মৌথিনী রাকা সংবেখিত রাক্ কার্পস
(তিবহ্মতা নিপি)

Maithili Sentence Aligned Speech Corpus
(Tirhuta Script)

MAITHILI SENTENCE ALIGNED SPEECH CORPUS (TIRHUTA SCRIPT)

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 Sentence Aligned Speech Corpus (Tirhuta Script)

Dinesh Mishra, Shantanu Kumar, Dr. Narayan Kumar Choudhary, Rajesha N., Prof. Shailendra Mohan

e-ISBN: 978-93-48633-51-4
CIIL Publication No.: 1507

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: Dinesh Mishra

TABLE OF CONTENTS

FIGURES

TABLES

1 SPEECH ANNOTATION

Narayan Kumar Choudhary, Rejitha K .S.

1.1 INTRODUCTION

Linguistic Data Consortium for Indian Languages (LDC-IL) has been acting as the repository of Indian language datasets since 2008. It has so far released 42 datasets covering 20 scheduled languages. The released datasets include raw text corpora and raw speech corpora. More details about the scheme and the datasets are available in Choudhary, 2021 and Choudhary and Rao, 2020.

This paper presents a documentation of sentence aligned speech corpora in various languages. This is in continuation of the earlier raw speech corpora in various languages. The difference between the raw speech corpora and the sentence aligned corpora can be easily guessed. As the title of the dataset suggests, the sentence aligned corpus is split at the 'sentence' or a meaningful 'utterance' boundary while the raw speech corpus had speech segments that were usually larger in size.

1.2 LDC-IL SPEECH ANNOTATION

LDC-IL speech corpora are collected using two methods (complete details about the raw speech corpus and the process of its creation are included in the raw speech corpus documentation of respective languages).

a. Read Speech: A vast majority of the speech dataset in LDC-IL is created by reading texts from various sources. These sources are usually part of the corresponding language taw text corpus.

b. Spontaneous Conversation: These have been used for data collection in some of the languages and conversational data for other languages will be collected in subsequent phases.

We have manually transcribed both of these kinds of data and aligned the transcriptions at the sentence level. Even though the read speech data is created by reading the texts, a manual transcription was necessitated because a lot of times speakers do not reproduce the texts verbatim and so the original texts do not necessarily correspond to the speech recordings. Moreover, the speech transcriptions represent the different variant forms in speech as well as other speech properties (viz false start, mispronunciation etc.) and non-speech properties (viz background noise) - all of these are meticulously marked during the annotation process.
LDC-IL speech data is annotated with the official script of a language, wherever such a script is specified. For example, Malayalam speech data is annotated in Malayalam Script in UTF-8 encoding; Hindi is annotated in Devanagari script and so forth. The languages which do not have an official script are annotated using the most commonly used script for writing that language. All annotations and transcriptions are carried out using the Praat software. The annotators were

provided with the audio files aligned with the corresponding text file for annotation of the read speech data - this greatly reduced the need of annotating from the scratch for such data.
We have provided the annotations at two levels -

a. Phonetically Normalised Speech Annotation
b. Orthographically Normalised Speech Annotation

These two levels are discussed in the following subsections below.

1.2.1 Phonetically Normalized Speech Annotation

In phonetically normalized speech annotation layer, we provide the narrow transcriptions of speech. At this layer, even though the script used for transcribing speech is the official script of the language, we use non-standard and even non-existent spellings to represent the speech exactly as it is spoken. The detailed guidelines used by the annotators for transcription and annotation at this layer are given in the following subsection.

1.2.1.1 Guidelines for Phonetically Normalized Speech Annotation

Annotation of this aligned data should be carried out as per the pronunciation of the speaker in the audio. The wrong pronunciation, that is, deviation from the text should be transcribed accordingly.

Following are the instances which should be marked in the annotation.

1. Use of '#'
   '#' symbol is used to mark excessively noisy or unnecessary parts of the audio - generally these portions also contain human speech but are not legible. The audio is post-processed to remove such portions from the audio files that are being released publicly. It is used within the sentence.

2. Use of '0'
   '0' is used to mark silences at the beginning and end of the audio files as well as long silences or non-speech sounds in the intermediate portions of the audio files - these are marked only for those portions which do not have any kind of human speech. These portions are also removed from the publicly released audio files in the post-processing step.

3. Silences are removed.
   Any silence longer than 50 ms is marked using #.

4. Cut-off speech and intended speech is marked.
   Example: [mini]*ster —shows that the speaker intended to speak minister but spoke mini in an unclear fashion and ster clearly.

5. Annotation of speech disfluency
   Restarts/false starts should be marked. For example, if the speaker intends to speak "bengaluru" but speaks "be bengaluru", this should be marked as be-bengaluru.

6. Numbers
   All number sequences should be spelled out. Years should be transcribed in spoken format.

7. Mispronunciations
   If a speaker mispronounces a word and the mispronunciation is not an actual word, transcription should be done as the word is spoken.

8. Utterances longer than 30 seconds should be further split into multiple parts - this split is made at the point of a long silence of around 500 ms.

1.2.2 Orthographically Normalized Speech Annotation

In Orthographically Normalized Speech Annotation, a broad transcription is given. This implies that the standardised spelling is used for transcription and speech segments such as cut-off speech, intended speech, repetition etc. are not marked. The complete guidelines are reproduced in the following subsection.

1.2.2.1 Guidelines for Orthographically Normalized Speech Annotation

Orthographically normalized textual layer is prepared over the phonetically normalized text by using the guidelines given below

1. Small portion of a word like a grammatical element or a letter is missed then it is corrected as per the correct writing pattern.
   For example, if the speaker speaks ‘avan viitti poyi’ ‘He went home’ in an informal way then it is annotated as ‘avan viittil poyi’ in the proper standard Malayalam sentence. There is no valid word ‘viitti’ in the language, so it is annotated as a proper word according to the context.
2. Any deviation in the phonetically annotated text is corrected according to standard writing form.
   For example, if the audio is “Maiyam Engineering”it must be corrected as “Marine Engineering”.
3. Restarts/false starts are removed. For example, the speaker intends to speak “keralam” but speaks “ke keralam”. If the word “ke” is a valid word morpheme, then it has been kept, otherwise “ke” is not marked.
4. Cut-off speech is written as standard form and remove [ ]* symbol
5. Incomplete sentence is standardized to the extent of available audio

Note: Indian English - Bengali Variant Speech Annotation and Indian English - Kannada Variant Speech Annotation are following a separate guideline which is available in its respective sections.

1.3 SPEECH ANNOTATION QUALITY ASSURANCE

Transcriptions are susceptible to human errors, such as typos or spelling mistakes. To ensure data accuracy, both phonetically normalized, and orthographically normalized textual layers undergo third party evaluation. The third-party evaluator must check the transcription against its corresponding audio to rectify any mistakes. The linguist may accept or reject the corrections by means of matching the original and evaluated transcriptions against the audio. The linguist's specialized knowledge is essential in bringing the transcriptions back in line with accuracy. It is ensured that the mistakes encountered by a third party evaluator are also corrected by arbitrating it further by a third linguist.

2 MAITHILI SPEECH ANNOTATION (TIRHUTA SCRIPT)

Dinesh Mishra, Narayan Kumar Choudhary

2.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Maithili Sentence Aligned Speech Corpus (Tirhuta Script) is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Maithili Speech Corpus is available [3]. LDC-IL Maithili Sentence Aligned Speech Corpus (Tirhuta Script) files contain an audio file and two textual layers in 'Tirhuta' script. Each File 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 Sentence Aligned Speech data is
'Maithili_Female_16To20_Contemporary_Text-T1_SP-0007_T1-0007-001.wav'

LDC-IL Sentence Aligned Speech corpus for Maithili contains read speech from four content types viz. contemporary text, creative text, sentences and date format. The contemporary text and creative text are sampled from news and essays/novels respectively. The sentences are a collection of phonetically balanced sentence lists - each speaker has typically recorded 25 sentences randomly selected from his set. Date format contains a question uttered by the investigator and the response of the speaker. The corpus consists of an audio file for each recording and corresponding textual layer consisting of the phonetically normalised and the orthographically normalised annotation.

LDC-IL Sentence Aligned Speech corpus for Maithili is available in both Devanagari and Tirutha scripts. The 'Tirhuta' script was standardised by Unicode in June 2015. Since then there are only a few available digitised Maithili fonts, most of which often representing certain letters/symbols in handwritten style to complete the typeface. As a result, Maithili relies on the Devanagari script. The Tirhuta script is used for producing genealogical records, manuscripts of religious texts, personal correspondence, and printing books. Recognizing the significance of Tirhuta, LDC-IL has annotated and made Maithili speech data available in the script to support its propagation and promotion.

2.2 OBSERVATIONS

LDC-IL sentence-level speech annotation strictly follows what the speaker pronounces to produce the phonetically normalised annotation. The text has been written in the official script of the language and the speech is transcribed as narrowly as the script supports. Even if it is read speech data, there are widespread variations in the pronunciation. For example, speakers from different regions speak the same word in different ways. For example, in Samastipur region few speakers never pronounce /ətʃʰ/ and instead of /ətʃʰ/ they consistently pronounce it as /ətʃʰɪ/.

There were also variations in how numbers were pronounced. For example, while reading sports news, the speakers read scores of different sports such as cricket, tennis, etc. in different ways and they deviated from the standardised way of pronouncing the scores. Similarly, there were some errors in reading large numbers such as thousands or lakhs and also in reading decimals,

fractions, etc. Most of the speakers faced difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely-used words also influenced the reader's fluency.

2.2.1 PHONETIC ALTERNATION IN MAITHILI SPEECH DATA

Read speech has disfluencies like unwanted pauses, elongated syllables, word fragments, self-corrections, and repeated words. Some such disfluencies in the recording are given below:

a. Repetition of words

While reading, if the informant observes that the word hasn't been pronounced in correct or effective manner then normally the speaker repeats a part of that word, the whole word or sometimes even the phrase. Sometimes the speaker also struggles to read the text and keeps repeating when the content seems unfamiliar to him or there may be instances of foreign words or such words which are difficult to pronounce. These are mainly instances of self-correction.

b. False start

False start is a common phenomenon in most of the speakers and for some speakers the frequency increases. Usually, it is the replacement of the first word or a syllable of the word but sometimes speakers start with some other letter as well instead of the actual letter.
E.g.: b-vɪɖ̲jaː nɪrɖ̲ʰaː - nɪrɖ̲ʰaːrɪʈ̲

c. Addition and Deletion

An extra vowel or a consonant or a syllable is sometimes added into a word. The sound which already exists in the word might be repeated or a different sound might be inserted into the word.
E.g.: dʒɪmmeːvaːrɪ > dʒɪmmvaːriː raːkʰəl > rəkkʰəl

Deletion or elision of a vowel or a consonant or a syllable from a word is also a common phenomenon attested in the corpus.
E.g.: eːgaːrəɦ > gjaːrəɦ səŋgəʈʰn > səgəʈʰn

d. Assimilation and Dissimilation

Speech is a continuous syllabic fragment, so the articulatory organs influence the preceding or following sound. Consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment called assimilation.
E.g.: tʃəkr > tʃəkk

Dissimilation is dropping out a syllable or a letter by the influence of adjacent speech segments.
E.g.: kərɪjaː > kəɳɪjaː

e. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.

E.g.: ɒpəreːʃən > aːpəreːsən ɖaːkʈər > ɖaːgəɖər
The original form has been kept in the transcription.

f. Substandard alternation

It has been observed that some speakers have consistently replaced the aspirated sounds with their unaspirated counterparts.
E.g.: ʃaːsən > saːsən ʃɪkʂɪt̪ > ʃɪtʃtʃʰɪt̪

g. Phone variation

It is the alternative pronunciation of the word, and which does not affect the meaning. Both pronunciations are considered to be in free variations.
E.g.: gã̠ːɖ̪ʱi : > gaːnɖ̪ʱiː

h. Metathesis

There are instances when the speaker reluctantly speaks with the metathesis alteration. In this case, avoiding the incorrect word, the standard correct word has been transcribed.
E.g. /geːləhaː/is the word pronounced by the speaker whereas the correct form of the word should be /geːlaːh/. So, while transcribing, the correct form has been kept.

2.3 SUMMARY OF THE CORPUS

The total duration of Maithili Sentence Aligned Speech Corpus is 41:54:30 (hh:mm:ss) comprising 21,412 audio segments from 300 speakers. Figures 1, 2 and 3 show the distribution of the corpus with respect to gender, age and content type, respectively. Figure 4 and 5 show gender and age distributions for each content type respectively. Table 1 gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details. Table 2 shows the age and gender-wise distribution of all the speakers.

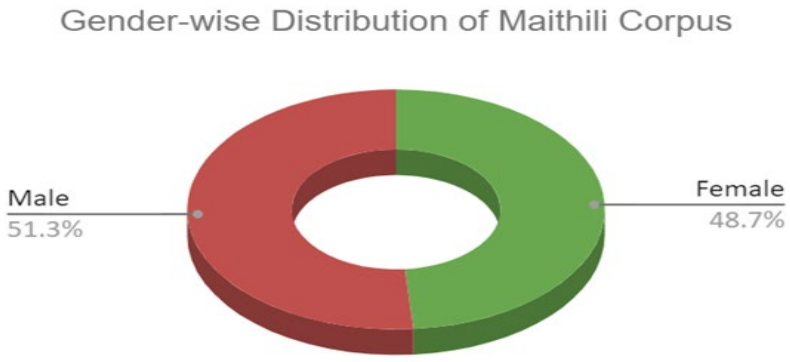

Figure 1: Gender-wise Distribution of Maithili Corpus

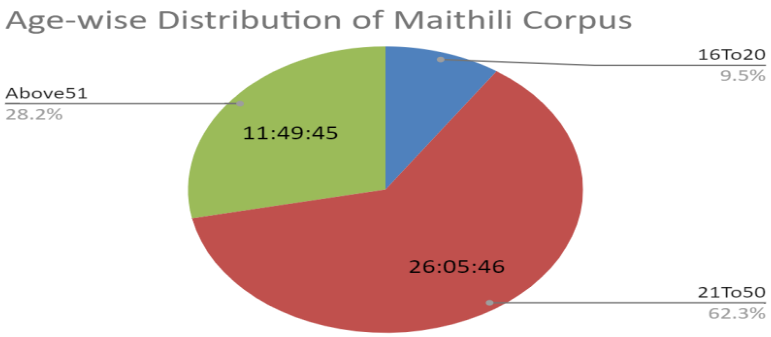

Figure 2: Age-wise Distribution of Maithili Corpus

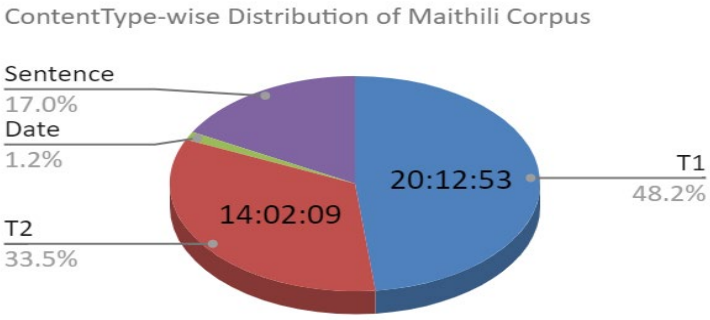

Figure 3: Content Type-wise Distribution of Maithili Corpus

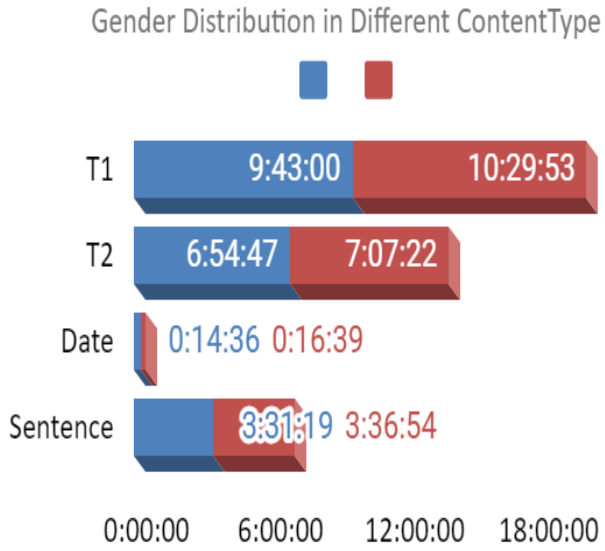

Figure 4: Gender Distribution in different Content Types of Maithili Corpus

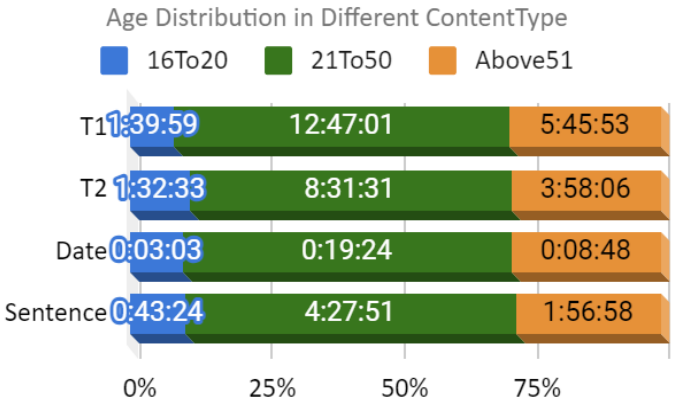

Figure 5: Age Distribution in different Content Types of Maithili Corpus

2.3.1 DURATION OF MAITHILI SENTENCE ALIGNED SPEECH DATA

The table below shows the duration of each of the content types and their distribution across a few factors in Maithili Sentence Aligned Speech Data.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:57:07.783085	09:42:59.775607	20:12:52.939966
		21To50	06:07:46.476309		
		Above51	02:38:05.516212		
	Male	16To20	00:42:51.628073	10:29:53.164359	
		21To50	06:39:14.185974		
		Above51	03:07:47.350312		
Creative Text-T2	Female	16To20	00:46:40.027174	06:54:46.948657	14:02:08.984780
		21To50	04:22:01.036929		
		Above51	01:46:05.884554		
	Male	16To20	00:45:52.533406	07:07:22.036123	
		21To50	04:09:29.594624		
		Above51	02:11:59.908092		
Date-D	Female	16To20	00:01:24.594480	00:14:36.302954	00:31:15.019064
		21To50	00:09:32.138534		
		Above51	00:03:39.569940		
	Male	16To20	00:01:38.264070	00:16:38.716110	
		21To50	00:09:51.571019		
		Above51	00:05:08.881021		
Sentence-S	Female	16To20	00:20:30.674311	03:31:19.119820	07:08:13.520495
		21To50	02:16:26.490076		
		Above51	00:54:21.955433		
	Male	16To20	00:22:53.650018	03:36:54.400675	
		21To50	02:11:24.829998		
		Above51	01:02:35.920659		

Table 1: Representation of Maithili Sentence Aligned Speech Data Duration

2.4 SUMMARY OF SPEAKERS

The table below shows the total number of speakers and their distribution in the Maithili Sentence Aligned Speech Data.

Age Group	Female	Male	Total
16To20	15	16	31
21To50	93	93	186
Above51	39	44	83
Total	14	153	300

Table 2 : Distribution of Speakers of Maithili Sentence Aligned Speech Data

2.5 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 Coordination 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 (ed.), Linguistic Resource For AI/NLP in Indian Languages. 2019. Central Institute of Indian Languages, Mysuru. ISBN No 978-81-7343-295-8
4. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh & Dinesh Mishra. 2019. A Gold Standard Maithili Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Arun Kumar Singh, Dinesh Mishra & Atuleshwar Jha. 2019. Maithili Raw Speech Corpus. Central Institute of Indian Languages, Mysore.
6. Shantanu Kumar, Dinesh Mishra, Rajesha N., Manasa G., Srikanth D., Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan. 2023. Maithili Sentence Aligned Speech Corpus Central Institute of Indian Languages, Mysore. 978-81-19411-96-2.