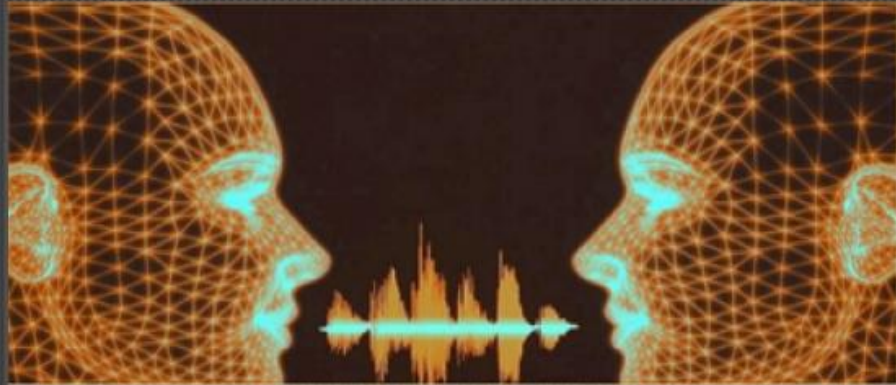


हिंदी मौलिक वाक् कॉर्पस



Hindi Raw Speech Corpus

Hindi Raw Speech Corpus

Authors:

Dr. Satyaendra Awasthi

Dr. Madhupriya Pathak

Mr. Rajesha N.

Ms. Manasa G.

Dr. Narayan Choudhary

Dr. L. Ramamoorthy



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

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

Hindi Raw Speech Corpus

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

For Further contacts: oic@ldcil.org

All rights reserved

© Central Institute of Indian Languages, Mysore-2018

Authors:

Dr. Satyaendra Awasthi

Dr. Madhupriya Pathak

Mr. Rajesha N

Ms. Manasa G

Dr. Narayan Choudhary

Dr. L. Ramamoorthy

ISBN 978-81-7343-221-7 Electronic publication of Corpus

Last updated: November 15, 2018

Publisher:

Linguistic Data Consortium for Indian Languages

Central Institute of Indian Languages

Mysore, India-570006

Contents

1	LDC-IL Raw Speech Corpora: An Overview	1
1.1	Introduction	1
1.2	LDC-IL Speech Corpus	2
1.3	T1: Contemporary Text	3
1.4	T2: Creative Text	4
1.5	D: Date	4
1.6	S: Sentences	4
1.7	W1: Command and Control Words.....	5
1.8	W2: Proper Noun (Person Names and Place Names).....	5
1.8.1	Person Names	5
1.8.2	Place Names.....	5
1.9	W3: Most Frequent Words	6
1.9.1	W3A: Most Frequent Words-Part	6
1.9.2	W3B: Most Frequent Words-Full	6
1.10	W4: Phonetically Balanced Vocabulary	6
1.11	W5: Form and Function words	7
2	Planning for Fieldwork	8
2.1	Dataset preparation and distribution	8
3	Field Work.....	10
3.1	Possible places for collecting data	10
3.2	Field work Ethics	10
3.3	Data Collection.....	11
3.3.1	Technical Specifications for collecting data	11
3.3.2	Metadata.....	12
3.3.3	Data Transferring and Storing.....	15
3.3.4	Observations	15
4	Organising and Archiving the Data	16
4.1.1	Text - Speech Mapping and Naming Conventions	16
4.1.2	Observations	16
5	Data verification and Quality Control	17
6	Hindi Raw Speech Corpus	18

6.1	Introduction	18
7	Dataset preparation for Hindi.....	19
7.1.1	Model 1 (T1, T2, S, D, W1, W2, W3, W4, W5).....	19
7.1.2	Model 2 (T1, T2).....	20
8	Transliterations in LDC-IL Hindi Read corpus.....	21
9	Summary of the Corpora.....	22
9.1	Summary of the Audio Segments	22
9.2	Duration of the Raw Speech Data.....	23
9.3	Distinct Set	23
9.3.1	Contemporary Text (News) T-1.....	23
9.3.2	The Creative Text-T2	24
9.4	Random Set.....	24
9.4.1	The Creative Text-T2	24
9.4.2	The Date-D	24
9.4.3	Sentences-S.....	25
9.4.4	Command and Control Words-W1	25
9.4.5	Person Names-W2	25
9.4.6	Place Names-W2	26
9.4.7	Most Frequent Words-W3A.....	26
9.5	Full Set.....	26
9.5.1	Most Frequent Words-Full-W3B.....	26
9.5.2	Phonetically Balanced Vocabulary-W4	27
9.5.3	Form and Function Words-W5.....	27
9.6	Native Speakers Distributions.....	27
9.7	Mother Tongue Distributions of Informants.....	28

Table of Figures

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

1 LDC-IL RAW SPEECH CORPORA: AN OVERVIEW

1.1 INTRODUCTION

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

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

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

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

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

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

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.

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

Table 1: LDC-IL Speech Data Content Types

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

2.1 T1: CONTEMPORARY TEXT

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

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

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

2.2 T2: CREATIVE TEXT

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

Creative 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 preceding phoneme. The pronunciation of initial position is different from middle and final positions. For example the phoneme ‘pa’ is used in different forms while pronouncing words like

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

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

2.9 W5: FORM AND FUNCTION WORDS

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

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

3 PLANNING FOR FIELDWORK

3.1 DATASET PREPARATION AND DISTRIBUTION

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

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

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

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

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

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

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

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

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

4 FIELD WORK

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

4.1 POSSIBLE PLACES FOR COLLECTING DATA

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

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

4.2 FIELD WORK ETHICS

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

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

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

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

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

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

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

4.3 DATA COLLECTION

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

4.3.1 Technical Specifications for collecting data

Recording Setup:	Sample Rate : 48.0 KH
Recording Mode:	wav -16bit
Date Setup:	Current date and time.
Storage:	SD Card
Power:	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Keep minimum 5 cm to 25 cm distance between the microphone and the speaker and if possible use microphone holder. • The recorder should not be placed orthogonally but it should be placed diagonally. • Do not move the recorder during recording • Fix the recorder upon a table/ fixed plane if possible. • Try to have fixed the distance between the recorder and speaker • The recorder should not be placed orthogonally but it should be placed diagonally

Table 4: Technical Specifications for collecting data

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

4.3.2 Metadata

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

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

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

Informant Data:

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

Investigator Data:

Name:
Date:
Place:
Region:
Environment:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 5: Metadata Legends and their Description

4.3.3 Data Transferring and Storing

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

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

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

4.3.4 Observations

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

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

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

5 ORGANISING AND ARCHIVING THE DATA

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

5.1.1 Text - Speech Mapping and Naming Conventions

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

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

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

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

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

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

5.1.2 Observations

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

6 DATA VERIFICATION AND QUALITY CONTROL

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

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

- The Audio recording of each speaker is checked against the mapped text.
- Each distinct text audio recording will be matched with automated entries of the same speaker to check for any mis-mapping of speaker.
- Metadata like Gender, age group, District etc. are selective part of manual data entry and could be prone to errors so verification is needed.
- Metadata like Dialect entry, place, etc. are keyed in by manual data entry and could be prone to errors like typo errors so verification needs to be done.
- The audio segments could be duplicated because of system/network errors and need to be checked.
- At the time of data segmentation, one might have saved the whole file instead of selected part. Such 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 HINDI RAW SPEECH CORPUS

7.1 INTRODUCTION

Hindi is an Indo-Aryan language, a descendent of *Sanskrit*, which is spoken in the central and northern India, in the states of *Bihar, Chhattisgarh, Delhi, Haryana, Himachal Pradesh, Jharkhand, Madhya Pradesh, Rajasthan, Uttarakhand* and *Uttar Pradesh*. It is the official language of the Union of India and is also *lingua franca* across India. Being the most intelligible language of India, it is currently reported to be spoken as the first language by 528.35 million people in India (as per 2011 census of India) i.e. a total of 43.63% of the populace of India speaks Hindi as their primary language.

According to the constitution of India the official languages are written in the Devanagari and English. Based on the provisions mentioned in the Official Language Act, Hindi is used for official activities such as communications between the Central Government and a State Government, judiciary and parliamentary proceedings.

Hindi is written in Devanagari script, a Left to Right script which is a descendent of *Brahmi* script. The script is also used to write several other languages of India and neighboring countries such as Nepali, Marathi, and Maithili etc.

Many ethnolects, sociolects and other varieties of Hindi are in practice in the Hindi-belt, which are very different from each other in terms of phonological nuances and structural features. However, only major varieties which are intelligible throughout the Hindi plain due to the variety continuum, and are accepted as the dialects/sister languages of Hindi by the academia are considered in the fieldwork.

As claimed by various linguists of repute, Hindi can be divided into following varieties based on the phonological nuances:

#	Regions	Dialects
1.	Eastern Hindi	Awadhi, Bagheli, Chhattisgarhi
2.	Western Hindi	Bangru, Braj, Bundeli, Haryanvi, Kannauji, Khariboli
3.	Bihari Hindi	Bhojpuri, Magahi
4.	Rajasthani Hindi	Marwadi, Mewati, Malvi, Dhundhandi
5.	Pahari Hindi	Pahari (Himachal Pradesh), Kumaoni, Garhwali

Table 6: Hindi Regions and Dialects

LDC-IL divided the Hindi speaking areas into these five regions and collected speech data from three regions. Hindi speech corpus is representative in the terms that it reflects all the nuances of dialects it was collected from. Many phonological nuances such as no-distinction in pronunciation of sibilants in ‘Awadhi speakers’ speech, simplification of consonant clusters in ‘Bihari speakers’ speech can be easily seen across the corpus.

8 DATASET PREPARATION FOR HINDI

Three fieldworks were conducted in Eastern Hindi and Bihari Hindi regions. Speech data comprising of 434 speakers was collected in these fieldworks. The fieldworks covered the areas as follows.

Eastern Hindi (Awadhi belt)

- Allahabad, UP
- Balrampur, UP
- Lucknow, UP
- Sitapur, UP

Bihari Hindi (Bhojpuri belt and Magahi belt)

- Bhabua, Bihar
- Bhojpur, Bihar
- Buxar, Bihar
- Chandauli, UP
- Gaya, Bihar
- Ghazipur, UP
- Gopalganj, Bihar
- Gorakhpur, UP
- Jehanabad, Bihar
- Kaimur, Bihar
- Nalanda, Bihar
- Varanasi, U

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

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

8.1.1 Model 1 (T1, T2, S, D, W1, W2, W3, W4, W5)

For the Regions of Awadhi belt, Bhojpuri belt and Magahi belt LDC-IL prepared the following dataset by which the prompt sheets were prepared.

Notation	Content Type	Count
T1	Contemporary Text (News)	400
T2	Created Text	6
D	Date	2
S	Sentences	500
W1	Command and Control Words	250
W2	Person Name	500
W2	Place Name	324
W3	Most Frequent Words	1000
W4	Phonetically Balanced Words	800
W5	Form and Function Words	200

Table 7: Content Types of Dataset Model-1

The command and control word list also includes such form and function words which is not covered in form and function word list. 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 typical prompt sheet had a distinct news item and selected part of the dataset as follows.

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

* selected by machine

Table 8: Prompt Sheet Distribution of Contents

The full set of

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

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

8.1.2 Model 2 (T1, T2)

For the fieldwork of Delhi-NCR region (Khariboli belt), LDC-IL attempted a different approach of dataset preparation, and tried to concentrate on capturing more distinct continues text recordings rather than word segments. The prompt sheet for Delhi-NCR was prepared as follows.

Content Type	Content in Each typical prompt sheet	Content selection type
News Text	1Text	Distinct Text
Created Text	1 text	Distinct Text

Table 9: Content Types of Dataset Model-2

Once all these preparations are made, the investigator started collecting the data. The Collection of data is carried out in four different regions as follows.

Region/Place	Year of data collection	Resource Person
Uttar Pradesh	2008	Anjali Sinha
Uttar Pradesh	2008	Jitendra Kumar Singh
Bihar	2008	Dheeraj Kumar Mishra
Delhi-NCR	2012	Satyaendra Awasthi

Table 10: Data Collection Fieldwork Details

9 TRANSLITERATIONS IN LDC-IL HINDI READ CORPUS

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

The LDC-IL transliteration scheme of Hindi (in Devanagari scripts) to Roman is given below.

LDC-IL Transliteration Schema											
Hindi-Devanagari characters to Roman and Hindi Numerals to Hindu-Arabic											
Vowels and Vowel Signs											
अ	आ	इ	ई	उ	ऊ	ऋ	ए	ऐ	ओ	औ	ऑ
	ा	ि	ी	ु	ू	ृ	े	ै	ो	ौ	ँ
a	A	i	I	u	U	x	E	ai	O	au	ao
Consonant					Signs						
क्	ख्	ग्	घ्	ङ्				ं	ः	ँ	
k	kh	g	gh	ng'				M	H	m'	
च्	छ्	ज्	झ्	ञ्		क	ख	ग	ज	फ	
c	ch	j	jh	nj'		k'a	Kh'a	g'a	j'a	ph'a	
ट्	ठ्	ड्	ढ्	ण्	ड्	ढ्					
T	Th	D	Dh	N	D'	Dh'					
त्	थ्	द्	ध्	न्							
t	th	d	dh	n							
प्	फ्	ब्	भ्	म्							
p	ph	b	bh	m							
य्	र्	ल्	व्	श्	ष्	स्	ह्				
y	r	l	v	sh	S	s	h				
Numerals (Devanagari to Roman mapping)											
०	१	२	३	४	५	६	७	८	९		
0	1	2	3	4	5	6	7	8	9		

10 SUMMARY OF THE CORPORA

In the sections below, we provide the tabular details of the different content types of the Hindi 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 Speech data has 70692 Audio segments with the duration of 118:40:03 (hh:mm:ss)

10.1 SUMMARY OF THE AUDIO SEGMENTS

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

LDC-IL Hindi 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	457	29	161	26	42	137	62
Created Text-T2	463	31	163	24	41	140	64
Date-D	764	29	277	45	55	238	120
Sentence-S	10182	375	3747	627	697	3125	1611
Command and Control Words- W1	12281	450	4605	771	866	3714	1875
Person Name-W2	8171	278	3058	517	559	2498	1261
Place Name-W2	4085	140	1524	260	279	1253	629
Most Frequent Word-Part-W3A	12315	449	4626	760	840	3756	1884
Most Frequent Word-FullSet- W3B	6994	0	2000	2000	0	1994	1000
Phonetically Balanced-W4	11986	0	3994	3200	0	1598	3194
Form and Function Word- W5	2994	0	999	795	0	400	800

Table 11: 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 Hindi 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	37:22:29	3:51:19	12:22:27	1:42:41	4:35:13	10:46:15	4:04:34
Created Text-T2	27:03:47	3:48:01	6:36:49	1:03:10	4:05:24	8:23:46	3:06:37
Date-D	0:47:01	0:01:55	0:16:21	0:03:16	0:02:48	0:14:23	0:08:18
Sentence-S	8:41:40	0:19:05	3:14:14	0:32:46	0:35:22	2:36:04	1:24:09
Command and Control Words-W1	8:34:41	0:16:56	3:12:46	0:30:59	0:33:56	2:40:02	1:20:02
Person Name-W2	9:55:25	0:20:42	3:42:04	0:38:05	0:38:23	3:04:13	1:31:58
Place Name-W2	3:14:44	0:06:31	1:11:12	0:12:10	0:12:38	1:01:21	0:30:52
Most Frequent Word-Part-W3A	8:09:10	0:16:45	3:01:19	0:29:06	0:31:59	2:34:13	1:15:48
Most Frequent Word-FullSet-W3B	4:30:14	0:00:00	1:23:23	1:18:39	0:00:00	1:10:44	0:37:28
Phonetically Balanced-W4	8:23:43	0:00:00	3:32:13	2:00:22	0:00:00	0:59:51	1:51:17
Form and Function Word-W5	1:57:09	0:00:00	0:47:30	0:28:53	0:00:00	0:13:57	0:26:49

Table 12: Duration of the Collected Data

10.3 DISTINCT 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

10.3.1 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	Total Audio Segments	Gender-wise Distribution of text		Region-wise Distribution							
				Awadhi belt		Bhojpuri belt		Magahi belt		Khariboli belt	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	71	29	42	0	13	8	8	6	7	15	14
21 To 50	298	161	137	51	38	49	46	49	39	12	14
50+	88	26	62	3	19	18	21	5	22	0	0
Total	457	216	241	54	70	75	75	60	68	27	28

Table 13: Distribution of Contemporary Text (News) Data

10.3.2 The Creative Text-T2

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

Age Group	Total Audio Segments	Gender-wise Distribution of text Khariboli belt	
		Female	Male
16 To 20	29	15	14
21 To 50	26	12	14
Total	55	27	28

Table 14: Distribution of Creative Text (Distinct)

10.4 RANDOM SET

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

10.4.1 The Creative Text-T2

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

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt (Random Set)		Bhojpuri belt (Random Set)		Magahi belt (Random Set)	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	43	16	27	0	14	8	7	8	6
21 To 50	277	151	126	51	38	49	46	51	42
50+	88	24	64	3	20	18	21	3	23
Total	408	191	217	54	72	75	74	62	71

Table 15: Distribution of Created Text (Random)

10.4.2 The Date-D

To get the varieties of date formats of the native speakers, two questions were answered. The distribution of data is as follows:0

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	84	29	55	0	25	16	16	13	14
21 To 50	515	277	238	101	76	98	92	78	70
50+	165	45	120	6	40	34	42	5	38
Total	764	351	413	107	141	148	150	96	122

Table 16: Distribution of Date Format

10.4.3 Sentences-S

The sentences content type contains a list of sentences that is a representation of almost all the phonemes occurring in Hindi. 25 Randomly selected sentences are recorded from a list of 500 sentences. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1072	375	697	0	347	200	200	175	150
21 To 50	6872	3747	3125	1280	949	1223	1150	1244	1026
50+	2238	627	1611	75	505	450	525	102	581
Total	10182	4749	5433	1355	1801	1873	1875	1521	1757

Table 17: Distribution of Sentences

10.4.4 Command and Control Words-W1

The command and control words content type contains a list of 250 words that is a representation of almost all the command and control words occurring in Hindi. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1316	450	866	0	417	240	239	210	210
21 To 50	8319	4605	3714	1523	1137	1469	1379	1613	1198
50+	2646	771	1875	90	564	540	630	141	681
Total	12281	5826	6456	1613	2118	2249	2248	1964	2089

Table 18: Distribution of Command and Control words

10.4.5 Person Names-W2

The person name contains a list of 500 popular Pan Indian and regional person name. 20 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	877	278	599	0	260	160	160	118	139
21 To 50	5556	3058	2498	1022	759	980	921	1056	818
50+	1778	517	1261	60	378	360	422	97	461
Total	8211	3853	4358	1082	1397	1500	1503	1271	1418

Table 19: Distribution of Person Names

10.4.6 Place Names-W2

The place name contains a list of 324 popular Pan Indian and regional place name. 10 randomly selected names are recorded from a list of names. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	419	140	279	0	130	80	80	60	69
21 To 50	2777	1524	1253	510	380	490	460	524	413
50+	889	260	629	30	190	180	211	50	228
Total	4085	1924	2161	540	700	750	751	634	710

Table 20: Distribution of Place Names

10.4.7 Most Frequent Words-W3A

The most frequent words-part contains a list of 1000 most frequent words. 30 randomly selected words are recorded from a list of words. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	1289	449	840	0	390	240	240	209	210
21 To 50	8382	4626	3756	1557	1141	1468	1381	1601	1234
50+	2644	760	1884	90	568	522	636	148	680
Total	12315	5835	6480	1647	2099	2230	2257	1958	2124

Table 21: Distribution of Most Frequent Words-Part (Radom Selection)

10.5 FULL SET

The full sets are the master set of certain datasets which are read completely from few selected speakers in each group. The 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	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution					
				Awadhi belt		Bhojpuri belt		Magahi belt	
		Female	Male	Female	Male	Female	Male	Female	Male
21 To 50	3994	2000	1994	0	0	1000	997	1000	997
50+	3000	2000	1000	2000	1000	0	0	0	0
Total	6994	4000	2994	2000	1000	1000	997	1000	997

Table 22: Distribution of Most Frequent Word (Full Set)

10.5.2 Phonetically Balanced Vocabulary-W4

The phonetically balanced vocabulary contains a list of words where almost all the phones of Hindi language have occurred in all the possible positions of a word. In full set all the 800 words is recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Awadhi belt		Bhojpuri belt	
		Female	Male	Female	Male	Female	Male
21 To 50	5592	3994	1598	3196	800	798	798
50+	6394	3200	3194	2400	3194	800	0
Total	11986	7194	4792	5596	3994	1598	798

Table 23: Distribution of Phonetically Balanced Vocabulary

10.5.3 Form and Function Words-W5

The form and function words content type contains a list of 200 words that is a representation of almost all the form and function words occurring in Hindi. All the words are recorded from the informant where they uttered those words three times. The distribution of data is as follows:

Age Group	Total Audio Segments	Gender-wise Distribution		Region-wise Distribution			
				Awadhi belt		Bhojpuri belt	
		Female	Male	Female	Male	Female	Male
21 To 50	1399	999	400	799	200	200	200
50+	1595	795	800	597	800	198	0
Total	2994	1794	1200	1396	1000	398	200

Table 24: Distribution of Form and Function Words

10.6 NATIVE SPEAKERS DISTRIBUTIONS

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

Region-wise Distribution of Native Speakers											
Age Group	Total Native Speakers	Gender-wise Distribution of Native Speakers		Region							
				Awadhi belt		Bhojpuri belt		Magahi belt		Khariboli belt	
		Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
16 To 20	75	31	44	0	14	8	8	8	7	15	14
21 To 50	317	174	143	56	39	49	46	57	44	12	14
50+	97	29	68	6	24	18	21	5	23	0	0
Total	488	234	255	62	77	75	75	70	74	27	28

Table 25: Regional Distribution of Native Speakers

10.7 MOTHER TONGUE DISTRIBUTIONS OF INFORMANTS

The following table shows the distribution of mother tongue of the native speakers in LDC-IL speech data.

Mother Tongue of the native speaker	Geographical Dialect Distribution of LDC-IL Hindi Speech Corpus				Total speaker
	Awadhi belt	Bhojpuri belt	Khariboli belt	Magahi belt	
Awadhi	62	-	-	-	62
Bhojpuri	-	150	-	-	151
Magahi	-	-	-	124	124
Hindi	77	-	55	20	152
Total	139	150	55	144	488

Table 26: Distribution of Mother Tongue