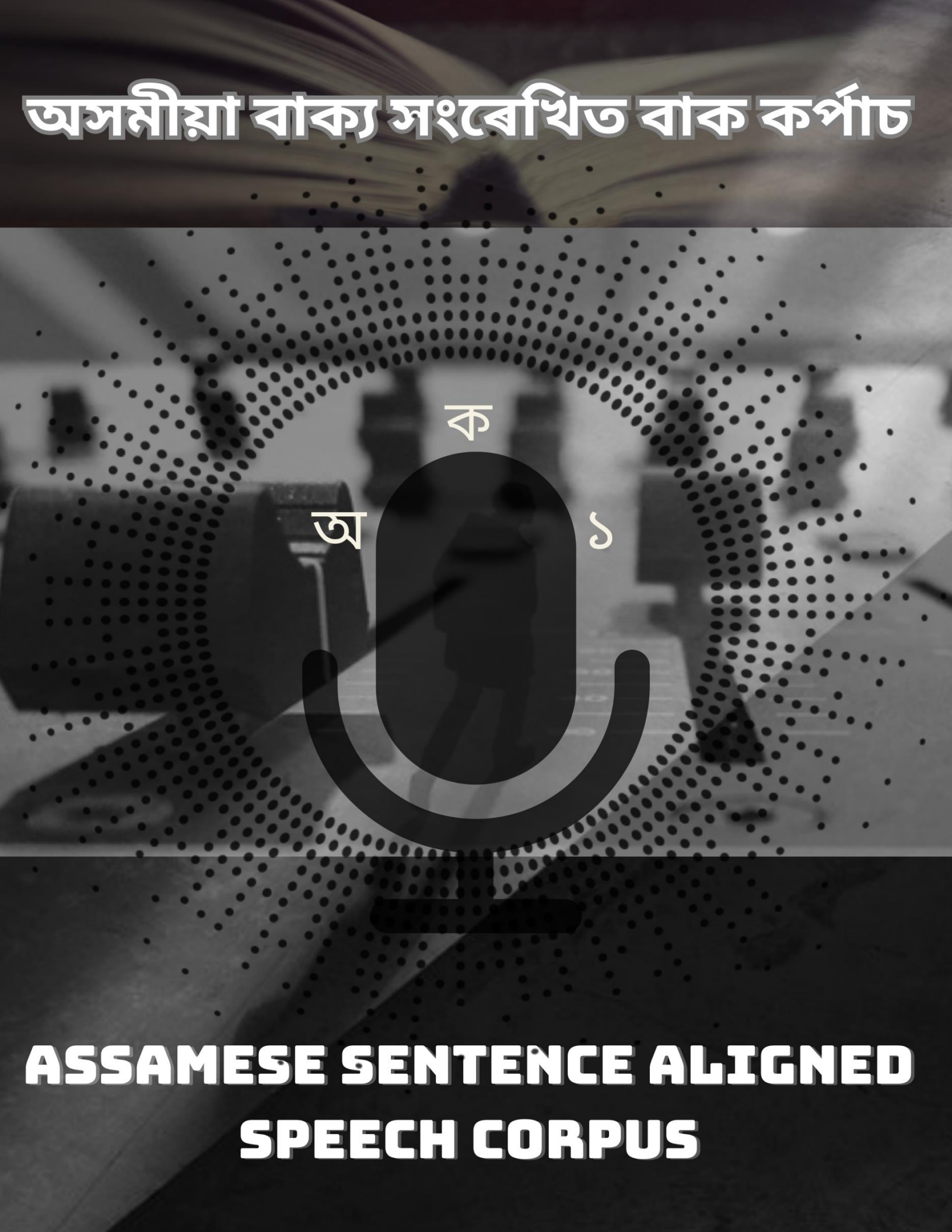
অসমীয়া বাক্য সংৰেখিত বাক কৰ্পাচ



ASSAMESE SENTENCE ALIGNED SPEECH CORPUS

ASSAMESE 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: Assamese Sentence Aligned Speech Corpus
Authors: Syeda Mustafiza Tamim, Priyanshe Adhyapak, Rajesha N., Manasa G., Srikanth D., Stephen Fernandes, Nithin S., Narayan Kumar Choudhary, Shailendra Mohan

ISBN: 978-81-19411-53-5
CIIL Publication No.: 1430

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: Syeda Mustafiza Tamim

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 ASSAMESE SPEECH ANNOTATION

Syeda Mustafiza Tamim, Priyanshe Adhyapak, Narayan Kumar Choudhary

2.1 OVERVIEW OF SENTENCE ALIGNED SPEECH CORPUS

Assamese Sentence Aligned Speech Corpus is created by annotating the speech data collected by LDC-IL. A detailed explanation of the Assamese Raw Speech Corpus will be available in the Assamese Speech Data Documentation (Ramamurthy, L. et. Al, 2019). LDC-IL Assamese Sentence Aligned Speech files contain an audio file and two textual layers in Assamese 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 shown below.

'Assamese_Female_16To20_Contemporary Text-T1_SP-0031_T1-0001.wav'

LDC-IL Sentence Aligned Speech corpus for Assamese 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 this set. Date format is kept as uttered by 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.

2.2 OBSERVATIONS

LDC-IL sentence level speech annotation strictly follows what the speaker pronounces rather than what is in the prompt sheet. The text has been written in the respective language script and the speech is transcribed as much as the script supports. Two or more different pronunciations can be uttered by the same or different speaker for the same word. Even if it is read from speech data, the dialect variation drastically influences the pronunciation. Therefore, speakers from different regions speak the same word in different ways. For eg. In Sivasagar and Jorhat region few speakers pronounce /x/ as /h/ and they have a tendency to omit or replace the /r/ pronunciation with a more of a like vowel /a/ /i/ depending on the word. The reading speed differs from person to person. Fast reading informants pose difficulty in annotation. Since news items contain sports news, it includes the informant reading all types of numbers. Speakers sometimes utter large digit numbers incorrectly like thousands or lakhs, decimal numbers, fractions etc. It is observed that speakers read Cricket score, Tennis score etc. in their own way and very few speakers read it properly. Most of the speakers show difficulty in pronouncing foreign names which frequently appear in sports news. Abbreviations and rarely used words interrupt the reader's fluency. All these factors contribute to the complexity in speech which makes it a rather difficult task. Since the dialect of the annotator can differ from that of the informant, the annotation process may need repetitive hearing in some cases. The annotation has to discard the data in particular places where the investigator has communicated with the informant. Some background noise like the sound of a bell, bus horn, other people's

conversation, baby crying etc. can be heard in the recording. Since this can be heard along with the voice of the informant, they have to be retained. This slows down the annotation process. Vocal noise of informants like coughing, sneezing etc. can also be observed.

2.2.1 PHONETIC ALTERNATION IN ASSAMESE SPEECH DATA

Reading speech has in-fluency like unwanted pauses, elongated syllables, word fragments, self-corrections, and repetitive words. When speakers notice what they utter then they suspend their speech and add, delete, or replace words they have already produced. Some fluctuated occurrences were detailed as follows:

a. Repetition of words

While reading, if the informant observes that the word has been pronounced incorrectly or not in an effective manner then the speaker normally repeats fragments of the word or sometimes the whole word or the phrase. Sometimes the speaker struggles to read the text and repeats when the content is a bit unfamiliar or there are many foreign words which are difficult to pronounce.

b. False start

False start is a common phenomenon in most of the speakers and some speakers it is frequent. Usually it is the replacement of the first word or syllable of the word but sometimes speakers start with some other letter as well.
E.g: *সোনকালে কাপোৰ ধোৱা, বৰষুণ আহিব এতিয়া।*
xunkaːle kaːpur ɖʱuwaː, bɔrɔxun aːhibɔ etijaː(Actual given sentence)
তুমি সোনকালে কাপোৰ ধোৱা, বৰষুণ আহিব এতিয়া।
tumi xunkaːle kaːpor ɖʱuwaː, bɔrɔxun aːhibɔ etijaː (Sentence with an extended letter or a false start)

c. Addition and Deletion

An extra vowel or a consonant or a syllable is added into a word. The letter which is existing in the word or different letter might be added into the word.
E.g.: *প্ৰকৃতি prɔkriti* (actual word) *প্ৰৰিকিতি prɔrikiti*

Deleting a vowel or a consonant or a syllable from a word is called deletion or elision. It is a common phenomenon when a natural language speaker speaks indistinctly.
E.g.: *উলিওৱা uliuwaː(actual word) উলিয়া ulijaː*
কৰিছো kɔrisu(actual word) কইছো kɔisu

d. Common phonetic variation

While pronouncing a word which has '/x/' in the initial, medial or at the end of the sentence the Assamese native speaker depending on the dialectal differences tends to invariably change it to voiced velar stop /h/.

e. Phone variation

It is the alternative pronunciation of the word and which does not affect the meaning. Both forms are correct and considered as the spelling variation of the same word.
E.g.: *যাই za:i > যায় za:j*
চিলনী silɔni: > চিলণি silɔni

f. Compound word splitting

Some compound words have been read in such a way that a pause is at the point of joining and that interrupt the natural flow of language.
E.g.: *পাকঘৰ pakgʰɔɹ > পাকঘৰ pak gʰɔɹ*

g. **Colloquial usage**

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

E.g.: *কৰিছা kɔɹisa (standard form) > কৰছা kɔɹsa (colloquial form)*

2.3 SUMMARY OF THE CORPUS

The total duration of Assamese Sentence Aligned Speech Corpus is 30:18:16 (hh:mm:ss) comprising 21,716 audio segments from 304 speakers. Figures 1, 2 and 3 show the distribution of the corpus with respect to gender, age and content type, respectively. Figures 4 and 5 show gender and age distributions for each content type respectively. Table 7 gives a break-up of the corpus in terms of recordings obtained from different kinds of texts and also other demographic details. Table 8 shows the age and gender-wise distribution of all the speakers.

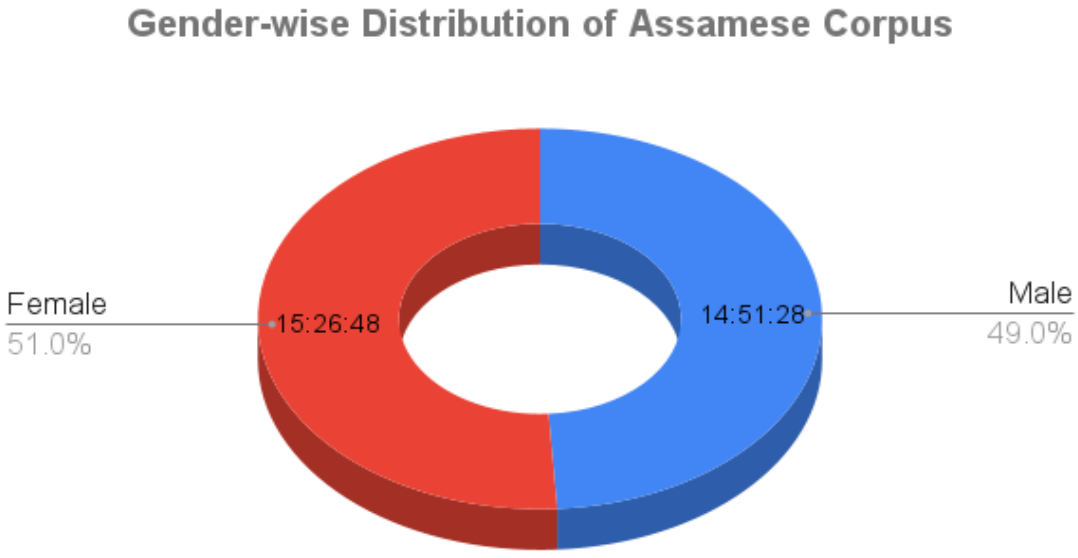


Figure 1: Gender-wise Distribution of Assamese Corpus

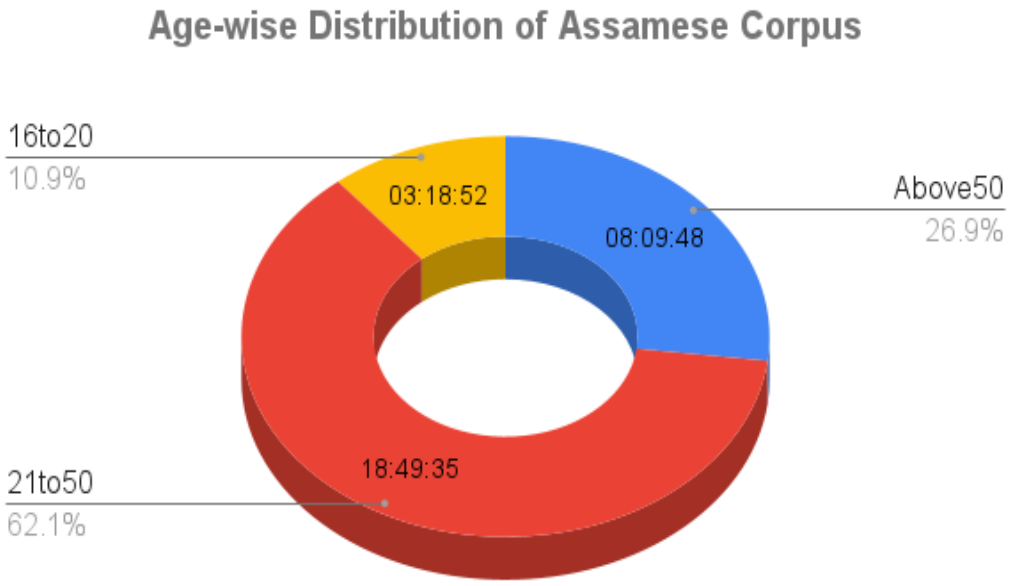


Figure 2: Age-wise Distribution of Assamese Corpus

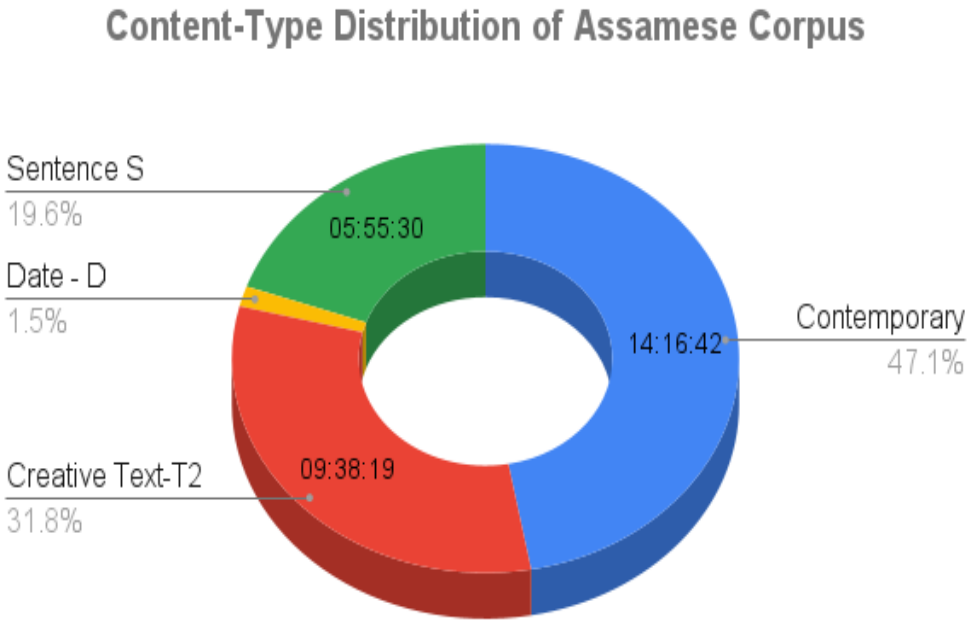


Figure 3: Content Type-wise Distribution of Assamese Corpus

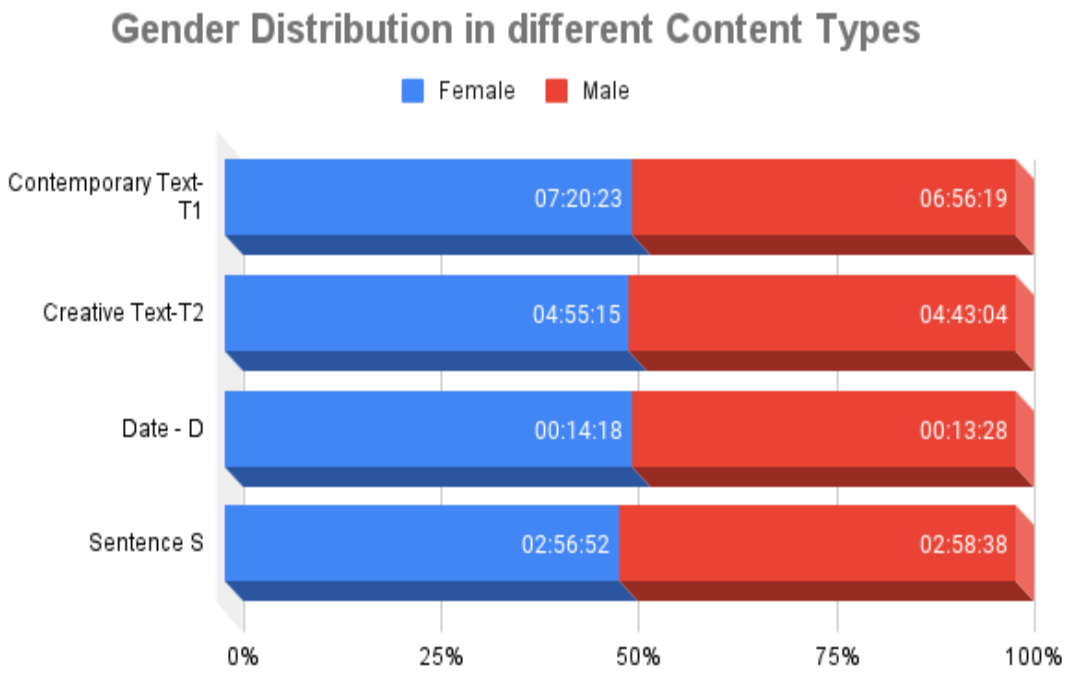


Figure 4: Gender Distribution in different Content Types of Assamese Corpus

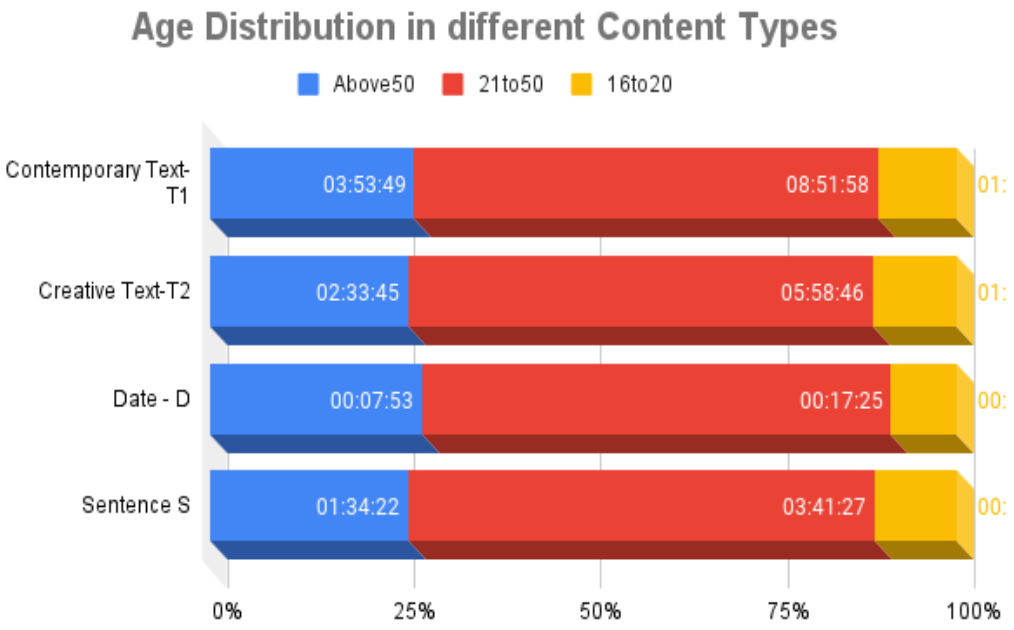


Figure 5: Gender Age Distribution in different Content Types of Assamese Corpus

2.3.1 DURATION OF ASSAMESE SENTENCE ALIGNED SPEECH DATA

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

Content Type	Gender	Age Group	Duration (hh:mm:ss.ms)		
Contemporary Text-T1	Female	16To20	00:45:51.532336	07:20:22.623416	14:16:41.300854
		21To50	04:36:23.356812		
		Above51	01:58:07.734268		
	Male	16To20	00:45:02.652097	06:56:18.677439	
		21To50	04:15:34.957759		
		Above51	01:55:41.067583		
Creative Text-T2	Female	16To20	00:32:26.605434	04:55:15.017747	09:38:18.644052
		21To50	03:03:29.021302		
		Above51	01:19:19.391011		
	Male	16To20	00:33:21.632878	04:43:03.626306	
		21To50	02:55:16.577616		
		Above51	01:14:25.415812		
Date-D	Female	16To20	00:01:19.494349	00:14:18.115419	00:27:44.900158
		21To50	00:09:03.717849		
		Above51	00:03:54.903220		
	Male	16To20	00:01:08.530930	00:13:26.784739	
		21To50	00:08:20.724579		
		Above51	00:03:57.529230		
Sentence-S	Female	16To20	00:19:52.059903	02:56:51.999937	05:55:30.902044
		21To50	01:49:53.549936		
		Above51	00:47:06.390098		
	Male	16To20	00:19:49.479959	02:58:38.902107	
		21To50	01:51:33.382227		
		Above51	00:47:16.039922		

Table 1: Representation of Assamese Sentence Aligned Speech Data Duration

2.4 SUMMARY OF SPEAKERS

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

Age Group	Female	Male	Total
16To20	16	16	32
21To50	97	94	191
Above51	41	40	81
Total	154	150	304

Table 2: Distribution of Speakers of Assamese 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, 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 Kumar Choudhary, Atreyee Sharma, Jahnobi Kalita, Samhita Bharadwaj, Taznin Hussain, Priyanshe Adhyapak, Syeda Mustafiza Tamim, Rajesha N., Manasa. G. 2021. A Gold Standard Assamese Raw Text Corpus. Central Institute of Indian Languages, Mysore.
5. Ramamoorthy L., Narayan Kumar Choudhary, Atreyee Sharma, Jahnobi Kalita, Samhita Bharadwaj, Plabita Bora, Priyanshe Adhyapak, Mustafiza Tamim, Rajesha N., Manasa G.. 2021. Assamese Raw Speech Corpus. Central Institute of Indian Languages, Mysore.