Transliteration of Devanāgarī

D. Wujastyk

25 June 1996

Contents

1 History of Writing in India

The earliest writing in India is in a script called Brahmī, which survives in rock inscriptions. From Brahmī, via intermediate scripts, are derived Devanagarī and all the major scripts subsequently used in South Asia, with the exception of the Perso-Arabic script used for Arabic, Persian, and Urdu, and of course the Roman script.

The Devanāgarī alphabet emerged in something like its present form in inscriptions from northern India in the latter part of the first millennium AD. It was used at that time, and since, for the writing of Sanskrit and of a number of related languages and dialects. Today, it is the script normally used for the writing of the languages Hindi, Marathi, and Sanskrit.[1]

The best study specifically of the Devanāgarī script is that of H. M. Lambert, *Introduction to the Devanagari Script for students of Sanskrit, Hindi, Marathi, Gujarati and Bengali* (Oxford, 1953).

2 The phonetic system

The rigorous phonetic system underlying the written form of the Indo-aryan languages was formulated extremely early in India.[2] Several hundred years before the beginning of the Christian era, Indian phoneticians had already analysed the sounds of Sanskrit speech, and had composed elaborate tracts on phonetic theory. The result of this early precision and linguistic awareness applied to phonetics is the extremely regular and phonetically-informed nature of all writing systems derived from Brahmī.

Thus, the Devanagarī script explicitly differentiates voiced and unvoiced consonants, as well as aspirated and unaspirated ones. The script also distinguishes the consonants according to the position of articulation in the mouth. Thus there are five series of letters: velar, palatal, retroflex, dental, and labial. Each series has voiced and unvoiced consonants, and each of these has aspirated and unaspirated forms. That makes twenty phonemes. Added to this there are vowels (short and long), semivowels, and sibilants, again all corresponding to well-defined phonetic qualities. In almost every case, each written letterform represents a distinct phoneme. This is probably unique for any script.

This phonetic regularity makes the issue of transliteration much easier

[1] For maps of script and language distribution in South Asia, and further documentation, see Joseph E. Schwartzberg (ed.), *A Historical Atlas of South Asia* (Chicago and London, 1978), p. 102 *et passim*. For a brief account of the evolution of Indian scripts, see Colin P. Masica, *The Indo-Aryan Languages* (Cambridge, 1991), chapter 6: "Writing Systems".

[2] See W. Sidney Allen, *Phonetics in Ancient India* (London, 1953).

than it might otherwise be. Indeed, European scholars of Sanskrit settled on a transliteration scheme in roman characters over a hundred years ago, and this scheme has been revised very little since that time.

3 Roman transliteration

Transliteration inherently presents multi-valent possibilities, and indeed Indic languages have historically been represented in numerous other scripts, especially Greek, Tibetan, Chinese, and Perso-Arabic. And non-Indic languages have also been represented in Devanāgarī and other Indic scripts, especially Greek, Persian, Arabic, and English. In contemporary India, English words frequently appear transliterated into Indic scripts. There is also a related, though less complex set of issues surrounding the representation of Indian languages in eachother's scripts. The Tamil script, for example, does not contain characters to represent all the phonemes of Sanskrit; Sanskrit manuscripts written in South India historically used an enriched version of the Tamil script, called Grantha.

In what follows, we restrict ourselves to considering the transliteration of the Devanāgarī script into roman characters.

3.1 Early history of standardization

In the introduction to his influential *Sanskrit-English Dictionary*,[3] Sir Monier Monier-Williams devotes a number of pages to the discussion of the scholarly transliteration of Sanskrit in Europe.[4] He outlines the early history of discussions on this subject, and the reader is referred to this work for this history. In particular, a Transliteration Committee was set up at the Geneva Oriental Congress in September 1894, the decisions of which have been broadly adhered to until today.

3.2 Contemporary conventions

The transliteration scheme in normal use in scholarly circles today, deriving directly from the 1894 Geneva committee's recommendations, represents the Devanāgarī script with the following conventions.

Normal roman alphabetic characters are used to represent the nearest similar sounding Devanagari letter. The macron is used to distinguish a vowel as being of long duration (as opposed to short). An underdot marks a consonant as being retroflex. A tilde is used to mark a palatal nasal. An overdot is used to mark a velar nasal. An "h" is appended to a consonant to mark it as being aspirated. An acute accent is used to distinguish the

[3] First edition: Oxford, 1899.

[4] Section IV, pp. xxii ff., and especially pp. xxviii–xxx.

palatal sibilant (from the dental). Vowels in Devanagari may be written as diacritical marks above or even before the consonsants after which they are pronounced. They are transliterated as the nearest-sounding roman vowel letter, and are written in the position in which they are pronounced.

An influential body in the establishment of transliteration standards today is the Library of Congress in Washington D.C. This body has endorsed the transliteration scheme that is in normal widespread use in the scholarly community, with very little alteration.[5]

3.3 Ligatures or conjunct consonants

Both in writing and in printing, the Devanagari alphabet uses an elaborate system of ligatures or "conjunct consonants". In general, any two consonants written together without an intervening vowel are replaced by a more or less elaborate ligature. (In roman typography, only the ligature f+i = fi is at all common today. In Dutch there is also i+j = (approximately) ij.)

For example, if the medial "a" in *papa* पप is removed, the resulting *ppa* is written प्प (horizontal ligature). Similarly, *kaka* कक becomes *kka* क्क (vertical ligature). Less obvious ligatures include *t* त् + *ra* र = *tra* त्र, *j* ज् + *ña* ञ = *jña* ज्ञ, and least obviously of all, *k* क् + *ṣa* ष = *kṣa* क्ष.

However, because the phonetics of the language being represented in the script has always been so clearly and distinctly understood in the Indian intellectual tradition, complexities of the writing system have never led to any complications of pronunciation, and therefore pose no special problem for transliteration. In transliteration from Devanāgarī to Roman script, ligatures are ignored.

4 The script and standard transliteration

With the above phonetic scheme in mind, the Devanagari alphabet may be represented as in Figures 1, 2, and 3.

Visarga and anusvāra represent a rough breathing and a nasalized vowel respectively.[6]

4.1 Some anomalies and variations

The Library of Congress (LC) scheme for transliterating Indic scripts uses an under-circle for vocalic r, ऋ which is given as *ṛ* (underdot, not undercircel) in Figure 3. The sign *r̥* is reserved by the LC for the representation

[5]The "All-India Roman Alphabet" devised in the early 1950s by Prof. J. R. Firth at SOAS, London, and referred to in the preface of Lambert's *Introduction to the Devanagari Script*, has not made the transition to widespread use.

[6]See W. S. Allen, *Phonetics in Ancient India*, pp. 40-46, for further details.

	Voiceless				Voiced					
	Unaspirated		Aspirated		Unaspirated		Aspirated		Nasal	
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

Figure 1: Stops (consonants) and nasals

	Short		Long	
Velar	अ	a	आ	ā
Palatal	इ	i	ई	ī
Retroflex	ऋ	ṛ	ॠ	ṝ
Dental	ऌ	ḷ	ॡ	ḹ
Labial	उ	u	ऊ	ū

Figure 2: Vowels

of a "retroflex flap" which occurs in New Indo-Aryan languages like Hindī.[7] This reflects a difference in usage between Sanskritists on the one hand (who generally use *ṛ*), and librarians and New Indo-Aryan language scholars on the other (who use the under-circle). This distinction of user communities is not a rigid one.

An underbar (kh) is used for the transliteration of some "dotted" Devanāgarī characters which are used to transliterate (in Devanāgarī) Persian, Arabic, and English sounds.[8]

5 Seven-bit transliteration

Now that the conventions for the scholarly transliteration of Devanāgarī have been described, we may turn to a more specialized system of transliteration, namely one restricted to the use of the characters from the ISO 646 character set, more widely known as ASCII.

It is not necessary here to describe the need for such a scheme. Internet protocols for email, for example, normally preclude the display of characters using the diacritical marks displayed in the Figures above.

[7]See R. S. McGregor, *Outline of Hindi Grammar* (Oxford, 1972), p. xviii.

[8]See McGregor, *Outline*, p. xxx.

	Semivowel		Sibilant	
Velar	ह	ha		
Palatal	य	ya	श	śa
Retroflex	र	ra	ष	ṣa
Dental	ल	la	स	sa
Labial	व	va		

"visarga"	ः	ḥ
"anusvāra"	ं	ṃ

Figure 3: Semivowels and sibilants

The community of students and scholars of South Asian studies have in the main adopted two conventions for representing the transliteration of Indic languages: the Velthuis and the Harvard-Kyoto systems. A third system, ITRANS, borrows from the Velthuis system. A fourth system, devised by Prof. Peter Schreiner of Zurich University is used by Schreiner himself and by his colleagues who use the Tübingen TUSTEP system for text processing.

All these systems have been used to encode substantial text corpora.

5.1 Velthuis transliteration

This scheme is named after Frans Velthuis, a scholar living in Groningen, The Netherlands, who has created a popular, high-quality software package for typesetting Devanagari. This package has been used for typesetting the Devanāgarī script used in the present document.

Another package for doing the same job, but with differently designed fonts, has been written by Charles Wikner of South Africa. It uses the same input scheme as Velthuis, but adds some characters for marking Vedic accents.

The Velthuis transliteration may best be represented by the table which forms part of Velthuis's documentation. It is given here as Figure 4.

As can be seen by inspection, the Velthuis scheme is based on the principle of using the ISO 646 repertoire to represent mnemonically the accents used in standard scholarly transliteration. Thus the retroflex consonant ड, commonly transliterated as *ḍa*, is represented in the Velthuis scheme by ".da". In fact, all retroflex sounds are transliterated using the underdot, which appears as a pre-dot in the Velthuis scheme.

a	अ or implicit	ch or C	छ	r	र	f	फ़	
aa or A	आ or ा	j	ज	l	ल	z	ज़	
i	इ or ि	jh or J	झ	L	ळ	.kh or .K	ख़	
ii or I	ई or ी	~n	ञ	v	व	.g	ग़	
u	उ or ु	.t	ट	¨s	श	q	क़	
uu or U	ऊ or ू	.th or .T	ठ	.s	ष	.o	ॐ	OM
.r	ऋ or ृ	.d	ड	s	स	.a	ऽ	avagraha
.R	ॠ or ॄ	.dh or .D	ढ	h	ह	~o	ऑ	English o
.l	ऌ or ॢ	.n	ण	R	ड़	~a	ॅ	English a
.L	ॡ or ॣ	t	त	Rh	ढ़	.m or M	ं	anusvāra
e	ए or े	th or T	थ	1	१	/	ँ	candrabindu
ai or E	ऐ or ै	d	द	2	२	.h or H	ः	visarga
o	ओ or ो	dh or D	ध	3	३	|	।	sentence end
au or O	औ or ौ	n	न	4	४	||	॥	paragraph end
k	क	p	प	5	५	@	॰	abbreviation
kh or K	ख	ph or P	फ	6	६	#	॑	elliptical dot
g	ग	b	ब	7	७	..	.	period
gh or G	घ	bh or B	भ	8	८	~r	ॅ	Marathi r
¨n	ङ	m	म	9	९			
c	च	y	य	0	०			

Figure 4: Velthuis Transliteration

5.2 Harvard-Kyoto transliteration

In describing this scheme some years ago, Prof. H. Nakatani[9] said the following:

> The following scheme may be adopted by those who are engaged upon putting a fairly large amount of Sanskrit textual material into machine readable form. It enables one to input texts with a minimum motion of the fingers on the keyboard. It is also easy to learn: all letters with an underdot are typed as the same letter capitalized; guttural and palatal nasals (ṅ, ñ) as the corresponding capital plosives (G, J); ḷ, l̤, l̤h are quite rare; the only transliteration that needs to be remembered is z for ś. Finally, it is easily readable with a little practice.

[9]Prof. H. NAKATANI,
Kobegakuin University,
Nishi-ku, Kobe 673, JAPAN.
Tel.: 078.974.1551, ext. 2359 (Tuesday afternoon).

A	B	A	B	A	B	A	B	A	B
a		k		t		y		anunāsika	**&**
i		kh		th		r		upadhmānīya	**f**
u		g		d		l		jihvāmūlīya	**x**
ṛ	**R**	gh		dh		v		udātta	**;**
ḷ	**L**	ṅ	**G**	n		ś	**z**	svarita	**:**
ā	**A**	c		p		ṣ	**S**		
ī	**I**	ch		ph		s		external	
ū	**U**	j		b		h		sandhi	**^**
ṝ	**q**	jh		bh		ṃ	**M**	compound	
ḹ	**E**	ñ	**J**	m		ḥ	**H**	junction	**.**
e		ṭ	**T**						
o		ṭh	**Th**						
ai		ḍ	**D**						
au		ḍh	**Dh**						
		ḻ	**W**						
		ḻh	**Wh**						
		ṇ	**N**						

Figure 5: Harvard-Kyoto transliteration

5.3 ITRANS transliteration

ITRANS is a system created by Avinash Chopde for printing various Indic scripts using a PC. The transliteration input scheme for Devanāgarī is similar to the Velthuis system. See Figure 6.

5.4 Schreiner transliteration

Prof. Schreiner's scheme is also based on the representation in ISO 646 characters of the diacritical marks used by scholars. He describes his scheme as follows, in the introduction of one of his machine-readable texts. (Some of this description concerns meta-characteristics of the text which do not impinge on transliteration.)

> Diacritical marks:
> Punctuation marks are used to code diacritics. All diacritics are typed in front of the letter to which they belong (which imitates the traditional "layout" of typewriters where accents etc. are placed on dead keys and need to be typed before the character is typed).
>
> . = subscript dot (e..g. k.rta.h)
> ; = superscript dot (e..g. a;nga)
> ? = tilda (superscript) (e..g. praj?n-a)

- = superscript hyphen (macron) (e..g. -atm-a)
' = aigu (superscript) (e..g. 's-astra)

(Where, as in this introductory document or in comments to variant readings within the text, the use of punctuation marks in their proper function has to alternate with their function as diacritical marks, their use as punctuation marks is distinguished by a following blank (or other signs of punctuation including parentheses) or by doubling where a blank is not possible (e.g. in abbreviations). Thus, the dots in ".r.s.i" are diacritics, but the dot after ".r.si. " is a full stop. Similarly, since no blank can be inserted after a hyphen, the actual hyphen is written by doubling it ("--").

Exclamation mark is used for single quotation mark (!quote!) and apostrophe ("author!s"). In text format it is also used as avagraha (e..g. so !ham, which would be "so{ {aham" in input format).)

The quarter of a verse (p-ada) is marked by a || (vertical bar). This bar (da.n.da) is not followed by a blank before verse quarters 2 and 4 (b and d) in anu.s.tubh metre ('sloka). After quarter 2 and 4 a new line begins.

[...]

Lacunae are indicated by using the letter x one time per missing ak.sara; the x–es are put in pointed parentheses (e..g. ⟨xxx⟩ for a lacuna of three syllables).

6 Other issues

6.1 Word boundaries and hyphenation

Writers of Sanskrit in Devanāgarī commonly do not mark word boundaries with a space except when certain grammatical operations related to *sandhi* apply. However, the convention used by writers of Sanskrit in roman transliteration is to mark word boundaries with spaces. This difference is not reversible by any simple algorithm or search-and-replace operation.

In ancient and medieval manuscripts, Devanāgarī is not hyphenated. At the end of the line, the scribe would simply break off writing anywhere in a word, and continue on the next line. This practice was imitated in the very earliest Devanāgarī printing and lithography, but soon the hyphen began to be used in printed books. However, the simple rule of hyphenation generally applied was – as in manuscripts – to break a word at the end of any syllable. This rule has been implemented in some modern Devanāgarī printing packages, and is satisfactory.

Again, in transliteration, writers do not follow the above hyphenation rule, but hyphenate much more self-consiously, with hyphens being dissal-

lowed except at etymologically acceptable points within words or between the words of compounds.

6.2 Case

Clearly the Harvard-Kyoto system relies on the case of roman letters to carry meaning about the characters being transliterated. One of the options for within the Velthuis system also uses case significantly.

If case carries transliteration information, then it cannot simultaneously be used for marking proper names, sentence beginnings, and the other uses that are normal within European usage of the Roman script.

If such usage is required, a system which does not invest case differences with transliterational significance must be used.

6.3 Ambiguity and language tagging

Take the following example. The Devanāgarī letter अ is transliterated as "a" in all systems. How is this "a" to be distinguished from an English "a"?

In mixed-language texts, it becomes important to tag all transliterated strings by language, if further processing on these strings is envisaged. A search-and-replace operation, for example, might be required for all Hindī strings, but not for English words. Only language-tagging (or perhaps script-tagging) will resolve the ambiguity of meaning of transliterated characters, and make this possible.

Not all transliterated characters are ambiguous. For example, ट *ṭa* may be 7-bit transliterated as ".ta" There is (probably) no normal English string containing this combination of characters, so a language-sensitive search-and-replace operation could be successful in such a case, even if no language tagging is present.

Language tagging is required if automatic retransliteration is envisaged, as in the use of the widely-distributed Indian GIST card, a system for the entry, display, and printing of Indian and roman transliterated scripts on IBM compatible personal computers.

Figure 6: