TDIL Programme: A Government Initiative

Ajit Kumar*, Vishal Goyal**
*Multani Mal Modi College, Patiala,
**Department of Computer Science, Punjabi University, Patiala
*ajit8671@gmail.com, **vishal.pup@gmail.com**

Abstract: India is a country with huge population of over hundred and twenty crore, who speak different languages. Only 5% of Indian population can effectively communicate in English and rest 95% are comfortable with their regional languages and are deprived of the benefits of information technology. To penetrate the benefits of Communication and Information Technology up to common masses, Ministry of Communication and Information Technology, Government of India, initiated a Programme known as Technology Development for Indian Languages (TDIL). The objective of the programme is to develop Information Processing Tools and Techniques to facilitate human-machine interaction without language barrier; to create and access multilingual knowledge resources; and integrate them to develop innovative user products and services. This paper surveys the various initiatives taken under TDIL for the promotion of Communication and Information Technology in Indian Languages. We are also providing an introductory list of tools/products developed by different institutions for Indian languages.

1. Introduction: The Technology Development for Indian Languages Programme was initiated by The Ministry of Communication and Information Technology, Government of India in 1991 with the objective to develop information processing tools to facilitate human-machine interaction in Indian Languages and to create multi-lingual resources to promote the use of information technology tools for languages research and development of information technology products for the use of common man in their native languages.

The Programme also promotes Language Technology standardization through active participation in International and national standardization bodies such as ISO, UNICODE, and World-wide-Web consortium (W3C) and Bureau of Indian Standards (BIS) to ensure adequate representation of Indian languages in existing and future language technology standards.

The TDIL Programme has taken up research and development work for all the 22 official languages i.e. Assamese, Bangla, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Manipuri, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu and Urdu. A lot of work related to font development, code converter, dictionary development, Optical Character Recognition, messenger, browser, corpora and editors is compiled in the form of CDs. These CDs containing various software tools like Bharteeya Open Office, Open Type Fonts, Keyboard Drivers, Firefox Web Browser, E-mailing Client, etc. are freely available at www.ildc.gov.in and www.ildc.in on request from the Ministry of Communication and Information Technology, Government of India.

Large number of Educational Institutions, Universities, Public and Private Business organizations are part of TDIL program and are involved in the development of various tools for Indian Languages.

2. Research Projects under TDIL Programme

2.1 Machine Aided Translation (MAT): In machine translation system one natural language is get translated into another natural language without human intervention or with minimal human efforts. Various softwares developed for machine translation are:-

A) Development of English to Indian Languages Machine Translation System (**Anuvadak**) for English to {Hindi, Marathi, Bangla, Oriya, Tamil, Urdu} Since majority of Indians cannot understand English so TDIL taken up machine translation project from English to Indian languages under consortium mode. Four machine translations technologies namely, Tree-Adjoining-Grammar (TAG) based MT, Statistical based MT (SMT), Analyze & Generate rules (AnalGen) based MT, and Example Based MT (EBMT) merged together to develop Anuvadak. Initially the software is domain specific for tourism and is being enhanced in phased manner.

B) Development of English to Indian Languages Machine Translation (MT) System with **Angla-Bharti** Technology for English to {Bangla, Punjabi, Malayalam, Urdu} AnglaBharti represents a machine-aided translation methodology specifically designed for translating English to Indian languages. Angla-Bharti uses pattern directed approach using context free grammar like structures. It analyses English only once and creates an intermediate structure called PLIL (Pseudo Lingua for Indian Languages). The PLIL structure is then converted to each Indian language through a process of text-generation. There is a provision for automatic pre-editing and paraphrasing, recognition of named-entities and incorporated an error-analysis module and statistical language-model for automated post-editing. The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form, which is more easily translatable. The project had been implemented in consortium mode with four institutions are participating to build the system.

C) Development of Indian Language to Indian Language Machine Translation System (**Sampark**) for Punjabi to Hindi, Hindi to Punjabi, Urdu to Hindi and Telugu to Tamil. The Sampark system is based on analyze-transfer-generate paradigm. Sampark uses Computational Paninian Grammar (CPG) approach for analyzing language and combines it with machine learning. Thus it uses both traditional rules-based and dictionary-based algorithms with statistical machine learning.

2.2 Development of Cross-lingual Information Access (CLIA): Cross Lingual Information Access (CLIA) systems make it possible for users to directly access sources of information which may be available in languages other than the language of query. The project is being implemented in consortium mode and eleven institutions are participating to build the system. At present, nine languages are being targeted under Tourism and Health domain: - Assamese, Bengali, Gujarati, Hindi, Marathi, Oriya, Punjabi, Tamil and Telugu.

2.3 Development of Robust Document Analysis & Recognition System for Indian Languages (OCR): OCR is required to convert scanned documents to editable text. The project is being implemented in consortium mode. The fifteen scripts/languages being targeted are: - Assamese, Bengali, Bodo, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Tamil, Telugu, Tibetan and Urdu

2.4 Development of On-line handwriting recognition system (OHWR): There are seven institutions participating to build the On-Line Handwriting Recognition System. The eight languages being targeted are: - Assamese, Bengali, Devanagari, Gurumukhi, Kannada, Malayalam, Tamil and Telugu

2.5 Development of Text to Speech System for Indian Languages (TTS): Consortium Mode Project has been initiated to develop Text-to-Speech system in six Indian Languages Hindi, Bengali, Marathi, Malayalam, Tamil and Telugu languages.

2.6 Development of Automatic Speech Recognition in Indian Languages (ASR): Consortium Mode project has been initiated for development of Automatic Speech Recognition system for accessing prices of agricultural commodities through telephone channel as an interface on NIC website, which is multilingual and provides information on agricultural commodities.

2.7 Development of Sanskrit Machine Translation System: In India, there have been several efforts in the development of computational tools for Sanskrit. Under the leadership of University of Hyderabad a consortium Mode project has been initiated with the objective to develop Sanskrit computational tools and use them to develop machine translation technology from Sanskrit to Hindi.

3. Software Products for Indian languages

Joint activities carried out by various Academic Institutions and Business Organizations under TDIL programme has resulted in the development of large number of tools and software's for people who like to work in Indian languages. An introductory survey of these tools is given in this paper.

3.1 Products developed by CDAC Pune: ALP (Apex Language Processor): It is a multilingual wordprocessor running under MS-DOS and UNIX; this allows typing of all Indian scripts through the common INSCRIPT keyboard overlay. It has Wordstar compatible commands. It can be used by user comfortable with Wordstar. ALP supports various Indian scripts such as Assamese, Bengali, Devanagri, Gujarati, Kannada, Malayalam, Oriya, Punjabi, Tamil, Telugu and other scripts such as Tibetan, Diacritic Roman, Bhutanese and Sinhalese. **Anveshak:** A Natural Language based Information Retrieval system which can efficiently and accurately provide explicit information in natural language text to the question intended to be queried on a certain document. **Chitrankan :** Chitrankan archives Indian Language content in electronic form through OCR. It enables the user to take a book, magazine or printed text in an Indian Language, feed it directly into an electronic computer file, and edit the file. **GIST CARD:** The C-DAC GIST Card is a PC add-on card that allows the use of Indian & other scripts, with English, in all existing text-based application packages like dBase, FoxPro, Lotus 1-2-3, QBasic, and compilers like C, C++ and Clipper on MS-DOS. It is available with its own WordStar compatible, customized multilingual word-processor called 'ALP For GIST Card'. The C-DAC GIST Card is ideal for multilingual database applications on MS-DOS, wherein the database can be entered, processed, sorted, displayed and printed in any Indian language with English. **GIST Card , Related Products & GIST Terminals :** A versatile plug-in card developed for IBM compatible PC/XT/AT for using any Indian language with English. It provides facility for instant transliteration between Indian languages. **iLEAP:** It is an Intelligent, Internet ready, Indian language Word Processor for Windows platform. iLEAP is from the LEAP range of products, developed in collaboration with Mithi.com Pvt. Ltd. This software has been created keeping in mind the requirements of individual users. This is an independent application with a very simple user interface. **iplugin :** iPlugin is an Web application development tool, used to develop interactive Indian language applications to be deployed over the internet or intranet offering technology solutions, iPlugin 3 from GIST enables a user to realize the full potential of the Internet in the creation of static as well as interactive Indian language web content for front-end & back-end Internet solutions. In this way it enables communication among users in the language of their choice. It is easy to enable existing English site for Indian Languages, without changing look and feel of the webpage. **ISM :** Intelligent Script Manager, a Windows based Software using which we can work through Indian languages on MS-Office, PageMaker, CorelDraw etc. Unicode Compliant and has Open Type fonts support. Supporting 19 Indian languages besides English, the ISM range has a colossal collection of designs, clip arts and is packed with a multitude of user friendly features. **ISM 2000 OFFICE :** A Word & Data Processer, On-Line communication used for Multilingual office. **ISM 2000 PUBLISHER** Advanced Multilingual Publishing Solutions. **ISM 2000 SOFT:** Used for Web publishing, database applications as well as for working on windows applications. **LEAP OFFICE 2000.** It is Web Base Indian language software for office applications. It provides multilingual Spellchecker, keyboard shortcuts, online keyboard and official language dictionaries tools. **LILA:** A tool for Non-Hindi speakers who wish to learn Hindi from the basic to the advanced stage. **Lila Hindi Parveen:** A multimedia based intelligent self-tutoring software package for learning Hindi as a second language. It is a specially designed for Government & Corporate employees. **Lila Hindi Pragya:** Self-tutoring software package for learning Hindi as a second language. It focuses on advanced functional Hindi and facilitates in handling the professional, official correspondence and transactions. **Lips:** The pioneering technology that enables Indian language subtitling for the entertainment world. **LISM:** Linux based application for Indian languages. **Mantra:** Translates the English text into Hindi in a specified domain. **Shaili:** A Rich collection of computer generated designs, inspired by the traditional Indian art forms. **Talaash:** Searches for content in Indian languages using English or Hindi over the Internet. These products are available at http://pune.cdac.in/html/gist/products/.

3.2 The products from CDAC Noida. Address Management System : It provides a very user-friendly interface with message and help commands in English as well as in Hindi for address management. **Bilingual Electronic Dictionaries** This electronic dictionary provides the corresponding meaning in Hindi and other information for Language Learning like synonyms, category information, example usage in source and target language and all variation of input word with ease of using. Also provides facility for entering new words in the database and facilities fast retrieval of information. **Lekhika :** A platform independent bilingual word processor in Hindi and English. **On-line Bilingual IT Terminology** : This terminology, categorized in various subjects for easy navigation and downloading is being made available on the Internet. It also contains tools and translation facility. **On Line Hindi Vishwakosh**: It collection of more than 12,000 topics in Hindi with search facility in the web site. The information is available in alphabetical order as well as category wise. These products are available at http://www.cdacnoida.in/

3.3 Products developed by CDAC Mumbai: INDIX: It is a tool for Indian script processing, to localize various applications in Indian languages, various GUI elements like menus, labels, messages, etc. appear in local language depending on the current locale setting, to provide multi-lingual support in various existing applications without any need to modify or recompile them, under X Window system on Linux platform. **MaTra**: Matra is a translation demonstration prototype which can perform translation (from English to Hindi) of certain classes of simple sentences. **Mulyaankan** :Mulyaankan is a Data Mining software aimed at detecting anomalies in valuation of imports. These products are available at http://www.cdacmumbai.in/projects/indix/

3.4 The products developed by VSOFT Services Pvt. Ltd. APS Corporate 2000++ : It is Windows based multilingual interface in Indian languages and provides support for Ms-Office (Word, Excel, Power Point), Lotus Smart Suite, FoxPro, PageMaker, CorelDraw etc. The product is developed to satisfy the office automation needs of Corporate, Government and Banks. **APS Designer 5.0** : It is multi-script Software for Windows with all language support, typing tools, spell check, diacritical and Vedic fonts, Border fonts etc. It is developed for print and publishing industry. It provides multilingual support for Page Maker, Photoshop, Indesign, QuarkXPress, CorelDraw, MS Word, Excel, PowerPoint, etc. **APS Desk.** : A set of tools developed by VSOFT Services Pvt. Ltd. for developers. It contains tools like Active X controls for language programming. **MAYA:** It is a dynamic font filter to provide indian language support. Usefull for web developers and web designers. **YAHI HAI INDIA:** A collection of Multimedia, clip arts, photographs. These products are available at: http://vsoftsolutions.in/.

3.5 The products developed by Modular Infotech Pvt. Ltd.:Ankur Professional: It is a Complete Multilingual Office Suite with user interface in all Indian languages. User Interface Screens, Menus, Messages, Hints, Helps etc. are in all local language as well as in English. User can select languages like Marathi, Assamese, Bengali, Gujrati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, and Telugu as User Interface language along with English. **Shree Lipi** : Shree Lipi is the set of fonts for Devanagari, Gujarati, Punjabi, Bengali, Assamese, Oriya, Tamil, Kannada, Telugu, Malayalam, Sindhi, Sanskrit, Sinhalese, Russian, Arabic, etc. **Shree-Lipi Samhita** It is a toolkit for application developers who want to develop applications in Indian languages on Windows platform. Shree-Lipi Samhita can be integrated with applications developed in VC++, Visual Basic, Visual FoxPro, Borland Delphi, Power Builder etc. The developers can very easily develop applications to give good Indian language interface to the users who are more comfortable with their mother tongue than English. **Vividha** : Vividha is a windows based package having a Word Processor, Indian fonts and a lot of Cliparts. When it comes to a multilingual word processor suitable for a variety of word processing requirements. Vividha package is available for English and Indian languages i.e. Devnagari, Gujarati, Bengali, Assamese, Punjabi, Oriya, Tamil, Kannada, Telugu, Malayalam etc. It contains multilingual word processor, clipart, wallpapers, keyboard generator and keyboard tutor. These products are available at http://www.modular-infotech.com/.

3.6 Products developed by Summit Information Technologies Pvt. Ltd.: **E-Indica:** A web based solution for typing in Indian languages. It incorporates Intellikeys that makes languages typing an intuitive process using English Keyboard. **Hindi Spell Checker :** Online and integrated spell check facility in Word, Quark and PageMaker for Hindi. Customization, Underlining of incorrect word and suggestions are available. **Indica :** Product includes Fonts, Tools, Keyboard Layout and Translation and Web based Solution. It provides supports for Page Maker & almost all window based programs. It also have Vedic fonts for Sanskrit. **Indica 2000** : Product includes Fonts, Tools, Keyboard Layout and Translation and Web based Solution for Indian language software that allows user organization to do text and data processing with indexing, searching and sorting capabilities - with a multilingual advantage. **Indica Unicode** : Product include Fonts and Keyboard Layout for Indian languages. It works with a wide range of application such as MS Office, MS Outlook and Outlook Express, Quark Xpress, Adobe Collections, Open Office, Final CutPro etc. These products are available at http://www.summitindia.com/

3.7 Products developed by Artech India. Akshar Naveen 3.0: The software has changed the way Indian language word processing is seen and done in India. The product has added features of Unicode/non Unicode editing and printing, enhanced font converter, dictionary, database creation application, web browser and other many such features. Akshar Naveen 3.0 supports eight Indian regional languages and provided more compatibility. It is available in English, Hindi, Bangla, Punjabi, Gujarati, Oriya, Tamil, Telugu and Malayalam. **Akshar Office:** Akshar Office is a complete office suite which includes key desktop applications such as word processor, spreadsheet program, presentation manager, and drawing program. It is compatible for small, medium and large businesses and works transparently with a variety of file formats. It has Inbuilt Email Facility, PDF Converter, Web Publishing , Graphics Facility and several other features. These and other many products from Artech India are available at http://www.artechinfo.in/ProdOurProducts.html.

3.8 Products from M/s. CK Technologies Pvt Ltd: **SOaccess:** It is database management software to create and modify tables, add or modify records, write queries etc. **SOcalc :** A spread sheet tool for easy formatting, cell referencing, easy sorting facility. **SOmail:** Web Based Solution to organize mails, search by name, subject, message etc., Encryption, import/export options, features to block spam. **SOnet :** A Web Editor, easy to insert images and hyperlinks. **SOshow :** A Presentation Program, insert tables, insert graphics, insert audio, insert video clips, image Preview. **SOwrite :** A word processor having Bilingual Spell Checker, Mail Merge, Sorting Facility, Add tables, pictures, objects. **Typing Tutor :** This Tamil typing tutor will teach how to input Tamil using Typewriter, Phonetic methods as defined by Tamilnet99 standards. **Shakti Office** : It has a word processor, spreadsheet, database management software, presentation, Webpage Design Tool and an email client. These products are available at http://www.shaktioffice.in/.

3.9 Tally.ERP 9 : New Product from Tally Solutions Pvt. Limited. It has grown from a basic accounting package into a simple-yet-sophisticated business management software product. The languages currently supported are Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu and English (UK).

3.10 Machine Aided Translation System (MAT) : Developed by IIT, Kanpur & CDAC, Noida. This translation tool based on Angla Bharti approach performs translation from English to Hindi in public health campaign domain. The package offer 85% parsing and about 60% correct translation output. It also provides the facility for editing incorrect sentences. It is available at http://cdacnoida.in/SNLP/machine_translation.asp

3.11 LooKeys :The Lookeys Pro software is developed by FTK Technologies Ltd. The software provides facilities for chat, e-mail and on-line word processing in Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam, Punjabi and Urdu. The software is available at http://www.lookeys.com/

3.12 ScriptMagic : ScriptMagic provides Indian Language interface for existing English software. It is developed by Image Point Technologies Private Limited. This software changes the interface of existing English software without touching their source code. It is available at http://www.imgpoint.com/

3.13 Line Matrix printer : The hardware printing solution for Hindi and regional languages is developed by Lipi Data Systems Pvt. Ltd. It uses GIST card technology. These printers can be ordered at http://www.lipidata.com/line_matrix.htm

3.14 TLX 10 : Teleprinters and Telex bilingual products for Indian army, Indian Air force and Indian railways are developed by Databyte Equipments Pvt. Ltd. Information about these products is available at http://www.databyteindia.com/telecom_products.html

3.15 FontSuvidha - Devnagari Font Converter : FontSuvidha supports over 60 Font Formats including UNICODE & ISCII and Convert documents from one format to another. Tool is developed by Cyber Shoppee (IL Infotech Pvt Ltd) and is available at http://www.cybershoppee.com/

3.16 DHVANI : DHVANI is the Text-to-Speech engine developed by Simputer Trust. The aim of this tool is to ensure that the knowledge of English is not essential for using this tool. The tool uses images in conjunction with voice output in local languages this makes the Dhavani accessible to a larger fraction of the Indian population. Dhvani has a Phonetics-to-Speech engine which is capable of generating intelligible speech from a suitable phonetic description in any Indian Language. It is also capable of converting UTF-8 text in Hindi or Kannada to this phonetic description, and then speaking it out using the Phonetics-to-Speech engine.

3.17 Lotus Notes in Hindi and Tamil: Developed by IBM India Ltd. And provides support for two primary scripts-Devanagari and Tamil.

3.18 Panini Keypad: The software developed by Luna Ergonomics Pvt Ltd for Java enabled phone, Android platform and iPhones. With the help of Panini Keypad the user can type in eleven Indian languages. The software is available at http://www.paninikeypad.com/

3.19 Gyandoot : Gyandoot is an intranet in Dhar district connecting rural cybercafe catering to the everyday needs of the masses. This web site of Gyandoot is an extension of Gyandoot intranet, for giving global access.

3.20 Surabhi Tools: It is a collection of tools developed by Apple Soft to support Indian Languages on the Tools of MS Office running on MS Windows. This includes tools such as Sorting, Text conversion, Auto correct, Date and Time, Numerals to text etc., for Publishers, DTP centers, Web content developers, Advocates, Bankers, Public sectors, Government Departments, Teachers, Students etc. These tools can be downloaded from **surabhi-*tools*.*software***.*informer.com/ but Apple Soft do not have their official website.*

3.21 www.webdunia.com: It is news portal in Hindi, Tamil, Telugu, Gujrati, Kannada, Marathi and Malyalam. It provides fatures like e-mailing, searching, providing information etc.

3.22 www.raftaar.com: It is Hindi search engine and provides all information in Hindi.

4. Publication: Technology Development for Indian Languages (TDIL) programme has been publishing TDIL Information Brochure / News Letter: Language Technology Flash: VishwaBhart@TDIL since the year 2000. The above Newsletter is widely circulated amongst the researcher, scientists and officials involved in Language Technology researchers in India as well as abroad. Currently it is published bi-annually and it consolidates in one place information about products, tools, services, activities, developments, achievements in the area of Indian Language software. It serves as a means of sharing ideas and creating awareness among technology developers. It has received wide appreciation for its useful contents.

5. Conclusion: With the constructive efforts of Government of India through Ministry of Communication and Information Technology a conducive environment for research in Indian Languages has been created. A large number of public and private institutions are involved in Technology development for Indian Languages and have produced large number of products, services, standards and tools. There is a lot of scope of improvement in existing technologies and serious collaborative efforts required.

6. References:

1) "Local Language Information Technology Market in India",Study Conducted by, Frost & Sullivan for MAIT- COIL Tech Under the aegis of TDIL, Department of IT, Ministry of Communications and Information Technology 2003 Frost & Sullivan, www.frost.com
2) http://tdil.mit.gov.in/
3) http://pune.cdac.in/html/gist/products/.
4) http://www.cdacnoida.in/
5) http://www.cdacmumbai.in/projects/indix/
6) http://vsoftsolutions.in/.
7) http://www.modular-infotech.com/.
8) http://www.summitindia.com/
9) http://www.artechinfo.in/ProdOurProducts.html.
10) http://www.shaktioffice.in/.
11) http://www.tallysolutions.com/website/html/tallyerp9/tallyerp9-main.php
12) http://cdacnoida.in/SNLP/machine_translation.asp
13) http://www.lookeys.com/
14) http://www.imgpoint.com/
15) http://www.lipidata.com/line_matrix.htm
16) http://www.databyteindia.com/telecom_products.html
17) http://www.cybershoppee.com/
18) http://www.simputer.org/simputer/downloads/software/dhvani/
19) http://www-01.ibm.com/software/lotus/products/notes/
20) http://www.paninikeypad.com/
21) http://gyandoot.nic.in/gyandoot/intranet.html