Indian languages, Unicode and automatic Transliteration

Girish Nath Jha
Associate Professor, Computational Linguistics
Special Center for Sanskrit Studies, J.N.U., New Delhi – 110067

Mukesh and Priti Chatter Distinguished Professor of History of Science,
University of Massachusetts Dartmouth, USA

Overview of Indian languages

In this presentation…

- Overview of Indian languages
- Indian languages scripts and spread of Unicode
- Multilingualism and multi-script languages
- Case for Automatic Transliteration
- Demo

- India has over 1600 languages from 5 language families
- 22 of these are classed as national (or 'scheduled') languages (have been listed in the constitution's 8th schedule)
- Many of these are also official languages of states they are spoken in
- Hindi is National as well as Official language of Indian union (50% plus people)
- English is not national language but has an 'Associate Official Language' status with 6% speakers

Transliteration standards

IAST

ITRANS

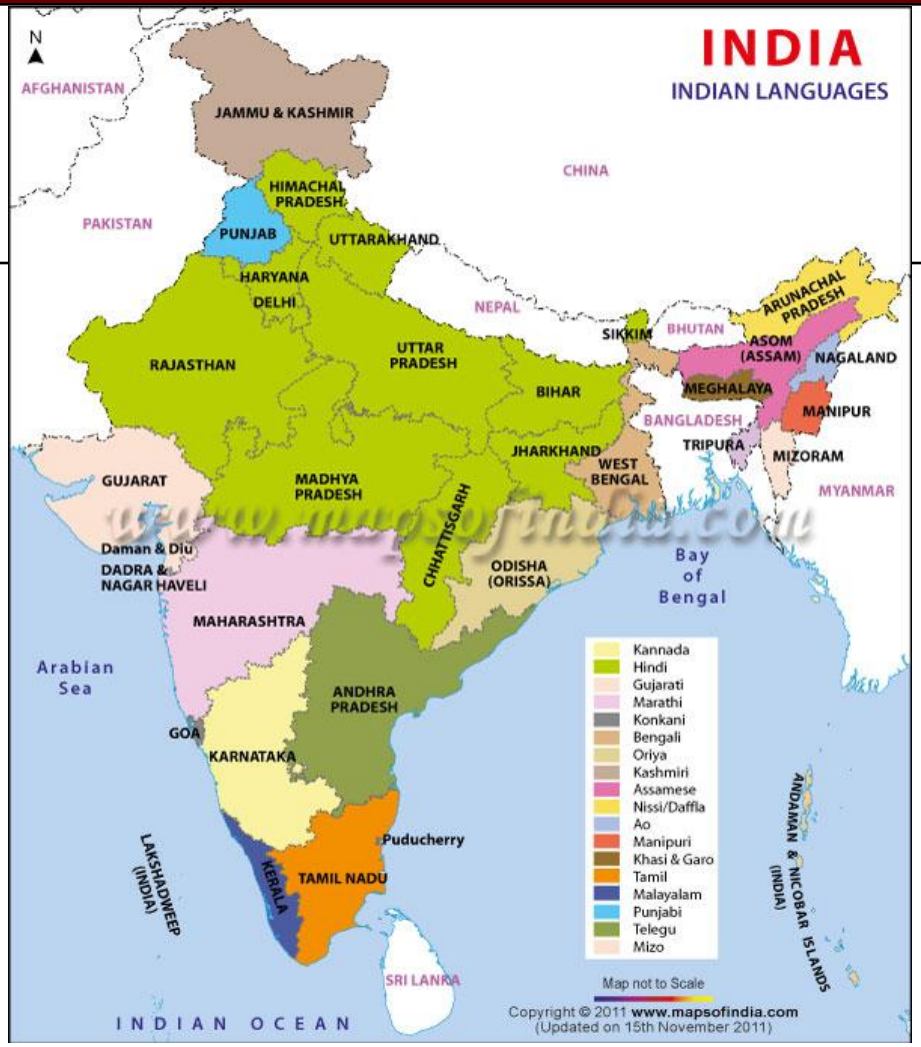
IPA

Unicode

National Languages & Scripts

Sl. No.	Language	Script
1.	Hindi	Devanagari
2.	Sanskrit	Devanagari
3.	Marathi	Devanagari
4.	Konkani	Devanagari
5.	Nepali	Devanagari
6.	Maithili	Devanagari
7.	Sindhi	Devanagari
8.	Bodo	Devanagari
9.	Dogri	Devanagari
10.	Santhali	Devanagari, Ol Chiki
11.	Bengali	Bengali
12.	Assamese	Bengali
13.	Manipuri	Bengali, Meithei
14.	Gujarati	Gujarati
15.	Kannada	Kannada
16.	Malayalam	Malayalam
17.	Oriya	Oriya
18.	Punjabi	Gurmukhi
19.	Tamil	Tamil
20.	Telugu	Telugu
21.	Urdu	Perso-Arabic
22.	Kashmiri	Perso-Arabic

Courtesy:mapsof india.com



Multi script languages

- Hindi (Devanagari, Perso-Arabic, Gurumukhi)
- Urdu (Devanagari, Perso-Arabic)
- Kashmiri (Devanagari, Perso-Arabic)
- Punjabi (Gurumukhi, Perso-Arabic)
- Assamese (Assamese, Bengali)
- Manipuri (Metei-Mayek, Bengali)
- Bodo (Bengali, Devanagari)
- Sanskrit (Devanagari, Bengali, Grantha, Brahmi, Maithili, Gujarati…..)

Indian constitution guarantees rights to each language

Challenges in language policy and planning

- Kashmiri vs Urdu
- Hindi vs Urdu
- The case of Sanskrit
- Promoting Hindi and opposition
- Multiple scripts
- Including a language in constitution
- 3-language formula

Various government agencies fund language development

- **MHRD (Ministry of Human Resource Development) MCIT(Ministry of Communications & Information Technology)**
- **MST(Ministry of Science & Technology)**
- **MC (Ministry of Culture)**

The IT Ministry has a Technology Development for Indian Languages (TDIL) program

Indian language scripts

Brahmi and IL scripts

Indian language scripts (except Perso-Arabic) have evolved from Brahmi script (4th Century BC)

Sanskrit sound system and Devanagari alphabet

Indo Aryan and Dravidian sound system

Govt's efforts to standardize scripts are not successful

ISCII - 1991

Indian Script Code for Information Interchange

(BIS No. IS:13194-1991)

ISCII-Devanagari

	A0	B0	C0	D0	E0	F0
0		ओ	ढ	ऱ	◌	EXT
1	ँ	औ	ण	ल	◌	०
2	ं	ऑ	त	ळ	◌	१
3	ः	क	थ	ऴ	◌	२
4	अ	ख	द	व	◌ ◌	३
5	आ	ग	ध	श	◌	४
6	इ	घ	न	ष	◌	५
7	ई	ङ	ऩ	स	◌	६
8	उ	च	प	ह	◌	७
9	ऊ	छ	फ	IIIV	◌	८
A	ऋ	ज	ब	ा	।	९
B	ऌ	झ	भ	ि		
C	ए	ञ	म	ी		
D	ऐ	ट	य	ु		
E	ऍ	ठ	य़	ू		
F	ऑ	ड	र	ृ	ATR	

Efforts by CDAC

The GIST group at CDAC Pune promoted ISCII

Unicode

- Devanagari chart
- Indian languages supported by unicode
  - Bengali and Assamese, Brahmi, Chakma
  - Devanagari, Devanagari Extended, Vedic Extensions
  - Gujarati, Gurmukhi, Kaithi
  - Kannada, Kharoshthi
  - Lepcha, Limbu
  - Malayalam, Meetei Mayek, Meetei Mayek Extensions
  - Ol Chiki, Oriya, Saurashtra
  - Sharada, Syloti Nagri, Takri
  - Tamil, Telugu, Thaana
- Ancient scripts – Brahmi, Grantha, Sharada
- How does unicode work
- How can we encode it (HTML, Java etc)

Automatic script conversion

- Comparing two scripts
- Understanding codepages
- Preparing charts of matches and misses
- Creating parallel arrays
- Use conditional statements and controlled string processing for nuances

Unicode

- How can we encode it (HTML, Java, databases etc)
- Practice Unicode in HTML
- HTML source
- Unicode editors
- Unicode in Java/JSP
- Uniode in RDBMS (backend)
- Unicode in RDBMS (front end)

Demo of the transliteration engine

- Devanagari – Roman
- Devanagari to north Indian (Indo Aryan language) scripts
- Devanagari to south Indian (Dravidian language) scripts
- Issues in connversion

Dziękuję !

Dhanyavaad !

Thank you !

girishjha@gmail.com