SRELS Journal of Information Management
Vol. 48, No. 4, August 2011, Paper AL. p.367-376

INFORMATION SYSTEMS IN MALAYALAM

R. Raman Nair [1]

Knowledge accumulated in a society is recorded in its language is the base for development, is a capital resource to be equitably distributed in democracy. ICT can enable this and empower and emancipate the backward people if their vital information needs can be met by it through their language and script. But our Information Infrastructures has disregarded 'information' part; especially content in local language. A cause is identified as non availability of Language Technology. But now Unicode and related research enables reviving original local scripts for use in ICT. Discusses the essential Information systems to be developed in local scripts like those on indigenous knowledge. Opines that development of local language technology with provision for use of OCR, spell check etc for original script is a priority area in ICT development.

KEYWORDS/DESCRIPTORS: Information and communication technology, Information systems, Local languages, Malayalam

1 INTRODUCTION

Information and Communication Technology (ICT) can contribute immensely to the empowerment and emancipation of people living in rural areas in economically backward conditions. This essentially requires developing digitized information in vernacular scripts like Malayalam, Kannada, and Tamil. People at large require vital information on hygiene, farming, subsidies, supports, educational facilities etc. Optical cables, computers and communications stacks should be capable of carrying this information to people in their language. We have never given due importance in developing digitized contents and applications in the language of our people.

2 INFORMATION INFRASTRUCTURES WITHOUT INFORMATION

For almost a decade we have been engaging in developing Information Infrastructures state wide and campus wide. We have invested millions for computers, cables, LANs, WANs and V-Sats. We claim that these are going to make information accessible to the un-reached. But from an Internet Kiosk at Panchayat office can a farmer or villager get information relevant to his needs? Look at the vastness of Information our people may seek: A particular section of Panchayat Act, a Government Order relevant to him, higher education options available for one's daughter completing +2, a Gazetteer of the region, the preventive measures for Denki, pesticides to control Mandari or Manjalippu disease of the coconut palms, the support available from government to safe guard life and properties if Mullapperiyar Dam falls, or some aspects of sustainable farming and animal rearing methods relevant to his/her locality. Is the villager today able to get these and like information in his Mother tongue, the only language he knows?

[1] *University Librarian*, Mahatma Gandhi University, Kottayam 686 560, INDIA. E-mail: ramannair_r@yahoo.com

Even in English it is unlikely to get. That is the current situation in Kerala. On the other hand a villager may get any government document of USA, the educational facilities in UK, any book from Library of Congress or market information of grains from a Japanese village in Japanese language within seconds at the Kiosk in the Panchayat. Is this the way a Kerala villager should be equipped with modern Information Technology to get the information vital for the sustainability and the development?

3 LACK OF APPROPRIATE LANGUAGE TECHNOLOGY

One of the reasons for this failure is that local specific information is naturally in Malayalam and today it can not be digitally disseminated using Malayalam. We haven't so far tried to make Malayalam script viable to digital information systems of the State and its institutions. We have been totally disregarding the development of 'information part' of our Information Systems while building up huge infrastructures. We hope computer and communication stacks automatically create Information Systems. We have never cared to develop applications for managing our content in our mother tongue. Even after two plan periods that initiated ICT development projects, even after the physical systems we acquired and installed became obsolete three or four times and got changed, even after millions of rupees was wasted for cables and computer stacks, we have not developed information content in the language of the people that should be the life blood to flow through these physical entities. In State Information Infrastructure major part of the content ought to be developed for catering to the requirement of our people. Without doubt this should be in Malayalam, in its own script.

Now the constraints in developing information systems in Malayalam stands removed with the development of the DBMS for search and retrieval in the local scripts, thanks to the attempts of Sri K H Hussain and others. Such ventures in which linguists, documentalists and ICT experts come together to develop Malayalam script based applications should be shaped for full functionality by the needs of people who have to use it. People's involvement and feedback becomes important components for these information systems.

The earliest attempts for building up of information systems in Malayalam or in other Indian languages were done successfully by librarians using traditional methods. The notable one in India was that of Sri. K.M Govi on documenting our literary heritage which has no parallel in any Indian language. His *Malayala Granthasoochi* published by Sahithya Akademi in print which runs into almost ten volumes and more than 10000 pages is a complete information system of the books ever published in Malayalam. It is the result of devotion and untiring efforts of a saintly person through almost forty years. In no other Indian language such an initiative was ever undertaken or completed. Even institutions spending crores of rupees and many years, could never successfully finish such a project. The State organization that has utilized this work has never acknowledged or rewarded him for the work, or supported the efforts in a befitting way. They have also not effectively utilized his work for the sake of the society.

In ICT based systems it is the library professionals who are first to seriously encounter the difficulties in using scripts of languages other than English for cataloging the documents. In computerised cataloguing systems transliteration method was applied by documentalist for decades. As we know, transliterated words will never carry the essence of words, sound and meaning as in the original language. If a standard scheme is not followed, a transliterated word can be written, read and interpreted in dozen ways. Development of Information Systems using regional script was found to be the only solution that can truly contain and reflect the linguistic and cultural characteristics of the documents and its authors.

4 DOS AND ASCII AGE

The earliest attempts on this line occurred in late eighties. C-DAC under government of India designed Gist Cards which are extra hardware components for Indian language processing. Some public libraries still use Malayalam databases capable of search and retrieval using Malayalam script. Thrissur Public Library is one such library. DC Books Kottayam had developed a Store inventory control using Gist and dBase2, for processing all their published Malayalam titles using Malayalam script.

5 MS WINDOWS AND ASCII AGE

In 2001, St. Mary's College, Thrissur used MISIS Version 1 based on New Lipi to develop their library catalogue which was the first attempt in Malayalam to localize CDS\ISIS software developed by UNESCO.

In 2001 Kerala Sahitya Academy, Thrissur started computerization of Malayalam Grantha Soochi prepared by Sri. K.M. Govi which contained the most authentic data on 50,000 books which consisted of more than 90% of the publications in Malayalam from the beginning of printing to the present. Akademi built up a database of nearly 50,000 records programmed in PowerBuilder. The program made use of new Malayalam Lipi of ISM-Gist fonts like Revathi and Kartika. Akademi is using this DBMS internally in the library as a digital catalogue, but could not convert it to a product to distribute widely among libraries and scholars.

In addition to Granthasooji, Kerala Sahithya Akademi developed an information system in Malayalam named Malayala Paramparyam. It consists of the biographical details of all the Malayalam writers in Malayalam Script, their pictures, archival film strips, audio on their talks or poetry recitations and many other related materials. It has been published in CD form also. This is one of the very interesting Malayalam Information systems in Multimedia that will be of interest to the whole population and can be installed in village libraries also. It needs Malayalam search Interface.

In 2003, Govt. Brennen College, Tellycherry published the complete catalogue of its Malayalam books in MISIS version 2 using Malayalam Old Script. It consisted of 21000 records. It was a CD publication meant for distribution. Technically speaking, Malayalam Granthasoochi

(2004) of Brennen College became the first published digital catalogue in original Malayalam script that can be searched and processed.

In 2004, Vimala College, Thrissur also developed Malayalam books' database using MISIS 2 for internal use in the library.

6 UNICODE AGE 2005

It is the research and development conducted by a small NGO named *Rachana Aksharavedi* that materialized major Information system development programmes in Malayalam.

State Central Library, Trivandrum used the program 'Nitya' for preparing an in-depth catalogue of 800 old rare books in Malayalam and English for digital archiving the collection. It is the first Unicode based Multilingual DBMS developed in Kerala. It used Original/Old Malayalam script applying Rachana font.

In 2005 DC Books Kottayam developed a cataloguing system for their collection of Malayalam books in Linux using Rachana font.

Centre for South Indian Studies (CSIS, Thiruvananthapuram) has initiated with technical support from CIRD (Centre for Informatics Research and Development, Thiruvananthapuram) a digital library project entitled 'Kerala Reference Library' consisting of selected authoritative rare and antique books on Kerala's history and culture in Malayalam and English. The project when completed will make an authentic and reliable reference collection on Kerala affordable to more than 6000 rural libraries of Kerala as well as to schools and colleges.

Centre for Development of Advanced Computing has under it a Resource Centre for Indian Language Technology Solutions of which a major section deals Malayalam. They have developed various information systems in Malayalam. They have a Malayalam Data base containing Malayalam version of the Constitution of India, forms and formats commonly used in Government, SSLC Question papers and Answers in Malayalam, a knowledge base in Malayalam for Rubber cultivators and a tourist aid package called "Explore Kerala". They have developed Software products (such as Sandesam – e-mail Server, Anweshanam-Search Engine and e-com application). The following are the Information Systems and products developed by C-DAC.

- *Nayana* - Malayalam OCR System;
- *Aksharamaala* - Malayalam Font Package and Script Manager;
- *Ezhuthachan* - The Malayalam Tutor;
- *Nerpadam* - Malayalam Spell Checker;
- *Subhashini* - Malayalam TTS System;
- *Aachaaryan* - The English Tutor;
- *Prakes* - Prakruti Estimation Software;
- *Sandesam* - Malayalam Web-based Mail Server;

- *Anweshanam* - Malayalam Web-based Search Engine; and
- *Kalakeralam* - Knowledge base of Traditional Art Forms & Culture of Kerala.

The computational linguistics team at the Centre for Development of Imaging Technology (C-DIT) Trivandrum is conducting research and development for incorporating or using Malayalam Language Technology in various programmes of interest to the people. It consolidates and coordinates the efforts for integrating Malayalam with ICT. Standardization of keyboard, production of Unicode supported Malayalam fonts, evolving a corpora of Malayalam are the major projects. They have developed the following systems:

- *Kaveri* - is an Open Office software;
- *Paithrukam* – A digital archiving software;
- *Webdarshini* - Malayalam interface browser;
- *Akshaya* - IT Literacy Resource CD;
- *Aksharamala* - Programme incorporating 25 Malayalam fonts; and
- *Kairali GNU/Linux* - Malayalam unicode fonts, paint/ DTP tools.

7 WHY OUR SCRIPT?

From early nineties onward attempts were made in libraries in Kerala to build up electronic catalogue of Malayalam books. Transliteration was the popular method applied to surmount the language barrier. Unfortunately there was no standardized transliteration scheme to follow. Entering data using one scheme and searching using another scheme result failure in retrieval. Past fifteen years of effort for transliterated data in computerized libraries in Kerala (University, College and some Public) is becoming useless due to its un-scalability to Unicode standards and will have to enter once again when Unicode based multilingual library management programs become available in the market in the near future.

Considering Malayalam language technology, there are confusions related to old and new scripts. Once Unicode takes a final decision on CHILLUs and Malayalam Code page is fixed for ever, the difference between Old and New Lipi will disappear for ever and rendering of Malayalam words using Malayalam script will become standardized which could not be dreamed in transliteration and ASCII era. Then only search and retrieval using Malayalam script will become as natural as English.

8 ESSENTIAL INFORMATION SYSTEMS IN MALAYALAM

There are many areas to be explored applying DBMS/ Information Systems using Malayalam script.

9 LIBRARY DATABASES

Our libraries store books in many languages (Malayalam, English, Hindi, Tamil, Arabic, etc). Only Unicode based Multilingual DBMS can solve our bibliographic problems encountered in digital applications. Once prominent academic and public libraries develop Unicode based catalogs, we can easily build up a cumulative/integrated / networked catalogue for public and scholarly use. One notable attempt is Malayalam Grantha Soochi CD 2007, which is being prepared by Centre for South Indian Studies (CSIS). Consulting editor of the project is Sri. KM Govi. It will be ready for distribution in 2007. It is a complete catalogue of Malayalam books published so far, available in different rare and antique collections with public and private organizations and individuals India and abroad. It covers a large number of banned books of erstwhile Travancore and Cochin kingdoms, which were available only in collections outside the State. It is a Unicode based Multilingual DBMS. It uses original / old Malayalam script applying Rachana font.

10 RIGHT TO INFORMATION

Government has to ensure online accessibility of basic information resources and services in the villages also if people's Right to Information is to be ensured. Without that in Malayalam script accessible at the village people can not understand the meaning of government actions and acquire and interpret administrative information that affect their life and properties under RTI Act.

11 LEXICOLOGICAL INFORMATION SYSTEM

It is the basic information system of all words in a language. Digital Dictionary of Malayalam will essentially be Multilingual since the etymology of a word can be described in the original scripts along with Malayalam. Already some attempts for preparing dictionaries have been done by Malayala Manorama Publications etc. DC Books have also brought out a searchable encyclopedia in Malayalam named Akhila Vijnana Kosam. But the script and related technology used in them are not scaleable.

12 INDIGENOUS KNOWLEDGE

Malayalam script is very important for building up information systems of our indigenous knowledge. Examples of some areas are *Nattarivu, Kattarivu, Kadalrivu*, etc. These are to be recorded from oral narrations or old manuscripts. For automating the process and conserving this knowledge content development should adhere to the old script.

13 INDIGENOUS MEDICINES

A vast ocean of information on indigenous medicines is under threat of extinction. Examples are *Otttamooli, Visha chikitsa* and remedies prescribed by local schools of Ayurveda / Unani / Siddha / *Marma sastra*. Malayalam language technology, especially that of Old Script becomes essential to materialise information systems of Indigenous Medicines. Conserving this knowledge; is important for our State as well as for the whole humanity since the modern science is becoming more and more tangible to the effectiveness and sustainability of our old health care system

14 AGRICULTURAL / FARMER INFORMATION SYSTEMS

More than 70% of our people depend on agriculture, animal rearing, fisheries and related activities for their livelihood. Their efficiency, well being and contributions are significant for national development. Agriculture and related information systems should address the farmers who can not understand any language other than their mother tongue. The agricultural information systems and virtual universities proposed by IIITK and other high tech groups have not yet started providing information to the real Malayali farmers. These agencies should apply an effective Malayalam language technology to really make profit the investments made in information system infrastructures.

15 FOLK ARTS

Developing information systems of our folk songs and arts having several thousand items with hundreds of local variations is important in preserving our cultural heritage. Take for example the ritualistic art of Theyyam. There are more than 800 Teyyams. Some rarely performed Teyyams like Mappilla Teyyam of Kasaragod are facing the threat of extinction. Audio, video and image technology along with the language technology not only to record but also to catalog is very important in applying ICT for conserving such art forms.

16 REVENUE RECORDS

For long developing Land Records Information System is under consideration. Attempts made so far have not come out successfully. Ninety nine percent of the documents to be digitized are in old Malayalam script. There indexing needs to be done in the local language so that searching and retrieval of the documents should serve the real purpose for the public.

17 REASONS FOR FAILURE OF INFORMATION SYSTEMS IN MALAYALAM

Malayalam is a language that evolved through thousands of years. It got its final shape four centuries back at the time of Tunjath Ezhuttachan. In 1821 Benjamin Bailey designed the first set of movable types of Malayalam. He preserved unique characteristics of Malayalam lipi and conjuncts in designing the first font. He strictly followed the pan-Indian style in designing the characters not forgetting that all the lipis in Indian languages are descendants of the ancient *Brahmi* lipi. The types and characters he (and late Herman Gundert) designed made our lipi more scientific and systematized, and that continued without any major modification in printing and writing for one and half centuries.

In the 1970s some scholars were engaged in devising a new script system to suite English typewriters and thus came in to existence the so-called 'New Lipi'. In the late eighties when the computers started to occupy our desktops the official language scholars still thought that computer was a new kind of typewriter and made no attempt to reconsider the original script for digital applications. They never cared to explore the functionality and aesthetics of the original script in the new environment. This has resulted in adapting the deformed alphabet called 'New' Lipi' for computers leading to total chaos and confusion. In 1999 Rachana Akshara Vedi led by Mr. R. Chitrajakumar launched the 'Rachana' linguistic campaign for the Malayalam Original script. The Rachana word processor proved the viability and superiority of original scripts in digital applications beyond any doubt.

18 ADVANCED RESEARCH FOR MALAYALAM INFORMATION SYSTEMS

In addition to the research and development going on under Rachana Malayalam research on the following aspects are going on at C-DAC which will be of great interest to Digital Library developers and archivists who have to deal with Malayalam documents.

Malayalam OCR: Already some experiments have been conducted on Malayalam Optical Character Recognition (OCR) system. The presence of two different scripts (Old Script and New Script) and a large number of characters (including conjuncts) makes the Malayalam OCR a complex system. The experiments are based on the Feature Extraction method of character recognition. The development of the post processor, a spell checker tuned for the OCR system, is in progress. The objective of the post processing is to correct errors in OCR output by using Malayalam grammar rules, etc. those who conducted the test claim that there is an accuracy of 97% at character level for good quality printouts

Web Malayalam: Attempts are also going on to solve the problems in the representation of phonetic symbols of Malayalam language in html page. These symbols are not the ones normally available on the computer keyboard. In many cases it is used to be written by hand and only few symbols were available in the insert mode in the word processors and many times they were not compatible with html packages. These experiments will be helpful in converting the printed documents to digital resources.

Malayalam Spell Checker: Research is also going on for the development of a Malayalam Spell Checker, which is a tool that will check the spelling of the words in a Malayalam text file, validate them and in case of error, list out the right spelling in the form of suggestions. The structure of Malayalam language is such that a large number of words can be derived from a root word and also any number of words can be concatenated with a root word based on their use. Hence a purely dictionary based approach for spell checking is not practical. The experiments have shown that a 'Rule cum Dictionary' based approach is to be followed in the Design of the Spell Checker. The grammatical behavior of the language, the formation of words with multiple suffixes, concatenation of multiple words with suffix/suffixes in between and the preparation of the language module are the aspect covered by related research. Different modules developed for the Spell Checker Engine was tested with 400 pages of data and average error detection rate is claimed to be around ninety percent.

19 ORIGINAL SCRIPT SHOULD SURVIVE

Society is fast becoming digitally networked. The concept of paperless society put forward by Lancaster may take many more years to come true. Digital archiving and digital library development will be major activities of government and organizations in the coming decades. Millions of documents / pages like note files, revenue records, books, journals, etc are in Malayalam script and they need to be digitized and indexed. Back files of government offices will need to be digitally archived. Majority of these documents are in Old Malayalam script. We can not discard the ocean of old knowledge coded in the Original Script. ICT will be forced to use Optical Character Recognition programs for our old / rare documents and manuscripts. ICT will eventually be forced to develop a language technology giving prime focus to our original script. Once Unicode standard is finalized, which will happen immediately, old script can be more effectively used in Malayalam applications. New Lipi, which is only a subset of the original script, may become obsolete. The efficiency, effectiveness and cultural content of the original Script will eventually revive our language from the present chaos created by thoughtless experiments.

20 TOOL SHOULD NOT DECIDE THE MASTER

Though logical, computer is only a tool. It presents unprecedented capability and speed in the hands of linguists, historians, mathematicians, documentalists, etc. to do research, teaching and development. One should not change his language and other systems of learning to fit them according to the needs of a machine, even though the machine happens to be the powerful, non-mechanical and artificially intelligent. Computer technology should be developed to contain all nuances and cultural specificities of our language.

The research and development projects discussed above on language technology clearly establish that our Original script evolved through centuries is more scientific and functional in developing indigenous information systems. The experiments conducted by NGO's like CIRD, CSIS, Rachana Akshara Vedi and professionals like K.H. Hussain are milestones in the history of modern Malayalam Language technology that will devise its future applications in preserving our heritage.

21 REFERENCES

1. Centre for Development of Advanced Computing. *Annual Report of C-DAC, 2003-2004.* Poona, the Author, 2004.
2. Geetha Kumary (V); Sharada (B A). Knowledge representation for web Malayalam speech sounds and their mapping to unicode symbols: a case study. **In** *Language Engineering Conference*, University of Hyderabad, India, Dec. 2002, Paper 5, p3.23
3. Govi (K M). Malayala Granthasooji. Thrissur, Kerala Sahithya Akademy, 1970.

4. Hussain (K H); Vijayakumaran Nair (P); Chitrajakumar (R); Ravindran Asari (K); Raman Nair (R). Creation of digital archives in Indian languages using CDS/ISIS: development of M-ISIS (Malayalam ISIS) and 'Nitya'. *Information Studies*. Vol. 11(1); 2005; p59-68.
5. Jithesh (K); Sulochana (K G); Ravindra Kumar (R). Optical Character Recognition (OCR) System, for Malayalam Language. **In**: *National Workshop on Application of Language Technology in Indian Languages. Hyderabad, March 6-8, 2003*: Paper 5, p1.15
6. Moodithaya (P N). Malayala Granthasoochi 2004 of Government Brennen College, Tellycherry: The first electronic catalogue in Indian languages. **In** *Proceedings Language Technology for Malayalam: Seminar*, pp. 1-5, Tellycherry, Government Brennen College, 2004.
7. Rachana Akshara Vedi. Rachana Manual. Trivandrum, the author, 2000.
8. Santhosh. (T); Varghese (K G); Sulochana (R); Ravindra Kumar. Malayalam Spell Checker. International Conference on Universal Knowledge and Language, RCILTS-Malayalam. Goa, November. 2002. Paper 5, p3.38.