multilingual multimodal resources - creating a leveling field in India

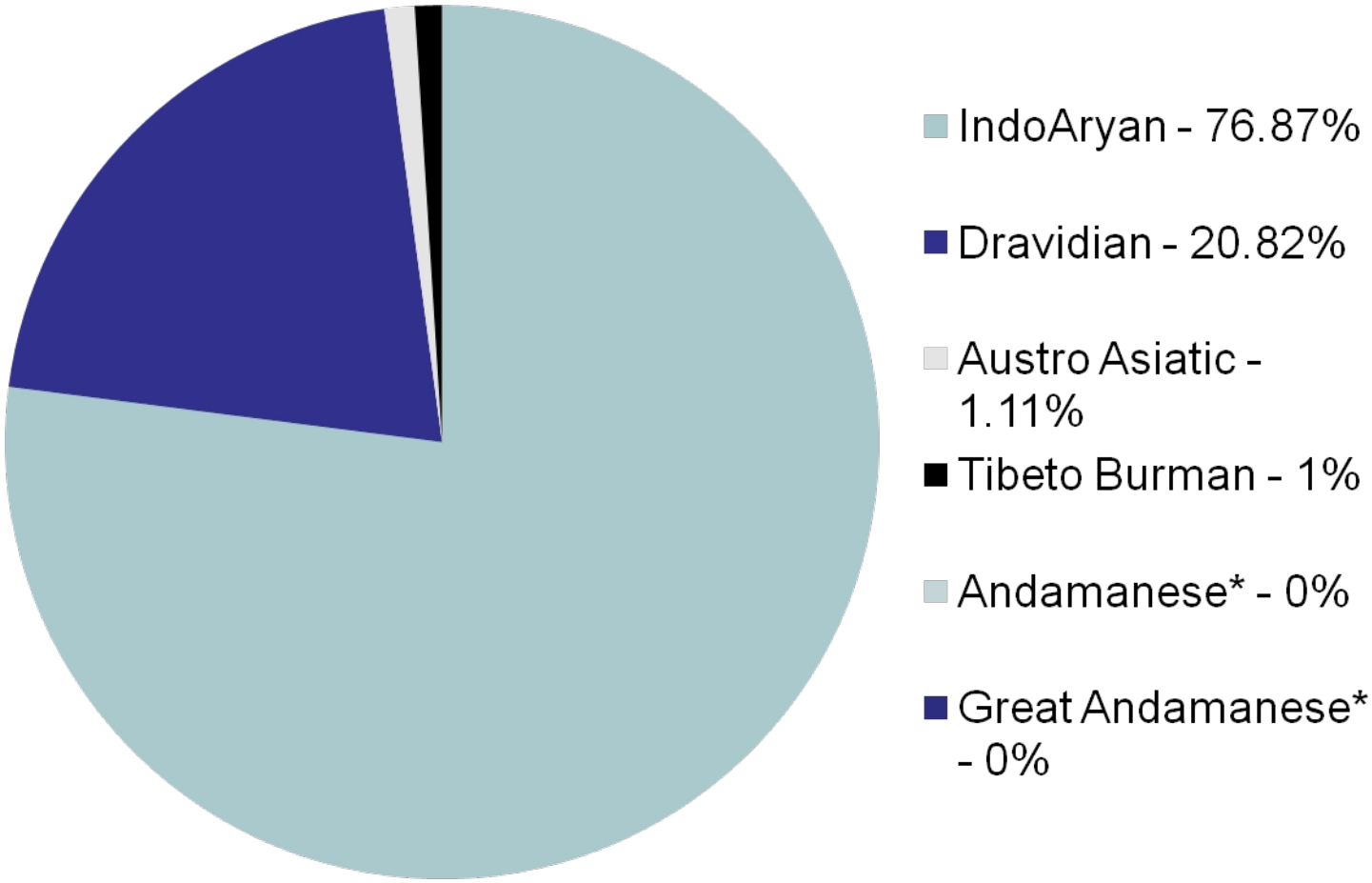
Girish Nath Jha
Special Center for Sanskrit Studies
Jawaharlal Nehru University, New Delhi-110067

India’s language situation

Unity in diversity

- 1 billion plus people, at least 7 races, 2000 castes, 43000 sub-castes
- Hinduism (with several sects and several hundred million gods), Islam, Christianity, Sikhism, Jainism, Buddhism and many more
- 5 language families - Indo Aryan, Dravidian, Austro Asiatic, Tibeto-Burman, Andamanese
- 25 states in India, more than 1600 languages

Indian Language Families and % Speakers



Status of languages

- 22 **national languages** and 12 scripts
- Most of these are also **official languages** of the states they are spoken in
- 100 **mother-tongues** reported in census 2001
- About 1000 **documented languages** and dialects
- Hindi – (**NOL**) 42% speakers, the official language of Union with English (**AOL**) as its associate (4 %)

Scheduled 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

Indian Constitution and language planning

- **448 articles, 12 schedules, 107 amendments (so far)**
- **Article III – Fundamental rights**
- **Article IV A – Fundamental duties**
- **Article XVII – Official Language**
- **Article XVII – Regional Languages**
- **Article XVII – Language of Supreme Court and High Court**
- **Article XVII – Special Directives**

Ministries which matter

- **MHRD (Ministry of Human Resource Development)**
  - Central Institute of Indian Languages (CIIL), Mysore
  - University Grants Commission (UGC) and Indian universities
- **MCIT(Ministry of Communications & Information Technology)**
  - Department of Technology (DIT)
    - Technology Development for Indian Languages (TDIL)
- **MST(Ministry of Science & Technology)**
  - Department of Science & Technology (DST)
- **MC (Ministry of Culture)**
  - Anthropological Survey of India

Ministry of HRD

- **Central Institute of Indian Languages (CIIL)**
  - New Linguistic Survey of India (NLSI)
  - National Translation Service
  - Linguistic Data Consortium for Indian Languages (LDC-IL)
  - Development and Promotion of Minor Indian Languages
  - National Testing Mission (NTM)

Technology Development for Indian Languages (TDIL) → Min. of Communication & IT

- **TDIL was established in 1991**
- **objectives**
  - **develop and promote** the information processing tools
  - **support R&D** efforts in the area of information processing in Indian Languages and to support research on Knowledge Tools
  - **consolidate IL technologies** thus developed for Indian Languages
- **activities** (in collaboration with CDACs and universities/institutes)
  - National Rollout plan
  - Funding Language and speech technology projects
  - Work on corpora and tool standards
  - Create and seed dedicated research groups in each region of India

National Rollout Plan (CDAC Pune)

- Software tools and fonts for all 22 Indian languages have been released in the public domain

- The CD-ROM typically contains basic software tools like
  - Fonts, Keyboard Drivers, converters, editors, typing tutors
  - Integrated Word Processor
  - Open source localized document editor - Bharateeya Open Office
  - Bilingual Dictionaries
  - Browser
  - Email Client
  - Messenger

Mission Mode Projects

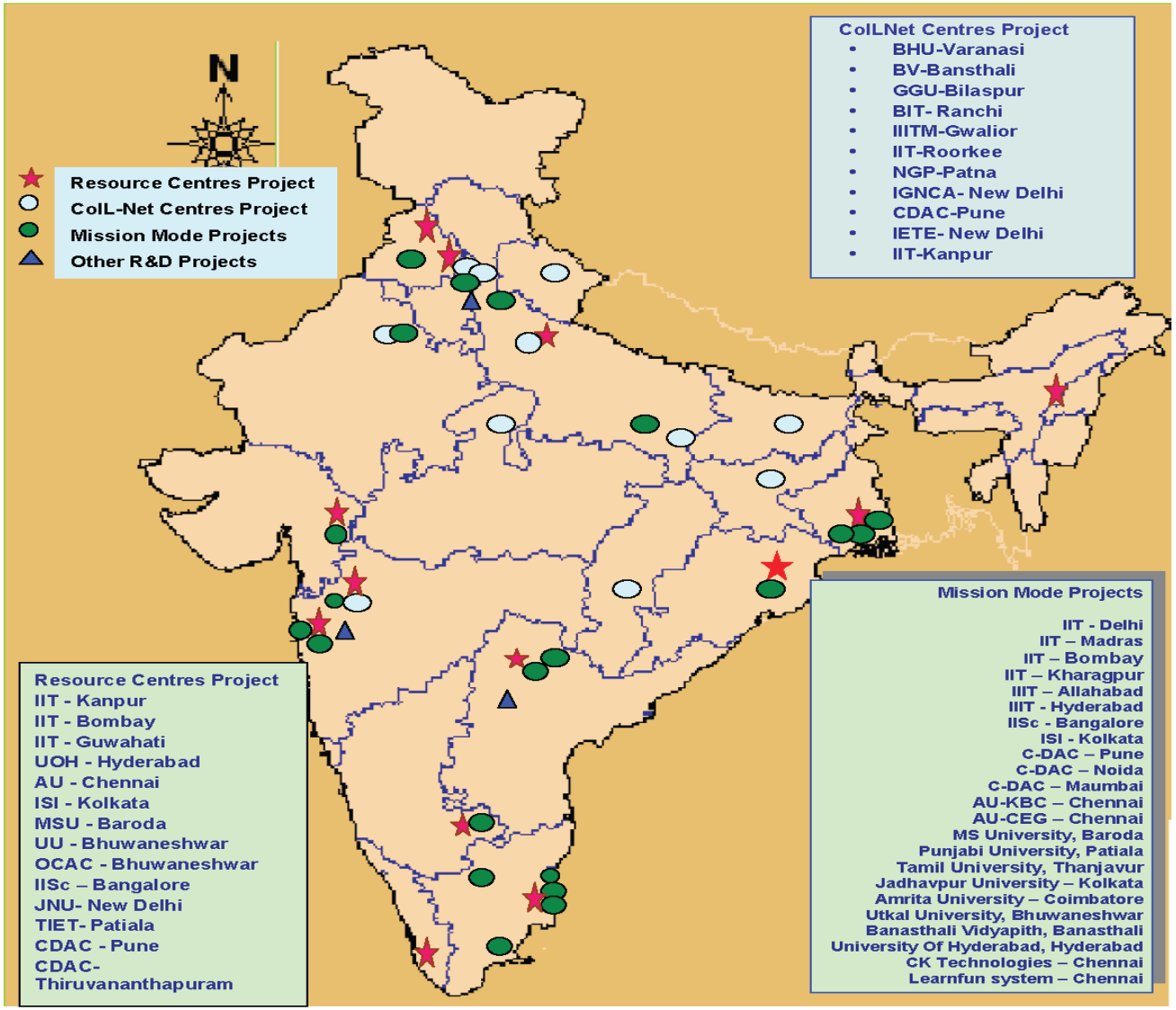
In the consortium mode 26 premier Institutes and R&D organizations are working together

- **English to Indian Languages Machine Translation (EILMT) → CDAC, Pune and others**
- **English to Indian Languages Machine Translation (AglaBharati) → IIT Kanpur and others**
- **Indian Language to Indian Language Machine Translation (ILILMT) → IIIT Hyderabad** and others
- **Sanskrit-Hindi Machine Translation (SHMT) → University of Hyderabad, JNU and others)**

Mission Mode Projects…

- **Document Analysis & Recognition System → IIT Delhi and others**
- **Online Handwriting Recognition System → IISc, Bangalore and others**
- **Cross-Lingual Information Access → IIT, Bombay**
- **Speech Corpora/Technologies → IIIT Chennai and others**
- **Indian Languages Corpora (ILCI) → JNU and others**

Institutions involved in Language Technology Development in India



The above figure is an indicative diagram of geographical spread of the projects supported by TDIL Programme

TDIL projects of near future

- Establishment of National Localization Research Centers (NLRC)
- Lexical Resource building
- Draft standards
- Young Investigator Program
- National Indian Languages Web Browser

Participation by private enterprise

- Industry
  - Microsoft
  - IBM
  - HP
  - Google
  - Other smaller players
- Societies/interest groups

Speech and OCR research groups

Speech research groups

- I.I.Sc. Bangalore
- I.I.T. Madras
- MSRI, Bangalore
- CDAC-Kolkata
- HP Labs
- Google
- ICSI-UC Berkeley
- IIT Guwahati

OCR research group

- I.I.Sc. Bangalore
- IIT Delhi
- Microsoft
- Univ. of Heidelberg (for Sanskrit)
- CDAC Noida
- IIT Kanpur

India’s LR needs in immediate future

- Language & cultural documentation
- e-governance
- Primary education and health
- Knowledge transfer and communication

Language & cultural documentation

- Documentation of fringe, neglected, and heritage languages
  - corpora (written and spoken) for minority and fringe languages
  - corpora (written and spoken) for classical languages important for heritage
  - corpora of digitized manuscripts

- MHRD, MCIT, Min of Culture, Min of S&T

e-governance

- **National e-Governance Plan (NeGP) vision**
  - Make all Government services accessible to **the common man in his locality**, through common service delivery outlets and ensure efficiency, transparency & reliability of such services at affordable costs to realize the basic needs of the **common man**

- NeGP started in 2006. comprises of 27 Mission Mode Projects (MMPs) and 8 components
- **Language Technology is a big component**
- Approximately **10,000 million USD** spread across 3 five-year plans

e-governance…

- land records
  - OCR (land transfers have been a major activity in rural India)
- handwriting samples of Indians
  - OHWR (difficulty in writing on computers)
- speech database
  - ASR/TTS (high illiteracy rates in rural India)
- names database
  - NER (cultural diversity in names)

Resources for e-governance

- database of judicial documents
  - Expert Systems/search/e-library (slow judicial system, language barrier)
- agri database of crop patterns, water tables, pests, nature of soil, climate changes
  - Expert systems (agriculture being the major activity in India)
- commodity prices databases
  - search engines (commodity buying/selling is major activity in rural India)
- localization database for popular software (English software will not sell)

Primary education and health

- Database of text books
  - e-library
- e-lessons
  - e-learning and LMS systems
- grapheme-phoneme text database
  - TTS
- corpora of the health domain
  - Expert systems/ translators
- database of Ayurvedic herbs and medicine system
  - expert system

Knowledge transfer and communication

- Translations from and to major languages
  - M(A)T (website, newspapers, publishing)
- Translations from major (including English) to minor languages
  - M(A)T (website, newspapers, publishing)
- Parallel corpora and dictionaries
  - M(A)T (domain specific translations)

Scope for International cooperation

develop/examine standards

- Examine suitability/extensibility of existing standards
- New standards
- Progressive involvement of global standard bodies

corpora development & data sharing

Corpora development

- A parallel corpora of major Indo European languages in specific domains (e.g. business, tourism, sports, culture)
- Spoken language corpora

Sharing

- Bilateral/multilateral agreements on linguistic data sharing

Business opportunities

India's advantages

- Low cost
- Good pool of trained linguists and computer scientists
- Un-paralleled linguistic diversity in **one** country
- Hindi is now a global language (inherits from Sanskrit – the oldest documented Indo-European language

Localization of software

- There is a huge market for it in India
- Build localization dictionaries for major European and Indian languages
- Evolve standards for localization

Conclusion

- India's diverse linguistic scenario, needs of literacy, mass education and social development demand multilingual resources and technologies
- Indian parliament has also committed to e-governance in local languages
- There is a political will, well planned constitution driven language planning and technology development program
- With low average literacy rate, only 4 % English users, the demand for multilingual, multimodal resources is expected to grow more
- A buoyant economy, low cost development, easy availability of quality human resource, India is certainly poised for a giant leap in language technology
- However, there are challenges for building such resources as per global standards. Therefore need for international cooperation
- There are tremendous opportunities for business.

धन्यवाद

Thank You