தமிழ் மூலப் பேச்சுத் தரவகம்

Tamil Raw Speech Corpus

Tamil Raw Speech Corpus

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 Language
Mysore, India-570006

CENTRAL INSTITUTE OF INDIAN LANGUAGES
Manasagangotri, Mysuru, Karnataka, India, 570006
www.ciil.org

Title :Tamil Raw Speech Corpus
Authors:Ramamoorthy L., Narayan Kumar Choudhary, Thennarasu S, Prem Kumar L R, Amudha R, Prabagaran R, Srikanth D.

e-ISBN: 978-93-91386-01-6
CIIL Publication No.: 1283

First published: AD 2021 May
Vaisakha 1943 Saka

Publisher: C.G. Venkatesha Murthy, Director, CIIL

Production Team
Head, Publication Unit: Umarani Pappuswamy
Officer-in-Charge, Publication Unit: Aleendra Brahma
Artist: H. Manohara
Staff-in-charge: R. Nandeesh
Compositor: M.N. Chandrashekar
Cover design: D.Srikanth

Contents

Table of Figures

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 study[1] 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 has 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 a subset 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	Conent 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 text was 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

- 'ˈ**pa**llaʊi'- 'pa' inherent vowel at initial position (CV initial)
- '**pra**kaɻa' - 'p' as pure consonant in conjunction with 'ra' in initial position, (CCV Initial)
- '**spa**ndana',- 'pa' with inherent vowel preceded by a consonant at medial position(CCV Initial)
- '**pa**rika**lpa**ne'- '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 sings, 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) Text	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 languages 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 W1to 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, and50+
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	ReadIn Script	The script in which the content has been provided to read in.
9	Recording Environment	A brief info on the environment in which the recording has been done.
10	Power Source	The source of the power using which the recording was done. It may be Li-ion, NiCd or Alkaline batteries.
11	Recorder Manufacturer	Manufacturer of the recorder.
12	Recorder Type	Type of the recorder. It is mostly 24-bit Linear PCM (R-09).
13	Sampling Frequency	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	Educational Qualification	Highest educational qualification of the speaker.
21	Place Of Elementary Education	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	Recording Date	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 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_Tamil_Female_16To20_Contemporary Text-T1_SP-0035_T1-0035.wav"
"LDC-IL_Scheduled_Tamil_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.

5.2 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 cases need 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 TAMIL RAW SPEECH CORPUS

7.1 INTRODUCTION

Tamil is one of the prominent languages among the Dravidian language family. Tamil is widely spoken in the state of Tamil Nadu, Union Territory of Pondicherry, Sri Lanka, in East-Asian countries like Burma, Malaysia, Singapore, Indonesia, Indio china, Fiji, in South-Africa, British Guinea and in islands like Mauritius and Madagascar etc. Within India, it is the official language of the Indian state of Tamil Nadu and the Indian Union Territory of Pondicherry. Tamil also has an official language Sri Lanka and Singapore. The Tamil language uses a script by its own name which has originated from the Brahmi script. The language is highly agglutinative in nature. Compared to Sanskrit and other Indian languages.

Scholars have described six broad categories of geographical dialects. They are Chennai Tamil (Chengalpet, Tiruvallur, Chennai, Kanchipuram, and Pondicherry) Kongu Tamil (Tirupur, Nilgiris, Coimbatore, Erode), Kumari Tamil (Tamil spoken in coastal region of Kanyakumari), Madurai Tamil (Madurai, Sivaganga, Ramanathapuram, Dindigul, Teni and Virudunagar), Nellai Tamil (Tirunelveli and Tuticorin), Salem Tamil (Dharmapuri, Karur, Namakkal and Salem), Thanjai Tamil (Thanjavur Trichy, Pudukottai, Perambaur, Aruiyalur, Nagapattinam, Tiruvarur and Cuddalore) and Vellore Tamil (Vellore, Tiruvannamalai and Viluppuram). Of course, each one of these consists of sub-dialects which have their own distinctive features. Many of these distinctions occur because the dialects are strongly influenced by their neighbouring languages.

Chennai Tamil is a combination of Tamil, Telugu, Kannada, Urdu, and English. Kongu Tamil is a unique variety of Tamil as this dialect is very respectful and uses a unique kind of vocabulary in a respectful manner as standard Tamil. Kumari Tamil will be very tough to understand for someone who speaks Madras Bashai or Kongu Tamil, but is understandable to speakers of Madurai Tamil, Tirunelveli Tamil, or even Malayalam and this involves a lot of tongue rolling i.e., retroflex sounds. Madurai Tamil is also considered one of the purest Tamil dialects spoken in Tamil Nadu along with the Central Tamil dialect. Tirunelveli Tamil is also famously known as Nellai Tamil and is easily understood if someone knows Madurai Tamil well. Salem Tamil and Thanjai Tamil are easily understood among all Tamil speakers

LDC-IL divided the Tamil speaking areas into these eight regions and collected speech data from six regions. Total speech data of 452 speakers, were collected from Kumari, Madurai, Nellai, Salem and Thanjai regions.

8 DATASET PREPARATION FOR TAMIL

For the selected regions, Kongu, Kumari, Madurai, Nellai, Salem and Thanjai. LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	8
Date	2
Command and Control Words	370
Most Frequent Words	1000
Form and Function Words	598
Phonetically Balanced Words	565
Person Name	572
Place Name	348
Sentences	228

Table 6: LDC-IL Speech Dataset

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 sleeted 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*
		*randomly selected by machine

Table 7: Table of Contents in LDC-IL Dataset

The Full Set of

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

were also carried to the field to get recorded by selected individuals. Once all these preparations were made, the investigator started collecting the data.

The Collection of data is carried out in three phases for different regions. First phase carried out in Madurai, Salem, Thanjai region in 2008 by S.Thennarasu. In Kumari, Madurai, Nellai region in 2009 by L R Prem Kumar. In Kongu, Madurai, Salem region from 2009 by R Prabagaran. Some of the data are collected by D Srikanth in 2018

9 TRANSLITERATIONS IN LDC-IL TAMIL READ CORPUS

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

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

LDC-IL Transliteration Scheme
Tamil characters to Roman and Tamil Numerals

Vowels and Vowel Signs

ஆ	இ	ஈ	உ	ஊ	எ	ஏ	ஐ	ஒ	ஓ	ஔ	ஃ
ா	ி	ீ	ு	ூ	ெ	ே	ை	ொ	ோ	ௌ	
A	i	I	u	U	e	E	ai	o	O	au	H

Consonants

க	ங	ச	ஞ	த	ண	ட	ந	ப	ம
ka	ng'a	ca	nj'a	ta	Na	Ta	na	pa	ma

ய	ர	ற	ல	வ	ள	ழ	ன
ya	ra	Ra	la	va	La	Za	n'a

Tamil Grantha Consonants

ஜ	ஷ	ஸ	ஹ
ja	sha	sa	ha

Numerals (Tamil to Hindu-Arabic)

௦	௧	௨	௩	௪	௫	௬	௭	௮	௯
0	1	2	3	4	5	6	7	8	9

10 SUMMARY OF THE CORPUS

In the sections below, we provide the tabular details of the different content types of the Tamil 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 139:11:41 (hh:mm:ss) comprising 60,287 audio segments.

10.1 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Tamil speech dataset.

LDC-IL Tamil 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	433	22	144	48	18	145	56
Creative Text-T2	429	22	143	48	18	143	55
Sentence-S	10764	552	3586	1199	449	3602	1376
Date-D	842	44	279	92	36	283	108
Command and Control Words-W1	12882	660	4291	1406	538	4341	1646
Person Name-W2	8755	457	2907	938	365	2985	1103
Place Name-W2	4002	204	1330	455	159	1310	544
Most Frequent Word-Part-W3A	12813	662	4213	1408	539	4341	1650
Most Frequent Word-FullSet-W3B	2000	0	1000	0	0	1000	0
Phonetically Balanced-W4	3860	564	563	435	991	871	436
Form and Function Word-W5	3507	593	1194	592	0	535	593

Table 8: Audio Segments and their Distribution

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 Tamil 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	58:01:47	03:08:11	21:03:01	04:54:00	02:14:05	21:17:56	05:24:34
Creative Text-T2	14:21:31	00:40:55	04:36:11	01:35:17	00:47:15	05:00:29	01:41:24
Sentence-S	14:51:03	00:46:12	05:01:18	01:24:44	00:37:10	05:15:49	01:45:50
Date-D	01:20:17	00:03:54	00:26:01	00:08:07	00:03:05	00:28:34	00:10:36
Command and Control Words-W1	12:57:06	00:40:59	04:22:40	01:14:16	00:32:53	04:30:28	01:35:50
Place Name-W2	03:57:29	00:11:59	01:19:08	00:24:55	00:09:59	01:19:15	00:32:13
Person Name-W2	10:34:38	00:35:04	03:31:43	01:02:22	00:26:55	03:39:47	01:18:47
Most Frequent Word-Part-W3A	11:14:05	00:35:39	03:44:22	01:02:48	00:29:47	03:58:46	01:22:43
Most Frequent Word-FullSet-W3B	02:26:05	00:00:00	01:14:45	00:00:00	00:00:00	01:11:20	00:00:00
Phonetically Balanced-W4	04:55:10	00:45:27	00:38:05	00:31:13	01:11:46	01:09:15	00:39:24
Form and Function-Word-W5	04:40:29	00:53:17	01:37:21	00:47:00	00:00:00	00:43:42	00:39:09

Table 9: Duration of the Collected Data

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 The Contemporary Text (News) - T1

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	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	38	9	6	62	46	108
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	0	25	23	12	7	39	30	69
Nellai Tamil	3	5	13	14	8	14	24	33	57
Salem Tamil	0	2	3	15	15	25	18	42	60
Thanjai Tamil	10	5	46	46	2	2	58	53	111

Table 10: Distribution of Contemporary Text (News) Data

10.4 RANDOM SET

The Random Set data comprises 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 The Creative Text-T2

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	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	37	9	6	62	45	107
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	0	23	23	12	7	37	30	67
Nellai Tamil	3	5	13	13	8	14	24	32	56
Salem Tamil	0	2	3	15	15	24	18	41	59
Thanjai Tamil	10	5	47	46	2	2	59	53	112

Table 11: Distribution of Tamil Creative Text

10.4.2 The Date-D

The answer to one randomly selected question from the list of 2 questions to get the date format of the informants. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	10	4	92	74	16	12	118	90	208
Kumari Tamil	4	8	18	18	4	4	26	30	56
Madurai Tamil	4	0	47	44	24	14	75	58	133
Nellai Tamil	6	10	24	25	16	26	46	61	107
Salem Tamil	0	4	6	30	30	48	36	82	118
Thanjai Tamil	20	10	92	92	2	4	114	106	220

Table 12: Distribution of Tamil Date Format

10.4.3 The Sentences-S

The Sentences contain a list of sentences that is a representation of all most all the phonemes occurring in Tamil. 25 Randomly selected Sentences are recorded from a list of 228 sentences. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	127	49	1186	950	224	150	1537	1149	2686
Kumari Tamil	50	100	225	225	50	50	325	375	700
Madurai Tamil	50	0	600	575	300	177	950	752	1702
Nellai Tamil	75	125	324	326	200	350	599	801	1400
Salem Tamil	0	50	75	374	375	599	450	1023	1473
Thanjai Tamil	250	125	1176	1152	50	50	1476	1327	2803

Table 13: Distribution of Tamil Sentences

10.4.4 Command and Control Words-W1

The Command and Control Words contain a list of 370 words that is a representation of all most all the command and control words occurring in Tamil. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	148	59	1409	1139	240	180	1797	1378	3175
Kumari Tamil	60	120	270	270	60	60	390	450	840
Madurai Tamil	60	0	720	691	360	210	1140	901	2041
Nellai Tamil	90	150	361	391	240	420	691	961	1652
Salem Tamil	0	60	90	449	446	716	536	1225	1761
Thanjai Tamil	302	149	1441	1401	60	60	1803	1610	3413

Table 14: Distribution of Tamil Command and Control Words

10.4.5 Person Names –W2

The Person Names contain a list of 572 popular Pan Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	100	38	930	760	159	120	1189	918	2107
Kumari Tamil	40	80	181	180	40	40	261	300	561
Madurai Tamil	40	0	490	475	240	141	770	616	1386
Nellai Tamil	60	100	256	261	159	280	475	641	1116
Salem Tamil	0	40	61	301	296	479	357	820	1177
Thanjai Tamil	217	107	989	1008	44	43	1250	1158	2408

Table 15: Distribution of Tamil Person Names

10.4.6 Place Names-W2

The Place Names contain a list of 347 popular Pan Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Kongu Tamil	50	20	458	380	80	60	588	460	1048
Kumari Tamil	20	40	90	90	20	20	130	150	280
Madurai Tamil	20	0	230	216	120	72	370	288	658
Nellai Tamil	30	40	131	130	76	136	237	306	543
Salem Tamil	0	20	29	149	143	239	172	408	580
Thanjai Tamil	84	39	392	345	16	17	492	401	893

Table 16: Distribution of Tamil Place Names

10.4.7 Most Frequent Words-PART-W3A

The Most Frequent Words-part contains a list of 1,000 most frequent words. 30 randomly selected words are recorded from the list. The distribution of data is as follows:

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Kongu Tamil	152	60	1373	1137	239	180	1764	1377	3141
Kumari Tamil	60	120	270	271	60	60	390	451	841
Madurai Tamil	60	0	719	690	360	210	1139	900	2039
Nellai Tamil	90	149	391	408	240	420	721	977	1698
Salem Tamil	0	60	90	450	449	720	539	1230	1769
Thanjai Tamil	300	150	1370	1385	60	60	1730	1595	3325

Table 17: Distribution of Tamil Most Frequent Words – Part

10.5 FULL SET

The Full Set is the master set of certain data set which is read completely from few selected speakers in each group. Full sets are as below.

10.5.1 Most Frequent Words-Full-W3B

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

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Thanjai Tamil	0	0	1000	0	0	1000	1000	1000	2000

Table 18: Distribution of Tamil Most Frequent Words – Full

10.5.2 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where all most all the phones of Tamil language has occurred in all the possible positions of a word. In full set all the 565 words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender	Female	Male	Female	Male	Female	Male			
Thanjai Tamil	564	991	563	871	435	436	1562	2298	3860

Table 19: Distribution of Tamil Phonetically Balanced Vocabulary

10.5.3 The Form and Function Words-W5

The Form and Function Words contain a list of 598 words which is a representation of all most all the form and function words occurring in Tamil. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Salem Tamil	0	0	598	0	0	0	598	0	598
Thanjai Tamil	593	0	596	535	592	593	1781	1128	2909

Table 20: Distribution of Form and Function words

10.6 SPEAKER DISTRIBUTIONS

The following table shows the distribution of native speakers of Tamil, across different regions.

Age Group →	16 to 20		21 to 50		Above 51		Total Female	Total Male	Total
Gender →	Female	Male	Female	Male	Female	Male			
Kongu Tamil	5	2	48	38	9	6	62	46	108
Kumari Tamil	2	4	9	9	2	2	13	15	28
Madurai Tamil	2	23	25	0	12	7	39	30	69
Nellai Tamil	3	5	13	14	8	14	24	33	57
Salem Tamil	0	2	4	15	15	25	19	42	61
Thanjai Tamil	12	7	51	51	4	4	67	62	129

Table 21: Distribution of Tamil Native Speakers