The TDIL program and the Indian Language Corpora Initiative (ILCI)

Girish Nath Jha
Special Center for Sanskrit Studies
Jawaharlal Nehru University
New Delhi-110067
girishjha@gmail.com

Abstract

India is considered a linguistic ocean with 4 language families and 22 scheduled national languages, and 100 un-scheduled languages reported by the 2001 census. This puts tremendous pressures on the Indian government to not only have comprehensive language policies, but also to create resources for their maintenance and development. In the age of information technology, there is a greater need to have a fine balance between allocation of resources to each language keeping in view the political compulsions, electoral potential of a linguistic community and other issues. In this connection, the government of India through various ministries and a think tank consisting of eminent linguistics and policy makers has done a commendable job despite the obvious roadblocks. This paper describes the Indian government's policies towards language development and maintenance in the age of technology through the Ministry of HRD through its various agencies and the Ministry of Communications & Information Technology (MCIT) through its dedicated program called TDIL (Technology Development for Indian Languages). The paper also describes some of the recent activities of the TDIL in general and in particular, an innovative corpora project called ILCI - Indian Languages Corpora Initiative.

1. The linguistic scene in India

India has a very complex and peculiar linguistic situation. There are 4 language families (Baldridge 96) – Indo Aryan (76.87 % speakers), Dravidian (20.82 % speakers), Austro-Asiatic (1.11 %), and Tibeto-Burman (1%) . These have 22 constitutionally recognized (scheduled) languages out of which Hindi has the 'official' status in addition to having the 'national' status. English which is not a national language of India has the status of 'associate official' language. Besides these, India has 100 mother tongues reported by the recent census (2001), and many more (running up to 1000) documented languages and dialects. A new language family called 'Andamanese' has been recently discovered (Abbi 2001), and the possibility of another – the 6th family called 'Great Andamanese'– is very likely. Of the major Indian languages, Hindi is spoken in 10 states of India with a total population of over 45 % followed by Telugu and Bangla. Not only the languages, there are multitude of scripts as well. India has more than 18 scripts in India which need to be standardized and supported by technology.

2. Constitutional provisions and the language policy of India

Indian has currently 25 states and 7 union territories (UTs). The Indian constitution adopted in 1950 lists 14 languages as scheduled languages of the union. In the schedule called THE EIGHTH SCHEDULE: (Articles 344 (1) and 351), the following languages are listed –

- Assamese
- Bengali
- Gujarati
- Hindi
- Kannada
- Kashmiri
- Konkani
- Malayalam
- Manipuri
- Marathi
- Nepali
- Oriya
- Punjabi
- Sanskrit
- Sindhi
- Tamil
- Telugu
- Urdu
- Maithili
- Bodo
- Santhali
- Dogri

The Indian states and UTs have exclusive rights on their regional languages. However if the language of the state is also listed as scheduled language (as above), then the union has a constitutional obligation to promote the language. Hindi, besides being a scheduled (national) language with 21 other languages, is also the official language of the union with English as its associate. Hindi is the official language of 10 out of 25 Indian states and spoken by more than 42% of Indian population.

Each of the 25 states and 7 UTs in India can have its official language (one from the 22 listed above) and several other minor languages the speakers of which have a fundamental right to maintain and promote these languages. The

constitution of India through following sections tries to do a delicate balancing act in the vast diversity of languages – (Language in India, 2002)

ARTIII: FUNDAMENTAL RIGHTS: Cultural and Educational Rights

- Protection of interests of minorities
- Right of minorities to establish and administer educational institutions

PART IVA: FUNDAMENTAL DUTIES

PART XVII: OFFICIAL LANGUAGE: CHAPTER I -LANGUAGE OF THE UNION

- Official language of the Union
- Commission and Committee of Parliament on official language

PART XVII: CHAPTER II - REGIONAL LANGUAGES

- Official language or languages of a State
- Official language for communication between one State and another or between a State and the Union
- Special provision relating to language spoken by a section of the population of a State

PART XVII:CHAPTER III - LANGUAGE OF THE SUPREME COURT, HIGHCOURTS

- Language to be used in the Supreme Court and in the High Courts and for Acts, Bills, etc
- Special procedure for enactment of certain laws relating to language

PART XVII: CHAPTER IV.-SPECIAL DIRECTIVES

- Language to be used in representations for redress of grievances
- Facilities for instruction in mother-tongue at primary stage
- Special Officer for linguistic minorities
- Directive for development of the Hindi language

The Indian government through various acts and policies tries to implement and enforce constitutional provisions.

3. Language promotion and maintenance by Ministry of HRD

The Ministry of Human Resource Development (MHRD) through its nodal agency called Central Institute of Indian Languages (CIIL) has a systematic program to maintain and promote Indian languages. The CIIL was established at Mysore, Karnataka to co-ordinate the development of Indian Languages, to bring about the essential unity of Indian languages through scientific studies, promote inter-disciplinary research, contribute to mutual enrichment of languages, and thus contribute towards emotional integration of people of India (CIIL website, 2010). The mandate of CIIL is given as –

- advise and assist central as well as state governments in the matters of language.
- contribute to development of all Indian languages by creating content and corpus.
- protect and document minor, minority and tribal languages
- promote linguistic harmony by teaching 20 Indian languages to non-native learners

Among some of the newer initiatives of the CIIL are

- New Language Survey of India (NLSI)
- National Translation Service
- Linguistic Data Consortium of Indian Languages (LDC-IL)
- Development and promotion of minor Indian languages
- National Testing Mission
- Development of Pali

4. The TDIL program of the MCIT

The MCIT started a program called TDIL in 1991 for building technology solutions for Indian languages. The stated objective of the TDIL is (i) to develop information processing tools and techniques, (ii) to facilitate human-machine interaction without language barrier, (iii) to create and access multilingual knowledge resources and integrate them to develop innovative user products and services. Among the major activities of TDIL have been –

Basic software tools for Indian languages (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 basic software tools for enabling the linguistic community in the digital age

Ongoing Language technology/corpora projects in the consortium mode

- 26 premier institutes and R&D organizations are working together on projects to develop the advanced technologies & applications.
- Development of English to Indian Languages Machine Translation (MT) System: (CDAC, Pune) o Development of English to Indian Languages Machine Translation (MT) System with Angla-Bharati Technology: (IIT Kanpur)
- Development of Indian Language to Indian Language Machine Translation System: (IIIT Hyderabad)
- Sanskrit-Hindi Machine Translation: (University of Hyderabad, JNU…)
- Development of Robust Document Analysis & Recognition System for Indian Languages: (IIT Delhi)
- Development of On-line handwriting recognition system: (I.I.Sc, Bangalore)
- Development of Cross-lingual Information Access (IIT, Bombay)

- Speech Corpora/Technologies: (IIT Chennai)
- Language Corpora (ILCI) : (JNU)

5. Indian Languages Corpora Initiative (ILCI)

ILCI (Indian Languages Corpora Initiative) started by Technology Development for Indian Languages (TDIL) program of Ministry of Communication and Information Technology (MCIT) for building parallel corpora for major Indian languages including English. The languages in the current phase include 8 languages from the Indo Aryan family (Hindi, Bangla, Punjabi, Oriya, Marathi, Gujrati, Konkani and Urdu), 3 languages from the Dravidian family (Tamil, Telugu, Malayalam) and English - the Indo Germanic language. The second phase is expected to include the remaining 11 major Indian languages (India has 22 constituent or national languages plus English as the associate official language). The main objective of the project is to build annotated parallel corpora in the domain of tourism and health with Hindi as the source language. The corpora encoding and annotation will be as per global standards which are currently being examined. The resource thus generated is supposed to feed in various MT and other LT projects in the country The following table lists the consortium partners, languages they are working on and the university/institute where the project is being carried out (name of the consprtium partner, language/s beng worked on, host institute)

- Girish Nath Jha, *Hindi, English*, J.N.U. New Delhi
- Sumanpreet Virk, *Punjabi*, Punjabi University Patiala
- Mazhar M Hussain, *Urdu*, J.N.U. New Delhi
- Niladri Shekhar Dash, *Bangla*, Indian Statistical Institute, Kolkata
- Gopal K Dash, *Oriya*, Utkal University, Bhubaneshwar, Orissa
- Malhar A Kulkarni, *Marathi*, IIT Mumbai, Mumbai
- Kirtida S. Shah, *Gujarati*, Gujarat University, Ahmedabad
- Jyoti D. Pawar, *Konkani*, Goa University, Goa
- S. Arulmozhi, *Telugu*, Dravidian University, Kuppam, AP
- S. Rajendran, *Tamil*, Tamil University, Thanjavur,Tamil Nadu
- Elizabeth Sherly, *Malayalam*, IIITM-K, Trivandrum,Kerala

5.1 The ILCI deliverables

The project has the following deliverables –

draft standards

- to evaluate existing current standards and to evolve a common corpora standards for all Indian languages. These include – Corpora Collection, Corpora Encoding and Markup, Corpora Annotation, Corpora Validation and Tools

Parallel aligned annotated corpora

- in 12 Indian languages including English with Hindi as the source
- Domain of usage – tourism and health
- 600,000 sentences (50,000 in each languages). On an average of 10 words in a sentence would mean a corpora of 60,000,00 annotated words

Tools

- Corpus annotation tool
- KWIC identifier
- Stemmer
- Affix list builder
- Frequency list builder
- Named Entity lists builder

5.2 Methodology and current status

The ILCI project is based on sound research methodology for Corpus Collection, Corpus Encoding and Marking, Corpus Annotation, Corpus storage, editing and search. The corpora will be hosted on a server with centralized code management systems like CVS. The consortium partners will be able to 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. The current developments so far include evaluation of corpora standards in India, parallel data development in 12 Indian languages, setting up of a centralized server and communication channels. Currently a discussion on evolving a national standard of linguistic annotation is going on under the banner of Bureau of Indian Standards (BIS). The next step will be annotation training to the linguists in each language group. This will be followed by corpora annotation and validation.

6. Conclusion

India has a distinct edge in linguistic diversity. The government through its various agencies backed by Indian constitution does a difficult balancing act in managing this diversity. There is a huge need in the country to build resources and tools for Indian languages, however due to a paucity of standard tagged lexical resources, the progress has been slower. The recent activities of the TDIL have been directed towards bridging this gap. The CIIL also has been its resources towards resource building, however at this point, there seems to be a focus on developing raw corpora for the language community than the annotated corpora which is needed by the language technology community. The ILCI project in this context becomes very important, as it can learn from the previous experiences and try to develop corpora based on global standards which can be used by both the language community as well as the language technology community.

7. References

Abbi, A., 2001, *A Manual of Linguistic Fieldwork and*

Structures of Indian Languages, Lincom Europa, Muenchen, Germany

Baldridge, J. (1996). *Reconciling Linguistic Diversity: The History and the Future of Language Policy in India*. University of Toledo Honors Thesis

EAGLES, 1996, *Recommendations for the Morpho-syntacticAnnotation of Corpora*, EAG – TCWG -- MAC/R

Language in India, 2002, *Indian constitution on languages and language policy,* Vol 2:2, April 2002

TEI, 1996, *Corpus Encoding Standard - Document CES 1. Part 0*. Version 1.

Thompson, P., 2004, *Developing Linguistic Corpora: a Guide to Good Practice Spoken language corpora*, University of Reading