6.0 Resource Centres for Indian Language Technology Solutions: Brief Achievements & Future Plans

Phase-I of Resource Centres :

Objectives:

Under the aegis of TDIL Programme thirteen (13) Resource Centres for Indian Language Technology Solutions (RCILTS) have been established at various educational institutes and R&D organizations covering all Indian Languages.

The Broad objectives of the phase – I of the Resource Centres are:

- To develop niche technologies (such as OCR, TTS, Text-editors, conversion utilities, spell-checkers etc) for providing IT solutions in Indian Languages
- To build repository of Information Processing tools and products for Indian Languages
- Technology dissemination through:
  - Specialized training programs
  - Interaction with State Government for e-governance

Technologies Developed During Phase – I

During the Phase-I of the project, the Resource Centres have developed several niche technologies and products. Some of the software tools and technologies are:

- On-line English to Hindi Machine translation system
- Localization of LINUX operating system for Six Indian Language (Sanskrit, Hindi, Marathi, Kannada, Tamil and Malayalam)
- Optical Character Recognition (OCR) for Seven Indian Languages (Hindi, Marathi, Bangla, Punjabi, Tamil, Telugu & Malayalam)
- Parallel corpora of one million pages covering 11 Indian languages
- Text-to-speech systems (Hindi, Bangla, Malayalam, Tamil, Oriya)
- Basic Information Processing tools e.g. text-editors, spell checkers, fonts, keyboard drivers, bi-lingual dictionaries
- Messaging Systems in 2 Indian Languages (Hindi, Malayalam)
- Standardization of UNICODE for all Indian Languages (proposed changes finalized in consultation with state governments and most of the changes are accepted and incorporated in UNICODE4.0)

Most of the products have been tested by STQC.

2. Packaging / Productization of the technologies (Intermediate phase):

Urgent Need was felt for productizing / packaging of the prototype technologies developed. In this context it was suggested that the Resource Centres may take help of C-DAC Pune for development of products from the prototype technologies. Based on this decision, the Project Review and Steering Group (PRSG) reviewed the performance of all RCs during Phase – I during Aug31-Sep01, 2004 and suggested extension of Phase – I to all RCs upto Dec 31, 2004 with an additional grant of Rs. 10.00 lakhs with the objective of productizing the prototype technologies.

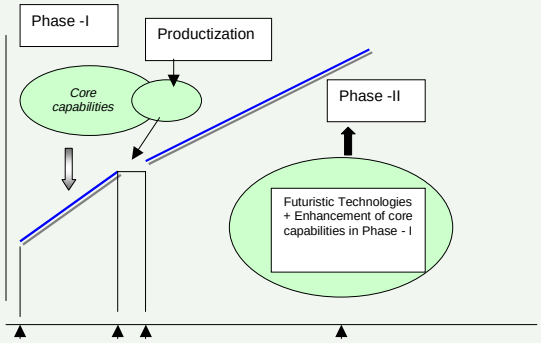
3. Conceptualization of the Phase –II of the Resource Centres for Indian Language Technology Solutions (RCILTS).

Objectives:

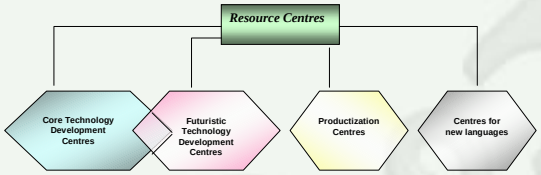
- Development of futuristic technologies in the fields of, Machine Translation Systems in Indian Languages, Speech-to-Speech translation system, Cross-lingual and Knowledge Management systems in multilingual environment.
- Development of innovative applications and services enabled with language interface towards bridging the existing Digital Divide.
- Deployment of Language Technology in thrust areas:

- e-gov
- e-learning
- e-rural prosperity

Schematic representation of RC-Phase –I & Ph. -II



Implementation Methodology for Phase -II:



In Phase –II , the Resource Centres would be classified into three categories:

(1) Centres responsible for development of Information Processing Tools developed in Phase - I

(2) Centres responsible for the development of futuristic technologies such as CLIR , Speech to Speech Translation and applications in thrust areas

(3) Centres responsible for prototyping and technology mentoring

(4) There would be common agenda for all centres to interact with state governments for developing tools for e-governance applications in local languages

(5) Collaborative development would be encouraged.

(6) Technological audit would be embedded in implementation strategy.

Moreover, new centres would be identified and development of core technologies would be initiated for four new constitutionally recognized languages namely Santhali, Maithili, Bodo, & Dogri.

4. Proposed Initiatives of Phase –II of Resource Centres (RCILTS):

Core Technologies:

1. Information Processing Tools for all Indian Languages

Enhanced Information Processing Tools for all Indian Languages comprising of:

- Fonts,
- KB drivers,
- Text editors,
- Spell Checkers,
- Tri-lingual dictionaries (Min. 60,000 words),
- Code/ File conversion utilities
- Messaging system,

2. Niche Technologies for e-gov. solutions

The Resource Centres would collaborate with state governments to develop necessary niche tools & technologies pertaining to e-gov. applications as per specific requirements of state governments

3. Optical Character Recognition System (OCR)

- Development of Optical Character Recognition Systems OCR for all languages with above 97% accuracy with multiple font support.
- Prototype for font independent OCR for 7 Languages
- Prototype for handwritten text OCR

4. Machine Translation Systems

- English to Hindi enhanced version (~ 95 % accuracy – simple sentences & 85 % accuracy

for complex sentences and support for large text files)

- Hindi to English (Example–based ; ~ 85 % accuracy)
- English to seven Indian languages

5. Development of open source Operating Systems and GUI tools in Indian Languages (LINUX platform)

Futuristic Technologies:

1. Cross-lingual Information Retrieval System

1.1 Resources

- *Indo-Word net* (60,000 words, 30,000 sync sets for each Indian Language) with Sanskrit as Indian Networking Language (INL).
- Interfacing with Universal Networking Language (UNL)
- *Tagged Corpora* of about 10 million words for all Indian languages.
- *Multilingual dictionaries* (at least Trilingual; e.g, English – Indian Language – Hindi) covering all Indian Languages.

1.2 Enabling Software tools

- Transliteration tools for Indian languages
- Automatic Summarization tools (10:1) for all Indian languages

1.3 Systems

- Machine Aided Translation systems (Text)
  - English to Hindi enhanced version (~ 95 % accuracy – simple sentences & 85 % accuracy for complex sentences and support for large text files)
  - Hindi to English (Example–based; ~ 85 % accuracy)
  - English to nine Indian languages

1.4 Specific Applications

- Cross-lingual database search engines for specific applications such as
  - e-governance
  - Crime-records
  - Railways
  - Medical transcription
- Cross-lingual support for Hand-held and wireless devices for SMS systems.
- Cross-lingual support for tourism application with GPS support.

2. Speech-to-speech translation system

2.1 Resources

- *Speech corpora* for all Indian Languages annotated with phoneme levels, Stress emphasis tags & phrase break tags (10Mn Words for each language)
- Task-oriented natural animated characters (Speech, lips, gesture) spoken dialogue for web (2000w)

2.2 Enabling Software tools

- Software to compute *prosodic pattern* (phoneme duration & pitch contour), to retrieve acoustics units (diphones)
- Speech Synthesis engine, Data sets (Acoustics & language models, vocabularies). Language model requires text data of over 50 Million words.
- Software tools for Morph analysis, disambiguation of homographs, letters to phonemes, assigning syllable stress, pitch contours, pitch accents & durations
- High quality competitive formant speech synthesis software in major Indian languages
- Software tools large coverage grammars with automatic acquisition for syntactic and semantic processing for limited applications.

2.3 Systems

- High Quality translation systems for limited domains (e.g. Business

Negotiations, Railway Enquiry, Sports, Tele-Medicine , Tourism) with machine learning capability (60,000 words)

- Medium-size vocabulary (3,000 W) conversational systems in limited domains
- Continuous speech understanding in workstations with standard dictionaries (5000 words) in major Indian Languages.
- Useful speech summarization systems in major Indian languages.

3. Knowledge-Management in Multilingual environment

- Development of tools for knowledge acquisition, representation and validation based on Indic logic system
- Development of UML type Object Oriented Software Engineering (OOSE) methodology for specifying information systems of a modern enterprise system
- Concept based networking Language
- To Develpment Intelligent Cognitive System, i.e. Software tools and products in multilingual environment through contextualization of Indic Knowledge.

Application Area :

Specific & customized applications with local language interface would be developed as per requirements of national level and state governments. The application domains are:

- E-Gov
- E-education &
- E-rural for village prosperity.

High Impact Applications:

- Voice enabled multilingual Rural Kiosks
- Voice-enabled Multilingual Telephone-based Information Access System for Railways & Tourism
- WebBharati for Trans-lingual Web Access.
- Multi-modal Digital Library Creations & Retrieval System.

Courtesty:
S.A. Kumar
skumar@mit.gov.in
Somnath Chnadra
schandra@mit.gov.in
Department of IT
New Delhi - 110003