India’s language diversity and resources of the future: challenges and opportunities

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

Abstract

The paper talks about India’s delicate balancing act in maintaining its huge linguistic diversity and planning for the future language technology resources and requirements. The paper highlights the challenges and opportunities for the government and the language technology community.

The paper begins with language situation in India and the constitutional provisions for maintaining and promoting Indian languages. The next section talks about the efforts by the Ministry of Human Resource Development (MHRD) and its agency for language development called CIIL (Central Institute of Indian Languages) and the Ministry of Communications and Information Technology (MCIT) with its agency TDIL (Technology Development for Indian Languages) for developing Indian languages through its various programs and initiatives. In this context, the paper talks about some major ongoing and past projects relating language corpora and technology.

The main focus of the paper however is the future language resources that Indian would need to build technologies in the key sectors of social development, governance, agriculture, education and health. In this context, the paper discusses the key technologies that need to be developed in each sector and the language resources that will be needed.

Introduction

India has 4 language families – Indo Aryan (76.87 % speakers), Dravidian (20.82 % speakers) being the major ones. These families have contributed 22 constitutionally recognized (‘scheduled’ or ‘national’) languages out of which Hindi has the ‘official’ status in addition to having the ‘national’ status. Besides these, India has 234 mother tongues reported by the recent census (2001), and many more (more than 1600) languages and dialects. Of the major Indian languages, Hindi is spoken in 10 (out of a total of 25) states of India with a total population of over 60 % followed by Telugu and Bangla. There are more than 18 scripts in India which need to be standardized and supported by technology. Devanaagari is the largest script being used by 6 languages.

Indian languages are under the exclusive control of respective states they are spoken in. Therefore every state may decide on measures to promote its language. However, since these 22 languages are national (constituent) languages, the center (union of India) also has responsibility towards each of them, though it has certain additional responsibility towards Hindi which is national as well official language of the Indian union. From time to time, minor/neglected languages claim constituent status. The situation becomes more complex when such a language becomes the rallying point for the demand for a new state or autonomous region.

This complex linguistic scene in India is a source of tremendous pressure 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.
Language promotion and maintenance by the Ministry of HRD

The MHRD through its language agency called CIIL was set up with the following goals –

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

Some of the ongoing initiatives of the CIIL are –

- 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) by the MCIT

The TDIL program of MCIT was started with the objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier; creating and accessing multilingual knowledge resources; and integrating 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)
- 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: (IIISc, Bangalore)
- Development of Cross-lingual Information Access (IIT, Bombay)
- Speech Corpora/Technologies: (IIIT Chennai)
- Language Corpora (ILCI) : (JNU)

Language Resource of the future

India has complex needs in terms of language technology resources. Among others, the following areas are expected to be the focus in future –
Language & cultural documentation
Indian has many undocumented languages, which may become extinct in the digital age. Many others may be documented but on the fringe and neglected. Such languages and those which are important for cultural heritage need to be comprehensively documented and preserved
- corpora (written and spoken) for minority and fringe languages
- corpora (written and spoken) for classical languages important for heritage
- corpora of digitized manuscripts

e-governance
E-governance is going to be a major activity with approximately 42,000 crore Indian rupees (10 billion USD) worth of budget from the government spread across 3 five-year plans. The language resources that would be required to be built are:
- land records Ô OCR (land transfers have been a major activity in rural India)
- handwriting samples of Indians Ô OHWR (difficulty in writing on computers)
- speech database Ô ASR/TTS (high illiteracy rates in rural India)
- names database Ô NER (cultural diversity in names)
- database of judicial documents Ô Expert Systems/search/e-library (slow judicial system, language barrier)
- agri database of crop patterns, water tables, pests, nature of soil, climate changes Ô Expert systems (agriculture being the major activity in India)
- commodity prices databases Ô 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 Ô e-library
- e-lessons Ô e-learning and LMS systems
- text readers Ô TTS
- corpora of the health domain Ô Expert systems/ translators
- database of Ayurvedic herbs and medicine system Ô expert system

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

Conclusion
The language technology and resources are of critical importance in present India. They will be more so in the future. Though there are tremendous challenges for building such resources as per global standards, there are immense opportunities for business. Cost-wise, it is cheaper to develop language resources and technologies in India. With over one billion people, emerging economy and only 4% people knowing English, one can very well imagine the market potential for non English resources and tools in India and the profit margin therein.