Overcoming the Language Barrier

India is the second largest in population in the world with one billion population. There are 18 constitutional languages with 10 scripts and over 1650 dialects. Development of the nation with such diversity depends on acquiring, absorbing and communicating knowledge seamlessly. Information Technology (IT) has emerged as an enabling technology in reducing the knowledge gap across different linguistic groups encompassing over 95% of India's population that is not English-literate. It is, therefore, necessary that people should be able to use computers and other IT systems in there own languages and derive benefits of enhanced productivity and better quality of life.

National excellence in the millennium shall be determined by the extent to which the Information Technology can deliver its potential in Local Languages. In a country like India, communication overcoming language barrier is crucial to the growth of society and in preventing the Digital Divide.

The first step in this direction was the launch of TDIL (Technology Development for Indian Languages) Programme in 1991 by MIT to develop information processing tools to facilitate human machine interaction in Indian Languages and to create and access multilingual knowledge resources. The next milestone has been the setting up of thirteen Resource Centres for Indian Language Technology Solutions. These centres will develop technologies for providing solutions with citizen interface in Indian languages selectively and thus covering all Indian languages. The centres will also disseminate these technologies through closer interaction with agencies in State Government, Industry and Academia.

Achievements

- The tagged corpora of texts in machine-readable form have been developed. This is useful as a basic research facility for linguists and computer scientists along with Tools for word level tagging, Word Count, Letter Count, Frequency Count, Spell checkers in various Indian Languages. [CIIL, Mysore]
- Computer Courseware in Hindi for DOEACC 'O' level courseware in machine-readable form is also developed and is being put on the web. [BV]
- Content creation in Electronic form Tagged corpus of Hindi, Hindi Vishwakosh, UN selected countries dictionary, Bharat Bhasha Kosh, SAARC dictionary, English to Hindi dictionary, Sanskrit to Hindi dictionary, and Bilingual (English, Hindi) IT terminology is under development. [ER&DC/N-CSTT]
- A Heritage Web site containing traditional Indian texts centered around the 'Upanishads and the Bhaagwadgita' is also hosted. [IITK]
- Java based Solutions for displaying Web Documents through Negotiation and Dynamic Rendering have been developed wherein client need not specially install any fonts or software on his system. [IITK]
- Hindi Search Engine for indexing and searching of Devanagari HTML documents for Linux platform has been developed. [IITK]
- CD Authoring Tools for Indian Language Documents has been developed. The development of Indian Language CD Publishers toolbox, 'site management' tools and searches integrated with a dictionary are underway. [C-DAC]
- Web based multilingual e-mail Solutions using Active -X provides a facility to type the text in Hindi language for sending an e-mail in Hindi which gets converted into HTML format. [C-DAC]
- Multi-lingual e-mail Client has also been developed. It's working prototype facilitate the clients for sending and receiving e-mails in Hindi without having need to have Internet connection provided sender and receiver both have this s/w. [CMC]
- Hindi Bulletin Board System is under development. This web-based application allows users to create topics for discussion and maintains threads within a topic. [IITK]
- Sanskrit word processor is under development, which will even handle special Sanskrit constructs. [C-DAC/B]
- Sanskrit Authoring System including a Sanskrit word processor for use by Sanskrit scholars in text processing etc. is being developed. [C-DAC/B]
- Desika Software package is a Natural Language Understanding System for Sanskrit. This software incorporates language generation and analysis modules for plain and accented written Sanskrit texts. It is based on the principles of ancient Indian Sciences. DESIKA aims to process all the words of Sanskrit. [C-DAC/B]
- Shabdhabodha is an interactive application built to analyze the semantic and syntactic structure of Sanskrit sentences. It works on MS-DOS Platform version 6.0 or higher with GIST shell and is being ported to Windows platform. [ASR Melkote]
- Spell checkers are useful for word processing and are mostly integrated with the word processing software's. Spell checkers in few Indian Languages are available. The development of Spell checkers is covered within the scope of the current projects for corpora development. [Punjabi Spell-checker at CEDTI, Mohali, C-DAC for all]
- An alpha version of "Hindi Vani" software which is PC based Unlimited Vocabulary Text-to-Speech Conversion Software for Hindi for DOS platform has been developed which is being ported to Windows platform. The quality of speech is also being improved upon in terms of pitch, tone, intonation with on-line screen reading capabilities. [CEERI]
- A Devanagari Optical Character Recognition software has been developed and using it approximately 95% accuracy has been obtained in Optical character Reading. [ISI/ Kolkata, C-DAC]

- Pocket Translator software has also been developed which is a tool for the foreign tourists to communicate with the locals. It offers instant translation in both script and voice from one language to another selected language. [C-DAC]
- Angalabharati, (at IIT Kanpur & ER&DCI /N) a Machine-aided Translation System (English to Hindi)for public health domain is being used for the Anti-Malaria Campaign. Domain is being extended with a test bed at Central Translation Bureau. [IITK]
- Anusaraka, (Indian Language to Indian Language), a Language Accessor, is a tool to overcome the language barriers. It also analyses the source language text and presents exactly the same information in a language close to the target language. It provides translation from other Indian Languages to Hindi. [UoH-IITK]
- Matra, a machine- aided translation system (English to Hindi) with a Prototype Vaakya system for web based translation service for English news stories to Hindi has been developed which is being enhanced and adapted for providing web translation service to the news agencies. [NCST]
- Mantra, a Machine-aided Translation System (English to Hindi) for Government notifications. [C-DAC]
- Anubharati, a machine - aided translation system, nascent prototype from Hindi to English. [IITK]
- Localization of Linux operating system at x-windows level is under development. [NCST]
- CASTLE software on DOS platform with GIST card has been developed for Sanskrit teaching and learning as a stand-alone application. Under this project, the synthesis aspect of Sanskrit phonology and word morphology has been handled.
- Trainers Training Programmes in Natural Language Processing were conducted.
- Standardization of 8 bit ISCII (Indian Script Standard Code for Information Interchange) was developed by erstwhile Department of Electronics, and subsequently published by the Bureau of Indian Standards. ISCII is subset of Unicode, which is 16-bit code. Unicode is emerging as future standard for multilingual information processing and MIT has now become a voting member of the Unicode consortium.
  Website: http://www.unicode.org/unicode/
  E-mail: magda@unicode.org
- Standardisation of keyboard layout in the form of INSCRIPT phonetic keyboard.
- Draft Standard of display codes in the form of ISFOC (Indian Standard for Font Code) is ready.
- Draft Standard of pager character code in the form of ISCLAP (Indian Standard for Language Pager) is ready
- Standard Terminology in Hindi for Information Technology is under development in collaboration with CSTT (Commission for Scientific & Technical Technology)

Resource Centres for Language Technology Solutions

The MIT has established thirteen Resource Centres for Indian Language Technology Solutions covering all the constitutional languages.

Organizations and associated Languages

Indian Institute of Technology, Kanpur. (Hindi, Nepali)

Indian Institute of Technology, Mumbai. (Marathi, Konkani)

Indian Institute of Technology, Guwahati. (Assamese, Manipuri)

Indian Institute of Science, Bangalore. (Kannada,Sanskrit , Cognitive Models)

Indian Statistical Institute, Kolkata. (Bengali)

Jawarharlal Nehru University, New Delhi. (Foreign Languages Japanese, Chinese & Sanskrit Language Learning Systems)

University of Hyderabad, Hyderabad. (Telugu)

Anna University, Chennai. (Tamil)

MS University, Baroda. (Gujarati)

Utkal University, Orissa & Orissa Computer Application Centre (OCAC) Oriya

Thapar Institute of Engg. & Tech., Patiala. (Punjabi)

ER&DC, Trivendrum. (Malayalam)

C-DAC, Pune. (Urdu, Sindhi, Kashmiri)

The core objectives of these Resource Centres are:

- To act as a repository of all knowledge tools and products concerned with computer processing of Indian Languages and bring out yearly resource documents.
- To develop the methodologies and tools for seamless integration of language processing tools with existing and evolving software development environment.
- To network with Centres concerned with computer processing of Indian Languages and potential user agencies.
- To create content and databases on the resource information available in Indian languages and to put at least 10 most respected books (related to Indian Heritage) in Indian language on the web. Also to work with local News Papers & to make it available on-line.
- To create awareness and organize training programmes for agencies and personnel concerned with the deployment of Indian language processing systems.
- To facilitate language technology research in Machine Aided Translation, Optical Character Recognition, Text-to-Speech and Speech Recognition for Hindi.
- To organize IT localization clinics for small business to provide consultancy on use of Indian language tools in developing IT solutions and to take up development of requisite niche technologies.