ꯃꯅꯤꯄꯨꯔꯤ ꯁꯦꯟꯇꯦꯟꯁ ꯑꯦꯂꯥꯏꯟꯗ ꯁ꯭ꯄꯤꯆ ꯀꯣꯔꯄꯁ
(ꯃꯤꯇꯩ ꯃꯌꯦꯛ)

Manipuri Sentence Aligned Speech Corpus (Meetei Mayek)

Manipuri Sentence Aligned Speech Corpus

MANIPURI SENTENCE ALIGNED SPEECH CORPUS (MEETEI MAYEK)

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

Manipuri Sentence Aligned Speech Corpus (Meetei Mayek)

Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N., Manasa G., Stephen Fernandes, Nithin S., Roopashri M. R., Dr. Narayan Kumar Choudhary, Prof. Shailendra Mohan

e-ISBN: 978-93-48633-96-5
CIIL Publication No.: 1504

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: Amom Nandaraj Meetei

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 MANIPURI SPEECH ANNOTATION

Amom Nandaraj Meetei., Yumnam Premila Chanu., Rajesha N., Manasa G., Nithin S.,

Dr. Narayan Kumar Choudhary., Prof. Shailendra Mohan

2.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Manipuri Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL (Ramamoorthy, L. et. Al, 2019). A detailed explanation of the Manipuri Raw Speech Corpus will be available in the Manipuri Speech Data Documentation (Nandaraj, A.M., et. Al, 2019). LDC-IL Manipuri Sentence Aligned Speech files contain an audio file and two textual layers in Manipuri scripts: Bengali and Meetei Mayek. 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
'Manipuri_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Manipuri 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 list - each speaker has typically recorded 25 sentences randomly selected from this 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 layers consisting of the phonetically normalised annotation and the orthographically normalised annotation.

2.2 OBSERVATIONS

The LDC-IL sentence-level speech annotation strictly adheres to the speaker's pronunciation to produce a phonetically normalised transcription. The text is written in the language's official script, with the transcription as precise as the script allows. Although the data consists of read speech, pronunciation varies widely. For example, speakers from different regions may pronounce the same word differently. In the Kakching dialect, the word ꯃꯗꯤ /mɑ-nə/ ('by him') exhibits a prominent low tone on the lexical root /mɑ/, which can be transcribed as /mɑ̀-nə/. In contrast, the Sekmai dialect features a long middle tone on the lexical root, transcribed as /mɑ̄-nə/.

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 MANIPURI SPEECH DATA

Read speech contains various disfluencies, including unintended pauses, elongated syllables, word fragments, self-corrections, and repetitions. In contrast, Manipuri also exhibits distinct phonological processes, including assimilation, dissimilation, consonant and vowel deletions and additions, as well as consonant strengthening. These differences are exemplified below.

(A) Some Disfluencies

a. Repetition of words

While reading, if the informant observes that the word has been pronounced not in correct or effective manner, then normally the speaker repeats part of the word, whole word or the phrase. Sometimes the speaker was struggling to read the text and repeats when the content is about unfamiliar subjects or there are many foreign words or words which are difficult to pronounce. These are mainly instances of self-correction.

b. False start

False starts are common among most speakers and occur frequently in some. They typically involve the repetition of the first word or syllable, though speakers may occasionally begin with a different letter.

E.g. ꯃꯤ-ꯃꯇꯝ /mə-mətəm/ ‘time’; ꯂꯥꯡ-ꯂꯥꯡꯇꯥꯀꯐꯗ /laŋ- laŋtəknəbə/

c. Intended speech

Intended speech occurs when the speaker slows down or fastens up their speech. Typically, it happens at the end of the sentence. It has resulted in inaudible speech in some instances.

For instance, if an audio segment is transcribed as **methəkıdu[də]**, it indicates that **[də]** is not clearly audible. In longer words, certain syllables or phonemes—particularly those in the middle—may be unclear to the listener or inadvertently skipped by the speaker. For example, in **ui[rum]nərəbə**, the middle portion **[rum]** is not distinctly audible.

(B) Phonological Processes

Phonological processes arise as articulatory or perceptual phenomena. When morphemes combine to form words, adjacent segments interact and may undergo changes.In Manipuri, based on the available corpus, these processes can be classified into four categories: assimilation, dissimilation, consonant and vowel deletions and additions, and consonant strengthening.

a. Assimilation

Speech is a continuous stream of syllabic fragments, where articulatory organs influence the preceding sound (progressive assimilation), the following sound (regressive assimilation), or cause two sounds to assimilate to each other (reciprocal assimilation).

Progressive Assimilation

ꯆꯥꯏꯅ /capə/ > ꯆꯥꯗ /cabə/ ‘eating’

ꯆꯥꯏꯪ /cate/ > ꯆꯥꯕ /cade/ ‘not eating’

Regressive Assimilation

ꯄꯦꯟꯂꯪ /penle/ > ꯄꯦꯎꯂꯪ /pelle/ ‘satisfied’

ꯖꯪꯟꯂꯪ /jenle/ > ꯖꯪꯎꯂꯪ /jelle/ ‘distributed’

Reciprocal Assimilation

ꯑꯃꯀꯣꯟ /nəmkən/ > ꯅꯋꯥꯡꯒꯟ /nəŋgən/ ‘back (body)

ꯁꯦꯟꯄꯦꯟ /sənpən/ > ꯁꯝꯕꯦꯟ /səmbən/

b. Dissimilation ‘fence’

Dissimilation is the opposite of assimilation. In Manipuri, certain suffixes beginning with aspirated stops change into unaspirated ones when following roots that end in an unaspirated stop, representing a case of progressive dissimilation.

Examples:

ꯍꯤꯍꯧꯗ /tʰəktʰok/ > ꯍꯤꯇꯧꯗ /tʰəktok/ ‘drink off’

ꯄꯨꯀꯨꯟ /pʰutkʰət/ > ꯄꯨꯀꯤꯟ /pʰutkət/ ‘begin to boil’

c. Deletion

The deletion of consonants and vowels simplifies the syllabic structure into a CVCV pattern. As shown below, the lateral /l/ is omitted after a syllable ending in the voiceless velar /k/, while the low vowel /ɑ/ is deleted before the mid vowel /ə/, further streamlining the syllable structure.

Consonant Deletion

ꯀꯏꯔꯗ /kəkləbə/ > ꯀꯏꯊꯗ /kək’əbə/ ‘the cut one’

ꯍꯦꯏꯔꯗ /hekləbə/ > ꯍꯦꯏꯊꯗ /hek’əbə/ ‘the plucked one’

Vowel Deletion

ꯀꯥ ꯊꯨ /kɑ + ədu/ > ꯀꯥꯨ /kɑdu/ ‘that room’

ꯆꯥ ꯊꯎꯃ /cɑ + ənɪ/ > ꯆꯎꯃ /cənɪ/ ‘two hundred’

d. Addition

The addition of sounds occurs in both consonants and vowels, though with varying frequency. In Manipuri, consonant addition is rare, whereas vowel addition is more common, particularly in loanwords. Vowels are frequently inserted in initial (prosthesis), medial (epenthesis), and final (epithesis) positions, shaping the phonological structure of borrowed words.

Prothesis

ꯁ꯭ꯀꯨꯜ /skul/ > ꯏꯁ꯭ꯀꯨꯜ /ɪskul/ ‘school’

ꯁ꯭ꯇꯦꯟꯗꯔꯗ /stendərd/ > ꯏꯁ꯭ꯇꯦꯟꯗꯔꯗ /ɪstendərd/ ‘standard’

Epenthesis

ꯖ꯭ꯔꯥꯁ /gjɑs/ > ꯖꯤꯔꯥꯁ /gɪjɑs/ ‘gyas’

ꯅ꯭ꯔꯀ /nərk/ > ꯅꯔꯀꯗ /nərək/‘hell’

Epithesis

ꯑ꯭ꯔꯃ /dhərm/ > ꯑꯔꯃꯁ /dhərmə/ ‘religion’

ꯇ꯭ꯔꯀ /tərk/ > ꯇꯔꯀꯗ /tərkə/ ‘argument’

e. Consonant strengthening

Consonant strengthening involves the reinforcement of a segment, often manifesting as the transformation of a non-geminate into a geminate or double consonant. In Manipuri, as observed in the speech corpus, when a suffix beginning with the lateral /l/ is added to roots ending in /p/, /m/, or /ŋ/, it undergoes gemination to strengthen the segment.

Examples:

ꯇꯔꯄꯤ /təple/ > ꯇꯔꯄꯄꯤ /təppe/ ‘have been slow’

ꯀꯨꯃꯤ /kʰumle/ > ꯀꯨꯃꯃꯤ /kʰumme/ ‘have cried’

ꯍꯥꯡꯤ /hɑŋle/ > ꯍꯥꯡꯡꯤ /hɑŋŋe/ ‘have opened’

(C) Colloquial usage

Some speakers have pronounced colloquial forms instead of the standardized forms provided in the prompt sheet. For example:

ꯑꯉꯚꯒꯤ /hɑɪbəsɪ/ → ꯑꯉꯚꯒꯍ /hɑɪbəse/ (‘that means’)

ꯍꯪꯗꯤꯡꯒꯤ /mətəmsɪ/ → ꯍꯪꯗꯤꯡꯒꯍ /mətəmse/ (‘this time’ or ‘in this modern time’)

ꯀꯔꯤꯅꯣ /kərɪno/ → ꯀꯦꯣꯅꯣ /kəɪno/ (‘what is it’)

The most frequently occurring colloquial functional morphemes in the speech corpus are demonstrative forms, as highlighted below:

Demonstrative Forms

Demonstratives	**Formal**	**Informal**	**Meaning**
Proximal	ꯑꯁꯤ/əsɪ/	ꯑꯁꯍ /əse/	‘this’
Distal	ꯑꯗꯨ /ədu/	ꯑꯗꯣ /ədo/	‘that’
Pronominal			
Proximal	ꯍꯁꯤ /məsɪ/	ꯍꯁꯍ /məse/	‘this’
Distal	ꯍꯗꯨ /mədu/	ꯍꯗꯣ/mədu/	‘that’

These variations reflect common colloquial tendencies in spoken Manipuri.

(D) Free Variation in Manipuri Phonology

Free variation occurs when a word has alternative pronunciations without affecting its meaning. In such cases, both variants are phonologically equivalent and interchangeable within a given context.

In Manipuri, this phenomenon is evident in lexical items ending in either /l/ or /r/, where both forms carry the same meaning. This phonetic flexibility, as reflected in the speech corpus, highlights a broader pattern of phonological variability in the language.

E.g.: ꯀꯦꯀꯣꯉ /ləɪkon/ > ꯀꯦꯀꯣꯜ /ləɪkol/ ‘garden’; ꯇꯨꯔꯦꯉ /turen/ > ꯇꯨꯔꯦꯜ /turel/ ‘river’

(E) Allophone Realization in Manipuri

Allophones, which are variant forms of a phoneme, occur in complementary distribution. In Manipuri, as observed from the analyzed corpus, four vowel phonemes—/ɪ/, /e/, /u/, and /o/—exhibit allophonic variations.

a. The Phoneme /ɪ/

The vowel phoneme /ɪ/ is realized as [jɪ] when it occurs as the initial sound of the second syllable following the vowels /ə/ and /o/:

ꯑꯉꯂ /əɪbə/ → ꯑꯌꯤꯂ [əjɪbə] (‘writer’)

ꯀꯣꯉꯩ /loɪn/ → ꯀꯣꯌꯤꯩ [lojɪn] (‘bodyguard’)

When /ɪ/ follows the vowel /u/, it is realized as [wɪ]:

ꯇꯣꯏꯉ /təuɪ/ → ꯇꯣꯏꯅꯤ [təuwɪ] (‘is done’)

b. The Phoneme /e/

The phoneme /e/ is realized as [je] when it appears at the beginning of the second syllable following the vowels /e/ and /u/:

ꯇꯩꯊ꯭ /tee/ → ꯇꯩꯌꯩ [teje] (‘is tame’)

ꯄꯨꯊ꯭ /pue/ → ꯄꯨꯌꯩ [puje] (‘have carried’)

Additionally, when /e/ follows the diphthongs /əu/ and /ɑu/, it is realized as [we]:

ꯀꯣꯎꯊ꯭ /kəue/ → ꯀꯣꯎꯅꯩ [kəuwe] (‘it is short’)

ꯆꯦꯎꯊ꯭ /cɑue/ → ꯆꯦꯎꯅꯩ [cɑuwe] (‘it is big’)

c. The Phoneme /u/

The phoneme /u/ is realized as [wu] when it occurs at the beginning of the second syllable after the central vowel /ə/:

ꯊꯎꯕꯗ /əubə/ → ꯊꯨꯕꯗ [əwubə] (‘one who sees’

When /u/ follows the vowel /ɪ/, it is realized as [ju]:

ꯊꯤꯎ /thiu/ → ꯊꯤꯨ [thju] (‘search’)

d. The Phoneme /o/

The phoneme /o/ is realized as [wo] when it appears at the beginning of the second syllable following the central vowel /ə/ and the low vowel /ɑ/:

ꯊꯊꯣꯏꯗ /əonbə/ → ꯊꯅꯣꯏꯗ [əwonbə] (‘something that changes’)

When /o/ follows the vowel /ɪ/, it is realized as [jo]:

ꯃꯤꯊꯣꯍ /mɪoŋ/ →ꯃꯤꯌꯣꯍ [mɪjoŋ] (‘type of man’)

This analysis highlights phonetic alternations in Manipuri speech data, following the LDC-IL sentence-level speech annotation system. The phonetic normalization process strictly adheres to the speaker’s pronunciation, ensuring an accurate transcription.

2.3 SUMMARY OF THE CORPUS

The **Manipuri Sentence Aligned Speech Corpus**, presented in both **Bengali Script** and **Meetei Mayek**, spans an impressive total duration of 123:29:55 (hh:mm:ss). This extensive dataset is composed of **60,819 meticulously curated audio segments,** contributed by **589 speakers,** showcasing the linguistic richness and diversity of Manipuri speech.

The table below provides a comprehensive breakdown of the duration across different content types, along with its distribution across various key factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition and structure.

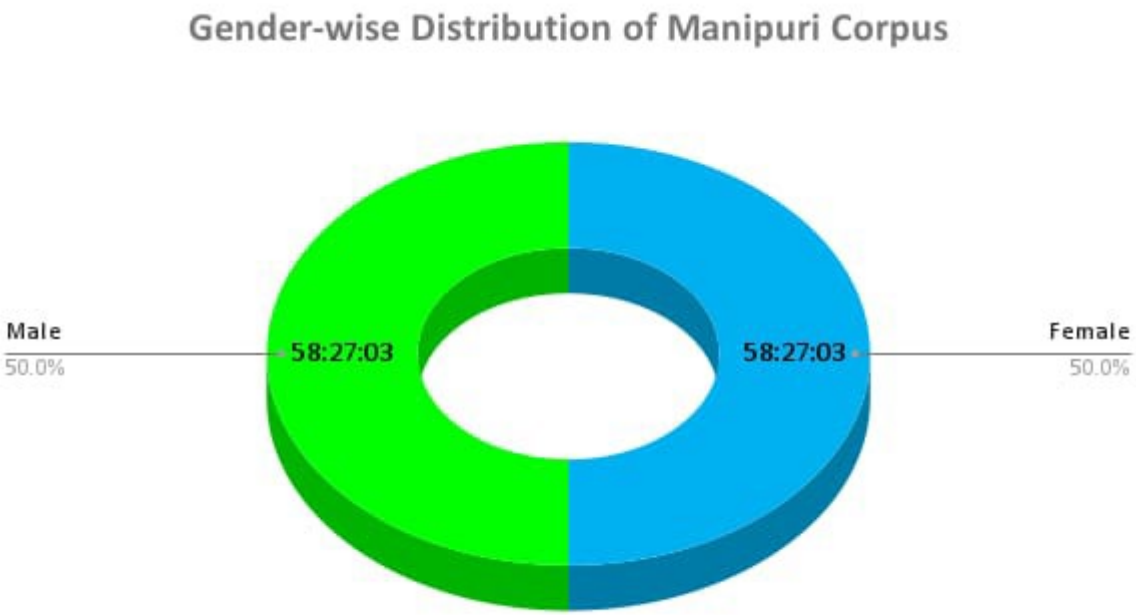

Figure 1: Gender-wise Distribution of Manipuri Corpus

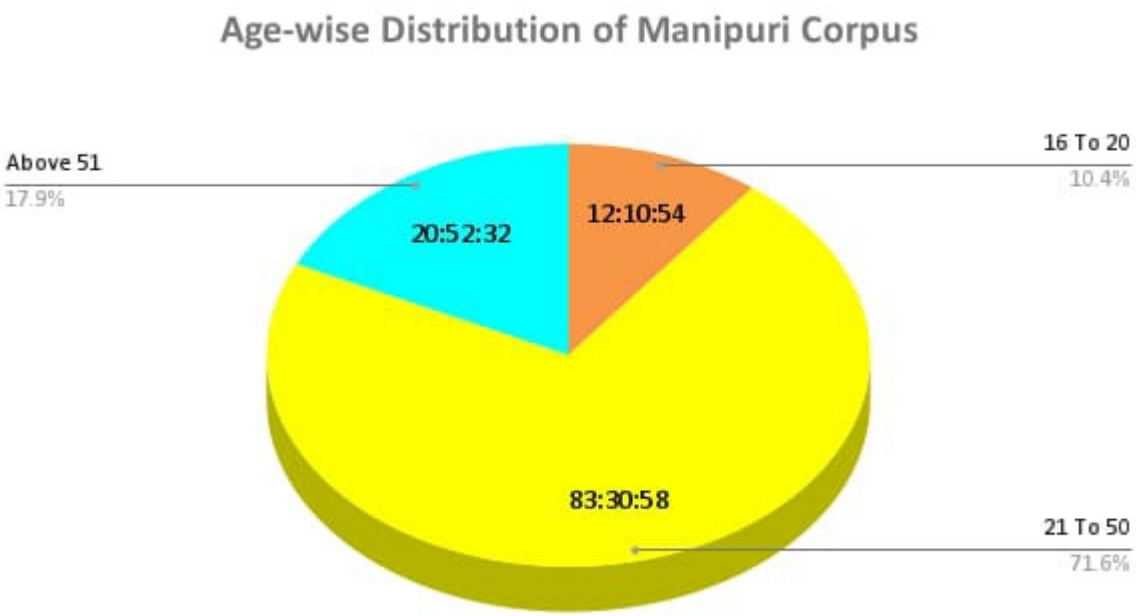

Figure 2: Age-wise Distribution of Manipuri Corpus

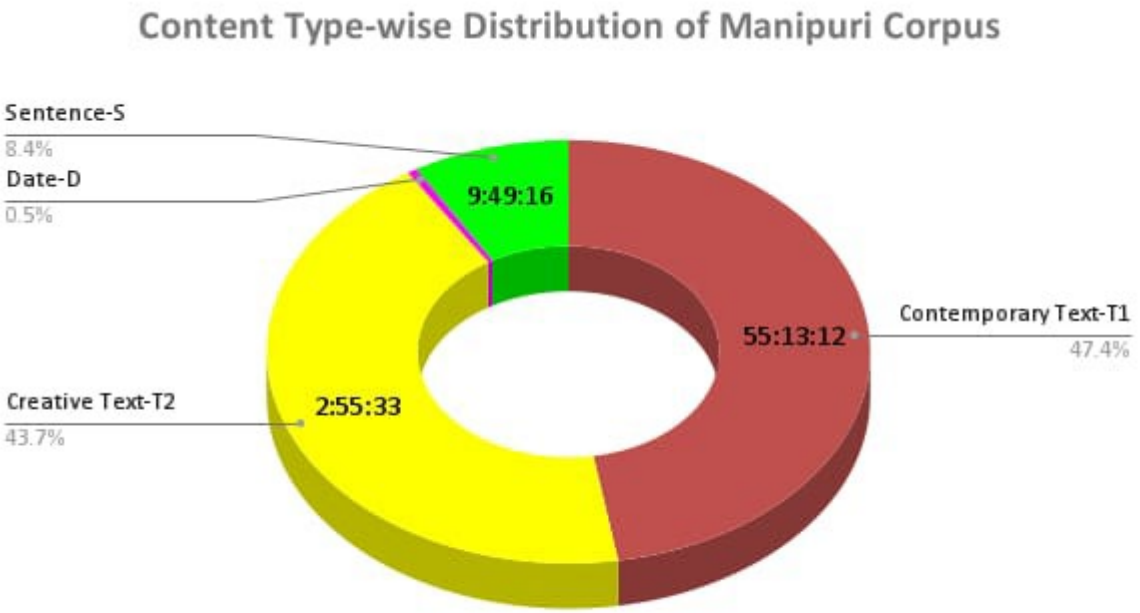

Figure 3: Content Type-wise Distribution of Manipuri Corpus

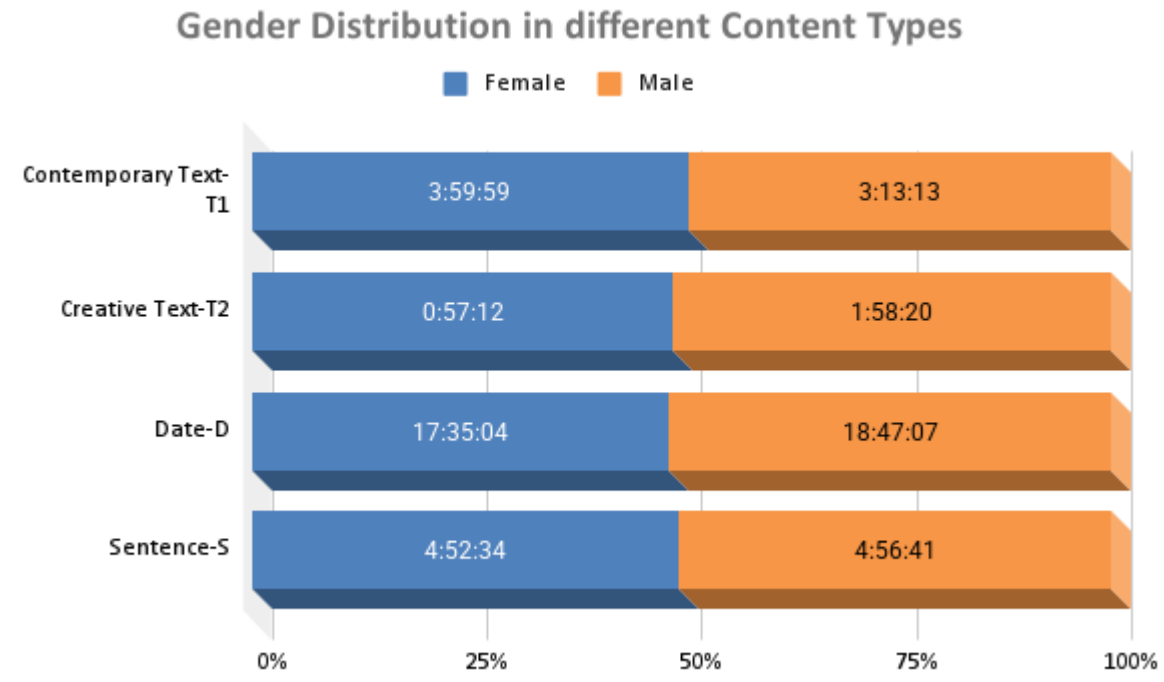

Figure 4: Gender Distribution in different Content Types of Manipuri Corpus

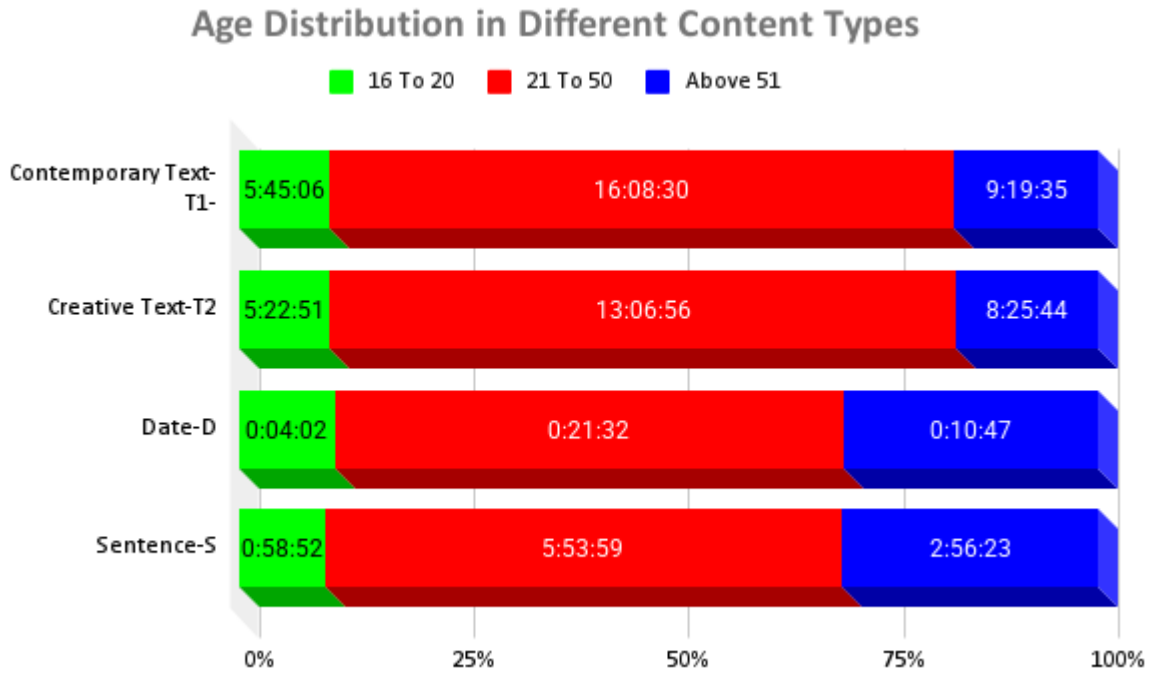

Figure 5: Gender Age Distribution in different Content Types of Manipuri Corpus

2.3.1 DURATION OF MANIPURI SENTENCE ALIGNED SPEECH DATA

The table below presents a detailed breakdown of the duration for each content type and its distribution across various factors within the **Manipuri Sentence Aligned Speech Data,** offering valuable insights into its composition.

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	02:35:05.004015	27:59:59.297727	55:13:12.425801
		21To50	21:00:47.830039		
		Above51	04:24:06.463674		
	Male	16To20	03:10:01.245593	27:13:13.128074	
		21To50	19:07:42.542715		
		Above51	04:55:29.339766		
Creative Text-T2	Female	16To20	02:05:28.540757	24:57:12.225164	50:55:33.008849
		21To50	18:53:48.854679		
		Above51	03:57:54.829729		
	Male	16To20	03:17:23.378449	25:58:20.783684	
		21To50	18:13:07.432186		
		Above51	04:27:49.973050		
Date-D	Female	16To20	00:01:58.661271	00:17:35.404161	00:36:23.094865
		21To50	00:10:08.222452		
		Above51	00:05:28.520438		
	Male	16To20	00:02:04.265103	00:18:47.690703	
		21To50	00:11:24.120978		
		Above51	00:05:19.304622		
Sentence-S	Female	16To20	00:28:32.915456	04:52:34.476693	09:49:16.126820
		21To50	02:58:55.249092		
		Above51	01:25:06.312145		
	Male	16To20	00:30:19.983125	04:56:41.650127	
		21To50	02:55:04.294377		
		Above51	01:31:17.372624		

Table 1: Representation of Manipuri Sentence Aligned Speech Data Duration

2.4 SUMMARY OF SPEAKERS

The table below provides a comprehensive overview of the total number of speakers and their distribution within the **Manipuri Sentence Aligned Speech Data.**

Age Group	Female	Male	Total
16To20	29	34	63
21To50	194	187	381
Above51	72	73	145
Total	295	294	589

Table 2: Distribution of Speakers of Manipuri 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 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. Amom Nandaraj Meetei, Yumnam Premila Chanu, Rajesha N, Manasa G, Narayan Choudhary & L. Ramamoorthy. 2019. "Documentation of LDC-IL Manipuri Raw Speech Corpus" in Linguistic Resources for AI/NLP in Indian Languages. Central Institute of Indian Languages, Mysore. pp. 91-103.
5. Ramamoorthy, L., Narayan Choudhary, Amom Nandaraj Meetei, Yumnam Premila Chanu &LongjamAnand Singh. 2019. Manipuri Raw Speech Corpus. Central Institute of Indian Languages, Mysore.