

Roman Transliteration of Indic Scripts

Kavi Narayana Murthy
Department of Computer and
Info.Sciences
University of Hyderabad
email: knmuh@yahoo.com

Srinivasu Badugu
Department of Computer and Info.
Sciences
University of Hyderabad
srinivasucse@gmail.com

Abstract

In this paper we analyze the need for Roman Transliteration for Indic scripts. We evaluate the pros and cons of various schemes in use today and argue for a scientifically designed standard scheme. We offer one such scheme for the consideration of all the experts. We believe the ideas we have presented here will also be of interest to people in many other countries where the language situation is similar.

1. Translation and Transliteration:

Language is all about systematization of mappings from sound patterns to meanings, through which we can think and also communicate our thoughts and feelings to others. A script is a systematization of rendering language in the written form. The relation between language and script is not one-to-one: a given language can be written in any script and a given script can be used for writing any language, although we may have usual preferences.

In translation, texts from one language are mapped to texts in a different language. This is intended to be meaning preserving transformation. The words used, the structure of words and sentences, etc. could change. The script used for rendering these texts is not important. The source language text as well as the target language text can be in any suitable script. They could even be in the same script. For example

raam ne siitaa ko kittab dee diya

is a Hindi sentence rendered in Roman script, while its translation in Kannada could be

raamanu siitege pustakavannu koTTubiTTanu

Transliteration is not translation, here the language does not change, only the script used to render the text is changed. Why should we even think of rendering one language in some script other than the default script? People may know a language but not the script - rendering the language in a script which they know

can help them to read and understand the texts. For example, if you know Hindi or Kannada language but you do not know the scripts, you can read and understand the Hindi and Kannada sentences above if you know the Roman script and the conventions used therein. If you do not know the language also, you can still read these sentences since you know the Roman script. The main goal of transliteration is to enable the reader to read, that is, pronounce the words as accurately as practically possible. Pronunciation is important, orthography is not the basis at all. After all, the most important thing about a word or sentence is its meaning and the next most important thing is its pronunciation. Transliteration has several other benefits as we shall see soon.

2. Romanization:

It is possible to think of transliteration from any script to any other. Here we shall focus mainly on the Indic script to Roman transliteration. Here we write texts in Indian languages using the Roman alphabet (which is the same as what we use for writing English. This whole article is in Roman script.) This process is known as Romanization. This can be bidirectional and if so, we can actually map from any script to any other script via Roman. The ideas presented in this paper are generic and applicable to other language and script scenarios anywhere in the world.

Romanization has several advantages. 1) To render any language in any fonts and a suitable rendering engine. If we need to type in these texts, we also need a keyboard driver. These may not always be available. Roman script is universally available on all computers. Even when all the required tools are available, typing in Indic scripts is still complicated and error prone. Typing in Roman could be much faster and less error prone. 2) Readers may know the language but not the script. Romanization helps. 3) When several different languages and scripts need to be used together, there will be frequent need for shifting from one system to the other. This can be cumbersome and error prone. Uniform use of Roman makes it simpler. 4) Typing in Roman can be simpler and faster than typing in their own script for many people. 5) Processing of texts

rendered in different scripts may require different techniques for dealing with idiosyncrasies if any. Rendering all texts in a uniform notation such as Roman can mitigate this problem. Computers have originated and evolved under the Roman script. Roman is the most natural and direct. It can be simpler and efficient too, especially from programming point of view. 6) Languages which do not have a script of their own can also be rendered in Roman. 7) When the language and script used is English but we wish to include bits of local languages, such as in English newspapers, notices, brochures, sign boards etc., Roman is natural. 8) All existing software would simply run on Roman rendered data. If data is encoded in any other non-standard form (such as in a TTF font), standard software may or may not work. 9) Rendering in Roman makes it readable by people speaking various languages whereas rendering the same in local language script restricts. Commercial advertisements such as product information and cinema posters would gain by rendering in Roman.

This whole article is written in the Roman script. The whole paper is basically written in English language and the usual spelling rules of English are sufficient to read these parts. However, we include examples from Indian languages and in order to be able to read them correctly and to understand and appreciate the issues raised and solutions proposed, you will need to know the conventions we are using in this paper. We have described the Romanization conventions we have followed in this paper in the appendix. Readers may please go through these conventions before reading on.

3. Need for Standards:

There are no well established or agreed standards for rendering Indic scripts in Roman. People use a wide variety of arbitrary, illogical, unscientific, highly confusing mappings, mostly driven by intuitions based on English spelling rules, making it difficult for readers to read. We may tend to think that since we are using a script that is used for English, we can as well use the English way of rendering Indic script. This would be the worst choice.

English writing system is alphabetic, we need to learn, store and use spelling rules to map spoken words to written form. These spelling rules are quite arbitrary and unscientific. English uses odd ways of writing long and short vowels (cut-caught, fit-feat, pull-pool, let-late, colour-collor, floatation-float), uses different spellings for the same sounds (meat-meet, their-there, too-two, are-or, sun-son etc.), same letters for different sounds (put-but, fat-fate, fit-fight, poll-pool, peg-page, pill-philosophy, case-chase, etc.), strange spellings (daughter, school, women, etc.),

silent letters (psychology, coup, etc.). Therefore, using English spellings as a basis for Romanization would only add to confusions. More importantly, it is not right to assume that all users would know English. We can actually develop a simpler, more uniform, more scientific, easy to learn and use system of transliteration.

English does not have a simple and uniform way of depicting long and short vowels. English uses 'oo' and 'ee' to indicate long vowels but 'aa', 'ii', 'uu' etc. are not found in English. Further, 'oo' is used for the 'uu' sound, not as a long counterpart of the 'o' sound. Similarly, 'ee' is used for the 'ii' sound, not for the long counterpart of 'e' sound. Therefore, mere intuition is not sufficient to render long and short vowels systematically using the conventions of English language. Further, there is no natural way in English to show the difference between 'd' and 'D' sounds, as also 't' and 'T' sounds. We need to have a systematic way of handling 't', 'th', 'd', 'dh', 'T', 'Th', 'D' and 'Dh' sounds or there will be too many confusions. English uses 'ch' for the 'c' sound and so intuition tells us that we should use 'chh' for the 'ch' sound but then we are not using the second letter 'h' consistently. The anusvaara ('M') is pronounced differently in different contexts and people use 'm' or 'n' arbitrarily. Since we have more than 26 basic varNa-s or phonemes in our chart, either we must use multi-letter mappings or we must use upper case and lower case letters differentially or we must use special characters beyond 'a-z' or two or more of the above. Special characters may occur in their own right in texts, leading to ambiguity. Also, people often fail to realize that the whole purpose of Romanization could be to help readers of other languages. They do not know or they do not care about other languages and use mappings which are very confusing to the readers. Since a given word may involve several of these situations, one word can actually be interpreted in a large number of ways and the readers will be left to guess. Thus, English spellings and our intuitions about them cannot be the right basis for Romanization.

Romanization has become a basic necessity in India. It is therefore imperative that we develop and use standards. This would make everybody's life so much better and we can avoid a whole lot of avoidable confusions and errors. The situation is similar in many other countries too.

It is not that there are no standards at all - several proposals exist and are being followed by various groups in pockets. There are many such proposals and suggestions around and it is time we take a careful look at all of them and come out with national/international standards that can be used by all. Even the

national/international standards sometimes have problems and issues. For example, Unicode is based on the ISCII standard but it does not seem to understand ISCII fully. In ISCII, two consonants in sequence will not automatically join to form a consonant cluster. We have a special symbol called 'halaMt' to form consonant clusters. 'halaMt' removes the implicit vowel 'a' in a consonant. In case we need to show two consonants in sequence without any vowel in between, and we wish to depict the first consonant as a pure consonant without any implied vowel in it, ISCII standard requires the use of two 'halaMt-s'. Perhaps with the idea of providing a generic solution to all the languages of the world with similar properties, Unicode has introduced the concepts of zero width joiner and zero-width non-joiner. Since 'halaMt' continues to exist, now we have more than one way of effecting the same thing, leading to ambiguities.

4. General Considerations

Indian scripts are highly phonetic in nature. That is, the writing depicts the pronounced phonemes quite closely and fairly accurately. Correct pronunciation is the right basis for transliteration of Indian scripts. Units of writing correspond to phonemes, not to narrow phonetic realization variations. Hence coding should be based on phonemic considerations, not the phonetic in a narrow sense. Allophones should not be coded.

The Roman alphabet has only 26 letters whereas many Indic scripts have more than 50 basic units and allographs as well. Therefore, we need to 1) use upper case and lower case letters to represent different units, 2) use two or more Roman letters per Indic script unit 3) and/or use other special symbols. Mixing of upper and lower case letters would not work in case-insensitive systems. If we try to avoid case mixing, we will find it hard to develop a simple, natural, readable scheme unless we use special characters.

Combining two or more letters is based on the assumption that those combinations never occur otherwise. This may be true of the language but a script can be used to write not only this language but other languages as well. For example, we have used 'kh', 'ph' etc. for aspirated sounds knowing that these could also mean a cluster of 'k' or 'p' and 'h' consonants. These latter possibilities are exceptionally rare in our native languages. However, in the rare case that we actually wish to depict the consonant cluster formed out of, say, 'k' and 'h', we can use an escape mechanism. The backslash symbol is widely used in computer science to form escape sequences and we could resort to the same or similar devices.

Capitalizing proper names and the first letter of the first word of a sentence are conversions of English language, these do not hold in Indian languages and we should be careful not to capitalize for these purposes.

If we need to use these notations on computers, it is better to use only those symbols which are readily available on the standard keyboard. Other special characters would be difficult to type. Also, simple, linear arrangements are easier and better compared to diacritic marks which may appear above or below a letter.

If the rendered texts need to be stored and processed by the computer, it is better to use one-to-one mappings where each phoneme is mapped to one single letter. However, this may force us to use mixture of cases and/or special symbols.

It is better to avoid the use of symbols which can appear naturally in texts. For example, symbols such as the colon, quote mark, hyphen, star and period should be avoided as they can occur in texts with different meanings.

Encoding should be preferably unique. This makes it possible to revert back to the original representation without loss or distortion. It is obvious that if the text includes Indian languages as well as English, we can Romanize but we cannot go back to the original. If we try, the original English texts will also be rendered in the chosen Indic script, unless we have a way of marking up English and Indic sections explicitly. If the original text does not include Roman letters, it should be possible to do round-trip conversion.

Most Indian scripts, with the exception of the few Perso-Arabic scripts, have a common origin in the ancient braahmi script, they all follow the same fundamental principles and conventions, they are largely common too. All are phonemic representations. Variations in the phoneme sets are minimal, although the visual appearance may vary widely. It would therefore be better to have a common or a largely common scheme for all Indic scripts rather than completely different schemes for each language. If a small super-set is built by taking into consideration of all the Indic scripts, we would then be able to map from any script to any script via Roman.

Tamil is one major exception - Tamil script has a much smaller number of units. Naturally, what a Tamilian would find natural would sound very strange for non-Tamilians and vice versa. For example, Tamilians tend to write dha for da, k for g, g for h, etc, causing major confusions. We should remember that the whole purpose of Romanization is to make it easy for

speakers of other languages to read our language. We should therefore take their views and requirements as more important rather than imposing our views and expectations on others.

Some languages including Sanskrit and Hindi have only long versions of 'ee' and 'oo' vowel sounds, there are no short counterparts. These are almost always pronounced as long vowels, they are written and pronounced as long vowels in other languages/scripts and must therefore be rendered as 'ee' and 'oo' respectively. Therefore, the correct rendering would be 'veeda' and 'yooga', not 'veda' or 'yoga'.

But some languages, notably Hindi, have the convention of dropping the last vowel in a word in pronunciation but we must remember that we are not doing a phonetic transcription. Here, we only need to ensure that what is written will be read accurately by speakers of other languages. If you want readers to read it as 'yoog', render it as such, not as 'yooga'. If you write 'yooga', the readers will read the final vowel after the 'g' sound. But if you are interested in rendering the text in as natural a way as possible in the target language/script, then you must render this word as 'yooga'. In any case, there must be a long vowel 'oo' in the standard.

Although mixing upper case and lower case letters is generally considered essential, this idea should not be carried too far. Too many upper case letters mixed up with lower case letters makes the text look ugly and typing becomes more difficult and error prone due to the need for frequent use of the shift keys.

While respecting the above considerations, shorter codes would be preferred as this would save storage space.

Most importantly, the rendered text should be natural, easy to read and type in. It should be easy to learn and easy to remember. It should not lead to an increase in typing errors. After all, computers are at our service, not the other way round. Making things difficult and unnatural to human users in the name of making things easier for the computer is no good. Some proposed schemes are especially notorious in this respect. For example, 'kRShNa' is written as 'kqRNa', 'kRti' is written as 'kqw', 'jnYaana' is written as 'jFAna', 'ceppaaru' is written as 'ceVppAru', 'rudruDu' is written as 'ruxrudu', 'peeru' is written as 'peru', 'RShi' is written as 'qRi', 'kaLa' is written as 'kaLYa' in the so called w-x scheme, making it highly unreadable.

5. Alternatives and Issues

We have long and short vowels in Indian languages and we need a simple and uniform way of depicting this difference. One choice is to use double letters. We do use double vowels to show long vowels in English too (cool, feet, etc.) but the point is the short form of a long vowel is not the same vowel used once (*cul, *fet). We should use a simple rule uniformly everywhere.

Some people use capital letters for long vowels. Capitalization is totally absent in Indic scripts and it carries no connotation at all. Even in English, capital letters are only indicative or proper names and beginnings of sentences, they have nothing to do with pronunciation. The letters 'a' and 'A' are pronounced the same way in English. Therefore, using capital letters for long vowels would be worst choice.

In many Indian languages, we do not have the lowered vowel as in the English word 'cat'. The letter 'a' is to be pronounced as in 'but'. The long counterpart of this vowel will be rendered as 'aa'.

We have two vowel sounds as in 'lit' and 'let'. We must consistently use these letters (i and e) for these two vowel sounds respectively. The long counterparts should be 'ii' and 'ee' respectively.

We can use 'u' for the vowel sound as in put. The long counterpart should then be 'uu'.

We can use the letter 'o' for the first vowel sound as in 'potato'. The long counterpart of this should then be 'oo'.

The vowel sound in 'kite', 'height' etc. begins with a 'a' sound and end in a 'i' sound and can so be rendered as 'ai'. There is no 'ei' sound in many Indian languages, although the written form may look closer to this. Likewise, we have 'au', not 'ou'.

We can reserve 'k' for the consonant sound as in 'king' and use the 'c' letter consistently for the 'ch' sound as in 'chase'.

We reserve 'g' for the sound as in 'get' and consistently use 'j' for the sound as in 'jar'.

We have phonemic status for aspirated sounds and our scripts depict this directly. We can use the widely used idea of adding a 'h' to indicate aspirated counterparts. Here the 'h' is indicative or a feature called aspiration. We are not uniformly depicting all features explicitly. Yet this is a good choice keeping readability in mind. Using capital letters for aspirated sounds is highly unintuitive and unreadable.

We also have retroflex consonants (such as the tea versus thee) and the widely used convention is to use capital letters for these. Examples are T, D, N, and L. Capital letters have no phonetic connotations and we should avoid their use as far as possible. Since it becomes impossible to map all the phonemes in our languages without using capital letters, we may be forced to use them in some cases like these with well documented conventions about their pronunciations.

The anusvaara and visarga are pronounced differently based on the context. These can be rendered as 'M' and 'H' respectively without worrying about their actual pronunciations since pronunciations are fully rule governed and here the concern is only to render written texts as they are, not their actual pronunciations.

The letters 'f', 'z', 'q', 'x' have no phonetic equivalents in many of our scripts and should therefore be avoided. There is no real difference in pronunciation between 'v' and 'w' from our language point of view and hence 'w' should be avoided and the more direct 'v' should be preferred.

Keeping English in the back of the mind, some people propose that 'oo' should also be allowed for 'uu', 'ee' should also be allowed for 'ii', 'w' for 'v', etc. These will make the schemes non-uniform and irreversible. If we develop widely accepted standards, we can also develop standard converters to automatically map between different languages, scripts and character encoding standards.

These ideas are generally accepted but there are a number of significant differences between various schemes proposed. In some of these schemes, a given idea is not used consistently for a given purpose. For example, capitalization may be used for indicating length of vowels, aspiration in consonants, retroflex consonants, etc. Let us take a quick look at some of the schemes in use today.

6. Various Schemes in use today:

Table 1 Part-A

UH	RIT	I TRA NS	WX	IT3	ADH	PH
a	a	a	a	a	a	a
aa	aa/ A/ a'	aa/ A	A	aa	aa/ A	aa /A
i	i	i	i	i	i	i
ii	ii/ ee/ I	ii/ I	i	ii/ I	ii/ ee/ ee'	ii /ee

	ia/ i'/ I				I	/I
u	u	u	u	u	u	u
uu	uu/ oo/ U/ ua/ u'	U/ uu	U	oo/ uu	oo/ uu	U /oo
R	R/ r'	RRi/ R^i	q	rx	-	zr
RR	Ru/ r'U	RRI/ R^I	Q	rx-	-	zR
-	-	LLi/ L^i	-	-	-	zl
-	-	LLI/ L^I	-	-	-	zL
e	e	-	eV	e	e	e
ee	E/ ea/ ae	e	e	ei	E	E
ai	ai	ai	E	ai	ai	ai
o	o	-	ov	o	o	o
oo	O/ oe/ oa	O	o	oo	O	O
au	au/ ou	au	O	au	au/ oq	au
M	-	-	z	-	-	AO
M	@ M	aM	M	an'	-	M
H	@ H	aH	H	a:	-	H
k	k	k	ka	k	k/ g/ h	k/ K
kh	Kh/ kh/ K	Kh	Ka	Kh a	-	kh/ Kh
g	g	g	ga	ga	-	g/G
gh	Gh/ G/ gh	gh	Ga	gha	-	Gh/ gh
nG	~m	~N	fa	ng~	n/ ng/ nG	NG
c	c/ ch	ch	ca	cha	ch/C	ch
ch	C/ Ch	chh	Ca	c hha	-	Ch
j	j	j	ja	ja	j	j/J
jh	jh/ J/ Jh	jh	Ja	jha	-	Jh/jh

nY	~n	~n/ JN	Fa	nj-	ny/ n^/ Ny	NY
T	T/ t'	T	ta	t'a	t/ d	T
Th	Th/ th'	Th	Ta	t'ha	-	Th
D	D/ d'	D	da	d'a	-	d
Dh	Dh/ dh'	Dh	Da	d 'ha	-	dh
N	N/ nh	N	Na	nd-	N	N
t	t	t	wa	ta	th/ dh	t
th	th	th	Wa	tha	-	th
d	d	d	xa	da	-	D
dh	dh	dh	Xa	dha	-	Dh
n	n	n	na	na	n/ N/ n^	n
p	p	p	pa	pa	p/ b	p
ph	ph/ f/ Ph	ph	Pa	pha	-	ph/f
b	b	b	ba	ba	-	b/B
bh	bh/ B	bh	Ba	bha	-	bh/Bh
m	m	M	ma	ma	m	m
y	y	y	ya	ya	y	y
r	r	r	ra	ra	r	r
l	l	l	la	la	l	l
v	v/ w	v	va	va	v	v/w
sh	S/ s'	sh	Sa	sh	-	sh
Sh	sh	Sh	Ra	s hha	S	Sh
s	s	s	sa	sa	sh	s
h	h/ H	ha	ha	ha	h	h
L	L/ Lh/ lh	-	IYa	l'a	-	L
kSh	X	x/ ksh	-	qs- a	x/ ksh	-
-	r`	-	rYa	r'	-	R

Table 1 Part-B

BH	LK	KH	SLP	VH
a	a	a	a	a

A/ aa	aa/ A	A	A	aa
i	i	i	i	i
I/ ee	ii/ ee/ I	I	I	ii
u	u	u	u	u
U/ oo	uu/ oo/ U	U	U	uu
Ru	R	R	f	.r
RU	Ru	RR	F	.rr
-	-	L	x	-.l
-	-	-	-	-
e	e	-	-	-
E	E/ ea	e	e	e
ai	ai	ai	E	ai
o	o	-	-	-
O	O/ oe	o	o	o
au/ ou	ou/ au	au	O	au
~M	-	M	M	.m
M	aM	M	~	“m
H	a@ h	H	H	.h
k	k	k	k	k
K/ kh	Kh/ K	kh	K	kh
g	g	g	g	g
G/ gh	G/ Gh	gh	G	gh
~g	~m	G	N	“n
c/ ch	c/ ch	c	c	c
Ch/ C	C/ Ch	ch	C	ch
j	j	j	j	j
jh/ J	J	jh	J	jh
~j	~n	J	Y	~n
T	T	T	w	.t
Th	Th	Th	W	.th
D	D	D	q	.d
Dh	Dh	Dh	Q	.dh
N	N	N	R	.n
t	t	t	t	t
th	th	th	T	th

d	d	d	d	d
dh	dh	dh	D	dh
n	n	n	n	n
p	p	p	p	p
ph	Ph/ f/ P	ph	P	ph
b	b	b	b	b
B/ bh	Bh/ bh	bh	B	bh
m	m	m	m	m
Y/ y	y	y	y	y
R	r	r	r	r
L	l	l	l	l
v/ w	v/ w	v	v	v
S/ sh	S	z	S	“s
Sh	sh	S	z	.s
S	s	s	s	s
h/ ~h	h	h	h	h
L	L/ lh	-	-	-
-	x/ ksh	-	-	-
Rx	~r	-	-	-

UH- LERC-UoH (See Appendix)
 ITRANS- Indian language transliteration [1, 2, 8, 9]
 RIT- Rice Transliteration [2, 3, 4]
 IT3- OM Mapping [5]
 ADH- Adhawan Tamil [2]
 W-X [9]
 BH- Baraha [12]
 LH- Lekhini [11]
 K-H- Kyoto-Harvard [7, 8, 9]
 PH- Phonetic Transliteration of Indian languages [10]
 VH- Velthuis' devnag [6, 9]
 SLP- Sanskrit Library Project [9]

7. Conclusions

In this paper we have discussed the need, alternatives and pros and cons for various alternatives we have for rendering Indian language texts in Roman notation. We have illustrated the problems with various schemes in use and suggested a scheme that we think is much better. We have been using our scheme effectively for many years now. Hope this paper will lead to further deliberations leading to widely accepted national and/or international standards.

References:

- [1] <http://www.aczoom.com/itrans/>
- [2] http://acharya.iitm.ac.in/multi_sys/ilttable.php
- [3] <http://rksanka.tripod.com/telugu/rts.html>
- [4] <http://www.learningtelugu.org/the-rice-transliteration-standard-for-telugu.html>
- [5] <http://www.cs.cmu.edu/~madhavi/Om/intro.html>
- [6] <http://sanskrit.inria.fr/faq.html>
- [7] <http://www.krishnamurthys.com/profvk/KHC.html>
- [8] http://www.sanskrit-sanscrito.com.ar/en/sanskrit_sanskrit3/transliterating2.shtml#ITRANS
- [9] <http://tdil-dc.in/san/>
- [10] <http://phtranslator.sourceforge.net/>
- [11] <http://lekhini.org/>
- [12] http://www.baraha.com/help/Keyboards/phonetic_keyboard.htm

Appendix: LERC-UoH Romanization Conventions for Indic Scripts:

The transliteration scheme used in this paper is explained here so that readers can pronounce these terms reasonably accurately irrespective of their language background.

In Sanskrit (and other modern Indian languages), the script is a more or less direct rendering of phonemes, the basic sound units in the language, unlike English which uses a set of alphabets and a fairly complex and involved system of spelling rules to map the written alphabet sequences to pronunciations. We give below the 'varNa-maala' or the chart showing the traditional arrangement of phonemes. Most modern Indian languages use a 'varNa-maala' similar to this one. Those who know Sanskrit or any other modern Indian language or other similar world languages will be able to quickly get a feel for the transliteration scheme by taking a look at this chart.

a aa i ii u uu R RR e ee ai o oo au M H

k kh g gh nG
 c ch j jh nY
 T Th D Dh N
 t th d dh n
 p ph b bh m

y r l v sh Sh s h L

Here the first line lists the vowel sounds. Note the correct use of i and e. The second part consists of 25 consonants arranged in a 5 by 5 matrix. The rows indicate the place of articulation. The sounds in the first row are produced at the glottis, at the back of the mouth. Moving gradually from back to front, we reach the last row where we have the sounds produced by closing the lips. The retroflex sounds T, Th, D, Dh, N are produced by lifting and curling up the tongue. The first column includes unvoiced unaspirated consonants, the second column has the unvoiced aspirated sounds, the third column has the voiced unaspirated sounds, the fourth column has the voiced aspirated sounds and the last column has the nasal consonants. Aspiration, an extra puff of air, is indicated by the addition of the letter h. The last line includes the remaining consonants including the semi-vowels, the laterals and the fricatives. We give below the pronunciation guide:

VOWELS:

a	son
aa	master
i	if
ii	feel
u	full
uu	fool
R	(Between ru and ri)
RR	(Long R)
e	let
ee	late
ai	lie
o	rotate
oo	rote
au	now
M	(Pronounced as the corresponding nasal consonant depending upon the following consonant. See notes below.)
H	(Pronounced as ha/hi/hu/he depending upon the previous vowel. See notes below)

CONSONANTS:

k	cart
kh	blockhead
g	goat
gh	ghost
nG	sing
c	chain
ch	catch-him
j	jet
jh	hedgehog
nY	(French n - sounds like nya)

T	ten
Th	ant-hill
D	dog
Dh	godhood
N	under
t	(French t)
th	thumb
d	then
dh	breathe
n	not
p	pot
ph	loop-hole
b	ball
bh	abhor
m	mother
y	yard
r	run
l	luck
v	avert
sh	(German reich - sounds like Sh)
Sh	show
s	son
h	hot
L	(Retroflex l)

M is pronounced as the nasal sound corresponding to the row of the consonant that follows it in the given word. Thus Mk is pronounced as nGk, Mc is pronounced as nYc, MT is pronounced as NT, Md is pronounced as nd, Mb is pronounced as mb. For the unclassified consonants, M is pronounced as m. H is pronounced as ha when preceded by the vowel a or aa, as hi when preceded by i, ii or ai, as hu when preceded by u, uu or au, as as he when preceded by e or ee.

Observe that there are more sounds than the 26 alphabets we have in the Roman script. We therefore use a combination of lower case and upper case letters. Running texts in English consist mostly of lower case letters. Upper case letters are used in acronyms, to start Proper names, and to start a new sentence. These conventions apply only to English, not to Sanskrit and other Indian languages. We shall use uppercase letters mainly to indicate what are called retroflex consonants. Also note that we have a very systematic use of short and long vowels. English is not so systematic in its spelling rules. We cannot depend on the usual spelling rules of English to render terms in Sanskrit, the English alphabet and spelling system is simply not adequate. We use here double letters systematically to indicate long forms of the corresponding short vowels. Sanskrit does not have a short 'e' or a short 'o' sound and many transliteration systems use 'e' and 'o' to indicate the

long forms. This can be very confusing to people who are not familiar with Sanskrit. Many modern Indian languages have long and short versions. Many languages of the world have long and short versions of these vowels. English has too, as in 'get-gate', 'let-late', 'quotation-quote', 'floatation-float'. Note that the scheme used here uses at most two characters per phoneme. Only alphabetic characters are used, no special symbols are used. Words can be formed by simply arranging these letters in a linear sequence as in English, without need for any diacritic marks which are usually marked on the top or bottom of the main characters. This scheme is therefore amenable for easy typing from the standard computer keyboard. The mapping from written words to the corresponding pronunciations will be simple and natural and fully rule governed, making it highly readable. This scheme is easy to learn, and once learned, easy to remember, easy to type-in and easy to read aloud. It can be used very effectively to render Sanskrit and modern Indian languages systematically and fairly accurately.

Note that 'ph' is not to be pronounced as 'f'. The pronunciation of 's', 'sh' and 'Sh' vary across Indian languages. We have used 'sh' and 'Sh', although these are not exactly aspirated counterparts of s. Note that ksha, tra, jnYa etc. are not phonemes, they should not be coded separately.

This scheme is well suited for Dravidian languages such as Kannada and Telugu, and can be easily extended for other Indian languages with minimal additions.

We give some examples below:

computer	'kaMpyuuTar'
Kannada	'kannaDa'
India	'iMDiya'
car	'kaar'
bus	'bas'
train	'Train'

Note that this scheme does not consider foreign sounds. It is quite common to write English and other foreign words in local scripts but when written, they get mapped to the nearest available codes. Thus 'f' becomes 'ph', 'z' becomes 'j', lowered 'a' becomes normal 'a' or 'ya' and so on.