

Localisation in South Asia

Pat Hall, Open University, UK

Venkatesh Hariharan, IndLinux, India

Tapan Parikh, Ekgaon Technologies, India

Durgesh Rao, DR Systems, India

LRC 2002, Dublin 12-13 November



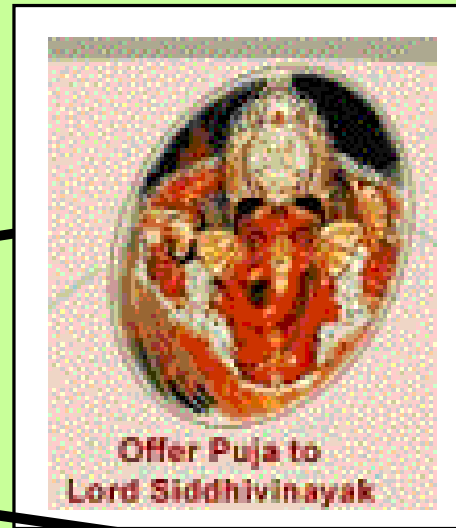
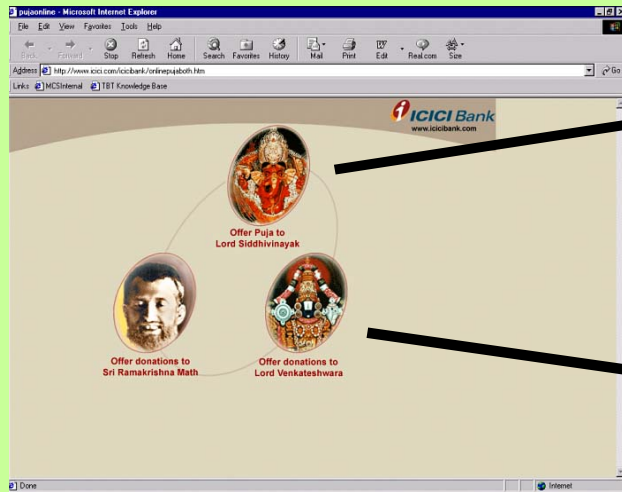
how come we are here?

Sharing Capability in Localisation and Human Language Technologies

funded by Asia IT&C within EU aid budgets www.asia-itc.org

- Open University, UK
- Lancaster University, UK
- European Language Resources Association, France
- National Centre for Software Technology, Mumbai.
- Indian Statistical Institute, Calcutta
- Conferences between Europe and South Asia
- Helping development projects use localised IT systems.

a small motivating example



a page on earlier version of a ICICI Bank web-site



outline

- **Background about South Asia**
 - languages, writing, culture, ...
- **Localisation issues**
 - encoding, rendering, input, formats, calendars
 - translation technology
 - speech
- **Localisation industry and opportunities**
 - available software
 - market
- **What an opportunity!**



Background

- **countries**
 - SAARC nations
- **populations**
 - well over a billion people
 - equal to China
- **languages**
 - more varied and numerous than in Europe
 - ancient writing systems



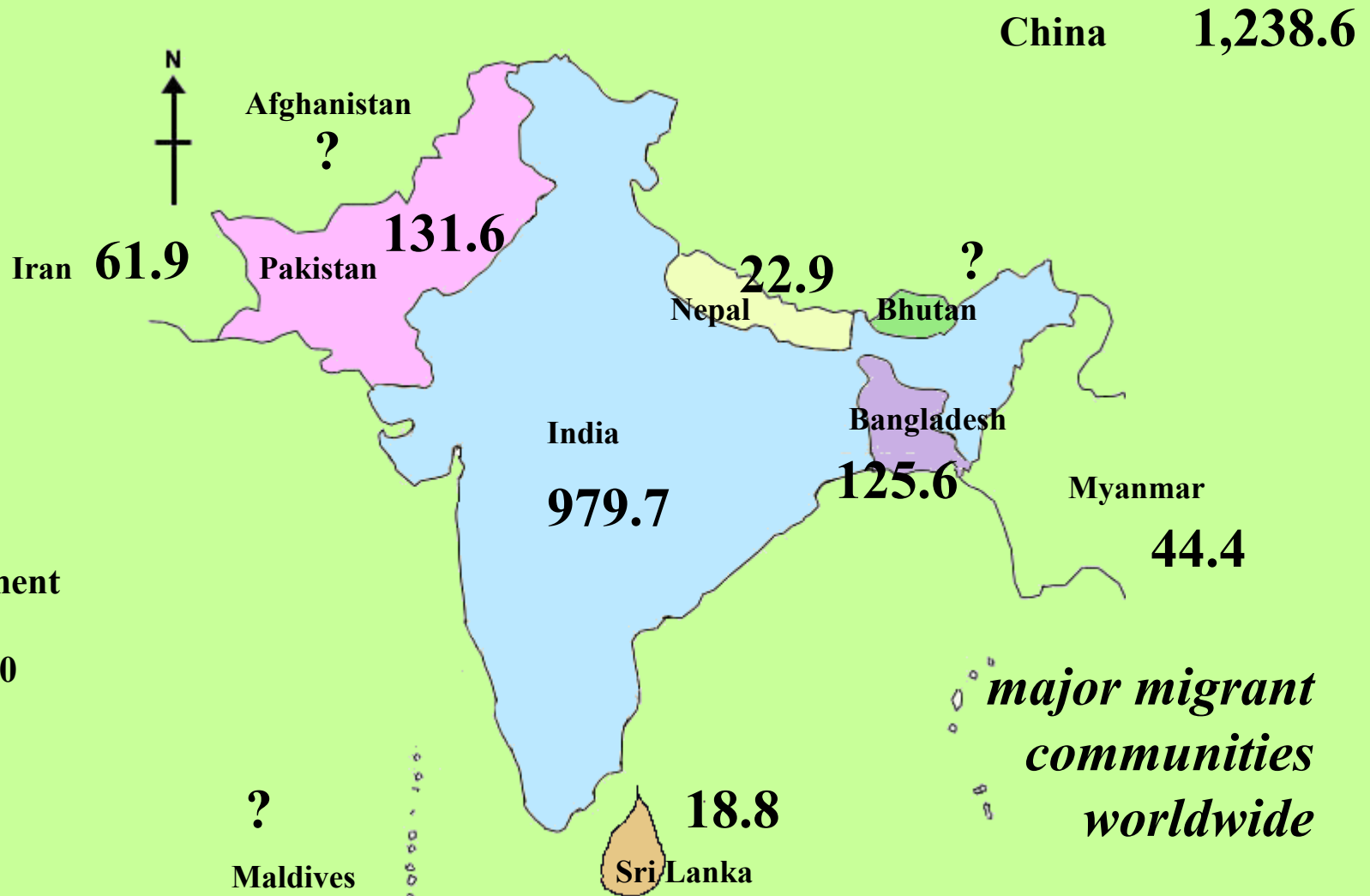
SAARC nations



cities



populations - 1998 millions

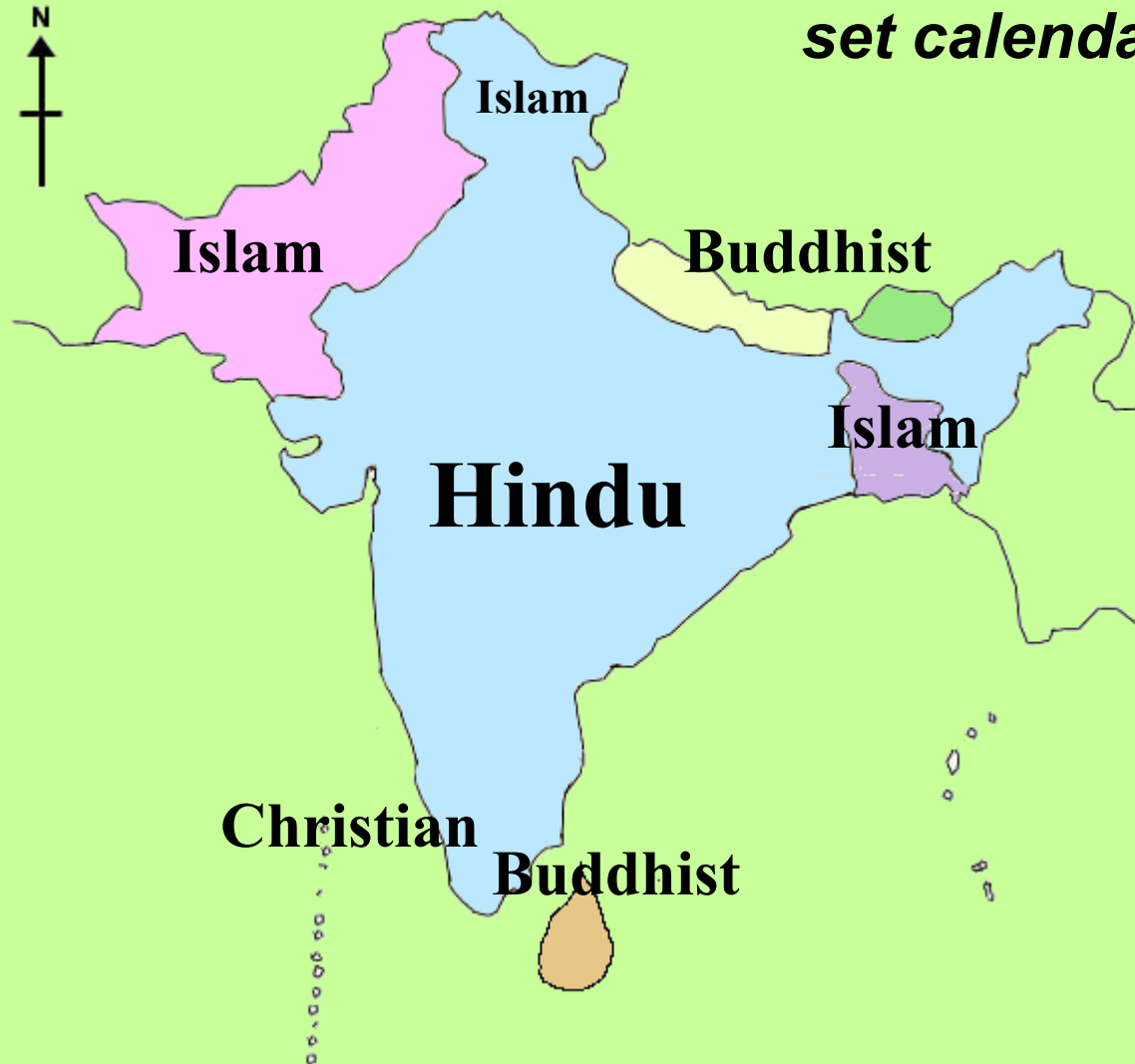


world
development
report
1999/2000

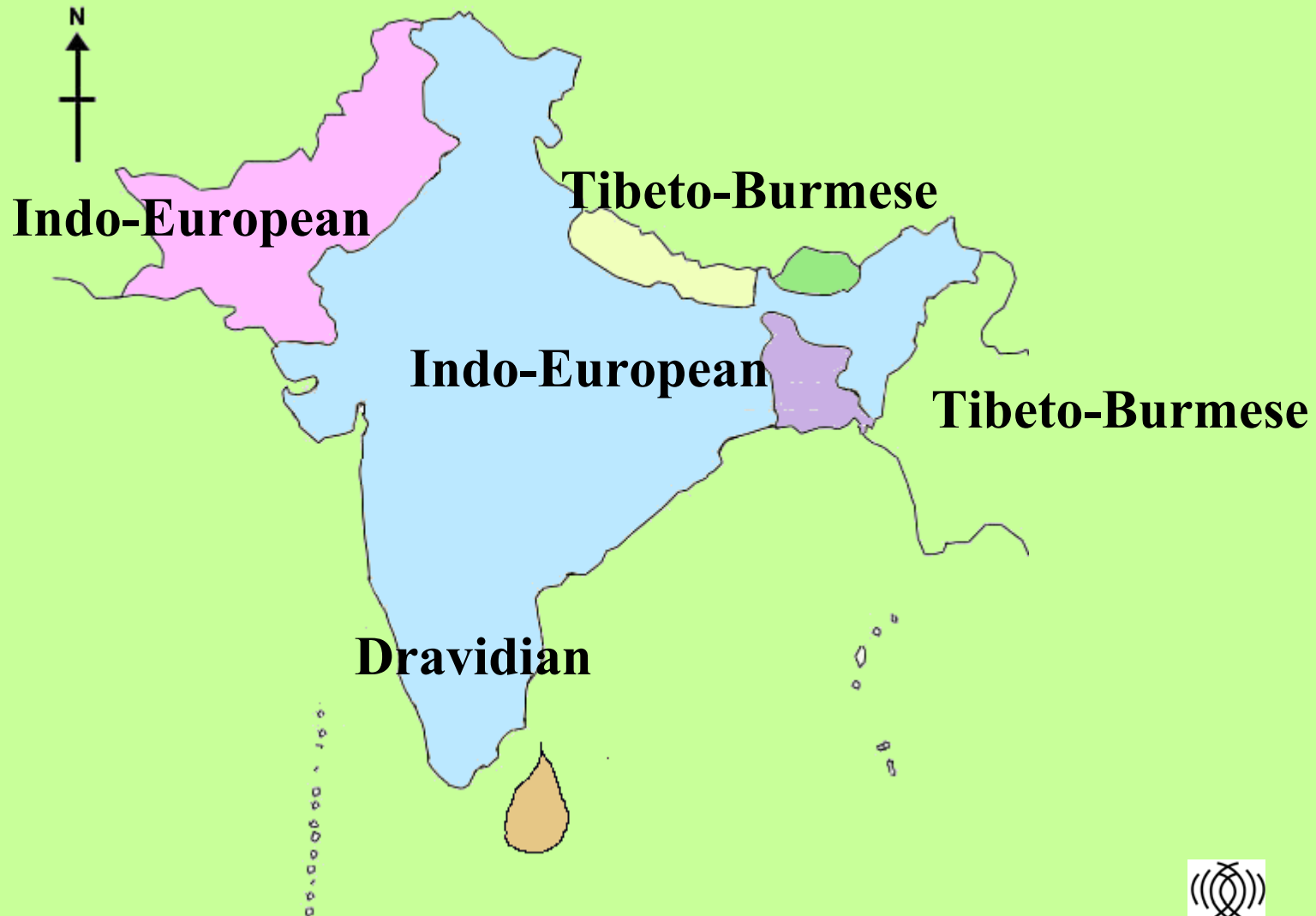
Majority Religions

intermingled

*pervade cultures
provide writing systems
set calendar system*

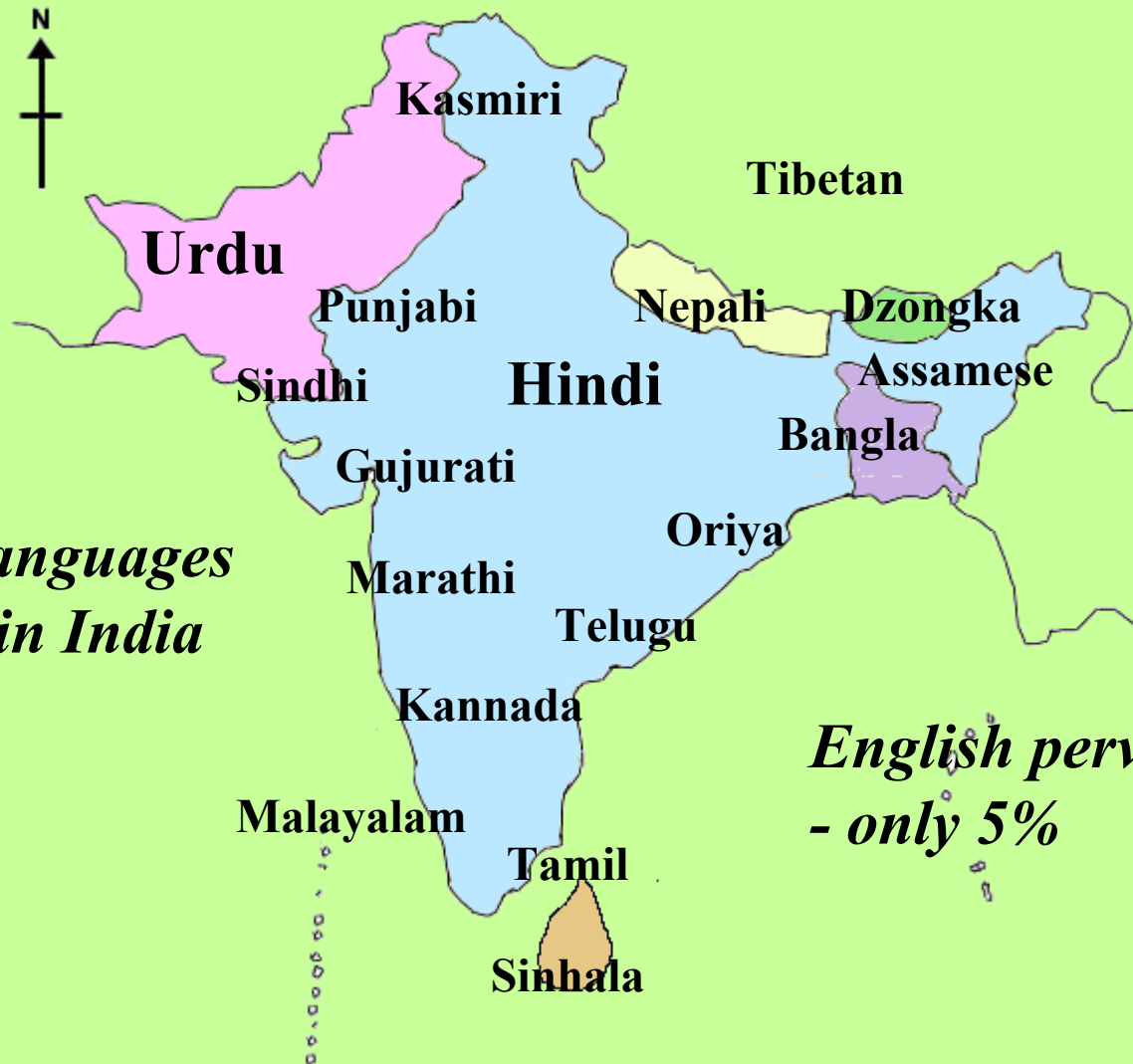


languages - groups



languages - major/official

scholarship
Sanskrit
Farsi



Over 500 languages
17 official in India

English pervasive
- only 5%

Writing systems

தமிழ்நாடு அரசு

Tamil

ಕನ್ನಡ ಸಾಹಿತ್ಯ ಪುಟ

Kannada

देवनागरि लिपि

Hindi (Devanagari)

- all derived from original script - Brahmi
 - alphabetical
 - largely phonetic - but phonology changes
 - multiple forms (conjuncts) $ख + य = ख्य$ $क + स = क्स$
 - maybe 50 languages are written
 - around 50% literacy
- new writing systems created for unwritten languages
 - by linguists, by missionaries (SIL)

a	ā	i	ī	u	ū
e		ai	o	-ṃ	
ka	kha	ga	gha		
ča	čha	ja	jha	ña	
ṭa	ṭha	ḍa	ḍha	ṇa	
ta	tha	da	dha	na	
pa	pha	ba	bha	ma	
ya	ra	la	ḷa	va	
sa, śa	ṣa	śa	sa	ha	

<- brahmi

evolved into:

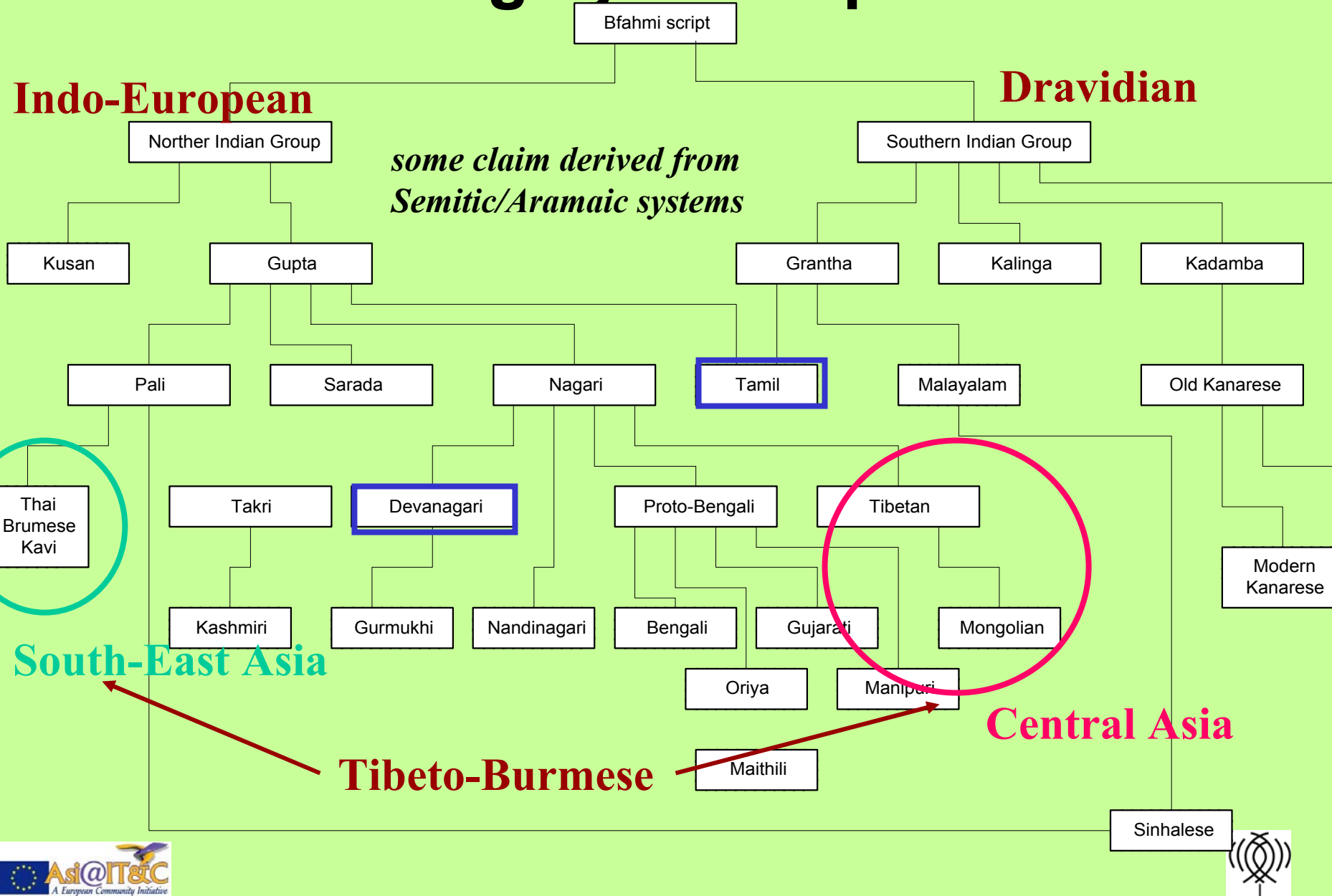
தமிழ்நாடு அரசு

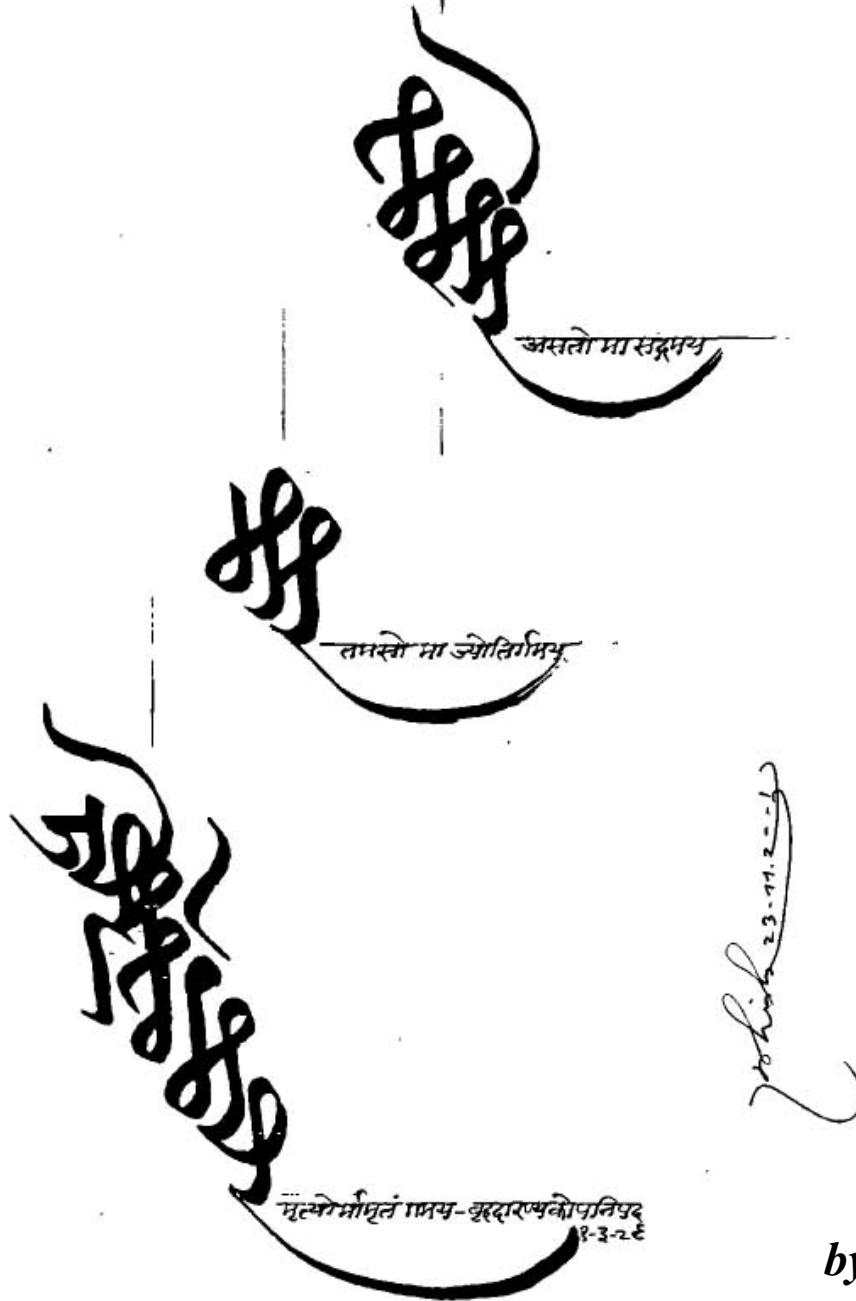
கனட சாஹித்ய பூரண

देवनागरी लिपि



how the writing system spread





**there is a
fine tradition
of
calligraphy
in South Asia**

by R.K.Joshi, Mumbai



cultural features

- **religions**
 - hindu, buddhist, islam, christian
- **numbers**
 - lakhs and crores
- **colours**
- **calendars**
 - now Gregorian
 - also Hindi, Buddhist, Islamic
- **time zones**
- **oral cultures**
- **high memory**
- **no do-it-yourself**
- **collectivist**
- **gurus**
- **...**



localisation issues

- **getting the language inside the computer**
 - **encoding**
 - **rendering**
 - **input methods**
 - **keyboards**
 - **OCR**
 - **stroke**
- **language technologies**
 - **translation**
- **Speech**



encodings

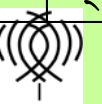
- **computer type setting and DTP**
 - enabled cheap publishing in minority languages
- **PC fonts**
 - glyphs, but only some of them
 - encodings usually accidental from keyboard
- *no data sharing*
- *but did get things started.*



sample PC font

Altsys Fontographer 3.5.2 13-Mar-95
Annapurna Devanagari, SIL Inc 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
00																
01																
02		!	"	#	\$	%	&	'	()	*	+	,	-	.	/
03	०	१	२	३	४	५	६	७	८	९	:	;	<	=	>	?
04	@	अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	क	ख	क्व	क
05	ऋ	क्ष	क्ष	ख	रु	ग	ग	ज्ञ	झ	ष	च	[\]	^	_
06	ड	डु	डु	डु	डु	डु	डु	च	च	च	छ	ज	ज	ज	झ	झ
07	भ	इ	ज	ज	ट	ट	ठ	ठ	ड	डु	{		}	~	□	
08	□	ण	प	त	त	...	त	त्र	थ	थ	द	द	द	द	□	द
09	ध	'	'	"	"	६	-	-	न	न	न्न	प	प	स	□	प
10		फ़	ब	ब	ब	भ	भ	म	म	स	य	र	र	च	च	र
11	०	८	रु	रु	ल	ल	॥	.	ल	व	०	व	¼	½	¾	श
12	श	श्र	श्र	श्र	ष	ठ	ष्ट	ष्ट	स	स	स	स	ह	ह	ह	ह
13	ह	ह	ह	ह		□	ॐ	ॐ	:
14	।	ि	ि	ि	ि	ि	ि	ि	ि
15	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु	ु



how PC fonts mess it up

three forms of the same letter, seen as part of font “design”

ः	ॆ	:	;	<	=	>	?
ए	ऐ	ओ	औ	क	क्	क्क	क्त
श्	घ	ह	[\]	^	_
च	च्च	छ	ज	ज्	ज़	झ	झ्
ळ	ड	ड्ड	{		}	~	□

**a good font takes one to three person years to design
these PC fonts may be knocked off in a few weeks.**



encodings

- **ISCII**
 - IIT Kanpur, then CDAC Pune
 - letters, not glyphs
 - like Arabic
 - needs renderer
 - based on Brahmi view
 - single code table with language switch to render “same” letter appropriately
 - sequence of characters as spoken
 - **Unicode**
 - based on ISCII but separate tables for each script
 - **still controversial**
 - **new Tamil standard**
- hardware GIST card
then software (DLL)
cost prohibitive



Unicode Devanagari

	090	091	092	093	094	095	096	097
0	ॐ	ऐ	ठ	र	ी	ॐ	ॠ	०
1	ॐ	ऑ	ड	र	ॐ	ं	ॠ	
2	ं	ओ	ढ	ल	ॐ	ॐ	ॠ	
3	ः	ओ	ण	ळ	ॐ	े	ॠ	
4		औ	त	ळ	ॐ	े	।	
5	अ	क	थ	व	ॐ		॥	
6	आ	ख	द	श	े		०	
7	इ	ग	ध	ष	े		१	

B	ई	घ	न	स	ै	क	२	
9	उ	ङ	न	ह	ॉ	ख	३	
A	ऊ	च	प		ो	ग	४	
B	ॠ	छ	फ		ो	ज	५	
C	ॠ	ज	ब	्	ौ	ड	६	
D	ँ	झ	भ	ऽ	्	ढ	७	
E	ऐ	ञ	म	ा		फ	८	
F	ए	ट	य	ि		य	९	



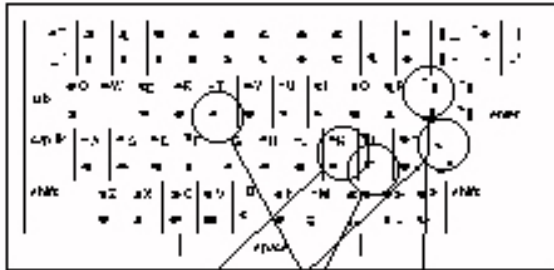
input

- **keyboards**
 - Remington
 - phonetic, English key that sounds similar
 - Inscript
- **OCR**
 - research moving into exploitation
- **hand writing strokes**
- **alternatives**
 - tap-a-tap - encoding on Simputer
 - Simpli - based on phonology of alphabet

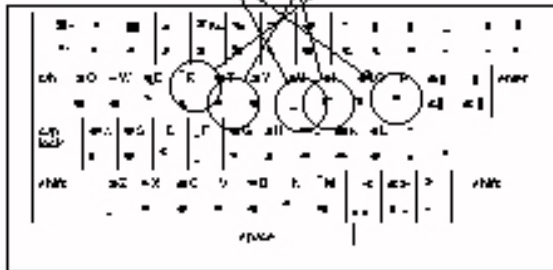


how Unicode fits together

remington keyboard



प रूर् त ि
typing sequence



phonetic keyboard

remington
key-map

font

2A 42 30 4D 24 34
internal codes

rendering

प रूर् ति

rendered
display
or print



Machine Translation Systems

- **rule based (syntax driven) systems**
 - commercial use in Europe and on web
 - research systems in India
- **translation memory systems**
 - standard support for translators in Europe
 - unused in South Asia?
- **Need language corpora**



Speech and literacy

- **Speech understanding**
 - Dragon Naturally Speaking excellent for English
 - Dragon Naturally Speaking now in India
 - few other systems
- **Speech generation - Text-to-Speech - TTS**
 - standard for European languages
 - systems available in India entering market
- **Speech dialogue**
 - commercial timetable enquiries, ...
 - why automate to save labour costs when labour is so cheap?
- **Killer application?**
 - support for the non-literate



localisation industry

- **platforms**
 - MS
 - Linux
- **applications**
 - purpose built
 - translated
- **potential market**



platforms

- **first computer at Indian Statistical Institute in 1950s**
- **Macs in early, overtaken by PCs.**
- **growing adoption of IT**
 - **6 million PCs**
 - **1 million internet subscriptions**
- **Operating systems**
 - **MS expensive**
 - **Indic renderer produced at NCST Mumbai**
 - **Linux official**
 - **many independent localisations**
 - **renderers from NCST,**
 - **ICU (International Components for Unicode) little used**
- **Simputer**



applications in local languages

- **specially produced software**
 - word processors
 - accounting
 - development project support
 - knowledge-sharing network
- **imported and localised software**
 - ??? is there any ???
 - apart from Linux platforms
 - is this an opportunity? or this an opportunity?



Market

- **growing adoption of IT**
 - 6 million PCs and growing very rapidly
 - 1 million internet subscriptions
 - Linux official
- **Indic software about to take off**
 - dominated by English - so far
 - national government has required data in local languages
 - TDIL <<<http://tdil.mit.gov.in/>>> gives useful data
 - people work in local languages
 - development projects using IT



South Asian Localisation

- **huge opportunity**
 - enormous population
 - technology now available
 - need lots more fonts
 - IT use expanding beyond English speakers
 - localisation industry embryonic



how to contact us

Pat Hall <p.a.v.hall@btinternet.com>

Venkatesh Hariharan <venky1@vsnl.com>

Tapan Parikh <tap2k@yahoo.com>

Durgesh Rao <drsistemas@vsnl.net>

