संताली

मूल रो़ड़ कॉर्पस

SANTALI

RAW SPEECH CORPUS

भारतीय भाषा संस्थान
उच्चतर शिक्षा विभाग, शिक्षा मंत्रालय, भारत सरकार, मानसगंगोत्री, हुन्सूर रोड, मैसूरु-570006, कर्नाटक
CENTRAL INSTITUTE OF INDIAN LANGUAGES
Department of Higher Education, Ministry of Education, Government of India
Manasagangotri, Hunsur Road, Mysuru-570006, Karnataka
www.ciil.gov.in :: e-✉ ciil.publication@gmail.com :: ☎ +91-821-234-5182

Santali Raw Speech Corpus

Authors: Dharmenjay Hembram, Ankita Tiwari, Dr. Narayan Choudhary, Prof. Shailendra Mohan

e-ISBN: 978-81-69175-04-3
CIIL Publication No.: 1689

First published: AD 2026 March :: Phalguna 1947 Śaka

Publisher: Prof. Shailendra Mohan, Director, CIIL

Printing & Publication Unit
Aleendra Brahma, Officer-in-Charge, Printing Press & Publication Unit
B Manju Bhargavi, Unit-in-Charge, Printing Press & Publication Unit
M.N. Chandrashekar, Compositor, Printing Press & Publication Unit

Cover Design: Ankita Tiwari

TABLE OF CONTENTS

1. INTRODUCTION

A major challenge in the development of language technology for Indian languages is the lack of necessary linguistic resources. Many languages in India do not have adequate speech data or digital text, which makes it difficult to design and develop language technologies. Research studies have shown that this shortage is one of the main reasons for the slow progress in this field. Due to the absence of such resources, the growth of language technology applications has been limited in both academic and industrial sectors. To overcome this problem, the Government of India has taken several steps to develop and share important linguistic resources. One important initiative is the Linguistic Data Consortium for Indian Languages, which was set up by the Ministry of Education at the Central Institute of Indian Languages in Mysore to support language technology.

The LDC-IL has initiated efforts to collect speech data for the Santali language as part of its mission to support language technology development in India. This initiative focuses on creating a structured speech corpus that can be used for research and the development of speech-based applications such as speech recognition, speech synthesis, and other language processing tools. LDC-IL aims to strengthen digital support for Santali and facilitate further research and technological innovation related to the language by building reliable linguistic resources. This step toward the development of language technology will enhance the resources available for Indian languages and help maintain their vitality and long-term use. As part of the preparation process, the corresponding text for the read speech data was provided in the Devanagari script, since a large number of Santali speakers in Jharkhand are familiar with and commonly use this script for writing the language.

As per Census 2011, there are 73,68,192 Santali speakers which cover 0.61 per cent of the country's total population. However, Santali speakers in the state of Jharkhand constitute 9.91 percent of the state's total population. The major share of Santali speakers are found in the states of Jharkhand (44.38%), West Bengal (32.97%), Odisha (11.71%) and Bihar (6.23%). They altogether account for 95.28 per cent of the total Santali speakers of the country. Santali speakers are also distributed in the States/Union Territories of Assam (2.89%), Maharashtra (1.40%), Chhattisgarh (0.24%), Tripura (0.05%) and Arunachal Pradesh and Mizoram (0.02% each). Below these, Uttar Pradesh, Himachal Pradesh, NCT of Delhi, Uttarakhand and Madhya Pradesh have registered 0.01

per cent each of Santali speakers. In 18 States/Union Territories their percentage share is below 0.01 per cent.[1]

The Santali (ISO 639-3: sat)[2] language is one of the most prominent languages of the Munda subgroup within the Austroasiatic language family. It is spoken by millions of people across eastern India, particularly in Jharkhand, West Bengal, Odisha, Bihar, and Assam. Traditionally, Santali was mainly transmitted through oral traditions such as folklore, songs, and storytelling, which played an important role in preserving the cultural heritage of the Santali community. Dr. Dhirendra Nath Baskey say about oral literature. “Form writing to speaking, the manner, style and tone of speech attract the audience. But is that why the indigenous people did not think of writing, did not try to all? Santali literature is written by word of mouth and has been preserved in the memory of ages and epochs. Jamsam Binti, Karam binti, countless rhyming, Folk tales, proverb, pranks, riddles, etc. Are prime example of this”. Santali texts have been written using different scripts such as the Roman, Bengali, Devanagari and Ol Chiki script depending on regional influences.[3]

To establish a standardized writing system that accurately represents the phonological structure of the language, the Ol Chiki script was introduced in 1920s by Raghunath Murmu. This script was specifically designed for Santali and consists of alphabetic characters that correspond directly to the distinct sounds of the language. Today, Ol Chiki is widely used in education, literature, and official publications related to Santali. However, the Roman, Bengali and Devanagari scripts continue to be used in certain contexts due to historical and regional influences. The use of multiple scripts creates both opportunities and challenges for linguistic research, particularly in the development of digital language resources and language technology for Santali.[4]

---

[1] Census of India, 2011
[2] https://iso639-3.sil.org/code/sat
[3] Gurudas Murmu, A Brief History of Santali Language and Literature
[4] Dr. Truptirekha Sahoo, Reviving Identity Through Script: A Study on Santali Language and Pandit Raghunath Murmu’s Contribution

2. CHALLENGES IN COLLECTING SANTALI SPEECH DATA

Since this was one of the first initiatives in which Santali speech data was collected through a mobile application by the Linguistic Data Consortium for Indian Languages (LDC-IL), introducing a relatively new approach to data collection. As a result, several challenges were encountered during the speech data collection process, which affected the efficiency and smooth execution of the task. One of the major difficulties was poor internet connectivity, particularly in remote and rural areas where many participants resided. Because the recording application required stable internet access to save and submit recorded segments, weak or unstable network connections often interrupted the uploading process. In some cases, speakers were unable to upload their recordings successfully, which led to the loss of a few recorded segments. Furthermore, unstable internet connectivity also caused delays in completing the assigned tasks. Some participants had to wait for better network availability and in certain cases a single speaker required two to three days to complete one recording task because they could only record and upload the data at night when the internet connection was relatively stable.

Another challenge was related to the limited storage capacity of mobile phones. Since sufficient storage space was required to record and temporarily save the audio files, many participants from remote areas who used mobile devices with limited storage faced additional difficulties during recording and submission.

Additionally, collecting spontaneous speech data proved challenging for some speakers. A few participants were not familiar with speaking spontaneously on a given topic, which affected the natural flow of their recordings. Another issue was the limited awareness among some community members about the importance of speech data collection and language technology. As a result, some individuals were initially hesitant to participate in the recording process and often felt uncomfortable or hesitant while recording, as this type of task was completely new to them. These challenges required additional explanation, guidance, and repeated efforts to help participants feel comfortable and to ensure the successful completion of the data collection process.

3. SANTALI RAW SPEECH CORPUS

The Santali Raw 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 type of speech are collected while building this corpus.

This speech data is recorded using the SDCP (Speech Data Collection Portal) mobile application. This application is developed by the Linguistic Data Consortium for Indian Languages (LDC-IL) to facilitate efficient and systematic collection of speech data for various languages. SDCP is a speech recording platform designed for speech data contributors and voice talents. The portal provides a mechanism to record and manage speech or audio recording tasks for various types of projects and ensures that the recording, review, and management of the speech data take place smoothly and systematically.

The Santali Raw Speech Corpus consists of recordings of native Santali speakers from various parts of the state of Jharkhand. The corpus contains multiple audio recordings for each content category, with each recording featuring different material. Speakers from different age groups recite prompted text, including literary excerpts and sentences. The sentences were translated by language experts and were selected from the Parallel Corpus developed by the Linguistic Data Consortium for Indian Languages (LDC-IL). The sentences were taken from the Parallel Corpus developed by the Linguistic Data Consortium for Indian Languages (LDC-IL).

These sentences cover the general grammatical categories like Adjective, Adverb, Compounding, Interrogative, Imperative, Speech type, Statement, Structural Question, Coordination, and Subordination etc. Moreover, it includes linguistic features such as Adposition, Anaphora, Derivational Morphology, Inflectional Morphology and language specific features like Emphasis, Equative, Posession, Reciprocal, Reflexive etc.[5]

A minimum of 2,000 words of read speech is recorded from each speaker, which includes approximately 1,000 words from PCP sentences and 1,000 words from literary texts. The data also includes spontaneous speech recordings based on various topics or questions related to day-to-day

[5] LDC-IL Corpus Insights

life, as well as themes from different domains. This approach helps maintain the authentic nature of everyday utterances and the use of native terminology. Below are detailed explanations of each of the content types:

3.1 CREATIVE TEXT (CT)

The Creative Text (CT) read speech data consists of recordings of various Santali literary texts. Descriptive and standard texts are selected as creative text, reflecting the linguistic styles and preferences of different authors. In addition, a few translated texts were also used for the data collection.

3.2 PCP SENTENCE (SE)

PCP sentences were used from the LDC-IL Parallel Corpus, which is designed to represent a wide range of grammatical structures and linguistic features. The Santali speech corpus uses PCP sentences from the LDC-IL Parallel Corpus to include a variety of grammatical structures in the recordings. In total, 5,332 PCP sentences were included in the dataset. To maintain consistency and balanced coverage, the same set of 5,332 sentences was provided to both male and female speakers for recording.

3.3 SPONTANEOUS SPEECH (SS)

Spontaneous speech consists of natural, unscripted utterances produced by speakers while responding to questions or topics. It captures authentic language use, including natural vocabulary, discourse patterns, and speaking styles used in daily communication. Spontaneous speech consists of natural, unscripted utterances produced by speakers while responding to questions or topics. It captures authentic language use, including natural vocabulary, discourse patterns, and speaking styles commonly used in everyday communication.

A total of 33 questions/topics were designed to collect spontaneous speech in the Santali speech corpus. Among these, the first five questions were mandatory for every speaker, as they focused on commonly used numerical expressions such as current date, PIN code, date of birth, time and distance formats. These prompts helped capture natural speech involving numbers and everyday informational expressions. In addition to these mandatory prompts, participants were asked to record their speech on any eight topics out of the remaining twenty-eight prompts according to their

choice. However, some speakers showed considerable enthusiasm during the recording process and voluntarily recorded responses for all 33 segments.

3.4 GENDER AND AGE BALANCE

Three age groups were selected for the recordings: 16–20 years, 21–50 years, and above 50 years. An effort was made to maintain the corpus balanced in terms of both age and gender.

3.5 COLLECTION OF DATA

Speech data from 60 speakers was collected using the SDCP (Speech Data Collection Portal) mobile application. The fieldwork was conducted between June 2025 and February 2026 across different regions of Jharkhand by five investigators: Dharmenjay Hembram, Ankita Tudu, Tanushree Singh Murmu, Joy Sagar Murmu, and Ankita Tiwari. Investigator Miss Tudu visited several locations to collect the recordings, while the other investigators gathered data by contacting speakers through email and phone calls, with assistance from local community members.

The dataset includes 60 speakers (30 male and 30 female). The speakers come from various locations including Dumka, Pakur, Jamtara, Dhanbad, Bardhaman, Purulia, Singhbhum, Sahibganj, East Singhbhum, Deoghar, Bokaro, Godda, and Sikaripara.

Consent was obtained from each speaker before recording the data. During data collection, basic requirements such as a suitable recording space, a power source for charging the phone, and internet connectivity were ensured wherever possible. Before obtaining consent, the informants were clearly informed about the purpose of the data collection and its potential benefits for the wider community. Once the informants understood this information and agreed to participate, their consent was obtained in writing along with certain personal details, which form part of the metadata.

Before recording, every effort is made to minimize background noise as much as possible. Informants are also allowed to read the dataset and topics/questions beforehand so that they can become familiar with the content.

Informants were asked to read the text naturally during the read-speech recording. For spontaneous speech, the informants were informed about the prompts related to various topics and daily

activities available in the mobile application, and their responses were recorded to observe natural conversational patterns in the language.

3.6 GUIDELINES FOLLOWED WHILE SPEECH RECORDING

- Select native speakers representing different age groups, genders, and socio-economic backgrounds to ensure diversity in the speech data.
- Obtain consent from each participant before recording and maintain detailed metadata for every speaker.
- Ensure that the recording device or mobile phone is fully charged and functioning properly prior to starting the recording process.
- While recording in indoor spaces like rooms or halls, it is important to ensure that echo must be avoided.
- Listen to a few recorded segments during the session to verify the quality.
- Clearly explain the recording procedure to the informant and encourage them to speak naturally at a normal speed and tone.
- Check the availability of sufficient mobile storage space and stable internet connectivity before beginning the recording process.
- After each recording session, the investigator should verify that the recorded data has been saved and successfully submitted.

4. 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 of 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.
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

5. TEXT-SPEECH MAPPING AND NAMING CONVENTIONS

Once the data collection process is completed, the recorded files must be stored on a server at the earliest to ensure safe preservation. Each speech recording is maintained along with its corresponding textual content and metadata information. The naming convention of the recordings follows the metadata information, such as language name, gender, age group, speaker ID, and content ID, ensuring systematic organization and easy access to the data.

A Typical LDC-IL naming convention for Santali Speech Data is shown below:

STF1SP001CT384633.wav / STM3SP023SS389282.wav / STM2SP050SE1184091.wav	
ST	Two characters of Language
F/M	**Gender Notation:** F- Female M- Male
F1, F2, F3 M1, M2, M3	**Age Group Notation:** 1 16 to 20 years 2 21 to 50 years 3 Above 50
SP001 / SP050 / SP023	**Speaker ID:** A unique identifier consisting of five alphanumeric characters used to represent a distinct speaker of the language.
CT384633 / SS389282 / SE1184091	**Content ID:** A unique alphanumeric characters containing eight or nine characters that indicates a specific content segment in the language corpus.
.wav	Filename Extension

Table 2: Description of Santali Speech Data Naming Convention

6. SUMMARY OF THE CORPUS

The speech dataset consists of recordings from a total of 60 distinct speakers, with an overall duration of 50:00:33 (hh:mm:ss). To ensure balanced representation in the dataset, the speakers were equally distributed by gender, with 30 male speakers and 30 female speakers participating in the recording process. The distribution of duration across each content type is as follows:

Content Type	Gender	Duration (hh:mm:ss)	Speech Segments
Creative Text (CT)	Female	05:11:53	89
	Male	05:25:38	87
Sentence (SE)	Female	06:09:51	5331
	Male	06:56:47	5332
Spontaneous Speech (SS)	Female	11:19:48	453
	Male	14:56:36	554

Table 3: Distribution of Duration across Each Content Type

REFERENCES

- Sahoo, Truptirekha. 2025. Reviving Identity Through Script: A Study on Santali Language and Pandit Raghunath Murmu's Contribution. Journal of Emerging Technologies and Innovative Research (JETIR), Vol. 12, Issue 12. 2349-5162. https://www.jetir.org/papers/JETIR2506337.pdf
- Murmu, Gurudas. 2022. A brief History of Santali Language and Literature. Journal of Emerging Technologies and Innovative Research (JETIR). Vol. 9, Issue 12. 2349-5162. https://www.jetir.org/papers/JETIR2212493.pdf
- https://www.ebsco.com/research-starters/language-and-linguistics/santali-language
- 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
- 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
- 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.
- Census 2011. https://censusindia.gov.in/nada/index.php/catalog/42561
- https://iso639-3.sil.org/code/sat