Informatics Studies. ISSN 2320 – 530X. Vol. 3, Issue 4,
Fourth Quarterly Issue. October – December, 2016. P 5-14

Information Systems using Malayalam Script: Problems and Solutions

Hussain K.H

Abstract

Unicode Language Technology has unleashed tremendous opportunities for building up information systems in Indian languages. Great efforts have been taken during the last ten years to build up bibliographic as well as non-bibliographic information systems using Malayalam script, all of them initiated and led by a few librarians in Kerala. The paper focuses on the problems faced/ facing in the creation of these systems. Dual encoding of Malayalam Unicode, multi-encoding originated from inflection and agglutination, non-systematization caused by script reform and negative attitudes of majority of librarians in Kerala towards these systems are discussed. Solutions are thought in many levels of linguistics, technological and educational. Soundex methodology is explained as an area to be investigated to contain the multitudes of spelling variations and compound formations in Malayalam.

Keywords: Computational Linguistics, India, Kerala, Unicode Malayalam, Malayalam Language Technology, Malayalam Information Systems, Malayalam DBMS, Malayala Lipi, Soundex, MLphone

1. Introduction

Unicode encoding of Malayalam in MS Windows and Linux operating systems in 2003 opened up a new era of databases using Malayalam script. Attempts like M-ISIS (Malayalam CDS/ISIS of Unesco) couldn't make an impact on information system building in Malayalam due to the natural limitations of ASCII encoding.

Unicode encoding attributes unique code points to basic characters of a language. It assigns unique positions (Range) to each language in the world having script. For Malayalam Unicode Range is 0D00–0D7F (Figure 1). Developments in Unicode based Malayalam Language Technology during the past one decade was tremendous, thanks to the commitment of youths in Swathanthra Malayalam Computing (SMC), a volunteer FOSS group. Malayalam computing now owns an envious position among Indian Language Computing.

Unicode offered incredible possibilities to build up Malayalam information systems (MIS). Technological advancements in Malayalam language computing were fully utilised to build up online archives and digital libraries during the past decade, which were acclaimed nationally and internationally. Some of them are pioneering works to become models for other Indian languages. It is interesting to note that all the digital archives and information systems discussed below which include non-bibliographic also,

Fig. 1 Malayalam Unicode Chart

are initiated, designed and successfully led by librarians.

2. Information Systems Using Malayalam Script (MIS)

2.1. **Rare Books Archive** (http://statelibrary.kerala.gov.in/rarebooks/index.php).

State Central Library (SCL), Thiruvananthapuram. Started in 2005 under the leadership of State Librarian Mr. Devadathan; Continuing in a phased manner. Nearly 1100 titles in which 110 are Malayalam. Digitised more than 600,000 pages

2.2. **MGU Thesis** (http://www. mguthe ses es.org/)

Mahatma Gandhi University Library, Kottayam. Initiated and led by University Librarian Dr. R. Raman Nair. Online archive launched in 2008. Archived 2186 doctoral theses out of which 113 are Malayalam. Digitised more than 600,000 pages.

2.3. **Malayala Grandha Vivaram** (http://www.grandham.org/)

Bibliography of 52,000 Malayalam books published up to 2000. Formerly compiled by K.M. Govi, Librarian, National Bibliography; Recompiled by Hussain K.H, Librarian, KFRI; Web hosted in 2009

2.4. **KFRI Digital Herbarium** (http://www.kfriherbarium.org/)

Kerala Forest Research Institute, Peechi; Launched in 2012. 10306 specimens representing more than 2040 species from 203 families from South India and Andaman Nicobar Islands. Locality and local names of species searchable using vernacular script.

2.5. **Kerala Legislative Assembly Archive** (http://klaproceedings.niyamasabha.org/)

Initiated and led by Mr. Sathikumar C.S, Librarian, Legislative Assembly. Digital Archives of Kerala Legislative Assembly Records from the year 1888 to 2011. Digitised more than 600,000 pages. Proceedings fully classified and indexed in Malayalam and English

2.6. **Gazette Archive** (http://statelibrary.kerala.gov.in/gazette/index.php)

State Central Library (SCL), Thiruvananthapuram. Initiated by State

Librarian Mrs. P. Suprabha in 2010; Continuing under the leadership of State librarian Mrs. P.K. Shobana; The collection starts with Travancore Gazattes from 1903 onwards. More than 20,00,000 (Twenty lakhs) of pages of government notifications and orders in English and Malayalam are digitised and indexed. A multilingual internal dictionary is built in to handle variations in place/personal names. This archive turned out to be the largest and most complicated one significantly contributing to the methodology of making MIS.

2.7. **Kerala Index** (http://www.kerala index.org/)

Initiated by Mrs. Girijamma R and Hashim E, Assistant Librarians, Kerala University Library, Thiruvananthapuram. Online bibliography of 60,000 articles published in 50 Malayalam newspapers, weeklies and monthlies since 1985. It is widely referred by scholars and journalists who research in Malayalam language and literature and Kerala studies

2.8. **N.V. Krishna Warrior Archive** (http://nvkrishnawarrior.org/)

Initiated by B. Krishnakumar (Atmaraman) it is the first complete archive of a laureate in Kerala. All published works of N.V. Krishna Warrior and studies by others on him are archived. Handwritten manuscripts, letters, paper clippings, translations, etc are classified and indexed for Malayalam retrieval.

2.9. **Kerala Vanasoochika** (http://vanasoochika.org/)

Initiated by Hussain KH, and led by Dr. M. Amruth, Scientist, KFRI. It is an archive of Malayalam articles, journals and books on forest and environment of Kerala and documentation of past forty years' environmental activism.

(All systems except Kerala Legislative Assembly Archive are designed by Hussain K.H., former librarian, KFRI)

3. Dual Encoding in Unicode

Considering some of the codes given to Malayalam Characters, general principle of Unicode (namely 'unique code points') is violated. Vowel sign of 'OU' is given two code points (0D4C and 0D57) for its two shapes, but one of them has been ceased to exist in printing and writing since 1950s. Similar dual encoding happens to 'EE' also (0D08 and 0D5F).

Encoding of CHILLS (0D7A to 0D7F) in 2008 has become a disastrous step by Unicode Consortium. CHILLS in Malayalam are exceptional characters in Indian scripts. Though the structure of Malayalam script follows the same patterns of all other Indian scripts, it takes a definite deviation in the case of CHILLS. Linguistically they are not basic characters in the alphabet. They are only 'other forms' of some combinations of basic characters taking different shapes and this makes them special to Malayalam. Many of the words with middle CHILLS can be expressed without them, forming conjuncts with basic ccharacters. Once the CHILLS were encoded in Unicode 5.1, the natural kinship with its basic components was broken. Two kinds of CHILLS now exists in Malayalam – shapes are same but codes are different (Figure 2). Some times it leads to absurd

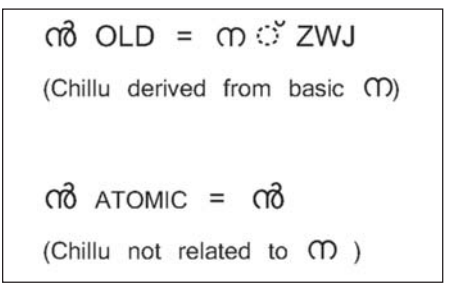

Fig. 2 Atomic CHILL

combinations (Figure 3). Millions of words, single as well as compound, have become victim to multi-encoding due to this erroneous encoding (Figure 4).

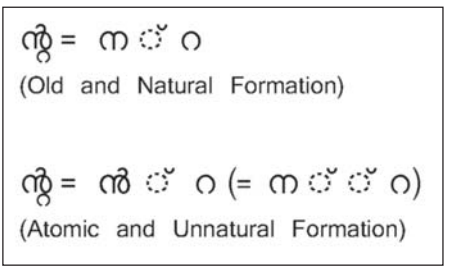

Fig. 3 Absurd combinations

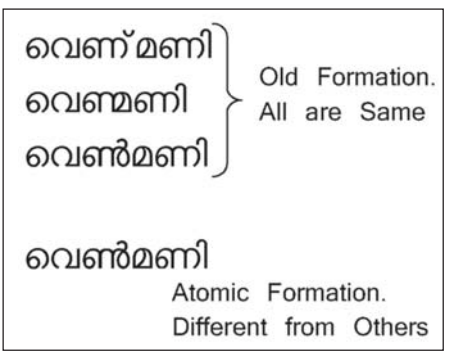

Fig. 4 Dual Encoding due to Atomic CHILL

4. Million Shades of Inflection and Agglutination

Prospective combinations of any number of basic single words to form compound words is a phenomena occurring in all Indo-Aryan and Dravidian languages. Malayalam is specially versatile in this compound formation. A short comparison of English and Malayalam dictionaries is sufficient to reveal this striking feature. All words English people use in speech, writing and printing can be found out in an English dictionary, while 95% of words Malayalee use in speech, writing and printing cannot be found in a Malayalam dictionary! This is because words in Malayalam dictionaries are basic words where as words in usage are compounds formed from dictionary words (Figure 5).

മലയാളം	മലയാളം
+ ഭാഷ	മലയാളഭാഷ
+ സാങ്കേതികം	മലയാളഭാഷാസാങ്കേതികം
+ പദാവലി	മലയാളഭാഷാസാങ്കേതികപദാവലി

Fig. 5 Formations of Compound Words

Even prepositions are combined with the preceding words. The word 'Stand' in English have only two other derivatives, where as its counter part in Malayalam have more than two hundred (or never ending?) derivatives (Figure 6)

Fig. 6 Never ending derivatives

Laws of formation of compound words are somewhat concrete, but breaking a lengthy compound word with space is a subject of dispute among grammarians. Various possibilities of separation of a compound word are not fixed in any style manuals of leading Malayalam publishers.

Combinations of components of a compound lead to 16 different expressions while its English counter part 'Malayalam Language Technology Glossary' has got only one expression (Figure 7). We all know that misspelling of a word with a single character or punctuation mark leads to a non-/alien

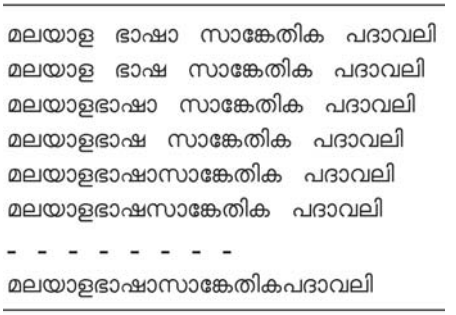

Fig. 7 Different Combinations of Components

retrieval in an information system. One can imagine the extent of multi-encoding and its danger in MIS from the above figures.

'Samvruthokaram' is a particular derivative of *Chandrabindu (Chandrakkala/ Meethal* Uni 0D01) which is the most elemental character in Malayalam alphabet without which other characters cannot exist. Changes occurred to words within a century, the twin conjuncts especially after CHILLS, encoding confusions of breaking a long word with spaces, usage of Samvruthokaram, etc. present bewildering instances of multi-encoding. '*Toddy Shop Auction*' in the Gazette Information system is a classic example of the situation. Each component of its Malayalam compound word takes variety of forms which can be joined or separated with or without spaces (Figure-8). Total possibility of rendering this expression at different stages of MIS (i.e. writing down keywords by indexer in the data sheet, typing data into database and typing query by users) open up 84 instances (Figure-9). No other language in the world can boast such an array of variance for a single expression leading to mismatching and non-retrieval of information!

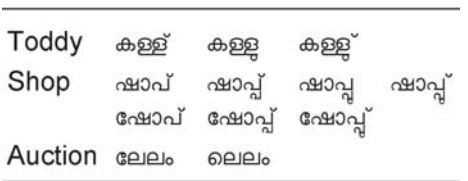

Fig. 8 Variations in components

Fig. 9 Possibilities of Combinations

How this menace can be conquered is the biggest question in MIS. The task involves combined efforts in linguistics, Informatics and education. Gazette Information System has attempted an internal dictionary of word-variant equivalences effectively. But the solution is only limited for two reasons. First, making of dictionary with all variations of keywords in a database like Gazette is a huge task and a single agency cannot complete it and maintain. Secondly, all possible derivatives and combinations of words in Malayalam are impossible to list. (Figure 6)

Standardisation of word formation and rendering should only be accomplished by a government agency like Kerala Language Institute. At the same time teaching of it should be part of the education, which might be a decade long process. Even then standardised compound words and their right usage at the user level will only be partially attained in Malayalam.

5. Perils of Script Reform

Reforming Malayalam script took place in early seventies. Distribution of typewriters in government offices and promoting Malayalam as language of governance necessitated the script reform. Nearly 900 types were in use at that time in hand-composing presses and it was obvious that all of them couldn't be accommodated on 90 keys of typewriter. Most of the conjuncts were therefore discarded and some important consonant-vowel joining critical to all Indian scripts were made disjoint by assigning uniform signs. Truncated character set for typewriters were praised as 'scientific', 'modern' and 'standardised' in the beginning. The reformed character set was hailed as 'New Lipi' (new alphabet) while the original character set rich in conjuncts was described as 'Old Lipi'. In 1973 new Lipi was introduced in the education system through textbooks. Slowly the dangers of 'New Lipi' started to surface. Devastating effects of non-systemization took place in all levels of teaching, learning and writing systems. Those disturbances and confusions continues even after four decades of its introduction.

In eighties desktop computers began to replace typewriters. Due to the limitation of 256 slots in the ASCII fonts New Lipi was adopted for Malayalam computing. Malayalam fonts were designed adding a few 'old' conjuncts for the expanding market of word processing, typesetting and desktop publishing (DTP). Different companies made fonts with different conjuncts and it is estimated that more than forty different mappings exist in Malayalam ASCII fonts which is unheard in any other languages.

In 1999 'Rachana Aksharavedi' started its campaign for the reintroduction of Old/ Original Lipi in Malayalam computing. 'Rachana' font was designed consisting 900 conjuncts. The exhaustive character set introduced by Rachana later happened to be the backborn of Unicode Malayalam. 'Rachana' and 'Meera' traditional Unicode fonts now own a superlative position in Internet Malayalam. Old Lipi is now being slowly adapted to typesetting and printing with the advent of Unicode complaint DTP packages like InDesign and Scribus. Latex has also shown its presence in typesetting Malayalam books. The momentum is on increase.

The changes and developments that occur are in the right direction, but forty years of new lipi has done all possible damages to learning and practicing of the language. This is visible now at the time information system building in Malayalam. Introduction of additional signs and breaking of conjuncts in New Lipi increased the length of compound words. To make them visually comprehensive there began the practice of separating compound words with spaces. Since the time of introduction of metal typefaces by Benjamin Bailey in 1824 Malayalam characters and words have been almost systemized and practiced for one and half centuries which is now in a broken state. The net result of script reform is the multiplication of multi-encoding and unpredictability in search and retrieval.

Chaotic state of content coding can be minimized to some extent by discarding new lipi. It may take years to return to the old structure of characters by reintroducing them in the education system but no other way is seen ahead. It is noted in the making of Gazette archive that the use of Rachana traditional font at the time of data entry reduces the errors in characters and word building. But when queries are made by users new lipi is followed. No one can guarantee that the present way of making words in queries may match to the right words in the database!

6. Negative Attitudes of Librarians

It is noted that all information systems in Malayalam, both bibliographic and non-bibliographic are built under the leadership of a handful of librarians. Unfortunately

majority of librarians in Kerala are not watching these contributions in a passionate way. Though Unicode has proven the feasibility of multilingual databases, librarians are not convinced of the role of Malayalam script. OPAC at almost all libraries in Kerala, including university libraries and major public libraries, are running with non-Malayalam script. Librarians' mindset work against MIS in many ways:

- Roman transliteration of Malayalam metadata is sufficient for e-catalog. It is easy to enter data and make query.
- Malayalam typing is very cumbersome compared to English
- Information systems like Gazette archive do not yield exact hits, and so on.

University libraries and public libraries are now switching over to Unicode compliant ILMS like Koha and Evergreen. Still professional librarians are reluctant to build up Metadata of Malayalam documents using Malayalam script. Although Malayalam has attained a formidable position in Indian language computing; librarians are ignorant of this fact. They are reluctant to investigate advanced querying and filtering system of 'Nitya' search engine provided in MG University Theses Repository, Gazette Archive and Kerala Index. It is interesting to note that they don't have any criticism against Google and its million irrelevant hits! But when they are getting more than one relevant hits while querying Gazette archive they cry for the inefficiency of the search system for not getting a single exact hit!

Librarians in Kerala who are ignorant of language technology and hesitant to accept the merits of already developed systems, are actually retarding the progress made in MIS. Departments of Library and Information Science in Universities and colleges have a major role in rectifying this malady. Curriculum and syllabus should be redesigned to accommodate topics like Malayalam Language Technology, its practice using right keyboards like INSCRIPT, building of Malayalam metadata and Unicode compliant ILMS. A center for Advanced Research in Library and Information Science established at a university in Kerala with language technology research as an important mandate could not start function due to cheap politics and personal complexes of some librarians. Unless the teachers and students become aware of the importance of language technologies for information retrieval and are equipped with related skills MIS will face continued threat from new generation of librarians also.

7. Spell Checker – Still A Dream

The case of inflection and agglutination together with many manifestations of compound word formations make spell checking system in Malayalam the hardest task to realize. It is not even rightly conceived at the government level, although projects on grammar checker and automatic translation is going on at government institutions and university departments! Volunteer organizations like SMC which contributed to much of the development of Unicode Malayalam Language Technology has not yet formulated a project due to lack of funds. Though a Herculean task, Malayalam should have a spell checker. It may be partially achieved by creating a 'User Dictionary' consisting of 'usable/used words/compounds/derivatives' instead of using vocabulary in an ordinary Malayalam dictionary. Author has proposed such a system to his friends at SMC and provided a set of words and personal names (Authors) extracted from the database of 'Malayala Grandha Vivaram' years ago. Making such a 'User Dictionary' is now more feasible since enormous corpus of Malayalam content in Unicode has been accumulated in the net and set of used words and phrases can be extracted from this. Proof reading and correction of each word in this set should be carried out meticulously before making it

available for spellchecking in word processors and MIS. In English spellchecking system usually a word is recognised by a unit of characters in between spaces and it is looked for in the internal dictionary. When a match is found in the dictionary the word is termed as 'right', otherwise 'wrong' and nearby words in the dictionary are suggested for correction. This simple method can be adapted for Malayalam also by the matching word (compound and derivatives) in the 'User Dictionary'. Apart from suggesting a right word two more functions should be added to Malayalam: 1. Suggesting possible splitting of the compound words, and 2. Possible combination with nearby words in the text to form right compound words. Malayalam Lexicon at Kerala University can accomplish such a task, since they have the experience of half a century of making one of the major dictionaries in Indian languages.

8. Soundex – A Possibility?

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English. It is used to search and retrieve words having similar pronunciation but slightly different spelling. A lot of variants may be found for English and Metaphone is one among them. MLphone by Kailash Nadh is a variant for Malayalam which is effectively used in his online dictionary OLAM. Santhosh Thottingal has attempted an IndicSoundex for Indian languages that can be applied to Malayalam.

Method of Soundex can be simply described as to eliminate vowels in a word to form a pseudo word/code to represent former. Discarding spaces and punctuation signs further in personal names eliminates slight variations in the rendering, producing a single unit of representation (Figure-10).

Soundex is not applied at the time of building actual data, but executed later at the time of building up index and query. Inverted File Index generation (usually by Lucene engine in modern DBMS like MySQL) is done on the set of pseudo words.

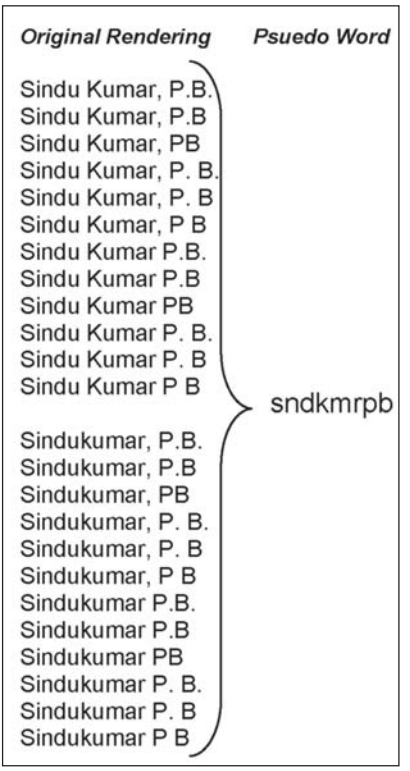

Fig 10 Single Unit of Representation

Search and retrieval is carried out by matching pseudo words of query with those in Inverted Index. Retrieved records (hits) are then formatted and displayed using original words (Figure-11).

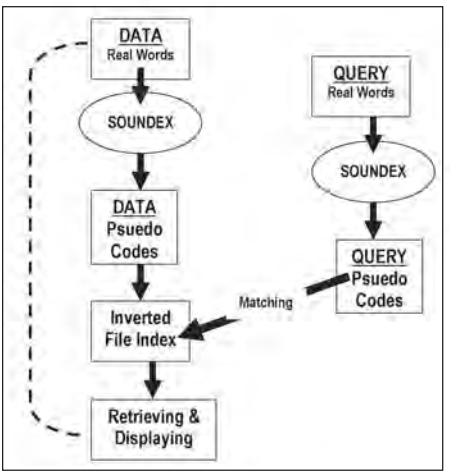

Fig. 11 Working of Soundex

Soundex is widely used in English textual databases effectively since elimination of vowels usually doesn't lead to the same pseudo representation for two words. But use of Soundex or MLphone may not be that much effective in the case of Malayalam (Figure-12).

Fig. 12 Same Pseudo Code for Different Words

Elimination of spaces in between words may be disastrous in representing combination of words. Joining of two words in Malayalam is not merely done by eliminating space in between, but according to definitive laws ('Sandhi' or 'Samasam'). The end character of the first word and the beginning character of the second word may some time be discarded or multiplied or replaced. Derivatives of a single verb shows the vast extension of expressions in Malayalam. Such shades of a single word are unthinkable in English. Principles for these formations are described by grammarians like Kerala Panini A.R. Raja Raja Varma, but any number of exemptions and patterns can also be found out. Without doubt it is a challenging task to discover an effective variant of Soundex for Malayalam, but once achieved it may be a way to reduce multi-encoding. Attempts of Kailash Nadh and Santhosh Thottingal are pioneering efforts towards this end.

9. Conclusion

Problems faced in MIS are more or less common to all Indian languages. This is because semantics of Indian languages are similar, and structure and shapes of scripts are inherited from ancient Brahmi script. Compared to other languages hurdles are more in Malayalam, especially caused by erroneous Unicode encoding and the new script. Even then Malayalam has made commendable advancements in building up digital libraries and Information systems. MGU Theses Archive is first of its kind among Indian universities. Even ShodGanga established by INFLIBNET (UGC) as a repository for theses from all universities is modeled after this archive, but inferior to MGU Theses in search and browsing using Malayalam and Devanagari scripts. KFRI Herbarium is the largest plant specimen digital collection in the country and no other digital herbariums have attempted retrieval using vernacular script for locality and local names of species. National Gazette of India and all other State Gazettes are uploading PDFs of their printed gazettes in web, but it is the Gazette Archive at Kerala State Central Library that ventured to index and retrieve its two million government orders and notifications using Malayalam script. No other language in India can boast of an exhaustive bibliography like 'Malayala Grandha Vivaram' for published works in their language.

OCR, Spell Checker, Standardisation of compound words and implementation of original script in education are prime areas for ICT research and development. Government of Kerala, policy makers and IT experts should focus on these themes urgently. Present lethargy and ignorance of librarians towards MIS will be fast disappeared once the topics of language technology are imbibed into Library and Information Science curriculum. They will then come to learn and appreciate works done

in this area by a few of their fellow professionals. Let us hope this will ignite their creativity and professionalism for better systems and solutions in their Mother tongue.

Acknowledgements

Author has got opportunities for building up MIS by the full support and freedom given by Mr. P.M. Abdulkadir, Managing Director of Beehive Digital Concepts. 'Nitya Digital Archive' was initially programmed by author in 2002 and later rewritten for web and configured to different MIS by Mr. Ajayan K.G. Rachana and Meera Unicode fonts designed by author are widely used in MIS and it is the continuous experiments and innovations of SMC friends Anivar Aravind, Santhosh Thottingal, Kavya Sathosh, Rajeesh K Nambiar and Praveen Arimprathody that keep these fonts and all MIS concurrent and functional. Mrs Suprabha P confided in me in creating Gazette Archive and Mr. Sathikumar C.S. taught me its organizational intricacies. Dr. Amruth at KFRI is continuing the building up of MIS Vanasoochika after my retirement from there. Finally, it is Dr. R. Raman Nair who is behind the continuous inspiration and guidance for my research on 'Bibliographic Information Systems Using Malayalam script'.

References

Anivar Aravind (2008). Localisation a Political process: A Case Study on Swathanthra Malayalam Computing. https://smc.org.in/presentations/localization-and-smc.pdf

Anivar Aravind (2016). Swathanthra Malayalam Computing. https://atlarge.icann.org/applications/smc-13may16-en.pdf

Arjun Babu and Sindhu L (2014). A Survey of Information Retrieval Models for Malayalam Language Processing. Proceedings of the International Journal of Computer Applications (0975 - 8887) 107 (14), December 2014.

Chitraja Kumar R; Gangadharan N (2005). Chandrakkala. Samvruthokaram. Chillaksharam. http://unicode.org/~emuller/iwg/p28/05210-malayalam.pdf

Hussain, K. H., et al. (2005) Creation of Digital Archives in Indian languages Using CDS/ISIS: development of M-ISIS (Malayalam ISIS) and 'Nitya'. Information Studies 11(1), 59-68.

Kailash Nadh (2012). MLphone. http://nadh.in/code/mlphone/

Metaphone. https://en.wikipedia.org/wiki/Metaphone

Muller, Eric (2005). Comments on PRI 66: Malayalam Cillaksarams. http://unicode.org/~emuller/iwg/p28/05148-muller-pri66.pdf

Rajeev J S, Chitrajakuma R, Hussain K H and Gangadharan N (2005). Multilingual Computing in Malayalam: Embedding the Original Script of Malayalam in Linux and Development of KDE Applications, Cochin, CALIBER, 2005.

Raman Nair, R. (2004). Malayala Granthasoochi 2004 of Government Brennen College, Tellycherry: The first electronic catalogue in Indian languages. http://hdl.handle.net/10760/8104

Raman Nair, R., and Hussain, K.H. (2010) Nitya Archive: Software for Full Text Digital Libraries in Indian Languages. http://hdl.handle.net/10760/15484

Santhosh Thottingal (2009). Inflection and Agglutination - Challenges to Malayalam Computing. http://thottingal.in/documents/MalayalamComputingChallenges.pdf

Santhosh Thottingal (2009). Phonetic Comparison Algorithm for Indian Languages. http://thottingal.in/blog/2009/07/26/indicsoundex/

Santhosh Thottingal (2011). Report on the Issues of Malayalam Language in Unicode. http://thottingal.in/documents/ReportonMalayalamUnicodeIssues.pdf

Sathikumar, C.S., Raman Nair, R. and Bhagi, N.K. (2007). Digital Archive of Kerala Legislative Assembly Proceedings. http://hdl.handle.net/10760/9291

Sobhana, P K (2016). Digitizing Heritage: An Account of the Digitization project of Kerala State Central University. National conference on Innovative Library Services in the Digital Age. 1-2, July 2016. Ernakulam, Toc H Institute of Science and Technology.

Soundex. https://en.wikipedia.org/wiki/Soundex

Suber, Peter, Raman Nair and Hussain, K.H. (2009). Open Access to Public Funded Research: A Discussion in the Context of Mahatma Gandhi University Open Access Archives of Doctoral Dissertations. http://ir.inflibnet.ac.in/handle/1944/998