

# Indian languages corpora and application development at JNU

**Prof R N Shrivastava Memorial Lecture**

Kendriya Hindi Sansthan, Agra, 23<sup>nd</sup> March 2010

**Girish Nath Jha**

Associate Professor, Computational Linguistics  
Special Center for Sanskrit Studies, J.N.U., New Delhi

Mukesh & Priti Chatter Professor, History of Science  
Center for Indic Studies, University of Massachusetts, Dartmouth, USA

**Sanskrit**

# Goals

**□ Explore Sanskrit as a common connection between Indian languages**

**□ The theory of India as a single linguistic area is being followed as a broad guideline**

**□ Currently active courses offered for this purpose are**

**□ Structure and history of Sanskrit (M.A.)**

**□ Sanskrit and Indian languages (M. Phil.)**

# Goals

**□ Train students in the theory and technique of Computational Linguistics for enabling them to do useful R&D for Sanskrit**

□ Currently active courses for this purpose are

□ Computer Awareness (M.A.) - **compulsory**

□ Computer Application for Sanskrit (M.A.) - **compulsory**

□ Introduction to Computational Linguistics (M.A.)

□ Computational Linguistics Toolkit (M.A.)

□ Text processing & storage (MA)

□ Indian Theories of Knowledge and Computational Sanskrit (M. Phil.)

# Goals

- Take away our kids from popular media (read Disney/cartoon etc)**
  - Provide multimedia/e-learning content for kids
  - Translation tools for rendering online content into Indian languages

**funded activities**

# projects

1. Online Multilingual Amarakosha  
(UPOE, UGC grant)
2. Sanskrit-Hindi MT and  
Multimedia (DIT funded)

# Consultancies

- CDAC consultancy for localization for software
  - Dictionary of Sanskrit words for computer domain
- Umass consultancy/collaboration for STAIT

# **Un-funded R&D**

- **J-TESS**
- **noun analyzer (M.Phil.)**
- **verb analyzer (M.Phil.)**
- **POS Tagger (Ph.D.)**
- **Sandhi splitter and analyzer (M.Phil.)**
- **Karaka Analyzer (Ph.D.)**
- **Gender evaluator for SaHiT (M.Phil.)**
- **Sanskrit e-learning (Ph.D.)**
- **Derivational morphology analyzer ( Krt - M.Phil.)**
- **Mahabharata indexer (adiparvan - M.Phil.)**
- **AD search and indexer (M.Phil.)**
- **Sanskrit-Hindi Translation (SaHiT)**
- **Linguistic resources & corpora (MA course projects)**

# **Description of Current projects**

# The Sanskrit Consortium

- Members
  - University of Hyderabad (leader)
  - JNU
  - IIIT Hyderabad
  - Tirupati Vidyapeeth
  - Sanskrit Academy Hyderabad
  - Poornaprajna Vidyapeetha Bangalore
  - Rajasthan Sanskrit University, Jaipur
- Budget → 3.16 crores
- Duration → 3 yrs (2008 – 2011)

# goals

- 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

# Major tasks

## **UoHyd: (Dept of Sanskrit)**

- Developing Engines for Sandhi, samaasa, Lexical Disambiguation, Transfer Grammar.
- Developing Annotated data for Sandhi, Samaasa, Kaaraka
- Develop Bilingual Dictionary
- Develop Lexical Disambiguation Rules
- Develop Transfer Grammar rules/data
- Develop Standards for different levels of tagging

## **JNU: (Spl Center for Sanskrit Studies)**

- Developing Rule based POS engine.
- Developing annotated data for POS, Sandhi, samaasa and kaaraka
- Develop Bilingual Dictionary
- Develop Lexical Disambiguation Rules
- Develop Transfer Grammar rules/data
- Develop Standards for different levels of tagging
- Developing multimedia CD for distribution: standalone as well as web version

## **IIT-Hyd**

- Testing/evaluation/integration
- Parser for Sanskrit

# Developments so far (2 yrs)

## **JNU: (Spl Center for Sanskrit Studies)**

- 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

Bringing the heads together...

**4<sup>th</sup> International Sanskrit Computational  
Linguistics Symposium**

**(4i-SCLS)**

**at JNU**

**Dec 10-12, 2010**

**<http://sanskrit.jnu.ac.in/conf/4iscls>**

**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 development

# Similar Initiatives elsewhere

- Titus
- Brown University → NSF funding
- Clay library
- Digital library of India
- INRIA
- SanskNet project, Tirupati Vidyapeeth  
→ MHRD funding
- others

# J-TESS developments so far

- Textual search for
  - Vedas
  - Ramayana
  - Mahabharata
  - Koshas (Nighantu-Nirukta, Amara, Medini, Mankha, Halayudha)
  - Dictionary (Apte)
  - Many other lexical resources (as part of student projects)
- Tagging schemes, tagged corpora
- Tools

# **Hindi and other Indian languages**

# Indian Languages Corpora Initiative (ILCI)

## WHY ILCI ?

- India has 22 constituent languages
- We do not have enough labeled lexical resources to carry out computational research
- From being 'resource poor' to 'resource rich' language communities
- Applications
  - Statistical NLP
  - Machine translation
  - Cross-linguistic research
  - Multilingual lexical resource building

# The ILCI Consortium

1. Hindi (and English) JNU (Center of Sanskrit Studies) → leader of consortium
2. Tamil – Tamil University
3. Punjabi – Punjabi University Patiala
4. Telugu – Dravidian University
5. Malayalam – IIITM-K, Trivandrum
6. Urdu – JNU (Center of Indian Languages)
7. Gujrati – Gujrat University
8. Konkani – Goa University
9. Bangla – ISI Kolkata
10. Marathi – IIT Mumba
11. Oriya – Utkal University

# Cost & Duration

- 2.83 crores
- 2009-2011

# Our deliverables

- **Draft Standards**
- **Parallel aligned corpora**
  - 12 languages (Hindi as the source)
  - Tourism & Health domain
  - 600,000 sentences (50 k in 12 languages)
- **Annotated corpora**
- **Tools**

# Searchable corpora

- Make this corpora available in a searchable mode on a server at JNU (<http://sanskrit.jnu.ac.in> )
- Allow online editing by members
- Allow automatic indexing, KWIC, dictionary making etc

# Corpora editing by members

- Something like centralized code management systems like VSS, GNU-CVS etc
- They can select a sentence directly from the server, edit its translation/annotation and save it
- The system will record who changed what at what date&time

# Corpora annotation

- Agree on a common annotation scheme – hierarchical or flat for all Indian languages. Remember despite language families there is one linguistic area in India
- Training in data formats, standards, tools
- POS annotation
- Server based annotation

# Tools development

- Corpus annotation tool
  - Online vs standalone
  - data repository and maintenance
- KWIC identifier
- Stemmer
- Affix list builder
- Frequency list builder
- Named Entity lists builder

**consultancies**

- Microsoft Research Consultancy for Indic languages tagset
  - A hierarchical scheme for all Indic languages.
- Microsoft Corp consultancy for Handwriting Recognition for Devanagari
  - implications for OCR of handwritten manuscripts
- ‘Hindi sangrah’ – database of Hindi varieties (MG Hindi Univ consultancy)

# R&D for Hindi and other Indian languages

- Hindi POS tagger based on MSRI tagset
- Hindi homonym marker
- Andamanese verb analyzer
- Oriya scrambling
- Hindi spellchecker
- Hindi POS evaluation
- Hindi Verb group identifier and analyzer
- Hindi grammar checker
- Politeness analysis in Hindi
- Magahi verb analyzer and generator
- Multilingual javascript keyboard
- Multilingual text editor

**Our real strength  
is our students**

# Current Research Students (PhD)

- Priti → e-learning
- Subash → Ontology (JNU & CDAC)
- Muktanand → wordnet (JNU & IIT Mumbai)
- Manji Bhadra → KnowledgeBase (JNU & BORI)
- Sachin → NER
- Surjit → Semantics (Amarakosha)
- Diwakar Mani → MT Divergence (JNU & CDAC (planned))
- Diwakar Mishra → TTS (JNU, MSR, IISc Bangalore planned)
- Narayan → chunking (Linguistics, JNU)
- Sanjeet → Scrambling (Linguistics, IIT Kanpur & JNU)
- Ritesh → Politeness (Linguistics, JNU, planned)

# **Current Research Students (M.Phil.)**

- Diwakar Mishra → Sandhi viccheda
- Prakash Kamble → Hindi-Marathi Homonyms  
(CIL, JNU)
- Mukesh Mishra → Amarakosha

**...and a large number of MA students  
from Sanskrit and Linguistics**

# Data encoding/annotation standards

- RDBMS/data files
- Looking to go for TEI standards, XML etc for reasons of portability, global mapping and uniform standards
- Unicode
- IL-POSTS scheme (follows EAGLES guidelines)

# 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

# How do we do what we do

- Rule based
  - But we will need some statistical help..  
Working in those..
- Lexical interfacing: lexicon, lists
- Annotated lexical resource
- [Sample](#)

# Demo

- [JNU Sanskrit website](#)

Please visit <http://sanskrit.jnu.ac.in>

Feedback ([girishjha@gmail.com](mailto:girishjha@gmail.com)) will be appreciated