A Survey on Technological development Resources of Malayalam

Rejitha K.S., PhD
Senior Resource Person,
Linguistic Data Consortium for Indian Languages,
Central Institute of Indian Languages, Mysore

ABSTRACT
Malayalam is in its developing stage to build structured linguistic and technological resources. These language resources can then be embedded in applications, devices and services that will eventually support as highly effective solutions for communication across language barriers in a flexible way. State and central government had funded for the language development and Malayalam has initiated it well. An enormous amount of effort will be required to produce language data and technology and the available resources listed may help the researchers for further development of Malayalam.

General Terms
Survey paper, Malayalam, Natural language processing,

Keywords
Language resources, Processing tools, Technological developments

1. INTRODUCTION
Language resources are essential for language study and analysis. A significant amount of work and time needed for high quality data. Writing system, printing press, computers have helped to store the language information to certain extends. There are many authentic language data and tools are available and listing those resources will be useful for further language progress. It is intended for knowing the developments and growth of a particular language; moreover these resources are required to advancing the research area and enhance the performance of language technology.

1.1 Malayalam as a Language
Malayalam, a high agglutinative and rich morphological language is the principal and administrative language of Kerala, Union territory of Lakshadweep and Mahé district, one of the districts of the Union Territory of Puducherry. In 7th century under the Kulashekhara dynasty Malayalam emerged as a distinct language. The Chera Dynasty was the first prominent kingdom based in Kerala. A series of Chera-Chola wars in the 11th century caused the decline of foreign trade in Kerala ports. Finally, the Kulasekhara dynasty was subjugated in 1102 by the combined attack of Later Pandyas and Later Cholas. In 14th century, Ravi Varma Kulashekhara, the southern Venad kingdom was able to establish supremacy over southern part. After his death, in the absence of a strong central power, the state was divided into thirty small warring principalities. The most powerful of them were the kingdom of Samuthiri in the north, Venad in the south and Kochi in the middle. In the 18th Century, Travancore King Sree Anizham Thirunal Marthanda Varma captured all the kingdoms up to Northern Kerala. The Kochi ruler sued for peace with Anizham Thirunal and Malabar came under direct British rule until India became independent. On 1 November 1956, when language came to be officially accepted as the basis for marking borders of states in India then Kerala was formed by the States Re-organisation Act and Travancore-Cochin state was merged with Malabar district of Madras and Kasaragod taluk of South Canara district.

1.2 Evolution of Malayalam Script
Malayalam designated as a Classical Language in India in 2013. The oldest available documents written purely in Malayalam are the Vazhappalli Copper plates from 832 AD and Tharisapalli Copper plates from 849 AD. The earliest extant work is a prose commentary on Chanakya's Arthasastra credited to the 13th century. The earliest script used to write Malayalam was the Vatteluttu alphabet, and later the Kolezhuttu, which derived from it. The current Malayalam script is based on the Vatteluttu script, which was extended with Grantha script letters to adopt Indo-Aryan loanwords. The Malayalam Script is a unicase script, means it does not have a case distinction. It is written from left to right direction. With the objective to simplify the script for print and typewriting technology of that time, the Government of Kerala reformed the orthography of Malayalam by a government order to the education department by reducing the number of glyphs required. The reformed script came into effect in 1971 thereby reducing the number of glyphs required. Print media almost entirely uses reformed orthography. Primary education introduces the Malayalam writing to the pupils in reformed script only and the books are printed accordingly. The script is also used to write Konkani and several minority languages such as Paniya, Betta Kurumba, Ravula etc.

1.3 Resource Development Challenges
Language technology is a profitable, sustainable and socially beneficial solution to overcome language barriers. Digitizing the language data make it easy to store, share and access. This conversion encounters many difficulties. Developing a highly accurate language model is difficult due to the complexity of human language. Especially Malayalam is a context sensitive language so making the robust language technology is a challenging task.

Only with the adoption of appropriate technology for each period the language can be sustained over time. Malayalam has completely implemented the changes that come from time to time. Natural Language Processing is the technology used to aid computers to understand the human's language. Computers interact with programming languages which are structured, unambiguous and precise. But human language has lot of ambiguity which makes sense contextually in natural language and human can comprehend it easily. The first requirement for the computer to understand human language is to encoding the language. Unicode is a scheme which

provides separate code for each character of all language. The Unicode range for Malayalam script is the code point 0D00–0D7F. Malayalam is a highly agglutinative language with an average of 10.56 UTF-8 characters per word [1].

"Average word length of Malayalam text is the highest among all the scheduled languages of India. Malayalam is highly agglutinative and morphologically rich language; hence the saturation level of Malayalam i.e. the new words coming into corpus for a unit amount of input is much higher compared to other languages." [2].

2. LANGUAGE RESOURCES

There are many language materials available to analyses the language data and develop different technologies and applications. A lot of advanced works are done for the improvement of Malayalam language. At present language resources mainly include language data and language processing software. Language data such as text and speech corpora, dictionary, ontologies, multimedia database etc. and software for their collection, preparation, annotation, analysis, management etc. are the resources. Moreover there are applications which help language study, electronic publishing, localization, language service industry are also included in it. There are many efforts happened to develop Malayalam language resources.

Linguistically rich huge data is the fundamental resource for any language technology. A corpus is the real time representation of the language. Corpus could be general, domain specific, parallel or annotated. Written corpus developing depend many factors like quality, quantity, representativeness, balance etc. It should represent wide range of texts and many varieties of text in consistent manner. Text corpus helps to make many supporting materials like frequency list, word list, lexicon, n-gram and which helps to build many higher level applications. Speech technology also needs large quality data for its development. Depending on the application read speech or spontaneous speech is recorded in different environment. Aligning text into the corresponding sound wave is an enormous task and speech annotation tools helps to make it in a proper format. Language analysis is done through tokenizers, morph analyser, part-of-speech taggers, parsers, information retrieval tools, machine learning tools.

Spell and grammar checking, speech recognition and synthesis, machine translation, information retrieval and extraction, text summarisation, question answering, dialogue systems are the main application area of language technology. Most of these tools depend on language resources. The ultimate aim of the language technology is to support human in such a way that it interact with its environment. It adapts the capacity to exchange information and communicate with the users and interact naturally. Different organisations, institutions and many public and private entities built tools and applications to proliferate the language.

Government of India is providing financial support through different funding agencies for language development. Central Institute of Indian Languages (CIIL), Technology Development for Indian Languages (TDIL), C.DAC Thiruvananthapuram, Indian Institute of Information Technology and Management–Kerala, etc. are some major institutions strive to make Malayalam language development possible.

2.1 Linguistic Data Consortium for Indian Languages

The Linguistic Data Consortium for Indian Languages (LDCIL) [3] is a scheme of the Department of Higher Education, Ministry of Human Resource and Development, Government of India implemented by and housed inside the Central Institute of Indian Languages, Mysore has developed Malayalam text and speech corpus. A Gold Standard Malayalam Text Corpus collected from books, magazines, newspapers, official documents etc. The data contains different domains like aesthetics, commerce, mass media, science and technology and social sciences. It consist of 63,70,954 words taken from 1,119 different titles. LDC-IL Malayalam Raw Speech Corpus contains read and spontaneous speech data. 164:01:02 hours of spoken data collected from various content types like words, sentences, continuous text etc. The data is collected from 458 speakers of male and female with different age group.

2.2 Swathanthra Malayalam Computing

SMC is a free software community and non-profit charitable society working on Malayalam and other Indian Languages [4]. It is launched on December 2001. SMC engaged in development, localization, standardization and popularization of various Free and Open Source Software in Malayalam language.

SMC created general text corpus, parallel corpus and word list for language processing. Text corpus collected from various free licensed sources and then curated and processed for general purpose usage. It has 98,15,533 words. English-Malayalam parallel corpora published by Wikipedia based on the article translations. Unique words extracted from Malayalam Wikipedia, Wictionary etc. The word list contains 14,27,392 words and which is listed as noun, verb, adjective etc.

A morphological analyser and generator built for Malayalam language using Finite State Transducer technology. They have created several Malayalam fonts like Manjari, Gayathri, Chilanka. They developed an Indic keyboard which currently supports 23 languages and 57 layouts and they set up a braille keyboard also. There are some other tools like Named Entity Recognition, Spell checker, Transliteration tool, Syllabifier, indic stemmer, Unicode converter, n-gram and Number spell out etc. are available. Dhvani is a text to speech system it is generating speech for Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Panjabi, Tamil, Telugu, Pashto.

2.3 Technology Development for Indian Languages

TDIL[5] is initiated by the Ministry of Electronics & Information Technology; Govt. of India has to develop Information Processing Tools, multilingual knowledge resources and standardization of Language Technology through active participation in International and national standardization bodies. The Central Institute of Indian Languages has coordinated for the development of 20,95,145 word corpora in Scheduled Languages under the scheme of TDIL. These are balanced corpora because it followed sampling methodologies. They are available in Indian Standard Code for Information Interchange (ISCII) format. Parallel Text Corpus contains millions of multilingual parallel text corpus in English and 11 Indian languages including Assamese, Bengali, Gujarati, Hindi, Kannada, Marathi, Malayalam, Oriya, Punjabi, Tamil & Telugu based on Unicode encoding. Under Indian Language Corpora Initiative

(ILCI) parallel corpus has been created. 50,000 sentences in Hindi have been translated from Hindi into Malayalam and have been POS tagged and chunked.

TDIL has Malayalam Image Corpus and it includes 500 scanned images of Malayalam language in TIFF format. Images are scanned at 600 dpi in grayscale mode. These scan images are the byproduct of OCR project.

Sampark [6] is an automated Indian Language to Indian Language Machine Translation System.

It is developed with the combined efforts of 11 institutions in India under the consortium project Indian language to India Language Machine translation. The project has developed Machine Translation technology for 18 language pairs. They are: 14 bi-directional pairs between Hindi vs Urdu, Punjabi, Telugu, Bengali, Tamil, Marathi, Kannada and 4 bidirectional between Tamil vs Malayalam, Telugu.

2.4 C-DAC Trivandrum

C-DAC [7] is a national resource center for computing in Malayalam and it is working on application oriented language research. C-DAC has been a pioneer in developing and proliferating the language technology. It includes multimedia and multilingual computing solutions covering a wide range of applications such as operating platforms, machine translation, language learning etc.

Developing TTS system has been undertaken for Indian Languages in the consortium project mode with the leadership of IIT Madras [8]. Other consortia members are: IIIT Hyderabad, IIT Kharagpur, IISc Bangalore, IIT Guwahati, IIT Mandi, CDAC Mumbai, CDAC Thiruvananthapuram, CDAC Kolkata, SSNCE Chennai, DA-IICT Gujarat and PESIT Bangalore. The project has been funded under TDIL Programme by Ministry of Electronics and Information Technology, Govt. of India. Text to Speech for 13 Indian Languages including Malayalam have been developed using the both Open Source FESTIVOX Framework and State-of-the art HTS based engine. C-DAC developed an efficient OCR for Malayalam. This was developed under e-Aksharayan [9] project of TDIL.

2.5 Cochin University of Science and Technology

Cochin University of Science and Technology (CUSAT) [10] has POS tagged corpus of 28,7588 words and the dataset is prepared using BIS tagset. Malayalam WordNet [11] '*padasrimkhala*' is developed by the Department of Computer Science, Cochin University. It could be used by common people and it works in a controlled crowd sourced manner in which the users would be able to add new and edit existing synsets or adding new relationships. There are 30,482 synsets.

2.6 Amrita Vishwa Vidyapeetham

Malayalam WordNet has been undertaken by Amrita University (2011-2015) under the financial support of MeiTy, Govt. of India. Malayalam WordNet has been constructed based on Hindi WordNet under Dravidian WordNet project funded by DietY. Nearly 30,000 systets have been created and linked to the four major Dravidian languages. The Dravidian WordNet is the part of the Indo WordNet project.

2.7 Indian Institute of Information Technology and Management -Kerala

IIITM-K [12] developed a Malayalam POS Tagger and this online tool tagging the by using BIS (Bureau of Indian Standards) tagset and IIITH (International Institute of Information Technology Hyderabad) tagset.

2.8 Commission for Scientific and Technical Terminology

CSTT [13] prepared Fundamental Administrative Terminology in English-Malayalam to localize the technical terms in regional languages. Maintaining the uniformity and popularizing the standard words in the respective language CSTT published text book oriented Glossary of Archaeology, Mathematics, Chemistry, Economics, Education, Geology, Political Science, Commerce, Geography and Sociology.

2.9 OPUS

OPUS [14] is a growing collection of translated texts from the web and aligned as a publicly available parallel corpus. The corpus is prepared in such a way that the text is aligned with sentences which allows searching and translation between all languages. OPUS is based on open source products and the corpus is also delivered as an open content package. All pre-processing is done automatically and no manual corrections have been carried out.

2.10 Open SLR

Open Speech and Language Resources provide a site [15] to download the software and language resources. Crowdsourced high-quality Malayalam multi-speaker speech data set is available in OpenSLR. "Each volunteer read around 100 sentences in an hour. The volunteers were asked to speak with neutral tone and pace" [16]. This data set contains transcribed high-quality audio of Malayalam sentences. 2,103 sentences recorded from female speakers and 2,023 sentences from male speakers. The data set consists of wave files, and a TSV file (line_index.tsv). The data set has been manually quality checked.

2.11 TC-11 Online Resources

Document image analysis and recognition helps other research area in a better way. There are some character image dataset available in [17]. This handwritten data collected from 77 (60 Female and 17 Male) native Malayalam writers between 20 to 55 age groups and all the writers have minimum graduation as the educational qualification. Fast global minimization algorithm for active contour models(ACM-FGM)employed for detecting the character objects in the collected document images. For converting the resultant image to a binary representation, Otsu's global image threshold algorithm is used. Each image is converted to 32*32 dimensions.

2.12 Kaggle

It is platform to share and publish dataset privately or publicly. It is not only a repository but also a community where one can discuss and discover the techniques and knowledge. Malayalam multi-speaker speech dataset is available in [18].

3. MALAYALAM LANGUAGE RESOURCE COMPARISON

A comparative overview of the available Malayalam language resources are given below:

Table 1. Malayalam language resource comparison

Resource/Tool	Resource Centre	Content
Text corpus	LDC-IL	63,70,954 words
	TDIL	20,95,145 words
	SMC	98,15,533 words
Parallel text corpus	TDIL	50,000 words
	SMC	98,15,533 words
	OPUS	Millions of tokens
Speech corpus	LDC-IL	164:01:02 hours
	Kaggle	Public data platform
	Open SLR	4,126 Sentences
POS	CUSAT	2,87,588 words
	IIITM-K	Online tagger
	ILCI	50,000 sentences
Chunking	ILCI	50,000 sentences
Word net	CUSAT	30,482 synset
	Amritha	30,000 synset
OCR	TDIL	Application software
	C-DAC	Image data
	TC-11	Handwritten data

4. CONCLUSION

Malayalam has done less work compared to some other Indian languages and European languages. Resources developing take huge amount of time and money so available resources have to be shared. Then repetition of the works needs to be avoided. There should be combined efforts for the technological development of Malayalam. Technology stands in turning point so forecast the natural language processing in the direction of research, development and industrial area. Innovative language processing techniques modify Malayalam as a full-fledged digitized language. There are many incomplete useful works the respective authority has to initiate to complete those works. Build language resource such as lexicons, huge data base for general and specific domain, large annotated corpora, NLP systems and application. Encourage cloud sourcing to minimize the data collection effort and improve language data management structure and right way of data distribution policy. Competent authority has to increase funding to develop Malayalam language technology and research. Researchers have to work more on goal oriented approaches.

5. REFERENCES

[1] K.S., Rejitha, Rajesha N. “Saturation of Indian Language Corpora - Malayalam vs. Hindi”. Working Papers on Linguistics and Literature, Department Of Linguistics, Bharathiyar University, Coimbatore. Volume: XIII No.2, 2019. ISSN 2349-8420

[2] K.S., Rejitha, Saritha S.L., Sajila S., Rajesha N., Manasa G., Narayan Choudhary & Ramamoorthy, L. “Documentation of LDC-IL Malayalam Raw Text Corpus". Central Institute of Indian Languages, Mysore. 2019.

[3] Data Distribution Portal of Linguistic Data Consortium for Indian Languages (https://data.ldcil.org/)

[4] Swathanthra Malayalam Computing (SMC) is a free software collective engaged in development, localization, standardization and popularization of various Free and Open Source Softwares in Malayalam language. (https://smc.org.in/)

[5] Technology Development for Indian Languages Programme, MeitY, Govt of India. (http://tdil.meity.gov.in/)

[6] Indian Language to Indian Language Machine Translation System (https://tdil-dc.in/index.php?option=com_vertical&parentid=74&lang=en)

[7] Centre for Development of Advanced Computing (C-DAC) (https://www.cdac.in/index.aspx?id=tvm)

[8] TTS Consortium, DeitY (https://www.iitm.ac.in/donlab/tts/)

[9] e-Aksharayan, Indian Language OCR (http://xn----ytdg2a7eme8k7bydbgb.xn--h2brj9c/eocr/index.html)

[10] CUSAT Tagged Malayalam Corpus dataset (https://cs.cusat.ac.in/mlpos.jsp)

[11] Malayalam WordNet (http://malayalamwordnet.cusat.ac.in/search.do)

[12] IITMK Malayalam PoS Tagger (https://www.iiitmk.ac.in/MalayalamPOSTagger/)

[13] Commission for scientific & technical terminology (http://www.csttpublication.mhrd.gov.in/english/result.php)

[14] OPUS Dataportal (http://opus.nlpl.eu/)

[15] Open Speech and Language Resources (https://openslr.org/63/)

[16] He, Fei and Chu, Shan-Hui Cathy and Kjartansson, Oddur and Rivera, Clara and Katanova, Anna and Gutkin, Alexander and Demirsahin, Isin and Johny, Cibu and Jansche, Martin and Sarin, Supheakmungkol and Pipatsrisawat, Knot. "Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems.", Proceedings of The 12th Language Resources and Evaluation Conference (LREC). European Language Resources Association (ELRA). 2020

[17] Malayalam Character Image Database (http://tc11.cvc.uab.es/datasets/Amrita_MalCharDb_1)

[18] Malayalam multi-speaker speech data set portal (https://www.kaggle.com/kurianbenoy/malayalam-multispeaker-speech-data-set/kernels)

[19] Menon, A.Sreedhara. 2007. A Survey of Kerala History. DC Books.