অসমীয়া মৌলিক বাক ভাষাংশ

অ স মী য়া

Assamese 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: Assamese Raw Speech Corpus

Authors: Ramamoorthy L., Narayan Kumar Choudhary, Atreyee Sharma, Ashmrita Gogoi, Jahnobi Kalita, Samhita Bharadwaj, Plabita Bora, Priyanshe Adhyapak, Mustafiza Tamim, Rajesha N., Manasa G.

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: Rupesh Rai

Contents

Table of Figures

1 LDC-IL RAW SPEECH CORPORA: AN OVERVIEW

INTRODUCTION

Lack of basic linguistic resources has 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 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

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 is 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.

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.

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.

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?

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.

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.

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.

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.

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

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.

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.

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.

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

- '**pa**llaʋi'- 'pa' inherent vowel at initial position (CV initial)
- '**pra**kaʈa' - '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)
- 'aːp**ta**' - '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.

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

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 Assamese 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), it is 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.

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.

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

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

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.

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 Assamese speakers may be Tulu, Konkani, Chitpavani etc. This influences the articulation of Assamese speech in these areas.

Place: Place gives better information about the speech data collected. For example, Assamese 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	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 profess 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 elementary education. This usually corresponds to the early childhood experiences which happen 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

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.

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   ORGANIZING 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.

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 below.

"LDC-IL_Scheduled_Assamese_Female_16To20_Contemporary Text-T1_SP-0035_T1-0035.wav"
"LDC-IL_Scheduled_Assamese_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.

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 or 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 ASSAMESE RAW SPEECH CORPUS

7.1 INTRODUCTION

Assamese, an official language of Assam and is a language that belongs to the Indo-Aryan language family. Its linguistic presence is widely presented in the state of Assam and some parts of Arunachal Pradesh and Nagaland. According to 2011 census, the Assamese Language is spoken by 15 million speakers. It is said to be developed from Kamrupi Prakrit and Old Assamese. At present the script used for Assamese writing is Assamese which is shared also with Bengali and a few other languages.

The Assamese phonemic inventory consists of total forty-one phonemes that include eight vowels, ten diphthongs and twenty three consonants. The velar fricative /x/ is known as the hallmark of the Assamese language. The language has a complex consonant cluster system. The morphological salience of Assamese lies on its rich case marking system. Known as a tripartite split ergative language, Assamese exhibits the subject+object+verb sentence structure. However, in some contexts, it shows some level of flexibility in its word order. Unlike many other Indo-Aryan languages of India, grammatical gender distinction is not present in Assamese. In this aspect and many others for that matter, Assamese shares many similarities with the Bengali Language.

Assamese a widely spoken language does encounter several dialectal variations. The regional dialects can be broadly divided into two parts - the Eastern Group and the Western Group. The Eastern Group essentially includes the region in and around Sivasagar District. This group is relatively homogenous in nature. On the other hand, the Western Group includes the Western region of the state. This group exhibits heterogeneity. According to some recent linguistic studies, the two dialectal groups can be further divided into four different varieties namely Eastern, also known as Xiboxagoria, Central, Kamrupi and Goalparia Group. Here, the Central Group is spoken in Nagaon, Sonitpur, Morigaon and other adjoining areas; the Kamrupi variety is spoken in Kamrup, Barpeta, Nalbari etc.; and the Goalparia Group is spoken in the Goalpara region.

LDC-IL divided the Assamese speaking areas into these four regions and have collected speech data from each speaker. Following are some of the regions from where the LDC-IL Assamese Speech Data is collected and is listed in the table below:

Region→	Xiboxagoria	Central Assam	Kamrupi	Goalparia
Places →	1. Sivasagar 2. Golaghat 3. Jorhat 4. Dibrugarh 5. Tinsukia	1. Sonitpur 2. Nagaon 3. Morigaon 4. Lakhimpur 5. Dhemaji 6. Darrang	1. Guwahati 2. Kamrup Rural 3. Nalbari 4. Baksa 5. Barpeta	1. Goalpara 2. Dhubri

Table 6: Regions and Places Covered for Assamese Speech Data

8   DATASET PREPARATION FOR ASSAMESE

For the selected regions, Xiboxagoria, Central Assam, Kamrupi and Goalparia LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Content Type	Count
Creative Text	9
Date	2
Command and Control Words	254
Form and Function Words	267
Phonetically Balanced Words	439
Person Name	500
Place Name	341
Sentences	200

Table 7: 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 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*
		*randomly selected by machine

Table 8: Table of Contents in LDC-IL Dataset

The Full Set of
1. Phonetically Balanced Vocabulary
2. Form and Function 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 by Atreyee Sharma in 2008 and 2009 in two different phases in the field for different regions and Plabita Bora brought in 6 speakers into the corpus in 2019.

9   TRANSLITERATIONS IN LDC-IL ASSAMESE READ

CORPUS

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

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

LDC-IL Transliteration Schema
Assamese characters to Roman and Assamese Numerals to Hindu-Arabic

Vowels and Vowel Signs										
অ	আ	ই	ঈ	উ	ঊ	ঋ	এ	ঐ	ও	ঔ
	া	ি	ী	ু	ূ	ৃ	ে	ৈ	ো	ৌ
A	A	i	I	u	U	x	E	ai	O	au

Consonants				
ক	খ	গ	ঘ	ঙ
Ka	kha	ga	gha	ng'a
চ	ছ	জ	ঝ	ঞ
Ca	Cha	ja	jha	nj'a
ট	ঠ	ড	ঢ	ণ
Ta	Tha	Da	Dha	Na
ত	থ	দ	ধ	ন
Ta	Tha	da	dha	na
প	ফ	ব	ভ	ম
Pa	pha	ba	bha	ma

Symbols		
ং	ঃ	ঁ
M	H	m'

য	ৰ	ল	ৱ	শ	স	ষ	হ	ড়	ঢ়	য়	ৎ
Ya	ra	La	wa	sha	sa	Sa	ha	D'a	Dh'a	Ya	t

Numerals(Bengali to Hindu-Arabic)									
০	১	২	৩	৪	৫	৬	৭	৮	৯
0	1	2	3	4	5	6	7	8	9

.

10 SUMMARY OF THE CORPUS

In the sections below, we provided the tabular details of the different content types of the Assamese raw speech corpus based on various yardsticks which can also be filtered out from the metadata sheet. These figures may be helpful in tuning the corpus for various purposes of the training, testing, and evaluating various algorithms as well as providing useful insights into the datasets. The total duration of the dataset is 54:21:12 (hh:mm:ss) comprising 37,570 audio segments.

10.1 SUMMARY OF THE AUDIO SEGMENTS

The table below shows the total number of Audio Segments and their distribution in the Assamese speech datas17et.

LDC-IL Assamese 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	304	16	97	41	16	94	40
Creative Text-T2	304	16	97	41	16	94	40
Sentence-S	7593	399	2423	1025	398	2349	999
Date-D	599	32	192	80	32	187	76
Command and Control Words-W1	9118	479	2910	1228	479	2823	1199
Person Name-W2	6081	320	1938	819	319	1886	799
Place Name-W2	3044	160	969	410	160	945	400
Phonetically Balanced-W4	6567	875	2188	876	875	875	878
Form and Function Word-W5	3960	528	1319	528	527	529	529

Table 9: 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 Assamese 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	17:23:25	0:50:39	05:18:38	2:24:40	0:56:35	5:26:53	2:26:00
Creative Text-T2	11:44:37	0:38:07	3:32:03	1:35:24	0:41:10	3:42:50	1:35:03
Sentence-S	5:55:29	0:19:56	1:49:53	0:47:06	0:19:49	1:51:35	0:47:13
Date-D	0:33:59	0:01:29	0: 10:57	0:05:08	0:01:21	0:10:13	0:04:51
Command and Control Words-W1	4:56:49	0:15:17	1:34:47	0:40:32	0:13:27	1:30:53	0:41:53
Person Name-W2	5:38:07	0:18:45	1:47:25	0:46:42	0:16:37	1:43:32	0:45:06
Place Name-W2	1:58:33	0:06:08	0: 37:27	0:16:28	0:05:54	0:36:27	0:16:09
Phonetically Balanced-W4	3:41:45	0:33:34	1:25:12	0:27:39	0:20:59	0:25:33	0:28:48
Form and Function-Word-W5	2:28:28	0:25:04	0:54:02	0:19:12	0:13:37	0:17:21	0:19:12

Table 10: 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 the each speaker.

10.3.1 The Contemporary Text (News) - T1

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	32	16	16	8	8	7	3	1	5	0	0
21 to 50	191	97	94	40	38	34	29	23	26	0	1
50+	81	41	40	14	13	15	15	10	9	2	3
Total	304	154	150	62	59	56	47	34	40	2	4

Table 11: Distribution of Assamese Contemporary Text (News) Data

10.4 RANDOM SET

The Random Set 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

The prepared dataset consists of 9 literary pieces. From this dataset, any one randomly selected text was recorded from each informant to get the speech data of creative text. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
		Female	Male	Xiboxagoria		Kamrupi		Central Assam		Goalparia	
				Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	32	16	16	8	8	7	3	1	5	0	0
21 to 50	191	97	94	40	38	34	29	23	26	0	1
50+	81	41	40	14	13	15	15	10	9	2	3
Total	304	154	150	62	59	56	47	34	40	2	4

Table 12: Distribution of Assamese Creative Text

10.4.2 The Date-D

The answers to 2 questions listed in the dataset to get the date format of the informants. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
		Female	Male	Xiboxagoria		Kamrupi		Central Assam		Goalparia	
				Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	64	32	32	16	16	14	6	2	10	0	0
21 to 50	379	192	187	78	76	68	58	46	51	0	2
50+	156	80	76	28	26	30	28	18	16	4	6
Total	599	304	295	122	118	112	92	66	77	4	8

Table 13: Distribution of Assamese Date Format

10.4.3 The Sentences-S

The Sentences contain a list of 200 sentences that is a representation of almost all the phonemes occurring in Assamese. 25 Randomly selected sentences are recorded from each speaker. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
		Female	Male	Xiboxagoria		Kamrupi		Central Assam		Goalparia	
				Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	797	399	398	199	200	175	75	25	123	0	0
21 to 50	4772	2423	2349	999	948	850	721	574	655	0	25
50+	2024	1025	999	350	325	375	374	250	225	50	75
Total	7593	3847	3746	1548	1473	1400	1170	849	1003	50	100

Table 14: Distribution of Assamese Sentences

10.4.4 Command and Control Words-W1

The Command and Control Words contain a list of 254 words that is a representation of almost all the command and control words occurring in Assamese. 30 randomly selected words are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	958	479	479	239	240	210	90	30	149	0	0
21 to 50	5733	2910	2823	1200	1139	1020	870	690	784	0	30
50+	2427	1228	1199	418	390	450	449	300	270	60	90
Total	9118	4617	4501	1857	1769	1680	1409	1020	1203	60	120

Table 15: Distribution of Assamese Command and Control Words

10.4.5 Person Names –W2

The Person Names contain a st of 500 popular pan- Indian and regional person names. 20 randomly selected names are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	639	320	319	160	159	140	60	20	100	0	0
21 to 50	3824	1938	1886	799	760	680	580	459	526	0	20
50+	1618	819	799	279	260	300	300	200	179	40	60
Total	6081	3077	3004	1238	1179	1120	940	679	805	40	80

Table 16: Distribution of Assamese Person Names

10.4.6 Place Names-W2

The Place Names contain a list of 341 popular pan- Indian and regional place names. 10 randomly selected names are recorded from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	320	160	160	80	80	70	30	10	50	0	0
21 to 50	1914	969	945	400	378	339	290	230	267	0	10
50+	810	410	400	140	130	150	150	100	90	20	30
Total	3044	1539	1505	620	588	559	470	340	407	20	40

Table 17: Distribution of Assamese Place Names

10.5 FULL SET

The Full Set is the master set of a certain data set which is read completely from a few selected speakers in each group. Full Sets are given below:

10.5.1 The Phonetically Balanced Vocabulary-W4

The Phonetically Balanced Vocabulary contains a list of words where almost all the phones of Assamese, have occurred in all the possible positions of a word. In full set all the 439 words are recorded from the informants and these words are uttered by them three times from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Xiboxagoria		Central Assam	
		Female	Male	Female	Male	Female	Male
16 to 20	1750	875	875	875	875	0	0
21 to 50	3063	2188	875	1749	875	439	0
50+	1754	876	878	876	0	0	878
Total	6567	3939	2628	3500	1750	439	878

Table 18: Distribution of Assamese Phonetically Balanced Vocabulary

10.5.2 The Form and Function Words-W5

The Form and Function Words contain a list of 266 words which is a representation of almost all the form and function words occurring in Assamese. All the words are recorded from the informants and words are uttered by them three times from the list. The distribution of the data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Xiboxagoria		Central Assam	
		Female	Male	Female	Male	Female	Male
16 to 20	1055	528	527	528	527	0	0
21 to 50	1848	1319	529	1057	529	262	0
50+	1057	528	529	528	0	0	529
Total	3960	2375	1585	2113	1056	262	529

Table 19: Distribution of Assamese Form and Function words

10.6 NATIVE SPEAKERS DISTRIBUTIONS

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

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Regions							
				Xiboxagoria		Kamrupi		Central Assam		Goalparia	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 to 20	34	17	17	9	9	7	3	1	5	0	0
21 to 50	196	101	95	43	39	34	29	24	26	0	1
50+	83	42	41	15	13	15	15	10	10	2	3
Total	313	160	153	67	61	56	47	35	41	2	4

Table 20: Distribution of Assamese Native Speakers

10.7 MOTHER TONGUE DISTRIBUTION OF THE NATIVE SPEAKERS

The following table shows the distribution of Mother Tongue of the Assamese native speakers in LDC-IL speech data.

Mother Tongue of the Native Speaker	Geographical Dialect Distribution of LDC-IL Assamese Speech Corpus				Total speaker
	Xiboxagoria	Kamrupi	Central Assam	Goalparia	
Assamese	125	101	73	6	305
Bodo	1	-	1	-	2
Miri-Mishing	-	1	1	-	2
Bengali	-	1	-	-	1
Maithili	1	-	-	-	1
Hindi	1	-	-	-	1
Punjabi	-	-	1	-	1
Total	128	103	76	6	313

Table 21: Representation of Mother Tongue Distribution of the Assamese Native Speakers