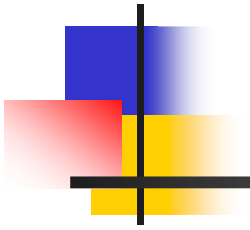


Language Processing and Corpus Linguistics

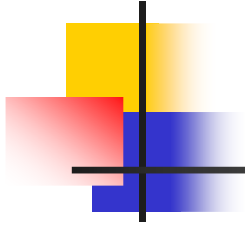


Prof R N Shrivastava Memorial Lecture
Kendriya Hindi Sansthan, Agra, 22nd March 2010

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



Language Processing



Communication

- **Machine-Machine Communication**
 - Artificial Language (AL)
- **Human-Human Communication**
 - Natural Language (NL)
- **Human-Machine Communication**
 - AL commands + Mouse/KB → very difficult, one has to learn commands, operating system behavior etc



Human Language Processing

- Is it possible to use NL for HM Communication ?
- The area of study that aims to make it possible is Computational Linguistics (COLING/ HLP/NLP/ Language Technology/ HCII/AI)



What is required?

- Machine must know a human language (KOL)
- Machine must know its valid usage in the real world (KOW)
- We have to build 'intelligent' machines
- Model human cognition and language behavior

Areas of R&D under NLP



- **Natural Language Analysis/Understanding (NLA/NLU)**

language → computer → understands/parses

- **Natural Language Generation (NLG)**

linguistic instructions → computer → language

- **Speech Synthesis (SS or TTS)**

written language → computer → speech

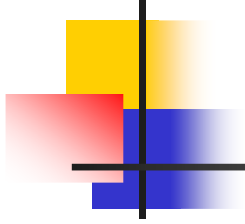
- **Speaker Recognition (ASR)**

spoken language → computer → recognizes speaker

Areas of R&D under NLP



- **Natural Language Interface (NLI)**
computer software operated by natural language
- **Machine Translation (MTS)**
language1 → computer → language2
- **Computational Lexicography**
Specific lexical databases for end-user access or for other software



Indian Language Computing



What is IL computing?

- Input/output mechanism in computer
- Text editors
- Spell checkers
- Grammar checkers
- Language processing
- Translation
- Standards and localization/internationalization



Input/Output mechanism

- Alphabet standardization
 - Govt of India efforts
- Unicode/font issues
 - TDIL efforts
 - JNU workshop
- Issue of pure consonant / syllabic consonant
 - Conjuncts, geminates, compound letters have to be formed by halanta infixing
- Hindi-Urdu specific letters, Sanskrit, Vedic Sanskrit, Marathi etc
- New languages using Devanagari



Unicode examples

- Writing in 'bare' unicode
 - Example
- Using unicode editors
 - Baraha
 - MS word
 - Other 'office' applications
- Hindi on web



Key-board layouts

- Problem with Qwerty KB
- Type writer format
- Phonetic writing
- Other formats
- Problem with multilingual documents



Other input mechanisms

- Handwriting
- OCR
- Speech (ASR)
- Gesture



Other output techniques

- Voice (TTS)
- Gesture



Language processing

- Word level
 - Spell checkers
 - MS word
- Sentence level
 - none
- Discourse level
 - none



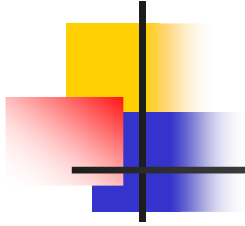
Translation

- To Hindi
 - Angla Bharati technology
 - Shakti translation system
 - Matra/Mantra
 - SaHiT
- From Hindi
 - Anu Bharati



Paucity of online content in Hindi

- Initial problems of Hindi typing
- Bilingual/multilingual data
- Fonts
- Now the basic tools are available
- Awareness
- Determination to use Hindi



What needs to be done



What needs to be done

- Evolve standards for Indian languages
- 22 national languages
- Numerous other languages
- Indo Aryan → 76.87%
- Dravidian → 20.82 %
- Hindi → 42% approx



What needs to be done

- The case of Hindi
- Dialect Dictionary (KHS project)
- Hindi's varieties → "Hindi Sangrah"
(MGAHU project)
- Sample data



What needs to be done

- Need for tag set

- Overview of tagging in Hindi
 - IIIT Hyderabad efforts
 - TDIL initiatives
 - MSRI
- Sanskrit tags
 - CSS, JNU tagset
- Sanskrit based tagset for Indian languages

- Need for corpus collection

- Need for nation-wide effort in creating tagged corpus
- Annotated speech corpora
- Names database
- Hindi slang database
- Fixed phrases / usages
- W3C standard for e-corpus collection?



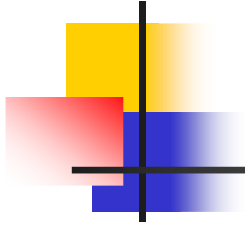
What needs to be done...

- Lexicons
 - General purpose
 - Domain specific
 - Ontological
- Grammar modeling
 - Generative models
 - Unification based models
 - Paninian karaka based models
 - Conceptual semantics/ Navya Nyaya based models



What needs to be done...

- Dialect dictionaries
- Language and usage modeling
 - numeral ranges
 - Measurements
 - date/time
 - Currency
 - Phone
 - person/place names
 - Addresses
 - pin codes
- Script grammars



Need for Corpora



Why have corpora at all?

- Rule based NLP
 - Innatist
 - Chomskyan???
- Data based NLP
 - Behaviorist???
 - statistical



Corpora

- What is it?
- Usage data in a domain and context
- its analysis
- Its abstraction, cross lingual parallelism
- Its modeling
- Evolving statistical models and algorithms
and
- Tools to manage, access



Corpora – various kinds

- Standard
 - Idealistic .. Do we have an ideal behavior?
 - Do only models behave ideally?
- non standard
 - Performance errors?
 - So, do we introduce correction?



Corpora – various kinds

- Corpora
 - they are generally monolingual
 - they have natural / minimally artificial data
- Natural
 - Speech/gesture corpora → text corpora
- Artificial
 - normalized
 - Corrections introduced (spellings, grammar, even style)
 - Translated (parallel corpora)
 - Monolingual corpora which is a translation



Corpora – various kinds

- Corpora
 - Unimodal
 - text
 - Bimodal
 - Text and speech
 - Multimodal
 - Text, speech and gesture



Corpora – plain or annotated

- Annotation
 - Do we need it?
 - What to annotate
 - How much
 - Coarse-grained
 - Fine grained
 - By how many annotators



Annotating Text Corpora

- POS
- morph
- Phrase
- Sentence
- Discourse
- Entire context
- Tier based architecture or all together
- Separation of info or merging



Annotating Speech Corpora

- Phoneme
- Syllable
- Supra-segmentals
- Complete Prosody



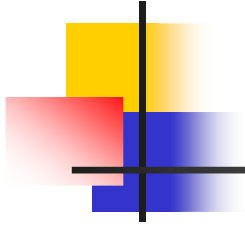
Annotating multimodal corpora

- We are not there yet
- What could be done
 - Evolve culture/race/gender/age specific gesture to text translation standards (Motorola project at UIUC)
 - Tag specific scenes with corresponding text transcriptions
 - Weave in speech-text transcriptions



How good are they?

- Use of corpora
 - Building dictionaries
 - Updating dictionaries
 - Linguistic analysis
 - NLP/AI
 - Information extraction
 - KWIC
 - Disambiguation
 - Machine learning



How to build good corpora

How to build good corpora?



- Corpora Collection Methodology (CCM)
- Most ideal method
 - non-monotonic data acquisition as humans do

 - We will have to first build an intelligent machine
 - Corpora for machine-intelligence and vice versa
 - Compares with newborn human-child with capacity to learn
 - This 'capacity' (LAD??) is certainly NOT data-deprived
 - So we have a confused situation...
 - This is because our knowledge of how we do things is not complete yet



Methodologies...

- Other more 'real' methods
 - Language documentation
 - SOAS, London
 - SOROSORO, Chirac Foundation, Paris
 - ASI, India
 - CIIL, India
 - Prof Anvita Abbi
 - Prof Vaishna Narang
 - Others...



Methodologies...

- Linguistic fieldwork
 - Informant must be a bilingual
 - or get help from another native bilingual
 - How natural is the data?
 - The informant is conscious of the 'outsider'
 - The linguist (fieldworker) may unconsciously introduce corrections
 - may frame the questionnaire based on what helps him/her in getting the desired rules/patterns
 - Certain other important aspects may be missed



Methodologies...

■ Current Best Practices

- prepare appropriate questionnaire/topics (for speech data)
- Train in suitable recording equipments (for speech data)
- Preparing a consent form for informants/data source persons/institutions.
- Emphasize on copyright and metadata issues.
- prepare a list of potential institutes/individuals which can contribute data
- call them to set up prior appointment



Methodologies...

Current Best Practices...

- be presentable, courteous and explain the purpose of the task to the data source
- obtain data, make sure the proper citations and meta data are recorded (in writing or speech)
- make sure the printed data is properly catalogued, grouped and labeled so that it is easier for the data entry person to key in all the relevant information about the data
- make photocopies if the safety of the data is not guaranteed



Corpora Encoding Standards (CES)

- How do we encode and markup the corpora
- Encoding
 - Unicode
 - ASCII based schemes
- Markup Standards
 - XML
 - SGML
 - Legacy standards like SSF



Corpora Annotation Standards (CAS)

- Often the most debated aspect of corpora building
- There are various standards/recommendations
- TEI
- EAGLES
- others



POS annotation

- Hierarchical
 - The parent-child (dominating-dominated) relations between categories is maintained
 - Flexible
 - Easily scalable
 - More tags
 - fine grained
 - Desired results from machine will require larger corpora



POS annotation

- Flat

- Based on 'as is' POS categories
- Easier to tag data (if the tagset is smaller)
- Generally coarse grained
- Smaller corpora needed to train the machine
- A single tagset may not apply on dissimilar languages
- India has five language families



Corpora Validation Standards (CVS)

- external validation
- Internal



Major corpora

■ Outside India

- **The BAF corpus of English-French Bitext** - Approximately 450,000 words in each language
- **INTERSECT** (International Sample of English Contrastive Texts) contains parallel bilingual corpus of French and English
- **The Lingua Project** - Multilingual corpora in six European languages for language pedagogy purposes
- **TRIPTIC** - Trilingual Parallel Text Information Corpus (English, French and Dutch) – 2000,000 words, with aligned paragraphs

■



Major corpora – outside India

- **Translation Corpus of English and German** - Includes EU materials, academic text books, modern fictions, tourist brochures (500,000)
- **JS-ELAN (Slovene-English Parallel Corpus)** - 1 million words from Slovene-English and English-Slovene texts
- **MULTEXT/MULTEXT EAST**

The LRE project MULTEXT is one of the largest EU projects in the domain of language tools and resources
- Prominent corpora organizations like -BNC, LDC, ECI, ELRA, FLaReNet



Initial projects (Dash 2003)

Language	Agency	Duration	Words
Hindi English Punjabi	IIT Delhi	1991-1994	3 million
4 Dravidian languages	CIIL Mysore	1991-1994	Do
Marathi Gujrati	Deccan College	1991-1994	Do
Oriya Bangla Assamese	IIALS Bhubaneshwar	1991-1994	do
Sanskrit	SSU, Varanasi	1991-1994	Do
Urdu Sindhi Kashmiri	AMU	1992-1994	do



Other projects (Dash 2003)

- IIT Kanpur → Hindi Nepali
- IIT Bombay → Marathi Konkani
- IIT Guwahati → Assamese, Manipuri
- IISc Bangalore → Kannada, Sanskrit
- ISI Kolkata → Bangla
- JNU → Sanskrit
- HCU → Telugu
- Anna Univ → Tamil
- MSU Baroda → Gujrati
- Utkal Univ → Oriya
- Thapar Instt Patiala → Punjabi
- ERDCI Trivandrum → Malayalam
- CDAC Pune → Urdu, Sanskrit, Kashmiri



Other projects

- CDAC-Gyan-nidhi
- LDC-IL (CIIL Mysore)
- IL-MT projects
- KHS- Hindi corpora
- ILCI – Consortium (JNU with other partners)



Thank you

Please send feedback to
girishjha@gmail.com
<http://sanskrit.jnu.ac.in>