Language Technology Resources (LTR) for Sanskrit and other Indian Languages at Jawaharlal Nehru University, India

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
In this presentation...

- India’s diversity and LTR development by various government agencies, particularly the IT ministry
- Role of Sanskrit and its current status
- LTR for Sanskrit and other Indian languages at JNU
- Corpora development in India
- Issues & challenges
- Opportunities for collaboration and Business
India’s cultural & linguistic diversity
1 billion plus people, 2000 castes, 43000 sub-castes

Hinduism, Islam, Christianity, Sikhism, Jainism, Buddhism, Parsis etc

5 language families - Indo Aryan, Dravidian, Austro Asiatic, Tibeto-Burman, Andamanese

25 states in India, more than 1600 languages
Indian Language Families and % Speakers

- IndoAryan - 76.87%
- Dravidian - 20.82%
- Austro Asiatic - 1.11%
- Tibeto Burman - 1%
- Andamanese* - 0%
- Great Andamanese* - 0%
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 %)
National Languages & Scripts

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

- 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
Role of government agencies
- 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)
Ministry of Communication & IT

- TDIL - Technology Development for Indian Languages (TDIL) established in 1991
- TDIL 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 technologies thus developed for Indian Languages
- TDIL activities
  - 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
Activities by TDIL

- National Rollout Plan

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

  - The CD-ROM typically contains the following software tools:
    - Fonts, Keyboard Drivers, converters, editors, typing tutors
    - Integrated Word Processor
    - Bharateeya Open Office
    - Bilingual Dictionaries
    - Spell checker
    - Transliteration tool
    - Browser
    - Email Client
    - Messenger
    - Text to Speech system
    - OCR
TDIL – Mission mode projects

In the consortium mode, 26 premier Institutes and R&D organizations are working on LTR projects
Major ongoing TDIL projects

- English to Indian Languages Machine Translation (MT) System (E-ILMT) → CDAC, Pune
- English to Indian Languages Machine Translation (MT) System with Angla-Bharti Technology (E-ILMT-ABT) → IIT Kanpur
- Indian Language to Indian Language Machine Translation System: (ILMT) → IIT Hyderabad
- Sanskrit-Hindi Machine Translation (SHMT) → University of Hyderabad, JNU...
Major ongoing TDIL projects

- Document Analysis & Recognition System for Indian Languages (DARSIL) → IIT Delhi
- On-Line Handwriting Recognition (OLHR) → I.I.Sc, Bangalore
- Cross Lingual Information Access (CLIA) → IIT, Bombay
- Speech Corpora & Technologies (SCT) → IIIT Chennai
- Indian Language Corpora Initiative (ILCI) → JNU, New Delhi
The above figure is an indicative diagram of geographical spread of the projects supported by TDIL Programme.
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
- 42K crores INR (approx 10,000 million USD) spread across 3 five-year plans
Budget….

Lot of money available with different ministries

- TDIL (MCIT) ➔ Approx 2 million USD per year
- E-gov (MCIT and other ministries) in 22 Indian languages ➔ approx 9000 m USD total for 15 yrs
- E-learning and content creation in Indian languages (MHRD) ➔ approx 217 m USD total for 5 yrs
Sanskrit
India...

- An ocean of diversity
- Has had weak governance
- Has had invasions, destructions, colonization, partitions….
- Still going on (for last 5000 years)
- and has maintained un-interrupted tradition without having to use any violence
Role of Sanskrit

- Language with negligible speakers
- No state (Uttaranchal has recently accepted it as official language of the state)
- Yet spoken in all the states
- Mother of most of Indo Aryan languages (77% speakers of India)
- Source of India’s intellectual, religious, spiritual power
- Common source for vocabulary for Indian as well many European languages
- Source of much of Indian languages literature, art, culture
- Language of Hinduism, Buddhism, Jainism
- Source of Indian language grammars
Jawaharlal Nehru University (JNU), New Delhi
India’s best research university

- One of the best language programs in Asia for Indian and foreign languages
- Known for post graduate research
- Students selected based on South Asia wide test, 10:1 student:teacher ratio
- Fully residential
- Semester system with grading scheme
Special Center for Sanskrit Studies at JNU
- A center for traditional as well as modernized Sanskrit studies
- Wide variety of courses including linguistics, computational linguistics, grammar, logic, philosophy, literature etc
- Research students from Sanskrit, linguistics
- Large number of undergraduate students of foreign languages
- Heavily funded by Ministry of IT
- Consultancies
- Collaborations with foreign universities
- Rated 8+ (out of 10) by University Grants Commission
Our Funded Activities
Sanskrit-Hindi MT (SHMT) project – TDIL funded

- Members
  - University of Hyderabad
  - JNU
  - IIIT Hyderabad
  - Tirupati Vidyapeeth
  - Sanskrit Academy Hyderabad
  - Poornaprajna Vidyapeth Bangalore
  - Rajasthan Sanskrit University, Jaipur

- Budget → 3.16 crores
- Duration → 3 yrs (2008 – 2011)
Goals of SHMT project

- Develop necessary tools and data leading to Sanskrit-Hindi machine translation in the domain of children stories

- Multimedia and e-learning tools for children

- Deliverable → online/standalone system, data
Developments at JNU

- Rule based POS engine.
- annotated data for POS, Sandhi, samaasa
- Bilingual Dictionary
- Lexical Disambiguation Rules
- Standards for POS annotation
- Multimedia/e-learning standalone as well as web version
Our technology

- Online system
- Front end → Java, JSP, HTML, JS
- Backend → RDBMS, data files
  - Stored procedures, security features
- Standards → w3C, Unicode, TEI, EAGLES
- Connectivity → MS-JDBC driver
- Hosting → Tomcat/Apache webserver
Indian Languages Corpora Initiative (ILCI)
- Funded by TDIL program of Ministry of C & IT
- Approx 3 crores
- 2 years
Languages & Consortia partners

1. Hindi (and English) – Special Center for Sanskrit Studies, JNU (consortium leader)
2. Punjabi – Punjabi University, Patiala
3. Urdu – Center for Indian languages, JNU
4. Tamil – Tamil University
5. Telugu – Dravidian University
6. Malayalam – IIITM-K, Trivandrum
7. Gujarati – Gujrat University
8. Konkani – Goa University
9. Oriya – Utkal University
10. Bangla – ISI Kolkata
11. Marathi – IIT Mumbai
ILCI user-groups

- Various Indian language technology projects in the country, in particular Machine Translation projects in the tourism and health domain

- Language research groups in the country
ILCI deliverables

- Help evolve national Standards

- Parallel aligned corpora
  - 11+1 languages (Hindi as source)
  - Tourism & health domain
  - 50 k sentences in 12 languages

- Annotated corpora

- Tools
National standard for POS

- National Standard for POS annotation has been evolved recently by Bureau of Indian Standards (BIS)
- It is a generic hierarchical scheme considered adaptable for diverse Indian languages
Corpora collection

☐ Collect 50 k sentences in tourism & health domain in Hindi

☐ Create parallel aligned corpora in 11 Indian languages

☐ An annotated corpora of approx 6 million words
Other funded activities - Consultancies

- Microsoft Corp consultancy for Handwriting Recognition for Devanagari
- Microsoft Research Consultancy for Indic languages tagset and annotation
- CDAC consultancy for localization for software
Collaborations

- With University of Massachusetts, Dartmouth (Center of Indic Studies) for the project STAIT (Science and Technology in Ancient Indic Texts)

- MoU with University of Wurzburg, Germany for Sanskrit, Hindi, Linguistics, Computational Linguistics
Unfunded activities
J-TESS:
JNU Text Encoding & Search for Sanskrit
An unfunded initiative by the Computational Linguistics group at Sanskrit Center

Goals -

- Creating a searchable database of Sanskrit texts
- Search through Indian language scripts, IAST, ITRANS, Wx and other schemes
- Linking searches with scanned images, with metadata, introduction and other useful links
- Reading help like sandhi viccheda, morph analysis, lexical look up, translation
- Multimedia texts, e-learning tools for younger generation
- Tools
What has been done

- Texts
  - Vedas
  - Ramayana
  - Mahabharata
  - Dictionaries (Nighantu-Nirukta, Amara, Medini, Mankha, Halayudha)
  - Dictionary (Apte)
  - Many other lexical resources (as part of student projects)

- Tagging schemes, tagged corpora
Tools

- Sandhi splitter (vowel), consonant (shortly)
- Subanta (noun inflections) analyzer
- Tinanta (verb inflections) analyzer
- Kridanta (primary derived nouns)
- POS tagger
- Generators (sandhi, subanta, tinanta)
- SHMT (in progress)
- MM animation
- Lexical resources and search
R&D by research students

- As part of M.Phil and Ph.D research by my students, a large number of tools have been developed mainly for Sanskrit.
- Their tools and dissertations are online on our server http://sanskrit.jnu.ac.in
Corpora development at other places
- **Kolhapur Corpus of Indian English (KCIE)**
  - 1 million words of English
- **TDIL (1991-94)**
  - Approx 3 million words in major Indian languages (done by individual institutes)
LDC-IL project → CIIL Mysore

- General purpose corpora of approx 3 million words for many Indian languages
- Annotation work is continuing

Gyan Nidhi → CDAC Noida

- parallel aligned corpus of 13 Indian languages including English.
MSRI

- Multilingual Systems initiated IL-POST discussions by leading experts in the country (I was one of them)
- The IL-POST scheme is a hierarchical scheme based on EAGLES guidelines
- 200 K words tagged in Hindi, Bangla, Tamil
- 50 K tagged in Sanskrit
- **IIIT Hyderabad**
  - annotated corpora in 9 languages including Hindi
  - Simple flat tagging scheme with 26 tags

- **TDIL MT projects**
  - Many of the consortium projects have developed corpora for specific tasks, may or may not be tagged
Issues and Challenges
Lack of uniform standards, now they are coming up
Earlier projects suffered from transparency
Too many languages, not enough trained manpower for all languages
Manpower training projects are being planned
Scope for collaboration
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)
Collaborate to develop/examine standards

- Examine suitability/extensibility of existing standards
- New standards
- Progressive involvement of global standard bodies
Collaborate in corpora development & data sharing

Corpora development

- Low cost corpora development for English and Indian languages in specific domains (e.g. business, tourism, sports, culture)

- Spoken language corpora

Sharing

- Bilateral/multilateral agreements on linguistic data sharing
Collaborate in Localization of software

- There is a huge market for it in India
- Localized versions of major software applications has huge market in India Evolve standards for localization
Demo

Sanskrit.jnu.ac.in

localhost
Dhanyavaada

Thank you

girishjha@gmail.com