छत्तीसगढ़ी
मूल वाक् कॉर्पस

Chhattisgarhi
Raw Speech Corpus

CHHATTISGARHI 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
Mysuru, India-570006

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

Title: Chhattisgarhi Raw Speech Corpus
Authors: Satyaendra Kumar Awasthi, Ankita Tiwari, Shantanu Kumar, Rupesh Pandey, Saurabh Varik, Rajesha N., Manasa G., Srikanth D., Nithin S., Narayan Kumar Choudhary, Shailendra Mohan

ISBN:978978-81-19411-78-8
CIIL Publication No.: 1436

First published: AD 2023September
Bhadrapada1945 Śaka

Publisher: Prof. Shailendra Mohan, 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: Ankita Tiwari

TABLE OF CONTENTS

FIGURES

TABLES

CHHATTISGARHI RAW SPEECH CORPUS

1.1 INTRODUCTION

LDC-IL has taken a positive step in its approach towards the mother tongues spoken in India in addition to 22 scheduled languages, which is an indication of greater efforts to support and promote linguistic variety in the nation. In order to acknowledge the significance of mother tongue. LDC-IL has stepped up its efforts to collect speech data of Chhattisgarhi. This step towards developing language technology for Indian mother tongues will contribute to the overall enrichment and empowerment of mother tongues and will ensure the continued vitality of the language.

As the mother tongue of an estimated 17 million people, Chhattisgarhi is of great cultural and historical significance to the state of Chhattisgarh. It also serves as a medium of expression for the people of Chhattisgarh, reflecting their rich traditions, folklore, and way of life. According to the Indian government, Chhattisgarhi is an eastern dialect of Hindi and is mostly spoken in the Indian state of Chhattisgarh. As of the 2011 Census, 1,62,45,190 people were estimated to speak Chhattisgarhi. Since 2007, Chhattisgarh has recognisedit as one of its official languages.

Central India's Chhattisgarh state was established on November 1st, 2000, following its split from Madhya Pradesh. The Chhattisgarhi language's traditional name is Kosali or Dakshin Kosali, which has historical importance because the Chhattisgarh area was formerly known as Dakshin Kosal. The language has a lengthy and complex history, and it has evolved and changed through time in a number of ways. It has continued to play a significant role in the Chhattisgarh region's cultural and linguistic environment despite these developments.

These recordings will offer a priceless window into the language's everyday use, capturing the colloquialisms, idioms, and regional vocabularies crucial for creating models for natural language processing. The Chhattisgarhi raw speech corpus is a wide-ranging dataset that represents the vibrant, dynamic linguistic legacy of Chhattisgarhi.

1.2 LDC-IL SPEECH CORPUS

The 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. To ensure that the corpus is good for an ASR, the continuous speech is recorded in a natural environment. Three different content types of speech are collected while building this corpus.

The Chhattisgarhi raw speech corpus is made up of recordings of native Chhattisgarhi speakers from various parts of the state of Chhattisgarh, and it represents a wide range of Chhattisgarhi varieties as they are spoken in various locations by diverse speakers. There are several audio for each content category, and each audio has different material. Each speaker from various agegroups recites prompt text extracts of literaryand news texts. A minimum of 2000 words of

read speech is recorded for each speaker. In terms of data size and time length, read speech makes up the majority of the speech corpora. Below are comprehensive explanations of each of the content types:

1.2.1 CREATIVE TEXT

The creative text read speech data includes the recording of a variety of Chhattisgarhi literary writings. In this content type the Chhattisgarhi short stories and essays are read by informants. Any standard text that is descriptive may be used as the creative text. It displays the linguistic preferences of several authors from various regions of Chhattisgarh, from where the content was obtained. Some of the stories were taken from textbooks used in public schools.

1.2.2 NEWS TEXT

To maintain an accurate representation of formal spoken Chhattisgarhi from native speakers, news text extracts were recorded. This content type includes recordings of Chhattisgarhi articles, editorials, and news that have been published from Chhattisgarhi newspapers, monthly magazines, and some online sources

1.2.3 SPONTANEOUS SPEECH (SS)

The spontaneous speech (SS) data includes recordings of responses to questions from native Chhattisgarhi speakers. The Investigator asked several questions to the informants in order to collect spontaneous speech; the answers to these questions had to be given in their own words. LDC-IL made an effort to capture a native speaker's response in a natural speaking style..

1.2.4 GENDER AND AGE BALANCE

Three age groups have been chosen. The categories are "16 to 20 years", "21 to 50 years" and "Above 50 years". An effort has been made to maintain the corpus is balanced in terms of age and gender.

1.2.5 COLLECTION OF DATA

Speech data comprising of 140 speakers was collected through fieldwork. The fieldwork covered the areas from Chhattisgarh (Central), during January 15 - 21, 2023, by six investigators namely*Satyaendra Kumar Awasthi,Srishti Singh, Saurabh Varik, Shantanu KumarAnkita Tiwari*and*RupeshPandey*. The Investigators visited several colleges, universities and organisations in order to gather the data. There are 140 speakers in the Chhattisgarhi speech data, The locations covered by the different speakers are mentioned are, Balod, Baloda Bazar, Bemetara, Bilaspur, Dhamtari, Durg, Gariaband, Janjgir-Champa, Kabirdham, Khairagarh-Chhuikhadan-Gandai, Korba, Mungeli, Mahasamund, Raipur, Sakti, Rajnandgaon, Sarangarh-Bilaigarh.

Informants are made aware of why exactly the data is being collected. The informants are also made aware of the potential benefits of the data to the wider community. Once the informant is

aware of all this information and is ready to give the data, consent is acquired in writing along with certain personal details which makes a part of metadata.

Every effort is made to reduce any additional noise as much as possible. Before the actual recording of the text, test recordings are conducted. Informants are allowed to read the dataset earlier before recording so that they can get familiar with the content of the text. While audio recording the read-speech content types the informants were told to read the Text as natural as reading a book or newspaper possible

1.2.6 TECHNICAL SPECIFICATIONS FOR COLLECTING DATA

The LDC-IL data is recorded using Roland EDIROL Recorder. It is a 24-bit Linear PCM (R-09) Recorder. Recording was done at the Sample Rate: 48.0 KH in 16bit wav format using rechargeable batteries.

1.2.7 GUIDELINES FOLLOWED WHILE AUDIO RECORDING

- Minimum 5 cm to 25 cm distance between the microphone and the speaker
- Recorder should not be placed orthogonally but it should be placed diagonally.
- Do not move the recorder during recording. Fix the recorder upon a fixed plane if possible.
- Maintaining fixed distance between the recorder and speaker

After each recording, it is suggested to the investigator to ensure that the recorded data is received properly. The investigator may oblige the informant's request if they also want to hear the data.

1.3 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. A brief of each of these 21 fields is given in the table below:

Sl.	Legend	Description
1.	Language	Name of the Language
2.	Speaker ID	Each speaker has a unique ID. However, this is within the language.
3.	Gender	Note gender, whether it is male, female or other.
4.	Age Group	Three age groups of 16 to 20, 21 to 50, and 50+
5.	Mother Tongue	Mother tongue of the informant.
6.	Dialect	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.
7.	State	Name of the Indian state/province of the speaker
8.	District	Name of the district of Indian state/province of the speaker
9.	Education	Highest educational qualification of the speaker.
10.	Place: Elementary Education	Place of elementary education. This usually corresponds to the early childhood experiences which happen to more than often affect the way a language spoken.
11.	Place	Place provides the information that where the speech data collected.
12.	Date	Date when the recording took place.
13.	Environment	Indoor/Outdoor environment in which the recording has been done.
14.	Content Type	This corresponds to the notation of the content types noted above.
15.	Content ID	This corresponds to the ID of the text being read out.
16.	Audio File Name	The name of every Audio file is distinct based on speaker and content types
17.	Recorded Text	Text (Datasets) of the recorded speech.

18.	Duration	This is related to the duration of audio in milliseconds.
19.	User By	Person who handled the data-metadata association in corpus.
20.	Script	The script in which the content has been provided to the informant for reading.
21.	Investigator	Name of the Investigator.

Table 1: Metadata fields and their description

1.4 TEXT-SPEECH MAPPING AND NAMING CONVENTIONS

After the completion of field work, the speech segmentation and mapping audio with corresponding text and metadata is done. Each recording is named in accordance with its metadata information like language name, gender, age group, speaker id and content id etc. A Typical LDC-IL naming convention for Chhattisgarhi Speech Data is shown below.

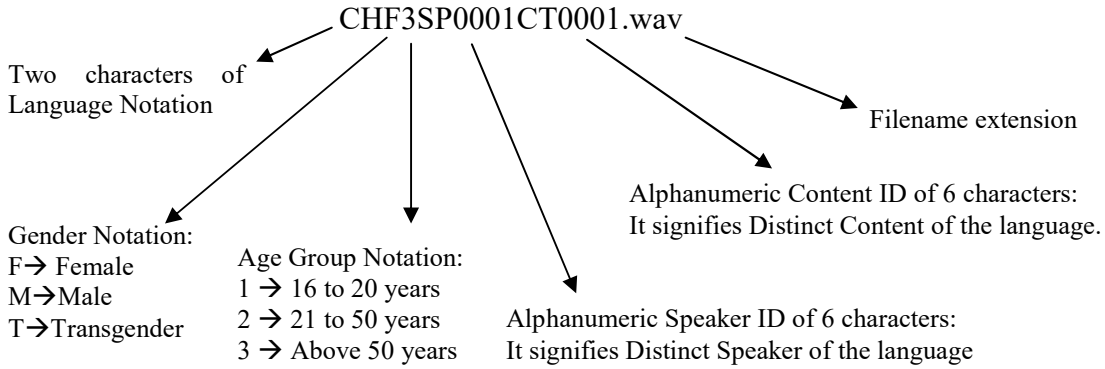

Figure 1: LDC-IL Naming Convention of Chhattisgarhi Speech Data

1.5 SUMMARY OF THE CORPUS

The Speech data has 140 distinct speakers with the duration of 138:09:27 (hh:mm:ss)
The distribution of duration across each content type is as follows

ContentType	Gender	AgeGroup	Duration (hh:mm:ss.ms)		
Creative Text	Female	16 to 20	00:39:35	21:45:55	49:00:23
		21 to 50	19:26:08		
		Above 50	01:40:10		
	Male	16 to 20	00:14:59	27:14:27	
		21 to 50	22:31:55		
		Above 50	04:27:33		
News Text	Female	16 to 20	00:43:36	10:05:56	28:57:53
		21 to 50	08:31:57		
		Above 50	00:50:22		
	Male	16 to 20	01:25:36	18:51:56	
		21 to 50	14:42:09		
		Above 50	02:44:10		
Spontaneous Speech	Female	16 to 20	01:15:30	24:06:35	60:11:11
		21 to 50	21:24:24		
		Above 50	01:26:41		
	Male	16 to 20	01:05:39	36:04:35	
		21 to 50	28:32:19		
		Above 50	06:26:37		

Table 2: Distribution of duration across each content type of Chhattisgarhi

Among 140 distinct speakers Creative Text content type is read by 129 speakers and its distribution is as follows.

Gender	AgeGroup	Speakers	Total Speakers
Female	16 To 20	3	54
	21 to 50	47	
	Above 50	4	
Male	16 To 20	2	75
	21 to 50	62	
	Above 50	11	

Table 3: Distribution of number of speakers across Chhattisgarhi Creative Text

Among 140 distinct speakers News Text content type is read by 94 speakers and its distribution is as follows.

Gender	AgeGroup	Speakers	Total Speakers
Female	16 To 20	3	34
	21 to 50	28	
	Above 50	3	
Male	16 To 20	3	60
	21 to 50	48	
	Above 50	9	

Table 4: Distribution of number of speakers across ChhattisgarhiNews

Among 140 distinct speakers Spontaneous Speech content type is recorded from 136 speakers and its distribution is as follows.

Gender	Age Group	Speakers	Total Speakers
Female	16 To 20	3	54
	21 to 50	48	
	Above 50	3	
Male	16 To 20	3	82
	21 to 50	65	
	Above 50	14	

Table 5: Distribution of number of speakers across Chhattisgarhi Spontaneous Speech

1.6 REFERENCES:

1 Choudhary, N. and D. G. Rao. 2020. The LDC-IL Speech Corpora. In Proceedings of 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), Yangon, Myanmar,2020. pp. 28-32, doi: https://doi.org/10.1109/O-COCOSDA50338.2020.9295011

2 Choudhary, N. 2021. LDC-IL: The Indian Repository of Resources for Language Technology. Language Resources & Evaluation. Springer, Vol. 55, Issue 1. doi: https://doi.org/10.1007/s10579-020-09523-3

3 Choudhary, Narayan, Rajesha N., Manasa G. & L. Ramamoorthy. 2019. "LDC-IL Raw Speech Corpora: An Overview" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 160-174.

4 Adil, Satyabhama 2003. Chhattisgarhi Bhasha aur Sahitya, Vikalp Prakashan Raipur pp. 2-31

5 Census 2011. https://censusindia.gov.in/