Language Technology Development in India

Dr. Om Vikas

Ministry of Information Technology, New Delhi, India.
Email: omvikas@mit.gov.in

Abstract

India is a multi-lingual country with 18 constitutional languages and 10 different scripts. Eighteen constitutional Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*); Assamese (*Assamese*), Manipuri (*Manipuri*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada (*Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) and Urdu (*Urdu*). These are in vogue in different states. There are less than 5 percent people who can work in English. Inspite of the plurality of languages and scripts, their script and language grammars are quite similar and they have 40 to 80 percent vocabularies in common. People of India are united over thread of cultural values.

Impact of Information Technology was felt as early in 1970s. Solutions towards adaptation of rapidly growing Information Technology for Indian languages were developed. Input-output problems and coding schemes were analysed. In 1990-91, Government launched the program on TDIL (Technology Development of Indian Languages) under which projects were supported for development of corpora, OCR, Text-to-Speech, machine translation and generic software for Information processing. Standards for keyboard layout and internal Code for Information Interchange were also evolved. This resulted into confidence in having solutions for Information processing in Indian languages. Rapid change in Information Technology (IT) - Operating systems, Generic packages, Peripherals, Internet and networking – made Indian solutions for IT adaptation drag behind. But demand by Government and people continued as thrust for developing Indian language technology solutions, especially, in the wake of establishing world – level par excellence by Indian Software Professionals and companies. In 2000-2001, Government launched mission-oriented program for Technology Development for Indian Languages (TDIL) with focus on seven major initiatives: Knowledge Resources, Knowledge Tools, Translation Support Systems, Human Machine Interface Systems, Localisation, Standardisation and Language Technology Human Resource Development. Thirteen Resource centres for Indian Language Technology Solutions (RC-ILTS) were supported covering all 18 Indian languages. Development of Indian language interface technology is also being promoted. Indian Language Technology Vision 2010 has been prepared with the Vision statement “ Digital Unite and Knowledge for All”. Technology audit focuses on peer-review, peer-technology sharing and product–oriented technology development. India has become voting member of UNICODE consortium. Industry consortium for Indian language technology has also been formed. In order to facilitate collaborative development or transfer of technology, Language Technology Business Meet is being held in October 2001 where technology developers and prospective technology takers will establish dialogue. In the wake of growing popularity of Internet, activities concerning e-Content creation, IT localisation, on-line gisting and summarisation, e-learning, Cross-Lingual Information Retrieval are being promoted to ensure information access in cyberspace in Indian languages.

1. Are We Losing More Knowledge than Gaining?

Sanskrit text in Devanagari Script -
xÉ Ê½ YÉÉxÉxÉ ºÉo¶É {ÉÊ´ÉjÉÊ¨É½ Ê´ÉtiÉ *

Sanskrit text in Roman Script -
na hi jnanen sadrasham pavitramih vidyate

English translation in Roman Script-
"Nothing is so pious like Knowledge."

- Shrimad Bhagwad Gita, 4.38

With the advent of technologies – Writing system, Steam engine, electricity, and computer – there had been rapid transformations in the Societies. Knowledge increased. The societies, which participated actively in the process of Knowledge generation, became advanced. Parity in sharing of Knowledge distances the societies.

" A people become poor and enslaved when they are robbed of the tongue left them by their ancestors; they are lost forever".

- Ignazio Bittira, Sicilia Poet

Tragically, all the environmental erosion comes at a time of equally unprecedented erosion in knowledge. From an estimated 10,000 language in 1900, the world has about 6,700 languages surviving today. Only 50 percent of those surviving ones are being taught to children. This means that half the current languages will be effectively extinct within a single generation. Some studies argue that 90 percent of language spoken in 1999 will be history by 2099. Half of today's languages are spoken by fewer than 10,000 people (half of these are actually used by fewer than 1,000 people). Already people who speak no indigenous tongue occupy one-third of the land area of South America. (Development Dialogue 1999).

Two percent of the world's languages are becoming extinct every year. Four European languages (English, German, French, Spanish) comprise more than 80 percent of all book translations. There is a worldwide unquantifiable erosion of cultural participation and innovation. With the loss of a language, we lose art and ideas, scientific information and innovative capacity, knowledge about medical plants and preparations that could cure maladies.

According to an UNESCO study involving world's 140 most published authors; 90 out of 140 were English writers in 1994 compared to 64 out of 140 in 1980. There is collapse in authorship, translation and quality in other languages. World-level literacy is improving, more people can read than ever before, but fewer people create stories or compose music. We have moved from being *creators to consumers* at a time when technology could have amplified our creative capacities. More than 80 percent of the information on the Internet is in English – even though only 8 percent of world population speak English as first language. [UNESCO, 1998]

Gap between scientific contributions in linguistic communities is widening. Every year, about 46,000 journals and over 80,000 books in science & technology are published. This amounts to over 20 million pages. Most of this is in English, and negligible in the languages of developing economies. [INSDOC, 2000]

In 1960, the world's poorest countries (20 percent of world population) accounted for 4 percent of global exports; by 1990 their share slipped to barely 1 percent. Predictions that the 'poor might not always be with us' have not come true. By 1998, percentage of absolute poor in the world (income below US $1 per day) was at 24 percent and the trendline had turned upward. Optimistic forecasts of gains of technology now seem illusory. Are we winning or losing? Is the world losing more knowledge than it is gaining? [Development Dialogue, 1999]

What is the relationship between Technology (T), Economic development (E) and Knowledge creation capacities (K)? T boosts up E; T erodes K; K-erosion reduces innovation in T. Gain in economic development (E) accompanies loss in cultural resources (K) and innovation capacities (T) in the long run. Constraint of sustainable development would therefore require Appropriate Technologies (T_a).

How can technology convert erupting “digital divide” into “digital unite”?

2. Linguistic Diversity:

Bio-diversity is the characteristic of nature in balance. Similarly the linguistic diversity is the characteristic of the evolving mankind that are geographically dispersed. Varieties of knowledge have grown in response to specific time and space situations.

India is a democratic country with 1 Billion population. There are 1650 dialects spoken by different communities. Linguistic-based division into states ensures use of the official language of that state in governance and education. There are 18 constitutionally approved languages, which are used in different states for citizen interface. There are 10 Indic scripts in vogue. All of these languages are well developed and rich in content. They follow similar Script and language grammars. Alphabetic order is similar. Some languages use common script, especially Devanagari. Hindi written in Devanagri script is the official language of the union Government. English is also used for government notifications and communications. Eighteen constitutional Indian Languages are mentioned as follows with their scripts within parentheses: Hindi (*Devanagari*), Konkani (*Devanagari*), Marathi (*Devanagari*), Nepali (*Devanagari*), Sanskrit (*Devanagari*), Sindhi (*Devanagari/Urdu*), Kashmiri (*Devanagari/Urdu*); Assamese (*Assamese*), Manipuri (*Manipuri*), Bangla (*Bangali*), Oriya (*Oriya*), Gujarati (*Gujarati*), Punjabi (*Gurumukhi*), Telugu (*Telugu*), Kannada *(Kannada*), Tamil (*Tamil*), Malayalam (*Malayalam*) and Urdu (*Urdu*). India’s average literacy level is about 52 percent. Less than 5 percent of people can either read or write English. Over 95 percent population is normally deprived of the benefits of English-based Information Technology. Interestingly, all Indian languages owe their *origin to Sanskrit,* hence they have in common rich cultural heritage and treasure of knowledge. Indic scripts have originated from Brahmi script. For an example, there are typically 19 prominent dialects/variations of Hindi language being used in different regions, e.g., Marwari, Jaipuri, Brijabhasa, Khari Boli, Avadhi, Chhatisgarhi, Bihari, Maithli, Bhojpuri, Magahi, Garhavali, Kumaunni.

Table 1. Distribution of language-based population in India

Language	Script	1991 Census based population	Percent
Hindi	Devanagari	33,72,72,114	41.6
Bangla	Bengali	6,95,95,738	8.6
Telugu	Telugu	6,60,17,615	8.1
Marathi	Devanagari	6,24,81,681	7.7
Tamil	Tamil	5,30,06, 368	6.5
Urdu	Urdu	4,34,06,932	5.4
Gujarati	Gujarati	4,06,73,814	5.1
Kannada	Kannada	3,27,53,676	4.0
Malayalam	Malayalam	3,03,77,176	3.7
Oriya	Oriya	2,80,61,313	3.5
Punjabi	Gurumukhi	2,33,78,744	2.9
Assamese	Assamese	1,30,79,696	1.6
Kashmiri	Urdu/Devanagari	32,00,000 ++	0.4
Sindhi	Urdu/Devanagari	21,22,848	0.3
Nepali	Devanagari	20,76,645	0.25
Konkani	Devanagari	17,60,607	0.20
Manipuri	Manipuri	12,70,216	0.15
Sanskrit	Devanagari	49,736	0.0006

++ estimate. **Total: 81,05,84,919**

Not much variation is expected in the percentage when calculated on the basis of the year 2001 census. According to a survey of 1997, Hindi is understood/used by 66 percent people in India. Hindi is spoken in other countries as well: USA(1,00,000); Germany(30,000); Nepal(80,00,000); New Zealand(20,000); Mauritius(6,85,170); South Africa(8,90,292); Yaman(2,32,760); Uganda(1,47,000); Singapore(5,000) . Besides Hindi, other Indian languages are also spoken in various foreign countries. Bangla is spoken in Bangladesh. Sinhalese, Myanmar and Tibbetan scripts also follow similar alphabetic order and the script grammar. Tamil has large population in Srilanka and Singapore.

During MT Summit XII, September 1999, Prof. Hozumi Tanaka presented estimated ranking of mother tongue based world populations in the following table [TANAKA,1999] :

Table 2. Language-wise world population

Language	2050 Population in Billion	1996 Population in Billion
Chinese	1.384	1.113
Hindi/Urdu	0.556	0.316
English	0.508	0.372
Spanish	0.486	0.304
Arabic	0.482	0.201
Portuguese	0.248	0.165
Bengali	0.229	0.125
Russian	0.132	0.155
Japanese	0.108	0.123
German	0.091	0.102
Malay	0.080	0.047
French	0.076	0.070

This ranking suggests that Chinese, Hindi, English, Spanish and Arabic will still remain the top major languages in 2050.

One-ness in Diversity:

Indian scripts may look different in shapes, but they follow similar alphabetic order. Script grammar is also similar. Alphabet consists of vowels and consonants. They are ordered on the basis of phonetic utterance’s. *What you write what you speak.* Pronunciation of a word is the concatenated string of pronunciation at letter-level. Vowels and consonants have distinct shapes. Pure consonant is a virtual consonant without vowel sound. When vowel follows the (Pure) consonant its modified shape may attach on top, on side or on bottom around the consonant. This modified vowel-grapheme is called MATRA or vowel modifier. Consonants can combine themselves. Indic Scripts are given at Annexure –1.

3. Script Grammar

A: Phonetic Alphabet; G: Graphemic Alphabet; V: Vowel; C: Pure consonant
C_1: C with basic vowel “a”; P: Post-fix; M: Matra or Vowel modifier; H: # Halant or “a” Vowel subtractor; C^+ : Consonant – Consonant – Vowel combine or syllable, W: Word

A = {V, C}; G = { V, C, M, P}
Illustration of Devanagari Script

Vowels:

<V> → + / +É / < / < / = / >ó / @ò / B / B / B / B / +É / +É / +É /+É
Cardinality * V * = 15

Consonants:

<C> → Eó / JÉ / MÉ / PÉ / Ró
SÉ / Ui / VÉ / ZÉ / \É
] / ` / b / f / hÉ
iÉ / lÉ / n / vÉ / xÉ
{É / ¡ó / ¤É / ¦É / ¨É
ªÉ / ® / ±É / ´É / ¶É / ¹É / ºÉ / ½

Cardinality * C * = 33

Matra or Vowel-Modifier:

* M * → #É / Ê# / #ò / # / # / # / # / # / # / # / #É / #É / #É / #É

Cardinality * M * = 14

Post-fix:

<P> → # / # / #&
Cardinality * P * = 3

Combining rules:

$<C> \rightarrow <C_1 + H> / <(C+M) \Theta M>$
$<C_1> \rightarrow <C + \text{“a”}>$
$<C^+> \rightarrow C_1/<C+M>/<C+M+P>/<C_1+P>/<C+C^*>$
$<W> \rightarrow <C^+>/<C^+ - (C^+)^*>$
$(C^+) \rightarrow <\text{null}>/<C^+>/<C^+(C^+)^*>$

- : concatenation + : combining Θ : subtracting

Illustrations of Combining rules of Devanagari script

- Consonant + # (HALANT) ⇒ Pure consonant

C_1 + H ⇒ C

Eó + # ⇒ Eó
{É + # ⇒ {É

- pure consonant + vowel ⇒ Consonant-Matra

C + V ⇒ C-M ⇒ C^+

Eó + + ⇒ Eó
Eó + < ⇒ ÊEó
Eó + = ⇒ Eó

- Consonant Combination

$C + C_1 \Rightarrow C^+$
Eó + ´É ⇒ C´É

- Consonant Combination followed by matra (Vowel-modifier)

$C + V \Rightarrow C\text{-}M \Rightarrow C^+$

C´É + < ⇒ ÎC´É

- Some consonant combinations appear as conjucts such as

$C + C_1 \Rightarrow C^+$
Eó + ¹É ⇒ lÉ
iÉ + ® ⇒ jÉ
VÉ + \É ⇒ YÉ

Script grammar enables in arriving at smaller code-set. Rendering mechanism takes care of shape changes as a result of CV/CC combining. Otherwise 33*14*3 code positions would have been required.

4. Technology Development for Indian Languages

4.1 Paradigm shift from data to information to knowledge processing

Science of computing began with numeric processing data. Data-base design, development and management were the research focus during 1960s & 1970s. Then there was paradigm shift in focus from *Data to Information* during 1980s & 1990s. Management Information System, Internet computing, Content creation were the areas of focus. After mid 1990s, another paradigm shift from *Information to Knowledge* is taking place. Knowledge engineering is emerging as an important discipline especially in the wake of convergence of computing, communication and content technologies. Issues such as Information Technology localization, Knowledge representation, understanding, summarization and integration will be of interest for research and technology development.

Paradigm shift in computing focus from data (in past) to information (at present) to knowledge (in future) has spurred research in Cognitive Science, Natural Language Processing, Speech Processing, Cross - lingual Information Retrieval, Machine Translation, Speech to Speech Translation. Newer information appliances will be wirelessly networked, cross-lingual speech oriented, and intelligent companion.

Development of modern Science and Technology concentrated in the west and English became lingua franca of Science and Technology. Technological transformations speeded up. The period for technology reach to people began shortening.

- Edison switched on the lights of Pearl Street in Manhattan in1882, but it was another 30 years before electrical appliances became widely available in USA.
- It took 38 years after the introduction of the first radio station before the new media was able to reach an audience of 50 million listeners.
- Television reached 50 million viewers 13 years after the first programs were commercialized.
- It took 16 years after introduction of personal computers before the technology could claim 50 million adherents.
- The early telegraph transmitted information at 0.2 bits per second. Today's fibre optic cables transfer data at over 1 billion bits per second.

Mankind has witnessed 1st Revolution with invention of writing system (5000 years ago), 2nd Revolution with invention of written book (1300 BC, China), 3rd Revolution with Gutenberg's invention of printing press (1450 AD) and 4th Revolution is the new information revolution since 1950s. Information Technology is rapidly changing technology with newer products and services showing constant trend of increase in performance and decline in prices. Moore's law (1965), that chip performance will double every 18 months, holds. According to Gilder's law (1993) communication bandwidth will triple every year until 2020 AD.

4.2 Knowledge defies economic principle of scarcity.

Knowledge defies basic economic principle of scarcity. Knowledge is not scarce in traditional sense: the more you use it and pass it on the more it proliferates. Economists view it as "infinitely expansible" or "non-rival" in consumption. It can be replicated cheaply and consumed over and over again. Knowledge is more difficult to measure than traditional inputs such as steel or labour. Government may encourage the creation and diffusion of knowledge by supporting basic scientific research and creating economic environment conducive to innovation by raising standards of education and skills.

At MCI, Vinton Cerf, one of fathers of Internet is developing "network-intensive communication" which would allow any device on the internet to communicate with the telephone network. According to Cerf, there will be a new driver: billions of devices will be attached to the internet.

By 2005, total telecommunication traffic will reverse from the present roughly 80% voice and 20% data. One fifth of all workers in advanced countries will be teleworking either part-time or full-time.

Future prosperity of rich economies will depend both on their ability to innovate and on their capacity to adjust to change. Knowledge based industries have three things in common.

- First, they have high fixed costs such as R&D, but low recurring cost.
- Second, network externalities that means that the more widely a software system is used, the more likely it is to become a standard for industry and more people would like to use it.
- Third, customer lock-in-effect. Many high-tech products are difficult to use. Once learnt, continues with it.

- If all three factors are present, the economist Brian Arthur argues "increasing returns will magnify the market leader's advantage".
- Knowledge – better ways to do things – has always been main source of long term economic growth from agriculture to the present day.
- Innovation themselves are only a first step. Beyond that lies the evolution of ways to use them, a much more gradual and unpredictable process.

4.3 A B C Technology Development Phases

India was aware of the technological changes and the local constraints. Development of Language Technology in India may be categorized in three phases:

- 1976-1990 : **A**-Technology Phase

  Focus was on **A**daptation Technologies; abstraction of requisite technological designs and competence building in R&D institutions.

- 1991-2000 : **B**-Technology Phase

  Focus was on developing Basic Technologies- generic information processing tools, interface technologies and cross-compatibility conversion utilities. TDIL(Technology Development for Indian Languages) programme was initiated.

- 2001-2010 : **C**-Technology Phase

  Focus is on developing Creative Technologies in the context of convergence of computing, communication and content technologies. Collaborative technology development is being encouraged to realise.

Government spending during FY 1991- FY 2001 was about US$ 3 Million.

4.4 Overcoming Language Barrier

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 Ministry of Information Technology 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.

4.5 TDIL Programme

Vision statement

Digital unite and knowledge for all.

Mission statement

Communicating without language barrier & moving up the knowledge chain.

Objectives

- To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access multilingual 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, Multilingual Libraries/Dictionaries, lexical resources)
- **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, Computational Linguistics)
- **Standardization**
  (ISCII, Unicode, XML, TMX, ISFOC etc.)

Long Term Goals

- Speech to Speech translation.
- Human Inspiring Systems

5. Achievements

5.1 Standardisation

Standardization of 8 bit ISCII (Indian Script Standard Code for Information Interchange) was developed by erstwhile Department of Electronics, Government of India, in 1988 and later on The revised version was published by the Bureau of Indian Standards in 1991. ISCII-1988 is subset of the Unicode which is 16-bit code. Unicode is emerging, as future standard for multilingual information processing Ministry of IT has now become a voting member of the Unicode consortium.
Website: http://www.unicode.org/unicode/

Feedback on all languages has been sought by Ministry of IT from language experts, state governments and Resource centers in order to pass on to UNICODE consortium.

Standardisation of keyboard layout was 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 code for Language Pager) is also ready.[at C-DAC]

Standard Terminology in Hindi for Information Technology is under development in collaboration with CSTT (Commission for Scientific & Technical Technology). Draft Standard of multi–lingual lexicon format has also been proposed.

5.2 Knowledge Resources

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. [at CIIL, Mysore]

Computer Courseware in Hindi for DOEACC 'O' level courseware in machine-readable form has been developed and is being put on the web. [at Banasthali Vidyapith]

Content creation in Electronic form, Tagged corpus of Hindi, Hindi Vishwakosh, UN selected countries dictionary, Bharat Bhasha Kosh, Pan-Indian Dictionary, SAARC dictionary, English to Hindi dictionary, Sanskrit to Hindi dictionary, and Bilingual (English, Hindi) are under development.[at ER&DC/Noida – CSTT/CHD]

A Heritage Web site containing traditional Indian texts centered around the 'Upanishads and the Bhaagwadgita' has been hosted. [at IIT, Kanpur]

IIT Mumbai is working on the UN funded project of Wordnet for Hindi in which the syntactic and semantic relationships between the words are represented. The project has just begun and enormous amount of effort is required to build a usable size of wordnet. This linguistic resource is very essential for building Hindi applications such as Machine Translation systems, linguistic analysis, OCR and speech applications

5.3 Knowledge Tools

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. Hindi Search Engine for indexing and searching of Devanagari HTML documents for Linux platform has been developed. Hindi Bulletin Board System is under development. This web-based application allows users to create topics for discussion and maintains threads within a topic. [at IIT, Kanpur]

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. 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. [at C-DAC]

Multi-lingual e-mail Client has also been developed. Its working prototype facilitates 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. [at CMC]

Sanskrit word processor to handle special Sanskrit constructs is under development. **Sanskrit Authoring System** including a Sanskrit word processor for use by Sanskrit scholars in text processing etc. is being developed. **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. [at C-DAC/Banglore]

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. [at 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. [Punjabi Spell-checker at CEDTI, Mohali, for all languages at C-DAC]

A number of Indian Language Processing software packages have been developed. [at C-DAC, Modular Infotech, etc.]

GIST Shell was software alternative of GIST card for DOS. ALP was multi-lingual word processor for DOS, UNIX & Novell. Leap Office 2000 is complete Indian language software applications on windows. iLEAP is internet ready Indian language word processor. ISM, Insfoc Script Manager, is font based interface to windows based software. ISM Publisher is suited for conventional & web publishing. ISM Soft is designed for software developers. GIST SDK is software Development Kit for applications on Windows 95/98/NT. iplugin enables development of Indian language applications such as text processing, e-mail messaging, chat, calendar, Scheduler of events, etc. N-Trans transliterates nouns from and into Indian languages. LIPS decoder sub-titles TV programmes in Hindi and other Indian languages simultaneously in the transmission. MOVE enables on-line video editing in TV broadcast. Teleprompter is a newsroom IL reading facilty in TV broadcast studios. DVD Authoring tool enables authoring of movies on DVD. UTRANS enables reading popular Hindi text for Urdu speaking community. [at C-DAC]

Font-based multilingual packages, multilingual word processor, transcription facility, Font based Indian script enabling DTP packages, Database packages, Indian script enabling packages, Data entry packages, e-mailing system, Machine Translation Systems, [E→H, IL→H, H→E], application software packages in Indian languages such as Address management system, Indian language learning system, Management Information Systems in Government, business management system, etc have been developed at various organizations [Modular, Sonata, Softek, Summit, NCST, CCE, ER&DC/N, NIC, TCS, IITK, IBM, Oracle, etc.]. Indian language support is also becoming available on operating systems, Windows 2000 and Linux [Microsoft, NCST, IITK].

5.4 Translation Support Systems

Pocket Translator software has 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. **Mantra** is a Machine-aided Translation System (English to Hindi) for Government notifications. [at C-DAC]

Angalabharati, (at IIT Kanpur & ER&DCI/N), a Machine-aided Translation System (English to Hindi)for public health domain is being developed for the Anti-Malaria Campaign. Domain is being extended with a test beds for offcialese, health and agriculture at Central Translation Bureau. **Anubharati** is a machine - aided translation system, a nascent prototype from Hindi to English. [at IIT, Kanpur]

Anusaraka a Language Accessor, (from one Indian Language to any other Indian Language), 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 rapid translation as language accessor 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. [at NCST, Mumbai]

5.5 Human Machine Interface Systems

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, New Delhi]. Speech Technology group at IIT Madras is also developing technologies for Indian languages.

A Devanagari Optical Character Recognition software has been developed with approximately 95% accuracy. OCRs for Hindi, Assamese, Punjabi are being developed. [at ISI/Kolkata, C-DAC, IIT, kanpur, IIT, Guwahati, Thaper Inst. Of Engg. & Tech.]

Line printers were enabled for printing Devanagari [at Lipi Data Systems & Transmetic Systems]. TVSE developed and marketed printers capable of printing Devanagari and other Indic scripts. INSCRIPT keyboards with engraved bi-scripts (Roman and Devanagari) are in vogue. Bilingual computer compatible electronic teleprinterss were manufactured [at Abacus, CMC, HTL, AEM, Databyte Equipments]. Gist terminal was developed [at C-DAC] that allows use of Indic scripts in UNIX environment. GIST add-on card supports Indian languages on DOS.

5.6 Localisation

Localisation of the existing generic software was carried out by designing Indic script enabling interface software. Indian script support is now being provided at Operating System level also in DOS, Windows and Linux. Font conversion utilities enable exchange of e-Content seamlessly. Localisation of software involves use of words/phrases of local languages. Localisation of e-Content involves use of local language enriched with locale specific cultural values.

Linux Technologies for Indian languages include Indian scripts based terminal software, print utility, spell checker, web browser capabilities for displaying Indian scripts, utilities library. It is now possible on Linux to give file names, domain names in Devanagari. Indian language support on Linux in open domain will also ensure Indian Language s support on all Open Source Tools Kits.

Unicode compliance will also ensure Indian language support on the software packages.

Simple Inexpensive Multi-lingual ComPUTER (**Simputer**) has been designed that enables use of Smartcard, Text-to-Speech, Information Markup Language for Internet applications (IMLi) is XML based. IMLi browser supports Indian languages. Its features include Linux OS, 32 bit CPU, 32 MB D-RAM, 320*240 display, Soft-modem, Touch Panel, MP3 Player, Stilus/tap-a-tap input. Its price is estimated about US$ 200. This may become a means for bridging Digital Divide.

Localization of Linux operating system at the X-windows level has been done [at NCST, Mumbai]

Indian language supports for Linux Products are being worked out at the Resource centers.

5.7 Language Technology HRD

CASTLE software on DOS platform with GIST card was developed for Sanskrit teaching and learning as a stand-alone application. Under this project, the synthesis aspect of Sanskrit phonology and word morphology have been handled.

Trainers Training Programmes in Natural Language Processing were conducted

Introduction of modular IT curricula in language studies and linguistics is under preparation.

Ministry of IT played proactive role of introducing Indian languages in IT curricula designed for Secondary and Senior Secondary schools of CBSE.

IT-enriched curricula for functional Hindi at BA & MA levels have also been prepared.

Courses on Computational Linguistics and Language Engineering are also being planned.

6. Resource Centers for Language Technology Solutions

The Ministry of IT has established thirteen Resource Centres for Indian Language Technology Solutions covering all the eighteen constitutionally approved official languages.

Organizations and associated Languages:

Indian Institute of Technology, Kanpur. (*Hindi & Nepali*)
Tel: 0512-597174 E-mail: rmk@iitk.ac.in

Indian Institute of Technology, Mumbai. (*Marathi & Konkani*)
Tel: 022-5767718 E-mail: pb@cse.iitb.ac.in

Indian Institute of Technology, Guwahati. (*Assamese & Manipuri*)
Tel: 0361-690321-28 E-mail: sbnair@iitg.ernet.in

Indian Institute of Science, Bangalore. (*Kannada & Sanskrit Cognitive Models*)
Tel: 080-3092377 E-mail: njrao@mgmt.iisc.ernet.in

Indian Statistical Institute, Kolkata. (*Bengali*)
Tel: 033-5778085 E-mail: bbc@isical.ac.in

Jawarharlal Nehru University, New Delhi. (*Foreign Languages: Japanese, Chinese and Sanskrit Language Learning Systems)*
Tel: 011-6107676 E-mail:gvs@jnuniv.ernet.in

University of Hyderabad, Hyderabad. (*Telugu*)
Tel: 040-3010500 E-mail: knmcs@uohyd.ernet.in

Anna University, Chennai. (*Tamil*)
Tel: 044-2351723 E-mail: rp@annauniv.edu

MS University, Baroda. (*Gujarati*)

Tel: 0265-792959 E-mail:sityash@satyam.net.in

Utkal University, Bhuvaneshwar (*Oriya*)
Tel: 0674-580216 E-mail: sangham@sanchar.net.in

Orissa Computer Application Centre, Bhuvaneshwar (*Oriya*)
Tel: 0674-543113 E-mail: akp@ocac.ernet.in

Thapar Institute of Engg. & Tech., Patiala. (*Punjabi*)
Tel: 0175-393137 E-mail: gslehal@mailcity.com

Electronics Research & Development Ccenter (ER&DC), Trivendrum. (*Malayalam*)
Tel: 0471-325897 E-mail: ravi@erdcitvm.org

Center for Development of Advanced Computing (C-DAC), Pune. (*Urdu, Sindhi & Kashmir*i)
Tel: 020-5652461 E-mail: rkarora@cdac.ernet.in

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 and 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 and other Indian languages.

- 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

7. Implementing Strategy

7.1 Consolidation, Integration, Embedding and Innovation

- Technology Integration and Localization of solutions through Resource Centres
- Focus on user products, services and total solutions
- Public Domain/General Public License (GPL) approach for faster development.
- IT localization clinics for wider dissemination and internship training.
- Bilateral/International cooperation in Language Technology and Applications.

7.2 Technology Audit

Technology Innovation Audit of the sponsored projects is essential in order to promote standardization and sharing of technologies. Audit steps may include:

- Concept, Design and Implementation audit
- Alpha Testing with Peer Developers

- Beta Testing with a small number of potential users
- Certification of IL Software

Peer review of the projects and enforcing Beta testing of products or services yield satisfactory results; Culture of collaborative technology development is also strengthened.

7.3 TDIL Web Site

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 and for other Indian languages.

Free Downloads

- Indian Language keyboard driver & fonts
- iLEAP, Akshar for Windows, Surbhi Professional
- Desika. Gita Reader, Shabdabodha
- ALP Personal, Spell Checkers

FAQ: Frequently asked Questions on Indian language technologies
Samadhan Seva: to answer user's queries
Gyan Nidhi Seva: to access to content, dictionaries, classic works

7.4 Information Dissemination through TDIL Newsletter

Events Under the TDIL program include

- Annual **TDIL Meet** in which national experts on language technology present their papers and discuss various techniques and tools for possible sharing and refinement of their technologies.
- **Language Technology Business Meet** will be the first event in October 2001 wherein technology developers from academia will have dialogue with prospective technology takers from industry for possible transfer of technology or for further collaborative development.[List of Indian Language Technologies for transfer are listed at Annexure 2]
- UNICODE Technical Committee requires feedback and detailed note on Indic Scripts for possible inclusion in their revised version. India participates in some of their meetings.
- UNESCO experts group on Multi lingualism invites India for participation.
- Resource Centres develop various language technology solutions.

All these events are published in the periodic newsletter VishwaBharat@tdil and disseminated widely.

7.5 Web Sites Supporting Indian Languages

The following Websites support composing in Hindi and other Indian languages and have all other features which occur in normal e-mailing 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

4. CDAC: www.cdacindia.com with multilingual support

Prominent industries developing Indian language Technologies are the following [refer VishwaBharat@tdil, Jan 2001 for details]

- Microsoft supports Hindi and Tamil on Windows 2000 and office XP.
- Oracle 8i, that is Relational Database Management System, supports Hindi.
- Lotus supports Hindi at menu level.

7.6 ZOPP Workshop

ZOPP is group problem solving approach for decision making. ZOPP means objectives Oriented Project Planning. ZOPP Workshop was organised on May 3-5, 2000 in Bangalore wherein all the Resource Centres for Indian Language Technology Solutions participated. A detailed participation analysis was made in smaller groups for discussing the concept of project planning matrix. Five common outputs were formulated as follows:

- Development of portals
- Training programs
- Knowledge and Database to create
- Development of Spell checker which is Unicode compliant
- 10 classic books on web

Upon the recommendation of the Zopp workshop a Technology Workshop at C-DAC was organised to enable the Resource Centres for speedy take up in respective activity and exposure to the C-DAC technologies for Indian languages.

7.7 Language Technology Marketing & IPR

Language/script is emotional issue and difficult for conversion. There is need for intervention by way of investment, market promotion, standardization. Participation of State Governments in Indian language technology development must be engineered. STQC and other organizations may be encouraged for certification of language technology products. Technology development and investment may be encouraged in parallel groups. 5% of all the IT purchases in the Government may be for Indian language products and services, which could be audited. Government funded project must have at least one Indian language Interface. There is need for a consortium/international cooperation in the field of Indian language technology.

So far, copyright applications for Corpora (CIIL, Mysore), and Desika (C-DAC) have been processed. Copyright for Lila (C-DAC, Pune) are in pipeline. Applications for patent/copyrights for the following products have been filed: Web search engine, Unicode font and encoding, ISCLAP and font conversion.

There is need to glamourize language technology products by organizing road shows and campaign. There is also need to involve State Governments to use Indian languages in e-governance and other socio economic projects/programs.

According to NASSCOM estimate, IT spending for e-governance by state governments in local languages is growing rapidly and language software market would grow from the current about US $20 million to US $ 50 million by end of FY2001.

7.8 Industry Consortium for Indian Language Technology

Ministry of IT has taken initiative to nucleate formation of an industry consortium for Indian Language Technology. This is being coordinated by MAIT. Involvement of

Industry will spur the language technology – development, dissemination and export. Language technology marketing is also essential. This consortium will conduct studies and survey and play a pivotal role in evolving national and international standards.

8. Major Language Technology Projects

8.1 On-going:

- Localization of Linux
- Content Creation (dictionary, encyclopedia)
- Machine Aided Translation (English to Hindi & other ILs)
- Text-to-Speech on windows
- Certification of Language software

8.2 New Initiatives under consideration are:

- COILNET (Content creation and IT Localisation Network)
- KUNDALINI (Knowledge Understanding, Acquisition of Languages, Inferencing and Interpretation)
- CLIR (Cross-Lingual Information Retrieval – text & speech)

9. Impediments in Proliferation of Indian Language Technologies

- Lack of industry involvement due to constrained demand;
- Unsustained demand in economically backward states: BIMARU and other States;
- R&D in language technology so far open ended, not product-driven in time-bound manner;
- Negligible software tools and re-usable components in public domain;
- Computer Scientists least interested in Natural Language Processing due to limited scope;
- No formal IT courses in the curricula for linguistics, language teaching and language studies. Lack of IT culture in language graduates;
- No strategy for language technology marketing;
- Unable to check import of IT products and services which don't support Indian language(s);
- No Consensus on standardisation Standards in use; ISCII-88, ISCII-91, UNICODE, many propriety code; Content is largely glyph-coded, not (ISCII) character-coded;
- Slow pace of transfer of language technology from academia to industry

Acknowledgement

Author expresses sincere gratitude to Prof. MGK Menon, then Secretary, Department of Electronics, Govt. of India for initiating him in 1979 into very new area of Informatics for Indian Languages.

Hon’ble minister Shri Pramod Mahajan constantly inspires him with his very focused vision and Shri Rajeeva R. Shah Secretary Ministry of Information Technology always encouraged smilingly to contribute innovatively. The author expresses sincere gratitude to them.

The author thanks the TDIL team-members – Prakash Chaturvedi, Swaran Lata, Manoj Jain, Vijay Kumar - for progressing collectively towards achieving the objectives of the TDIL program.

References

1. Development Dialogue, 1999:1-2, Dag Hammarskjold Centre
2. World Culture Diversity, UNESCO, 1995
3. World Culture Report - Culture, Creativity and Markets, UNESCO, 1998
4. TANAKA, Hozumi, “What should be do next for MT system development?”, MT Summit VII, 13-17 September 1999, at Kent Ridge Digital Labs, Singapore.
5. Hindi in the Republic of India - status and direction (in Hindi), Nehru Memorial Museum, 2000.
6. Indian Script Standard Code for Information Interchange (ISCII), IS 13194:1991
7. VishwaBharat @tdil, Jan 2001 & May 2001; Published by Ministry of Information Technology, India

Annexure - 1
IS 13194 : 1991 Indian Script Alphabet Correspondence

	RMN	DEV	PNJ	GJR	ORI	BNG	ASM	TLG	KND	MLM	TML
#Ä	MÄ	#Ä		# Æ	#Ü	#c	#c				
#Æ	Mï	#Æ	#}	#Å	#Õ	#e	#e	Lıı	#M	#w	
#ç	Hï	#ç		#ç	#Ó	#f	#f	M	#N	#x	ç
+	A	+	A	+	@	%	%	@	@	A	Ü
+É	ā	+É	As	+É	A	%oç	%oç	A	A	B	Ý
<	(	<	uB	<	B	+	+	B	B	C	Þ
<ç	(	<ç	Bv	>	C	<	<	C	C	Cu	ss
=	()	=	Cx	A	D	=	=	D	D	D	À
>ó	()	>ó	Cy	C	E	>	>	E	E	Du	Á
@ò	Rï	@ò	uj	Eì	F	@	@	ÊÁ	Fss	E	

								VV			
Bà	E							Fs	C	F	Â
B	Ē	B	B~	+à	H	A	A	C	H	C	Ã
Bä	Ai	Bä	A¤	+ä	I	B	B	H	I	sF	Ä
Bì	Eˆ	Bì									
+Éà	O	+Éà						I	J	H	Å
+Éä	O	+Éä	D	+É à	J	C	C	J	K	Hm	Æ
+Éè	au	+Éè	A¬	+É ä	K	D	D	K	L	Hu	Å÷
+Éì	Oˆ	+Éì									
Eó	ka	Eó	E	Hí	L	Eõ	Eõ		Oú	I	È
Fó	ka	Fó	E								
JÉ	kha	JÉ	C	LÉ	M	F	F	ÅÁ	R	J	
KÉ	kha	KÉ									
MÉ	ga	MÉ	I	NÉ	N	C	C	CRI	VÚ	K	
NÉ	gha	NÉ	J								
PÉ	gha	PÉ	K	PÉ	O	H	H	xm NsV	Yú	L	
	RMN	DEV	PNJ	GJR	ORI	BNG	ASM	TLG	KND	MLM	TML
Ró	nÆa	Ró	L	Rî	P¼	Iø	Iø	ÃÁ	\	M	É
sÉ	ca	sÉ	M	sÉ	Q	»Jô	$Jô	¿R Á	^ú	N	Ê
Uú	cha	Uú	N	Uï	R	»K÷	$K÷	¿³R Á	aÚ	O	
VÉ	Ja	VÉ	O	WD	s	L	L	ÇÁ	D	P	Ü
WÉ	za	WÉ	P								

	jha	ZÉ	Q	]Ñ	Tb	Mõ	Mõ	LRI ¢V	ÁÚ Hss	Q	
\É	N˜a	\É	R	_É	U	Aõ	Aõ	ÄÁ	J	R	Ë
]	TÏa	]	s	`ò	V	»O ô	$O ô	ÈÁ	L	s	Ì
`	+Ïha	`	T	Có	W	Pö	Pö	HRI	PÚ	T	
B	dÏa	B	U	Eô	X	Qö	Qö	²R¶ ¶	sÚ	U	
C	Ďˆa	C	[			QÍö	QÍö				
F	dÏha	F	W	Hõ	Y	»R ô	$R ô	²³R ¶	VÚ	V	
G	Ďˆha		[-			»RÍ ô	$RÍ ô				
HÉ	nÏa	HÉ	X	IÉ	Z	s	s	ssÁ	y	W	Í
IÉ	+a	IÉ	Y	LÉ	[	Tö	Tö	»R ½	}Ú	X	Î
LÉ	+ha	LÉ	\	OÉú	\	U	U	´R¶	¢Ú	Y	
NÙ	da	NÙ	]	Qö	]	V	V	µR¶	¥Ú	Z	
VÉ	dha	VÉ	_	yÉ	^	Wý	Wý	µ³R ¶	¨Ú	[	
xÉ	na	xÉ	`	{É	_	X	X	©«s	«Ú	\	Ï
xÉ	na	xÉ									Ù
{É	pa	{É	A	~É	`	Y	Y	xMs	®Ú	]	D
¡ò	pha	¡ò	B	£í	$¼	Zõ	Zõ	xMN s	±Ú	^	
¢ò	fa	¢ò	C								
¤É	ba	¤É	D	¥É	A	[ý	[ý	ÊÁ	¶	_	
¦É	bha	¦É	E	§É	BD	\ö	\ö	Ë³Ï	•Ú	`	

								Á			
¨É	ma	¨É	C	©É	C	]	]	ª«ʃ V	ÈÚ ʃʃ	A	Ñ
	RMN	DEV	PNJ	GJR	ORI	BNG	ASM	TLG	KND	MLM	TML
ªÉ	ya	ªÉ	H	«É	¯Æ	Ì^	Ì^	¸R¶ V	¾Ú ʃʃ	B	Ò
ªÉ	yÆ a				¯	Ì^	Ì^				
®	ra	®	J	-÷	EÞ	Ì[ý	»	LRI	ÁÚ	C	Ó
®	r̲a	®						ÀÁ		D	Ø
±É	la	±É	K	±É	MÞ	_	_	ÌÁ	Ä	E	Ô
	)ÏÆa			³	F				×Ú	F	÷
	z̲a									Ç	Ö
´É	va	´É	M	´É	C	[ý	¾	ª«ʃ	ÈÚ	H	Õ
¶É	ś a	¶É	O	¶É	H	`	`	AR Þ	ËÚ	I	
¹É	ʃÏA	¹É		•É	I	B	B	×Tʃ Q	ÎÚ	J	Û
•É	sa	•É	N	»É	J	A	A	×Qʃ	ÑÚ	K	Ú
½	ha	½	Þ	¾	KÞ	Ç÷	Ç÷	×¤¦ ¦¦	ÔÚ	L	Ý
#É	ā	#É	#ʃ	#É	#D	#ç	#ç	%ʃ	#Û	#M	#£
Ê#	i	Ê#	U#	Ê#	#Þ	×#	×#	%T Á	#Ý	#N	#¤
#Ò	ī	#Ò	#V	#Ò	#Ñ	#Ý	#Ý	%U Á	#Ý Þ	#O	#¦
#Ö	u	#Ö	#×	#Ö	#Ê	#Ç	#Ç	%Á V	#Ú ʃʃ	#Þ	#8

#ú	ū	#ú	#y	#Ú	#Ë	#É	#É	%Á W	#Ú à	#Q	#¨
#Þ	rÏ	#Þ		#Þ	#ó	#Ê	#Ê	%Á X	#Ú ä	#R	
#à	e	#à						Z % Á	#Æ	ʃ#	ª#
#ä	ē	#ä	#~	#à	Ò#	Ò#	Ò#	z% Á[	#Æ Þ	«#	«#
#è	aì	#è	#¤	#ä	Ò# ʃʃ	É#		\Z %Á	#Æ ç	ʃʃ#	¬#
#ì	ê	#ì									
#É à	o	#É à						%]	#Æ à	ʃ#M	ª#£
#É ä	ō	#É ä	#¨	#Éà	Ò# Ð	Å# ç	Å# ç	%][	#Æ àÞ	«# M	«#£
#É È	au	#É È	#¬	#Éä	Ò#×	Ò# ì	Ò# ì	%_	#è	#U	ª#÷
#Éì	ô	#Éì									
	RMN	DEV	PNJ	GJR	ORI	BNG	ASM	TLG	KND	MLM	TML
#Â		#Â	#±	#Ã	#ç	#Ë	#Ë	%` Á	#é	#V	#¢
#Ã		#Ã	#Q		#Ï	#Ì	#Ì				
*		*	*	*	¼¼ Ð	*	*	.	.	.	.