TDIL MEET 2001'
on January 23-25 at
Ministry of Information Technology

TDIL Meet 2001, was held at Ministry of Information Technology recently on January 23-25. It was organized into twelve sessions spread over three days. Each session was chaired by an eminent person in the field of Indian Language Technology. It was attended by over 150 experts from Academia, Industry and Government. There was an Exhibition also in which over 25 institutions displayed various language technology products. Besides the presentations by Industry and 13 Resource Centres for Indian Language Technology Solutions, there were technical sessions on Standardization, Lexical resources, Machine Aided Translation, Knowledge Processing Tools, OCR and Speech Technology, Language Technology Marketing. The Panel Discussion focussed on the theme: India as a Global Player in Language Technology.

Session 1:Inaugural Session

The TDIL Meet was inaugurated by Secretary MIT and the News letter VishwaBharat@tdil was also released by him. This News letter disseminates information about the TDIL programme and the parallel developments in the area of IT for Indian languages in the Private sector and also the efforts being made by the Multi-nationals in this direction. The enhanced version of TDIL website, http://www.tdil.gov.in with the Hindi version was launched by Secretary, MIT.

Dr. A.K. Chakravarty, Adviser, MIT welcomed the participants. Dr. Om Vikas Senior Director, MIT, presented the Genesis of the TDIL programme highlighting the objectives and achievements of the programmes since its inception in 1991. He also informed that the programme has been upgraded to the Technology Mission with long term and intermediate goals. He categorised the development of Indian Language Technologies in 3 phases, namely –

1976-1990 : A-Technology phase with focus on language technology acquisition and adoption;

1991-2000 : B-Technology phase with focus on development of basic language-technologies;

2001-2010 : C-Technology phase with focus on convergence and creative technologies and collaborative development approach.

Secretary, MIT, during his inaugural speech emphasized that 21st century commences with information revolution leading to knowledge based communities, hence, collaborative development is a norm of the future. The role of Indian languages has become important in promoting IT in various sectors of economy. IT in local languages is a tool for ensuring the technology absorption by the society for sustainable development. He felt that the TDIL programme has matured over decade, and he requested the experts to deliberate on the circulated draft paper on Indian Language Technology Vision 2010, and come out with the comprehensive and implementable document.

In his key note address, Dr. S. Ramani, Director R&D, SilverLine Technologies highlighted the importance of the Technology for Machine Translation. He also gave parallel examples of translation results obtained from web based tools for European languages which were quite satisfactory. However, a huge amount of funding and manpower effort has gone into development of these systems. The Cross Language information retrieval on the web is the goal in this direction. But it involves intensive R&D and 15-20 times the funding in Machine Translation for achieving visible results. The need for patenting cannot be undermined but the patenting procedures in India are long and costly. Also the same software ported on a different language or platform does not get due protection under the patent. If the Indian language technologies are patented abroad, it does not provide the adequate protection in India hence, it is pre-mature to focus on this issue. We need to focus on customer centric developments and open systems conversion and integration to products.

Session 2 : Visit to Exhibition

Secretary, MIT inaugurated the Exhibition put up by the Industry, TDIL projects, Resource Centres for Indian Language Technology Solutions and the professionals working in this area. He took keen interest in various products displayed in the exhibition and interacted with the developers.

TDIL Working Group Expert Members critically saw the demonstration of the TDIL Webserver and desired continuation of this project.

Session 3 : Unicode, National Standard for Indian Scripts and Certification

Dr. S.P. Mudur Chaired the Session and highlighted the

need for interoperability and data transfer between ISCII and Unicode and their co-existence.

Unicode - Shri Sunil Khullar, VP, Summit India

E-mail : skhullar@summitindia.com Tel : 011-6517994

Shri S. Khullar briefly described the Unicode, the storage formats supported by it, the Unicode support in the Indian language applications and its benefits in the globalization of the existing applications. He also clarified some of the prevailing misconceptions about Unicode and other issues involved in migrating ISCII applications, existing Indian language databases and web applications to Unicode.

ISCII - Dr. S.P. Mudur, Associate Director, NCST, Bombay

E-mail : mudur@ncst.ernet.in Tel : 022-6201606

Dr. Mudur explained the anomalies regarding the representation of ISCII in Unicode as the changes made during 1991 in the ISCII standard have not been incorporated in the Unicode. He informed the participants that MIT have become a voting member of Unicode and the suggestions regarding any changes to be made in the new version of Unicode are welcome by MIT.

Certification of Indian Language Products - Shri U.K. Nandwani, STQC, MIT

E-mail : nandwani54@hotmail.com Tel : 011-4363378

Shri Nandwani briefed about the various parameters which are tested in the case of software products and such a testing is being carried out at STQC lab Chennai. The product certification is done as per the claimed specifications and the industry can approach STQC for this purpose.

Session 4 - Industry Presentations

The session was chaired by Shri Viney Deshpande, President, MAIT. He emphasized the role of Indian language industry in the mass penetration of computers in our multilingual country. Recognizing this, MAIT is taking steps to organise this industry to increase their productivity.

Indian language Industry consortia - Shri Vinnie Mehta, Director, MAIT

E-mail : mait@vsnl.com Tel : 011-6855487

Shri Vinnie Mehta, Director, MAIT, presented the concept of Indian Language Industry consortia and made an appeal to the industry to join hands so that the issues specific to the Indian languages could be sorted out under this umbrella. This will ensure uniformity in following standards, addressing new applications like e-governance in Indian languages and also spreading the use of IT in the education using local languages. He also floated the concept of virtual R&D through collaborative efforts with the cooperation of MIT & DSIR.

Modular InfoTech, Pune - Dr. M.N. Cooper, MD

E-mail : modular@giaspn01.vsnl.net.in Tel : 020-4226614

Dr. Cooper elaborated his research efforts for the last 15 years and expressed difficulty in sustaining in the area of IT products for Indian languages due to small market size and piracy problems. He also briefed about wide variety of application offered by their company ranging from word processing, database applications to web based applications and customized applications.

Summit India - Shri Rakesh Kapoor, MD

E-mail : rkapoor@summitindia.com Tel : 011-6517994-98

Shri Kapoor described their latest product INDICA 2000 which is fully Unicode compliant and is a very versatile application which can be used as a single substitute for many of the dedicated applications currently available in the Market. He also made an appeal to the industry to move to Unicode to ensure that all the global applications become available for Indian languages without re-inventing the wheel.

Microsoft India - Shri Nitin Singal

E-mail : nitinsin@microsoft.com Tel : 011-6294600

Shri Singal demonstrated the Windows 2000 in Hindi and also described the ease with which the languages can be switched by changing the locale. With the availability of Windows 2000 in Hindi, the other development tools automatically get local language support through the operating system. Hence, it has become very easy to work on Microsoft Word, Excel and Access in Hindi and also Tamil. It is also possible to programme Visual Basic in Hindi.

IBM India - Shri Abhijit Dutta

E-mail : dabhijit@in.ibm.com Tel : 080-5262355

Shri Dutta presented their recent development of Lotus notes in Hindi. He navigated through the various menu options and demonstrated the total display of menus, errors, commands and messages in Hindi. He thanked MIT for helping IBM in improving upon the Hindi translation which still needs further improvement and is being worked on for the new version. He informed that IBM has a special group to address to the localization needs of India with regard to IBM products and other customized solutions.

24th January, 2001 (Wednesday)

Session 5 - Knowledge Resources: Lexware & Corpora

Shri K.B. Saxena, Secretary, Department of Official Language (DOL), chaired this Session. He applauded the efforts being made under the TDIL programme and assured full support from DOL for this cause.

Collaborative Development of Lexware - Shri Durgesh Rao, NCST, Mumbai.

E-mail : durgesh@ncst.ernet.in Tel : 022-6201606

Lexware for Indian Languages is a very useful resource for R&D in the area of IT for Indian Languages but it is a gigantic task. A conference was held up at IIIT, Hyderabad from Jan. 5-8 in which an initiative to develop Lexware for Hindi through collaborative effort has been launched. NSCT is co-ordinating and the other participating agencies are UOH, Hyderabad, IIT, Hyderabad, NCST Bombay, IIT, Bombay etc. It was felt desirable to have a format of Lexware such that the Machine Translation Groups in India can use this data for their developments.

WORDNET - Dr. Pushpak Bhattacharaya, IIT Mumbai

E-mail : pb@cse.iitb.ernet.in Tel : 022-5767718

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.

Natural Language Processing and Vishvakosh - Shri V.N. Shukla, ER&DCI, Noida

E-mail : shuklavn@hotmail.com Tel : 011-914587717-27

He briefed about the activities of ER&DCI, Noida, especially in the field of Natural Language Processing. ER&DCI at Central Translation Bureau have deployed the Machine Translation System from English to Hindi developed at IIT Kanpur so that system can be tuned to their needs. Hindi Vishvakosh is being made available online. IT terminology in Hindi is also under development.

Session 6 - OCR & Speech Technology

Prof. B. Yagnanarayana (E-mail : yegna@speech.iitm.ernet.in Tel : 044-4458338) IIT Chennai Chaired the Session and apprised about the importance of Speech Technology in common man's life and also the importance of integrating OCR & Speech Technologies for the handicapped.

OCR in Hindi - Prof. B.B. Choudhary, ISI Kolkata

E-mail : bbc@isical.ac.in Tel : 033-5778085

Prof. B.B. Choudhary from ISI Kolkata presented the status report that OCR in Hindi with about 95 percent accuracy has been developed for few fonts and it is being discussed with C-DAC for commercialization.

OCR in Hindi - Ms. Rashmi Sharma, C-DAC

E-mail : rsharma@cdac.ernet.in Tel : 020-5652461

C-DAC had developed algorithms for Hindi OCR and the work got discontinued due to lack of funds and now they are tying-up with ISI Kolkata to bring to product stage.

Text to Speech Synthesis System - Dr. S.S. Aggarwal, CEERI, Delhi

E-mail : sagrawal@ceerid.ernet.in Tel : 011-5781467

He briefly described the achievement in the area of voice recognition system dedicated for the Wheel Chair Application, Speech Data Bases. He also informed that the Windows version with online reading capability text to speech synthesis system in Hindi is in advanced stage of development.

Session 7 - Machine Aided Translation

Shri N. Gopalaswami, Secretary, Human Rights Commission Chaired the session and emphasized the need for larger funding and focused approach in this area.

Angla Bharti and Anubharti, Multilingual Machine Aided Translation Systems - Prof. R.M. K. Sinha, IIT Kanpur

E-mail : rmk@iitk.ac.in Tel : 0512-598254

Prof. R.M.K. Sinha explained the pattern directed rule based approach with context free grammar which generate a pseudo-target which is applicable to most of the Indian languages for translation from English. Right now the system is functional for English to Hindi for specific domain of public health campaign, however, with incremental effort the system can be extended to other languages, and domains.

Anubharti approach is an example-based approach. The corpus consists of examplebase of source language to target language translation pairs along with the mappings between the words of the source language sentence and the target language sentence.

Hybridising both the approaches can give very promising results, Prof. Sinha mentioned.

ManTra : English to Hindi Machine Translation System - Dr. M K Pandey, C-DAC

E-mail : mahendra@cdac.ernet.in Tel : 020-567009

Mantra has been developed based on tree adjoining grammar-based approach for a specific domain of administration. It has a good graphical user interface and is being tested at five Ministries.

MaTra : English to Hindi Machine Translation System - Shri Durgesh Rao, NCST

E-mail : durgesh@ncst.ernet.in Tel : 022-6201606

Matra is being ported to web so that it can be put to use by the news agencies for translation of English news stories to Hindi.

Anuvadak : English to Hindi Machine Translation System - Dr. Anjali Rai Choudhary

E-mail : anjalir@ndf.vsnl.net.in Tel : 011-6465016

The Machine Translation outputs demonstrated were of very high quality and the approach as elaborated is to have context sensitive dictionaries and also a general-purpose dictionary. The context sensitive dictionary is looked up first and if word is not found, then the general-purpose dictionary is used for translation. This however needs MT evaluation.

Session No.8: Knowledge Processing Tools

Shri N Gopalaswami, Secretary, Human Rights Commission chaired the session also.

Gita Supersite : A Case Study - Prof. T V Prabhakar, IIT Kanpur

E-mail : tvp@iitk.ac.in Tel : 0512-590725

Gita Supersite aims at Indian Scriptures on the Internet. It is on Bhagavadgita, together with translations in Hindi and English. The text can be viewed in any one of ten Indian scripts or roman. Gita Supersite version-1 has business logic on client and is re-engineered on the server with dynamic fonts. This supports both static and dynamic content. The future work will include fly Font conversion, more audio and integration of this web site with ISCII search engine.

Development of Sanskrit Authoring System (VYASA) P Ramanujan, C-DAC, Bangalore.

E-mail : rama@cdacb.ernet.in Tel : 080-5584205

This includes multilingual editor, Vedic character set and accent-marker handling, complete database of all original source texts with elaborate retrieval facilities, quotation extraction, tracing and inserting texts/source information, morphological and syntactic analysis, search, sort, index and concordance utilities etc. More of the proven concepts are OCX based applications using GIST/SDK, Vedic inputs also handled, Online Gita reader in multiple scripts based on Java servlets, in Red Hat Linux 5.02.

25th January, 2001 (Thursday)

Session No.9 Presentations by ILTS Resource Centers

Shri Gautam Soni, Adviser, MIT chaired the session.

ZOPP : Resource Centre for Indian language Technology- Prof. N J Rao, IISc., Bangalore

E-mail : njrao@mgmt.iisc.ernet.in Tel : 080-3092377

ZOPP is group problem solving approach for decision making. ZOPP Workshop was organised on May 3-5, 2000 in Bangalore wherein Resource Centres for Indian language technology solution participated. A detailed participation analysis was made in smaller groups for discussion the concept of project planning matrix. Five common outputs were formulated as follows:

1. Development of portals
2. Training programs
3. Knowledge and Database to create
4. Development of Spell checker which is Unicode compliant
5. 10 books on web

This workshop recommended a Technology Workshop of C-DAC to enable the Resource Centres for speedy take up in respective activity.

Presentations by Resource Centers

Resource Centre for Indian Language Technology Solutions- Hindi & Nepali-Dr. Sanjay Dhande, IIT, Kanpur

Tel : 0512-598570

IIT, Kanpur Resource Centre described the development activities in the field of content generation, Text encoding, Linux tools, OCR, MT on Web and extension activities. Resource Centre interacts with UP Government for e-governance and plans to organise conference in February.

Resource Centre for Indian language technology solutions- Kannada-Prof A G Ramakrishnan, IISc Bangalore

E-mail : ramkiag@ee.iisc.ernet.in Tel : 080-3092378

The activities included the design of website knowledge base on Indian aesthetics. Knowledge base on Indian system, Sudarshana (Shad Darshanas in Sanskrit), OCR Technology for south Indian languages, Meta font for Kannada. Algorithms for Kannada speech system, OCR for Kannada.

Session 10 : Presentations by Resource Centres (Continued)

Shri R P Sinha, Principal Adviser, Planning Commission chaired the session.

On the lines of the above, other Resource Centers also made the presentations on languge technologies being developed for their assigned languages, i.e., Punjabi by Prof. G.S. Lehal; Urdu, Sindhi & Kashmiri by Prof. S.K. Mohanti; Telugu by Prof. Narayan Murthy; Malyalam by Prof. Ravindra Kumar; Tamil by Prof. Rajani Parthasarthi; Bengali by Dr. Ujjwal Bhattacharya; Oriya by Prof. A.K. Pujari and Prof. Sanghmitra Mohanti; Marathi & Konkani by Prof. P. Bhattacharya; Gujrati by Prof. Mandar Mehta; Assamese & Manipuri by Prof. G.B. Nair; Foreign Languages (Japanese & Chinese) by Prof. G.V. Singh.

At the end of all the Resource Centers presentations, Shri R P Sinha made a remark that MIT should prepare a road map and ensure of know-how transfer and sharing of results. There should be no duplication of work. Resource Centres should stress on development of Language technologies rather than open-ended research. These Centres should also share results and collaborate for R&D.

Panel Discussion on theme: *"India as global player in Language technology"*

The panel discussion was Chaired by Prof. V S Ramamurthy, Secretary, Ministry of Science & Technology. The panelist included Prof. R M K Sinha, (IIT, Kanpur), Dr. M.N. Cooper, (Modular Infotech, Pune), Shri Suraj Bhan Singh, Ex-Chairman, CSTT, (Prof N J Rao, IISc. Bangalore), Shri R K Arora, (C-DAC, Pune), Shri Vivek Singhal, (ESC Ltd.), and Shri Vinay Chhajlani, (Webdunia, Indore).

Shri Arora from C-DAC was of the view that Indian language technology development activities are possible only in India and nowhere else in the world. Ministry of IT and Department of Official Language must continue funding the Indian language technology R&D projects. There is need to involve third party software in Indian languages. Indian languages are derived from Sanskrit and Hindi is spoken in several parts of world. Although we are not today even the national player but we must take advantage of Indian languages being spoken abroad. He clarified that C-DAC fonts are available on TDIL Website.

Prof R M K Sinha, IIT, Kanpur emphasized that India has become national player and is poised to play the role as regional player and global player. India has skilled persons. We remained problem solver but now we should provide solutions. There are certain hurdles which need to be carried out with regard to standardization and availability of tools in public domain. Some major initiative such as Cross Lingual Information Retrieval should be launched.

Dr. M N Cooper of Modular Infotech felt that we are not yet national player and lot more needs to be done especially with regard to the standardization of glyphs and fonts. Common man will benefit from standardization. He observed that India has funded very little in terms of R&D for Indian language in comparison to language technology R&D in other countries such as Europe, Canada, Japan & China.

Prof. Suraj Bhan Singh opined that language develops on technology demand. There is need for standardization of language.

Prof N J Rao optimistically said that India can emerge as global player of IT industry. US$ 2B out of US$ 1 trillion is the share of language technology products. It could be used for wealth creation. Over 95% people feel alienated in our country because of their non-literacy in English. Hence there is need to develop language technology and make them available. There is vast scope for applications of speech and language products. MNCS are exploring the vast rural market. Creating content on heritage and tools for day to day work may be promoted. Implementable road map needs to be prepared.

Shri Vinay Chhajlani, (Webdunia) felt that India is getting IT rich and wealthy. Information is not distributed to the needed ones and hence digital devide. CISCO was Technology Company and it is now eventually consumer driven company. Can we use all such technologies for the people at large? It is to note that about 300 Indian words have been added in the latest edition of Oxford dictionary. There is need to disclose technologies developed at various institutions/organizations and to permit collaboration with industry. There is also need to get feedback from market. He advocated for 20% of venture capital for technology development for Indian technology.

Shri Vivek Singhal felt need for standardization and policy paper from Ministry of IT. Government should fund Private sector as well. Technology business meets need to be organized for sharing technology and transfer of know-how. Database of resource persons, Centres and technologies need to be prepared.

There were comments on the Universal Language project involving 18 countries to enable information exchange over Internet. Presentations are organized in the participating countries. Language technology has different aspects. Government has not invested in primary resources, such as lexical resources, search engine, standardization of fonts and evaluation of language technology.

Prof Ramamurthy concluded that, language technology is an evolving technology we are not lagging behind we have technology groups. There is need for synergising technologies. There may be duplication in technology development until the technology matures. Role of Government is as facilitator only. India emerges as a potential global player in language technology and can provide technological tools and consultancy in multilingual computing to other countries as well.

Session 12: Language Technology Marketing and IPR

Shri Vinnie Mehta, MAIT chaired the Session.

Dr Mukul Sinha said that 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 should be encouraged for certification of language technology products. Technology development and investment should be encouraged in parallel groups. 5% of all the IT purchases in the Government should 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.

Dr. R C Tripathi explained about copyright issues in Language technology products/services. 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.

Shri Vinnie Mehta at the end felt that there is need to glamourize language technology products by organizing road shows and campaign. There is need to involve State Governments to use Indian languages in e-governance and other socio economic projects/programs.

UNESCO Expert Group on "Promotion and use of Multilingualism and Universal access to Cyberspace" on April 9-10, 2001 at UENSCO Headquarter, Paris

The experts from 24 countries from Africa, Arab States, Asia Pacific, Europe-I, Europe-II and Latin America participated in the meeting. Dr Om Vikas of Ministry of Information Technology participated as Indian representative. This was a follow up to the recommendation of Info-Ethics 2000 for developing mechanism to provide affordable and equitable access of information in public domain and access to telematics in all countries. 30 recommendations were made under four broad categories.

The report is available on the UNESCO website www.unesco.org. The report contains a preamble stating the genesis of this Workshop subsequent to the general conference of UNESCO on October 22 to November 9, 2000 at its 31st session. The technical terms such as Cyberspace, Digital devide, Telematics, Universal access have been defined in the chapter on definition. Basic principles are covered in the Chapter 2 wherein four key aspects in Cyberspace have been identified, namely:

(a) Provision of access to the telematics networks and services

(b) Promotion of multilingualism

(c) Provision of access to information in the public domain

(d) Application of exemption to copyright

Chapter 3 recommended measures : 9 recommendations have been made under category (a), 8 under category (b), 10 under category (c) and 3 under category (d). The recommendations under the category on Promotion of multilingualism are as follows:

M-10 Member State and intergovernmental international organizations should reaffirm and promote the respect and use of all languages in cyberspace, to contribute to the preservation of the richness and diversity of the universal human heritage and to peaceful coexistence, objectives and are enshrined in many international declarations and conventions and in many national constitutions.

M-11 Broaden language diversity in cyberspace by creating contents, and means to find, access and process them, in all widely used language as well as in other languages at the regional, national and local levels, including less used languages.

M-12 In order to prevent all forms of linguistic segregation in access to cultural and scientific information and knowledge, technical, financial and education resources should be provided by the public and private sectors at local, national regional and international levels to ensure the creation, preservation and maintenance of national and multilingual Web sites.

M-13 Member States and intergovernmental and non-governmental international organisations should adopt strategies to develop, and disseminate on-line, freely accessible language education materials.

M-14 Member States, international intergovernmental and non-governmental organisations and industries should encorage the participation of specialists in collaborative research and development on, and localisation (adaptation) of, operating systems, search engines and Web browsers with extensive multilingual capabilities, as well as the development of on-line dictionaries and terminologies, Software should preferably be developed and made available in an open source environment.

M-15 Member States and international intergovernmental and non-governmental organisations should support international cooperative efforts to develop automated translation services accessible to all, free or at a nominal charge, and to encourage the development of intelligent linguistic systems such as those performing multilingual information retrieval, summarizing/ abstracting and speech recognition.

M-16 Governments should formulate strong national policies on the crucial issue of language survival in cyberspace. International assistance to Member States in framing and implementing language policies, designed to promote mother tongues and language teaching, should be strengthened while respecting cultural diversity on the global information networks and reinforcing national and international solidarity.

M-17 International organisations, in particular UNESCO, should maintain and promote an international collaborative on-line observatory on the different existing policies and regulations relating to multilingualism and multilingual resources and applications. Such a portal