विश्वभारत@tdil

L2/00-407

Language Technology Flash

Moving up the knowledge chain...

Introduction

India is the second largest in population in the world with one billion populations. There are 18 constitutional languages with 10 scripts and over 1650 dialects. Development ɔf 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 multi-lingual knowledge resources. The next milestone has been the setting up of thirteen Resource Centers for Indian language Technology Solutions. These centers will develop technologies for providing solutions with citizen interface in Indian languages selectively and thus covering all Indian languages. The centers will also disseminate these technologies through closer interaction with agencies in State Government, Industry and Academia.

TDIL Programme

Vision statement
Digital unite and knowledge for all.

Mission statement
Communicating & moving up the knowledge chain overcoming language barrier.

Objectives

- To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access multi-lingual knowledge resources/content.
- To promote the use of information processing tools for language studies and research.
- To consolidate technologies thus developed for Indian languages and integrate these to develop innovative user products and services.

Major Initiatives

- **Knowledge Resources**
  (Parallel Corpora, Multi Lingual Libraries/ Dictionaries)
- **Knowledge Tools**
  (Portals, Language Processing Tools, Translation Memory Tools)
- **Translation Support Systems**
  (Machine Translation, Multilingual Information Access, Cross Language Information Retrieval)
- **Human Machine Interface System**
  (Optical Character Recognition Systems, Voice Recognition Systems, Text-to-Speech System)
- **Localization**
  (Adapting IT Tools and solutions in Indian Languages)
- **Language Technology Human Resource Development**
  (Manpower Development in Natural Language Processing)
- **Standardization**
  (ISCII, Unicode, XML, TMX, ISFOC etc.)

TDILPROGRAMME, Ministry of Information Technology, Electronics Niketan 6, CGO Complex New Delhi-110003
Telefax : 011-436 3076 E-mail : info@tdil.gov.in

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. (Banasthali, Vidyapeeth)

- 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 BI-lingual (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. (IIT-K)

- 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. (IIT-K)

- Hindi Search Engine for indexing and searching of Devanagari HTML documents for Linux platform has been developed. (IIT-K)

- 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. (IIT-K)

- 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 at C-DAC, Bangalore. (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. (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 has been developed at CEDTI, Mohali. (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/Calcutta, 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. (IIT-K)

- **Anubharati,** a machine - aided translation system, nascent prototype from Hindi to English. (IIT-K)

- **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)

- Localization of Linux operating system at x-windows level is under development. (NCST)

- CASTLE s/w 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 Programs 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 Script Font Code) is ready.

- Draft Standard of pager character code in the form of ISCLAP (Indian Script Code 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 the following thirteen Resource Centres for Indian Language Technology Solutions covering all the constitutional languages..

Organizations & 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, Calcutta *(Bengali)*
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)*
Jawarharlal Nehru University, New Delhi *(Japanese, Chinese & Sanskrit (Language Learning System)*

The Core objectives of these Resource Centers 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 language and potential user agencies.

- To create content and databases on the resource information available in Indian language 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 Paper to make it available online.

- To create awareness and organise 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 localisation 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.

Potential Products & Services

Multi-lingual Dictionaries, Thesauri, Educational Software, Encyclopedia, Gyan-nidhi Creative Writing System, Translation Support System, OCR, Text-to-speech & Speech Recognition System, Pocket Translator, Personal Digital Assistants, Speech Engine Reading machine for blinds & deafs, Portals, e-governance/e-business/e-skills, cross lingual information access.

TDIL Web Site

http://VishwaBharat.tdil.gov.in *or* http://www.tdil.gov.in

This Web Site contains information for various TDIL activities, achievements and provides access to a variety of content and downloadables in Hindi for other Indian languages.

Free Downloads

- Indian Language keyboard driver & font
- iLEAP
- Desika
- Gita Reader
- Akshar for Windows
- Surabhi Professional
- Pocket Translation
- ALP Personal
- Punjabi Spell Checker
- Shabdabodha

FAQ: Frequently asked Questions on Indian language technologies

Samadhan Seva: To answer user's queries.

Gyan Nidhi Seva : Access to content dictionaries etc.

Implementation Strategy

Consolidation, Integration, Embedding and Innovation

- Technology Integration and Localisation of solutions through Resource Centers

- Focus on user products, services and total solutions

- Technology Innovation Audit to promote standardisation and sharing of technologies.

- Public Domain/General Public License (GPL) approach.

- IT localisation clinics for wider dissemination and internship training.

- Bilateral/International co-operation in Language Technology and Applications.

International Programmes in Multi-lingual Computing

1. HLT (Human Language Technologies) programme of European Union aims developing competitive technologies to facilitate seem less trade and commerce, access to educational and health care aids across their 11 European languages.

   Website: http://www.linglink.lu/hlt/
   http://www.hltcentral.org/

2. TIDES (Translingual Information Detection, Extraction and Summarisation)

   The mission of the TIDES of USA is to develop the technology to enable use of English to locate, access, and utilize network-accessible text documents in other languages, without requiring any knowledge of the target languages.

This will require advances in component technologies of information retrieval, translation, document understanding, information extraction, and summarisation.

Website: http://www.darpa.mil/ito/research/tides/index.html

3. UNL (Universal Networking Languages)

UNL is being developed by UN University, Tokyo. This covers 15 world languages. Hindi is also covered. This aims at development of enconverter and deconverter software for each language into/from UNL using a concept based dictionary and knowledge base. This will be used for on-line translation of documents, and for promoting information exchange without language barrier in the cause of peace among nations.

Website: http://www.unl.ias.unu.edu/
E-mail: uchida@unl.ias.unu.edu

4. Some other Websites on language Technology programmes in other countries -

China	http://www.flce.org/china.htm
Canada	http://www.landfield.com/faqs/natural-lang-processing-faq/ http://eslu.cse.ogi.edu/HLTSurvey/Chinode10.html
USA	http://eslu.cse.ogi.edu/HLTSurvey/HLTSurvey.html http://www.csli.stanford.edu/
Japan	http://iit.msu.edu/vol3num1/nunan
Russia	http://reenie.utexas.edu/reenie/countries/Russia/russia.html
Europe	http://cslu.cse-cgi-edu/HLTSurvey/HLTSurvey.html

MNC Products support Indian Languages

Microsoft supports Hindi and Tamil on Windows 2000, that is a popular Operation System.

Oracle 8i, that is Relational Database Management System, supports Hindi

Lotus supports Hindi at menu level

Tentative List of Indian Language Products

C-DAC

- GIST card: Add-on card to support Indian Languages
- GIST Shell: Software alternative of GIST card
- GIST Terminal: Allow use of Indian scripts in UNIX environment
- ALP: Word Processor for DOS & UNIX
- ISM: Font based package
- LEAP Office: Multilingual Word Processor for Windows
- GIST SDK: Software Development Kit for applications on Windows 95/98/NT
- Lila Hindi Prabodh / Praveen / Pragya: Hindi Learning Software

Modular Systems Pvt. Ltd.

- ShriLipi : Font based multilingual package for Windows 3.11/95
- Rupa: Software for advertisements
- Suchika: Data Entry with transcription facility
- Ankur: Multilingual Word processor for Windows

Sonata Software Pvt. Ltd.

- Parkashak: Fonts base Indian script enabling DTP package for Windows, Macintosh

S C I S Consultants

- Aakriti : Font base package for Windows and Networking

Softek Pvt Ltd.

- Devbase: Database package like dBASE III+ (Hindi & English)
- Akshar: DOS based Word processor similar to Wordstar
- Akshar for Windows: Devanagari font base package for Windows compatible with MS Word

Soft Research Pvt. Ltd.

- ShabadRatna Super: DOS based Word processor
- Vinky: Font base Indian script enabling package for Windows

Sumit Data Products Pvt. Ltd.

- Indica : Indian script enabling package for DOS & Windows

R.K Computer Research Foundation

- Sulipi : Word processor
- Su Window-2.0: Window based Word processor

Tata IBM Ltd

- Hindi PC DOS: Bilingual Disk Operating System (DOS)

Vsoft Pvt. Ltd.

- APC 2.0: Font base Indian script enabling package for Windows

Vedica Software Pvt. Ltd.

- FACT: Accounting package

Abacus Computers Limited

- Mosiac: Complete DTP package

Indian Computers and Option

- Chitrlekha: Font base Indian script enabling package for Windows

Ankit Information Pvt. Ltd.

- Bahubhashi Notepad: Data Entry Package for Windows

Natural Technologies Pvt. Ltd.

- Bank Mitra: Bilingual Word Processor for Windows

Centre for Computer Education

- Aao Hindi Paden: Multimedia based Hindi learning package
- Anuvadak: Machine Aided Translation System (English to Hindi)

Magic Software Pvt. Ltd.

- Guru: Multimedia based Hindi learning package

Microsoft corporation India Pvt. Ltd.

- Windows 2000: Operating System for Hindi & Tamil fonts with Unicode

Secom Solutions India Pvt. Ltd.

- Sulipi: Software utility for Page Maker & Corel Draw

Tentative List of Indian Language Portals

These Websites support composing in Hindi and other Indian languages and have all other features which occur in normal emailing sites like inbox, addresses, compose, folders etc.

1. Web Dunia: www.epatra.com supports 11 languages
2. Mithi.com : www.mailjol.com supports 12 languages
3. Langoo: www.langoo.com supports 12 languages

Contact Person :

Dr. Om Vikas	omvikas@mit.gov.in
Sh. S.A. Kumar	kumar@mit.gov.in
Smt. Swaran Lata	slata@mit.gov.in
Sh. Vijay Kumar	vkumar@mit.gov.in