



2. Reports

2.1 TDIL Vision 2010

Vision statement

Digital unite and knowledge for all.

Mission statement

Communicating & moving up the knowledge chain overcoming language barrier.

Objectives

- ▶ To develop information processing tools to facilitate human machine interaction in Indian languages and to create and access 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)
- ▶ *Knowledge Tools*
(Portals, Language Processing Tools, Translation Memory Tools)
- ▶ *Translation Support Systems*
(Machine Translation, Multilingual Information Access, Cross Language Information Retrieval)
- ▶ *Human Machine Interface System*
(Optical Character Recognition Systems, Voice Recognition Systems, Text-to-Speech System)
- ▶ *Localization*
(Adapting IT Tools and solutions in Indian Languages)
- ▶ *Language Technology Human Resource Development*
(Manpower Development in Natural Language Processing)
- ▶ *Standardization*
(ISCI, Unicode, XML, TMX, ISFOC etc.)

TDIL Programme Goals

Short Term Goals

- ▶ Standardization of code, font, keyboard etc.

- ▶ Fonts and basic software utilities in public domain.
- ▶ Corpora creation and analysis
- ▶ Smart content creation.
- ▶ Language Technology be integrated into IT curricula.
- ▶ Collaborative development of Indian language lexical resources.
- ▶ Writing aids (Spell checks, grammar checks and text summarization utilities).
- ▶ Sharing of standardized lexware & development of lexware tools.
- ▶ Training programs on ILT awareness, lexware development, and computational linguistics.

Medium Term Goals

- ▶ Indian language speech database
- ▶ Multilingual, multimedia, content development with semantic indexing, classical and multi font and decorative fonts, offline/online OCR.
- ▶ Cross lingual information retrieval (CLIR) tools.
- ▶ Human speech encoding
- ▶ Speech Engine : Speech recognition, specific speech I/O.
- ▶ Indian language support on Internet appliances.
- ▶ Understanding and Acquisition of languages, knowledge representation, gisting and interfacing.
- ▶ Distinguished achievement awards for M.Tech/ MCA/Ph.D. level in Indian Language Technologies.
- ▶ Machine aided translation: English to Indian languages, among Indian languages, Indian languages to English and other foreign languages.
- ▶ On line rapid translation, gisting and summarization.

Long Term Goals

- ▶ Speech to speech translation.
- ▶ Human Inspiring Systems.



Resource Centres for Language Technology Solutions

The MCIT has established thirteen Resource Centres for Indian Language Technology Solutions covering all the constitutional languages.

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 some 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.
- ▶ 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.

Organizations and associated Languages

- ▶ Indian Institute of Technology, Kanpur. (Hindi, Nepali)
Prof. R.M.K.Sinha
E-mail : rmk@iitk.ac.in
- ▶ Indian Institute of Technology, Mumbai. (Marathi, Konkani)
Prof. Pushpak Bhattacharya
E-mail : pb@cse.iitb.ernet.in

- ▶ Indian Institute of Technology, Guwahati. (Assamese, Manipuri)
Prof. Gautam Barua
E-mail : gb@iit.ernet.in
- ▶ Indian Institute of Science, Bangalore. (Kannada, Sanskrit Cognitive Models)
Prof. N.J. Rao
E-mail : njr@mgmt.iisc.ernet.in
- ▶ Indian Statistical Institute, Kolkata. (Bengali)
Prof. B.B. Chaudhary
E-mail : bbc@isical.ac.in
- ▶ Jawaharlal Nehru University, New Delhi. (Foreign Languages Japanese, Chinese & Sanskrit Language Learning Systems)
Prof. G.V. Singh
E-mail : gvs@jnuiv.ernet.in
- ▶ University of Hyderabad, Hyderabad. (Telugu)
Prof. K. Narayan Murthy
E-mail : knmcs@uohyd.ernet.in
- ▶ Anna University, Chennai. (Tamil)
Dr. T.V. Geetha
E-mail : rp@annauniv.edu
- ▶ MS University, Baroda. (Gujarati)
Shri Sitanshu Y. Mehta
E-mail : rciltg@satyam.net.in
- ▶ Utkal University and Orissa Computer Application Centre (OCAC), (Oriya)
Prof. Ms. Sanghmitra Mohanty
E-mail : sangham1@sancharnet.in
Dr. A.K. Pujari
E-mail : akp@ocac.ernet.in
- ▶ Thapar Institute of Engg. & Tech., Patiala. (Punjabi)
Prof. G.S. Lehal
E-mail : gglehal@mailcity.com
- ▶ ER&DC, Trivendrum. (Malayalam)
Prof. Ravinder Kumar
E-mail : ravi@erdcitvm.org
- ▶ C-DAC, Pune. (Urdu, Sindhi, Kashmiri)
Sh. M.D. Kulkarni
E-mail : mdk@cdac.ernet.in



2.2 Proceedings of the Language Technology Business Meet (LTBM) and Technology Handshakes, on Nov.7-8, 2001

at Department of Information Technology
Ministry of Communications & Information Technology,
Electronics Niketan, 6 C.G.O Complex,
New Delhi



The participants consisted of 80 members from the Industry, 39 from Academia and 33 from Government.

I. Inaugural Session

Shri Rajeeva Ratna Shah, Secretary, Ministry of IT, declared the LTBM open on Nov. 7, 2001 by clicking the mouse displaying the message of bringing together the Government, the Industry and the Academia for "Digital Unite & Knowledge for all". Dr. Om Vikas presented the genesis of LTBM and mentioned the chronological development of Language Technology in India, which could be categorized into three phases:

- *1976-1990: A-Technology Phase*

Focus was on Adaptation of 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 during this phase.

- *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 realize.

Shri Rajeeva Ratna Shah released the 3rd issue of VishwaBharat@TDIL, which details over 50 technologies for possible transfer/ collaborative development.



Release of Newsletter VishwaBharat@tdil

During the inaugural address, Sh. Rajeeva Ratna Shah, Secretary made the following remarks:

- The formation of the Indian Language Technology Industry Consortium is a welcome sign. It will help in commercialization of the technologies developed under TDIL programme. The consortium could also provide a platform to facilitate a dialogue between technology providers and technology takers in the industry.
- Indian language technology development should emerge soon to become at par with similar technology available for English. All literary content in different Indian languages may be made available in electronic form. The works of Thiruvallavar and Tulsidas should be accessible to all without any language barrier with the help of Machine Aided Translation (MAT) Tools. Indian multi-lingual culture is much superior unlike American multi-lingual culture which has merged into Americanism. We retain our basic culture, which has two portions - literature and communication. IT will facilitate the communication part and in this context text-to-speech and speech-to-text are important.
- MAT (Machine Aided Translation) especially with multi-media and the scope for translation of Indian literature for making these available in



other Indian languages. This will accelerate the process of communication. Britanica has a site which contains originals of the “Great Books of the World” on 50 subjects and they claim to have covered the entire knowledge of the world with cross references. But we note that it does not contain works of any of the Indian thinkers or other thinkers like Confucius, and hence this has missed the precious knowledge and literature of Oriental Nations. Our vision for TDIL programme should include creation of a series of “Great Books of India” in Hindi with conversion facility from one Indian language to another as well as English. All this great work should be available on the internet.



Opening of Exhibition-संयोजन : Combining Academia, Industry & Government efforts.

- Sh. R.R. Shah, Secretary MIT opened the exhibition by joining the Red, Green and Blue ribbons representing Government, Industry and Academia respectively, and Symbolizing a joint endeavor. Theme of the exhibition was **Digital Unite and Knowledge for all**. Secretary interacted closely with the technology exhibitors. He appreciated the efforts of 21 organizations who demonstrated over 50 different language technologies.



II. Technology Presentations

Date: Nov. 7, 2001

Session II Machine Aided Translation (MAT)

It was chaired by Dr. K.P.A Menon, Chancellor, LBSS Vidyapeeth (formerly Secretary of Ministry Defence), who is considered to be the fastest translator of Sanskrit text in the world. He moderated the presentations and discussions for about three technologies presented during the session on Machine Aided Translation systems. These presentations include that of ER&DCI- Noida, IIT-Kanpur, NCST- Mumbai and C-DAC- Pune.

Date: Nov. 7, 2001

Session III : Optical Character Recognition (OCR)

It was chaired by Prof. K.K. Aggarwal, vice-chancellor, IP University. In this session five institutions discussed their OCR technologies for Hindi, Bangla, Punjabi, Oriya and Telugu. The Industry raised the question about the TOT fees and time for completion to reach a prototype.

Date: Nov. 7, 2001

Session IV: Text to Speech

It was chaired by Shri R. Ravindra Gupta, Secretary, Ministry of Heavy Industry. In this session, five institutions presented their technologies specially for Hindi, Bangla, Oriya. Voice recognition s/w were also demonstrated.

Date: Nov. 8, 2001

Session V: Linux & Tools

It was chaired by Shri M. Shankar, Secretary, Dept. of Official Language. In this session Linux for Indian

Languages was presented by NCST & IIT Kanpur. Four institutions presented various language technology tools for Bangla, Tamil, Telgu and Devanagari.

Date: Nov. 8, 2001

Session VI: TOT Dialogue & Resolution for Collaborative Development : Technology Handshakes

It was chaired by Prof. P.V.S. Rao of Tata Infotech who is a pioneer in the



area of Computer Systems & speech technology in the country. He reiterated that Government, academia and industry have joined together to make concerted efforts towards taking the language technology to the masses at the initiative of



Technology Handshake

Ministry of Communications & Information Technology. The success of this meet will address much more complicated technology development process to bring Indian languages at par with English in terms of the availability and affordability of IT to the common-man in India. Intensive technology presentations, dialogue and technology transfer negotiations over these two days have facilitated tie-ups of mutual interest for technology transfer, identification of gap technologies and collaborative development in the areas of Machine Aided Translation (MAT), OS, HUMIS, Tools and e-Content. The impact of these technologies will be far reaching:

- The translation tools will help the translators in reducing their translation effort in terms of man-hour effort by orders of magnitude. The machine aided translation for alien language pair i.e. from English to Hindi for simple sentences today has become a reality.
- The Hindi enabled Linux will be in public domain for proliferation of it's usage.
- (i) Optical Character Recognition(OCR) of Devanagari is going to revolutionize the content creation effort in Hindi by reducing it from 2000 hrs efforts to 2 hrs. effort. This will help in making available archival data and heritage data easily on the Internet.
(ii) The Devanagari Text-to-Speech systems will help many of visually impaired children in education.

- For various operating system platforms, language tools ranging from office suite, Authoring Systems, Search Engines, Multilingual communication Tools, and Natural Language Processing Tools will become available.

- Multilingual knowledge

resources are also available in public domain facilitating it's use by educationists, researchers, publishers, etc.

About 40 Technology Handshake agreements were signed for Transfer of Technology(TOT) or collaborative technology development. MIT will take necessary action to expedite the negotiations with regard to these technology handshakes for TOT aiming at early commercialization. On Technology transfer, certain issues by the Industry were raised which were clarified during the discussion. These issues mainly included the following points:

- TOT fee should be reasonable in view of the promotional role of MIT and MIT should mediate in finalising the details of Technology Transfer with time targets.
- The methodology of supporting the technology development for future upgrades with the concerned R&D organizations should be worked out.
- The negotiations with regard to the Technology Handshakes will be finalized with the help of Indian Language Industry Consortia.

MAIT Language Technology Consortium

Background

The MAIT – IIIT Bangalore Study (December 1999) on potential for local language applications and software revealed that there existed a significant demand for such applications (to the tune of Rs.500-



600 crores). However, the market has been severely constrained due to lack of standards for characters, fonts and keyboard layout.

With increased focus on e-Governance by the Centre and the various State Governments it has become critical to develop standards for Indian languages as without standards it would not be possible to share the various databases. For e-Governance to succeed, it is necessary to promote and develop the market for local language applications, solutions and products.

With this background an increasing need was felt by the IT Industry in India to form a Consortium that would focus attention on development of Local Language Technology standards and other issues of market development.

The MAIT Consortium on Local Language Technology since its inception in September 2001, has been actively co-ordinating various activities with the Industry and the TDIL Programme of the Department of the Information Technology. Some of the highlights of the activities of Consortium are stated below:

(i) Recommendations for Modification of UNICODE

A 2-day meeting of Consortium was held under the aegis of Ministry of Communications & Information Technology on 10th and 11th of September, 2001. UNICODE is a 16 bit standard and is by default emerging as the global standard for local language computing. However, it needs to be modified in context of Indian language computing. The Consortium recommended modifications of font layouts of the following languages:

Devanagari (Hindi, Konkani, Marathi, Nepali, Sanskrit, Sindhi), Gurmukhi (Punjabi), Gujarati, Bengali (Bengali, Assamese, Manipuri), Oriya, Malayalam, Kannada, Telugu, Tamil, Arabic (Urdu, Sindhi, Kashmere).

The Ministry of Communications & Information Technology is a voting member of the UNICODE Consortium and will take up these recommendations with the UNICODE Consortium.

(ii) Font Standards for Indian Languages

The Consortium has been able to successfully develop font standards for most of the Indian languages, an issue that has been pending for the past 10 years. Bi-lingual and mono-lingual font layouts have been developed for Devnagari, Gujarati, Malayalam and Punjabi. The font layouts for these languages have been developed by taking into account the support required for Linux. Font layouts for Bengali, Assamese and Oriya are under development while layout standards for Tamil and Kannada have already been developed by the State Governments of Tamil Nadu and Karnataka respectively.

The Consortium alongwith the Ministry of Communications & Information Technology will take up this font standards with respective State Governments for evolving a national consensus and early adoption of the same.

(iii) Transfer of Technology

At the Language Technology Business Meet 2001 organized by the Ministry of Information Technology on 7-8 November, 2001, 40 MoU's were signed by various companies for transfer of technology. The MAIT Local Language Technology Consortium was given the responsibility to facilitate and interface between the organizations for delivery of the technologies. Follow-up with the companies revealed that most of these technologies are in an advanced stage of development and would be ready for transfer soon.

Contact:

Mr. Vinnie Mehta
PHD House, 4th Floor
Opp. Asian Games Village
New Delhi – 110 016.
Ph: 011-6855487/6854284/6866976
Fax: 011-6851321
E-mail: mait@vsnl.com
Website: <http://www.mait.com>



Transfer of Technology Handshakes

1. Technology Developed

Optical Character Recognition System for Devanagari

► Technology Developer

Indian Institute of Technology
Dept. of Industrial and Management Engineering
Kanpur -208016
Tel: 0512-597743 Fax: 0512-597553
Dr. Veena Bansal
veena@iitk.ac.in

► Technology Description

The system works with the help of a Scanner. User puts a piece of paper document printed in Devanagari (Hindi) script under the scanner, runs the OCR software and gets all the text from that document available inside the computer just as if it was typed in. The data is stored in ISCII code. The system is developed using C programming language. The technology can be used with LINUX platform. It can be easily ported to Windows platform. The OCR software can be integrated with a Hindi Speech Synthesis System to make a Text to Speech system in Hindi. It can be used as front end for a Machine Aided Translation System. Potential beneficiaries are Newspaper (printed in Devanagari script) Houses, Libraries, Offices looking for office automation, Linguistic Community (for creating Corpus), Blind People, etc.

► Technology Recipients

- Shri Ashish Pandey,
Maitrey Infonet Private Limited
34/1, Lower Ground Floor, Yusuf Sarai
New Delhi 110016
Tel: 6535915, 6534896
ashish@maitreyinfonet.com
Technology : Optical Character Recognition System for Devanagari
- Shri Manoj R. Annadurai
Chennai Kabigal,
2, Reddy Colony, Ramalingapuram
Chennai 600012
Tel: 6449139/6601515
manoj@chennaikavigal.com
Technology : Optical Character Recognition System for Devanagari

2. Technology Developed

OCR system for Devanagari

► Technology Developer

Indian Statistical Institute
Computer Vision and Pattern Recognition Unit
203, Barrackpore Trunk Road, Kolkata 700035
Tel: 033-5778085, 5777694
Fax: 5776680, 5773035
Prof. B.B. Chaudhary, Head
bbc@isical.ac.in

&

Technology Promoter

Centre for Development of Advanced Computing
Pune University Campus, Ganesh Khind Road
Pune-411007
Tel: 020-5694060(D)
Fax: 5694059
Shri Aditya Gokhale

► Technology Description

The system works with the help of a Scanner. User puts a piece of paper document printed in Devanagari (Hindi) script under the scanner, runs the OCR software and gets all the text from that document available inside the computer just as if it was typed in. The data is stored in ISCII code. The system is developed using C programming language. The technology can be used with LINUX platform. It can be easily ported to Windows platform. The OCR software can be integrated with a Hindi Speech Synthesis System to make a Text to Speech system in Hindi. It can be used as front end for a Machine Aided Translation System. Potential beneficiaries are like Newspaper (printed in Devanagari script) Houses, Libraries, Offices looking for office automation, Linguistic Community (for creating Corpus), Blind People, etc.

► Technology Recipients

- Shri N. Srikumar
Pyramid Cyberway Pvt Ltd.
909, Ansal Bhawan
16, Kasturba Gandhi Marg
New Delhi 110001
Tel: 3314156, 3351804, 3355552
Fax: 3712769
harminder@pyramidcyberway.com
Technology : OCR system for Devanagari



3. Technology Developed

MAT (English to Hindi) : Machine Aided Translation System

► Technology Developer

Indian Institute of Technology

Department of Computer Science & Engineering
Kanpur-208 016 Tel: 0512-598254(R),
597174(O)

Fax:0512-597553

Dr. Ajay Jain

ajain@iitk.ac.in

► Technology Description

It is a Machine aided Translation System based on AnglaBharti approach of IIT, Kanpur. The AnglaBharti is a rule based system which currently can handle simple sentence translation from English to Hindi. The lexicon of 20000 words relating to Health and IT domain has been integrated into this system. Besides this, it can also handle general purpose translation. User friendly Pre-processor and Post-editing modules are being integrated. The system has been developed using ISCII Code. It is being expanded to handle complex Sentences. Efficiency of the system can be drastically improved by integrating customised domain specific dictionaries. It can be integrated into many commercial products. Installation Guide and User's Guide is ready. Patent is being applied by IIT, Kanpur. Feed back from Beta Site testing at Central Translation Bureau (CTB), New Delhi is being continuously used to improve the system. Potential beneficiaries will be like Translators, Linguists, and Govt. Offices, CTB and other Translation Units.

► Technology Recipients

- Shri N. Srikumar
Pyramid Cyberway Pvt Ltd.
909, Ansal Bhawan
16, Kasturba Gandhi Marg, New Delhi 110001
Tel: 3314156, 3351804, 3355552
Fax: 3712769
harminder@pyramidcyberway.com
Technology : MAT (English to Hindi)
- Shri Kaushal Pandey Crystal Hues Ltd
546-A, Chiragh Delhi,
New Delhi 110017
Tel: 6226255, 6401506, 6425710
Fax: 6221573
chues@nda.vsnl.net.in
Technology : MAT (English to Hindi)

4. Technology Developed

MAT (English to Hindi) : Machine Aided Translation System

► Technology Developer

Centre for Development of Advanced Computing

Pune University Campus, Ganesh Khind Road
Pune-411007

Tel: 020-5694060(D)

Fax: 5694059

Shri Mahendra Kumar Pandey

mahendra@cdac.ernet.in

► Technology Description

MANTRA, Machine assisted Translation tool: It translates the English text into Hindi in a specified domain of Personal Administration, specifically Gazette Notifications, Office Orders, Office Memorandums and Circulars. The strategy adopted in ManTra is Not Word To Word..... Not Rule To Rule But Lexical Tree To Lexical Tree. Mantra uses Lexicalized Tree Adjoining Grammar (LTAG) formalism to represent the English as well as Hindi Grammar. The storage code is ISCII. The MANTRA Technology is being expanded for translating the English texts into other Indian languages such as Gujarati, Bengali, and Telugu. Potential beneficiaries are Translators, Linguists and Govt. Offices, Central Translation Bureau and other Translation Units.

► Technology Recipients

- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167 Fax 080-3357167
applesoft@vsnl.com
Technology : MAT (English to Hindi)
- Shri N. Srikumar
Pyramid Cyberway Pvt Ltd.
909, Ansal Bhawan
16, Kasturba Gandhi Marg
New Delhi 110001
Tel: 3314156, 3351804, 3355552
Fax: 3712769
harminder@pyramidcyberway.com
Technology : MAT (English to Hindi)



5. Technology Developed

MaTra: Human Aided Machine Translation Tool from English-Hindi

➤ Technology Developer

National Centre for Software Technology

Gulmohar Cross Road No. 9, Juhu,

Bombay – 400 049

Tel. 022-7579935/7812120(D)

Fax: 022-6232195/6210139

Shri Durgesh Rao

durgesh@ncst.ernet.in

➤ Technology Description

A software system for translating English to Hindi. There are two versions of the MaTra based on the amount of interaction they expect from the user. MaTra Lite–Fully Automatic On Line Translator, it is simple web based interface and MaTra Pro- Professional Translators Tool with Auto, Semi-Auto and Manual Modes, GUI and Customizable lexicon. At present it supports NCSTs format and unicode. It can be made to support ISCII also. The system GUI is designed in Java for portability. In addition to the internal system lexicon, there is a user-defined lexicon, which can be enhanced and modified by user. Potential beneficiaries are like Media News Agencies, Translation Bureaus and Educational Institutions involved in long distance and Online Education.

➤ Technology Recipients

● Shri N. Anbarasan

Chief Executive Officer

Applesoft (Software Developer for Indian Languages)

No. 39, 1st Cross, 1st Main Shivanagar, W C Road,

Banglore – 560 010

Phone 080-3386167 Fax 080-3357167

applesoft@vsnl.com

Technology : MAT (English-Hindi) MaTra:

Human Aided Machine Translation Tool from English-Hindi

6. Technology Developed

Anusaaraka – Text to Text converter from one Indian Language to other & Telugu Spell Checker

➤ Technology Developer

University of Hyderabad

Department of Computer and Information Sciences

University of Hyderabad, P.O. Central University Hyderabad-500 046

Tel: 040-3010500(O),3010846(R)

Fax: 040-3010145/3010120

Dr.K.N.Murthy

knmcs@uohyd.ernet.in

➤ Technology Description

Anusaaraka is a computer software which renders text from one Indian language into another. It produces output which is comprehensible to the reader, although at times it might not be grammatical. For example, a Telugu to Hindi anusaaraka can take a Telugu text and produce output in Hindi which can be understood by a Hindi reader, but which is not fully grammatical. Therefore, the reader will require some amount of training for reading the output. Anusaarakas have been built from Telugu, Kannada, Bengali, Marathi, and Punjabi to Hindi. Beta versions of all of these have been released for use over the internet as e-mail servers. The storage code is ISCII. It is open source code under GPL therefore users can easily adopt it for their use. Anusaaraka can be used in various scenarios. For example, A reader might be accessing a web site containing Indian language texts. He comes across a site of interest, and wants to read material on it. However, he does not know the language. He can run anusaaraka and read the text. Normally, the reader motivation is high and he is willing to put in some effort.

➤ Technology Recipients

● Shri Manoj R. Annadurai

Chennai Kabigal,

2, Reddy Colony, Ramalingapuram

Chennai 600012

Tel: 6449139/6601515

manoj@chennaikavigal.com

Technology : Anusaaraka, Spell Checker Components (Telugu & Kanada)

● Dr. M.N.Cooper

The Managing Director

Modular Infotech Pvt. Ltd.

26A, Electronic Cooperative Estate,

Pune - Satara Road Pune 411 009

Tel. 020- 4223342/4226614/4225896

modular@giaspn01.vsnl.net.in

Technology : Telugu Spell Checker



7. Technology Developed

Text to speech processing system for Oriya
Oriya Spell Checker

► Technology Developer

Utkal University, Bhubaneswar
Department of Computer Science &
Applications
Vani Vihar, Utkal University
Bhubaneswar – 751 004
Tel. 0674-580216(O), 540865(R)
Fax: 0674-581850
Prof. (Mrs) Sanghmitra Mohanty
sangham1@sancharnet.in

► Technology Description

It is Oriya Speech Recognition System, it converts Oriya Text into Speech. In it Pulse Code Modulation technique is applied to speech signals then least mean square technique is applied to the signals to find the desired parameter and thus the speech database is prepared. For normalization average of a set of samples is taken to standardize the database. For data base creation pure phonetic nature of Oriya language is taken into account and thus words and sentences are uttered. The storage code is ISCII. Potential beneficiaries are like Blind people, Illiterate people.

► Technology Recipients

- Shri Ambika Prasad Das
Avon Technologies,
6-3-563/26/A, Somavarapu Heights,
Hilltop Colony, Erramanzil
Hyderabad 500082 A.P.
Tel: 040-6580836/37
Fax: 040-3307025
ambika@avontechnologies.com
Technology : Text to speech processing system
for Oriya
- Dr. M.N.Cooper
The Managing Director
Modular Infotech Pvt. Ltd.
26A, Electronic Cooperative Estate,
Pune - Satara Road
Pune 411 009
Tel. 020- 4223342/4226614/4225896
modular@giaspn01.vsnl.net.in
Technology : Oriya Spell Checker & TTS

8. Technology Developed

Speech Processing Technology

► Technology Developer

Tata Infotech Ltd.,
Technology Cognitive Systems Research Lab,
Sanpada,
Navi Mumbai - 400705.
Tel. 022-7903251/56, 022-7682379 Ext-127
Fax: 022-7619926
Dr.P.V.S.Rao
pvsrao@tatainfotech.com

► Technology Description

Isolated word recognition system specially tuned for Indian speakers. Easily scalable for larger number of words. Useful for fixed vocabulary speech recognition applications such as speech enabled Interactive Voice Response (IVR) systems, Telebanking, tourism information kiosk, airline reservation, voice portals, voice commands' recognition for automobile control etc.

► Technology Recipients

- Prof. (Ms) Sanghmitra Mohanty
Department of computer science & application
Vani Vihar,Utkal University
Bhubaneswar – 751 004
Tel. 0674-580216 (O), 540865(R)
Fax: 674-581850
Sangham1@sancharnet.in
Technology : Speech Processing Technology



9. Technology Developed

Hindi Speech Recognition System

► Technology Developer

IBM India Research Lab
Block – I, IIT Campus,
Hauz Khas, New Delhi-16
Tel. 6861100 Fax : 6861555
Shri Ashish Verma &
Dr. P.V. Kamesan, Senior Manager
pkamesam@in.ibm.com

► Technology Description

This is a Hindi Speech recognition system for a large vocabulary speaker-independent dictation task. For any given language, the computer first needs to learn the sound of spellings in various context which is a training phase. After this system is ready for speech recognition. Complex signal processing and statistical techniques are used to make the recognition robust to speaker speech variations and to make it work on continuous speech of a large vocabulary. Compliant with IBM ViaVoice Standards will soon be using Unicode standard. The system can be customized to different tasks for further improving the accuracy. Expandability (in terms of increasing new words to the system) is available. Potential beneficiaries are Hindi speaking population.

► Technology Recipients

- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167
Fax 080-3357167
applesoft@vsnl.com
Technology : Hindi Speech Recognition System
- Shri Achinto Rakshit
Cyberspace Multimedia,
101, Mahalakshmi Mansion
941, 21st Main Road, 22nd 'A' Cross
Opp: BDA Complex Banashankari 2nd stage
Bangalore 560070
Tel: 080-6710925, 6714446
info@cyberscapeindia.com
Technology : Hindi Speech Recognition System

10. Technology Developed

Hindi Speech Recognition System & Adaptation for Bengali

► Technology Developer

Megasoft India
921, Sector-14, Sonipat
Haryana-131001
Tel : 01264-47807 Fax : 01264-42583
Shri Anil Aggarwal
megasoftindia@rediffmail.com

► Technology Description

Hindi dictation and PC control is dynamic speech recognition application for two languages, Hindi (first time) and Indianized English by Megasoft India. It is capable of taking continues dictation and controlling PC functions in both languages by simple phrases and commands which can be extended to any limit like telephony-PC communication, voice query systems and lots more using VB, VC++ applications. Hindi dictation and PC control uses L&H Dragon Naturally Speaking professional platform. It supports Unicode. Many speech applications can be made with business potential using Hindi/English speech recognizer. Present research work regarding vocabulary and Hindi language (context and vocabulary) can be implemented on other recognizer platform very quickly with very effective results.

► Technology Recipients

- Shri Vivek Siegell
HCL Info Systems Ltd.
E 4, 5 &6, Sector 11
Noida 201301 UP
Tel: 4550862 (D), 4520977 Extn. 3208
Fax: 4533877, 4544527
seigell@hclinsat.com
Technology : Hindi Speech Recognition System
- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167
Fax 080-3357167
applesoft@vsnl.com
Technology : Hindi Speech Recognition System
- Biswajit Saha
ER&DCI,
Plot E-2/1, Block GP, Sector-V
Bidhannagar, Kolkata 700091
Tel: 3579846, 3575989, 3573581
Fax: 033-3575141
erdcical@giascl01.vsnl.net.in
erdccal@giascl01.vsnl.net.in
Technology : Adaptation for Bengali



11. Technology Developed

Morphological Analyser and Generators
TTS for Tamil
Spell Checker for Tamil

► Technology Developer

Anna University
School of Computer Science & Engineering
Chennai - 600 025
Tel. 044-2351723 Fax: 044-2350397
Dr. T.V.Geetha/Ms. Ranjani Parthasarthi
rp@annauniv.edu

► Technology Description

This software is a complete morphological Generator & Analyzer for Tamil nouns, verbs, adjectives and Proposition. Morphological Generators is one of the two important tools for language processing. Morphological analyzer is another basic tool required for language processing. Given any word or a group of words, it analyzes the word and determines the root and all the other add ones that it has taken. The code has been developed using Java. It uses a specially designed internal code instead of storing string. The data storage is Unicode compatible. The system is developed in Java hence fully portable. The tool can be used in any of the available platforms. The tool is developed using object oriented concept so it can be expanded as per requirement. This tool is a collection of modules such as verb generator and noun generator. Modules can be added to this tool and also it can be added as a module to any other language tool which needs it.

► Technology Recipients

- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167
Fax 080-3357167
applesoft@vsnl.com
Technology : Morphological Analyser and Generators
- Shri Manoj R. Annadurai
Chennai Kabigal,
2, Reddy Colony, Ramalingapuram
Chennai 600012
Tel: 6449139/6601515
manoj@chennaikavigal.com
Technology : Spell Checker for Tamil
- Dr. M.N.Cooper
The Managing Director
Modular Infotech Pvt. Ltd.
26A, Electronic Cooperative Estate,
Pune - Satara Road
Pune 411 009
Tel. 020-4223342/4226614/4225896
modular@giaspn01.vsnl.net.in
Technology : Spellchecker & TTS for Tamil

12. Technology Developed

Text to Speech Synthesis System for Hindi,
Speech Recognition & Voice Synthesizer

► Technology Developer

Central Electronics Engineering Research Institute
CSIR Complex,
NPL Campus, Hillside Road,
New Delhi-110 001
Tel. 011-5781467/5783172 Fax: 5788347
Dr. S. S. Aggarwal, Scientist G and Head
sagrawal@ceerid.ernet.in

► Technology Description

The HindiVani is the Windows based software for converting Hindi Text files into Speech. The text document is generated using Hindi editor, which supports ISCII standard. The input words are split into syllables, using a parser. An acoustic-phonetic database of all these syllables is available in the database, which is subsequently used to create words. The concatenation of syllables into words and the superimposition of quality features is done by developing rules. A cascade-parallel format synthesizer developed at CEERI is used to synthesize the speech. The storage code is ISCII. The system can be used with any Pentium machine with Windows Operating System and multimedia facility. This system can be expanded to other spoken languages of India. It can be integrated with OCR system. It is very useful product for Hindi speaking visually handicapped people, Information retrieval in spoken form, Text Reading Machines.

► Technology Recipients

- Prof. (Ms) Sanghmitra Mohanty
Department of computer science & application
Vani Vihar, Utkal University
Bhubaneshwar – 751 004
Tel. 0674-580216 (O), 540865(R)
Fax: 674-581850
Sangham1@sancharnet.in
Technology : TTS for Hindi
- Shri Kanta Verma
Aarkay Computer Research Foundation,
8/53, Birbal Road, Jangpura Extension,
New Delhi – 110014
Phone 6440940/1/2/3 Fax: 4326051
Aarkay@del2.vsnl.net.in
Technology : TTS for Hindi
- Dr. Mukul Kumar Sinha
President,
Expert Software Consultants Pvt Ltd
Almora Bhawan
C-17 South Extension Part I



New Delhi 110 066
Tel. 011- 4642675/6963146
expert@vsnl.com
Technology : TTS for Hindi

- Shri Achinto Rakshit
Cyberspace Multimedia,
101, Mahalakshmi Mansion
941, 21st Main Road,
22nd 'A' Cross
Opp: BDA Complex Banashankari 2nd stage
Bangalore 560070
Tel: 080-6710925, 6714446
info@cyberscapeindia.com
Technology : TTS for Hindi
- Shri Keyur Shroff
National Centre for Software Technology,
Gulmohar Cross Road No. 9, Juhu,
Bombay – 400 049
Tel. 022-6201606/6201574/6249817
022-7579935/7812120(D)
Fax: 022-6232195/6210139
keyur@ncst.ernet.in
Technology : TTS for Hindi
- Dr. M.N.Cooper
The Managing Director
Modular Infotech Pvt. Ltd.
26A, Electronic Cooperative Estate,
Pune - Satara Road
Pune 411 009
Tel. 020- 4223342/4226614/4225896
modular@giaspn01.vsnl.net.in
Technology : TTS for Hindi
- Shri Ambika Prasad Das
Avon Technologies,
6-3-563/26/A, Somavarapu Heights,
Hilltop Colony, Erramanzil
Hyderabad 500082 A.P.
Tel: 040-6580836/37
Fax: 040-3307025
ambika@avontechnologies.com
Technology : Hindi TTS & Speech Recognition
- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167
Fax 080-3357167
applesoft@vsnl.com
Technology : Vice Synthesizer

13. Technology Developed

Linux with Indian Language Support
ISPELL

► Technology Developer

Indian Institute of Technology
Department of Computer Science &
Engineering
Kanpur-208 016
Tel: 0512-597652
Fax: 590725
Prof. Rajat Moona
moona@iiit.ac.in

► Technology Description

Using this software it is possible to execute any text based utilities in Unix environment and use the Indian language support. It is possible to name the files and see the directory listing in Hindi. The Hindi files can be edited using terminal based standard Unix utilities such as vi, sed, etc. The configuration files supplied with the ITERM are written to support ISCII files, Inscript Key board layout and Devanagari TeX fonts. Characters can be coded using ISCII or any other standard. It is possible to support the ISFOC fonts or any other fonts. It supports inscript/ phonetic keyboard layout. Few configuration files can be written so that it can be used with wide variety of scripts other than Devanagari. Potential beneficiaries are Users of Unix platform.

► Technology Recipients

- Shri V.K. Gupta
VXL INSTS LTD.
IInd floor, No. 1 Mohammadpur,
Behind Bhikaji Cama Place
New Delhi 110066
Tel: 6190061/62/63 Fax: 6190064
medleyindia@vsnl.com
Technology : Linux with Indian Language Support
- Shri Manoj R. Annadurai
Chennai Kabigal,
2, Reddy Colony, Ramalingapuram
Chennai 600012
Tel: 6449139/6601515
manoj@chennaikavigal.com
Technology : ISPELL



14. Technology Developed

INDIX (Localized Linux)

► Technology Developer

National Centre for Software Technology

Gulmohar Cross Road No. 9, Juhu,

Bombay – 400 049

Tel. 022-7579935/7812120(D)

Fax: 022-6232195/6210139

Shri Keyur Shroff

keyur@ncst.ernet.in

► Technology Description

Graphical User Interface in Indian Languages for Linux Operating System. Indic Script Shaping Engine and open Type Font support has been built into the System. System is supporting Unicode encoding and backward compatibility is provided through UTF-8 encoding. The system has Unicode support at the core level, so it is highly portable. The system can be localized for other Indian Languages apart from Devanagari. Installation guidelines and User's guide to use and configure the system is ready. It is being developed under public domain (GPL model). Alpha testing is in Progress. Potential beneficiaries include end users interested in Linux with localized Applications

► Technology Recipients

- Prof. K. Narayan Murthy
Dept. of Computer Science & Information Science
University of Hyderabad
P.O. Central University
Hyderabad 500 046 (AP)
Tel. 040-3010500/518 Extn. 4017 (O),
3010064(D), 3010374(R)
Fax: 040-3010120, 3010145
knmcs@uohyd.ernet.in
Technology : INDIX (Localized Linux)

- Shri N. Anbarasan
Chief Executive Officer
Applesoft (Software developer for Indian Languages)
No. 39, 1st Cross, 1st Main Shivanagar, W C Road,
Banglore – 560 010
Phone 080-3386167 Fax 080-3357167
applesoft@vsnl.com
Technology : INDIX (Localized Linux)

- Shri V.K. Gupta
VXL INSTS LTD.
IInd floor, No. 1 Mohammadpur,
Behind Bhikaji Cama Place
New Delhi 110066
Technology : INDIX (Localized Linux)

15. Technology Developed

Punjabi Spell Checker

Official Dictionary

► Technology Developer

Thapar Institute of Engineering & Technology

Department of Computer Science &

Engineering

(Deemed University), Patiala 147 001

Tel: 0175-214868, 393137

Fax: 0175-214498, 216391, 212012, 212002

Dr. G.S. Lehal

Tel: 393374(D), 283502(R)

gslehal@mailcity.com

► Technology Description

Punjabi spell checker can be integrated with a word processor. It also supports both phonetic and typewriter keyboard layout for Punjabi. Spell checker can operate on all documents typed in both Punjabi and English. The system encodes the output in ASCII encoding format. The system encodes the output in ASCII encoding format. The technology can be used with any windows platform. Potential beneficiaries are like Government sector, Typist, Educationist.

► Technology Recipients

- Dr. M.N. Cooper
The Managing Director
Modular Infotech Pvt. Ltd.
26A, Electronic Cooperative Estate,
Pune - Satara Road
Pune 411 009
Tel. 020- 4223342/4226614/4225896
modular@giaspn01.vsnl.net.in
Technology : Punjabi Spell Checker & Official Dictionary
- Shri Manoj R. Annadurai
Chennai Kabigal,
2, Reddy Colony, Ramalingapuram
Chennai 600012
Tel: 6449139/6601515
manoj@chennaikavigal.com
Technology : Punjabi Spell Checker



16. Technology Developed

Hindi Encyclopedia (Vishwakosh)

► Technology Developer

Electronics Research & Development Centre
C-56/1, Anusandhan Bhawan, Sector-62,
Noida – 201301

Tel. 91-4587717-25 Fax: 91-4587726

Shri V.N.Shukla (ERDCI/N)

&

Kendriya Hindi Sansthan (KHS)

Agra-282005

Tel. 91-562-530683/684 Fax:91-562-530159

Prof. Thakur Das (KHS)

&

Nagari Pracharini Sabha

Varanasi

Tel. 91-542-331277 Fax:91-542-331488

Shri Sudhakar Pandey (NPS)

► Technology Description

Hindi Encyclopedia is the only encyclopedia published way back in 60's by Nagari Pracharini Sabha, Varanasi. It consists of 12 volumes of data covering almost all details of 1500 topics from various fields of life. The task of digitizing the information & putting on CD and internet was assigned to ERDCI/N & KHS Agra, by MIT & MHRD as a joint project. The information has been made available in such a way that one can find the information in alphabetical way or by categories. The storage code is ISCII. Expandable to other domains/ Lexicons. The system can be integrated with various windows based word processors. Useful to Translators, Linguists and Office Assistants, Educational Institutes and organization working in Hindi.

► Technology Recipients

● Shri N. Srikumar

Pyramid Cyberway Pvt Ltd.

909, Ansal Bhawan

16, Kasturba Gandhi Marg

New Delhi 110001

Tel: 3314156, 3351804, 3355552

Fax: 3712769

harminder@pyramidcyberway.com

Technology : Hindi Encyclopedia

2.3 Intellectual Property Rights (IPR)

Email : rct@mit.gov.in Tele:4363648, 4363123

न हि ज्ञानेन सदृशं पवित्रमिह विद्यते ।

तत्स्वयं योगसंसिद्धः कालेनात्मनि विन्दति ॥ गीता ४.३८ ॥

“On earth, there is no purifier as great as knowledge, he who has attained purity of heart through a prolonged practice of Karmayoga automatically sees the light of truth in the self in course of time (Geeta 4.38)”

Ministry of MC&IT intends to increase awareness about the IPR matters amongst all class of concerned professionals so as to make India participate on equal footing with Member States of the World Trade Organisation (WTO) since it has joined the Trade Related Intellectual Properties (TRIPs) Agreement and also the Patent Cooperation Treaty (PCT) Geneva. The IPR Cell in Ministry of Communications and Information Technology is pursuing the following objectives:

- Promotion of IPR activities in the field of electronics and Information Technology
- Create awareness and provide promotional & facilitator support
- Provide value-added patent information support—search and analysis—to assist technology Assessment, Development, Acquisition and Investment Decisions.
- Respond to IPR needs of digital era.

It is a Single Window Facility in MC&IT which assists S/W Developers, Scientists, Researchers, R&D Managers, Chief Executives of IT and Electronics Units, Policy Makers and Funding Organisations for Protection of all Intellectual Property Rights in India and Abroad for Electronics and Information Technology for obtaining registration of :

►S/W Copyrights ►Patents ►Trademarks ►Designs

It is evolving plans, developing Tools and infrastructure for international/national IPR protections, facilitation, providing services of IPR Awareness, Technology Assessment and Alert based on Patent Search Services in Electronics and IT (TAPs), and is responding to the needs of the digital era.

IPR Cell, MCIT has already filed 43 Patents, 32 copyrights, 9 Trade Marks and one Design to assist its PSUs, Scientific Societies and Grantee Institutions secure IPR Protection for meaningful transfer of technology in the current products & services. It has ventured to evolve suitable forms and procedures for research organisations to disclose their patents, licence the patents and copyrights, enter into a contract with industry & Govt.collaborations and related services required by the inventors/creators of IPR on one hand and industry/service organisations/society on the other hand.



2.4 International Symposium on “Language in Cyberspace”

on 26-27 September-2001 at Seoul, Korea

Knowledge and information sharing through the global information networks by every country and every community in the world is of vital importance for the economic participation, Social cohesion and enrichment of linguistic and cultural diversity of mankind;

Multilingualism, is of strategic importance to ensure the right to freedom of opinion and expression, the right to participate in the cultural life of the community and to have access to information;

The ICT (Information and Communication Technologies) can play an important role in extending and preserving multilingualism for sustainable economic growth, human advancement and cultural diversity.

UNESCO, General Conference at its 31st session (October,2001) got inputs from regional workshops on, “ the promotion and use of multilingualism and universal access to cyberspace” and the implementation of an intersectional programme, the Initiative [B@bel](#). The international symposium on “ Language in Cyberspace” was organized by Korean National Commission for UNESCO on 26-27 Sept-2001, at Seoul. Dr. Om Vikas, MCIT participated and presented the country paper on Language Technology.

The Recommendations of the Symposium :

I. Policy challenges of multilingualism : bridging the language divide

- Access to the enormous legacy of knowledge available in cyberspace is too often limited by the use of languages not known to the user. This causes a disparity in the access to information between those who speak that language and those who do not, a “ language divide”. Although the trend is not a new one, it is accelerated by the use of new technologies leading to the “globalization” of the most important languages of the world to the detriment of a large number of less-used languages. Expertise and resources must be mobilized in the Asia-Pacific region to formulate and implement a coherent strategy to address this issue, such as use of vernacular languages and the development of translation technologies,

not forgetting other means of communications.

- Another phenomenon, not new by itself, but increasing exponentially with the introduction of multimedia technologies of communication is the general decrease in linguistic diversity in the world. Languages are often used as instruments of division between the people. Their survival depends on their capacity to resist the dominance and standardized use of other languages. Any loss of language is an impoverishment of global cultural heritage and research capacity. Strategies should be designed to put ICT to the service of preservation and “ resurrection” of languages in danger of disappearance.
- The absence or insufficiencies of national policies on the use of languages on the global networks need also to be looked into in order to increase the number of vernacular languages and reinforce the development of national standards for internal representations, character glyphcoding and transcription schemes. In this connection, the respect and use of all languages in cyberspace should be reaffirmed in the different national and international formally adopted texts and compatible international norms and principles should be adopted to increase their accessibility on-line;
- Issues of ownership of translated and published works on-line must also be studied and principles meeting general consensus must be derived in order to facilitate access to these works while preserving the authors rights;
- The policies for the advancement of multilingualism in cyberspace must be underpinned by the development of language education and ICT user training strategies and tutorial materials freely accessible on-line and responding to the demands of the information society. Policies, support and incentives are needed to provide all citizens with the opportunity to nurture linguistic and ICT literacy through formal, informal and life-long education;
- The challenge of bridging the language divide in the access to and use of ICT cannot be effective without the full participation of the developing countries of the region and consideration of their particular constraints and needs: Member states should plan action to bridge language divide and to



ensure promotion of knowledge creation capacities.

- In Particular, regional cooperation among nations that share cultural and linguistic similarities must be strongly encouraged by governments, regional and international organizations such as UNESCO.

II. Technological challenges of multilingualism: ensuring language interoperability

- Language cannot be promoted on the global networks without accessibility to the ICT and services they can render. It is therefore important to give due attention to the economic constrains to the Internet connectivity particularly in developing countries; the international community should support universal access to telematics networks and services as a contemporary of the human right.
- It is equally important that people at large have access to the information available on the networks in their mother tongue. Countries of the Asia-Pacific region must be encouraged to support the development of multimedia community centres that would reach out all segments of the society in their local language;
- Ensuring the “intercommunication” of languages on the networks is an issue for which technological solutions are being actively sought. Numerous research and development projects are being carried out to ensure the automatic quality translation of languages and machine use and recognition of different scripts such as the development of smart fonts, cross-lingual conversion utilities, multilingual search engines, multilingual voice searching, speech recognition etc. These efforts should be based on solid theoretical foundations and supplemented by objective translation evaluation and accreditation systems.
- In particular, the lexicon, morphology, syntax and semantic of different languages should be described systematically in a manner suitable for automatic word form recognition, parsing, and generation. Furthermore, these components should be functionally integrated to model mappings from language-specific surfaces to universal content representations (hearer mode) and vice versa (speaker mode). International, regional and national governmental and non-governmental organizations should seek to

harmonize these individual efforts into a consolidated worldwide approach and to ensure the provision of freely accessible translation portals;

- Machine translations require the creation of numerous aid tools such as specific vocabularies and multilingual dictionaries, virtual terminology networks and terminology data banks. On-line specialized glossaries, lexicons multilingual indexing of contents, MT software dedicated to the less common languages, etc. Countries of the Asia-Pacific region should support efforts by public and private sectors, international organizations, as well as the civil society, to pool intellectual and financial resources in core multilingual tools development projects;

III. Content challenges of multilingualism: Digitizing and providing e-contents

- Language diversity should be broadened by increasing the number of languages and scripts on-line and by the creation of multilingual e-contents, websites and means of maintaining, accessing, retrieving and preserving them through the use ICT. All the public domain information (laws, regulations, statistics etc) locally relevant and informative to the citizens for their health, security and participation in their public life should be accessible in the national and/or local language. Countries of the Asia-Pacific region should adopt programmes to develop these freely accessible Web sites;
- Moreover, a great part of the national physical, written and oral heritage which is already freely accessible by all should be produced in digitised form and made available through the application of ICT in the various local languages. Partnership approach involving governments, international organizations, the private sector and NGOs should considered to build-up these massive content;
- In this regard, countries should mobilize resources to assist their major cultural institutions such as libraries, archives and museums in preserving and making their collections accessible with appropriate measures of security in several languages on the global information networks, though multilingual portal conception and digitisation.
- The creation of reliable, re-usable and interoperable content being highly labour intensive and costly, most cost-efficient tools and



methods should be used on the basis of international standards for the preparation of these contents. It should include open system development (i.e. CDS/ISIS software) with respect to data modeling and system design for net-based distributed collective work, their evaluation as well as content validation, user empowerment and promotion of international standards.

- Special attention should be given in the Asia-Pacific region to the issues arising from the protection of the intellectual property rights and the preservation of legal exemptions to the copyright especially when it concerns reuse of information in general and particularly the translated public domain information;

IV. International challenges : Promoting and disseminating policy experiences.

- A reliable international comparative survey on the use of language on the internet and more particularly on the related policies, norms and standards adopted in different countries is much needed. International organization, in particular UNESCO, should work out a scheme for the collection and sharing of such information.
- UNESCO should support mechanisms preventing exploitation of knowledge from indigenous language e-contents.
- The UNESCO statistical Office and the UNESCO Observatory on the Information Society should be mobilized to collect, maintain and diffuse information on the multilingual resources and services produced in the countries and on existing policies related to these resources and to disseminate information on the best practices;
- The principles and measures envisaged in the Recommendation on the promotion and use of multilingualism and universal access to cyberspace prepared within the UNESCO Programme Information for All reflecting the above-mentioned proposals should be given support from the countries of the region; these principles should be presented during the World Summit on the Information Society to take place in 2003.
- The Asian-Pacific countries should participate actively in the implementation of this recommendation and take part in the development of Initiative [B@bel](#) through intellectual and financial input.

● What is B@bel Initiative?

Public domain information is a global public good. With this in mind, UNESCO's main goal consists in redefining universal access to information in all languages in cyberspace by encouraging (1) the development of tools (translation mechanisms; terminology; protocols; etc.) that will facilitate multilingual communication in cyberspace (2) the promotion of fair allocation of public resources to public information providers; and (3) the promotion of access to multilingual public domain information and knowledge.

The programme "Initiative B@bel" proposes to do this by implementing concrete activities at national and international levels, with the objective to develop multilingualism on the information networks and to encourage full partnership between governments, industry and civil society. The programme could be oriented in several directions:

Creation of the infrastructure: establishment of UNESCO Chairs, associating universities with industry, for strengthening research in and development of multilingual search engines, multilingual gateways, virtual libraries and archives, etc.;

Development of multilingual tools: adapting multilingual indexing of websites, thesauri, standards, lexicons and terminology existing in the European Union, UNESCO, ISO, UNU, Union Latine, Infoterm, etc., to other languages including local ones;

Strengthen interoperability: supporting the development of automatic translation tools, including the production of translation free software, the application of translation schools work to the webpages, the on-line development of multilingual encyclopedia, upgrading of routers, etc.;

Formulation of national and international policies and regulations: encouraging the use of many languages on the information networks, the on-line teaching of foreign languages in the education systems, the development of multilingual websites (with a web prize), etc.



2.5 SCALLA 2001 “Sharing Capability in Localisation and Human Language Technologies”

on Nov. 21-23, 2001 at NCST, Bangalore

The SCALLA 2001 working conference was held at NCST, Bangalore on November 21-23, 2001. SCALLA is a joint project of NCST Mumbai, Open University UK and ISI Kolkata, with funding from the European Union under the Asia IT&C program. Its aim is to organize a series of 3 workshops in consecutive years from 2001 to 2003, to bring together experts in the field of localisation and human language technology from Europe and South Asia, with a view to exchange ideas in the field of localisation and human language technology, that have an important role to play in bridging the digital divide.

SCALLA 2001 featured about 20 invited experts from India and the UK with focus on the languages of the South Asian region.

Day I

Session 1 : “Introduction”

Chair : Dr S.P. Mudur

Prof. Pat Hall gave an overview of the conference.

Session 2 : “Localisation Needs”

Chair : Mr S Ramakrishnan

A prototype of the Simputer was demonstrated which is a low-cost portable and sharable device that can take IT to the common man. Honey Bee Networks project disseminates information about grassroots innovations in technology in several Indian states and languages.

Session 3 : “Localisation Practices”

Chair : Prof. B B Chaudhuri

Dr Reinhard Schaler spoke about the current scenario of localisation in Europe. Dr Mudur spoke about the issues and status of localisation of Indian languages. Prof. Pat Hall spoke about Software components and APIs, which are useful technologies for localisation.

Session 4 : “Writing Systems, Input and Output”

Chair : Prof. Pat Hall

Prof. R K Joshi spoke about the uniqueness of the Writing Systems of India, and Prof. B B Chaudhuri presented OCR for Bangla.

Session 5 : “Cultural aspects of localisation”

Chair : Dr Ostler

The main issue raised was the need to be aware of

cultural differences that merit investigation of alternative paradigms of user interfaces, but it was felt that this question has not yet been adequately studied.

Day II

Session 1 : “Language Models”

Chair : Prof. B N Patnaik

Prof. Boyd Michailovsky spoke about developing a lexicon and syntax for the Hoya Language of Nepal, and Prof. Harold Somers spoke about Developing linguistic resources from corpus material.

Session 2 : “Language Generation”

Chair : Prof. Rajeev Sangal

In this session, Dr Donia Scott gave a presentation on Multilingual Natural Language Generation in specific domains.

Session 3 : “Lexicography and Translation”

Chair : Prof. U N Singh

Mr Durgesh Rao spoke about the issues in translating between English and Hindi. This was followed by a presentation by Prof. Rajeev Sangal on building Lexical Resources for Indian Languages, such as a free electronic English-Hindi dictionary.

Session 4 : “Lexicography”

Chair : Prof. Harold Somers

Dr Niladri Sekhar Das spoke on the contribution of language corpora on the development of dictionaries. Prof. U N Singh and Dr B Mallikarjun spoke on making a traditional dictionary into an electronic lexicon.

Day III

Session 1: “Speech and Literacy”

Chair: Dr Reinhard Schaler

Dr Asoke Kumar Dutta spoke on Disbursing spoken language technology in regional dialects, and Dr Gautam Sengupta talked on Voice-enabled Machine Readable dictionaries. Dr Roger Tucker presented a pure speech personal digital assistant.

Session 2: “Concluding Session”

Chair: Dr Om Vikas

Dr Om Vikas highlighted the role of the Indian Government in promoting growth of Indian language technologies, through the Technology Development for Indian Languages (TDIL) initiative of the Department of IT. There is possibility of collaboration between universities in Europe and institutes in India to offer courses in Computational Linguistics. Finally, Prof. Pat Hall summed up the conference.



2.6 Asia-Pacific Regional Consultation on UNESCO's Medium Term Strategy for "The Major Programme on Communication and Information" from 2002-2007,

on 18-19 December 2001 at New Delhi

Dr. Om Vikas, DIT and Dr. N. Vijayaditya, NIC participated in the discussion workshop

Four areas of activities and priority deliverables were identified as follows:

- Content
- Access
- Capacity Building
- Policy

Content

Challenges

- Lack of content development templates/guidelines
- Lack of content development experiences
- Lack of tools to digitize local content.

Actions

- Develop content templates/guidelines for community needs
- Share content development experiences
- Identify and promote technology to provide tools to digitize local content.
- Form partnerships for content development and sharing.
- Partners: Extension services (Education, Agriculture, Health) Govt. Agencies, NGOs, Community organizations and leaders, Media organizations, Professional organizations, Private entrepreneurs.

Flagship Programmes

- Developing Local language processing capability.

Access

Challenges

- Lack of connectivity (telecom, internet) High cost of connectivity and access.

- Lack of appropriate low-cost devices (eg. Computers, FM, radios, energy sources)
- Lack of policy support and political will for providing and ensuring universal access.

Actions

- Promote low cost devices (eg. Handhelds, FM radio sets)
- Promote multi-purpose community centres and access points, especially for marginalized groups.
- Advocate universal access policies.

Flagship Programmes

- Develop a business/operation model for sustainable multipurpose community centres and Test this model by creating five community centres in different countries/cultures/languages/settings.

Capacity Building

Challenges

- Lack of trained human resources in the field of ICT,
- Lack of localized tools, technologies and methodologies in the field of ICT,
- Lack of adequate financial resources for ICT enabled development
- Lack of awareness and management of changes—economic, political and social
- Lack of awareness among professionals and decision-makers about the role of ICT in development.

Actions

- Training of trainers
- Training and retraining human resources in the field of ICT at various levels.
- Promoting ICT for high quality e-skills development
- Organizing dialogues on ethics and social responsibility in ICT.



- Sharing of basic tools, technology and methodologies for ICT enabled development
- Promoting national and international partnerships to respond to various challenges in the field of ICT.
- Establishing UNESCO collaborative virtual institute for ICT enabled development.

Flagship Programmes

- Training the trainers for ICT for development— equitable, sustainable and peaceful
- Integrating ICT into curricula in schools and universities.
- Promoting ICT enabled entrepreneurship
- Collaborable Virtual Institute for ICT enabled Development.

Policies

Challenges

- Lack of specific policies and strategies and institutional infrastructure for implementation.

Actions

- Formulation of holistic and national & regional policies & strategies. (HRD, Access and Application) through participative process & enactment. Preparation of legal and institutional framework for executing them.

Flagship Programmes

- Assisting in formulation of national/sub regional policy and promote sub regional and regional cooperation.
- Development of guidelines for policy and strategy formulation.

2.7 The Asia Pacific Development Information Programme (APDIP) - UNDP

APDIP seeks to promote and establish information technology (IT) for social and economic development throughout Asia-Pacific. The Programme serves 42 countries.

APDIP is funded by the United Nations Development Programme (UNDP) and implemented by the UN Office for Project Services (UNOPS), Asia office.

Information Technology for Developing Countries

IT supports social and economic development by:

- Reducing geographic isolation;
- Providing access to information and knowledge resources and a means for exchanging information;
- Creating opportunities for expanded trade and economic growth; and,
- Enabling greater participation of Civil society and transparency in governance.

APDIP Strategies

- ▶ **Building capacity** at all levels: APDIP sensitises decision-makers and assists in developing an enabling environment for IT and provides technical training.
- ▶ **Identifying IT champions** That provide a vision for what IT can do, and have the influence to make it happen.
- ▶ **Identifying new opportunities for development** is critical as countries implement IT/ Internet-related services.
- ▶ **Co-operating with the private sector** helps spur investment in the development world.
- ▶ **Encouraging South-South partnerships** as one way of building capacity.
- ▶ **Researching appropriate technologies** to minimise access costs and extend the reach of the Internet to rural areas of developing countries.

Capacity Building

Seminars on IT Policy and Infrastructure Development

The APDIP-Cisco Networking Academies

An innovative partnership with Cisco Systems



established to counter the severe shortage of network specialists in the Asia-Pacific.

The Mobile Internet Unit(MIU)

Equipped with Internet-ready computers, this electronic classroom on wheels travels to rural and marginalised urban secondary schools, training teachers, and students alike on how to fully benefit from the global information infrastructure.

Workshops for Technical Personnel

APDIP frequently conducts Webmaster Development Workshops.

Technical Assistance

Wiring the World: Building Connectivity

APDIP assists countries lacking an Internet connection to design and establish low-cost links via satellite so that applications of IT-telemedicine, distance education, or e-commerce- can take off and prosper.

Networking People

APDIP assists regional organisation and countries to improve their co-operation by creating and establishment systems that are designed to meet specific needs.

Hosting Information

One of APDIP's major objectives is assisting other development agencies to implement information services.

Research & Development

Expanding Connectivity to Rural Areas.

APDIP is working with a number of partners to devise ways of making satellite connections to the Internet affordable in rural areas.

Application Development

APDIP promotes customisation of software for developing countries.

Global Issues

The Internet Governance Information Service

APDIP hosts the on-line Asia-Pacific Internet Governance Information Service.

Contact :

Asia Pacific Development Information Programme (APDIP) of UNDP

P.O Box 12544,50782 Kuala Lumpur, Malaysia

Tel:603-255-9122 Fax:603-2539740

E-mail:info@apidip.net URL : www.apidip.net

2.8 The Workshop on Corpus-based Natural Language Processing on 17-31 Dec 2001 at Anna University

The Workshop on Corpus-based Natural Language Processing was held from 17th Dec 2001 to 31st Dec 2001 at AU-KBC Research Centre, MIT campus, Anna University jointly with The Language Technology Research Centre, IIT Hyderabad, The National Centre for Software Technology, Mumbai, The Resource Center for Indian Language Technology Solutions-Tamil, Anna University & The Tamil University, Thanjavur.

The objectives of the workshop were to provide an understanding of NLP in the context of Machine-Translation, Multilingual Information Retrieval and other applications related to Indian Languages and English, to introduce the usage of Statistical Techniques on Lexical Resources to refine rule-based methods. These would be used to develop the NLP applications and also to provide training in the use of Tools and Resources in the domain of Statistical processing.

The topics covered in detail in lectures, tutorials and labs were Finite State Automata and Finite State Transducers, Linguistic Formalisms and Computational Grammar, Hidden Markov Models, Parsing, Machine Learning, Word Sense Disambiguation, Statistical Machine Translation Information Retrieval.

In addition to the hands-on lab classes, the teams of participants were also involved in executing real-life Projects as a part of the workshop, broadly towards realizing an English-Indian language MT solution, and/or a cross-lingual Information Retrieval system. Necessary multilingual corpora and other resources are being created in Tamil, Hindi and Telugu, in addition to English and that was made available at the workshop.

There were 35 participants who belonged to different streams : Language and Linguistic streams and Science and Engineering streams such as Computers, Communications, Maths and Statistics.

The Resource persons of the Workshop were :

- Prof. Aravind K. Joshi, University of Pennsylvania, USA; Dr. B. Srinivas, AT&T Research, New Jersey, USA; Dr. Anoop Sarkar, University of Pennsylvania, USA; Dr. Rajeev Sangal, IIT, Hyderabad; Durgesh D. Rao, NCST, Mumbai; Sushma Bendre, IIT, Hyderabad
- Five Resource centers from Punjab, Anna University, Bangalore, Assam and Kerala presented their work to the participants at a specially arranged Session on 29.12.2001.



2.9 1st Workshop on Indian Language OCR on February 1-3, 2002 at University of Hyderabad

The Resource Centre for Indian Language Technology Solutions (Telugu), established by the Ministry of Information Technology, Govt. of India, conducted a three-day workshop on Indian Language OCR Systems.

Researchers from various centres including ISI (Kolkata), IISc (Bangalore), C-DAC (Pune), Thapar Institute (Patiala), IIIT(Hyderabad), Vicisoft Technologies, Secunderabad, DRDL Hyderabad, University of Mysore, and University of Hyderabad participated in the workshop. Each of the centres explained in depth all the technical details of their own systems.

Status report and plan of action resulted from the workshop as summarized below:

Status Report and Plan of Action

Various centres working on OCR systems for Indian scripts have been doing very well and full-fledged OCR systems can be developed in six to twelve months from now. These centres have been taking their own approaches which hold promise. There is still a lot of scope for further experimentation and fine tuning.

OCRs for different scripts being developed at various centres is as follows,

1. Optical Character Recognition System for Devanagari

Dr. Veena Bansal, IIT, Kanpur
Tel: 0512-597743 E-mail: veena@iitk.ac.in

Prof. B.B.Chaudhary, ISI, Kolkata
Tel: 033-5778085 E-mail: bbc@isical.ac.in

Shri M.D.Kulkarni, C-DAC, Pune
Tel: 020-5694000 E-mail: mdk@cdac.ernet.in

2. Optical Character Recognition System for Bangla

Prof. B.B.Chaudhary, ISI, Kolkata
Tel: 033-5778085 E-mail: bbc@isical.ac.in

3. Optical Character Recognition System for Oriya

Prof. B.B.Chaudhary, ISI, Kolkata
Tel: 033-5778085 E-mail: bbc@isical.ac.in

Prof (Mrs.) Sanghmitra Mohanty, Utkal University
Tel: 0674-580216 E-mail: sangham1@sancharnet.in

Sh. A.K. Pujari, OCAC Bhubaneswar
Tel. : 0674-543113 E-mail : akp@ocac.ernet.in

4. Optical Character Recognition System for Gurumukhi

Prf. G.S.Lehal, TIET, Patiala
Tel: 0175-214868 E-mail: gslehal@mailcity.com

5. Optical Character Recognition System for Telugu

Prof. K.Narayan Murthy, Univ. of Hyderabad, Hyderabad
Tel: 040-3010500 E-mail: knmcs@uohyd.ernet.in

6. Optical Character Recognition System for Tamil, Kannada

Prof. N.J.Rao, IISc. Bangalore
Tel: 080-3092222 E-mail: njrao@mgmt.iisc.ernet.in

7. Optical Character Recognition System for Malayalam

Prof. Ravinder Kumar, ER&DCI Trivendrum
Tel: 0471-320116 E-mail : ravi@erdcitvm.org

Various centres have also been planning for post processing techniques suitable for their own languages and scripts. However, it was stressed that many of the pre-processing modules are general purpose tools that can be exchanged for mutual benefit. This would save a lot of time and effort and avoid duplication of effort.

It was agreed that another technical workshop cum open competition may be planned sometime in August 2002 wherein various OCR systems can be tested thoroughly on standardized test data including laser printed documents as well as printed books. Newspapers can also be tried out as a challenge. The output of the OCR systems will be in ISCII/UNICODE.

The IIIT Hyderabad volunteered to host a mailing list with name as ILOCR so that the community could all be in close touch.



2.10 1st International Conference on Global WordNet on January 21-25, 2002 at CIIL Mysore

The Central Institute of Indian Languages, Mysore organized the 1st International Conference on Global WordNet from January 21-25, 2002 in collaboration with Global WordNet Association, Netherlands, IIT Bombay and IIIT Hyderabad. The total number of registered delegates was 81 from 19 countries including India.

There were 12 academic sessions apart from 2 sessions of tutorials, Introductory Session and Business Meeting.

In the introductory session, Prof. Udaya Narayana Singh (Director, CIIL) welcomed delegates, informed them about the Institute's activities and the possible contribution of the Institute in the area of Building WordNets in Indian Languages. Dr. Piek Vossen and Dr. Christiane Fellbaum in their introductory remarks gave an account of the development of WordNets with emphasis on Princeton and Euro WordNet.

The twelve academic sessions discussed the following themes-

- Building WordNets
- Aligning WordNets/Cross Linguistic Work
- Semantic Relations and Lexical Semantics
- Assigning Domain Labels
- Ontologies, Concepts, Top Levels
- Lexical Semantics
- Disambiguation and Semantic Annotation
- Sublanguages
- Applications
- Interfaces

During academic sessions, a total of 50 papers were presented. The accepted papers went through a two-way blind review.

There were two tutorial sessions. One was given jointly by Dr. Christiane Fellbaum, Dr. Piek Vossen and Dr. Palmira Marrafa and the other was given by Eneko Agirre and German Rigau together. Both tutorials discussed various issues in Building

WordNets based on the experiences while building Princeton, Euro and Basque WordNets.

The business meeting was open to all the participants and was conducted by three board members of Global WordNet Association, namely, Piek Vossen, Christiane Fellbaum and Palmira Marrafa. The issues discussed during business meeting were the activities and membership of of GWA, Communication among members, Standardization of WordNets and Future meetings.

The following are the recommendations:

1. Membership fee for the GWA for the period upto and including the next GWA Meeting will be waived for all the participants of the First International Conference and membership is already available through the GWN website.
2. Prof. Udaya Narayana Singh offered to establish a website exclusively for the WordNet related activities from the CIIL, Mysore site.
3. Standards must be developed with respect to following points:
 - (a) lexical and semantic relations (both content and labelling);
 - (b) representation (such as XML, as developed by IRST and Brno);
 - (c) the database
 - (d) shared tools
4. The available Industrial tools and interfaces should be shared as much as possible.
5. A tentative projected date for the next GWN conference is January, 2004. Interest in organizing the next meeting had been expressed informally during the meeting by Sofia Stamou (Patras, Greece) and Pavel Smrz (Brno, Czech Republic).

The Proceedings of the conference papers are published with an introduction by Dr. Piek Vossen & Dr. Christiane Fellbaum and Foreword by Prof. Udaya Narayana Singh. It contains all the 50 papers accepted for the conference.

The publication is priced at Rs. 160 (\$18.00). The Proceedings are also available in a CD form, which costs the same amount. These could be ordered for at CIIL, Mysore (Contact: rajya@ciil.stpmysoft.net).