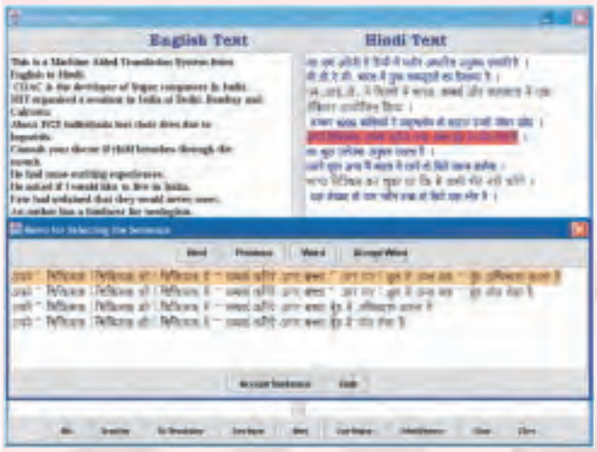
6.3 Language Technology Initiatives at C-DAC Noida

CDAC Noida (formerly ER&DCI, Noida) has been working on development of various language technology products, resources and tools with TDIL programme. A brief description of various language technology products developed by CDAC Noida are presented here.

1. Translation Support System

A software for translation from English to Hindi based on Angla Bharati approach of IIT, Kanpur Paninian framework fused with modern Artificial Intelligence Techniques Karak Theory used in choosing right prasarg in target text.

Built in intelligence to deal with unknown words, correct preposition resolution. User friendly post editing facility. No restriction on input text envisaged.



2. GyanNidhi

1 Million Pages Parallel Multilingual Corpus

1 Million Pages Multilingual Parallel Corpus in English and 12 Indian languages.

Text in UNICODE format (International Standard).

State of the art GUI.

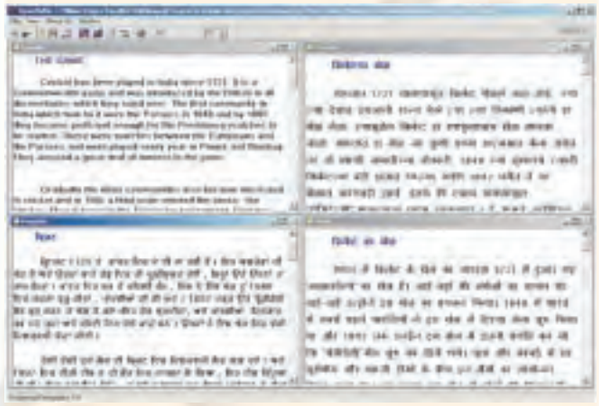
Corpus collected from various domains.

Data availability in XML / HTML format.

Corpus management tools.

Useful for applications such as Improving translation systems, Translation memory, Spell checkers, Dictionaries, Statistical text analyzer, Language related research, Writing style analysis, Morphological analyzer.

Major contributors of text are National Book Trust India, Sahitya Akademi, Publication Division



3. Dware Dware Gyan Sampada - Mobile Digital Library

Internet Enabled Mobile Digital library brought to use of common citizen for spreading literacy.

C-DAC Noida (DIT, MC&IT, Govt. of India) contributes in bringing the digitized books at the doorsteps of common citizens. It makes use of Mobile van with satellite connection for connectivity to Internet. The van is fitted with printer, scorer, cutter and binding machine for

providing bound books to the end user. Different places such as schools in villages and other remote areas will be covered under this programme to promote literacy and demonstrate use of technology for masses.

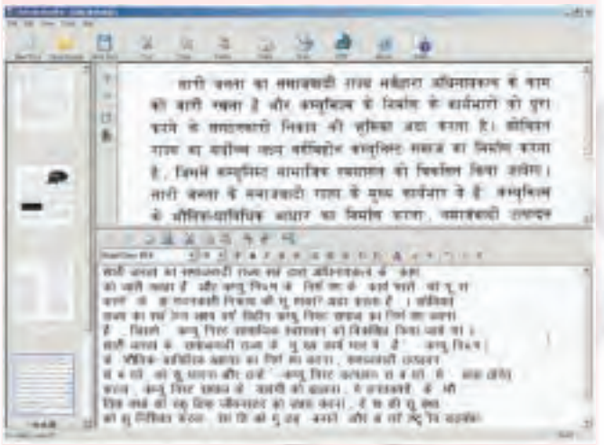
Books formatted for book printing can be selected depending upon

Language, Author & Title. Useful for Children, Villagers, Students and Public at large.

4. Chitraksharika

Optical Character Recognition for Devanagari Script

An efficient way to convert optically scanned images of printed materials into computer processable data files.



Based on technology by ISI Kolkata.

Recognizes Hindi, Marathi & Nepali.

Scanning images via TWAIN interface.

Auto Image segmentation, de-skewing, detection of text, table & pictures

Image editing features.

Embedded spell checker for Hindi.

Save text & scanned images in different formats.

5. Lekhika

Platform Independent Word Processor (Linux, Solaris, Windows, etc.)

Multilingual document support

संशोधिका : Language sensitive Spell Checking

शब्दिका : Authenticated glossaries

Text Formatting, Paragraph, Font, Style etc.

परिवर्तिका : Font Conversion utility

Conversion between various file formats

Automated / Interactive Transliteration for

lAny line l Any selected portion lAny Document

Import & Export for various file & encoding formats

Embedded utilities like Calculator etc.

Drag & Drop support.

6. On-Line Hindi Vishwakosh

A joint project of KHS, Agra (Ministry of HRD, Govt. of India) & C-DAC, Noida, (DIT, Ministry of Communications & IT, Govt. of India) for bringing out Hindi Encyclopedia (Vishwakosh Published by Nagari Pracharini Sabha, Varanasi) on Internet in public domain.

Collection of approximately 12,000 topics.

Search facility in Hindi within the site

Information in Alphabetical as well as Categorised form.

7. On Line IT Terminology

A joint project of CSTT, New Delhi (Ministry of HRD, Govt. of India) & C-DAC, Noida (DIT, Ministry of Communications & IT, Govt. of India) for bringing out the Information Technology Terminology in Hindi on Internet in public domain.

Collection of approximately 10,000 standardised terms with their Hindi equivalents.

Search facility within the site in English/Hindi.

Categorisation of terms in various fields of Information Technology.

8. *स्वरणाकृति* (Swarnakriti)

A Unicode Based Word Processor with integrated OCRs & TTS

C-DAC, Noida is developing "Swarnakriti"- A Unicode Based Word Processor with integrated OCRs & TTS, under the umbrella of TDIL Programme of Ministry of Communication and Information Technology.

The support for typing in Inscript has been extended for Marathi, Tamil & Gurmukhi languages as well apart from Hindi.

Swarnakriti has special features besides having common features of a Word Processor, like an integrated powerful Spell-Checker for Hindi which works on Unicode format, Dictionary based Transliteration, Calendar and Scientific Calculator.

Spellchecker has a Unicode format dictionary of around 34,000 root words, Swarnakriti allows user to keep updating this dictionary through spellchecker interface. English to Hindi transliteration currently works for Unicode strings and is dictionary based.

Optical Character Recognition (OCR) Systems developed by Indian Language Technology Solution Resource Centres are being integrated with the word processor.

9. Gyanaudyog

Advances in computing technology are enabling people to interact with one another and on a global scale. It has also made far-reaching impact on every aspects of human life. The Internet has added a new dimension to IT, allowing free flow of Information across boundaries of space, time and language

cultures. In the information age of today and knowledge era, if we have to target to make India a hub for R & D in the area of ICT and ICT related services internationally as well as at National level our aim need to be on bridging the digital divide. Digital divide can be narrowed by developing the technologies supporting Indian languages and developing contents in Indian languages so that the information is available to common man as well as these efforts results in skilled human resource generation.

Creating awareness about IT tools for Indian languages will act as a catalyst in making IT a powerful tool for empowerment of the masses and bridging the digital divide. In Indian context as India is a pluralistic multicultural society; it is necessary to promote tools for different languages relevant to different regions of the country.

A Gyanodyog workshop is a three day workshop aimed at creating awareness about tools for using Indian languages on Computers and the possibility of self-employment with the help of the IT catering the need of local population. The workshop will cover topics related to

- Computer Fundamentals
- Business and Management aspects
- Language Tools
- Financial Assistance Available for Entrepreneurs
- Possible Enterprise setups such as content Creation, Design, Remote customer interaction, Educational Services, and other Small Office, Home & Education (SOHE) applications.

This will involve the designing a proper course material, training the workshop trainers and identifying participants at regional level in order to proliferate the benefits of the IT tools and services to the masses.

One such workshop has already been conducted as pilot at TDIL, Department of IT, MC&IT, New Delhi

The Gyanaudyog centre will also act as nodal centres necessary Technological Mentoring, Financial support & Marketing information from time to time to the enter preneures. The Workshops will be conducted with a two Tier structure first tier will have basic level workshop whereas second tier will focus on specialized information to participants interested in setting up their own enterprise.

Why Gyanaudyog ?

Gyanaudyog workshops will serve the purpose of information dissemination of Indian Language Technologies and other information tools developed in the country and their access & gainful deployment towards entrepreneurial initiatives. Further it will motivate educated unemployed youth from different knowledge domains to utilize IT and IT enabled services in various fields such as content development, translation, trans-creation, Educational services and also setting up their own enterprises. Awareness and education spread by Gyanudyog series will also help in SOHE initiatives for housewives, senior citizens and other such persons who wish to utilize their knowledge and expert to do work but have limitations in leaving home for longer duration. Thus it will help in tapping the vast pool of efficient but dormant human resources who can prove to be the best skilled resource for technology development and content creation in Indian Languages as well as in English on our cultural heritages and other aspects.

These workshops will also result in dissemination of benefits of IT and ITES upto the grass root level. It is envisaged that Gyanudyog series would also facilitate to improve the low Human Development Index from 0.571 for India as reported by the UNDP HDI report of 2001.

Proposed Methodology for conducting the workshop:

The workshops will be conducted on a two-tier structure

First Tier: Basic awareness workshop for the educated persons in the region for three days giving details of computer fundamentals, various tools, SOHE enterprise setup and their management along with hands-on experience on computers and workshops on project proposal development catering the needs of region . These workshops will be conducted once in a month.

Second Tier : Specialized workshops will be conducted for the participants of basic workshops who show their interest in starting their own enterprise or setting up SOHE. The workshop will be conducted to provide details of financial assistance, marketing information, and technological requirement for individual projects prepared by the participants. These workshops will be conducted once in four months.

Forecast

Information technology and its related services has resulted in information explosion world-wide. As per UNDP reports India has become a Super-Power in IT to the same time Digital divide is also at its maximum. Still PC and Internet penetration is low in comparison to other countries who are still far-away from us in IT race. Multi-Language culture of the country has been a major barrier in harnessing the benefits of IT and ITES in the country. For diminishing the digital divide, availability of IT tools in Indian Languages and their easy access to common masses is necessary. These workshops will create awareness about technology benefits, its usage along with increasing skilled human resources at SOHO level.

Gyanudyog Workshops will be able to create and share knowledge of Indian culture, heritage, technology development with local masses as well as International fraternity in the language of their choice along with empowerment of masses and bridging the digital divide.

Target Segments

- Housewives, Senior Citizens
- Unemployed youth willing to set up their enterprise
- Government officials
- Information & Technology Professionals – responsible for distributing, managing and organizing content
- Entrepreneurs/Content buyers – who evaluate, acquire, manage information/ external contents
- Content creators & owners – thinking globally but acting locally.
- Digital Asset Managers – providing and distributing media assets
- Web Managers – involved in content design and delivery
- Management Experts, Administrators
- Hindi Users

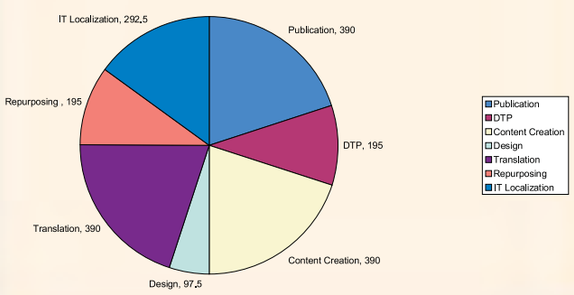
According to NASSCOM Survey Report the total market revenue for IT enabled services in year 1999-2000 was Rs. 2400 Crores and in year 2000-2001 was 4100 Crores. Employment opportunity for 70,000 people existed.

According for this growth pattern market opportunity for year 2003-2004 in IT enabled services is about Rs.6500 Crores out of which 30% (Rs. 1950 Crores) can be assumed for Hindi and Indian language (IL) and opportunity exists for 21,000 people.

Break up of opportunities in various sectors of IL

S. No.	Sector	%age	Revenue in Crores (Rs.)
1	Publication	20	390
2	DTP	10	195
3	Content Creation	20	390
4	Design	05	97.5
5	Translation	20	390
6	Repurposing	10	195
7	IT Localization	15	292.5
	Total	100	1950

Breakup Revenue in Various Sectors



Reference Paper

Indian software industry to grow by 65 per cent: Nasscom

The National Association for Software and Services Companies, an apex body representing the Indian software industry and the Boston Consulting Group will soon come out with a survey on the domestic e-commerce scenario and the global e-commerce solution space.

June 14, 2001

This was disclosed by Phiroz Vandrevala, chairman, Nasscom, while announcing the 65 per cent annual growth of the software and services industry.

During the year 2000-01, the Indian IT software and services industry has shown gross annual revenue of Rs 37,760 crore (US$ 8.26 billion), the annual industry survey Nasscom stated.

"In 2000-01 fiscal, the Indian software exports have gone up to gross Rs 28,350 crore (US$ 6.2 billion) of revenue and registered a 65 per cent growth in rupee terms over revenues of Rs 17,150 crore (US$ 4 billion) in 1999-2000. The growth rate in dollar terms was 55 per cent," said Vandrewala.

The survey indicated that the industry registered an overall growth of 55 per cent during 2000-01, up from the existing Rs 24,350 crore (US$ 5.7 billion) in 1999-2000. Out of the total revenue of Rs 37,760 crore (US$ 8.26 billion) during 2000-01, software exports increased a total of Rs 28,350 crore (US$ 6.2 billion) and the domestic software market fetched a total of Rs 9,410 crore (US$ 2.06 billion).

The IT software and services industry currently accounts for almost 2 per cent of India's GDP. As per the -McKinsey Study 1999, the Indian software industry will account for 7.7 per cent of India's GDP by 2008.

The report said that the domestic software market grew at the rate of 31 per cent as opposed to 45 per cent in the year 1999-2000. However, the proliferation of the Internet, e-business, WAP-enabled technologies and growth in the SOHO market will result in higher growth rates in the domestic market in the years ahead.

The survey revealed that the major factors which continue to hinder growth of the Indian software industry are continuance of physical bonding at STP, EOU and EPZ units, lack of global parity in telecom tariff, inadequate telecom infrastructure and issues of clarity in Section 10A/ 10B.

As per the Nasscom survey the other important area that has emerged to be tapped is the IT Enabled Services or "Remote Processing". This covers a wide gamut of services including Customer Interaction Services, Help Desks, Medical Transcription/Translation Localisation Services, Data Digitisation, Legal Databases, Data Processing, Back Office Operations, Digital content development, Animation, Remote Network Management, Specialised Knowledge Services etc.

According to Vandrevala, the Indian IT-enabled services sector has clearly emerged as a key driver of growth for the Indian IT Industry.

The IT-enabled industry currently employs 70,000 people and accounts for 10.6 per cent of the total IT software and services industry revenues.

Courtesy : Sh. V.N. Shukla
Director Special Applications
Centre for Development of Advanced Computing
Anusandhan Bhawan,
C-56/1 Sector 62,
Noida -201301
Tel. : 0120-2402551
E-mial : vnshukla@cdacnoida.com