6. Institutional Initiatives

6.1 A Brief History of Language Technology Research at I.I.T. Kanpur

The work on Computer Processing of Indian Languages and scripts started in early seventies at IIT Kanpur. The dominant role that the computer could play in solving complex problems posed by the plurality of languages and scripts in the country, was very well visualized by the researchers at IITK and some ice-breaking work done in this area provided the foundation to the present R & D effort. The work started on optical character recognition (R.M.K. Sinha and H.N. Mahabala), and on developing keyboarding and coding schemes (P.V.H.L.Narasimham, V.Rajaraman and B.Prasada; R.M.K. Sinha and H.N.Mahabala). However due to the exorbitant cost of the computers which could be afforded only by a few in those days, these remained more or less an academic exercise. With the advent of microprocessors in mid seventies, it became economically viable, to translate some of these ideas into a stand-alone Indian language terminal design (R.M.K. Sinha and Arjun Raman). A number of B.Tech and M.Tech project works were devoted to this.

In 1978, with the initiation of IIT Kanpur (R.M.K. Sinha), a National Symposium was organized on Linguistic Implications in Computer based Information Systems by Department of Electronics (DOE) (Om Vikas), Govt. of India. This triggered widespread activity in the area in the country. The work at IIT Kanpur got a real fillip when a project on design and development of 'Integrated Devanagari Computer (IDC)' terminal was sponsored by DOE, Govt. of India in 1983 (R.M.K. Sinha, S.K. Mullick). The IDC terminal was designed in a record time of about 8 months and was demonstrated at the Third World Hindi Convention at Delhi. It was developed using Intel 8086 processor with multitasking firmware. The IDC project was further extended to implement the same technology using the 32-bit 68000 microprocessor and the outcome was named as GIST (Graphics and Indian Script Terminal) technology. A number of companies bought this technology for manufacturing multilingual computer terminals. This GIST technology was adapted by the Centre for Development of Advanced Computing (C-DAC). In 1984, Journal of Institution of Electronics and Telecommunication Engineering published a special issue on Computer Processing of Indian languages and scripts. This special issue carried several articles on results of research and development works at IIT Kanpur. Prof. Sinha's research on comparison of different possible coding schemes, keyboarding schemes, pros & cons of phonetic keyboarding and internal representation, its inherent transliteration capability, schema for machine translation based on interlingua, etc were presented in the lead article. Some of the other articles presented, for the first time, strategy for English to Hindi and Hindi to English transliteration (R.M.K. Sinha, B. Srinivasan), spell-checker (R.M.K. Sinha and K.S. Singh), segment display (R.M.K. Sinha). This special issue became a reference material for researchers in this area.

The GIST technology represented a major breakthrough in solving our complex problem of man-machine linguistic interface for Indian languages. This technology incorporated several desirable features. A natural phonetically oriented keyboarding scheme directly converting to internal codes called ISSCII-8 (8-bit Indian Standard Script Code for Information Interchange), a human engineered keyboard layout, a display which dynamically changes as the input progresses, built-in intelligence to disallow illegal compositions such as attaching two vowel modifiers on the same character, automatic transliteration from one Indian script to another, are some of the key attractive features making it user friendly. ISSCII-8 is an extension of ASCII which has been designed during early 1980's with active inputs from IIT Kanpur, caters to the entire set of Indian scripts in an uniform way. ISSCII-8 has undergone further modifications and a modified version has been accepted by Bureau of Indian Standards as ISCII (8-bit Indian Script Code for Information Interchange) code in 1991. ISCII forms the basis for UNICODE code assignments for Indian scripts.

In 1992, UNESCO and UNDP sponsored the Second Regional Workshop on Computer Processing of Asian Languages (CPAL-2) at IIT Kanpur under the Chairmanship of Prof. R.M.K.

Sinha. (The first CPAL was held at Asian Institute of Technology, Bangkok in 1989.) This workshop was attended by several international experts and a set of recommendations on current issues were generated out of the panel discussions which were submitted to UNESCO.

In 1993 IITK designed and developed a Machine Aided Translation system for translation from English to Indian Languages under the leadership of Professor R.M.K. Sinha with support from MHRD funds. This system was named as ANGLABHARTI and the underlying methodology named as ANGLABHARTI Technology or ANGLABHARTI Approach. In 1994, IITK (R.M.K. Sinha) implemented the Anglabharti system on Sun OS environment for translation from English to Hindi. All the modules of the systems were implemented, tested and demonstrated. In 1995, Department of Electronics, Govt. of India, sanctioned a grant-in-aid for implementation of the project titled "Machine Aided Translation from English to Hindi for standard documents (domain of Public Health Campaign) based on ANGLABHARTI approach" for which ERDC (with its office at Lucknow and now moved to NOIDA) was associated for implementation and commercialization of this software on a PC platform in the domain of public health campaign. The ANGLABHARTI software already developed by IITK on SUN system was used in this project and was implemented (re-engineered) on PC under Linux jointly by IITK and ERDC under the supervision of IITK (R.M.K. Sinha, Ajai Jain). In 1996, IITK also designed and developed an Example-based approach for Machine Aided Translation for similar (Indian languages) and dissimilar (English and Indian Languages) under the leadership of Professor R.M.K. Sinha. This approach has been named as ANUBHARTI approach. A system to translate from Hindi to English has been implemented based on ANUBHARTI approach by IITK (R.M.K. Sinha, Ajai Jain and Renu Jain) .

Currently, AnglaHindi, the English to Hindi MAT based on Anglabharti methodology, which accepts unconstrained text, has already been made available to the users and is very well received. AnglaUrdu which is based on AnglaHindi has also been demonstrated. HindiAngla, the Hindi to English MAT based on Anubharti methodology, has been demonstrated for simple sentences and further work is going on to handle compound and complex sentences. The current research at IITK is focused towards development of more efficient machine translation strategies with user friendly interfaces for these systems. Another dimension of diversification for future, is to cater to all other Indian languages by implementing AnglaSanskrit, AnglaBangala, AnglaPunjabi, and so on; SanskritAngla, BangalaAngla, PunjabiAngla, and so on; and HindiSanskrit, HindiBangala, and so on; based on hybridization of Anglabharti and Anubharti methodologies.

R & D Activities at Language Technology Laboratory, Indian Institute of Technology, Kanpur

URL: http://www.cse.iitk.ac.in/users/langtech

Machine Translation

Chief Investigator: Dr. R.M.K. Sinha
Co-Investigator: Dr. A. Jain

The work on machine translation started in early eighties when we proposed using Sanskrit as interlingua for translation to and from Indian languages (See the paper on "Computer processing of Indian languages and scripts - Potentialities and Problems", Jour. of Inst. Electron. & Telecom. Engrs., vol.30, no.6, 1984). This was further elaborated in CPAL-1 paper presented at Bangkok in 1989.

Later in 1991, the concept of a Pseudo-Interlingua was developed which exploited structural commonality of a group of languages. This concept has been used in development of machine-aided translation methodology named ANGLABHARTI for translation from English to Indian languages.

Anglabharti is a pattern directed rule based system with context free grammar like structure for English (source language). It generates a 'pseudo-target' (Pseudeo-Interlingua) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bangla, Asamiya, Punjabi, Marathi, Oriya, Gujrati etc.), Dravidian family

(Tamil, Telugu, Kannada & Malayalam) and others. A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the 'pseudo-target' is constructed. Within each group the languages exhibit a high degree of structural homogeneity. We exploit the similarity to a great extent in our system. A language specific text-generator converts the 'pseudo-target' code into target language text. Paninian framework based on Sanskrit grammar using Karak (similar to case) relationship provides an uniform way of designing the Indian language text generators. We also use an example-base to identify noun and verb phrasals and resolve their semantics. An attempt is made to resolve most of the ambiguities using ontology, syntactic & semantic tags and some pragmatic rules. The unresolved ambiguities are left for human post-editing. Some of the major design considerations in design of Anglabharti have been aimed at providing a practical aid for translation wherein an attempt is made to get 90% of the task done by the machine and 10% left to the human post-editing; a system which could grow incrementally to handle more complex situations; an uniform mechanism by which translation from English to majority of Indian languages with attachment of appropriate text generator modules; and human engineered man-machine interface to facilitate both its usage and augmentation. The translation system has also been interfaced with text-to-speech module and OCR input.

This project also received funding from TDIL programme of Govt. of India during 1995-97 and 2000-2002.

The English to Hindi version named AnglaHindi, of Anglabharti machine aided translation system has been web-enabled and is available at URL: http://anglahindi.iitk.ac.in

The technical know-how of this technology has been transferred on a non-exclusive basis to ER&DCI/CDAC Noida for commercialization.

A system for translating English to Urdu, named AnglaUrdu, has also been developed using our AnglaHindi system and Urdu display software of CDAC, Pune.

In 1995, we developed another approach for MT which was example-based. Here the pre-stored example-base forms the basis for translation. The translation is obtained by matching the input sentence with the minimum 'distance' example sentence. In our approach, we do not store the examples in the raw form. The examples are abstracted to contain the category/class information to a great extent. This makes the example-base smaller in size and further partitioning reduces the search space. The creation and growth of the example-base is also done in an interactive way. This methodology, named ANUBHARTI, has been used for Hindi to English translation and further details of this approach can be seen in the Ph.D. thesis of Renu Jain.

The Anubharti approach works more efficiently for similar languages such as among Indian languages. In such cases the word-order remains the same and one need not have pointers to establish correspondences.

Currently, we are working towards developing an Integrated Machine-aided translation system (with funding from TDIL programme of Govt. of India, 2003 onwards) hybridizing the rule-based approach of Anglabharti, example-based approach of Anubharti, corpus/statistical based approaches to get the best out of these approaches. This is also being explored to be used for translation engine of speech to speech translation system.

In parallel, we are also developing MAT system for Hindi to English translation system, HindiAngla, based on our Anubharti approach with funding from CoILNET project of Govt. of India (2001 onwards). AnglaHindi and HindiAngla have been used to demonstrate the two way reverse translation for simple sentences.

Some Relevant Publications

- R.M.K. Sinha and A.Jain, AnglaHindi: An English to Hindi Machine-Aided Translation System, MT Summit IX, New Orleans, Louisiana, USA, Sept. 23-27,2003.
- R.M.K. Sinha, Translating News Headings from English to Hindi, 6th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2002), Banff, Canada, July 17-19, 2002.

- R.M.K. Sinha, Towards Speech to Speech Translation, Key-note presentation at Symposium on Translation Support Systems (STRANS2002), March 15-17, 2002, Kanpur, India.
- Vartika Bhandari, R.M.K. Sinha and Ajai Jain, Disambiguation of Phrasal Verb Occurrence for Machine Translation, Proc. Symposium on Translation Support Systems (STRANS2002), March 15-17, 2002, Kanpur, India.
- Ajai Jain, R.M.K. Sinha and Renu Jain, On Translating Unconstrained Text, Proc. Symposium on Translation Support Systems (STRANS2002), March 15-17, 2002, Kanpur, India.
- R.M.K. Sinha, Multilinguality and Global Digital Divide, Joint IAMCR/ICA International Symposium on the Digital Divide, November 16-17, 2001, Austin, USA.
- R.M.K. Sinha, Dealing with Unknown Lexicons in Machine Translation from English to Hindi, Proc. of IASTED International Conference on Artificial Intelligence and Soft Computing, May 21-24, 2001, Cancun, Mexico, pp 333-336.
- R.M.K. Sinha, Renu Jain and Ajai Jain, Translation from English to Indian Languages: ANGLABHARTI Approach, Proc. Symposium on Translation Support Systems (STRANS2001), February 15-17, 2001, Kanpur, India.
- Renu Jain, R.M.K. Sinha and Ajai Jain, ANUBHARTI: Using Hybrid Example-Based Approach for Machine Translation Proc. Symposium on Translation Support Systems (STRANS2001), February 15-17, 2001, Kanpur, India.
- R.M.K. Sinha, Hybridizing Rule-Based and Example-Based Approaches in Machine Aided Translation System, 2000 International Conference on Artificial Intelligence (IC-AI'2000) June 26-29, 2000, Las Vegas, USA.
- R.Jain, R.M.K. Sinha, A.Jain, Translation between English and Indian Languages, Journal of Computer Science and Informatics, March 1997, pp 19 -25.
- R.M.K. Sinha, Machine Translation, Key-Note Plenary Address at the International Multi-Conference on Systematics, Cybernetics and Informatics (SCI'97) at Caracas, Venezuela, July 7-12, 1997.
- Renu Jain and R.M.K. Sinha, 'Machine Translation using Examples for Similar and Dissimilar Languages', International Conference on Information Systems Analysis and Synthesis (ISAS'96), Orlando, 1996.
- R.M.K. Sinha, 'R & D on Machine Aided Translation at IIT Kanpur: ANGLABHARTI and ANUBHARTI Approaches', Invited paper at Convention of Computer Society of India, (CSI'96), Bangalore, 1996.
- R.M.K. Sinha, 'Strategies for Machine Translation for Application in Research, Education and Science Popularization' Invited paper at INSA National Expository Workshop on Information and Communication Technology (INSA NEW-ICTE'96), New Delhi, 1996.
- R.M.K. Sinha and Ajai Jain, 'Relevance and Strategies of Machine Translation in Global Environment and an Integrated Approach to MT in Indian Context' Theme paper at Symposium for Machine Aids for Translation and Communication (SMATAC96), New Delhi 1996.
- Renu Jain, R.M.K. Sinha and others, 'Some Experiences in Development of ANGLABHARTI and ANUBHARTI Systems', Symposium for Machine Aids for Translation and Communication (SMATAC96), New Delhi 1996.
- Renu Jain and R.M.K. Sinha, 'On Multi-lingual Dictionary Design', Symposium for Machine Aids for Translation and Communication (SMATAC96), New Delhi 1996.
- R.M.K. Sinha and others, 'ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi', 1995 IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada, 1995, pp 1609-1614.
- Renu Jain, R.M.K. Sinha and A. Jain, 'Role of Examples in Machine Translation' 1995 IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada, 1995, pp 1615-1620.

- R.M.K. Sinha, R. Srivastava and A. Agrawal, 'Designing Hindi Text Generator for Machine Translation' SNLP'95 - Symposium on Natural Language Processing, Bangkok, Thailand, 1995, pp 286-296.
- Renu Jain, R.M.K. Sinha, A. Jain and R. Srivastava, 'HFSM: A Finite State Machine for Analyzing Hindi Sentences' SNLP'95 - Symposium on Natural Language Processing, Bangkok, Thailand, 1995, pp 317-324.
- Renu Jain, R.M.K. Sinha and A. Jain, 'A Pattern Directed Hybrid Approach to Machine Translation through Examples' SNLP'95 - Symposium on Natural Language Processing, Bangkok, Thailand, 1995, pp 325-335.
- R.M.K. Sinha, 'Machine Translation: The Indian Context', Invited paper at the International Conference on Applications of Information Technology in South Asian Languages, AKSHARA'94, New Delhi 1994, pp 275-284.
- R.M.K. Sinha, 'Correcting ill-formed Hindi sentences in machine translated output' Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS'93), Fukuoka, Japan, 1993, pp 109-119.
- R.M.K. Sinha, 'A Sanskrit based Word-expert model for machine translation among Indian languages', Proc. of workshop on Computer Processing of Asian Languages', Asian Institute of Technology, Bangkok, Thailand, Sept.26-28, 1989, pp. 82-91.
- R.M.K.Sinha, 'CALP: Some Perspectives' Symposium on Computer Aided language processing, New Delhi 1987.
- R.M.K. Sinha,'Computer processing of Indian languages and scripts - potentialities and problems', Jour. of Inst. Electron. & Telecom. Engrs., vol.30,no.6, 1984,pp. 133-49.
- A.K. Bansal and R.M.K. Sinha, 'Some aspects of pronoun disambiguation using real world knowledge', Comp. Soc. of India 1984.
- R.M.K. Sinha and G.C. Pathak, ' A heuristic based question answering system in natural Hindi', IEEE-SMC International conference, Delhi-Bombay, Dec.30,1983-Jan.7,1984, pp 1009-13.
- R.M.K. Sinha. 'Computers for Indian languages', Annual convention of Computer Society of India (invited paper), 1982, pp 163-174.
- R.M.K. Sinha, 'Computer processing of Indian languages', Fourth International Conference on Computer in Humanities', Hanover, NH (USA), Aug. 19-22, 1979.
- R.M.K.Sinha,, 'Some thoughts on computer processing of natural Hindi', Annual convention of Computer Society of India, 1978,pp 151-165.
- R.M.K. Sinha, K. Sivaraman, Aditi Agrawal, T. Suresh and C. Sanyal, 'On logical design of multi-lingual lexicon for machine translation', Technical Report TRCS-93-174, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- R.M.K. Sinha and K. Sivaraman, 'Ambiguity resolution in ANGLA-BHARTI', Technical Report TRCS-93-175, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- K. Sivaraman and R.M.K. Sinha, 'On Tamil text generator', Technical Report TRCS-93-176, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- R.M.K. Sinha, Aditi Agrawal and C. Sanyal, 'Morphological Analyzer', Technical Report TRCS-93-177, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- Aditi Agrawal and R.M.K. Sinha, 'On Hindi text generator', Technical Report TRCS-93-178, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- T. Suresh and R.M.K. Sinha, 'On Telugu text generator', Technical Report TRCS-93-179, Department of Computer Science and Engineering, IIT Kanpur, 1993.
- T. Suresh and R.M.K. Sinha, 'On Man-machine interface in ANGLA-BHARTI', Technical Report TRCS-93-180, Department of Computer Science and Engineering, IIT Kanpur, 1993.

Ph.D.Thesis supervision topics:

- Renu Jain, 'HEBMT: A Hybrid Example-Based Approach for Machine Translation (Design and Implementation for Hindi to English)'.

Speech to Speech Translation

The speech to speech (S2S) translation requires a tight coupling of the automatic speech recognition (ASR) module, MT module, and the target language text to speech (TTS) module. A mere interfacing of ASR, MT and TTS modules does not yield an acceptable S2S translation. S2S requires an integration of these modules such that the hypotheses are cross verified and appropriate parameters get generated. In our environment, it has to cater to bi-lingual (Hindi mixed with English) speech with commonly encountered Indian accent variations. The MT also needs be a chunk translator with multiple translation engines. Our investigations are directed to domain specific applications in Indian environment.

Some Relevant Publications

- R.M.K. Sinha, Towards Speech to Speech Translation, Key-note presentation at Symposium on Translation Support Systems (STRANS2002), March 15-17, 2002, Kanpur, India.

Lexical Knowledge-Base Development

Lexical knowledge base is the fuel to the translation engine. It contains various details for each word in the source language, like their syntactic categories, possible senses, keys to disambiguate their senses, corresponding words in target languages, ontology and word-net information/linkages. We are also working towards development of Indian language wordnet named ShabdKalpTaru in association with Dr. Om Vikas and Dr. Pushpak Bhhattacharya.

Some Relevant Publications

- Renu Jain and R.M.K. Sinha, 'On Multi-lingual Dictionary Design', Symposium for Machine Aids for Translation and Communication (SMATAC96), New Delhi 1996.
- R.M.K. Sinha, K. Sivaraman, Aditi Agrawal, T. Suresh and C. Sanyal, 'On logical design of multi-lingual lexicon for machine translation', Technical Report TRCS-93-174, Department of Computer Science and Engineering, IIT Kanpur, 1993.

Optical Character Recognition

The work on Devanagari OCR started in early seventies. Devanagari script is a logical composition of symbols in two dimensions as opposed to mere juxtaposition of symbols in Roman. A methodology for segmentation of words into composite characters and decomposition into constituent symbols were developed. A pattern description language called PLANG was developed and used in syntactic recognition of Devanagari symbols. A script composition grammar and confusion matrix obtained through training were used to recompose the script from the recognized symbols. This was part of a Ph.D. thesis work in 1973.

Subsequently, use of higher level knowledge layers interacting with each other, in the form of word level dictionary, language model and confusion matrices obtained through training, primarily formed the basis for disambiguation, word hypothesis generation and verification, and for tackling the problem of character fusions and fragmentation. This technique was used both for English OCR and for Devanagari.

Further work on Devanagari OCR was carried out with TDIL, Govt. of India, sponsored project named, DEVDRISHTI, on Recognition of Handprinted Devanagari script. The investigations were carried on in developing new features and in integrating decision making taking into account large variations in shape. Further, an automated strategy for training for construction of prototypes and confusion matrices, from true ISCII files was developed. This had to be very much distinct from their Roman counterpart due to script composition being involved in case of Devanagari script. This work was further expanded incorporating blackboard model for knowledge integration in Ph.D. thesis of Veena Bansal titled "Integrating Knowledge Sources in Devanagari Text Recognition"

Some work has also been carried out on On-line character recognition for Roman using handwriting modeling. Investigations on on-line isolated Devaganagi characters have also been carried out and further investigations are in progress on the subject.

Some Relevant Publications

- Veena Bansal and R.M.K. Sinha, Partitioning and Searching Dictionary for Correction of Optically Read Devanagari Character Strings, Int. Jour. On

Document Analysis and Recognition, Vol. 4, 2002 pp 269-280 (Presented at 5th International Conference on Document Analysis and Recognition 1999, Bangalore, India.)

- Veena Bansal and R.M.K. Sinha, Segmentation of touching and fused Devanagari characters, Pattern Recognition, Vol. 35, 2002, pp 875-893.
- Veena Bansal and RMK Sinha, A Devanagari OCR and A Brief Overview of OCR Research for Indian Scripts, Proc. Symposium on Translation Support Systems (STRANS2001), February 15-17, 2001, Kanpur, India.
- Veena Bansal and R.M.K. Sinha, Integrating Knowledge Sources in Devanagari Text Recognition System, IEEE Transaction on Systems, man and Cybernetics , Vol. 30, 4, 2000.
- Scott D. Connell, R.M.K. Sinha and Anil K. Jain, Recognition of Unconstrained On-Line Devanagari Characters, International Conference on Pattern Recognition, (ICPR2000), Sept 3-8, 2000, Barcelona, Spain.
- Veena Bansal and R.M.K. Sinha, On how to Describe Shapes of Devanagari Characters and Use Them for Recognition, 5th International Conference on Document Analysis and Recognition(ICDAR '99), 1999, Bangalore, India.
- Veena Bansal and R.M.K. Sinha, Segmentation of Touching characters in Devanagari, Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP'98), pp. 371 - 376, December 21-23, New Delhi, 1998.
- Veena Bansal and R.M.K. Sinha, 'On Integrating Diverse Knowledge Sources in Optical Reading of Devanagari Script' International Conference on Information Systems Analysis and Synthesis (ISAS'96), Orlando, 1996.
- Veena Bansal and R.M.K. Sinha, 'Designing a Front End OCR System for Machine Translation - A Case Study for Devanagari', Symposium for Machine Aids for Translation and Communication (SMATAC96), New Delhi 1996.
- R.M.K. Sinha and V. Bansal, 'On Devanagari Document Processing', 1995 IEEE International Conference on Systems, Man and Cybernetics, Vancouver, Canada, 1995, pp 1621-1626.
- R.M.K. Sinha, B. Prasada, G. Houle and M. Sabourin, 'Hybrid Contextual Text Recognition with String Matching', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No. 9, September 1993, pp. 915-925.
- Bruno Simard, Birendra Prasada and R.M.K. Sinha, 'On-line character recognition using hand-writing modeling', Pattern Recognition, Vol. 26, No. 7, 1993, pp. 993-1007.
- R.M.K. Sinha, 'On using syntactic constraints in text recognition' Proc. Second International Conference on Document Analysis and Recognition, Tsukuba Science City, Japan, 1993, pp 858-861.
- R.M.K. Sinha, 'On partitioning dictionary for visual text recognition', Pattern Recognition, Vol. 23, No.5, 1990, pp 497-500.
- Subodh Harmalkar and R.M.K. Sinha, 'Integrating word level knowledge in text recognition', 10th International Conference on Pattern Recognition, Atlantic City, NJ, June 17-21, 1990.
- R.M.K. Sinha, and H.C. Karnick, 'PLANG based specification of patterns with variations for pictorial data bases', Computer Vision, Graphics, and Image Processing, Vol 43, 1988, pp. 98-11.
- R.M.K. Sinha and Birendra Prasada, 'Visual text recognition through contextual processing', Pattern Recognition, Vol. 21, No.5, 1988, pp 463-479.
- R.M.K. Sinha, 'Some characteristic curves for dictionary organization with digital search' IEEE Trans. on Systems, man and Cybernetics, 1987, Vol SMC-17, No.3,1987,pp 520-527.
- R.M.K. Sinha, ' Rule based contextual post-processing for Devanagari text recognition' Pattern Recognition, 1987, Vol 20,No.5, 1987, pp.475-485.
- R.M.K. Sinha, 'Role of Context in Devanagari Script Recognition', Jour. of Inst. Electron & Telecom. Engrs., Vol 33, No.3,1987, pp 86-91.
- R.M.K. Sinha, 'A width-independent algorithm for character skeleton estimation', Computer Vision, Graphics, and image Processing, 1987, vol 40,1987,pp 388-397

- R.M.K. Sinha, Comments on 'Fast thinning algorithm for binary images', Image and Vision Computing, vol.4, 1986, pp 57-58.
- R.M.K. Sinha, 'PLANG - a picture language schema for a class of pictures', Pattern Recognition, vol. 16, 1983, pp 373-383.
- R.M.K. Sinha, 'A knowledge based script reader', Seventh International Conference on Pattern Recognition Montreal. 1984, pp 763-765.
- R.M.K. Sinha, 'Primitive recognition and skeletonization via labeling', International Conference on Systems, Man and Cybernetics, Halifax, Canada, 1984, pp 272-279.
- R.M.K. Sinha, 'A parallel architecture for recognition of pictorial patterns' IEEE International Conf. on Computers, Systems and Signal processing, Bangalore, 1984, pp 1523-1526.
- H.Karnick and R.M.K.Sinha,'A representational framework for recognizing patterns with variations', IEEE International Conf. on Computers, Systems and Signal processing, Bangalore, 1984,pp 19-22.
- S.S.Marwah, S.K.Mullick and R.M.K.Sinha, 'Recognition of Devanagari characters using hierarchical binary decision tree classifier', International Conference on Systems, man and Cybernetics, Halifax, Canada,1984, pp 414-420.
- R.M.K. Sinha, 'Methodology for computer recognition of Devanagari script', IEEE-SMC International conference, Delhi-Bombay, Dec.30,1983 - Jan.7,1984, pp 1220-1224.
- R.M.K. Sinha and H.N. Mahabala, 'Machine recognition of Devanagari script', IEEE Trans. on Systems, Man and Cybernetics, 1979, pp 435-441.
- R.M.K. Sinha and H.N. Mahabala, 'Towards design of a natural picture description language', IEEE Conf. on Pattern Recognition and Image Processing, Chicago, Ill., USA, 1978, pp 416-420.
- R.M.K.Sinha, 'Primitive recognition via labeling schemata', Annual convention of Computer Society of India,1975.
- R.M.K. Sinha and H.N. Mahabala, 'On design of a syntactic pattern analysis system', Annual convention of Computer Society of India, 1975.
- R.M.K. Sinha et.al., 'DEVADRISHTI: A Devanagari text reader - version I', Technical Report TRCS-93-181, Department of Computer Science and Engineering, IIT Kanpur, 1993.

Ph.D. Thesis supervised :

- Harish C. Karnick, 'On Learning Recognizing Patterns with Natural Variations'.
- Veena Bansal, 'Role of Knowledge in Document Recognition- A case study for Devanagari Script'.

Development of ISCII and INSCRIPT Keyboarding

In order to be able to process data in Indian languages, the first task was to develop mechanism for inputting the Indian script and then represent it internally for editing, storage, retrieval and further processing. Romanization of Indian script was not a desirable approach as this may not be error-free. Indian scripts are phonetic in nature (i.e. you write the same way as you speak) and is a logical composition of constituent symbols in two dimensions. Our scheme for keyboarding and internal representation exploited the phonetic nature of our scripts. We proposed the phonetic order of the symbols to be sequence of symbols for inputting rather than the visual order as followed in the mechanical typewriters. In fact the phonetic order is the way in which children learn writing the script. Since this phonetic ordering was applicable to all Indian scripts, our keyboarding scheme became universally applicable to all Indian scripts. Indian scripts typically have 12-15 vowels, 35-40 consonants and a few diacritical marks. Besides this for each vowel, there is a corresponding modifier symbol and for each consonant, there is a corresponding pure consonant form (called half-letter). This makes the total set of symbols to be larger than what a normal keyboard could accommodate. Our keyboarding layout and the keyboarding scheme utilized the phonetic groupings and derivational property to limit the number of physical keys and achieve a logical layout based on frequency analysis of the symbols of the script. This became possible as the task of inputting was kept distinct from the task of rendering for display. This gave birth to the INSCRIPT keyboard and keyboarding scheme.

Similarly, for the internal representation, an 8-bit Indian Standard Script Code for Information Interchange (ISSCII-8) was designed utilizing phonetic properties of the script. ISSCII-8 is an extension of ASCII which has been designed during early 1980's and it caters to the entire set of Indian scripts in an uniform way. Editing operation became very much like that for English. Also, transliteration among Indian scripts became simply switching the script rendering device as the ISSSCII code for all Indian scripts texts remained the same. ISSCII-8 has undergone further modifications and a modified version has been accepted by Bureau of Indian Standards as ISCII (8-bit Indian Script Code for Information Interchange) code in 1991. ISCII forms the basis for UNICODE code assignments for Indian scripts.

The paper on "Computer processing of Indian languages and scripts - Potentialities and Problems", Jour. of Inst. Electron. & Telecom. Engrs., vol.30,no.6, 1984, carries a detailed discussion on various aspects of coding.

Some Relevant Publications

- R.M.K. Sinha, 'Standardizing Linguistic Information - An Overview' Proceedings of Second Regional Workshop on Computer Processing of Asian Languages, Tata McGraw-Hill, New Delhi, 1992, pp 272-290.
- R.M.K. Sinha, 'Non-Latin Information Systems: Some Basic Issues', in Information Processing, 1986, H. Kugler (Ed.), Elsevier Science Publishers, 1986. Conference Proceedings.
- R.M.K. Sinha, 'Computer processing of Indian languages and scripts - potentialities and problems', Jour. of Inst. Electron. & Telecom. Engrs., vol.30,no.6, 1984,pp. 133-49.
- R.M.K. Sinha and A. Raman, 'A modular Indian language data terminal', Computer Graphics, Vol.14, 1980, pp. 39-72.
- M.P. Sastri, A.Raman and R.M.K. Sinha, 'An universal I/O device for Indian scripts', Annual convention of Computer Society of India, 1978, pp 151-165.
- R.M.K. Sinha, 'Machine oriented Devanagari script (MODS) from information theoretic viewpoint', Symposium on Linguistic Implication of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- P.V.H.M.L. Narasimham and R.M.K. Sinha, 'Phonetically coded keyboarding in Indian languages', Symposium on Linguistic Implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- R.M.K. Sinha and H.V. Sahasrabuddhe, 'Hyphenation in Indian scripts for computer aided printing', Symposium on Linguistic Implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- A.K. Pathak, A. Raman and R.M.K. Sinha, 'A modular Indian language I/O terminal', Symposium on Linguistic Implication of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- R.M.K. Sinha, H.V. Sahasrabuddhe and V.K. Vaishnavi, 'Mechanization of Indian scripts', Symposium on Linguistic implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- R.M.K. Sinha, 'Teaching script on a digital computer ', Jour. of Inst. Telecom. Engrs., Nov. 1976, pp 720-22.
- R.M.K. Sinha and H.N. Mahabala, 'MODS - machine oriented Devanagari script' Jour. of Inst. Telecom. Engrs., vol.19. no.3, 1973, pp 623-28.
- R.M.K Sinha (Chief Investigator), 'Integrated Devanagari Computer', Project Report, Dept. of Elect. Engg., I.I.T., Kanpur 1984.
- R.M.K. Sinha. 'Character Code Standardization', a report prepared for UNESCO, Paris, 1992.

Development of Integrated Devanagari Computer (IDC) / Graphics and Indian Script Terminal (GIST) multilingual technology

In 1983, Department of Electronics, Govt. of India, sponsored a project on design and development of 'Integrated Devanagari Computer(IDC)' terminal. In this project we implemented our basic strategies for phonetic keyboarding scheme for Devanagari inputting, used our ISCII code for internal representation of the script, and a script composition module for rendering the script on the display and

other output devices. The IDC terminal was designed in a record time of about 8 months and was demonstrated at the Third World Hindi Convention at Delhi. It was developed using Intel 8086 processor with multitasking firmware. The IDC project was further extended to implement the same technology using the 32-bit 68000 microprocessor and the outcome was named as GIST (Graphics and Indian Script Terminal) technology. Since ISCII was designed to cater to all Indian scripts exploiting commonality and phonetic nature of the scripts, the GIST technology could easily cater to all Indian scripts by merely incorporating specific script composition rendering module. A number of companies bought this technology for manufacturing multilingual computer terminals.

The GIST technology represented a major breakthrough in solving our complex problem of man-machine linguistic interface for Indian languages. This technology incorporated several desirable features. A natural phonetically oriented keyboarding scheme directly converting to internal representation codes (ISSCII-8), a human engineered keyboard layout, a display which dynamically changes as the input progresses, built-in intelligence to disallow illegal compositions such as attaching two vowel modifiers on the same character, automatic transliteration from one Indian script to another, are some of the key attractive features making it user friendly.

More details of IDC/GIST technology can be seen in the Project Report on Integrated Devanagari Computer, Dept. of Elect. Engg., I.I.T., Kanpur 1984.

Some Relevant Publications

- R.M.K. Sinha, 'Standardizing Linguistic Information - An Overview' Proceedings of Second Regional Workshop on Computer Processing of Asian Languages, Tata McGraw-Hill, New Delhi, 1992, pp 272-290.
- R.M.K. Sinha, 'Non-Latin Information Systems: Some Basic Issues', in Information Processing, 1986, H. Kugler (Ed.), Elsevier Science Publishers, 1986. Conference Proceedings.
- R.M.K. Sinha, 'Computer processing of Indian languages and scripts - potentialities and problems', Jour. of Inst. Electron. & Telecom. Engrs., vol.30, no.6, 1984, pp. 133-49.
- R.M.K. Sinha and A. Raman, 'A modular Indian language data terminal', Computer Graphics, Vol.14, 1980, pp. 39-72.
- M.P. Sastri, A. Raman and R.M.K. Sinha, 'An universal I/O device for Indian scripts', Annual convention of Computer Society of India, 1978, pp 151-165.
- R.M.K. Sinha, 'Machine oriented Devanagari script (MODS) from information theoretic viewpoint', Symposium on Linguistic Implication of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- P.V.H.M.L. Narasimham and R.M.K. Sinha, 'Phonetically coded keyboarding in Indian languages', Symposium on Linguistic Implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- A. Raman, P.V.H.M.L. Narasimham and R.M.K.Sinha, 'System modules for business machines, computer terminals and printing in Indian languages',Symposium on Linguistic Implications of Computer based information Systems, Delhi Nov. 10-12, 1978.
- A.K. Pathak, A. Raman and R.M.K. Sinha, 'A modular Indian language I/O terminal', Symposium on Linguistic Implication of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- K.P. Laturkar and R.M.K. Sinha, 'Devanagari script composition from phonetically coded symbol strings', Symposium on Linguistic Implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- M.P. Sastri, A.Raman and R.M.K. Sinha, 'An Indian language script generator for CRT terminals and matrix printers', Symposium on Linguistic implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.

- R.M.K. Sinha, H.V. Sahasrabuddhe and V.K. Vaishnavi, 'Mechanization of Indian scripts', Symposium on Linguistic implications of Computer Based Information Systems, Delhi, Nov. 10-12, 1978.
- R.M.K. Sinha, 'Teaching script on a digital computer ', Jour. of Inst. Telecom. Engrs., Nov. 1976, pp 720-22.
- R.M.K. Sinha and H.N. Mahabala, 'MODS - machine oriented Devanagari script' Jour. of Inst. Telecom. Engrs., vol.19. no.3, 1973, pp 623-28.
- R.M.K Sinha (Chief Investigator), 'Integrated Devanagari Computer', Project Report, Dept. of Elect. Engg., I.I.T., Kanpur 1984.
- R.M.K. Sinha. 'Character Code Standardization', a report prepared for UNESCO, Paris, 1992.

Transliteration

Transliteration among Indian scripts is easily achieved using ISCII (Indian Script Code for Information Interchange). ISCII has been designed using the phonetic property of Indian scripts and caters to the superset of all Indian scripts. By attaching an appropriate script rendering mechanism to ISCII, transliteration from one Indian script to another is achieved in a natural way.

However, transliteration from Indian script requires use of heuristics to convert the non-phonetic script to its probable intended spoken form before it could be transliterated. Similarly, transliteration from an Indian script to Roman requires using a standardized mapping table to easily readable. In our work on transliteration, we have suggested heuristics and tables. Several other workers have come up with their own suggestions. Recently, TDIL has come up with a standardization of this table called INSROT which uses only lower case letters to facilitate standard search.

Some Relevant Publications

- R.M.K. Sinha, 'Computer processing of Indian languages and scripts - potentialities and problems', Jour. of Inst. Electron. & Telecom. Engrs., vol.30, no.6, 1984, pp. 133-49.
- R.M.K. Sinha and B. Srinivasan, 'Machine transliteration from Roman to Devanagari and Devanagari to Roman', Jour. of Inst. Electron. & Telecom. Engrs., vol.30, no.6, 1984, pp 243-45.

Spell-Checker design

For Indian scripts, there is a very loose concept of a spelling. Writing in Indian scripts is a direct mapping of the inherent phonetics and you write as you speak. There are geographical variations in the spoken form and so the spellings vary. Our approach to design of a spell checker is to develop an user error model for each class of user where the source of error may the due to incorrect phonetics, inaccurate inputting or other influences. The spell-checker uses this error-model in making suggestions for the error.

Some Relevant Publications

- R.M.K. Sinha, 'Computer processing of Indian languages and scripts - potentialities and problems', Jour. of Inst. Electron. & Telecom. Engrs., vol.30, no.6, 1984, pp. 133-49.
- R.M.K. Sinha and K.S. Singh, 'A program for correction of single errors in Hindi words' , Jour. of Inst. Electron & Telecom. Engrs., vol.30, no.6, 1984, pp 249-51.

Courtesy : Prof. R.M.K. Sinha
Department of CSE
Indian Institute of Technology,
Kanpur 208 016 (UP)
Tel. : 0512-2597174 (O), 0512-2590260
0512-2590007
E-mail : rmk@iitk.ac.in