca • ଛ cha → ଚ୍ଛ ccha

ଜ ja • ଜ ja → ଜ୍ଜ jja

ଜ ja • ଞ ña → ଜ୍ଞ jña

ଞ ña • ଚ ca → ଞ୍ଚ ñca

ଞ ña • ଛ cha → ଞ୍ଛ ñcha

ଞ ña • ଜ ja → ଞ୍ଜ ñja

ଞ ña • ଝ jha → ଞ୍ଝ ñjha

ଟ ṭa • ଟ ṭa → ଟ୍ଟ ṭṭa

ଣ ṇa • ଡ ḍa → ଣ୍ଡ ṇḍa

ଣ ṇa • ଢ ḍha → ଣ୍ଢ ṇḍha

ତ ta • କ ka → ତ୍କ tka

ତ ta • ପ pa → ତ୍ପ tpa

ତ ta • ସ sa → ତ୍ସ tsa

ଦ da • ଦ da → ଦ୍ଦ dda

ଦ da • ଧ dha → ଦ୍ଧ d'dha

ଧ dha • ୟ ya → ଧ୍ୟ dhya

ନ na • ଥ tha → ନ୍ଥ ntha

ନ na • ଦ da → ନ୍ଦ nda

ନ na • ଧ dha → ନ୍ଧ ndha

ବ ba • ଦ da → ବ୍ଦ bda

ମ ma • ପ pa → ମ୍ପ mpa

ମ ma • ଫ pha → ମ୍ଫ mpha

ଶ śa • ଛ cha → ଶ୍ଛ ścha

ସ sa • ପ pa → ସ୍ପ spa

ସ sa • ଫ pha → ସ୍ଫ spha

¶298 In the absence of a fully-joined form, the core of a typographic cluster is displayed in conjoined form, where all but the first consonant are in half form and the last consonant is in its full form.

¶299 The reduced forms of ଢ ḍha, ତ ta, ଥ tha, ମ ma and ବ va are ୧ ḍha, ୬ta, ୰ tha, ୩ma and ୴ va. respectively.

¶300 The reduced forms of କ ka, ଜ ja, ନ na, ଲ la and ଳ ḷa is obtained by dropping their loop: ୩ ka, ୷ja, ୬ na, ୰ la and ୰ ḷa.

¶301 The reduced form of remaining consonants is a smaller version of them.

¶302 [Examples]

¶303 [visible halant] is displayed with a *halant* sign under it:

ଖ kha • କ ka → ଖ୍କ khka

¶304 ***Satellite Consonants.*** ର ra is a satellite consonant in initial position. It is then displayed as a small mark, called a *repha* on the top right side of the cluster core.

ର ra • କ ka → ର୍କ rka

¶305 ର ra is a satellite consonant in final position todo is this correct? examples..

¶306 Where should the reph be placed if there is an a sign? where should the reph be placed if there is a postjoined ya? clean up the examples above.

¶307 ତ ta is a satellite consonant in initial position.

¶308 ର ra is a satellite consonant in final position.

¶309 ଯ ya is a satellite consonant in final position.

¶310 In a cluster formed of ର ra and one of the consonants above, the initial ର ra adopts its satellite form and the second consonant adopts its full form. To encourage the initial ର ra to have its full form and the second consonant to have its satellite form, one inserts a U+200D ZERO WIDTH JOINER between the ର ra and the consonant linker.

ର ra, 0B4D, *y* ର ra, 200D, 0B4D, *y*

ର୍ଣ rṇa	ର୍‌ଣ raṇa
ର୍ତ rta	ର୍‌ତ rata
ର୍ନ rna	ର୍‌ନ rana
ର୍ବ rba	ର୍‌ବ raba
ର୍ଭ rbha	ର୍‌ଭ rabha
ର୍ମ rma	ର୍‌ମ rama
ର୍ଯ rya	ର୍‌ଯ raya
ର୍ର rra	ର୍‌ର rara
ର୍ଲ rla	ର୍‌ଲ rala

ର୍ଳ rḷa ଳ୍ର raḷa

¶311 ***Satellite Vowels.*** The vowel sign attaches arond the cluster core:

		on କ ka	on ର୍କ rka
ଆ ā	◌ା -ā	କା kā	ର୍କା rkā
ଇ i	◌ି -i	କି ki	ର୍କି rki
ଈ ī	◌ୀ -ī	କୀ kī	ର୍କୀ rkī
ଉ u	◌ୁ -u	କୁ ku	ର୍କୁ rku
ଊ ū	◌ୂ -ū	କୂ kū	ର୍କୂ rkū
ଏ ē	◌େ -ē	କେ kē	ର୍କେ rkē
ଐ ai	◌ୈ -ai	କୈ kai	ର୍କୈ rkai
ଓ ō	◌ୋ -ō	କୋ kō	ର୍କୋ rkō
ଔ au	◌ୌ -au	କୌ kau	ର୍କୌ rkau
ଋ r̥	◌ୃ -r̥	କୃ kr̥	ର୍କୃ rkr̥

¶312 There is no vowel sign for ଅ a, ୠ r̥̄, ଌ l̥, ୡ l̥̄.

¶313 ***Typographic Fragments.*** The following half forms can be represented as follows:

Full form	Half form	Represented by
କ	୍କ	<0020, 200D, 0B4D, 0B15>
ଜ	୍ଜ	<0020, 200D, 0B4D, 0B1C>
ନ	୍ନ	<0020, 200D, 0B4D, 0B28>
ବ	୍ବ	<0020, 200D, 0B4D, 0B2C>
ଲ	୍ଲ	<0020, 200D, 0B4D, 0B32>
ଳ	୍ଳ	<0020, 200D, 0B4D, 0B33>
ଵ	୍ଵ	<0020, 200D, 0B4D, 0B35>

¶314 The list above comes from the TDIL document, p30, bottom of left column "KA, .. LLA are presented in their half form". Peter's doc, L2/04-279 says "most consonants" have a (subjoined) half form (table 6 on page 5.

¶315 However, look at JA. In the TDIL doc, Table 1 on page 54 gives only three oher consonants after which JA subjoins. If that is the case, it is really worth saying that JA has a half form, or is more appropriate to say that there are three conjuncts with JA in second position for which one can identify the fragment that corresponds to JA? Personally, I think the later.

¶316 Furthermore, the three consonants listed by TDIL are KA, LA and LLA. First, we can't tell what they possibly have in mind, because they have KA instead of JA in their examples. Second, I checked utkal.ttf (which is apparently well accepted in the Linux community) and I don't see any subjoined JA.

¶317 Starting from utkal, I see that NNA, TA, NA, BHA, MA, RA, LA, LLA and VA have subjoined half forms. RA and TA are not in TDIL's list only because they are treated separated. That leaves KA and JA in TDIL's list but not in utkal; and NNA, BHA, MA in utkal but not in TDIL. And all of that is short of Peter's "most". Given the disparity of sources, the likely that we can reliably render any set of half forms seems pretty small, unless Unicode set a list.

¶318 The left side of the vowel signs 6◌ -ai, 6◌ -ō and 6◌ -au can be represented in isolation by <0020, 0B47> → 6.

¶319 The top side of the vowel sign 6◌ -ai can be represented in isolation by <0020, 0B56> → ◌. The coded character U+0B56 ◌ ORIYA AI LENGTH MARK does not correspond to an abstract character and is encoded solely to provide a canonical decompositon for U+0B48 6◌ ORIYA VOWEL SIGN AI, and to represent its top side in isolation.

¶320 The right side of the vowel sign 6◌ -ō can be represented in isolation by <0020, 0B3E> → ।

¶321 The right side of the vowel sign 6◌ -au can be represented in isolation by <0020, 0B57> → ◌. The coded character U+0B57 ◌ ORIYA AU LENGTH MARK does not correspond to an abstract character and is encoded solely to provide a canonical decomposition for U+0B4C 6◌ ORIYA VOWEL SIGN AU, and to represent its right side in isolation.

6.6 Tamil

¶322 The Tamil script is a South Indian script. South Indian scripts are structurally related to the North Indian scripts, but they are used to write the Dravidian languages of southern India and of Sri Lanka, which are genetically unrelated to the North Indian languages such as Hindi, Bengla, and Gujarati. The shapes of letters in the South Indian scripts are generally quite distinct from the shapes of letters in Devanagari and its related scripts.

¶323 The Tamil script is used to write the Tamil language of the Tamil Nadu state in India, as well as minority languages such as the Dravidian language Badaga and the Indo-European language Saurashtra. Tamil is also used in Sri Lanka, Singapore, and parts of Malaysia. The Tamil script also lacks conjunct consonant forms.

¶324 Tamil is written horizontally from left to right, and words are separated by space.

¶325 ***Standards.*** The Tamil block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶326 ***Encoding.*** The encoding of the Tamil script in Unicode follows the consonant linking model. The consonant linker coded character of Tamil is U+0BCD ◌ TAMIL SIGN VIRAMA.

¶327 ***Letters.*** The Tamil script has fewer consonants than the other Indic scripts, no vocalic vowels, but has all the other vowels presented in section *2.1, Letters*. In addition, Tamil as the full complement of additional letters presented in section .

¶328 The consonant nasalization sign ◌ is not used in Tamil, but it encoded as U+0B82 ◌ TAMIL SIGN ANUSVARA.

¶329 The sign ◌ is known as the *aytham* in Tamil. It is historically related to the visarga in other Indic scripts, but has become an ordinary spacing letter in Tamil. It is used to modify the sound of other consonants and, in particular, to represent the spelling of words borrowed into Tamil from English or other languages. It is encoded as U+0B83 ◌ TAMIL SIGN VISARGA; despite its position in the code charts and its name, this sign is not similar to the unvoiced aspiration sign of the other scripts.

¶330 When representing the "missing" consonants in transcriptions of languages such as Sanskrit or Saurashtra, superscript European digits are often used, so ⊔² = pha, ⊔³ = ba, and ⊔▌ = bha. The characters U+00B2 ² SUPERSCRIPT TWO, U+00B3 ³ SUPERSCRIPT THREE, and U+2074 ▌ SUPERSCRIPT FOUR can be used to preserve this distinction in plain text.

¶331 ***Digits and Numbers.*** Tamil has a full set of digits, encoded at U+0BE6 ൦ TAMIL DIGIT ZERO .. U+0BEF ൯ TAMIL DIGIT NINE.

¶332 Traditionally, the digits ௧ 1 through ௯ 9 are used in a positional system, along with ௰ 10, ௱ 100 and ௲ 1000, which are encoded as U+0BF0 ௰ TAMIL NUMBER TEN, U+0BF1 ௱ TAMIL NUMBER ONE HUNDRED and U+0BF2 ௲ TAMIL NUMBER ONE THOUSAND.

¶333 ***Punctuation and Signs.*** Tamil uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶334 [Amd3] The *om sign* is encoded as U+0BD0 ௐ TAMIL OM.

¶335 ௳ ... ௸

¶336 ௺ is used to [...] and is encoded as U+0BFA ௺ TAMIL NUMBER SIGN.

¶337 ***Typographic Clusters.*** The Tamil script is notable for the quasi absence of typographic interactions between the consonant letters which form a typographic cluster. About the only case is that of க ka + ஷ ṣa which results in க்ஷ kṣa but even that form is disappearing in modern usage. As a result, clusters are essentially always rendered by strategy 3.

¶338 On the other hand, there are rich interactions between the consonants of a cluster and the diacritic marks for the vowels:

ட ṭa • ◌ி -i → டி ṭi
ட ṭa • ◌ீ -ī → டீ ṭī
ல la • ◌ி -i → லி li
ல la • ◌ீ -ī → லீ lī
க ka • ◌ு -u → கு ku
க ka • ◌ூ -ū → கூ kū
ங ṅa • ◌ு -u → ஙு ṅu
ங ṅa • ◌ூ -ū → ஙூ ṅū
ச ca • ◌ு -u → சு cu
ச ca • ◌ூ -ū → சூ cū
ஜ ja • ◌ு -u → ஜு ju
ஜ ja • ◌ூ -ū → ஜூ jū
ஞ ña • ◌ு -u → ஞு ñu
ஞ ña • ◌ூ -ū → ஞூ ñū
ட ṭa • ◌ு -u → டு ṭu
ட ṭa • ◌ூ -ū → டூ ṭū
ண ṇa • ◌ு -u → ணு ṇu
ண ṇa • ◌ூ -ū → ணூ ṇū
த ta • ◌ு -u → து tu
த ta • ◌ூ -ū → தூ tū
ந na • ◌ு -u → நு nu
ந na • ◌ூ -ū → நூ nū
ன ṉa • ◌ு -u → னு ṉu
ன ṉa • ◌ூ -ū → னூ ṉū
ப pa • ◌ு -u → பு pu
ப pa • ◌ூ -ū → பூ pū
ம ma • ◌ு -u → மு mu
ம ma • ◌ூ -ū → மூ mū

யு ya • ு -u → யு yu
யூ ya • ூ -ū → யூ yū
ரு ra • ு -u → ரு ru
ரூ ra • ூ -ū → ரூ rū
று r̲a • ு -u → று r̲u
றூ r̲a • ூ -ū → றூ r̲ū
லு la • ு -u → லு lu
லூ la • ூ -ū → லூ lū
ளு ḷa • ு -u → ளு ḷu
ளூ ḷa • ூ -ū → ளூ ḷū
ழு l̲a • ு -u → ழு l̲u
ழூ l̲a • ூ -ū → ழூ l̲ū
வு va • ு -u → வு vu
வூ va • ூ -ū → வூ vū
ரு ra • ு -u → ரு ru
ரூ ra • ூ -ū → ரூ rū
ஷு ṣa • ு -u → ஷு ṣu
ஷூ ṣa • ூ -ū → ஷூ ṣū
ஸு sa • ு -u → ஸு su
ஸூ sa • ூ -ū → ஸூ sū
ஹு ha • ு -u → ஹு hu
ஹூ ha • ூ -ū → ஹூ hū
க்ஷு kṣa • ு -u → க்ஷு kṣu
க்ஷூ kṣa • ூ -ū → க்ஷூ kṣū

¶339 The consonant ர ra is simplified when this does not lead to confusion; for example:

ர ra • ் → ர் r
ர ra • ி -i → ரி ri
ர ra • ீ -ī → ரீ rī

¶340 However, various governmental bodies mandate that the basic shape of ர should be used for these ligatures as well, especially in school textbooks. Media and literary publications in Malaysia and Singapore mostly use the unchanged form of ர.

¶341 Some changes are no longer in common use:

ணா ṇa • ா -ā → ணா ṇā
னா n̲a • ா -ā → னா n̲ā
றா r̲a • ா -ā → றா r̲ā
ணொ ṇa • ொ -o → ணொ ṇo
னொ n̲a • ொ -o → னொ n̲o
றொ r̲a • ொ -o → றொ r̲o
ணோ ṇa • ோ -ō → ணோ ṇō
னோ n̲a • ோ -ō → னோ n̲ō
றோ r̲a • ோ -ō → றோ r̲ō
ணை ṇa • ை -ai → ணை ṇai

ன n̲a • ை -ai → னை n̲ai

ல la • ை -ai → லை lai

ள l̤a • ை -ai → ளை l̤ai

¶342 The coded character U+0BD7 ◌ௗ TAMIL AU LENGTH MARK does not correspond to an abstract character, i.e., it is not used for the representation of text. It is encoded to represent the right side of the vowel sign ◌ெள -au in isolation, and to provide a canonical decomposition for U+0BCC ◌ொள TAMIL VOWEL SIGN AU.

¶343 ***Named sequences.*** Tamil is less complex than some of the other Indic scripts, and both conceptually and in processing can be treated as an atomic set of elements: consonants, stand-alone vowels, and syllables. The table below shows these atomic elements, with the corresponding Unicode characters or sequences. These elements have been approved as Tamil named character sequences: see NamedSequences.txt in the Unicode Character Database.

¶344 In implementations such as natural language processing, where it may be useful to treat such Tamil text elements as single code points for ease of processing, they can be mapped to a segment of the Private Use Area.

¶345 In the table below, the first row shows the transliterated representation of the Tamil vowels in abbreviated form, while the first column shows the transliterated representation of the Tamil consonants. These short strings can then be concatenated to form unique names for Tamil pure consonants and syllables: K, KA, KAA, KI, KE, KU, and so on. Details on the complete names for each element can be found in NamedSequences.txt.

		A	AA	I	II	U	UU	E	EE	AI	O	OO	AU
	◌ஂ 0B83	அ 0B85	ஆ 0B86	இ 0B87	ஈ 0B88	உ 0B89	ஊ 0B8A	எ 0B8E	ஏ 0B8F	ஐ 0B90	ஒ 0B92	ஓ 0B93	ஔ 0B94
K	க் 0B95 0BCD	க 0B95	கா 0B95 0BBE	கி 0B95 0BBF	கீ 0B95 0BC0	கு 0B95 0BC1	கூ 0B95 0BC2	கெ 0B95 0BC6	கே 0B95 0BC7	கை 0B95 0BC8	கொ 0B95 0BCA	கோ 0B95 0BCB	கௌ 0B95 0BCC
NG	ங் 0B99 0BCD	ங 0B99	ஙா 0B99 0BBE	ஙி 0B99 0BBF	ஙீ 0B99 0BC0	ஙு 0B99 0BC1	ஙூ 0B99 0BC2	ஙெ 0B99 0BC6	ஙே 0B99 0BC7	ஙை 0B99 0BC8	ஙொ 0B99 0BCA	ஙோ 0B99 0BCB	ஙௌ 0B99 0BCC
C	ச் 0B9A 0BCD	ச 0B9A	சா 0B9A 0BBE	சி 0B9A 0BBF	சீ 0B9A 0BC0	சு 0B9A 0BC1	சூ 0B9A 0BC2	செ 0B9A 0BC6	சே 0B9A 0BC7	சை 0B9A 0BC8	சொ 0B9A 0BCA	சோ 0B9A 0BCB	சௌ 0B9A 0BCC
NY	ஞ் 0B9E 0BCD	ஞ 0B9E	ஞா 0B9E 0BBE	ஞி 0B9E 0BBF	ஞீ 0B9E 0BC0	ஞு 0B9E 0BC1	ஞூ 0B9E 0BC2	ஞெ 0B9E 0BC6	ஞே 0B9E 0BC7	ஞை 0B9E 0BC8	ஞொ 0B9E 0BCA	ஞோ 0B9E 0BCB	ஞௌ 0B9E 0BCC
TT	ட் 0B9F 0BCD	ட 0B9F	டா 0B9F 0BBE	டி 0B9F 0BBF	டீ 0B9F 0BC0	டு 0B9F 0BC1	டூ 0B9F 0BC2	டெ 0B9F 0BC6	டே 0B9F 0BC7	டை 0B9F 0BC8	டொ 0B9F 0BCA	டோ 0B9F 0BCB	டௌ 0B9F 0BCC
NN	ண் 0BA3 0BCD	ண 0BA3	ணா 0BA3 0BBE	ணி 0BA3 0BBF	ணீ 0BA3 0BC0	ணு 0BA3 0BC1	ணூ 0BA3 0BC2	ணெ 0BA3 0BC6	ணே 0BA3 0BC7	ணை 0BA3 0BC8	ணொ 0BA3 0BCA	ணோ 0BA3 0BCB	ணௌ 0BA3 0BCC
T	த் 0BA4 0BCD	த 0BA4	தா 0BA4 0BBE	தி 0BA4 0BBF	தீ 0BA4 0BC0	து 0BA4 0BC1	தூ 0BA4 0BC2	தெ 0BA4 0BC6	தே 0BA4 0BC7	தை 0BA4 0BC8	தொ 0BA4 0BCA	தோ 0BA4 0BCB	தௌ 0BA4 0BCC
N	ந் 0BA8 0BCD	ந 0BA8	நா 0BA8 0BBE	நி 0BA8 0BBF	நீ 0BA8 0BC0	நு 0BA8 0BC1	நூ 0BA8 0BC2	நெ 0BA8 0BC6	நே 0BA8 0BC7	நை 0BA8 0BC8	நொ 0BA8 0BCA	நோ 0BA8 0BCB	நௌ 0BA8 0BCC
P	ப் 0BAA 0BCD	ப 0BAA	பா 0BAA 0BBE	பி 0BAA 0BBF	பீ 0BAA 0BC0	பு 0BAA 0BC1	பூ 0BAA 0BC2	பெ 0BAA 0BC6	பே 0BAA 0BC7	பை 0BAA 0BC8	பொ 0BAA 0BCA	போ 0BAA 0BCB	பௌ 0BAA 0BCC
M	ம் 0BAE 0BCD	ம 0BAE	மா 0BAE 0BBE	மி 0BAE 0BBF	மீ 0BAE 0BC0	மு 0BAE 0BC1	மூ 0BAE 0BC2	மெ 0BAE 0BC6	மே 0BAE 0BC7	மை 0BAE 0BC8	மொ 0BAE 0BCA	மோ 0BAE 0BCB	மௌ 0BAE 0BCC
Y	ய் 0BAF 0BCD	ய 0BAF	யா 0BAF 0BBE	யி 0BAF 0BBF	யீ 0BAF 0BC0	யு 0BAF 0BC1	யூ 0BAF 0BC2	யெ 0BAF 0BC6	யே 0BAF 0BC7	யை 0BAF 0BC8	யொ 0BAF 0BCA	யோ 0BAF 0BCB	யௌ 0BAF 0BCC
R	ர் 0BB0 0BCD	ர 0BB0	ரா 0BB0 0BBE	ரி 0BB0 0BBF	ரீ 0BB0 0BC0	ரு 0BB0 0BC1	ரூ 0BB0 0BC2	ரெ 0BB0 0BC6	ரே 0BB0 0BC7	ரை 0BB0 0BC8	ரொ 0BB0 0BCA	ரோ 0BB0 0BCB	ரௌ 0BB0 0BCC
L	ல் 0BB2 0BCD	ல 0BB2	லா 0BB2 0BBE	லி 0BB2 0BBF	லீ 0BB2 0BC0	லு 0BB2 0BC1	லூ 0BB2 0BC2	லெ 0BB2 0BC6	லே 0BB2 0BC7	லை 0BB2 0BC8	லொ 0BB2 0BCA	லோ 0BB2 0BCB	லௌ 0BB2 0BCC
V	வ் 0BB5 0BCD	வ 0BB5	வா 0BB5 0BBE	வி 0BB5 0BBF	வீ 0BB5 0BC0	வு 0BB5 0BC1	வூ 0BB5 0BC2	வெ 0BB5 0BC6	வே 0BB5 0BC7	வை 0BB5 0BC8	வொ 0BB5 0BCA	வோ 0BB5 0BCB	வௌ 0BB5 0BCC
LLL	ழ் 0BB4 0BCD	ழ 0BB4	ழா 0BB4 0BBE	ழி 0BB4 0BBF	ழீ 0BB4 0BC0	ழு 0BB4 0BC1	ழூ 0BB4 0BC2	ழெ 0BB4 0BC6	ழே 0BB4 0BC7	ழை 0BB4 0BC8	ழொ 0BB4 0BCA	ழோ 0BB4 0BCB	ழௌ 0BB4 0BCC
LL	ள் 0BB3 0BCD	ள 0BB3	ளா 0BB3 0BBE	ளி 0BB3 0BBF	ளீ 0BB3 0BC0	ளு 0BB3 0BC1	ளூ 0BB3 0BC2	ளெ 0BB3 0BC6	ளே 0BB3 0BC7	ளை 0BB3 0BC8	ளொ 0BB3 0BCA	ளோ 0BB3 0BCB	ளௌ 0BB3 0BCC

RR	ற் 0BB1 0BCD	ற 0BB1	றா 0BB1 0BBE	றி 0BB1 0BBF	றீ 0BB1 0BC0	று 0BB1 0BC1	றூ 0BB1 0BC2	றெ 0BB1 0BC6	றே 0BB1 0BC7	றை 0BB1 0BC8	றொ 0BB1 0BCA	றோ 0BB1 0BCB	றௌ 0BB1 0BCC
NNN	ன் 0BA9 0BCD	ன 0BA9	னா 0BA9 0BBE	னி 0BA9 0BBF	னீ 0BA9 0BC0	னு 0BA9 0BC1	னூ 0BA9 0BC2	னெ 0BA9 0BC6	னே 0BA9 0BC7	னை 0BA9 0BC8	னொ 0BA9 0BCA	னோ 0BA9 0BCB	னௌ 0BA9 0BCC
J	ஜ் 0B9C 0BCD	ஜ 0B9C	ஜா 0B9C 0BBE	ஜி 0B9C 0BBF	ஜீ 0B9C 0BC0	ஜு 0B9C 0BC1	ஜூ 0B9C 0BC2	ஜெ 0B9C 0BC6	ஜே 0B9C 0BC7	ஜை 0B9C 0BC8	ஜொ 0B9C 0BCA	ஜோ 0B9C 0BCB	ஜௌ 0B9C 0BCC
SH	ஶ் 0BB6 0BCD	ஶ 0BB6	ஶா 0BB6 0BBE	ஶி 0BB6 0BBF	ஶீ 0BB6 0BC0	ஶு 0BB6 0BC1	ஶூ 0BB6 0BC2	ஶெ 0BB6 0BC6	ஶே 0BB6 0BC7	ஶை 0BB6 0BC8	ஶொ 0BB6 0BCA	ஶோ 0BB6 0BCB	ஶௌ 0BB6 0BCC
SS	ஷ் 0BB7 0BCD	ஷ 0BB7	ஷா 0BB7 0BBE	ஷி 0BB7 0BBF	ஷீ 0BB7 0BC0	ஷு 0BB7 0BC1	ஷூ 0BB7 0BC2	ஷெ 0BB7 0BC6	ஷே 0BB7 0BC7	ஷை 0BB7 0BC8	ஷொ 0BB7 0BCA	ஷோ 0BB7 0BCB	ஷௌ 0BB7 0BCC
S	ஸ் 0BB8 0BCD	ஸ 0BB8	ஸா 0BB8 0BBE	ஸி 0BB8 0BBF	ஸீ 0BB8 0BC0	ஸு 0BB8 0BC1	ஸூ 0BB8 0BC2	ஸெ 0BB8 0BC6	ஸே 0BB8 0BC7	ஸை 0BB8 0BC8	ஸொ 0BB8 0BCA	ஸோ 0BB8 0BCB	ஸௌ 0BB8 0BCC
H	ஹ் 0BB9 0BCD	ஹ 0BB9	ஹா 0BB9 0BBE	ஹி 0BB9 0BBF	ஹீ 0BB9 0BC0	ஹு 0BB9 0BC1	ஹூ 0BB9 0BC2	ஹெ 0BB9 0BC6	ஹே 0BB9 0BC7	ஹை 0BB9 0BC8	ஹொ 0BB9 0BCA	ஹோ 0BB9 0BCB	ஹௌ 0BB9 0BCC
KSS	க்ஷ் 0B95 0BCD 0BB7 0BCD	க்ஷ 0B95 0BCD 0BB7	க்ஷா 0B95 0BCD 0BB7 0BBE	க்ஷி 0B95 0BCD 0BB7 0BBF	க்ஷீ 0B95 0BCD 0BB7 0BC0	க்ஷு 0B95 0BCD 0BB7 0BC1	க்ஷூ 0B95 0BCD 0BB7 0BC2	க்ஷெ 0B95 0BCD 0BB7 0BC6	க்ஷே 0B95 0BCD 0BB7 0BC7	க்ஷை 0B95 0BCD 0BB7 0BC8	க்ஷொ 0B95 0BCD 0BB7 0BCA	க்ஷோ 0B95 0BCD 0BB7 0BCB	க்ஷௌ 0B95 0BCD 0BB7 0BCC

6.7 Telugu

¶346 The Telugu script is a South Indian script used to write the Telugu language of the Andhra Pradesh state in India, as well as minority languages such as Gondi (Adilabad and Koi dialects) and Lambadi. The script is also used in Maharashtra, Orissa, Madhya Pradesh, and West Bengal. The Telugu script became distinct by the thirteenth century CE and shares ancestors with the Kannada script.

¶347 Many Telugu letters have a v-shaped headstroke, which is a structural mark corresponding to the horizontal bar in Devanagari and the arch in Oriya script.

¶348 ***Standards.*** The Telugu block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶349 ***Encoding.*** The encoding of the Telugu script in Unicode follows the consonant linking model. The consonant linker coded character of Telugu is U+0C4D ◌్ TELUGU SIGN VIRAMA.

¶350 ***Letters.*** Telugu has the full complement of basic letters presented in section *2.1, Letters*. In addition, Telugu as the full complement of additional letters presented in section , except for న n̲a and ళ l̲a.

¶351 The vowel nasalisation sign ◌ఁ is known as అరసున్న arasunna and is encoded as U+0C01 ◌ఁ TELUGU SIGN CANDRABINDU.

¶352 The consonant nasalization sign ◌ం is known as సున్న sunna and is encoded as U+0C02 ◌ం TELUGU SIGN ANUSVARA.

¶353 The unvoiced aspiration sign ◌ః is known as విసర్గ visarga and is encoded as U+0C03 ◌ః TELUGU SIGN VISARGA.

¶354 Telugu does not have a nukta.

¶355 The letters చే and జే have been used historically for the voiceless alveolar affricate [ts] and the voiced alveolar affricate [dz] respectively. They are encoded as U+0C58 చే TELUGU LETTER TSA and U+0C59 జే TELUGU LETTER DZA. These characters are commonly found in old grammar books and dictionaries. Historically, various glyphs were used to render these two characters but the notation used currently was proposed by Charles Philip Brown in the 19th century. These letters are not in current use in contemporary Telugu.

¶356 ***Digits and Numbers.*** Telugu has a full set of decimal digits, encoded at U+0C66 ౦ TELUGU DIGIT ZERO .. U+0C6F ౯ TELUGU DIGIT NINE.

¶357 A set of digits for fractions is encoded at U+0C78 ౸ TELUGU FRACTION DIGIT ZERO FOR ODD POWERS OF FOUR .. U+0C7E ౾ TELUGU FRACTION DIGIT THREE FOR EVEN POWERS OF FOUR.

¶358 ***Punctuation and Signs.*** Telugu uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶359 The *avagraha* is encoded as U+0C3D ఽ TELUGU SIGN AVAGRAHA.

¶360 The sign ౏ is used to denote a unit of volume or weight and is encoded as U+0C7F ౏ TELUGU SIGN TUUMU.

¶361 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ఓ	0C13	<0C12, 0C55>
ఔ	0C14	<0C12, 0C4C>

¶362 Similarly, some vowel signs are atomically encoded:

To represent	Use	Do not use
ొ	0C42	<0C41, 0C3E>
ౌ	0C44	<0C43, 0C3E>
ో	0C47	<0C46, 0C55>
ో	0C4B	<0C4A, 0C55>

¶363 ***Typographic Clusters.*** Telugu consonant clusters are most commonly represented by a subscripted, and often transformed, consonant glyph for the second element of the cluster:

గ ga • గ ga	→ గ్గ gga
క ka • య ya	→ క్య kya
క ka • ష ṣa	→ క్ష kṣa
క ka • క ka	→ క్క
మ ma • మ ma	→ మ్మ
ల la • ల la	→ ల్ల
త ta • త ta	→ త్త
య ya • య ya	→ య్య
వ va • వ va	→ వ్వ
ర ra • ర ra	→ ర్ర
న na • న na	→ న్న
స sa • త ta • ర ra	→ స్త్ర stra

¶364 ***Satellite Consonants.*** Historically, a ర ra in initial position could be rendered is a satellite consonant. It was then displayed on the right side of cluster core by the sign ర, known as వలపలగిలక valapalagilaka.

¶365 ***Satellite vowels.*** Vowel signs attach either above, above and below or the right; when they attach above, and the consonant has a v-shaped headstroke, that headstroke disappears. Here are example on క ka:

		క ka
ఆ ā	◌ా -ā	కా kā
ఇ i	◌ి -i	కి ki
ఈ ī	◌ీ -ī	కీ kī
ఉ u	◌ు -u	కు ku
ఊ ū	◌ూ -ū	కూ kū
ఋ ṛ	◌ృ -ṛ	కృ kṛ
ౠ ṝ	◌ౄ -ḷ	కౄ kḷ
ఎ e	◌ె -e	కె ke
ఏ ē	◌ే -ē	కే kē
ఐ ai	◌ై -ai	కై kai
ఒ o	◌ొ -o	కొ ko
ఓ ō	◌ో -ō	కో kō
ఔ au	◌ౌ -au	కౌ kau

¶366 A few combinations show a more complex interaction of the vowel signs:

	◌ా -ā	◌ి -i	◌ీ -ī	◌ు -u	◌ూ -ū	◌ొ -o	◌ో -ō	◌ౌ -au
ఙ ṅa				ఙు	ఙూ			
ప pa				పు	పూ			
ఫ pha				ఫు	ఫూ			
మ ma	మి	మీ				మొ	మో	
య ya	యి	యీ				యొ	యో	
వ va				వు	వూ			
హ ha	హా			హు	హూ	హొ	హో	హౌ

¶367 ***Clusters without vowels.*** For Sanskrit or foreign words ending in a consonant sound, the absence of a final vowel sound is indicated by the presence of the sign ◌్, called పొల్లు pollu, above the consonant.

¶368 ***Non-joined Forms of Clusters.***

¶369 ***Typographic fragments.*** The coded character U+0C55 ◌ౕ TELUGU LENGTH MARK does not correspond to an abstract character and is encoded solely to represent the second element of U+0C47 ◌ే TELUGU VOWEL SIGN EE and of U+0C4B ◌ో

TELUGU VOWEL SIGN OO in isolation.

¶370 The coded character U+0C56 ◌ౖ TELUGU AI LENGTH MARK does not correspond to an abstract character is encoded solely to provide a canonical decomposition for U+0C48 ◌ై TELUGU VOWEL SIGN AI, and to represent its bottom side in isolation.

6.8 Kannada

¶371 The Kannada script is a South Indian script. It is used to write the Kannada (or Kanarese) language of the Karnataka state in India and to write minority languages such as Tulu. The Kannada language is also used in many parts of Tamil Nadu, Kerala, Andhra Pradesh, and Maharashtra. This script is very closely related to the Telugu script both in the shapes of the letters and in the behavior of conjunct consonants.

¶372 Many Kannada letters have a headstroke formed of a horizontal line and a hook corresponding to the horizontal bar in Devanagari and the arch in the Oriya script.

¶373 ***Standards.*** The Kannada block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard*, for more details on the relation to ISCII.

¶374 ***Encoding.*** The encoding of the Kannada script in Unicode follows the consonant linking model. The consonant linker coded character of Kannada is U+0CCD ◌್ KANNADA SIGN VIRAMA.

¶375 ***Letters.*** Kannada has the full complement of basic letters presented in section *2.1, Letters*. In addition, Kannada as the full complement of additional letters presented in section except for ಱ ṉa and ಠ ṭa.

¶376 The consonant nasalization sign ◌ಂ and is encoded as U+0C82 ◌ಂ KANNADA SIGN ANUSVARA.

¶377 The unvoiced aspiration sign ◌ಃ is encoded as U+0C83 ◌ಃ KANNADA SIGN VISARGA.

¶378 The tongue-root sibilant □ is encoded as U+0CF1 □ KANNADA SIGN JIHVAMULIYA.

¶379 The labial sibilant □ is encoded as U+0CF2 □ KANNADA SIGN UPADHMANIYA

¶380 ೞ ḻa is actually an obsolete Kannada letter that is transliterated in Dravidian scholarship as ẕ, ḻ or ṛ. It is encoded as U+0CDE ೞ KANNADA LETTER FA. This coded character should have been named "LLLA" rather than "FA", so the name in this standard is simply a mistake. This letter has not been actively used in Kannada since the end of the tenth century. Collations should treat U+0CDE ೞ KANNADA LETTER FA as following U+0CB3 ಳ KANNADA LETTER LLA.

¶381 U+0CD5 ◌ೕ KANNADA LENGTH MARK and U+0CD6 ◌ೖ KANNADA AI LENGTH MARK are provided for the encoding of fragments of vowel signs. However, those two length marks have no independent existence in the Kannada writing system and do not play any part as independent codes in the traditional collation order.

¶382 ***Digits and Numbers.*** Kannada has a full set of decimal digits, encoded at U+0CE6 ೦ KANNADA DIGIT ZERO .. U+0CEF ೯ KANNADA DIGIT NINE.

¶383 ***Punctuation and Signs.*** Kannada uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶384 The *avagraha* is encoded as U+0CBD ಽ KANNADA SIGN AVAGRAHA.

¶385 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of

multiple parts. The table below 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.

To represent	Use	Do not use
ಔ	0C94	<0C92, 0CCC>

¶386 Similarly, some vowel signs are atomically encoded:

To represent	Use	Do not use
◌ಾ	0C42	<0C41, 0C3E>
◌ೈ	0C44	<0C43, 0C3E>
◌ೋ	0C47	<0C46, 0C55>

¶387 ***Typographic Clusters.*** Kannada consonant clusters are most commonly represented by a subscripted, and often transformed, consonant glyph for the second element of the cluster:

ಗ ga • ಗ ga → ಗ್ಗ gga

ಕ ka • ಕ ka → ಕ್ಕ kka

ಕ ka • ಯ ya → ಕ್ಯ kya

ಕ ka • ಷ ṣa → ಕ್ಷ kṣa

¶388 ***Satellite Consonants.*** In Kannada, ರ ra is a satellite consonant in initial position. It is then displayed on the right side of the cluster core: ರ್ಕ rkna, ರ್ಕಾ rknā, ರ್ಕೀ rknī. Alternatively, an initial ರ ra can keep it full form, and this rendering can be imposed by the use of U+200D ZERO WIDTH JOINER: ರ್‌ಕ rka

¶389 ***Satellite vowels.*** Vowel signs attach either above, above and below or the right; when they attach above, and the consonant has a v-shaped headstroke, that headstroke disappears. Here are example on ಕ ka:

			ಕ ka
ಆ ā	◌ಾ	-ā	ಕಾ kā
ಇ i	◌ಿ	-i	ಕಿ ki
ಈ ī	◌ೀ	-ī	ಕೀ kī
ಉ u	◌ು	-u	ಕು ku
ಊ ū	◌ೂ	-ū	ಕೂ kū
ಋ ṛ	◌ೃ	-ṛ	ಕೃ kṛ
ೠ ṝ	◌ೄ	-ḷ	ಕೄ kḷ
ಌ ḷ	◌ೢ	◌ೢ	ಕ್ಲ kḷa
ೡ ḹ	◌ೣ	◌ೣ	ಕ್ಲ kḷa
ಎ e	◌ೆ	-e	ಕೆ ke
ಏ ē	◌ೇ	-ē	ಕೇ kē

ಐ ai ◌ೈ -ai ಕೈ kai

ಓ ō ◌ೋ -ō ಕೋ kō

ಔ au ◌ೌ -au ಕೌ kau

¶390 A few combinations show a more complex interaction of the vowel signs:

◌ೀ -ī ◌ು -u ◌ೂ -ū

ಮ ma ಮೀ

ಪ pa ಪು ಪೂ

ಫ pha ಫು ಫೂ

ವ va ವು ವೂ

¶391 ***Non-joined forms of clusters..***

¶392 ***Clusters without vowels..*** For Sanskrit or foreign words ending in a consoant sound, the absence of a final vowel sound is indicated by the presence of the sign ◌್ above the consonant.

¶393 ***Typographic fragments.*** The coded character U+0CD5 ◌ೕ KANNADA LENGTH MARK does not correspond to an abstract character and is encoded solely to represent the second element of U+0CC7 ◌ೀ KANNADA VOWEL SIGN EE in isolation.

¶394 The coded character U+0CD6 ◌ೖ KANNADA AI LENGTH MARK does not correspond to an abstract character and is encoded solely to represent the second element of U+0CC8 ◌ೈ KANNADA VOWEL SIGN AI in isolation.

6.9 Malayalam

¶395 The Malayalam script is a South Indian script used to write the Malayalam language of the Kerala state. Malayalam is a Dravidian language like Kannada, Tamil, and Telugu. Throughout its history, it has absorbed words from Tamil, Sanskrit, Arabic, and English.

¶396 ***Standards.*** The Malayalam block of the Unicode standard is based on ISCII-1988. See section *4.1, The* ISCII *Standard* , for more details on the relation to ISCII.

¶397 ***Encoding.*** The encoding of the Malayalam script follows the consonant linking model. The consonant linker coded character of Malayalam is U+0D4D ◌് MALAYALAM SIGN VIRAMA.

¶398 ***Letters.*** Malayalam has the full complement of basic letters presented in section *2.1, Letters* , except for the vowel sign for ഌ ഌ . In addition, Malayalam as the full complement of additional letters presented in section , except for ഩ ṉa.

¶399 The consonant nasalization sign ◌ം and is encoded as U+0D02 ◌ം MALAYALAM SIGN ANUSVARA.

¶400 The unvoiced aspiration sign ◌ഃ is encoded as U+0D03 ◌ഃ MALAYALAM SIGN VISARGA.

¶401 ***Chillu Characters.*** The letters ൺ ൺ , ൻ ൻ , ർ ർ , ൽ ൽ , ൾ ൾ , and ൿ ൿ are know as *chillu* or *cillakṣaram* characters.

¶402 In Malayalam-language text, chillu letters never start a word. The chillu letters ൺ ൺ , ൻ ൻ , ർ ർ , ൽ ൽ , and ൾ ൾ are quite

common; ക് ക് is not very common.

¶403 Prior to Unicode 5.1, the representation of text with chillus was problematic, and not clearly described in the text of the standard. Because older data will use different representation for chillus, implementations must be prepared to handle both kinds of data. The following table shows the relation between the representation in Unicode Version 5.0 and earlier and the new representation in Version 5.1, for the chillu letters considered in isolation.

Character	Representation in 5.0 and prior	Preferred 5.1 and later representation
ൺ ൺ	NNA, VIRAMA, ZWJ (0D23, 0D4D, 200D)	U+0D7A ൺ MALAYALAM LETTER CHILLU NN
ൻ ൻ	NA, VIRAMA, ZWJ (0D28, 0D4D, 200D)	U+0D7B ൻ MALAYALAM LETTER CHILLU N
ർ ർ	RA, VIRAMA, ZWJ (0D30, 0D4D, 200D)	U+0D7C ർ MALAYALAM LETTER CHILLU RR
ൽ ൽ	LA, VIRAMA, ZWJ (0D32, 0D4D, 200D)	U+0D7D ൽ MALAYALAM LETTER CHILLU L
ൾ ൾ	LLA, VIRAMA, ZWJ (0D33, 0D4D, 200D)	U+0D7E ൾ MALAYALAM LETTER CHILLU LL
ക് ക്	undefined	U+0D7F ക് MALAYALAM LETTER CHILLU K

¶404 The letter റ റ is normally read r. Repetition of that sound is written by two occurrences of the letter: ററ. Each occurrence can bear a vowel sign.

¶405 Repetition of the letter, written either ററ or റ്റ, is also used for the sound tt. In this case, the two റ fundamentally behave as a digraph. The digraph can bear a vowel sign in which case the digraph as a whole acts graphically as an atom: a left vowel part goes to the left of the digraph and a right vowel part goes to the right of the digraph. Historically, the side-by-side form was used until around 1960 when the stacked form began appearing and supplanted the side-by-side form. The same situation is common in many other orthographies. For example, *th* in English can be a digraph (*cathode*) or two separate letters (*cathouse*); *gn* in French can be a digraph (*oignon*) or two separate letters (*gnome*).

¶406 The sequence <0D31, 0D31> represents ററ, regardless of the reading of that text. The sequence <0D31, 0D4D, 0D31> represents റ്റ. In both cases, vowels signs can be used as:

പാറ	0D2A 0D3E 0D31 0D31	paatta	cockroach
പാറ്റ	0D2A 0D3E 0D31 0D4D 0D31		
മാറൊലി	0D2E 0D3E 0D31 0D46 0D31 0D3E 0D32 0D3F	maattoli	echo
മാറ്റൊലി	0D2E 0D3E 0D31 0D4D 0D31 0D46 0D3E 0D32 0D3F		
ബാററി	0D2C 0D3E 0D31 0D31 0D31 0D3F	baattari/	battery
ബാറ്ററി	0D2C 0D3E 0D31 0D4D 0D31 0D31 0D3F		
സൂററ്	0D38 0D42 0D31 0D31 0D31 0D4D	suu̲ratt	(name of a place)
സൂറ്റ്	0D38 0D42 0D31 0D31 0D4D 0D31 0D4D		
ടെംപററി	0D1F 0D46 0D02 0D2A 0D31 0D31 0D3F	tempa̲ra̲ri	temporary (English loan word)
ലെക്ചററോട്	0D32 0D46 0D15 0D4D 0D1A 0D31 0D31 0D4B 0D1F 0D4D	/lekca̲ra̲root/	to the lecturer

¶407 A very similar situation exists for the combination of ൻ ൻ and റ റ . When used side by side, ൻറ can be read either /nr/ or /nt/, while ൻ്റ is always read /nt/.

¶408 The sequence <0D7B, 0D31> represents ൻറ, regardless of the reading of that text. The sequence <0D7B, 0D4D, 0D31>

represents ൻറ. In both cases, vowels signs can be used as appropriate:

ആന്റോ	0D06 0D7B 0D47 0D31 0D3E	aantoo (proper name)
ആൻ്റോ	0D06 0D7B 0D4D 0D31 0D47 0D3E	
എൻറോൺ	0D0E 0D7B 0D31 0D47 0D3E 0D7A	/enrool/ enroll (English loan word)

¶409 ***Digits and Numbers.*** Malayalam has a full set of decimal digits, encoded at U+0D66 ൦ MALAYALAM DIGIT ZERO .. U+0D6F ൯ MALAYALAM DIGIT NINE.

¶410 Additional numbers and fraction characters are encoded at U+0D70 ൰ MALAYALAM NUMBER TEN .. U+0D75 ൵ MALAYALAM FRACTION THREE QUARTERS.

¶411 ***Punctuation and Signs.*** Malayalam uses the danda । and double danda ॥ encoded as U+0964 । DEVANAGARI DANDA and U+0965 ॥ DEVANAGARI DOUBLE DANDA.

¶412 The *avagraha* is known as the *praslesham* in Malayalam and is encoded as U+0D3D ഽ MALAYALAM SIGN AVAGRAHA.

¶413 The date mark ൗ is encoded at U+0D79 ൹ MALAYALAM DATE MARK.

¶414 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ഈ	0D08	<0D07, 0D57>
ഊ	0D0A	<0D09, 0D57>
ഐ	0D10	<0D0E, 0D46>
ഓ	0D13	<0D12, 0D3E>
ഔ	0D14	<0D12, 0D57>

¶415 ***Typographic Cluster Cores.***

¶416 ***Satellite consonants.***

¶417 ***Satellite vowels.*** In modern times, the dominant practice is to write the dependent form of the ഔ ഔ vowel using only ൗ which is placed on the right side of the consonant it modifies; such texts are represented in Unicode using U+0D57 ◌ൗ MALAYALAM AU LENGTH MARK. In the past, this dependent form was written using both ൌ on the left side and ൗ on the right side; U+0D4C ◌ൌ MALAYALAM VOWEL SIGN AU can be used for documents following this earlier tradition. This historical simplification started much earlier than the orthographic reforms mentioned above.

¶418 ***Conjoined and Non-Joined Forms of Clusters.***

¶419 ***Typographic Fragments.*** todo

6.10 Sinhala

¶420 The Sinhala script, also known as Sinhalese, is used to write the Sinhala language, the majority language of Sri Lanka. It is also used

to write the Pali ans Sanskrit languages. The script is a descendant of Brahmi and resembles the scripts of South India in form and structure.

¶421 ***Encoding.*** The encoding of the Sinhala script follows the consonant linking model. The consonant linker coded character of Sinhala is U+0DCA ◌් SINHALA SIGN AL-LAKUNA.

¶422 ***Letters.*** Sinhala has the full complement of basic letters presented in section *2.1, Letters* .

¶423 The consonant nasalization sign ◌ං is known as the *ansuvaraya* in Sinhala, and is encoded as U+0D82 ◌ං SINHALA SIGN ANUSVARAYA.

¶424 The unvoiced aspiration sign ◌ඃ is known as the *visargaya* in Sinhala, and is encoded as U+0D83 ◌ඃ SINHALA SIGN VISARGAYA.

¶425 Sinhala differs from other languages of the region in that it has a series of prenasalized stops that are distinguished from the combination of a nasal followed by a stop. In other words, both forms occur and are written differently – for example, AB <U+0D85, U+0DAC> a8}a [a:;a] "sound" versus ACDE <U+0D85, U+0DAB, U+0DCA, U+0DA9> aV}a [a9;a] "egg." In addition, Sinhala has separate distinct signs for both a short and a long low front vowel sounding similar to the initial vowel of the English word "apple," usually represented in IPA as U+00E6 æ latin small letter ae (ash). The independent forms of these vowels are encoded at U+0D87 and U+0D88; the corresponding dependent forms are U+0DD0 and U+0DD1.

¶426 Because of these extra letters, the encoding for Sinhala does not precisely follow the pattern established for the other Indic scripts (for example, Devanagari). It does use the same general structure, making use of phonetic order, matra reordering, and use of the virama (U+0DCA sinhala sign al-lakuna) to indicate conjunct consonant clusters. Sinhala does not use half-forms in the Devanagari manner, but does use many ligatures.

¶427 ***Digits and Numbers.*** While the Sinhala script has a full set of decimal digits, those are not encoded at the present time..

¶428 ***Punctuation and Signs.*** 𑇅 was formerly used as a full stop or period and is encoded as U+0DF4 𑇅 SINHALA PUNCTUATION KUNDDALIYA.

¶429 ***Atomicity of Vowel Letters.*** Vowel letters are encoded atomically in Unicode, even if they can be analyzed visually as consisting of multiple parts. The table below 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.

To represent	Use	Do not use
ඦ	0D86	<0D85, 0DCF>
ඇ	0D87	<0D85, 0DD0>
ඈ	0D88	<0D85, 0DD1>
ඌ	0D8C	<0D8B, 0DDF>
ඎaa	0D8E	<0D8D, 0DD8>
ඐ	0D90	<0D8F, 0DDF>
ඵ	0D92	<0D91, 0DCA>
ඓ	0D93	<0D91, 0DD9>
ඖ	0D96	<0D91, 0DDF>

¶430 ***Typographic Cluster Cores.*** A few consonant clusters are

¶431 ***Satellite Consonants.*** ර් ර in initial position is a satellite consonant. It is then displayed as a small hook, called *repaya*, on the top right side of the cluster.

¶432 todo examples

¶433 ◌్ ◌్ in final position is a satellite consonant. It is displayed as a curve below the cluster core, called a *rakaaraansaya*.

¶434 todo examples

¶435 ◌ు ◌ు in final position is a satellite consonant. It is displayed by a *yansay* on the right side of the cluster.

¶436 todo examples

Appendix A. Acknowledgments

¶437 Thanks to the Editorial Committee, Asmus Freytag.

Appendix B. Things to do

¶438 This section is of course not part of the text; it's just a place to remember the things to do to this text.

- ¶439 in the description of the models, introduce VIRAMA (small caps) as the generic name for the consonant linker coded character?
- ¶440 set baseline-to-baseline to some fixed value

Appendix C. References

References

[Snell]	Beginner's Hindi Script *London: Hodder & Stoughton, 2003. ISBN 0-07-141984-5*
[Arden]	A Progressive Grammar of the Telugu Language *Madras: The Society For Promoting Christian Knowledge, 1905*
[Brown]	A Grammar of the Telugu Language *Madras: The Christian Knowledge Society's Press, 1857*
[Feedback]	http://www.unicode.org/reporting.html *For reporting errors and requesting information online.*
[Reports]	Unicode Technical Reports http://www.unicode.org/reports/ *For information on the status and development process for technical reports, and for a list of technical reports.*
[Unicode]	*The Unicode Standard, Version 4.0* (Boston, MA, Addison-Wesley, 2003. ISBN 0-321-18578-1), as amended by *Unicode 4.0.1* (http://www.unicode.org/versions/Unicode4.0.1) and by *Unicode 4.1.0* (http://www.unicode.org/versions/Unicode4.1.0).
[Versions]	Versions of the Unicode Standard http://www.unicode.org/versions/ *For details on the precise contents of each version of the Unicode Standard, and how to cite them.*

Appendix D. Modifications

¶447 This section indicates the changes introduced by each revision.

¶448 ***Revision 1.***

- ¶449 First version