An Introduction to Indic Scripts

Richard Ishida
W3C
ishida@w3.org

HTML version: http://www.w3.org/2002/Talks/09-ri-indic/indic-paper.html
PDF version: http://www.w3.org/2002/Talks/09-ri-indic/indic-paper.pdf

Introduction

This paper provides an introduction to the major Indic scripts used on the Indian mainland. Those addressed in this paper include specifically Bengali, Devanagari, Gujarati, Gurmukhi, Kannada, Malayalam, Oriya, Tamil, and Telugu.

I have used XHTML encoded in UTF-8 for the base version of this paper. Most of the XHTML file can be viewed if you are running Windows XP with all associated Indic font and rendering support, and the Arial Unicode MS font. For examples that require complex rendering in scripts not yet supported by this configuration, such as Bengali, Oriya, and Malayalam, I have used non-Unicode fonts supplied with Gamma's Unitype. To view all fonts as intended without the above you can view the PDF file whose URL is given above.

Although the Indic scripts are often described as similar, there is a large amount of variation at the detailed implementation level. To provide a detailed account of how each Indic script implements particular features on a letter by letter basis would require too much time and space for the task at hand. Nevertheless, despite the detail variations, the basic mechanisms are to a large extent the same, and at the general level there *is* a great deal of similarity between these scripts. It is certainly possible to structure a discussion of the relevant features along the same lines for each of the scripts in the set. It is these common themes that this discussion will attempt to highlight, although we will also mention some of the more important deviations from the common path.

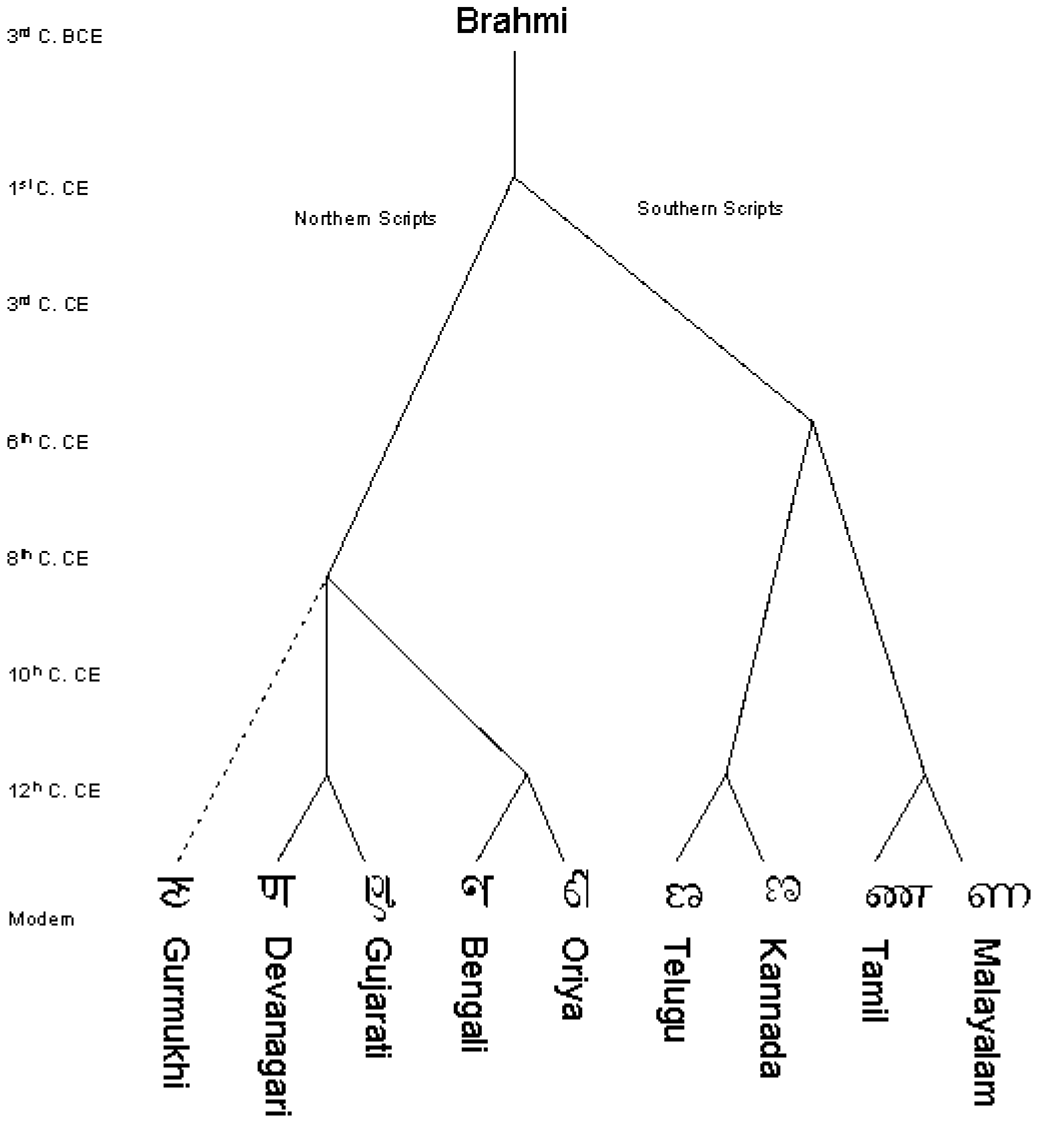
This paper will tackle the subject in two parts:

1. In the first part of this paper I will survey the *visual characteristics* of these scripts.
2. In the second part I will make brief reference to some of the *practical implications* of supporting Indic scripts using Unicode.

A note on terminology: The names of letters and diacritics with a particular function (and location relative to the beginning of the code block) are largely standardised in Unicode across all the Indic code charts. For example, although the 'vowel killer' diacritic may be called a 'pulli' in Tamil, it is still referred to by the Unicode character names as a 'virama'. In order to simplify the explanations and show better the commonality between the scripts, we will use the generic names for characters provided by Unicode. In addition, when it is occasionally necessary to refer to a specific letter by name part of the Unicode name will be used in upper case.

Historical background

The similarity of features across all these scripts is not surprising if you consider their history. As shown in the illustration below, they all derive from a common ancestor. Note also that these scripts are used for two distinct major linguistic groups, Indo-European languages in the north, and Dravidian languages in the south.



An illustration of the derivation of the character NNA, showing how from a common source (Brahmi) all the different forms arose for the modern scripts. The diagram shows an early divergence between north and south indian scripts. (Adapted from Daniels and Bright, *The World's Writing Systems*.)

Sounds of Indic languages

One of the defining aspects of a script is the repertoire of sounds it has to support. Because there is typically a letter for each of the phonemes in an Indic language, the alphabet tends to be quite large. The table below shows a superset of Indic consonant sounds in a traditional articulatory arrangement. It is meant to be illustrative rather than exhaustive, so as to give you an idea of the number of sounds most Indic scripts must support. The table also provides an *approximate* idea of how Unicode character names map to actual sounds. The IPA transcription is shown to the left, followed by the standard Unicode name for that sound. Note the following:

- retroflex variants of a basic sound are found in most Indian languages,
- a plosive sound typically has an aspirated and unaspirated version,
- many languages also recognise one or more combinations as a single unit for sorting or other purposes, eg. [kʃ],
- it is common for consonant sounds in particular locations to be held for longer than usual (or in the case of plosives, slightly delayed) - these geminated consonants are typically shown by writing two consonants together, although the actual visual appearance can become quite complicated (this is described in more detail later).

A very generalised table, showing the range of typical Indic sounds.

			Uvular		**Velar**		**Palatal**		**Retroflex**		**Dental**		**Labial**	
Plosives	**Voiceless**	**Unaspirated**	q	QA	k	KA	c	CA	ʈ	TTA	t	TA	p	PA
		Aspirated			k^h	KHA	c^h	CHA	ʈ^h	TTHA	t^h	THA	p^h	PHA
	Voiced	**Unaspirated**			g	GA	ɟ	JA	ɖ	DDA	d	DA	b	BA
		Aspirated			g^h	GHA	ɟ^h	JHA	ɖ^h	DDHA	d^h	DHA	b^h	BHA
Nasals					ŋ	NGA	ɲ	NYA	ɳ	NNA	n	NA	m	MA
Fricatives		**Voiceless**			x	KHHA	ʃ	SHA	ʂ	SSA	s	SA	f	FA
		Voiced			ɣ	GHHA					z	ZA		
Flapped & tapped sounds									ɽ	DDDHA	ɾ	RA		
									ɽ^h	RHA				
Aspirate, semi-vowels and liquid					h	HA	j	YA	ɭ	LLA	l	LA	v	VA

There are up to 18 Unicode code points dedicated to vowels in each script block, although fewer than this are actually needed on a per language basis. Nearly always these are simple vowel sounds, although occasionally a symbol may represent a diphthong (especially AI and AU).

Indic languages are syllabic in nature, and the *inherent vowel* is an important concept of a syllable (see below). This vowel can vary in pronunciation from script to script, and examples include [ə], [ʌ], and [ɔ].

Nasalisation of vowels is also an important phonetic feature that affects the written form of several South Asian languages. The effect is similar to the nasalisation of words like 'en' in French.

Characteristic script features

Direction and positioning of script

All Indic scripts run left to right, although some combining glyphs appear to the left of their base character for display (see the discussion of vowel signs below).

In a number of scripts, characters commonly have a headstroke and a high baseline. Such characters typically hang from the line when written.

Consonants and inherent vowels

These scripts are often called *abugidas* or *alpha-syllabaries*.

In this type of script consonant characters represent a consonant+vowel syllable. The consonant is associated with an *inherent vowel* that has to be overridden if it is not the required vowel sound for a particular spoken syllable (see the following section). For example, the character क in Hindi (Devanagari script) is pronounced [kə] rather than just [k]. The [ə] sound is the inherent vowel, and is usually transcribed as 'a'.

Note that the inherent vowel is not always pronounced. For example in Hindi it is not usually pronounced at the end of a word, although a ghost echo may appear after a word-final cluster of consonants, eg. योग्य [jogjᵊ], or राष्ट्र [ɾəstɾᵊ]. In addition Hindi has a general rule that when a word has three or more syllables and ends in a vowel other than the inherent a, the penultimate vowel is not pronounced, eg. समझ [səməɟʰ] but समझा [səmɟʰaː], रहन [rəhən] but रहना [rəhnaː]. (For a number of reasons, however, this rule does not always hold.)

Nonetheless, on the whole, Indic scripts are close to phonemic transcriptions. The pronunciation of consonants is typically quite regular and predictable, although there is the occasional exception.

> Example 1: aspirated voiced plosives and the non-initial letter HA in Gurmukhi are used to indicate tones rather than sounds. For example a voiced, aspirated plosive in word-initial position represents an *unvoiced, unaspirated* plosive sound with a low tone on the syllable, eg. ਘੋੜਾ [kòɽɑ] (The primary use of all voiced aspirated plosives in Gurmukhi is to express tone information.)
>
> Example 2: in Tamil, consonants such as க are typically phonemic rather than phonetic. This consonant may be any of [kʌ, gʌ, xʌ, ɣʌ, hʌ].

Most scripts supplement a basic set of letters with additional letters used to represent the sounds of other languages, such as Sanskrit and English. These additional letters are commonly formed by adding a diacritic to an existing letter. This diacritic is called a *nukta* in The Unicode Standard, although the name used by speakers of different Indian languages may vary. Some scripts use this diacritic with several basic letters (eg. Devanagari), others not at all (eg. Kannada).

Examples:

Devanagari, क़ [qə] (cf. क [kə])

Gurmukhi, ਲ਼ [ɭə] (cf. ਲ [lə])

Oriya, ଢ଼ [ɽʰ] (cf. ଢ [ɖʰ])

It is possible to 'kill' the inherent vowel sound where it would normally be pronounced. This is achieved by attaching a small diacritic mark, called a *virama* in The Unicode Standard, to the consonant in question.

Examples:

Gujarati, ક્ [k] (cf. ક [kə])

Tamil, க் [k] (cf. க [kʌ])

Telugu, క్ [k] (cf. క [kʌ])

In the examples that follow, consonants without the inherent vowel are depicted with a virama.

Vowel signs

Where a consonant is followed by a vowel other than the inherent vowel, the change is produced by adding a *vowel sign* to the base consonant (called a *matra* in Sanskrit). A consonant can only support one vowel (and one vowel sign) at a time (unlike Thai). A vowel sign may appear to the left or right, above or below the base consonant, and sometimes surrounds the base consonant on more than one side.

The following illustrates the use of vowel signs with the क consonant in Hindi, and the resultant sounds:

की [kiː] के [ke] कू [kuː]

Vowel signs may also appear to the left of the base consonant they are related to. For example:

Gujarati, ક + ◌િ -> કિ [ki]

Tamil, க + ◌ை -> கை [kʌy])

Occasionally a vowel sign may be composed of multiple parts. In some cases such a split vowel sign may have parts on both the left and right of the base character simultaneously, eg. in Tamil க + ◌ொ -> கொ [ko]. In Kannada there are no vowel signs that surround a base character on both left and right, but there are some that have multiple parts above and following the base character, eg. ಕ + ◌ೋ -> ಕೋ [koː]. Another alternative is top and bottom, eg. Telugu క + ◌ై -> కై [aj]. In some cases, the additional parts can be viewed as lengthening marks.

Often the pairing of base character and vowel sign produces a change in the basic shape of either base character or vowel sign or both. Tamil provides many such examples, especially with

[u] and [uː]. For example, the following are most of the Tamil consonants, each followed by the same vowel sign, ◌ு [u]:

Without vowel sign	க ங ச ஜ ஞ ட ண த ந ன ப ம ய ர ற ல ள ழ வ ஷ ஸ ஹ
With vowel sign	கு ஙு சு ஜு ஞு டு ணு து நு னு பு மு யு ரு று லு ளு ழு வு ஷு ஸு ஹு

Independent vowels

Vowels that appear at the beginning of a word or after a preceding vowel with no intervening consonant are typically rendered using independent vowel letters. The following table illustrates the correspondence between the most common independent vowels and vowel signs in Telugu:

Unicode name	A	AA	I	II	U	UU	vR	E	EE	AI	O	OO	AU
Independent vowel	అ	ఆ	ఇ	ఈ	ఉ	ఊ	ఋ	ఎ	ఏ	ఐ	ఒ	ఓ	ఔ
Vowel sign	-	◌ా	◌ి	◌ీ	◌ు	◌ూ	◌ృ	◌ె	◌ే	◌ై	◌ొ	◌ో	◌ౌ
Pronunciation	ʌ	aː	i	iː	u	uː	rɨ/ru	e	eː	aj	o	oː	aw

Note that there is no vowel sign for the sound associated with the inherent vowel A. Vowel signs are only needed to *change* the inherent vowel.

Because a consonant (or consonant cluster) can only support one vowel at a time, note the difference in Devanagari between की [kiː] and कई [kəiː].

Gurmukhi is unusual in that (with the exception of ਅ [ə]) there are no independent vowels. Instead there are special 'vowel-bearer' glyphs (of which ਅ is one) that are used to support the vowel signs.

ਅ is used for [ɑ] ਆ, [æ] ਐ, [ɔ] ਔ,

ੲ is used for [ɪ] ਇ, [i] ਈ, [e] ਏ,

ੳ is used for [ʊ] ਉ, [u] ਊ, [o] ਓ,

Consonant clusters

Where consonants appear together without intervening vowels special steps need to be taken to indicate that the inherent vowels have disappeared. There are many ways in which this is achieved in Indic scripts, and the specifics of how each character behaves are too many to catalogue here in detail for each script. There are, however, two main approaches: either (a) change the shape of the consonants or merge them together in some way (a *conjunct* form), or (b) use a special diacritic to indicate the absence of intervening vowels.

A number of strategies are used to show consonant clusters by merging or changing shapes, and nearly all scripts employ more than one of these approaches. The following are a few examples:

- For 60% of Devanagari conjuncts the consonants that lose the vowel typically lose their characteristic vertical bar (which is historically associated with the sound of the inherent vowel). Such glyphs are referred to as *half-forms*. For example, स् [s] + म [mə] -> स्म [smə].
- Sometimes the two consonant glyphs may be combined vertically. For example, certain combinations in Gujarati such as ટ્ [ʈ] + ઠ [ʈʰə] => ટ્ઠ[ʈʈʰə]. Note that the choice of vertical vs. horizontal combination may be a stylistic preference. For example, the result of क् [k] + क [kə] in Devanagari could be rendered as either a vertical or horizontal combination.
- Clustering may produce a conjunct that does not have easily recognisable parts, such as Bengali † [kʃɔ] (= N [k] + k [ʃɔ]), or that expresses the conjunct by extension of a glyph such as the Malayalam ¥ [kːa] (= P} [k] + P [ka]).
- Another common approach is to reduce and simplify one of the consonants in the cluster, and then attach it to the other like a diacritic. In Kannada most combinations are formed by reducing non-initial consonant glyphs in a cluster to a simplified form, joined beneath and/or to the right of the initial consonant, eg. ತ್ + ಯ -> ತ್ಯ [tjʌ]. Oriya also often reduces the second consonant, but in some combinations will reduce the *first* consonant and attach it to the bottom of the second.

Another approach is to simply show the virama we introduced earlier. This is really the standard approach for modern Tamil, eg. இந்த [intʌ] (the dot is the virama), but may also be used for any other script if the font being used does not support all the necessary ligatures and alternative glyph forms. Oriya tends to use the virama specifically for borrowed words.

There is actually a third approach, and that is to simply rely on the user to recognise contexts where the inherent vowel is dropped. This occurs in some specific situations such as at the end of a word or those examples described earlier. The only script that does this as a *general rule* is Gurmukhi. Very few letter combinations are handled as conjuncts in Gurmukhi, most of the time the reader just has to know where the inherent vowel is not pronounced, eg. ਉਤਸੁਕ [utsuk].

A common feature of Indic scripts is the gemination or lengthening of consonants. For example, note the lengthening of the ल [l] sound in चिल्लहट [cilːʌhʌʈ] (Devanagari). Such consonant lengthening is typically handled just as a normal consonant cluster. Gurmukhi, again, is somewhat non-standard in that it uses a special diacritic called *addak* in this situation. To indicate a geminated consonant, the addak sits above the *preceding* syllable, eg. ਪੁੱਤਰ [putːər].

The letter RA is a very common example of a letter that behaves quite idiosyncratically in consonant clusters, and typically quite differently depending on whether it appears at the beginning or end of the cluster. Its placement also often involves apparent reordering. This can be illustrated with the Devanagari RA र. When in initial position in the cluster the letter र is typically displayed as a small mark above the *right* shoulder of the last letter in the syllable, eg. शार्मा [ʃaːrmaː]. This is called a *repha*. A र in final position in a conjunct cluster is displayed as a small diagonal mark, but precisely where it appears depends on the shape of the previous member, eg. प्र [pra], त्र [tra], ह्र [hra]. With TTA and DDA it needs a little supporting line, eg ट्र [ʈra], ड्र [ɖra].

Note that a cluster is not limited to two consonants, eg.

र् [r] + ग् [g] + घ [dʰʌ] -> र्ग्घ [rgdʰʌ]

This use of the repha, appearing as it does to the right of the whole cluster, demonstrates the syllabic nature of the Indic scripts. The following extension of the example shows even more clearly that the repha is actually positioned to the right of the syllable, rather than just the cluster, since it appears above the vowel sign.

र् [r] + ग् [g] + घ [dʰʌ] + ◌ी [iː] -> र्ग्घी [rgdʰiː]

(Note that syllable boundaries in spoken text do not equate to those in written text. For example, 'Hindi' is spoken as 'hin-di', but written as 'hi-ndi'.)

Vowel signs used with consonant clusters

Where the vowel following a consonant cluster is rendered with a vowel sign, the placement of the vowel sign may need attention. As with the examples of the repha at the end of the last section, the syllabic nature of the script becomes apparent with the use of reordrant vowel signs attached to consonant clusters.

In Devanagari, where a non-inherent vowel is pronounced immediately after the cluster and is normally rendered to the left of a character, it will be rendered to the left of the whole cluster, eg. in मुश्किल [muʃkil], the ि◌ is pronounced after the क.

Vowel signs in a script like Kannada are visually attached to the first consonant in the cluster. Note how the vowel sign appears over the [k] in ಕ್ರಿ [kri], since the [r] is rendered as a reduced appendage at the bottom right of the first consonant in the cluster.

Nasalisation and alternative nasal letter representations

There are three diacritics associated with the nasalisation of vowels or the alternative representation of nasal consonants as part of a consonant cluster. Which diacritic is used for which purpose varies from script to script. The following are a few examples of usage. (As usual, although these diacritics have their own names in the various languages represented by the scripts, I will refer to them using the genericised names used in The Unicode Standard):

- In Devanagari, nasalisation of vowel sounds is indicated using the *candrabindu* ◌ँ or *anusvara* ◌ं diacritics, eg. अँग्रेज़ [ʌ̃grez], नहीं [nʌhĩː]. The anusvara is commonly used in conjunction with a vowel sign that extends above the headstroke.

  Nasal consonants in initial place in a conjunct may also be expressed using the anusvara over the previous vowel, rather than as a half-glyph attached to the following consonant. The anusvara is written above the headstroke, at the right-hand end of the preceding character. In the list below both spellings are correct and equivalent, although the anusvara is preferred in the case of the first two: रंग = रङ्ग [rʌŋg], □̇□ ाबी = □ञ्□ ाबी [pʌɲɟaːbiː], हिंदी = हिन्दी [hindiː], लंबा = लम्बा [lʌmbaː]. Note that the anusvara is still applied when the previous character has its own vowel sign. If the vowel sign is AA, the anusvara appears over the AA, eg. फ़्रांसीसी or आंदोलन.

- In Kannada the anusvara ◌ಂ is mostly used for nasal consonants that are homorganic with a following stop, eg. ಅಂಗ [ʌŋga]. When followed by a consonant other than a stop or when word final, the anusvara is pronounced [m], eg. ಸಂಹ [simha], ಲಗಾಂ [lʌgaːm].
- In Oriya, homorganic nasal+stop clusters are usually written with distinctive conjunct letters. However, the nasal may also be written with anusvara, eg. ଅଂକ [ɔŋkɔ].

  Nasalised vowels use the bindu, eg. ଆଁ [ã], କାଁ [kã].
- Gurmukhi is unusual in that it has its own special diacritic for indicating nasalisation of vowel sounds. The tippi ◌ੰ is used over the preceding syllable with [a, ɪ, ʊ] and final [u], eg. ਮੁੰਡਾ [mʊɳɖɑ].

  All other vowels use the anusvara ◌ਂ (called bimdi in Panjabi), eg. ਸ਼ਾਂਤ [ʃɑ̃t].

The visarga

The visarga is commonly required for transcribing Sanskrit, but occasionally has more specific uses too. It is not used at all in Gurmukhi.

The pronunciation of the visarga may vary. In Kannada it is commonly pronounced [ha], eg. ಪುನಃ [punəha]. In Gujarati it is typically silent.

In Tamil, the visarga is used to create additional fricative sounds. Before PA it creates [f], and before JA it creates [z], eg. ◌ஃபீசு[fiːsɯ], ◌ஃஜிராக்ஸ்[ziroks]. [Note: the glyphs for the visarga should not have the dotted circle before them. This is 'feature' of the font inherited from the fact that incorrect semantics were applied to this Unicode character prior to Unicode version 3.2 (see below).]

Numbers

All scripts have their own number shapes. While some scripts, such as Tamil, tend to favour European numerals over their own in modern text, other scripts, such as Hindi, still make heavy use of their native shapes.

The following table shows the number symbols:

European	0 1 2 3 4 5 6 7 8 9
Devanagari	० १ २ ३ ४ ५ ६ ७ ८ ९
Bengali	০ ১ ২ ৩ ৪ ৫ ৬ ৭ ৮ ৯
Gurmukhi	੦ ੧ ੨ ੩ ੪ ੫ ੬ ੭ ੮ ੯
Gujarati	૦ ૧ ૨ ૩ ૪ ૫ ૬ ૭ ૮ ૯
Oriya	୦ ୧ ୨ ୩ ୪ ୫ ୬ ୭ ୮ ୯
Tamil	௧௨௩௪௫௬௭௮௯ (no zero)
Telugu	౦ ౧ ౨ ౩ ౪ ౫ ౬ ౭ ౮ ౯
Kannada	೦ ೧ ೨ ೩ ೪ ೫ ೬ ೭ ೮ ೯
Malayalam	൦ ൧ ൨ ൩ ൪ ൫ ൬ ൭ ൮ ൯

Tamil number shapes are not based on the decimal system, and so there is no zero. There are however additional symbols to represent 10 ௰, 100 ௱, and 1000 ௲. Modern Tamil typically uses European numerals.

Text units & punctuation

Sub-sentence units (words) are separated with spaces.

Modern text commonly uses western punctuation, but some traditional punctuation is used in some scripts. For example, the DANDA । may still be used in Devanagari to mark the end of a sentence, or the DOUBLE DANDA ॥ in Telugu for certain abbreviations.

Implementation notes

Character sets

The characters in an Indic Unicode script block are a superset of the ISCII (Indian Standard Code for Information Interchange) character sets. The ISCII standard includes separate encodings for each of the scripts discussed here, using escape sequences to shift between them. The Unicode blocks were originally based on the 1988 version of ISCII encodings. ISCII published a new version of the standard in 1991 with a few changes to order and repertoire of characters. Unicode, nevertheless, remains a superset of all the ISCII codes, with the exception of a few Vedic extension characters.

The first 85 characters in each Unicode block are in the same order and position, on a script by script basis, as the ISCII characters for the respective script. Every script block orders analogous characters in the ISCII range in the same relative locations to the start of the block across all 9 scripts under consideration in this paper. The next 27 characters are additional Unicode characters, where each analogous character across the scripts is also assigned to the same code point relative to the beginning of the block. The final column is reserved for script specific characters. There is no special ordering here.

The zones described above are illustrated here using the Devanagari script block. The yellow background shows the ISCII-derived range, the orange is the coordinated Unicode extension range, and the grey is the range of additional script specific characters.

	ऐ	ठ	र	◌ी	ॐ	ॠ	॰
◌ँ	ऑ	□	ऱ	◌ु	◌॑	ॡ	
◌ं	ऒ	ढ	ल	◌ू	◌॒	◌ॢ	
◌ः	ओ	ण	ळ	◌ृ	◌॓	◌ॣ	
	औ	त	ऴ	◌ॄ	◌॔	।	
अ	क	थ	व		◌	॥	
आ	ख	द	श		◌ॆ	०	
इ	ग	ध	ष		◌े	१	
ई	घ	न	स	◌ै	क़	२	
उ	□	ऩ	ह	◌ॉ	ख़	३	
ऊ	च	□		◌ॊ	ग़	४	
ऋ	छ	□		◌ो	ज़	५	
ऌ	□	ब	◌़	◌ौ	ड़	६	
ऍ	झ	भ	ऽ	◌्	ढ़	७	
ऎ	ञ	म	◌ा		फ़	८	
□	ट	य	ि◌		य़	९	

The fact that ISCII and Unicode attempt to use the same code point relative to the start of the code block for analogous characters across all nine Indic scripts theoretically allows for easy transliteration between the various scripts, however in practice there are quite a few exceptions, so specific tables have to be developed anyway. For a more detailed discussion of this topic see the paper by Ram Viswanadha, *Transliteration of Indic Scripts Implementation in ICU and Lessons Learned*.

Each script block has a different number and distribution of characters. The following table contrasts the allocation of characters for Devanagari, Bengali and Tamil.

Devanagari								Bengali								Tamil							
	ऐ	ठ	र	◌ी	ॐ	ॠ	॰		ঐ	ঠ	র	ী		ৠ	ৰ		ஐ		ர	◌ீ			௰
◌ँ	ऑ	□	ऱ	◌ु	◌॑	ॡ		ঁ		ড		ু		ৡ	ৱ				ற	◌ு			௱
◌ं	ऒ	ढ	ल	◌ू	◌॒	◌ॢ		ং		ঢ	ল	ূ		ৢ	৲	◌ஂ	ஒ		ல	◌ூ			௲
◌ः	ओ	ण	ळ	◌ृ	◌॓	◌ॣ		ঃ	ও	ণ		ৃ		ৣ	৳	◌ஃ	ஓ	ண	ள				
	औ	त	ऴ	◌ॄ	◌॔	।			ঔ	ত		ৄ			৴		ஔ	த	ழ				
अ	क	थ	व		◌	॥		অ	ক	থ					৵	அ	க		வ				
आ	ख	द	श		◌ॆ	०		আ	খ	দ	শ			০	৶	ஆ				ெ◌			
इ	ग	ध	ष		◌े	१		ই	গ	ধ	ষ	ে	ৗ	১	৷	இ			ஷ	ே◌	◌ௗ	க	
ई	घ	न	स	◌ै	क़	२		ঈ	ঘ	ন	স	ৈ		২	৸	ஈ		ந	ஸ	ை◌		உ	
उ	□	ऩ	ह	◌ॉ	ख़	३		উ	ঙ		হ			৩	৹	உ	ங	ன	ஹ			௩	
ऊ	च	□		◌ॊ	ग़	४		ঊ	চ	প				৪	৺	ஊ	ச	ப		ெ◌ா		ச	
ऋ	छ	□		◌ो	ज़	५		ঋ	ছ	ফ		ো		৫						ே◌ா		௫	
ऌ	□	ब	◌़	◌ौ	ड़	६		ঌ	জ	ব	়	ৌ	ড়	৬			ஜ			ெ◌ள		௬	
ऍ	झ	भ	ऽ	◌्	ढ़	७			ঝ	ভ		্	ঢ়	৭						◌்		எ	
ऎ	ञ	म	◌ा		फ़	८			ঞ	ম	া			৮		எ	ஞ	ம	◌ா			அ	
□	ट	य	ि◌		य़	९		এ	ট	য	ি		য়	৯		ஏ	ட	ய	◌ி			கூ	

Other notes:

1. We saw earlier how the Tamil visarga was incorrectly rendered in word-initial position (◌ஃஜிராக்ஸ்). This derives from the fact that the Unicode Standard initially classified the Tamil visarga as a combining character. An erratum issued in September 2001 corrected this, changing the General Category from "Mc" (Mark, combining) to "Lo" (Letter, other). The font I am using (Latha) now needs to be updated too.
2. The Kannada letter U+0CDE KANNADA LETTER FA ೞ was incorrectly named. A more appropriate name would be LLLA, rather than FA. Because of the rules for Unicode naming, the current name cannot, however, be changed, although fortunately this letter has not been actively used in Kannada since the end of the 10th century.

Combining characters

It is important to order Indic characters correctly in memory where combining characters are involved. The Unicode Standard requires that all combining characters be stored in memory after the base character they are combined with. This is a fundamentally important concept. It means that, even if a combining glyph appears to the left of its base character, it is stored *after* the base character in memory. (This is often referred to as *logical* ordering.)

For example, the characters in the word 'Hindi' हिंदी are stored in memory as:

ह [h] + ि [i] + ं [n] + द [d] + ी [iː]

Lets see a slightly more complicated example from Kannada. The visual sequence ಕ್ರಿ is pronounced [kri]; the RA is rendered as a subscript to the bottom right and the vowel is rendered as a diacritic above the symbol for KA. The order of characters in memory (ignoring the virama, which is introduced in the next section) is:

ಕ [k] + ರ [r] + ಿ [i]

It is common that multiple characters are combined with a single base character. Examples of such combinations from Devanagari include an anusvara with a vowel sign, eg. हिंदी [hindiː], or a visarga with a vowel sign, eg. दुःख [duhkʰ]. Where there are multiple combining characters The Unicode Standard provides rules about relative ordering that should be observed. If there is a nukta it must immediately follow the base character. Next comes any virama or any vowel sign, then any bindus, and finally the svaras. Observing these rules improves operability and simplifies operations such as search, sorting, character indexing, and the like.

The treatment of combining characters in Indic scripts also necessitates the use of context-based rules in the font to ensure the correct positioning and behaviour of displayed glyphs (a *glyph* being the visual representation of an underlying character). The position of a glyph for a combining character will commonly vary according to the shape and position of its base character, and any other combining characters associated with that base. In a number of cases, the combination of base character and combining character produces a fused shape that must be rendered by use of a ligature or special context-sensitive glyph forms (see also the next section for use of the virama).

Note also that reordering is not limited to displaying certain vowel signs to the left of the immediately preceding base consonant. In a consonant cluster a vowel sign that appears to the left may need to be displayed to the left of the whole consonant cluster, not just the preceding character. Similarly, the symbol for the consonant RA may be rendered as a diacritic at the far right of a syllable involving a consonant cluster that it logically begins.

In addition, because the base character is typed first during normal keyboarding, the base character will typically need to be 'moved' slightly to the right to accommodate the combining character glyph that joins to the left. In practice, the entire word is redrawn with every Indic letter. This is typically done in an off-screen buffer and blitted to the screen. The effect is that characters appear to move and change shape considerably while one is typing.

As mentioned earlier, a number of scripts (Bengali, Oriya, Tamil, Telugu, Kannada, Malayalam) have vowel signs that are composed of more than one part. Such multiple-stroke vowel signs can normally be represented using a single character. For example, Tamil கௌ [kʌʊ] is represented using க and ◌ௌ. Kannada ಕೋ [koː] is composed of ಕ and ◌ೋ. The Unicode charts typically do provide separate codes that can be used to represent multiple-part vowel signs. If these parts are not already available as simple vowel signs, they are provided as special 'length marks' such as ◌ௗ and ◌ೕ. Note also that if two codes are used to represent a split vowel sign both combining characters must *follow* the base character in memory (eg. க + ◌ெ + ◌ௗ for the Tamil example கௌ above).

Although Unicode typically provides single code points for letters formed by the addition of a nukta (eg. क़, ਲ਼, and ଡ଼), these almost all have canonical decompositions to base character plus nukta diacritic.

Since there are alternative ways of representing multi-part vowel signs and consonants created using the nukta, the question arises, "which approach should be used when entering Indic text?" One answer may be to follow the rules of The Character Model for the World Wide Web (see the current working draft at http://www.w3.org/TR/charmod/), which recommends the use of NFC (Normalization Form C) for all web content. NFC represents all multi-part vowels as single characters, but all combinations of consonant plus nukta as two separate characters (apart from the the following exceptions in Devanagari: ऱ RRA, ऴ LLLA, and ऩ NNNA).

It was mentioned above that Gurmukhi is somewhat unusual in that vowel signs are carried by special 'vowel bearer' letters to create independent vowels. While Unicode does provide character codes for these vowel bearers, ੲ and ੳ (plus, of course, ਅ), their use isn't recommended. Instead Unicode provides precomposed codes for all the independent vowel sounds needed, eg. ਆ, ਇ, ਉ, etc.

Note, however, that for the general case whereas some other encoding systems for Indic represent an II SIGN, for example, by VIRAMA + VOWEL II, Unicode does not do that. It considers these two sequences to not be equivalent, and not have the same rendering.

Variant glyph forms

Unicode follows the rule of 'encode characters not glyphs'. This is another fundamentally important concept relating to the support of Indic scripts in Unicode. Even though there are many potential shapes for a character when displayed (half-form, conjunct, ligature, diacritic, etc.), the rule means that there is only one code to represent that character. This is a major advantage for conducting operations on the text such as string comparison, collation, etc. It also allows for a much simpler keyboard, and simpler correspondence between the keyboard input and the stored text. The task of producing the right shape for printing or display of a character according to its context falls to the rendering algorithms of the font, application or system.

We have already mentioned that combining character glyphs may sometimes adopt different shapes or merge with and alter the shape of the base consonant. Another key area where intelligent glyph shaping is required is the display of consonant clusters.

Consonant clusters are invariably indicated in a sequence of Unicode characters by the presence of a virama, whether or not the glyph for the virama will be visible on display. Thus the virama is the trigger for any complex glyph shaping that may be applied to a conjunct by the font or rendering algorithms.

The outcome of a *consonant+virama+consonant* sequence will vary according to the characters, scripts, and fonts involved. Some possibilities are:

- the initial consonant is rendered as a 'half-form' alternative glyph and no virama is shown, eg. क + ् + क = क्क
- the two consonants and virama are represented by a single glyph (a ligature), eg. क + ् + ष = क्ष
- one of the consonants is represented as a combining diacritic (that may or may not be spacing), eg. ತ + ್ + ಯ -> ತ್ಯ.

  Note that in some cases the diacritic may appear in a very different position visually than the character it represents appears in the text stream. For instance, we have already seen the example of the repha in र्ग्घी [rgd^{h}iː]. Even though the RA appears visually at the top right of the cluster, the sequence of characters underlying the cluster is:

  र् [r] + ् + ग् [g] + ् + घ [d^{h}ʌ] + ी [iː]
- the virama may be displayed as a combining glyph. Note that in some scripts (eg. Tamil) this is very much the norm, eg. ந + ் + த = ந்த. In other scripts (such as Devanagari) this is an optional scenario that depends on the preference of the user or the richness of the font - a font that has few ligatures and special glyph forms will resort to simply displaying the virama instead.

In Unicode it should be possible to force a consonant + virama sequence to display the virama (rather than convert the consonant to a half-form or ligature) by adding a *zero width non-joiner* U+200C immediately after the virama of the dead consonant, eg.क्‌क rather than क्क.

To force a dead consonant to assume a half-form rather than combine as part of a ligature, place a *zero width joiner* U+200D immediately after the virama, eg. क्‍ष rather than क्ष. The zero width joiner can also be used to produce an example of a half-form on its own, eg. क्‍ for illustration purposes. This also enables you to create half-forms of combining ligatures, eg. ्र‍.

Other practical considerations

Where scripts use glyphs that hang from the baseline, rather than sitting on the baseline, it is important to ensure that any glyphs from another, intermixed script (eg. Latin script letters) are correctly aligned with the Indic script. It is also important to ensure that the glyphs are aligned as expected with other elements, such as table cells, graphic elements, and the like. For a detailed treatment of the issues for alignment of such scripts with other fonts see Steve Zilles' talk, 'Internationalized Text Formatting in CSS and XML'.

There are other practical considerations related to enabling Indic script input and display. Keyboards must, of course, provide access to all needed characters, but consider standardisation of layout. On-screen display must support adequate resolution and line height, as well as proportional spacing.

For information about collation of Indic scripts, see Unicode Technical Note #1, at http://www.unicode.org/notes/tn1/.

Glossary

abugida

A word used to describe scripts where consonant letters represent syllables with an inherent vowel. See Consonants and inherent vowels.

addak

A diacritic used in Gurmukhi to lengthen the following consonant sound. See Consonant clusters.

anusvara

A diacritic used to represent nasalisation of vowels and/or to represent nasalised consonants. See Nasalisation and alternative nasal letter representations.

articulatory

Related to the production of speech sounds.

aspirated, aspiration

Aspirated consonants are those produced with an audible expulsion of breath. Note that a non-aspirated consonant, such as a [b], is produced with much less aspiration than a similar sound in English.

bindu

A diacritic used to represent nasalisation of vowels and/or to represent nasalised consonants. See Nasalisation and alternative nasal letter representations.

conjunct forms

A special graphical representation used to display a combination of consonants without intervening vowels. See Consonant clusters.

diphthong

A pair of vowels considered to be a single phoneme where the tongue moves from one to the other in such as way as to cause continual change in vowel quality.

glyph

The visual representation of one or more underlying characters. A font is made up of a set of glyph images.

half-form

A reduced version of a consonant glyph (typically missing the vertical stem) used to represent a consonant without a following vowel. See Consonant clusters.

homorganic

A consonant articulated at the same point in the vocal tract as a consonant in another class. For example, [ŋ] is the homorganic nasal of [k].

independent vowel

A vowel used at the beginning of a word or within a word immediately after another vowel sound. See Independent vowels.

inherent vowel

In Indic scripts a consonant character represents a syllable that includes the consonant followed by a default (inherent) vowel sound. This vowel sound varies by language and script. See Consonants and inherent vowels.

ligature

A glyph representing a combination of two or more characters.

nukta

A diacritic used in several indic scripts to extend the range of sounds covered by the alphabet. See Consonants and inherent vowels.

plosive

A sound produced by the mouth in such as way as to temporarily block the passage of the air, eg. a [p].

phoneme

A minimally distinct sound in the context of a particular spoken language. For example, in UK English /p/ and /b/ are distinct phonemes because 'pat' and 'bat' are distinct.

repha

A glyph representing the character RA as the initial consonant in a cluster. The repha appears to the right of the consonant cluster. See Consonant clusters.

retroflex

Retroflex sounds are those made with the tongue being curled upwards.

tippi

A diacritic used in Gurmukhi to represent nasalisation of vowels and/or to represent nasalised consonants in the following syllable. See Nasalisation and alternative nasal letter representations.

virama

A combining mark used to indicate a consonant without a following vowel. See Consonants and inherent vowels.

visarga

A character used most commonly to transcribe Sanskrit, but also sometimes having additional uses such as in Tamil where it is used in conjunction with other characters to create sounds not in the basic repertoire. See Visarga.

vowel sign

A combining character used to indicate the replacement of the inherent vowel associated with a consonant with another vowel sound. See Vowel signs.

References

Sources

1. The Unicode Consortium, *The Unicode Standard -- Version 3.0*, ISBN 0-201-61633-5. (See http://www.unicode.org/unicode/standard/versions/Unicode3.0.html.)
2. P Daniels, W Bright, *The World's Writing Systems*, ISBN 0-19-507993-0
3. R Gillam, *Unicode Demystified*, ISBN TBD
4. R Snell, S Weightman, *Teach Yourself Hindi*, ISBN 0-340-42464-8
5. Kalra, Purewall, *Teach Yourself Panjabi*, ISBN 0-340-70129-3

Related talks in IUC 22

1. S Zilles, *Internationalised Text Formatting in CSS and XML*
2. S Urs, *Unicode for Encoding Indian Language Databases: A Case Study of Hindi and Kannada Scripts*
3. R Viswanadha, *Transliteration of Indic Scripts Implementation in ICU and Lessons Learned*
4. M Karnati, *Developing Telugu Unicode Fonts: Practical Problems and Possible Solutions*
5. N Sato, *Thai and Hindi Support in Sun's Java 2 Runtime Environment*

Other references

1. C Wissink, *Unicode Technical Note #1*, http://www.unicode.org/notes/tn1/
2. K Zia, *Mapping of National Urdu Standard to Unicode*, Proceedings of 22nd International Unicode Conference, 2002
3. M Dürst, F Yergeau, M Wolf, R Ishida, T Texin, *Character Model for the World Wide Web 1.0*, (See http://www.w3.org/TR/charmod/.)

Acknowledgements

Many thanks to Cathy Wissink, Mark Davis and Joe Becker for reviewing this paper at very short notice and still making numerous useful comments and suggestions.

$Id: Overview.html,v 1.284 2002/08/29 18:09:21 rishida Exp $