Book Shelf

- Internationalisation with Visual Basic
  – the authoritative solution

by Michael S. Kaplan, SAMS Sept. 2000
(a division of Machmillan, USA)
ISBN : 0-672-31977-2

This book introduces authoritative solutions to build localised Visual Basic applications, international websites, to support international forms & formats, to create locale-friendly user interfaces & to handle regional calender issues. This helps in understanding how to globalise components, to create localizable Visual Basic applications, handle unicode support properly and to implement multilingual database systems. Indian scripts are represented in the UNICODE. Part 1 deals with globalization issues part 2 with multinationalization issues, part 3 with localisation tools, part 4 with Data Bases, part 5 with internet applications and components in multilingual context & part 6 with application content & help systems.

Resource Center(s) for Indian Language Technology Solutions
1st Year (2000 - 2001) Progress

1. RCs - ILTS : IIT, Kanpur ::Hindi, Nepali

- Two sites // www.gitasupersite.org// and // www.bimari-jankari.org// launched;
- A printed text OCR error processor being developed. OCR Performance improved from 83 to 95 %. Dictionary enriched from 25,000 to 1,00,000 words;
- Processor for recognizing Hindi Chhand ("doha, chaupayi, …") has been developed;
- Tutorials on Building IL websites, I/O in ILs and Lexical data-base design conducted; and
- e-governance site for UP Trade Tax being developed.

2. RCs - ILTS : ISI, Kolkata :: Bengali

- Work on a web site on Eastern Indian Language and language technologies being progressed;
- Work initiated for preparation of Electronic corpus of document images, text, and speech data;
- OCR for Bangla and Hindi printed text ready. Performance about 97 %;
- Work initiated on Bangla font generation, OCR system for Oriya, adoption of Bangla OCR to Assamese, and Information retrieval system for Bangla documents;
- Work initiated on Neural Networks based printed document processing, Automatic processing of hand printed table-form documents, and script identification and separation in Bangla, English, and Urdu.

3. RCs - ILTS : IIT, Guwahati :: Assamese & Manipuri

- A web site having linguistic map of the NE, with relevant information on 12 different languages of the region, is under progress;
- Work is progressing on Assamese fonts, on-line Assamese dictionary, language processing, speech recognition, and a search engine;
- Implementation of a GIS Knowledge Base for NE road map is progressing; and
- e-Shopping and e-Tourism sites are being progressed

4. RCs - ILTS -1 : Utkal Univ. :: Oriya

- Work initiated for printed text Oriya OCR. Literature survey completed;
- Research work on Oriya speech processing initiated;
- Literature survey initiated for preparation of a Knowledge Base System on Indian philosophy; and
- Work also initiated on Font design, and Machine Translation.

5. RCs - ILTS -2 : OCAC :: Oriya

- Work is in progress for Electronic Dictionary/ Thesaurus, Spell checker on Linux platform;
- Towards Net Education in Oriya language, preparation of course material has been initiated, and a Chat server with voice and text facility is nearing completion;
- Work initiated for CBT package for schools teaching in the Oriya medium; and
- Work initiated for multimedia and GIS based content creation, NLP base interface to Databases, and creation of web documents in XML.

6. RCs - ILTS : Univ. of Hyd. :: Telugu

- A bi-lingual English-Telugu web site is being designed, and a Database for a Telugu network is being created;
- Work on Telugu lexicon and Morphology being progressed;
- Two books identified for creating on-line knowledge base on Telugu History, Society, and Culture. Electronic reproduction aspects being progressed;
- Electronic reproduction rights for about 50 Telugu classics obtained, and 30 identified with expired rights for making literature on-line;
- Work initiated for printed text OCR and Spell checker;
- Preliminary design of a Virtual Class Room completed; and
- Work initiated for a text editor and Neural Network based Natural Language understanding.

7. RCs - ILTS : IISc. Bangalore :: Kannada

- Kannudi - a bilingual English-Kannada comprehensive web site is being progressed;
- Knowledge bases on Indian Aesthetics to facilitate creation of India brand, on Indian Logic (Nyaya Sastra), and Sudarshana (Darshan Sastra) are being progressed;
- Resource Information on European History & Geography created;
- A S/W tool (OCR) for converting printed information in South Indian Languages into text is being progressed;
- High quality Kannada fonts created for public domain use;
- A course package for learning German and Russian through Hindi is being progressed;
- Research work on Kannada speech synthesis and automatically classify language/accent in 4 South Indian Languages and Indish (Indian English) progressing.

8. RCs - ILTS : Anna Univ. :: Tamil

- A demo web site has been created;
- Work is progressing for Tamil word processor, on-line dictionary, search engine, selection/ identification of books for on-line availability;
- Study of Nyaya theory and Tholkappiam initiated; and
- Work on Tamil Text-to-Speech, Automatic Software Document Generation, and Tamil Office Suite are additionally being proposed.

9. RCs - ILTS : ER&DCI (T) :: Malayalam

- Design of a Malayalam web site and portal completed. Equipment being readied for web site hosting. A number of sites having Malayalam content have been surveyed;
- MOU signed with a local Newspaper "Pradeepam" for on-line publication. Action initiated for web hosting of some Ayurvedic and Literary classics;
- A Malayalam spell checker, e-commerce site, and a Knowledge Base of traditional Kerala Art Forms, Literature and Culture are being progressed;
- 4 Malayalam fonts, a sorting algorithm, a tutor package to learn Malayalam easily, and an Expert System package for estimation of " Prakruti" as per Ayurveda system have been developed.

10. RCs - ILTS : IIT, Mumbai :: Marathi, Konkani

- Indo Wordnet - massive lexical knowledge base for MT and NLP - for Hindi being developed;
- Content creation and Search Engine for Marathi being progressed. Works of Saint Tukaram have been web enabled. Work on web enabling Marathi Grammar by Damle in progress;
- Towards MT among Marathi-Hindi-English a lexical base with 60,000 entries prepared; and work on an Universal Networking Language (UNL) is progressing;
- High school level web books to teach Algebra, Geometry and English through Marathi are being progressed;
- Marathi Speech Interface for Internet being Progressed;
- A Symposium on UNL has been organized.

11. RCs - ILTS : MSU, Vadodra :: Gujarati

- A Gujarati website with search engine designed & tested;
- Sarasvati Chandra - the 19th century classic in the authors own handwriting - to be put on web site. Also talks and lectures by Sri Auribindo proposed to be put on the site;
- A workshop for website writers in Gujarati held and finalized overall structure of database in Gujarati;
- A collaborative network with 6 other institutions established; and
- Work on fonts, Unicode and lexicography also progressing.

12. RCs - ILTS : TIE&T, Patiala:: Punjabi

- A Gurumukhi printed text OCR with over 96% accuracy developed, and tested with 8 fonts;
- A Punjabi-English Dictionary with 30,000 words/ phrases and having search and other features developed;
- An on-line technical dictionary of English terms in Punjabi developed;
- Four Punjabi classics have been put on-line and Punjabi dynamic fonts created; and
- Work on a Punjabi literature site, a Punjabi teaching site, Punjabi word processor with built in Spell-checker, Enhancement of Punjabi-English dictionary with sound facility and off-line English-Punjabi/ Punjabi-Hindi dictionaries being progressed.

13. RCs - ILTS : C-DAC, Pune :: Perso-Arabic (Urdu/ Sindhi/ Kashmere)

- A workshop was conducted for usage of GIST series of Indian Language Technologies, Tools and Products for the Resource Centers;
- Creation of a network of persons and organizations active in IT applications in Perso-Arabic initiated;
- Initiated work on Glyph standard for Font development. A Nastaliq style digital font designed for printing and display;

- Towards development of Script Core Technology a Glyph Property Editor and a Rule Engine has been completed; and
- Work is in progress for Urdu Spell Checker, Urdu-Hindi and Hindi-Urdu Dictionary, Muslim names and Technical Dictionary, and Converters.

14. RCs - ILTS : JNU : Chinese, Japanese & Sanskrit

- Teaching learning system for Sanskrit under development. 5 lessons with multi-media support designed & implemented;
- Data structure for Sanskrit lexicon designed. Implementation and data entry being progressed;
- A web site to facilitate Chinese tourist being progressed; and
- Web lessons for learning Chinese and Japanese languages are being progressed.

Impediments in IT Localization & Penetration

- Lack of industry involvement due to constrained demand;
- Subcritical and Unsustained demand in BIMARU & other States;
- R&D in language technology so far open ended, not product-driven in time-bound manner;
- Negligible software tools and re-usable components in public domain;
- No incentive for Computer Scientists working in the area of Indian Language Processing and also limited scope of their career advancement;
- No formal IT courses in the curricula for linguistics, language teaching, and language studies. Lack of IT culture in language graduates;
- Unable to check import of IT products and services which don't support Indian language(s);
- No consensus on National Standards: ISCII-88, ISCII-91, UNICODE, and other propriety codes;
- Content is largely glyph-coded not (ISCII) character-coded;
- Slow pace of transfer of language technology. No strategy for language technology marketing;
- Industry consortium for language technology yet to form.

Feedback on UNICODE Standard 3.0

- UNICODE is emerging as international de-facto industry standard. It is 2-byte code. The Unicode consortium provides a number of on-line resources for obtaining information and data about the Unicode standard, as well as updates and corrigenda. These are:

  Unicode Web site: http://www.unicode.org;
  Unicode Public Mailing List:unicode@unicode.org
- The following are of our relevance in the UNICODE standard version 3.0.

(a) IPA extensions (0250-02AF) on PP 359-367

(b) Devanagari (0900 - 097F), Bengali (0980 - 09FF), Gurmukhi (0A00 - 0A7F), Gujarati (0A80 - 0AFF), Oriya (0B00 - 0B7F), Tamil (0B80 - 0BFF), Telugu (0C00 - 0C7F), Kannada (0C80 - 0CFF), Malayalam (0D00 - 0D7F), on pp 211-235 and 401-425

- Observations and comments :

(i) Sinhala (0D80 - 0DFF), Tibbetan (0F00 - 0FBF), and Myanmar (1000 - 109F) follow the alphabetic order similar as in DEVANAGARI and other Indian scripts.

(ii) Elaborate description for DEVANAGARI (U+0900 - U+097F) is given on pp 211 - 223. This needs review in view of some additions/revisions. This includes sections on Standards, Encoding Principles, Rendering mechanism, Consonant letters, Independent Vowel letters, Dependent Vowel signs (matras), Viram, consonants conjuncts, Rules for Rendering (R1 - R14), Memory Representation and Rendering order, sample half-forms and ligatures, combining marks, digits and punctuation symbols.

(iii) Use of VIRAM in Unicode 3.0 is wrong. This must be replaced by HALANT. VIRAM means full stop as punctuation mark. Halant suppresses inherent vowel, and thus used with consonant to represent pure consonant.

(iv) Rendering mechanism is suggested as follows :

(a) Linear conjunction

क	्	व	= क्व
0915	094D	0935	
क	्	ष	= क्ष
0915	094D	0937	
त	्	र	= त्र
0924	094D	0930	
ज	्	ञ	= ज्ञ
091C	094D	091E	

(b) Vertical conjunction

क	ZW J	व	= क्व
0915	200D	0935	