Workshop #13 **Bringing diverse linguistic communities together for a unified IDN ruleset**

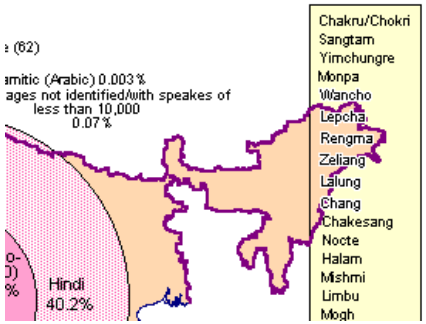
An APRiGF Event

3-5 Aug, 2014 New Delhi

ASIA PACIFIC REGIONAL INTERNET GOVERNANCE FORUM

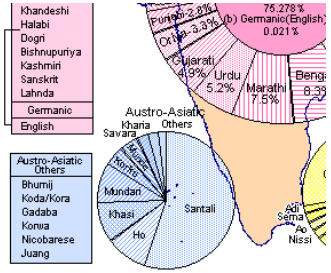
http://www.APrIGF.asia

APrIGF.ASIA



Multilingual Multiscriptal India

Bringing Diversity on a Unified Platform



Udaya Narayana Singh

VISVA-BHARATI, SANTINIKETAN

E-mail: unsciil@yahoo.com

Chair, Centre for Endangered Languages

Centre for Development of Advanced Computing, Pune

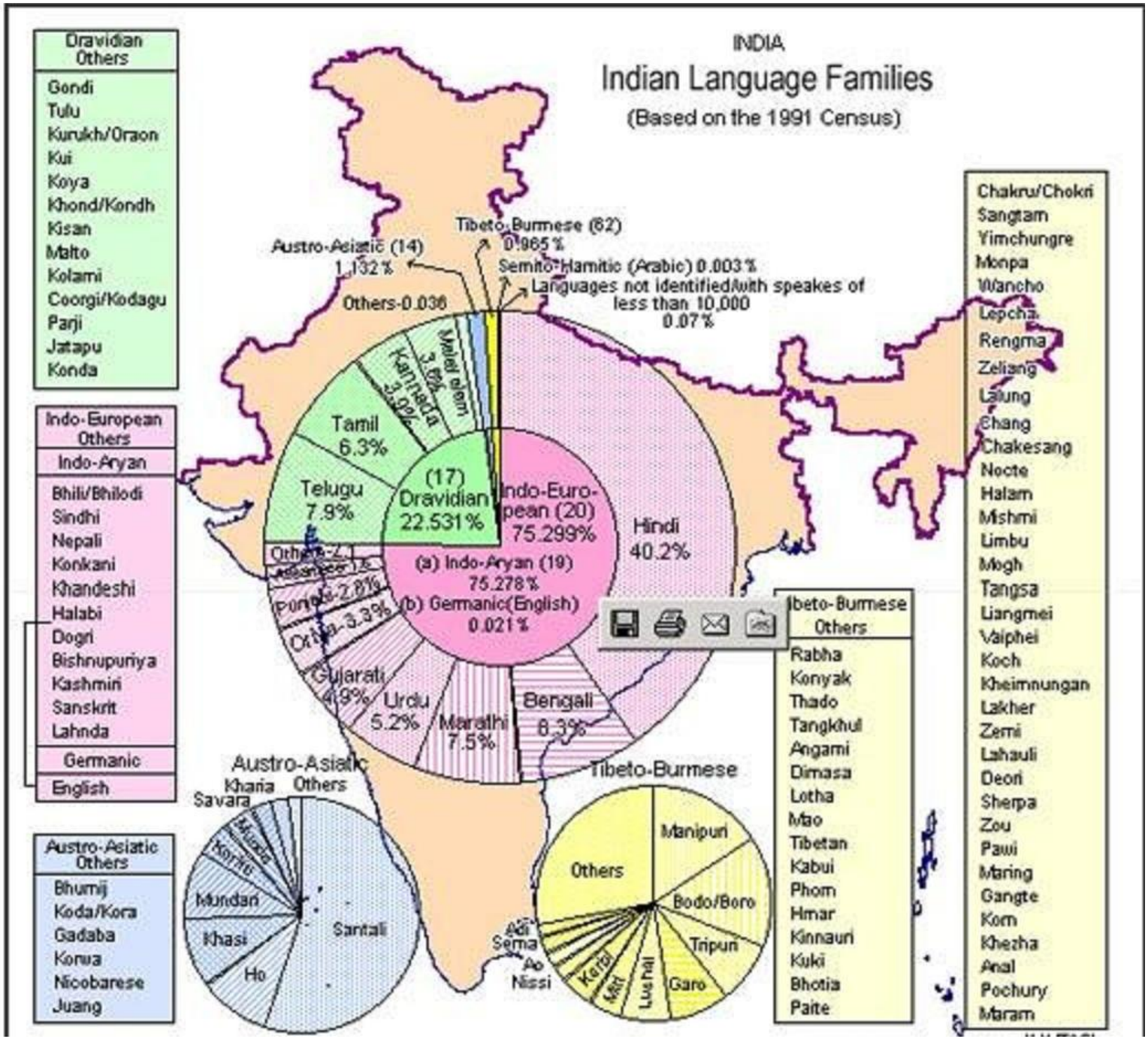
- 1576 rationalized mother-tongues (MTs) as per Census 2001
- 122 languages with 10,000 plus speakers;
- 146 used in radio network now, although before 1939 only English, Hindustani, Bangla and later, Telugu, Tamil, Marathi and Gujarati were aired.
- 35 of them publish 3954 newspapers as in 1971 (doubled in '03, with Hindi accounting for 2507, Urdu 534 & English 407.
- 22 Constitutional languages;
- 29 have one mill + speakers (Bhili, Gondi, Khandeshi, Khasi, Tulu, Mundari,Ho)
- 24 have highest/ Akademi awards

The Big Picture?

ROM	DEV	GUJ	PUN	BEN	ORI	TEL	KAN	TAM	MAL	SINH	URD	SIND	SAN	MAN
k	क	ક	ਕ	ক	କ	క	ಕ	க	ക	කා	ک	ڪ		
kh	ख	ખ	ਖ	খ	ଖ	ఖ	ಖ	–	ഖ	ඛ	کھ	ڪھ		
g	ग	ગ	ਗ	গ	ଗ	గ	ಗ	–	ഗ	ග	گ	ڳ		

- Four Major Language Families & a few minor ones
- 14 major writing systems in use but 66 scripts in all.

- 1796 other Mother-Tongues not yet classified – MTSI and SPPEL Projects should reveal more details;
- Large variation: Over 41% Indians speak Hindi with its 49 varieties & 422 million to Maram (Manipur): 10,144;
- Large non-scheduled languages; eg Bhili - 9.58 million speakers;
- 97.7% speak only 20-odd Indo-Aryan & Dravidian languages
- 2.32% of Indian population speak abt 97% of other languages


INDIA
Indian Language Families
(Based on the 1991 Census)
Dravidian Others
Gondi
Tulu
Kurukh/Oraon
Kui
Koya
Khond/Kondh
Kisan
Malto
Kolami
Coorgi/Kodagu
Parji
Jatapu
Konda
Indo-European Others
Indo-Aryan
Bhili/Bhilodi
Sindhi
Nepali
Konkani
Khandeshi
Halabi
Dogri
Bishnupuriya
Kashmiri
Sanskrit
Lahnda
Germanic
English
Austro-Asiatic Others
Bhumij
Koda/Kora
Gadaba
Korwa
Nicobarese
Juang
Austro-Asiatic (14)
1.132%
Others-0.036
Tibeto-Burmese (62)
0.965%
Semito-Hamitic (Arabic) 0.003%
Languages not identified/with speakes of less than 10,000
0.07%
Malayalam
3.6%
Kannada
3.9%
Tamil
6.3%
Telugu
7.9%
(17) Dravidian
22.531%
Indo-European (20)
75.299%
Hindi
40.2%
(a) Indo-Aryan (19)
75.278%
(b) Germanic(English)
0.021%
Punjabi-2.8%
Orissa-3.3%
Gujarati
4.9%
Urdu
5.2%
Marathi
7.5%
Bengali
8.3%
Austro-Asiatic
Kharia
Others
Savara
Munda
Korku
Mundari
Khasi
Ho
Santali
Tibeto-Burmese
Manipuri
Others
Bodo/Boro
Tripuri
Garo
Lushai
Mizo
Karbi
Adi
Sema
Ao
Nissi
Tibeto-Burmese Others
Rabha
Konyak
Thado
Tangkhul
Angami
Dimasa
Lotha
Mao
Tibetan
Kabui
Phom
Hmar
Kinnauri
Kuki
Bhotia
Paite
Chakru/Chokri
Sangtam
Yimchungre
Monpa
Wancho
Lepcha
Rengma
Zeliang
Lalung
Chang
Chakesang
Nocte
Halam
Mishmi
Limbu
Mogh
Tangsa
Liangmei
Vaiphei
Koch
Kheimnungan
Lakher
Zemi
Lahauli
Deori
Sherpa
Zou
Pawi
Maring
Gangte
Kom
Khezha
Anal
Pochury
Maram

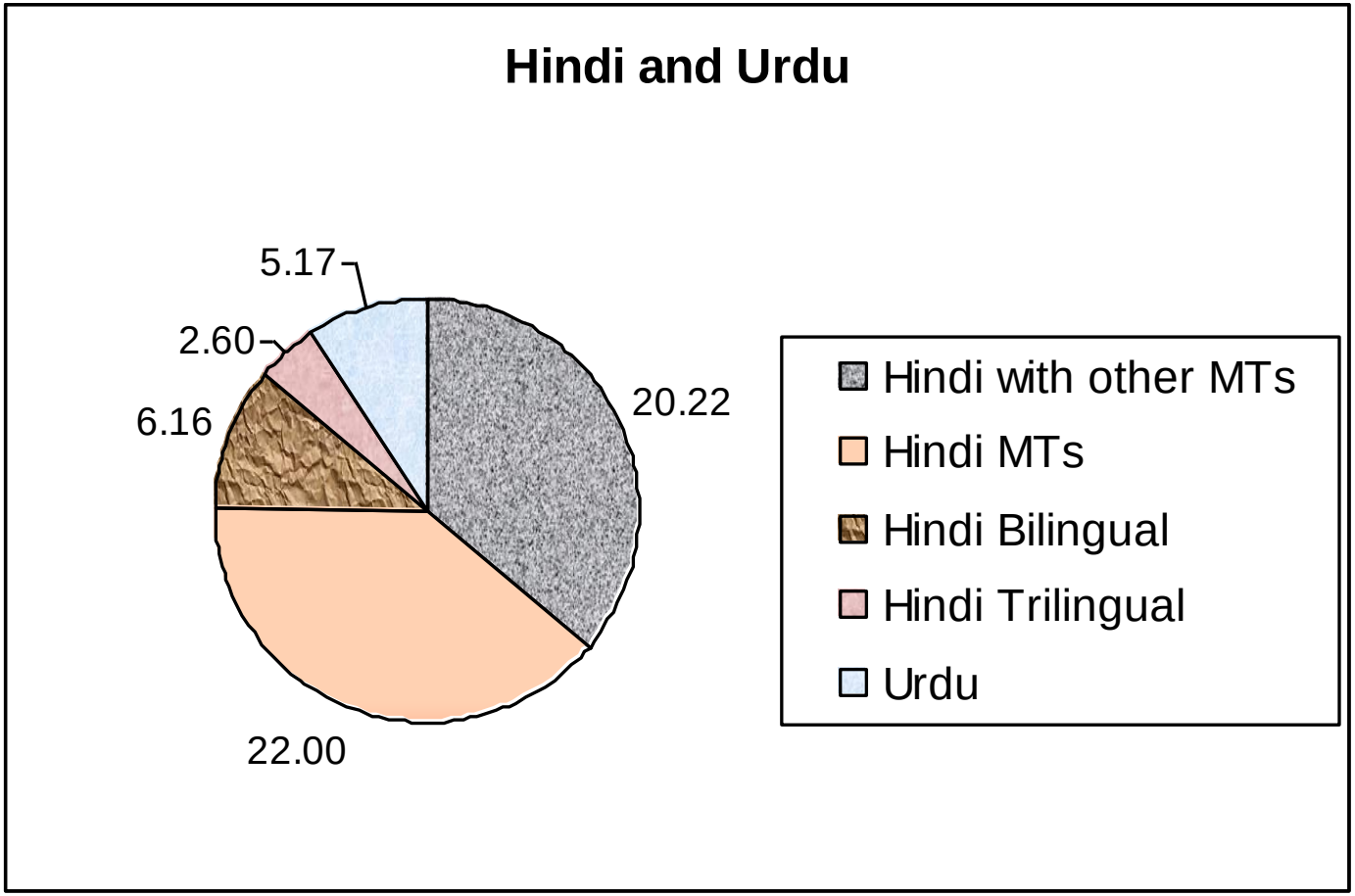
The Linguistic Landscape

Census1901: 179 languages & 544 Dialects
Census1951: 14 lgs in the Constitution, 23 major tribal lgs, 22 other minority languages & 722 smaller languages/ Dialects
Census1961: 193 classified lgs & 1,652 mother-tongues

S. No.	Items	Number
1.	'Raw returns of Mother Tongues	6661
2.	Rationalized grouping of Mother Tongues	3592
3.	Classified Mother Tongues	1635 (out of 3592)
4.	Unclassified Mother Tongues	1957 (out of 3592)
5.	Identifiable Mother Tongues (returned by more than 10,000 speakers)	234
6.	Unpublished Mother Tongues (1976 CCPA & Other restrictions)	68

Thus, **Census 2001** with 234 Mother Tongues were grouped under 122 languages – a huge challenge for their preservation

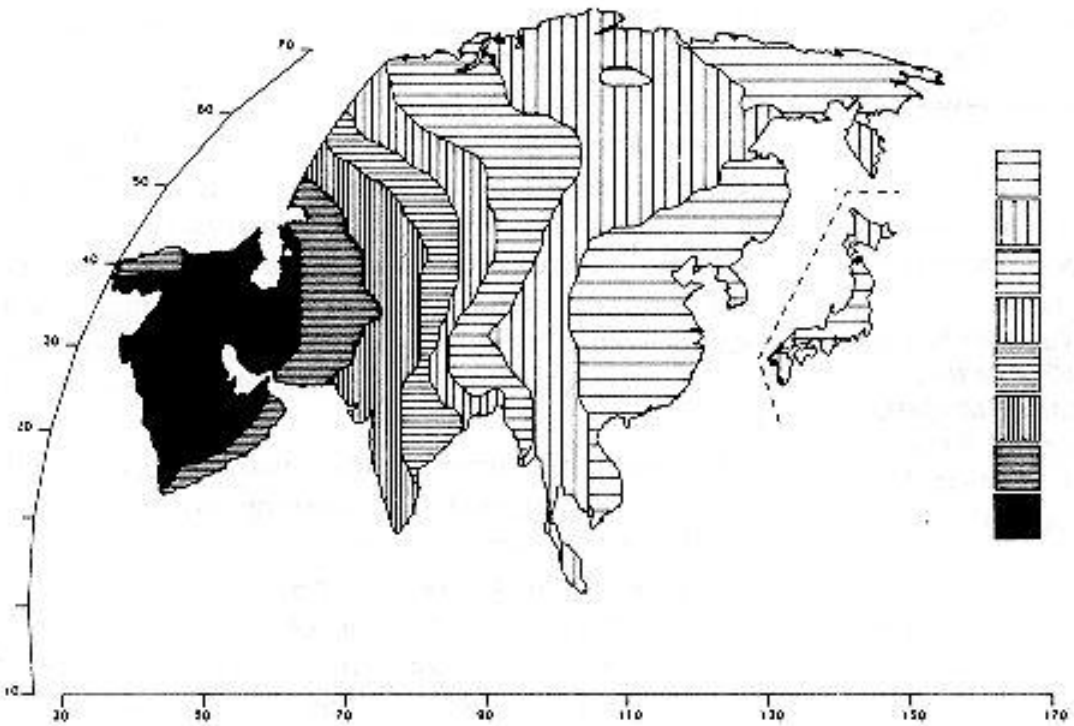
An Example of Large Variation: Only Hindi-Urdu has 49 mother-tongues



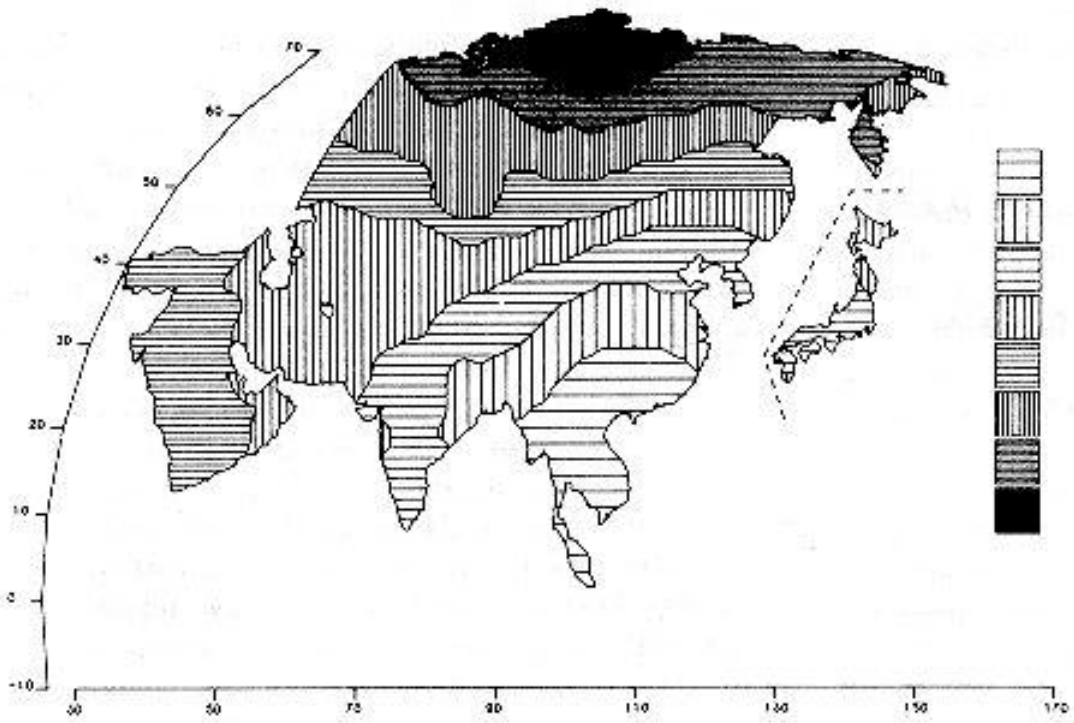
78.61% are in rural areas, 21.39% are urbanites.

THE SOUTH ASIAN SCENE: PATTERNS OF MIGRATION - FROM THE WEST TO THE ORIENT

- The expansions of Indo-Europeans towards India may have started around 3800 years ago.
- Perhaps 1000 years (or earlier), Indo-European speakers from the oases south of the Urals, north of the Black Sea and in western Kazakstan moved west ward & east-ward.
- They moved to the south, mixing with (presumably) Dravidian speakers of the region including Iran, India an Pakistan.
- Later, they also mixed with the Austric and th Tibeto-Burmans, or the Kiranti group.



Defining South Asia



Ref: *History and Geography of Human Genes* by Menozzi, Piazza and Cavalli-Sforza, Chapter 4

- **Widely believed: The first inhabitants were Negretoids who almost vanished in course of time, leaving only traces.**
- **By the time the nomadic Indo-Europeans came to India, the space already had other residents – the *NiSaada*s, the *Dravida*s and the *Kiraata*s who, later, got assimilated.**
- **Even if we consider the genetic gradation of Asian people, South Asia will emerge as a natural socio-historical entity that has retained all kinds of variations.**

From *Clashes* to *Comingling:* Evolving a Common Script Base

- In the political entity called India, there is no doubt "there have been occasional clashes of interests and ideals or of attitudes which ... linked up with the desire to wield power and control pelf...."
- All the latecomers also (such as the Indo-Mongoloids, or the Turks or the Perso-Arabic stocks) fell into this pattern of assimilation established through a 3,000 years of contact and convergence.
- One of the best examples of that is structurally and genealogically different languages accepting a common script base.

CONSTITUTIONAL LANGUAGES IN DESCENDING ORDER

S No	Languages	Mother-Tongue Speakers in mill	% to Total Population
1	**Hindi**	422.05	**41.03**
2	**Bengali**	83.37	**08.11**
3	**Telugu**	74.00	**07.19**
4	**Marathi**	71.94	**6.99**
5	**Tamil**	60.79	**5.91**
6	**Urdu**	51.54	**5.01**
7	**Gujarati**	46.09	**4.48**
8	**Kannada**	37.92	**3.69**
9	**Malayalam**	33.07	**3.21**
10	**Oriya**	33.02	**3.21**
11	**Punjabi**	29.10	**2.83**

S No	Languages	Mother-Tongue Speakers in mill	% to Total Population
12	**Assamese**	13.17	**1.28**
13	**Maithili**	12.18	**1.18**
14	**Santali**	6.47	**0.63**
15	**Kashmiri**	5.53	**0.54**
16	**Nepali**	2.87	**0.28**
17	**Sindhi**	2.54	**0.25**
18	**Konkani**	2.49	**0.24**
19	**Dogri**	2.23	**0.22**
20	**Manipuri**	1.47	**0.14**
21	**Bodo**	1.35	**0.13**
22	**Sanskrit**	0.01	**N(egli)**

CONSTITUTIONAL LANGUAGES : Decadal increase

S No	Languages (% of Increase)	1971-1981	1981-1991	1991-2001
1	Hindi	27.12	27.84	28.08
2	Bengali	14.52	35.67	19.79
3	Telugu	13.11	30.41	12.10
4	Marathi	18.41	26.35	15.13
5	Tamil	NA	NA	14.69
6	Urdu	22.08	24.23	18.73
7	Gujarati	27.83	23.02	13.32
8	Kannada	18.36	27.46	15.79
9	Malayalam	17.15	18.20	8.85
10	Oriya	15.90	21.89	17.66
11	Punjabi	39.00	19.21	24.48

S No	Languages (% of Increase)	1971-1981	1981-1991	1991-2001
12	Assamese	NA	NA	0.68
13	Maithili	22.71	3.25	56.81
14	Santali	14.41	20.40	23.03
15	Kashmiri	27.31	NA	NA
16	Nepali	-4.17	52.62	38.29
17	Sindhi	21.92	3.84	19.44
18	Konkani	4.09	12.13	41.37
19	Dogri	17.82	NA	NA
20	Manipuri	13.86	40.91	15.47
21	Bodo	NA	NA	10.52
22	Sanskrit	176.04	714.54	-71.58

Speech variation is the order of the day

"Father was posted in the 'true' Telugu districts; it was but natural that we siblings spoke purest Telugu. Grandma, being a true Tamilian, made sure we also spoke Tamil, and tutors gave us the basics of English. Mother, being a very wise lady, engaged a Hindi tutor. To complicate matters, father went to Calicut and Mangalore, making us reasonably proficient in Kannada and Malayalam.

Santosh Kshirsagar Calligraphy Design

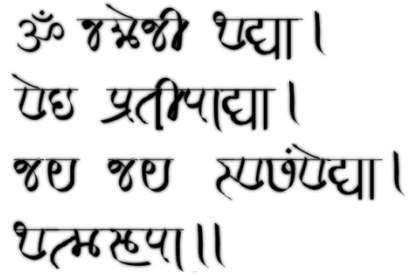
This multi-lingual upbringing can be a boon and a curse."

In a *Times of India*'s web-portal, Sandy Sundaram (April 16, 2003) in her '*Speaking Free*', recollects:

But Scriptal Variation is also a reality!

One Script for Many Languages

- **Devanagari used by official languages : Hindi, Marathi, Konkani, Nepali, Sanskrit, Sindhi, Dogri and Santhali are written, not to mention around 30 more languages**
- **Eastern Brahmi Base (close to Bengali) has been used by Assamese, Bangla, Bodo, Manipuri and Maithili.**
- **Modi was used primarily by Marathi until 19th century and also by Urdu, Kannada, Gujarati Hindi and Tamil.**

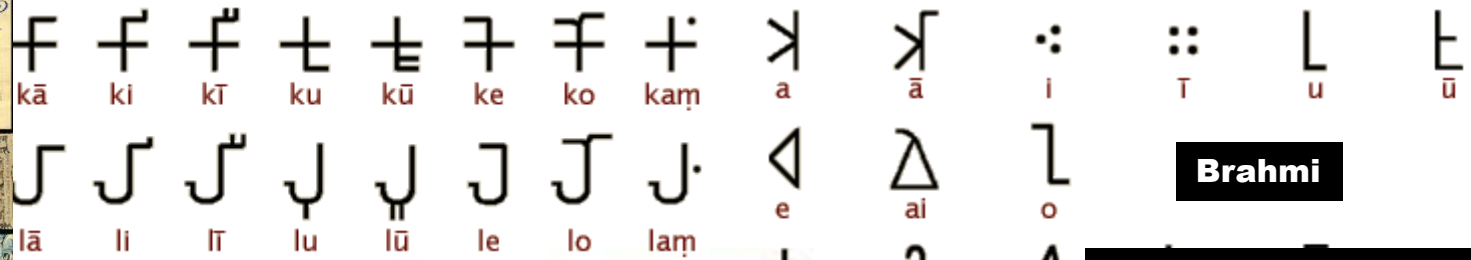


Modi Script sample

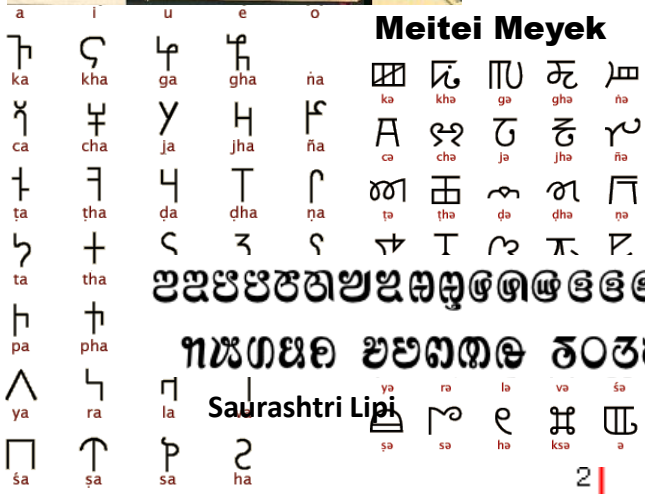
Many Scripts for One Language

In contrast, there could also be languages like Konkani written in Roman, Devanagari, Kannada, and Malayalam scripts. Earlier Santali written in Roman, Ol Chiki, Devanagari, Oriya ans Bangla scripts. Or, Bodo in Devanagari, Roman and Assamese.

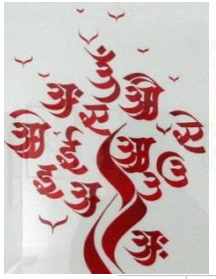
NEW SCRIPTS EMERGING AS ALWAYS



Brahmi



Meitei Meyek

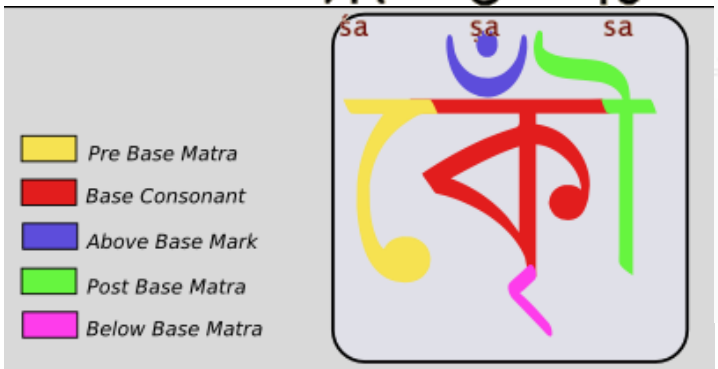


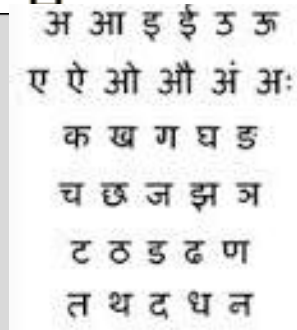
Saurashtri Lipi

Kharosthi

Akkhara Muni (Pali)

Rejang<Kawi<Brahmi



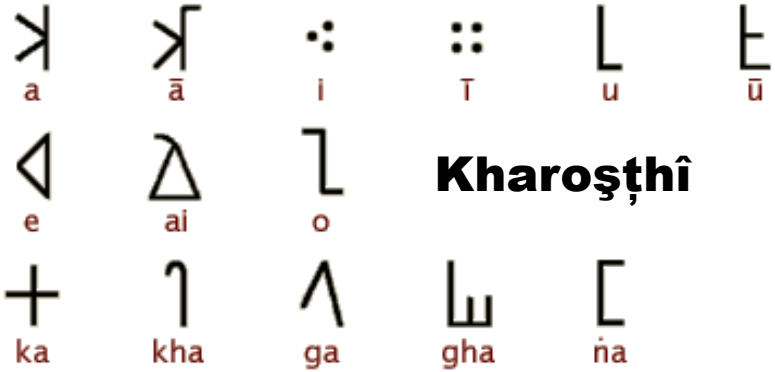


Nagari Lipi

Brâhmî & Kharoṣṭhî

- **Georg Bühler (1895) in his *On the Origin of the Indian Brahma Alphabet* argued that the mother script to the most present day Indic scripts - Brâhmî originated in the 8th cent BC from the Phoenician.**
- **David Diringer (1953: *The Alphabet*) questioned this date, and thought the origin was perhaps Aramaic. Another non-Indologist - Hans Jensen (1935) relied heavily on Buhler.**
- **The thesis that Brâhmî was developed out of a Semitic prototype in pre-Mauryan India (before 7th Cent BC?) was accepted by the western scholars but rejected by the South Asian experts, who derive it from the Indus Valley Writing system (3300-1300 BC).**

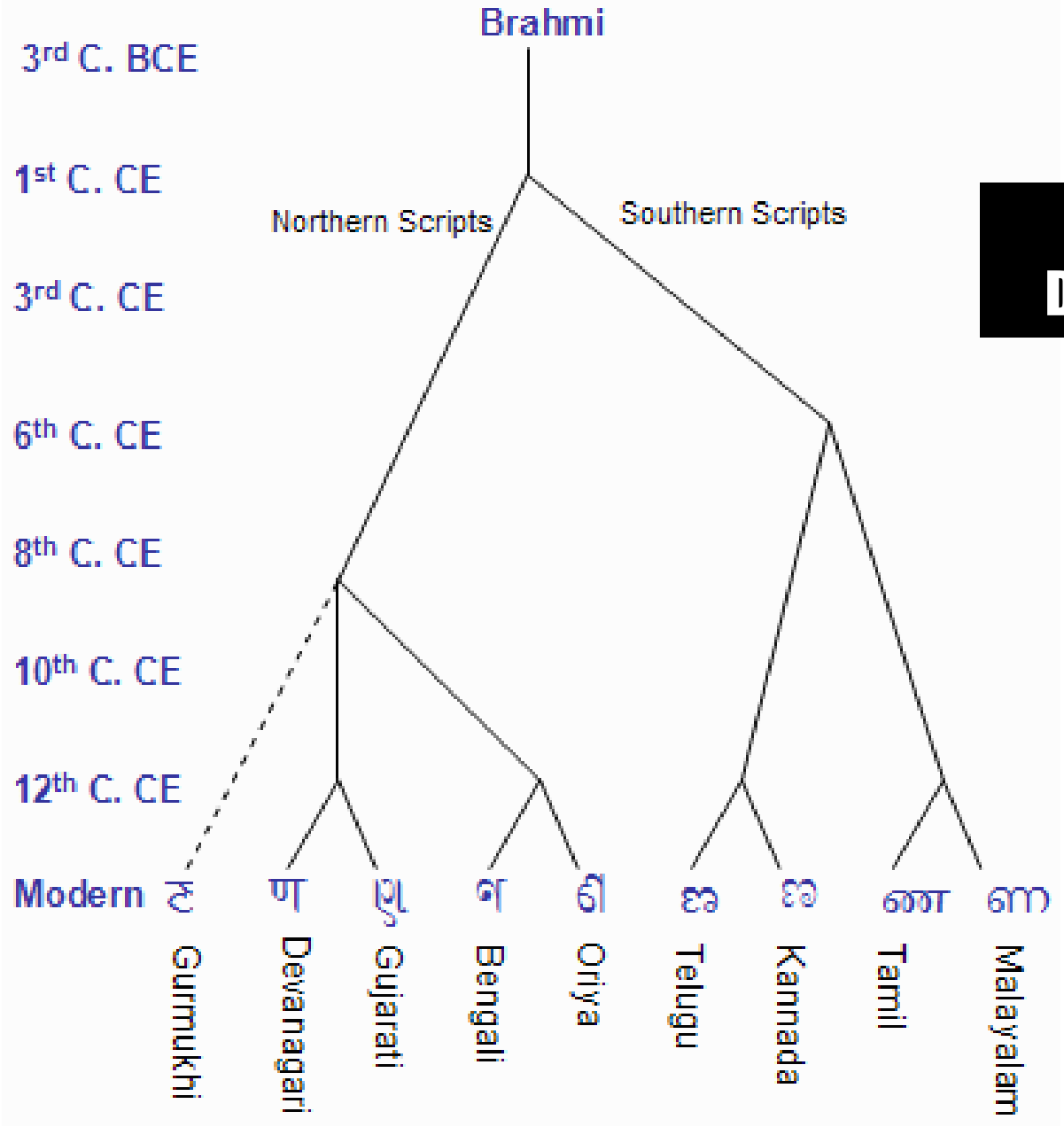
Brâhmî



Kharoṣṭhî

A greater clarity on the Indus valley script & discovery of Greek and Aramaic inscriptions of Ashokan times from Pakistan/ Afghanistan, and the studies by Fussman, von Hinüber, and Falk show that Brâhmî, &Kharoṣṭhî emerged in the Mauryan period (321-232 BC).

An Early Divergence



The diagram shows an early divergence between North and South Indian scripts as shown by Peter T. Daniels & William Bright (1996) *The World's Writing Systems*, OUP.

Internal Complications; eg অংক vs অঙ্ক vs অঙক

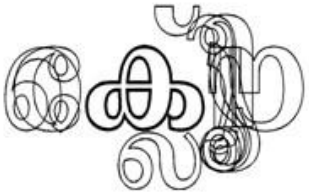
- **The variety of conjuncts and the accent markings in old script-variants of Brâhmi, known as Vedic, Sâmavedic, and Grantha make these heritage scripts a challenge.**
- **Besides, each of the modern Brâhmi systems have numerous glyphs and conjoined or combined letter-characters. While Telugu alone has 720 glyphs, Bengali-Assamese have 520 of them, Manipuri (meitei-Meyek) 509, Gujarati 436, Devanagari & Malayalam have 425, Oriya 410, Kannada 399, Punjabi 358, and Tamil 316.**
- **Bodo, drawing largely on Devanagari has 678 glyphs. Santali Ol Chiki, a newly developed script also has 678 glyphs, and so had the older script of Maithili – known as Tirahuta, or Dogri using 678 glyphs.**
- **There are a total 1520 glyphs are in the Vedic Sanskrit Open Type Font, whereas Sâmavedic has 1489 glyphs, and Grantha 1143 glyphs.**
- **Urdu Nataaliq has 563 glyphs, Kashmiri 584& Sindhi 411 glyphs.**

স sə + ত tə + র rə = স্ত্র strə,
ম mə + প pə + র rə = ম্প্র mprə,
জ jə + জ jə + ব্ব wə = জ্জ্ব jjwə,
ক kə + ষ ṣə + ম mə = ক্ষ্ম kṣmə

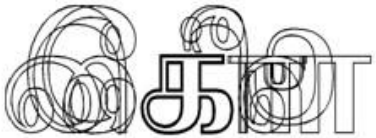
এত [æt̪ɔ] "so much",
এ্যাকাডেমী [ækademi] "academy",
অ্যামিবা [æmiba] "amoeba"

Cf. Mahesh Kulkarni's 2008-Report on 'Development of Open Type Fonts' C-DAC

Complications



Malayalam



Tamil

Gujarati

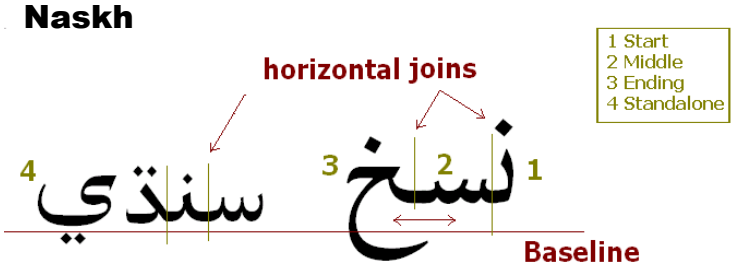
Devanagari

Oriya

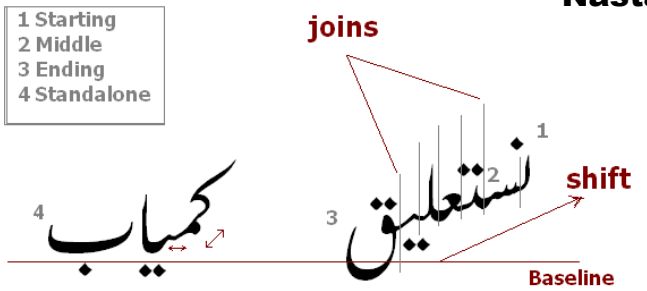
Irrespective of the South-North division, each of these writing systems has numerous difficulties for the font developers because of the way their syllabic structures are captured by the graphics – protruding below, above, and on the aisles as well as hanging letters. The Vowels & their Mâtrâ or shades are equally complicated.

One requires to know/write a Script Grammar for all Indic writing systems in order that their font-face designs are able to capture the aesthetics of the same. Only Sindhi uses Perso-Arabic based systems – Naskh (linear-seamless) and Kashmiri & Urdu use Nastaliq (Right-to-left plus Top-to-bottom):

Naskh



Nastaaliq



Concluding Remarks

- What is required is an **early implementation of IDN** in the 14 major Indian scripts for which NORMALIZATION and VARIANT TABLES exist.
- Having done that, we require some careful **Registrars to monitor** this implementation.
- **Localization** for the purposes of E-Governance would become an issue when these scripts are **used alone or in combination**, as we often need a mixture of two scripts or transliteration of large portions in another script in documents.
- **Spelling standardization** is another challenge faced by those developing **Style Manuals** for each writing system/language, and without that, search on the internet, especially multilingual search would be hampered.
- Many major search engines cannot handle these issues at present.
- Recognizing that हिन्दी and हिंदी in Devanagari are indeed the same thing is obvious for human readers but not so obvious for the machines. That such combinations could also be written as हिँदी or हिन्-दी could puzzle many.
- The present workshop only flags some of these issues for the interested researchers and developers in the field.