Portals Supporting Indian Languages

The following Websites support composing in Hindi and other Indian languages and have all other features which occur in normal e-mailing sites like inbox, addresses, compose, folders etc.

1. Web Dunia: www.epatra.com supports 11 languages
2. Mithi.com : www.mailjol.com supports 12 languages
3. Langoo: www.langoo.com supports 12 languages
4. CDAC: www.cdacindia.com with multilingual support

Other Efforts

1. A survey has already been conducted on "Study on Potential for Local Language Computing in India" by IIIT, Bangalore and Manufacturers Association for Information Technology (MAIT). This covers Hindi and Tamil as case study. It was found that the potential for multilingual computing solution is estimated at around Rs. 120-125 crs over a 5 year time frame (Rs. 38 crs. in the States of Hindi belt and Tamil Nadu.)
2. The following has been accomplished through the initiatives taken by the Ministry of Information Technology (MIT), MIT societies, Private Sector & Multinational Companies. The efforts in pipeline are also listed with suitable remarks:

a. Operating Systems:

- PC-DOS (Disk Operating System) in Hindi
- Windows 2000 with Hindi support
- Language Tool Kit for Apple Computers

b. Environment for Hindi:

- GIST Card for UNIX and DOS Platforms
- GIST Shell for DOS
- GIST Terminal for Multi-user Platforms
- GIST SDK
- Toolkits for S/W developers for developing Indian Language solutions

c. Commonly used Packages in Hindi:

- Word Processing Packages
- Spreadsheet
- Database Management
- Desk-Top Publishing
- Fonts & Clip Art S/W
- Plug-ins for various Windows based S/W
- Accounting Packages
- Banking S/W
- Spell Checkers
- Electronic Dictionary (English-to-Hindi)
- Transliteration
- Internet Content Creation Tools
- Mailing software

d. Customised Applications with Hindi support

- Doordarshan Applications like video subtitling
- Pagers
- Indian Railway Reservation & Ticketing
- Election ID Cards
- Land Records Management System
- Hindi Teaching Packages

e. Internet Based Services in Hindi

- E-mail
- Chat
- Newspapers and a variety of content

f. Training material in Hindi

- DOEACC 'O' Level Courseware

g. Hindi S/W in Public Domain

- Two Fonts with keyboard drivers (downloadable)
- Basic word processors

h. Introduction of Hindi in IT curricula

- MIT played proactive role of introducing Hindi in IT curricula designed for Secondary and Senior Secondary schools of CBSE.
- Also prepared IT-enriched curriculum for functional Hindi at BA & MA levels.

i. Consortium of Indian Language Technology Developers

- MIT has taken initiative to form a consortium which is due to be announced

j. Hindi in UNL (Universal Networking Language) of UN

- IIT, Mumbai is nodal agency for developing dictionary, enconverter and deconverter for the Universal Networking Language - a major initiative of UN University that encompasses 15 world languages. Hindi is one of these.

Major Events of Year 2000

- Meeting of UTC #85: Unicode Technical Committee on 7-10 November, 2000 at NCR Office at San Diego

There are 25 full members and 43 associate members of the Unicode Consortium. 13 full members participated. About 100 different issues/ motions were discussed. Separate discussion on Indic scripts took place in a subgroup consisting of Mr. Ken Whistler (representing Apple & Sybase), Dr. Mark E. Davis (President Unicode, also representing IBM), Mr. Michael Kaplan (representing Trigeminal Software) and Mr. Hideki Hiura (representing Sun Microsystems) and Dr. Om Vikas (representing Ministry of Information Technology, Govt. of India).

On 10th November, 2000 a separate time slot was allocated for discussion on Indic Scripts in the full Unicode meeting. The following points were highlighted:

- Unicode included the Indian Script Standard Code for Information Interchange (ISCII) proposed in 1988 and made some arbitrary modifications in various Indic scripts resulting into omissions of certain characters or inclusion of non-used symbols and change in alphabetic orders. ISCII (Indian Script Standard for Standard Code for Information Interchange) was further revised in 1991 and published by BIS as IS 13194:1991. However, Unicode has not revised their code as per their laid down policy of Unicode Consortium regarding character encoding stability :
  a) Once a character is encoded, it will not be moved or removed.
  b) Once a character is encoded, its character name will not be changed.
  c) Once a character is encoded, its canonical combining class and decomposition (either canonical or compatibility) will not be changed in a way that would affect normalization.
  d) Once a character is encoded, its properties may still be changed, but not in such a way as to change the fundamental identity of the character.
  e) The structure of certain property values in the Unicode Character Database will not be changed.
- There is no separate codes for Assemmese script. Two characters of Assemmese needs to be added to Bengali code block if that is used for Assemmese as well.
- The Currency signs need to be added in each Indic script code set. Currency sign in Gujarati is also used as a symbol in the text.
- In Devnagari code block, proper interpretation is necessary for some of these symbols, for example, "Halant" is mentioned as Viram this is wrong and needs to be corrected. Some glyph-shapes need to be redone. Devanagari is used for Hindi , Marathi, Konkani and Sanskrit . Hence, enhanced Devnagari code block includes a few additional characters for phonetic transcription.

  The interpretation of combining 'Consonant' with 'Matra' in graphic way and 'Consonant', 'Halant' and 'Vowel' in grammatical sense leading to the same result, the syllable (Consonant plus Vowel Modifier) needs to be added as detailed interpretation. Nukta is a diacritical mark there is no need to provide separate code points for consonants with Nukta as at 0958 …. 095F of Devnagari Code block on page 401 of Unicode 3.0.
- There is need to have a separate code block for Vedic Sanskrit.
- In Tamil, Visarga at Code point OB8/3 is not used, this is a separate character called ayutham. Character and the Code point OB8/2 is not in use and can be removed. On page 239 of Unicode 3.0 last three occasions may be deleted. 384 Tamil Syllable may be added as Hangul Syllables as in CJK.
- In Oriya symbol corresponding to Va in Devnagari is missing. This needs to be added.
- Proper interpretation of formation of conjuncts using Zero-width joiner, Non-Zero-width joiner, Halant and Invisible character may also be added.
- Addition of EXT code point for extending the code block may be necessary. It was, however, not agreed a point in view of the Unicode policy.
- UTC Working Group unanimously appreciated the problems of Indian scripts and agreed that India-representative should submit a separate proposal for addition of missing symbols, and updates on interpretation. Moreover, detailed text, as available for Devnagari, in the Unicode, should be made available for other scripts as well. A separate proposal for Vedic code block/points should also be submitted.
- UTC Working Group also agreed for sending Unicode experts to the Unicode Workshop in India.

ZOPP Workshop

Under TDIL programme have been 13 Resource Centres for Indian Language Technology Solutions. The activities of these resource centres are broadly classified into the following four categories.

- Core Activities
- Services and Knowledge Bases
- Product Development
- Applied Research

ZOPP workshop was held during the period of 3rd -5th may, 2000 at IISC, Bangalore (one of the Resource Centres) in order to arrive at a common understanding of the goals of Project and enable each Resource Centres to plan and implement their activities accordingly. ZOPP stands for Ziel Orientierte Projekt Planning that means Objectives Oriented Project Planning.

The participants identified all the agencies connected with the project and classified them. Subsequently, a detailed Participation Analysis (PA) was made in smaller groups taking up the agency, its interests and expectations, potential and the project approach to utilize the potential. For example, the group of users and their needs were identified and then the project approach towards these users was formulated.

The participants then discussed the concept of Project Planning Matrix (PPM) containing the hierarchy of objectives namely, overall goal, project purpose, results/outputs, and activities. After considerable discussions the goal and purpose were identified. With regard to outputs, the consensus was that there would be two sets, one consisting of common outputs, for all Resource Centres and the other set will be centre specific. In consonance with this approach five outputs were formulated.

- Design, development and deployment of portals in Indian languages to demonstrate technologies that are developed.
- Design and organization of training programmes for agencies and persons concerned with deployment of Indian language processing systems.
- Knowledge and databases on the resource information on Indian languages created.
- Development of a spell checker, which is Unicode compliant.
- At least 10 most respected books, related to Indian heritage, in Indian languages on the web.

For each of these outputs specific Indicators and Means of verification were also worked out. With regard to centre specific activities, it was agreed that each centre would do a similar exercise and submit a PPM.

The role of C-DAC in transferring the technology already developed was recognized. It was agreed that a technology workshop would be conducted with C-DAC in Pune by August 2000 to enable the Resource Centres for a speedy take off in the respective activity. It was also felt that the experience of IIT, Kanpur would need to be shared with other Resource Centres for successful execution of the project.

Discussions also took place on the review mechanism of the project. It was agreed that the indicators evolved from the Project Planning Matrix would be utilized for monitoring the progress. Six monthly reviews were recommended and each centre giving the progress based on the indicators would make a report.

GIST Technology Workshop

A technology workshop for Resource Centres was organized by C-DAC during the period 17-7-2000 to 21-7-2000. The need for this workshop was suggested during the ZOPP workshop held in Bangalore. Main idea of this workshop was to provide training to all the Resource Centres on the tools and technologies developed by C-DAC, and also to evolve a collaborative development approach among the Resource Centres.

The main topics that were covered are

- Spell checker architecture.
- How to build spell checkers/dictionaries.
- Issues related to Indian language keyboard management.
- Introduction to GIST SDK.
- Using GIST SDK for VC/VB/Web applications.
- Methodology of development of search engines.
- Development and rendering of Indian Language Fonts.
- Development of web applications using iPlugin.

OCR Technology Presentation

On 21-7-2000 Presentations were made on the development of OCR technology for Devnagari script by C-DAC, Pune and ISI, Kolkata. ISI demonstrated the product that was developed. It was felt the ISI product is capable of delivering 98% accuracy and exhibit font invariance .The means of making it a commercial product need to be explored.

Indian Languages Technology Vision 2010

Tentative Guidelines evolved during TDIL WG Meeting on 19th October 2000

Vision statement

Digital Unite and Knowledge for all

Mission Statement

Communicating and Moving up the knowledge chain overcoming Language barrier.

Intermediate Goals

- Standardization of code, font, keyboard, and Indian Script Roman Code.
- Coexistence of Unicode and ISCII. Code conversion utilities be available.
- 3-4 fonts should be standardized and be in public domain. The font and content development should be encouraged.
- IT Localization Information Services Cell should be setup that should provide information on phone and through e-mail of Indian Language Technology products and services.

Short-term Goals

- Corpus creation (about 10 million words) and analysis: Unilingual corpus, parallel corpus automated tagging and statistical dictionary.
- Web content creation.
- Applications: land records, voters list, reservation chart, cooperative stores, telephone dictionaries on web.
- Language Technology should be integrated into IT curricula and the projects. Curricula for functional languages and linguistics should cover the issues related to language technologies.
- Collaborative development of Indian language lexical resources.
- Development of web enabled courseware.
- Writing aids (Spell checkers, grammar checkers and text summarization utilities).
- Sharing of standardized lexware (written text and speech).
- Development of lexware tools.
- 1-4 weeks training programmes on ILT awareness, lexware development.

Medium-term Goals

- Indian language speech database
- Multilingual, multimedia, content development with semantic indexing, classical and multi font and decorative fonts, off-line/on-line 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 projects/dissertations in the field of Indian Language Technology.
- Machine aided translation: English to Indian languages, among Indian languages, Indian languages to English and other foreign languages.
- Web based translation.

Long-term Goals

- Speech to Speech Translation.
- Human Inspiring Systems.

Readers are requested to give their feedback on Indian Language Technology Vision 2010

E-mail : info@tdil.gov.in