

മലയാള മൂലഭാഷണ ദത്തശേഖരം



MALAYALAM
RAW SPEECH CORPUS

Malayalam Raw Speech Corpus

Authors:

Dr. Rejitha K.S.

Dr. Saritha S.L.

Dr. Sajila S.

Rajesh N.

Manasa G.

Dr. Narayan Choudhary

Dr. L. Ramamoorthy



34

Annotated, quality language data (both-text & speech) and tools in Indian Languages to Individuals, Institutions and Industry for Research & Development - Created in-house, through outsourcing and acquisition.

**Linguistic Data Consortium for Indian Languages
Central Institute of Indian Languages
Mysore, India-570006**

Malayalam Raw Speech Corpus

This Documentation is a part of LDC-IL Malayalam Raw Speech Corpus.

For Further contacts: oic@ldcil.org

All rights reserved

© Central Institute of Indian Languages, Mysore-2018

Authors:

Dr. Rejitha K.S.

Dr. Saritha S.L.

Dr. Sajila S.

Rajesh N.

Manasa G.

Dr. Narayan Choudhary

Dr. L. Ramamoorthy

Last updated: October 10, 2018

ISBN 978-81-7343-242-2 Electronic publication of Corpus

Publisher:

Linguistic Data Consortium for Indian Languages

Central Institute of Indian Languages

Mysore, India-570006

Contents

1	LDC-IL Raw Speech Corpora: An Overview	1
1.1	Introduction	1
1.2	LDC-IL Speech Corpus	2
2	Content Type Descriptions.....	3
2.1	T1: Contemporary Text	3
2.2	T2: Creative Text	4
2.3	D: Date	4
2.4	S: Sentences	4
2.5	W1: Command and Control Words.....	5
2.6	W2: Proper Noun (Person Names and Place Names)	5
2.6.1	Person Names	5
2.6.2	Place Names.....	5
2.7	W3: Most Frequent Words	6
2.7.1	W3A: Most Frequent Words-Part	6
2.7.2	W3B: Most Frequent Words-Full	6
2.8	W4: Phonetically Balanced Vocabulary	6
2.9	W5: Form and Function words	7
3	Planning for Fieldwork	8
3.1	Dataset preparation and distribution	8
4	Field Work	10
4.1	Possible places for collecting data	10
4.2	Field work Ethics	10
4.3	Data Collection.....	11
4.3.1	Technical Specifications for collecting data	11
4.3.2	Metadata.....	12
4.3.3	Data Transferring and Storing.....	15
4.3.4	Observations	15
5	Organising and Archiving the Data	16
5.1.1	Text - Speech Mapping and Naming Conventions	16
5.1.2	Observations	16
6	Data verification and Quality Control	17

7	Malayalam Raw Speech Corpus	18
7.1	Introduction	18
8	Dataset preparation for Malayalam	21
8.1	Model 1 (T1, T2, S, D, W1, W2, W3, W4, W5).....	21
8.2	Model 2 (T1, T2).....	22
8.3	Data Collection Details	22
9	Transliterations in LDC-IL Malayalam Read Corpus	23
10	Summary of the Corpora	24
10.1	Summary of the Audio Segments	24
10.2	Duration of the Raw Speech Data.....	25
10.3	Distinct Set	25
10.3.1	Contemporary Text (News).....	25
10.3.2	Creative Text	26
10.4	Random Set.....	26
10.4.1	Creative Text	26
10.4.2	Date.....	26
10.4.3	Sentences	27
10.4.4	Command and Control Words	27
10.4.5	Person Names	27
10.4.6	Place Names.....	28
10.4.7	Most Frequent Words - Part	28
10.5	Full Set.....	28
10.5.1	Most Frequent Words - Full	28
10.5.2	Phonetically Balanced Words	29
10.5.3	Form and Function Words	29
10.6	Native Speakers Distributions.....	29

Table of Figures

Table 1: LDC-IL Speech Data Content Types	3
Table 2: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-1 Dataset	8
Table 3: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-2 Dataset	9
Table 4: Technical Specifications for collecting data	11
Table 5: Metadata Legends and their Description.....	14
Table 6: Representation of Model 1 Content Type.....	21
Table 7: Representation of Model 1 Prompt Sheet	21
Table 8: Representation of Model 2 Prompt Sheet	22
Table 9: Representation of Data Collection Details	22
Table 10: Representation of Audio Segments of Malayalam Raw Speech Data	24
Table 11: Representation of Duration of Malayalam Content	25
Table 12: Distribution of Malayalam Contemporary Text (News)	25
Table 13: Distribution of Malayalam Creative Text in Distinct Set	26
Table 14: Distribution of Malayalam Creative Text in Random Set.....	26
Table 15: Distribution of Malayalam Date Format	26
Table 16: Distribution of Malayalam Sentences	27
Table 17: Distribution of Malayalam Command and Control Words	27
Table 18: Distribution of Malayalam Person Name	27
Table 19: Distribution of Malayalam Place Names.....	28
Table 20: Distribution of Malayalam Most frequent Words Part	28
Table 21: Distribution of Malayalam Most Frequent Words - Full	28
Table 22: Distribution of Malayalam Phonetically Balanced Words.....	29
Table 23: Distribution of Malayalam Form Function Words	29
Table 24: Distribution of Malayalam Native Speakers.....	29

1 LDC-IL RAW SPEECH CORPORA: AN OVERVIEW

1.1 INTRODUCTION

Lack of basic linguistic resources have been the first and foremost bottleneck in development of language technology for Indian languages. When text data itself has been available for most of the Indian languages, one could not even think of the speech data. India is one of the foremost multilingual country where multilingualism is ingrained and most people speak more than one language with more than 75 languages having more than one million speakers (as per 2011 Census of India data). As per a study¹ of KPMG and Google released in 2017, the internet user base grew at a compound annual growth rate (CAGR) of 41% between 2011 and 2016 to reach 234 million users at the end of 2016 and this trend is likely continue. It is also estimated that internet users in Indian language will account for nearly 75% of India's internet user base by 2021.

Despite this, the availability of technology in Indian languages have been on close to null. This is mainly due to the reason that the technology developing agencies find it either too difficult to come up with the language support on various applications for Indian languages or it is economically not a viable solution. However, recent analyses from various quarters have shown that the latter is not correct and the major issue is availability of the linguistic resources based on which language technology and language support for various types of applications proves to be a bottleneck for the developing community, be it industry or otherwise.

Considering this as an issue, the Government of India has taken several initiatives to provide the basic ingredients which may prove as a catalyst for the development of language technology in Indian languages. As part of the this initiative, a scheme named Linguistic Data Consortium for Indian Languages (LDC-IL) was established by the Ministry of Human Resource and Development at Central Institute of Indian Languages, Mysore.

The goal of LDC-IL was to develop linguistic resources for all Indian languages with the initial focus more on the scheduled languages of India. These linguistic resources may be as deemed fit by the language technology developing community.

Based upon the recommendations of the Project Advisory Committee which includes ex-officio members from MeitY, IITs Ministry of HRD, Director and other academicians from reputed Institutes/Universities working in this area as well as major and minor industrial entities working in this area, the LDC-IL decided to embark upon developing the text and speech corpus for the scheduled languages of India.

There have been several types of datasets prepared under LDC-IL. This document serves as a generic documentation for the raw speech corpus of the LDC-IL being released for several languages.

¹<https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/04/Indian-languages-Defining-Indias-Internet.pdf>

1.2 LDC-IL SPEECH CORPUS

LDC-IL speech corpus is collected after careful deliberations on what type of speech corpus is required for various types of speech based linguistic analysis that may suit multifarious needs of the research and development community.

After several meetings with the experts from around India and abroad, it was decided that LDC-IL should focus on not just developing a speech corpus for a particular need, rather to get the data that would be useful for various tasks such as ASR, STT, linguistic analysis, speech therapy and so on.

Keeping this in mind, various types of content were created a priori before the speech recordings took place. The content of these datasets have been prepared in consultation with the experts from the language as well as linguists giving inputs to ensure that no specific sound patterns are missed out.

For example, it has been ensured that the speech datasets contain all the phones and allophones of the language and ample examples are available in the language to prove their phonemic status in the language. To ensure that the corpus is good for an ASR, the continuous speech is recorded in natural environment.

2 CONTENT TYPE DESCRIPTIONS

Each content type has a number of files with each file containing standard content. A sub-set of these files in each of the content types selected randomly constitute subsets that are given to a speaker for reading out in natural flow. A few full sets (namely W3B, W4, and W5) are also read in full by certain selected speakers in each age group.

There are three age group ranges selected for LDC-IL datasets. These are ‘16 to 20 years’, ‘21 to 50 years’ and ‘above 50 years’. Attempt has been made to collect equal number of male and female data from each of the age groups.

The list of the datasets and their notation is given in the table below:

SL	Notation	Content Type
01.	T1	Contemporary Text (News)
02.	T2	Creative Text
03.	S	Sentence
04.	D	Date
05.	W1	Command and Control Words
06.	W2	Place Name
07.	W2	Person Name
08.	W3A	Most Frequent Word-Part
09.	W3B	Most Frequent Word-FullSet
10.	W4	Phonetically Balanced-Fullset
11.	W5	Form and Function Word-Fullset

Table 1: LDC-IL Speech Data Content Types

Detailed descriptions of each of the content types are given in the following sub-sections.

2.1 T1: CONTEMPORARY TEXT

The Contemporary Text (news) data is given the notation of T1. News items have been selected from the LDC-IL news corpora. The text is contemporary in nature as the news items such have been picked over a period from 2005 to 2012, either from news websites or from the print editions newspapers of the respective language.

The domain information is present in the news items as well as the news items deal various topics such as political news, editorials, sports news and so on. Given that the news items have been collected from local news reported for each language, the style may be considered as colloquial or belonging to the news reporting style.

Each LDC-IL dataset ‘Contemporary Text’ contains minimum of 500 words per speaker, which is rarely repeated. Since it is the continuous text, it constitutes the largest part of the speech corpora, in terms of data size and time duration.

2.2 T2: CREATIVE TEXT

‘Creative Text –T2’ is extracted mainly from literary sources. It is used to capture literary terms. Creative Texts are stories or essays collected from books. The text may be any standard text which is descriptive in nature. It exhibits the language style of a particular period from which the text is taken.

Creative texts were prepared in two types. In the first 6 or 8 essays or short stories were prepared. One of these selected randomly from the set, is assigned to one speaker for reading out. The same story may be read by multiple speakers.

In the other approach a distinct text is given to each individual

The creative text section of the LDC-IL Speech dataset comprises of mostly six essays or short stories. One of these essay or short story, selected randomly from the set of the six stories, is assigned to one speaker for reading out. The same story may be read out by multiple speakers.

2.3 D: DATE

Languages tend to speak out the date in a specific and many a times in a particular manner which may not always conform to the grammatical structure of the language. To capture it, LDC-IL tried to document how a date is spoken in each of the languages.

The normal way is put a question before the informant the answer of which must be in a date format. Normally the following six questions were placed before the informant and the informants would answer minimum one of the questions.

1. What is tomorrow’s date?
2. When is Gandhi Jayanthi observed?
3. What is the date today?
4. When do we celebrate our Independence Day?
5. What is your date of birth?
6. On which date you go to market?

2.4 S: SENTENCES

To ensure that all the types of syntactic structures are covered in the speech data, a set of sentences have been constructed with the help of language experts and linguists for each of the languages. It is ensured that all possible sentence structures are covered including all types of tenses, aspects, moods, compound and complex sentences and so on.

These sentences are isolated sentences and not part of a continuous speech. While care has been taken to extract sentences from the text corpus of the corresponding language, sometimes sentences have also been modified to ensure that the specific valid sentence structure of the language is present.

Very long sentences are avoided while selecting or constructing the sentences, so that the informant can read the sentences easily. The words used in these sentences are common words which are found in day-to-day life. Each sentence in the list contains minimum four words. The sentences are not too long so that each sentence does not span for more than a line in the prompting sheet. Care is taken to avoid abusive or taboo words.

Each speaker is given 25 sentences out of this sentence list for reading out.

2.5 W1: COMMAND AND CONTROL WORDS

Spoken language usually contains a lot of sentences that are commands or use a lot of control words. This happens mostly in the conversational speech. Even though the LDC-IL speech corpus at present does not contain the conversation speech, an attempt has been made by including common command and control sentences/phrases carefully crafted with the help of respective language experts and linguists.

These include imperative sentences, optative sentences as well as other controlling phrases which may come as a reply to an interrogative sentence. Each of the languages has a set of command and control sentences created before the speech data is recorded. Each speaker is given a list of 30 command and control sentences randomly selected from the set. Each of these phrases/sentences is repeated three times by each speaker while recording.

2.6 W2: PROPER NOUN (PERSON NAMES AND PLACE NAMES)

Recognizing proper nouns by using an ASR system is a complex task. For example, voice recognition application in mobile phone may have a few hundreds of names to distinguish when placing a call through voice command. Native speakers use different pronunciations depending on their language of origin and familiarity with the language. The speakers use different pronunciation for native and foreign names ranging from a nativised pronunciation to a totally foreignised pronunciation. All this adds to the complexity of an ASR system in recognizing proper nouns. To address this issue LDC-IL speech data has been collected to include person names and place names.

2.6.1 Person Names

Person names are included to capture the native pronunciations. The names are taken from people from different walks of life like Politicians, Film Actors and Directors, Writers, Kings and Queens, Astrologers, Historical Personalities, Scientists, Sports persons etc.

2.6.2 Place Names

Place names are included to capture the native pronunciations. This data set contains Indian place names. These include main cities, district names and popular tourist destinations from all over India. Some local place names are also included like names of villages, taluk headquarters, district names, local forest reserves, local tourist and pilgrimage destinations etc.

Each speaker typically pronounce 20 person names and 10 place names, out of the total Proper Noun wordlist of the particular language. Each word is uttered three times

2.7 W3: MOST FREQUENT WORDS

Most frequent word list is the regularly and repeatedly used list of words. Since these words are used most frequently in a language, it is imperative to have these words in our dataset.

The most frequent words dataset is derived from LDC-IL Corpus. However, it may be noted that when the most frequent word list was extracted, the text corpus was rather small. So, the frequency list might change if it is compared to the current LDC-IL text corpus.

2.7.1 W3A: Most Frequent Words-Part

The most frequent words of a language are randomly picked from a list of around 1000 most frequent wordlist of a language. Each speaker records randomly selected 30 words from this list. Each word is uttered thrice.

2.7.2 W3B: Most Frequent Words-Full

Two speakers, one male and one female, pronounces the full set of 1000 most frequent words. This is done for each dialect of the language, if available.

2.8 W4: PHONETICALLY BALANCED VOCABULARY

To cover all possible phonemic occurrences of a language, the “phonetically balanced Vocabulary” is prepared. It is a list of words in which the occurrence of a phoneme in initial medial and final positions of that language can be represented.

The pronunciation of the phoneme is varied according to the position of the phoneme in a word and the influence of the following and proceeding phoneme. The pronunciation of initial position is different from middle and final positions. For example the phoneme ‘pa’ is used in different forms while pronouncing words like

- ‘**pallavi**’- ‘pa’ inherent vowel at initial position (CV initial)
- ‘**prakata**’ - ‘p’ as pure consonant in conjunction with ‘ra’ in initial position, (CCV Initial)
- ‘**spandana**’, - ‘pa’ with inherent vowel preceded by a consonant at medial position (CCV Initial)
- ‘**parikalpane**’- ‘pa’ inherent vowel at initial position (CV initial) and ‘pa’ with inherent vowel preceded by a consonant in the medial position (CCV Medial)
- ‘a:**pta**’ - ‘p’ with followed by a consonant in the final position (CCV medial)

Using the articulation score as the measure, phonetically balanced lists have been used to test differences among transmission systems and to test the effects of noise. The phonetically balanced words used in word recognition testing contain speech sounds that occur in the same frequency as those of conversational speech.

2.9 W5: FORM AND FUNCTION WORDS

Form and function words dataset is a closed class list of words. They are quite limited in a language. These constitute mostly the indeclinable words of a language. Form words are static, bearing some content with them. They are meaningful and are actually the building blocks of a language.

The Form and Function dataset includes Grammatical function words, numerals, kinship terms, measurement terms, list of colours, days, months, seasons, directions, zodiac signs, body parts, planets etc. These words are included to the native words which may not be frequent in the overall corpus, but needs representation.

3 PLANNING FOR FIELDWORK

3.1 DATASET PREPARATION AND DISTRIBUTION

To ensure representativeness of the speech corpora, a conscious effort has been made to balance the speech data by taking varieties of styles into consideration. The first and foremost among at LDC-IL has been to take an expert view on the varieties of languages. For example, for Kannada it is ensured that speech varieties from different regions such as Hyderabad Karnataka, Bombay Karnataka, Coastal Karnataka and Old Mysore get proportionate weightage.

LDC-IL collected the data using two approaches, with two different kind of Dataset Models They are as follows

- Dataset Model 1 (T1, T2, W1, W2, W3, W4, W5, S, D)
- Dataset Model 2 (Distinct Texts of T1 and T2)

Some Languages followed Model-1 only, and some Languages followed both Model-1 and Model-2 After the regions are identified, speech samples are collected as per the criteria shown in the tables below:

Standard Speech Dataset Distribution for Each LDC-IL Fieldwork Dataset Model 1							
Content type	Content size#	Content to be read by one speaker	Total No. of speakers	Age group wise no. of speaker; Female & Male equally distributed#			Content selection type
				16-20	20-50	50+	
Contemporary Text	150 Texts	1 Text	150	18	90	42	Distinct Text
Creative Text	6 Texts	1 text	150	18	90	42	Random set*
Sentences	142 Sentences	25 Sentences	150	18	90	42	Random set*
Command and Control Words	82 Words	30 Words	150	18	90	42	Random set*
Person Names	489Words	20 Words	150	18	90	42	Random set*
Place Names	511 Words	10 Words	150	18	90	42	Random set*
Most Frequent Words	1144Words	30 Words	150	18	90	42	Random set*
Phonetically Balanced Vocabulary	390 words	Full set	6	2	2	2	Full set to be read by the speaker
Form and Function Words	432 words	Full set	6	2	2	2	Full set to be read by the speaker
1000 Most Frequent Words	1000 Words	Full set	2	0	2	0	Full set to be read by the speaker
*picked randomly by machine #The figures shown are for illustration purpose only. The numbers may differ for each language. Please refer specific Language documentation for actual figures.							

Table 2: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-1 Dataset

Speech dataset distribution for fieldwork Dataset Model 2						
Content type	Content size	Content to be read by one speaker	Total No. of speakers	Age group wise no. of speaker; Female & Male equally distributed		Content selection type
				16-20	21-50	
Contemporary Text (News)	150 Texts	1 Text	150	75	75	Distinct Text
Creative Text	150 Texts	1 text	150	75	75	Distinct Text

Table 3: Standard Speech Dataset Distribution for Each LDC-IL Fieldwork with Modle-2 Dataset

As the data is collected from different cities across India (as per the demand of the language), its imperative that proper preparation is made before proceeding towards the field such that day-to-day necessities of field are met with. Investigators had to make that s/he had sufficient charged batteries as well as alkaline batteries if so required, empty SD cards, laptops in proper condition, sufficient number of random and full datasets (prompt sheets) and so on. These formed as the daily routine for the linguists while in the field.

4 FIELD WORK

Some common guidelines and instructions were provided to the field workers before proceeding to the field. A brief of it is noted below.

4.1 POSSIBLE PLACES FOR COLLECTING DATA

Once the dataset is prepared and taken to the field, the next step is to determine places where there is an availability of speakers who can read fluently. The best possible places are schools, colleges, universities, govt. offices etc.

The Head of these organizations have to be briefed and asked permission for recording data from students, faculties etc. Certain infrastructural requirements like space, if possible power source for charging batteries etc. has to be requested from the institutions. The speakers from whom we collect data are referred as informants.

4.2 FIELD WORK ETHICS

The informants are briefed about the procedures, nature and purpose of speech data collection. Informants are informed about the funding agency behind the data collection. In case of LDC-IL, the data collection is funded by Govt. of India. Informant are made aware of who exactly is carrying out the data collection process and what will be done with the data collected before they give their consent.

There have been situations where the informant would find it distressing that the data given by them will be segmented and further processed. In such cases, their opinions have to be respected and the investigators have to refrain from taking their data. The informants are made aware of the degree of confidentiality and anonymity that will be maintained after collecting the data. The informants are also made aware of the potential benefits of the data to the wider community. Once the informant is aware of all these information and is ready to give the data, consent is acquired in written along with certain personal details such as their educational qualification, mother tongue, place of elementary education etc.

Informants are allowed to read the dataset earlier before recording so that they can get familiar with the content of the text. It is ensured that the informants do not have any objection to the content they are about to read. For example, the informant may have objection regarding the political, social views expressed in the content. In such cases, a different dataset is offered to the informant. There are certain texts in the data set, which may pose difficulty for a certain informant to read. For Example, some informants may have difficulty in reading contents which involve dialogues between people. Such contents may differ in dialects spoken by the informant which may pose a difficult situation for them while reading. In such cases, a different dataset is offered to the informant. Complex datasets are given only to the informants who are capable of reading them and state likewise.

An attempt is made to reduce the extra noise as much as possible before recording. If necessary, test recordings are conducted before the actual recording on certain portions of the text.

Brief introduction about the informant and investigator along with details like place, time, region etc. are collected at the beginning of each recording. The conversation between investigator and informant is done in their native language so that the informant is comfortable and the natural flow of language is established.

Care is taken while recording the words, so that there is a pause between two words or between utterances of the same word. All the words of the content type W1 to W5 (i.e. ‘Command and Control words’, ‘Proper Nouns’, ‘Most Frequent Words’, ‘Phonetically Balanced Vocabulary’ and ‘Form and Function words’) are repeated three times in a sequence. A pause is maintained between two sentences as well while recording.

While recording the News Item and Creative Text, the informants are briefed to read the text given, as naturally as possible. It should be as natural as reading a book or newspaper. Informants answer to a particular question themselves regarding date format. This is done to capture how people usually pronounce the date. The investigator does not prompt any particular format

4.3 DATA COLLECTION

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder.

4.3.1 Technical Specifications for collecting data

Recording Setup:	Sample Rate : 48.0 KH
Recording Mode:	wav -16bit
Date Setup:	Current date and time.
Storage:	SD Card
Power:	• Always use rechargeable batteries (Ni-MH) for recording. Otherwise line hum will come. Never use Ni-CD battery type as it is potential for ‘memory effect’. • Rechargeable batteries need to be thoroughly recharged before recording (minimum 16 hrs continuous charging).
Peak	While recording please be aware that it should not reach the peak i.e. PEAK (in the recorder) should not glow.
Recording Distance	• Keep minimum 5 cm to 25 cm distance between the microphone and the speaker and if possible use microphone holder. • The recorder should not be placed orthogonally but it should be placed diagonally. • Do not move the recorder during recording • Fix the recorder upon a table/ fixed plane if possible. • Try to have fixed the distance between the recorder and speaker • The recorder should not be placed orthogonally but it should be placed diagonally

Table 4: Technical Specifications for collecting data

After each recording, it is recommended to verify the recorded data, whether it is recorded in the right way. If the informant also wishes to hear the data, the investigator may oblige.

4.3.2 Metadata

The value of speech data can be determined according to the quality of metadata obtained. It is imperative to maintain metadata of the entire data collection for linguistic analysis.

After the recording is taken from the informant, personal details are collected. Care should be taken so that the signature and other formalities are completed as required.

The metadata of the speech corpus is made through the personal details taken from the informants. A typical copy of metadata sheet contains information as noted below:

Informant Data:

Name:
Dataset ID:
Address:
Gender:
Age Group: (with three options of 16 to 20, 21 to 50, and 50+)
Educational Qualification: (with three options of School/Bachelors/Masters)
Place of Elementary Education:
Mother Tongue:
Dialect (if any):

Investigator Data:

Name:
Date:
Place:
Region:
Environment:

It is to note that the name and the address of the informants are discarded while archiving metadata to keep the confidentiality and anonymity.

Dataset ID: It is a unique ID given to each speaker.

The following fields are considered for the distinctiveness of each data item recorded. Each field contributes certain features which pave way for diverse research.

Gender: The Speech data is taken from both male and female to capture the difference in intensity and pitch. The difference in vocal folds size between men and women makes them different in their pitched voices. Male voice usually has low pitch whereas a female voice is of high pitch. Pitch and intensity are proportional to each other.

Age Group: Different age groups exhibit difference in pitch and loudness. As the human body ages, it undergoes changes such as lessening strength, slower movements, degeneration of body tissues etc. these factors impact the voice as well. As people age their speech slows down, syllables and words are elongated, sentences are punctuated with more pauses for air. Scientific studies also show that as male and female age, the changing larynxes changes pitch and intensity. Age also affect the hearing process, which may make a person speak louder.

Educational Qualification: This determines the fluency and speed of reading speech data.

Place of Elementary Education: This parameter determines the effects of environment and dialect of a particular place of childhood which impacts the articulation of the speech.

Mother Tongue: Mother Tongue is one of the influential factors of a native speaker, for example In Karnataka, mainly in Canara region; it can be observed that the mother tongue of native Kannada speakers may be Tulu, Konkani, Chitpavani etc. This influences the articulation of Kannada speech in these areas.

Place: Place gives better information about the speech data collected. For example, Kannada spoken in Kundapura has its own distinct variety even when it belongs to Canara region.

Date: Date describes the timeline of data collected. It becomes useful information for historic research and language evolution in time line. It also dates the technology being used in that age.

Region: Region is an influencing factor of the language. Hence keeping the information about it with the data is always useful.

Environment: The recording environment information's like Indoor, Outdoor, School, Office, etc is useful, and its marking could be helpful in determining the noise level and the kind of noise that can be expected with the data.

Each of the datasets released contain a metadata sheet which has information about each of the audio files. A total of 25 fields are there in the metadata sheet.

A brief of each of these 25 fields/legends is given in the table below:

SL	Legend	Description
1	Language	Name of the Language
2	SpeakerID	Each speaker has a unique identity language. However, this is within the language. If one is working on speech corpus from more than one language, the IDs may get repeated.
3	ContentType	This corresponds to the notation of the content types noted above.
4	ContentID	This corresponds to the ID of the text being read out.
5	Gender	Notes gender, whether it is male, female or other.
6	AgeGroup	Three age groups of 16 to 20, 21 to 50, and 50+
7	Dialect	Notes the dialect of the language. An attempt has been made to cover all the dialects of the language as agreed upon in the academia of the language experts and linguists.
8	ReadInScript	The script in which the content has been provided to read in.
9	RecordingEnvironment	A brief info on the environment in which the recording has been done.
10	PowerSource	The source of the power using which the recording was done. It may be Li-ion, NiCd or Alkaline batteries.
11	RecorderManufacturer	Manufacturer of the recorder.
12	RecorderType	Type of the recorder. It is mostly 24-bit Linear PCM (R-09).
13	SamplingFrequency	Sampling frequency. It's mostly 48.
14	BitPerSample	Bit per sample. It is mostly 16-bit.
15	Channel	How many channels. All of LDC-IL data is stereo. Only data set is mono which is segregated and constitutes a separate dataset of its own.
16	State	Name of the Indian state/province to which the speaker belongs to.
17	District	Name of the Indian district to which the speaker belongs to.
18	Place	Name of the place to which the speaker belongs to.
19	MotherTongue	Mother tongue of the speaker. It is note that data has been taken from people who professo to speak the language. However, it may be that the speaker uses the target language as a second or third language. However, as long as the speaker confidently says (and it is also verified by other speakers of the community), some samples have been taken from these types of users as well. This adds to the variety of the speech data collected.
20	EducationalQualification	Highest educational qualification of the speaker.
21	PlaceOfElementaryEducation	Place of the elementary education. This usually corresponds to the early childhood experiences which happens to more than often affect the way a language spoken.
22	RecordingDate	Date when the recording took place.
23	Investigator	Name of the Investigator.
24	RecordedText	Text of the recorded speech (in the script of the language).
25	TextInRoman	Text of the recorded speech (in the Roman transliteration as per the LDC-IL transliteration scheme. If the text is long (as is the case with T1 and T2 content types), a reference of the corresponding file is given.)

Table 5: Metadata Legends and their Description

4.3.3 Data Transferring and Storing

After the data is collected for the day or when the SD card is full, the data needs to be transferred to the PC. It is important, to take certain precautions in this process so that the data is safely transferred. The data should be copied and pasted in the PC rather than cut and pasted. After successful transfer and rechecking the copied data, the SD card can be cleared.

For easier maintenance and organization of the data in PC, folder system is recommended for saving data. Each recorded wave file has to be labelled with corresponding speaker ID.

The investigator should try to get the required number of speakers/data before completing the field work within their schedule.

4.3.4 Observations

One of the reasons for error prone reading could be the over consciousness of the informant about being voice recorded. Despite being informed, the informant may try to read the data in a dramatic way, but may eventually lead to normal reading after few sentences. Even after the informants give consent and their data, they may later change their mind or express their concern about the text they have read. Some may even request to discard their recordings. In such cases, the investigator has to reassure them about their given data. If they still want their data to be discarded, they have to be accommodated. It is preferable to provide complete information to the informants to avoid such situations. In many instances informants assume that they are giving auditions for Radio Jockey vacancies, or some reality shows. They should be briefed about the purpose of data collection beforehand to avoid such situations.

The investigator may be in not so hospitable environments depending upon the region they are visiting. Proper precaution and aid is to be acquired before visiting such places.

The investigator may have to face challenges in food and accommodation since he/she travels in unfamiliar places. It is recommended to be prepared for such situations. The investigator should be prepared for all such hardships and take proper measures to minimize them beforehand.

5 ORGANISING AND ARCHIVING THE DATA

After the field work is completed, the data has to be stored in a server as soon as possible for safe keeping. Taking a backup of the saved data is also recommended as the data collected is of vital importance.

5.1.1 Text - Speech Mapping and Naming Conventions

After the data is stored, it is segmented and mapped with its corresponding text and metadata. Each recording is named in accordance with its metadata information like language name, speaker id, content id, gender, age, content type etc.

A Typical LDC-IL naming convention for recorded data is shown bellow.

“LDC-IL_Scheduled_Kannada_Female_16To20_Contemporary Text-T1_SP-0035_T1-0035.wav”

“LDC-IL_Scheduled_Kannada_Male_21To50_Sentence-S_SP-0001_S-0004.wav”
--

Wave Surfer, free software, is used for segmentation of LDC-IL Speech data. It is an open source tool which can be downloaded freely from the web. While segmenting the speech data file for archiving, the introduction, content headings and any unnecessary speech are discarded. Only the dataset content is retained.

The ASR data is prepared keeping in the view, the stochastic systems such as HMMs or neural networks that do not use explicit rules for speech recognition. On the contrary, they rely on stochastic models which are trained using some statistical optimization procedure, with very large amounts of speech corpus.

5.1.2 Observations

While segmenting a single large recording containing all the content types, there may be instances where an informant has made an error and later corrected it. In such cases, it is always a good approach to segment a recording from the end of the file in reverse order so that the correct utterance can be found before incorrect utterance; hence the incorrect utterance can be discarded/ignored. One may observe the interventions of investigator or other people between reading two data items which may also need to be discarded.

6 DATA VERIFICATION AND QUALITY CONTROL

Since mapping audio recordings with its corresponding text and other metadata information is a manual task. The process is prone to human errors; the data verification process will be done

Much of the audio text mapping is automated, but in case of distinct set texts, and other metadata entry is done by human needs verification. In this,

- The Audio recording of each speaker is checked against the mapped text.
- Each distinct text audio recording will be matched with automated entries of the same speaker to check for any mis-mapping of speaker.
- Metadata like Gender, age group, District etc are selective part of manual data entry and could be prone to errors so verification is needed.
- Metadata like Dialect entry, place, etc are keyed in by manual data entry and could be prone to errors like typo errors so verification needs to be done.
- The audio segments could be duplicated because of system/network errors and need to be checked.
- At the time of data segmentation, one might have saved the whole file instead of selected part. Such case needs to be checked.
- Some audio segments may not get migrated to the system because of wrong naming conventions. Such segments will be handpicked and corrected and migrated into the system.

7 MALAYALAM RAW SPEECH CORPUS

7.1 INTRODUCTION

On July 1, 1949, Travancore and Kochi joined to form the unified Travancore-Cochin state. But Kerala continued to be politically divided till the 1950s, even with the geographic similarities and solidarity of language. On 1 January 1950, Travancore-Cochin was recognised as a state. On 1 November 1956, the state of Kerala was formed by the States Reorganisation Act merging the three distinct areas such as Malabar district, Cochin and Travancore and taluk of Kasargod which is in South Canara. Four southern taluks like Thovala, Agastheswaram, Kalkkulam and Vilavancode separated from Travancore-Cochin which was merged with Tamilnadu.

Malayalam is the official language of Kerala and Laccadive Islands. It belongs to the Dravidian language family. Malayalam is closely related to Tamil and it is more influenced by Sanskrit than Tamil. After Independence, the state governments of Kerala started using regional languages more and more in administration. Greater use of Malayalam has contributed to the growth of the language in terms of vocabulary and the number of styles and registers.

Language is the collection of more or less similar idiolects. The fundamental fact about language is its diversity. Change in language is found when we move from country to country, region to region, class to class and caste to caste. Bloomfield (1933) says that linguistic diversity is related to the density of communication or to the amount of verbal interaction among speakers. In India dialect studies in a broad sense have been initiated by G.A. Grierson, who collected evidences to understand the linguistic situation in India and to group the regional dialects into families of Language such as the Austric, Tibeto - Chinese, Indo European and Dravidian.

Dialect variation in a language is not random but systematic. There are two types of dialects; regional dialects and social dialects. Regional dialects are geographically based and social dialects originate among social groups, class ethnicity, religion etc.

Language variation reflects the language change over time and people who live in the same geographical area or maintain the same social identity share language norms. Language change happens through three parameters like spatial, temporal and social. People never speak the same way in all time. They exploit the nuances for different purposes. People of different social classes, different occupations or different cultural groups in the same community will show variations in speech. People of different occupation have their own dialects and they use their own technical terms for better understanding. Education brings a greater difference in language style. History has contributed its own compliments to language. During wars people acquire words used by military people and in course of time these words spread through generations. Dialect variation is also due to political reasons like people are tried to preserve the dialects of their ruling kings. It correlated with geographical factors such as un-bridged rivers, impenetrable forests, valleys, mountains, deserts etc. Marshals and artificial political barriers divide speech communities.

Language variation is due to different internal factors like semantics, vocabulary, grammar, phonological features, intonation patterns etc. along with other external factors region, cast, religion, education, occupation, social stratum, style, register etc. Various levels of linguistic structures show variations in different regional varieties of a language. Different groups of people who are living in two different areas show considerable differences in their language patterns. Malayalam spoken by any group of the northern region and that of the southern region shows that significant feature is almost uniform for any group of the respectable regions. There are many lexical items with purely regional connotations and the same forms in two areas have two different meanings and also there are forms which are considered as taboos in one region, may not be perceived as same in another region. Similarly certain verbs and nouns have co-occurrence restrictions at regions.

All dialects of a language are equally efficient and expressive. In the case of Malayalam the socio economic and political status of the speech community has nothing to do with the standardization of the dialect. Irrespective of the socio-economic factors, all people use both the high and low varieties of Malayalam for different purposes. In Malayalam speech community, more of the lexical codes of the regional and caste dialects interfere with standard Malayalam.

Kerala formerly into Travancore, Cochin and Malabar resulted in lesser communication among the people of these three kingdoms. The sea separates Lakshadweep from Kerala and to lesser communication resulted in a distinct Malayalam dialect called Lakshadweep Malayalam. The Malayalam dialects show both regional and social variations and sometimes they overlap.

Malayalam has a number of social varieties depending on caste and religion. There are differences in the speech of Christians, Hindus, and Muslims within a single geographic area. The language of the high castes in Hindus is more influenced by Sanskrit than the language of the lower castes. Dialects spoken by Christians have more loan words from Portuguese, Latin, and English than other dialects. Dialects spoken by the Muslim population have many borrowing words from Arabic and Urdu. Lots of lexical items and many idiomatic expressions in modern Malayalam are of English origin. The influence of some other languages like Tamil, Prakrit, Pali, Marathi, Hindi, Persian, Dutch and French can be seen in the course of its evolution and transformation. In Kasargod, people are using 7 other languages such as Tulu, Kannada, Beary, Konkani, Urdu, Hindi, Gujarati, Marathi and Tamil along with Malayalam.

The existence of Southern, Central and Northern dialects of Malayalam is fairly obvious even to common people. But under each broad group there are a number of sub varieties, as shown by the dialect survey of Malayalam (Ezhava-Tiyya, Subramaniam 1974) which identified 12 dialect areas of Malayalam spoken by Ezhavas and Tiyyas. In all probability the other caste dialects also closely follow the geographical stratification found in Ezhava-Tiyya dialects. In other words, the twelve dialect areas identified by the survey can be generalized to include all Malayalam regional dialects. The twelve areas are (1) South Travancore, (2) Central Travancore, (3) West

Vempanad, (4) North Travancore, (5) Cochin, (6) South Malabar, (7) South Eastern Palghat, (8) North Western Palghat, (9) Central Malabar, (10) Wayanad, (11) North Malabar and (12) Kasargod. Another significant dialect is the Malayalam spoken in Lakshadweep which is not connected to the dialect of the mainland.

According to the formation of Kerala and the language of Travancore, Cochin and Malabar regions are influenced by different internal and external factors so LDC-IL considered Malayalam has three specifically different varieties, thus collected speech data from Thiruvananthapuram, Ernakulam and Kozhikode.

LDC-IL divided the Malayalam speaking areas into these three regions and collected speech data from each. After determining the regions for fieldwork, the datasets were prepared for each region.

8 DATASET PREPARATION FOR MALAYALAM

LDC-IL collected the Malayalam speech data using two approaches. Two different kind of Dataset Models were prepared as follows

- Dataset preparation Model 1 (T1, T2, W1, W2, W3, W4, W5, S, D)
- Dataset preparation Model 2 (Distinct Texts of T1 and T2)

8.1 MODEL 1 (T1, T2, S, D, W1, W2, W3, W4, W5)

For the Regions of Thiruvananthapuram and Ernakulam LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Notation	Content Type	Count
T1	Contemporary Text	300
T2	Creative Text	6
D	Date	2
S	Sentences	303
W1	Command and Control Words	58
W2	Person Name	599
W2	Place Name	300
W3	Most Frequent Words	1000
W4	Phonetically Balanced Words	518
W5	Form and Function Words	545

Table 6: Representation of Model 1 Content Type

Distinct News Items were prepared to get the audio recording of contemporary text. It was made sure that each selected news item had minimum 500 words. Each prompt sheet had a distinct news item and selected part of the dataset prepared as follows.

Content Type	Content that Each typical prompt sheet had	Content selection type
Contemporary Text	1 Text	Distinct Text
Creative Text	1 text	Random Text selected from dataset*
Sentences	25 Sentences	Random set selected from dataset*
Command and Control Words	30 Words	Random set selected from dataset*
Person Names	20 Words	Random set selected from dataset*
Place Names	10 Words	Random set selected from dataset*
Most Frequent Words	30 Words	Random set selected from dataset*
* selected by machine		

Table 7: Representation of Model 1 Prompt Sheet

The full set of

1. Phonetically Balanced Words
2. Form and Function Words
3. 1000 Most Frequent Words

were also carried to the field to get recorded by selected individuals.

8.2 MODEL 2 (T1, T2)

For the fieldwork of Kozhikode LDC-IL attempted a different approach of dataset preparation, and tried to concentrate on capturing more distinct continues text recordings rather than word segments. The prompt sheets for Kozhikode were prepared as follows.

Content Type	Content in Each typical prompt sheet	Content selection type
Contemporary Text	1Text	Distinct Text
Creative Text	1 text	Distinct Text

Table 8: Representation of Model 2 Prompt Sheet

Once all these preparations are made, the investigator started collecting the data. All the speakers who provided their recordings Malayalam Speech Corpus to LDC-IL are native speakers of Kerala and of Malayalam as their first language.

8.3 DATA COLLECTION DETAILS

The Collection of data is carried out in three phases for different regions as follows.

Region/Place	Year of data collection	Resource Person
Thiruvananthapuram	2008-09	Saritha S.L.
Ernakulam	2009-10	Saritha S.L.
Kozhikode	2012-13	Rejitha K.S. & Midhun P.G.

Table 9: Representation of Data Collection Details

9 TRANSLITERATIONS IN LDC-IL MALAYALAM READ CORPUS

For easy reference and uniformity, the recorded text in the metadata file, is transliterated from Malayalam to Roman letters. Numeric characters were transliterated from Malayalam to Hindu-Arabic system.

The LDC-IL transliteration scheme of Malayalam to Roman is given below.

LDC-IL Transliteration Schema															
Malayalam characters to Roman and Malayalam Numerals to Hindu-Arabic															
Vowels															
അ	ആ	ഇ	ഈ	ഉ	ഊ	ഋ	ൠ	ൡ	ൢ	എ	ഏ	ഐ	ഒ	ഓ	ഔ
	ഠ	ി	ീ	ു	ൂ	ൃ	ൄ	൶	൷	െ	േ	ൈ	ൊ	ോ	ൌ
A	A	i	I	u	U	x	X	q	Q	e	E	ai	o	O	au

Consonants				
ക	ഖ	ഗ	ഘ	ങ
Ka	kha	ga	gha	ng'a
ച	ഛ	ജ	ഝ	ഞ
Ca	cha	ja	jha	nj'a
ട	ഠ	ഡ	ഢ	ണ
Ta	Tha	Da	Dha	Na
ത	ഥ	ദ	ധ	ന
Ta	tha	da	dha	na
പ	ഫ	ബ	ഭ	മ
Pa	pha	ba	bha	ma
യ	ര	ല	വ	ശ
Ya	ra	la	va	sha

Symbols			
ഌ	ഥ	വ	ശ
m`	m'	M	H

ഷ	സ	ഹ	ള	ഴ	റ	ട്ര
Sa	sa	ha	La	Za	Ra	TTTa

Chillu Letters								
ൺ	൹	ൺ	ൺ	ൺ	ൺ	ൺ	ൺ	ൺ
N'	n'	R'	I'	L'	k'	M'	y'	Z'

Numerals (Malayalam to Hindu-Arabic)									
൦	൧	൨	൩	൪	൫	൬	൭	൮	൯
0	1	2	3	4	5	6	7	8	9

The greyed out characters are obsolete. They may rarely present in the current LDC-IL corpus.

10 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Malayalam raw speech corpus based on various yardsticks which can also be filter out from the metadata sheet. These figures may be helpful in tuning the corpus for various purposes of training, testing and evaluating various algorithms as well as provide useful insights into the dataset. The data size is of total duration 164:01:02 (hh:mm:ss) comprising 43,670 audio segments.

10.1 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of utterances and their distribution in the Malayalam speech dataset.

LDC-IL Malayalam Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Contemporary Text (News-T1)	449	51	123	53	48	122	52
Creative Text-T2	449	51	124	52	48	122	52
Date-D	598	26	172	106	22	168	104
Sentence-S	7452	323	2146	1342	275	2081	1285
Command and Control Words-W1	8923	388	2559	1608	330	2501	1537
Person Name-W2	5819	259	1679	1032	219	1603	1027
Place Name-W2	2906	128	826	515	110	810	517
Most Frequent Word-Part-W3A	8763	387	2509	1570	330	2425	1542
Most Frequent Word-FullSet-W3B	1979	0	990	0	0	989	0
Phonetically Balanced-W4	3096	0	1552	0	0	1029	515
Form and Function Word-W5	3236	0	2158	0	0	1078	0

Table 10: Representation of Audio Segments of Malayalam Raw Speech Data

10.2 DURATION OF THE RAW SPEECH DATA

The table below shows the duration of each of the content type and their distribution across a few factors.

LDC-IL Malayalam Speech Data Status	Gender →	Female			Male		
	Age Group →	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
Content Type	Total Duration	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)	Duration (hh:mm:ss)
Contemporary Text (News-T1)	71:29:21	10:02:57	18:18:49	07:04:59	09:48:49	18:38:36	07:35:11
Creative Text-T2	54:41:20	11:37:21	14:27:14	02:16:48	10:22:47	13:38:40	02:18:30
Date-D	00:53:45	00:02:12	00:15:10	00:09:51	00:01:42	00:14:38	00:10:12
Sentence-S	06:56:46	00:19:09	01:59:41	01:18:50	00:15:48	01:51:56	01:11:22
Command and Control Words-W1	07:09:37	00:18:56	02:07:02	01:18:35	0:13:05	02:00:16	01:11:43
Person Name-W2	05:26:33	00:14:45	01:37:40	01:00:15	00:10:29	01:27:31	00:55:53
Place Name-W2	02:28:24	00:06:50	00:43:10	00:27:42	00:04:52	00:40:24	00:25:26
Most Frequent Word-Part-W3A	06:51:31	00:18:51	02:01:13	01:16:18	00:12:40	01:52:08	01:10:21
Most Frequent Word-FullSet-W3B	02:08:58	00:00:00	00:56:10	00:00:00	00:00:00	01:12:48	00:00:00
Phonetically Balanced-W4	02:40:09	00:00:00	01:24:12	00:00:00	00:00:00	00:59:27	00:16:30
Form and Function Word-W5	03:14:38	00:00:00	02:12:37	00:00:00	00:00:00	01:02:01	00:00:00

Table 11: Representation of Duration of Malayalam Content

10.3 DISTINCT SET

The Distinct Set usually contains data which is distinct to each speaker and is rarely repeated. The LDC-IL speech data set contains newspaper extracts which are read by each speaker.

10.3.1 Contemporary Text (News)

Distinct Text Extracts from Newspapers are recorded from the informants to get the speech data of contemporary text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution of text		Region-wise Distribution					
				Thiruvananthapuram		Ernakulam		Kozhikode	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	99	51	48	5	1	8	10	38	37
21 To 50	245	123	122	43	37	42	47	38	38
50+	105	53	52	30	35	23	17	0	0
Total	449	227	222	78	73	73	74	76	75

Table 12: Distribution of Malayalam Contemporary Text (News)

10.3.2 Creative Text

Distinct Text Extracts from literary books are recorded from the informants to get the speech data of literary text. These types of distinct creative text are collected from Kozhikode Region. The distribution of data is as follows:

Age Group	Total Audio Segments	Kozhikode Region Distribution	
		Female	Male
16 To 20	75	38	37
21 To 50	76	38	38
Total	151	76	75

Table 13: Distribution of Malayalam Creative Text in Distinct Set

10.4 RANDOM SET

The Random Set data composes of content types which are sampled by machine for each speaker. They are sampled from collection of master data sets available. The random sets are given below:

10.4.1 Creative Text

One randomly selected text of literature out of 6 texts from the prepared dataset is recorded from the informants to get the speech data of Creative Text. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	24	13	11	5	1	8	10
21 To 50	170	86	84	44	37	42	47
50+	104	52	52	29	35	23	17
Total	298	151	147	76	75	73	74

Table 14: Distribution of Malayalam Creative Text in Random Set

10.4.2 Date

Each informant answers two questions to get the date format. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	48	26	22	10	2	16	20
21 To 50	340	172	168	88	74	84	94
50+	210	106	104	60	70	46	34
Total	598	304	294	158	146	146	148

Table 15: Distribution of Malayalam Date Format

10.4.3 Sentences

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Malayalam. 25 Randomly selected Sentences are recorded from a list of 303 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	598	323	275	123	25	200	250
21 To 50	4227	2146	2081	1096	907	1050	1174
50+	2627	1342	1285	768	860	574	425
Total	7452	3811	3641	76	75	1824	1849

Table 16: Distribution of Malayalam Sentences

10.4.4 Command and Control Words

The command and control words content type contains a list of 58 words that is a representation of almost all the command and control words occurring in Malayalam. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	718	388	330	149	30	239	300
21 To 50	5060	2559	2501	1299	1093	1260	1408
50+	3145	1608	1537	918	1027	690	510
Total	8923	4555	4368	2366	2150	2189	2218

Table 17: Distribution of Malayalam Command and Control Words

10.4.5 Person Names

The person name contains a list of 599 names. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	478	259	219	99	20	160	199
21 To 50	3282	1679	1603	834	679	845	924
50+	2059	1032	1027	592	687	440	340
Total	5819	2970	2849	1525	1386	1445	1463

Table 18: Distribution of Malayalam Person Name

10.4.6 Place Names

The place name contains a list of 300 popular Pan Indian and regional place name. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	238	128	110	48	10	80	100
21 To 50	1636	826	810	406	348	420	462
50+	1032	515	517	295	347	220	170
Total	2906	1469	1437	749	705	720	732

Table 19: Distribution of Malayalam Place Names

10.4.7 Most Frequent Words - Part

The most frequent words-part contains a list of 1144 most frequent words. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Thiruvananthapuram		Ernakulam	
		Female	Male	Female	Male	Female	Male
16 To 20	717	387	330	148	30	239	300
21 To 50	4934	2509	2425	1266	1035	1243	1390
50+	3112	1570	1542	881	1033	689	509
Total	8763	4466	4297	2295	2098	2171	2199

Table 20: Distribution of Malayalam Most frequent Words Part

10.5 FULL SET

The full sets are the master set of certain data sets which are read completely from few selected speakers in each group. The full sets are given below:

10.5.1 Most Frequent Words - Full

The most frequent words contain a list of 1000 most frequent words. In full set all the 1000 words is recorded from the informant. The distribution of data is as follows:

Total Audio Segments from 21-50 Age group	Gender-wise Distribution		Gender-wise Distribution			
			Thiruvananthapuram		Ernakulam	
	Female	Male	Female	Male	Female	Male
1979	1979	0	990	0	989	0

Table 21: Distribution of Malayalam Most Frequent Words - Full

10.5.2 Phonetically Balanced Words

The phonetically balanced words contain a list of words where almost all the phones of Malayalam language have occurred in all the possible positions of a word. In full set all the 518 words is recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution of words in Thiruvananthapuram	
		Female	Male
21 To 50	2581	1552	1029
50+	515	0	515
Total	3096	1552	1544

Table 22: Distribution of Malayalam Phonetically Balanced Words

10.5.3 Form and Function Words

The form and function words content type contains a list of 545 words that is a representation of almost all the form and function words occurring in Malayalam. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Total Words three utterance each from Speakers of 21-50 Age group	Gender-wise Distribution of words in Thiruvananthapuram	
	Female	Male
3236	2158	1078

Table 23: Distribution of Malayalam Form Function Words

10.6 NATIVE SPEAKERS DISTRIBUTIONS

For Malayalam speech data a total of 458 speakers were collected in which 231 female speakers and 227 male speakers from three different regions. The distribution of data is as follows:

Region-wise Distribution of Native Speakers									
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions					
				Thiruvananthapuram		Ernakulam		Kozhikode	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	98	50	48	5	1	8	10	38	37
21 To 50	253	127	126	52	42	42	47	38	38
50+	107	54	53	31	36	23	17	0	0
Total	458	231	227	82	78	73	74	76	75

Table 24: Distribution of Malayalam Native Speakers