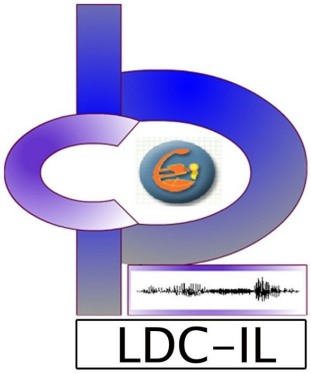
मैथिली
मूल वाक् कॉर्पस
खंड - II




Maithili
Raw Speech Corpus
Volume - II

MAITHILI RAW SPEECH CORPUS VOL. II



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

Maithili Raw Speech Corpus Vol. II

Shantanu Kumar, Ankita Tiwari, Rajesha N., Manasa G., Dr. Narayan Kumar Choudhary, Prof. Shailendra Mohan

e-ISBN:
CIIL Publication No.:

First published: AD 2025 March :: *Phalguna 1946 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

TABLES

1 MAITHILI RAW SPEECH CORPUS VOL. II

1.1 INTRODUCTION

The main barrier to language technology development for Indian languages has been a lack of fundamental linguistic resources. One could not even consider speech data when most Indian languages already had text data available. India is one of the foremost multilingual countries 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 by 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 to continue. It is also estimated that internet users in the Indian language will account for approximately 75% of India's internet user base by 2021.

Even though this is the case, very little technology is available in Indian languages. This is primarily because the institutions responsible for developing new technologies find it either too difficult or not economically feasible to provide linguistic support for a wide range of technology-based applications for Indian languages.

The Indian government has initiated several efforts in response to this problem to supply the core elements that could accelerate the development of language technology in Indian languages. As part of this initiative, the Linguistic Data Consortium for Indian Languages (LDC-IL) was established by the Ministry of Education at the Central Institute of Indian Languages, Mysore.

The goal of LDC-IL was to develop linguistic resources for all Indian languages, with an initial focus on the scheduled languages of India. The language technology development community may deem these resources suitable.

In addition to 22 scheduled languages, LDC-IL has also taken a positive step in its approach towards the mother tongues spoken in India, which is an indication of greater efforts to support and promote linguistic variety in the nation. To acknowledge the significance of the mother tongue, LDC-IL has stepped up its efforts to collect speech data of Maithili. 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.

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 the multifarious needs of the research and development community.

1 https://assets.kpmg.com/content/dam/kpmg/in/pdf/2017/01/Impact-of-internet-and-digitisation.pdf

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 but rather getting 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 has 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 a natural environment.

1.3 CONTENT TYPE DESCRIPTION

The Maithili raw speech corpus is made up of recordings of native Maithili speakers from various parts of the state of Bihar and Jharkhand, mainly Bhagalpur and neighboring districts, and it represents a wide range of Maithili varieties as they are spoken in various locations by diverse speakers. Though the region is primarily claimed as the Angika-speaking region by several works (Grierson 1907, Sharma 2001) and a few other individual research works, still, based on the government documents and the Census report of 2011 by the Govt. of India, the area is considered as the language variety of Maithili. Hence, the area is referred to as the Maithili-speaking region in this document hereafter.

Below are comprehensive explanations of each of the content types:

1.3.1 **Creative Text (CT)**

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

1.3.2 **Spontaneous Speech (SS)**

The spontaneous speech (SS) data includes recordings of responses to questions from native speakers. The Investigator asked several questions to the informants related to their daily lives to collect the real-time natural speech of a speaker as a 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. As any language has a dynamic nature, it tends to change its forms and features within a distance of a few kilometers. To capture the different forms of the language across the geographical regions, the spontaneous speech has been collected as a part of corpus development. 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.

Three age groups have been chosen for LDC-IL datasets: 16 to 20 years, 21 to 50 years, and above 50 years. An effort has been made to maintain a balanced corpus in terms of age and gender.

1.4 DATASET PREPARATION AND DISTRIBUTION

Each speaker from various age groups recites prompt text extracts of literature, i.e., Creative Text (CT). A minimum of 2000 words are recorded for each speaker in each recording. In terms of data size and time length, read speech makes up the majority of the speech corpora. The dataset also includes Spontaneous Speech (SS) recordings based on conversations and some random topics and questions about government policies, agriculture, state, districts, day-to-day life, culture, historical places, folk tales, and several other things from various fields to maintain the authentic essence of everyday utterances and native terminologies.

A deliberate attempt has been made to maintain distinctiveness in speech corpora by taking a variety of styles into consideration to ensure representativeness in speech corpora. After the regions are identified, speech samples are collected as per the criteria shown in the table below:

Content type	**Content size**	**Content read by each speaker**	**Age group-wise no. of speaker**			**Content selection type**
			16-20	21-50	50+	
Creative Text	85 Texts	1 Text	07	62	16	Distinct Text
Spontaneous Speech	NA	NA	08	79	34	Prompt questions from various domains and day-to-day life

Table 1: Content-wise Speech Dataset Distribution

1.5 DATA COLLECTION

Fieldwork was conducted in Bihar and Jharkhand, where the Angika variety is spoken. Speech data comprising 122 speakers was collected in these fieldworks. Six investigators namely, Shantanu Kumar, Rupesh Kumar Pandey, Akanksha Tiwari, Jyoti Kumari, Nikhil Kumar, Mukesh Kumar have collected the Data, during November 06 - 13, 2022 from Bhagalpur and neighboring areas of Bihar & Jharkhand.

1.6 SUMMARY OF THE CORPORA

LDC-IL Maithili raw speech corpus vol.II has 206 audio segments with a duration of 109:09:50 (hh:mm:ss:ms).

The data distribution of the speech corpora is shown in the table below:

Content Type	Gender	Female			Male		
	Age Group	16-20 Years	21-50 Years	50+ Years	16-20 Years	21-50 Years	50+ Years
	Total Segments	Segments	Segments	Segments	Segments	Segments	Segments
Creative Text-CT	85	06	28	04	01	34	12
Spontaneous Speech-SS	121	07	35	07	01	44	27

Table 2: Audio Segments and their Distribution

Content-wise data distribution and their duration are shown in the table below:

Content Type	Gender	Age Group	Duration (hh:mm:ss)		
Creative Text	Female	16To20	3:19:17	18:59:14	42:46:40
		21To50	13:41:49		
		Above51	1:58:08		
	Male	16To20	0:42:33	23:47:26	
		21To50	17:37:15		
		Above51	5:27:38		
Spontaneous Speech	Female	16To20	3:32:58	26:17:15	66:23:11
		21To50	18:59:32		
		Above51	3:44:45		
	Male	16To20	0:25:02	40:05:57	
		21To50	24:55:06		
		Above51	14:45:48		

Table 3: Content-wise Speech Dataset Distribution and their duration

1.7 REFERENCES

1. Grierson, GA. 1909, An introduction to the Maithili dialect of the Bihari language as spoken in North Bihar (2 ed.). Calcutta: Asiatic Society of Bengal. pp. xi-xiii.
2. Jha, S. 1958, "The formation of the Maithili language." PhD diss., Luzac London. pp. 5-7
3. Sharma, RM. 2007, Angika Bhasha Ka Dhwanivaigyanik Adhyayan (In Hindi) (II ed.). Shri Tara Geeta Printing Press, Bhagalpur, pp. 130-133
4. https://www.angika.com/p/angika-language.html
5. Choudhary, N. (2019). Linguistic resources for AI/NLP in Indian languages. Mysore: Central Institute of Indian Languages.
6. Ethnologue. (n.d.). hne|ISO-639-3. Retrieved 09 15, 2023, from SIL:https://iso639-3.sil.org/code/hne

7. Office of the Registrar General & Census Commissioner, India (ORGI). (2022, 04 22). Language. Retrieved from Census of India 2011: https://censusindia.gov.in/nada/index.php/catalog/42458
8. Brass, P. R. (2005). Language, Religion, and Politics in North India. iUniverse, Lincoln, NE
9. Choudhary, N. K. (2021). The Indian Repository of Resources for Language Technology. Language Resources & Evaluation (I ed., Vol. 55). Springer. https://doi.org/10.1007/s10579-020-09523-3
10. Rejitha K. S. and Narayan Kumar Choudhary. (ed.). 2023. Compendium of LDC-IL Sentence Aligned Speech Corpus. Central Institute of Indian Languages, Mysore. ISBN: 978-81-19411-34-4.