বাংলা বাক্যের সমরেখাবদ্ধ বাক্ কর্পাস

Bengali Sentence Aligned Speech Corpus

BENGALI SENTENCE ALIGNED 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: Bengali Sentence Aligned Speech Corpus
Authors: Sonali Sutradhar, Poulami Das, Rajesha N., Manasa G., Srikanth D.,Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan

ISBN: 978-81-19411-48-1
CIIL Publication No.: 1426

First published: AD 2023 September
Bhadrapada 1945 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: Sonali Sutradhar, Poulami Das

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 ‘avanviittipoyi’ ‘He went home’ in an informal way then it is annotated as ‘avanviittilpoyi’ 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 “kekeralam”. 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 BENGALI SPEECH ANNOTATION

Sonali Sutradhar, Poulami Das, Narayan Kumar Choudhary

2.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Bengali Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Bengali Raw Speech Corpus will be available in the Bengali Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Bengali Sentence Aligned Speech files contain an audio file and two textual layers in Bengali 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
'Bengali_Female_16To20_Contemporary Text-T1_SP-0031_T1-0031-001.wav'

LDC-IL Sentence Aligned Speech corpus for Bengali 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 the response of the speaker to the investigator. The corpus consists of an audio file for each recording and corresponding two textual layers consisting of the phonetically normalised annotation and the orthographically normalised annotation.

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 a 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, few speakers from North Bengal region have been found to use /s/ instead of /ʧʰ/ as in /e maduli to tar babari silo/ "এমাদুলি তো তার বাবারি সিলো" instead of /e maduli to tar babari ʧʰilo/ "এমাদুলি তো তার বাবারি ছিলো".

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. For example:
/sʈibʱ oagʱ/ "স্টীভ ওয়াঘ" for Steve Wagh, /ɾiki pɔenʈiŋ/ "রিকি পয়েন্টিং" for Ricky Ponting, /orʤun ɾɔnotuŋgo/ "অর্জুন রনতুঙ্গো" for Arjuna Ranatunga etc.

2.2.1 PHONETIC ALTERNATION IN BENGALI 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 has been pronounced not in correct or effective manner then normally the speaker repeats a 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 were many foreign words or words which are difficult to pronounce. These are mainly instances of self-correction. For example:
/bʱarot konokrome potidʒogaite potidʒogitae ʈikia atʃʱe/ "ভারোত কুনুক্রোমে পোতিজোগাইতে পোতিজোগিতায় টিকিয়া আছে।"

b. False start

False start is a common phenomenon in most of the speakers and in some speakers it is frequent. Usually, it is the repetition of the first word or syllable of the word but sometimes speakers start with some other letter as well.
E.g.: /ʃe ʃei ʃɔkol o-obʱinoe pʰiria dekʰibaɾ puɾbe ekɔtʰa ʃikaɾ koria lɔoa uʧit/ "শে-শেই শকোল ও-ওভিগ্নতায় ফিরিয়া দেখিবার পুর্বে একথা শিকার কোরিয়া লওয়া উচিত।"

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 example,
/ʃe kʰub taɽataɽi poũʧʰe gælo/ "সে খুব তাড়াতাড়ি পৌঁছে গেল" has been pronounced as /ʃe kʰub taɽataɽi poũʧʰeælo/ "সে খুব তাড়াতাড়ি পৌঁছেঅ্যালো".
/ʃe ɾodʒ ɾate baɽi pʰeɾe/ "সে রোজ রাত করে বাড়ি ফেরে" has been pronounced as /ʃe ɾodʒ ɾate baɽi pʰe/ "সে রোজ রাত করে বাড়ি ফে".

d. Addition and Deletion

An extra vowel or a consonant or a syllable is sometimes added into a word. Instances of anaptyxis, that is, the phonological process where a vowel is added or inserted between two consonants, has been found. For example:
/gɾameɾ pɔɾ gɾam udʒaɽ ɦoe gælo/ "গ্রামের পর গ্রাম উজাড় হয়ে গেল" has been pronounced as /geɾameɾ pɔɾ geɾam udʒaɽ ɦoe gælo/ "গেরামের পর গেরাম উজাড় হয়ে গেল".
Apocope, that is, the phenomenon of deletion or elision of a vowel or a consonant is also common in the given corpus. For example:
/æmoni ʃɔbʰab taɾ/ "এমন ইস্বভাব তার" is pronounced as /æmoni ʃɔbʰab ta/ "এমন ইস্বভাব তা" where the word final /ɾ/ 'র' has been deleted.

e. Assimilation and Dissimilation

Speech is a continuous syllabic fragment, so the articulatory organs influence the preceding or the following sound. When a consonant or vowel is changed to a similar sound because of the influence of a nearby speech segment, then it is called assimilation. Examples of assimilation have been found as a result of mispronunciation, which are as follows:

/matɾo kɔekʈi deʃei ʃudʱu kiketʼ kʰæla ɦoibe/ "মাত্র কয়েকটি দেশেই শুধু কিকেট খেলা হইবে" has been pronounced as /matto kɔekʈi deʃei ʃudʱu kiketʼ kʰæla ɦoibe/ "মাত্তো কয়েকটি দেশেই শুধু কিকেট খেলা হইবে"

The phonological process in which one of the two identical or closely related sounds in a word is changed or omitted is called dissimilation. For example:

/sɔmman/ "সম্মান" has been pronounced as /sɔnman/ "সন্মান" in most of the cases, which is also a free variation in Bengali.

f. Colloquial usage

Some of the speakers have pronounced colloquial forms instead of the standardised form written in the prompt sheet.
E.g.: /biɦɔttɔɾ paŋgone ʃapollo aʃile kunuo kʰælaɾ poʃareɾ pɔtʰ ʃugam ɦɔe/ "বিহত্তোর পাঙ্গনে শাপোল্লো আশিলে কুনুও খেলার পোশারের পথ শুগম হয়।"

g. Substandard alternation

It has been observed that some speakers have consistently replaced the voiceless and aspirated sounds with their voiced and unaspirated counterparts and vice-versa.
For example:
/ʃɔʤag/ "সজাগ" becomes /ʃɔʤak/ "সজাক". /ɾakʰʧʰi/ "রাখছি" becomes /ɾakʧʰi/ "রাকছি" or /ɾakʰʧi/ "রাখচি" and /obʱibhabok/ "অভিভাবক" becomes /obʱibabok/ "অভিবাবোক" or /obʱibʱabok/ "অবিভাবোক".

Phone variation
It is the alternative pronunciation of the word which does not affect the meaning. Both pronunciations are considered to be in free variations. For example:

E.g.: /kak/ "কাক" is pronounced as /kag/ "কাগ", /hat/ "হাত" as/hatʰ/ "হাথ" etc.

h. Final vowel modification

In continuous speech, at times the final vowel gets modified in the speech of some of the speakers: For example:
/labʱ/ "লাভ" has been pronounced as /lab/ "লাব" and /gaʧʰ/ "গাছ" as /gaʧ/ "গাচ" (because aspirated sounds at the word final position are generally not realised).

i. Compound word splitting

Long agglutinated words have been read in such a way that a pause is at the point of joining and that interrupts the natural flow of language.
/æto keleŋkaɾi eboŋ kutʃit noŋɾa pokkʰopatduʃʈo ampajaɾiŋ kɾiketʈ maʈʰ dækʰeni/ "অ্যাতো কেলেঙ্কারি এবং কুতশিত নোংরা পোক্খোপাতদুশ্টো আম্পায়ারিং ক্রিকেট মাড দ্যাখেনি" has been pronounced as /æto keleŋkaɾi eboŋ kutʃit noŋɾa pokkʰopat duʃʈo ampajaɾiŋ kɾiketʈ maʈʰ dækʰeni/ "অ্যাতো কেলেঙ্কারি এবং কুতশিত নোংরা পোক্খোপাত দুশ্টো আম্পায়ারিং ক্রিকেট মাড দ্যাখেনি".
/ɔpɾottaʃitobʱabe ʃe kadʒʈa koɾte ʃɔkkʰɔm ɦolo/ "অপ্রত্যাশিতভাবে সে কাজটা করতে সক্ষ মহল" has been pronounced as /ɔ pɾottaʃitobʱabe ʃe kadʒʈa koɾte ʃɔkkʰɔm ɦolo/ "অ প্রত্যাশিতভাবে সে কাজটা করতে সক্ষম হল".

Spelling Parameters according to pronunciation:

Following are the spelling parameters which were followed while dealing with the thrid party evaluation of transcription process:

1. Bengali has two graphemes for the sound /ʤ/, which are 'জ' and 'য'. 'য' occurs only in the word initial position in Bengali while, 'জ' occurs in all the three positions. But, while dealing with the sentences at the pronunciation level only 'জ' was considered as the symbol for representing the sound /ʤ/. For example:

   For both /ʤaɦaʤ/ জাহাজ and /ʤatɾa/ যাত্রা 'জ' is used at the pronunciation level.

2. An instance of over-differentiation which is present in Bangla has been observed over here and have dealt in the following manner:

   The grapheme 'এ' has two different realisations at the pronunciation level - /e/ and /æ/ which have been denoted with 'এ' and 'অ্যা' respectively.
   এদিন remains /edin/ এদিন but, এমন becomes /æmon/ অ্যামোন.

   The grapheme 'অ' also has two different realisations at the pronunciation level - /ɔ/ and /o/ which have been denoted with 'অ' and 'ও' respectively. For example:
   অরুণ is pronounced as /oɾun/ওরুন and অনেক is pronounced as /ɔnek/ অনেক.

2.3 SUMMARY OF THE CORPUS

The following section provides a tabular detail of the various content types of the Bengali Sentence Aligned Speech Corpus. These figures may be helpful in tuning the corpus for various purposes like training, testing and evaluating various algorithms as well as for providing useful insights into the dataset. The total duration of Bengali Sentence Aligned Speech Corpus is 69:10:03 (hh:mm:ss), comprising 40,240 audio segments from 450 speakers.

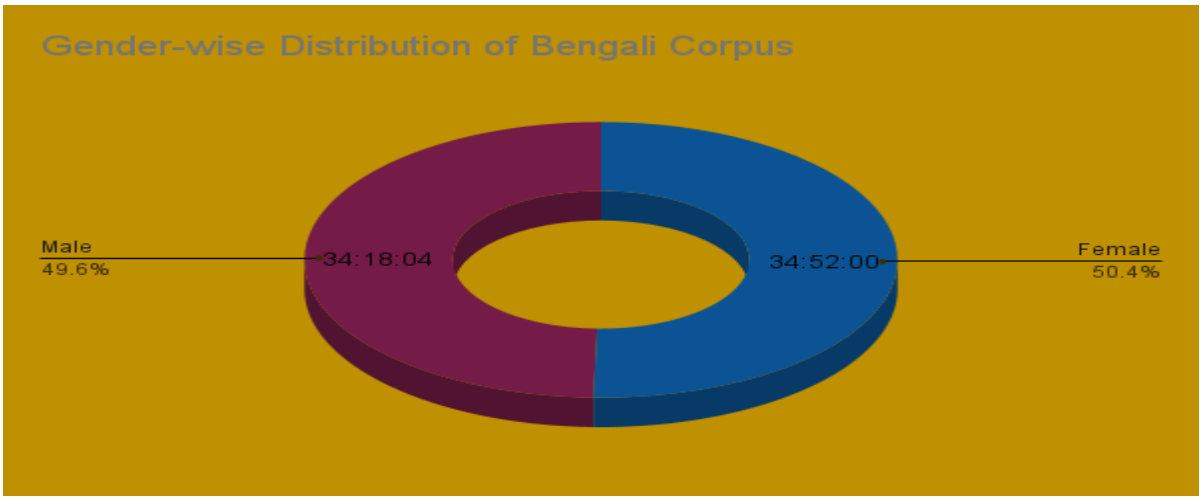

Figure 1: Gender-wise Distribution of Bengali Corpus

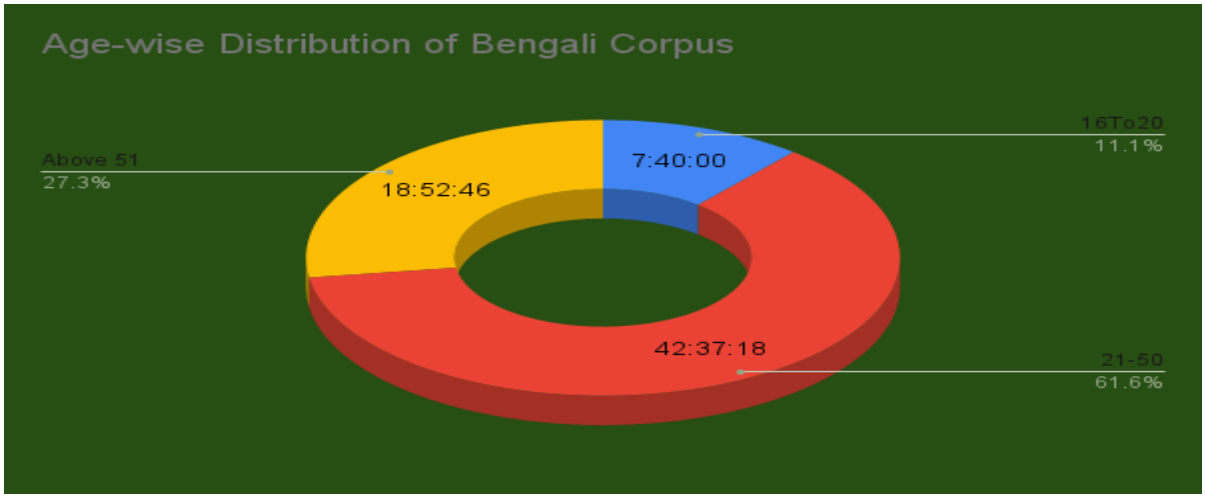

Figure 2: Age-wise Distribution of Bengali Corpus

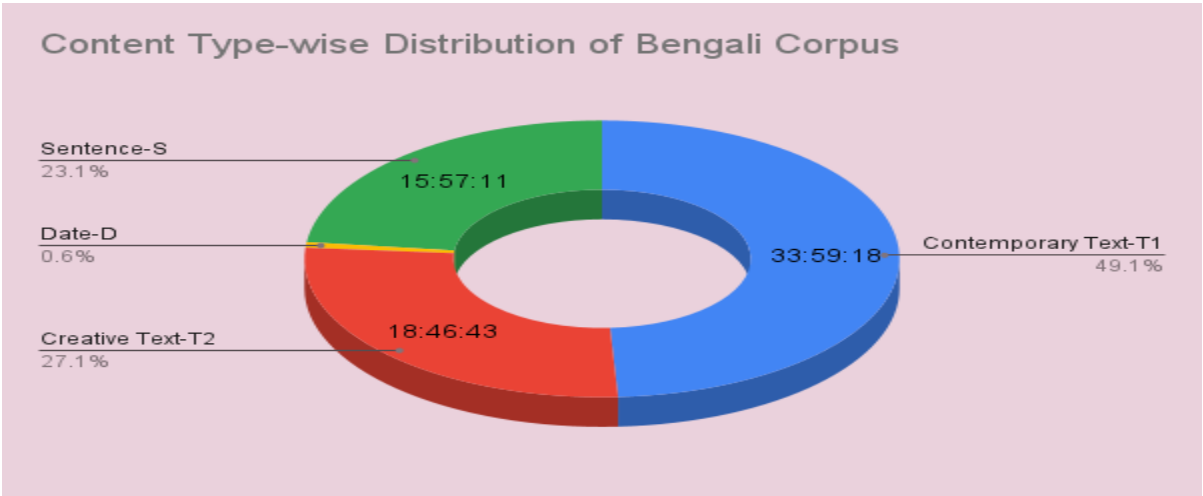

Figure 3: Content Type-wise Distribution of Bengali Corpus

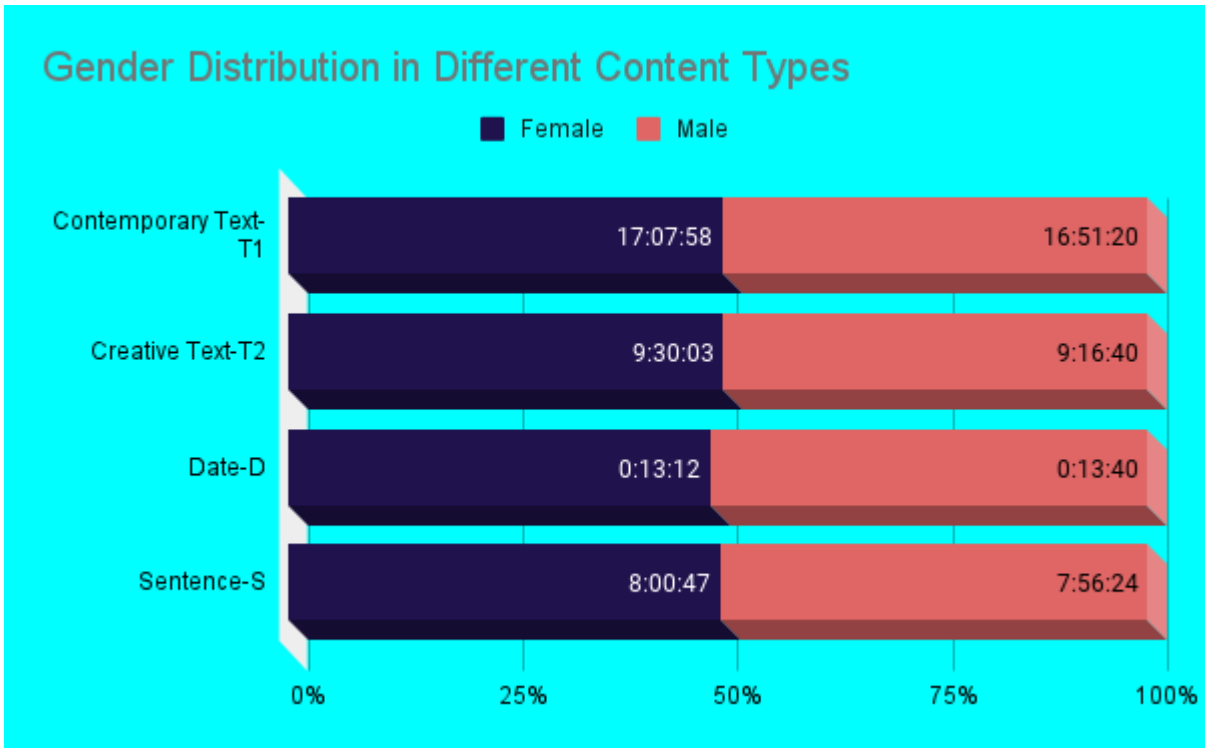

Figure 4: Gender Distribution in different Content Types of Bengali Corpus

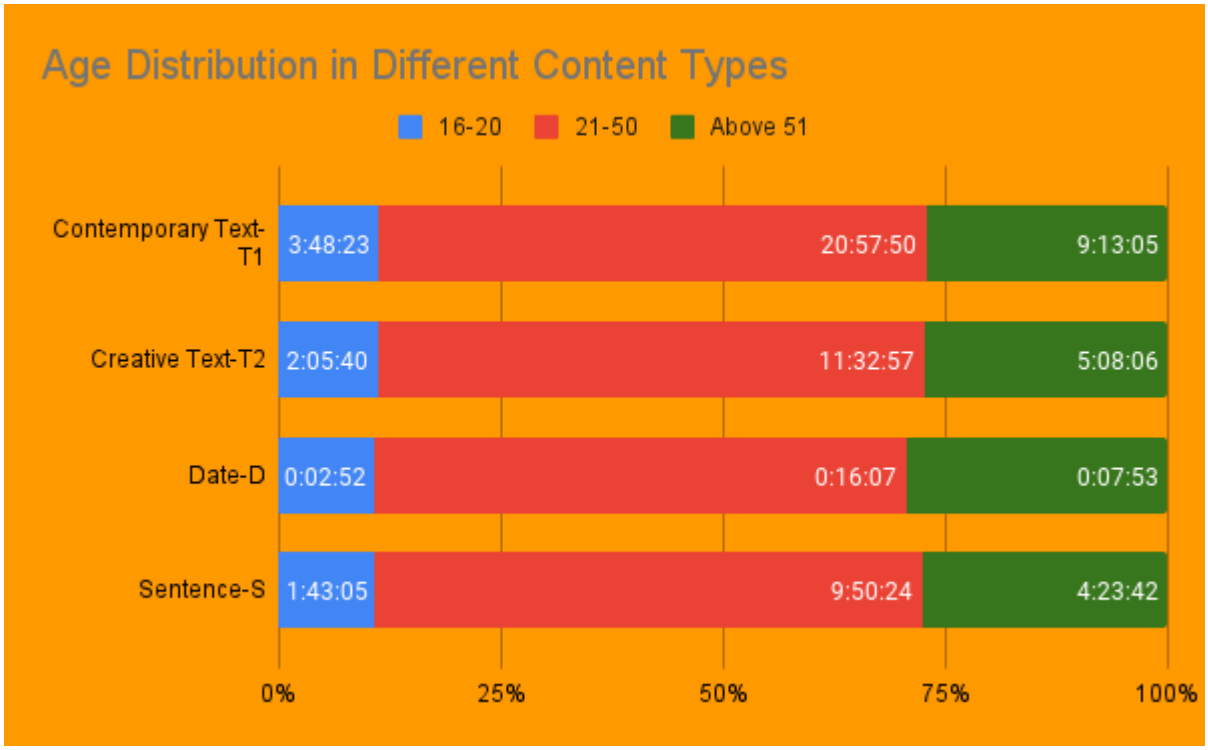

Figure 5: Age Distribution in different Content Types of Bengali Corpus

2.3.1 DURATION OF BENGALI SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	01:48:55.629907	17:07:58.214299	33:59:18.367872
		21To50	11:01:32.310071		
		Above51	04:17:30.274321		
	Male	16To20	01:59:27.462572	16:51:20.153573	
		21To50	09:56:17.504105		
		Above51	04:55:35.186896		
Creative Text-T2	Female	16To20	00:59:34.200795	09:30:02.949692	18:46:43.237346
		21To50	06:04:42.212154		
		Above51	02:25:46.536744		
	Male	16To20	01:06:06.272982	09:16:40.287654	
		21To50	05:28:15.134284		
		Above51	02:42:18.880389		
Date-D	Female	16To20	00:01:16.300000	00:13:12.039920	00:26:51.318884
		21To50	00:08:21.999960		
		Above51	00:03:33.739960		
	Male	16To20	00:01:35.529080	00:13:39.278964	
		21To50	00:07:44.909884		
		Above51	00:04:18.840001		
Sentence-S	Female	16To20	00:48:18.069773	08:00:46.585997	15:57:10.353052
		21To50	05:07:02.533275		
		Above51	02:05:25.982949		
	Male	16To20	00:54:46.890784	07:56:23.767055	
		21To50	04:43:21.308463		
		Above51	02:18:15.567808		

Table 1: Representation of Bengali Sentence Aligned Speech Data Duration

2.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	24	27	51
21To50	142	135	277
Above51	57	65	122
Total	223	227	450

Table 2: Distribution of Speakers of Bengali 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. Ramamoorthy, L., Narayan Choudhary, Sonali Sutradhar, Arundhati Sengupta, Sankarshan Dutta, Priyanka Das & Saswati Karmakar. 2019. A Gold Standard Bengali Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy, L., Narayan Choudhary, Sonali Sutradhar, Priyanka Biswas, Arundhati Sengupta, Sankarshan Dutta & Priyanka Das. 2019. Bengali Raw Speech Corpus. Central Institute of Indian Languages, Mysore.